FEDERAL UNIVERSITY OF TECHNOLOGY - PARANA (UTFPR)

EDERSON CARVALHAR FERNANDES

UM SISTEMA DE PRODUÇÃO FLEXÍVEL E INTELIGENTE PARA PLANEJAMENTO DE PROCESSOS E OTIMIZAÇÃO DE DESEMPENHO EMPRESARIAL

CURITIBA
2023

EDERSON CARVALHAR FERNANDES

UM SISTEMA DE PRODUÇÃO FLEXÍVEL E INTELIGENTE PARA PLANEJAMENTO DE PROCESSOS E OTIMIZAÇÃO DE DESEMPENHO EMPRESARIAL

A Flexible and Intelligent Production System for Process Planning and Enterprise Performance Enhancement

Thesis presented to the Post-Graduation Program in Mechanical and Materials Engineering from the Federal University of Technology - Parana, as a partial requirement for the achievement of the title of Doctor of Engineering.

Principal Advisor: Milton Borsato, PhD
Co Advisor: Liam Brown, PhD

CURITIBA
2023

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Curitiba

EDERSON CARVALHAR FERNANDES

UM SISTEMA DE PRODUÇÃO FLEXÍVEL E INTELIGENTE PARA PLANEJAMENTO DE PROCESSOS E OTIMIZAÇÃO DE DESEMPENHO EMPRESARIAL

Trabalho de pesquisa de doutorado apresentado como requisito para obtenção do título de Doutor Em Engenharia da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Engenharia De Manufatura.

Data de aprovação: 30 de Março de 2023

Dr. Milton Borsato, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Alan Ryan, Doutorado - University Of Limerick

Dra. Fathima Fazleena Badurdeen, Doutorado - University Of Kentucky

Dra. Janaina Mascarenhas Hornos Da Costa, Doutorado - Universidade de São Paulo (Usp)

Dr. Leandro Magatao, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Liam Francis Brown, Doutorado - Technological University Of The Shannon

I dedicate this thesis to my dear and blessed mother,
Aparecida Carvalhar Martin, for her encouragement,
understanding, constant support, and infinite love,
every day of my life.

ACKNOWLEDGEMENTS

To God for all His protection and for guiding, enlightening, and giving me the courage and strength to never give up on fighting to make this dream come true.

To my beloved wife Silvia, who has always been by my side through all moments, with great patience and understanding. She always made me smile and brought positive energy, even in the most challenging situations. I wouldn't have been able to get this far without her.

To my loving mother, Aparecida, and my wonderful aunt Neribe, who is like a second mother to me. To my Nii-san Everton, my good brother, always present and available when I need him. Both form my precious family foundation. To my lovely nieces Sofia and Julia for bringing their contagious joy wherever they go. I love you all!

To my advisors Milton Borsato and Liam Brown, who helped and guided me with mastery towards the best paths. The guidance was essential in overcoming all the challenges in developing this thesis. I am very grateful for everything I have learned from you.

To my brave companions in the LAMIS research group at UTFPR: Sato, Jaqueline, Suewellyn, Felipe, Samuel, Lucas, Rafael, Kaio, Fabrão, Raffs, Juliane, Paulo, Ruan, and Gisele.

To my longtime friends, who are my chosen family: Estela Ávila and Eric Alexander Hoch. I know you always have my back!

To my good friend Jair Deon, who patiently taught me many advanced concepts of Javascript that were vital to achieving the results I needed on the web application's front end.

To my international friends whom I had the privilege of meeting during this long journey of my doctorate, outside or within Brazil, and who are today exceptional friends: Kathy, Jairus, Jacopo, Alan, Gerard, and Danika.

To the lecturers of PPGEM for the brilliant knowledge they could share. To the research colleagues at the Hartnett Enterprise Acceleration Centre at Limerick Institute of Technology (currently Technological University of Shannon), who brought me valuable knowledge during the time we could share: Barry, Kelly, Chris, and Jenna. To the TUS staff, especially Dr. Patrick Murray, who was always willing to help me with whatever I needed throughout the year I was developing my thesis in Ireland.

To the engineers at the partner company who made their time available to contribute to this research, especially Élcio Jesus, Luiz Romanel, and Jessé Tabor. To all the reviewers who generously offered their time and expertise and gave me excellent feedback to improve the submitted articles.

And to all those who directly or indirectly contributed to achieving a great accomplishment in my life.

You all live forever in my heart.

To the GOIIES (Government of Ireland International Education Scholarships) Program 2017/2018, and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) under the Academic Doctorate for Innovation Program (nº 157253/2019-8), for the funding provided during this period in both countries.

“All our dreams can come true, if we have the
courage to pursue them”.
(Walt Disney)

RESUMO

Muitas empresas têm dedicado esforços consideráveis para melhorar continuamente seus processos de fabricação, a fim de garantir sua competitividade e permanecer no mercado. O Mapeamento do Fluxo de Valor é uma ferramenta estratégica que permite a visualização do macro da produção para auxiliar no planejamento e na tomada de decisões. É um mapeamento de processos que considera o fluxo de trabalho de um produto desde a chegada da matéria-prima até o resultado entregue ao cliente. Apesar dos benefícios que essa ferramenta tem proporcionado às organizações, o tempo para seu desenvolvimento ainda é muito elevado, muitas vezes variando de semanas a meses, pois envolve entrada manual de dados. Este processo manual aumenta a probabilidade de erros na análise devido ao potencial inerente de erros humanos. É importante ressaltar que o desenvolvimento manual do VSM acarreta o risco de ocorrência de erros simples, que podem impactar na precisão e confiabilidade dos resultados obtidos.. A utilização da ciência de dados tem trazido diversas melhorias aos métodos e ferramentas das organizações. Assim, para obter contribuições para a engenharia, este estudo apresenta o desenvolvimento de um sistema web dinâmico para o mapeamento de processos no estado atual e futuro, utilizando análise de dados e aprendizado de máquina. Através da metodologia DSR (Design Science Research) que visa desenvolver soluções inovadoras para problemas complexos, é construído um sistema de virtualização Docker, por meio de uma abordagem ETL (Extração, Transformação e Carregamento), para o gerenciamento de seis contêineres com aplicativos que se comunicam entre si em um ambiente isolado, a fim de reunir todos os recursos necessários para a organização, visualização e atualização dinâmica do mapeamento atual, bem como simulações e previsões para a construção de mapeamentos futuros para auxiliar na tomada de decisão gerencial. Como resultado, o desempenho, segurança, interoperabilidade, generalidade, flexibilidade e eficiência do sistema desenvolvido são demonstrados por meio de sua aplicação e avaliação em um fluxo de produção de uma empresa de equipamentos agrícolas em Curitiba, Brasil. O mapeamento do estado atual foi realizado com os dados mais atuais fornecidos pela empresa, e foram previstas quatro alternativas para melhoria do processo, com a avaliação e aprovação dos engenheiros da empresa. O sistema de previsão proporciona maior agilidade e suporte à tomada de decisão. Combinar um método já consolidado na indústria de manufatura com tecnologias digitais contribui para identificar novos padrões, áreas críticas e relações complexas entre variáveis, fornecendo *insights* adicionais sobre o processo produtivo.

Palavras-chave: Produção, Monitoramento, Aprimoramento, Mapeamento de Fluxo de Valor, Docker, Sistema web

ABSTRACT

Many companies have been dedicating considerable efforts to continuously improving manufacturing processes to ensure their competitiveness and remain in the market. Value Stream Mapping is a strategic tool that enables the visualization of the macro of production to assist in planning and decision-making. It is a process mapping that considers the workflow of a product from the arrival of raw material to the result that is delivered to the customer. Despite the benefits this tool has provided to organizations, the time for its development is still very high, often ranging from weeks to months, as it involves manual data entry. This manual process increases the likelihood of errors in the analysis due to the inherent potential for human mistakes. It is important to note that the manual development of VSM carries the risk of simple errors occurring, which can impact the accuracy and reliability of the obtained results. Using data science has brought several improvements to the methods and tools of organizations. Then, to obtain contributions to engineering, this study presents the development of a dynamic web system for current and future state process mapping using data analysis and machine learning. Through the DSR (Design Science Research) methodology that aims to develop innovative solutions for complex problems, a Docker virtualization system is built, through an ETL (Extraction, Transformation, and Loading) approach, for the management of six containers with applications that communicate with each other in an isolated environment, in order to bring together all the necessary resources for the organization, visualization, and dynamic updating of the current mapping, as well as simulations and predictions in building future mappings to aid management decision-making. As a result, the performance, security, interoperability, generality, flexibility, and efficiency of the developed system are demonstrated through its application and evaluation in a production flow of an agricultural equipment company in Curitiba, Brazil. The current state mapping was carried out with the most current data provided by the company, and four alternatives for improving the process were predicted, with the evaluation and approval of the company's engineers. The prediction system provides greater agility and support for decision-making. Combining a method already consolidated in the manufacturing industry with digital technologies contributes to identifying new patterns, critical areas, and complex relationships between variables, providing additional insights into the production process.

Keywords: Production, Monitoring, Enhancement, Value Stream Mapping, Docker, Web System

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ACID	Atomicity, Consistency, Isolation, Durability
AGV	Automated Guided Vehicle
AI	Artificial Intelligence
AM	Additive Manufacturing
API	Application Programming Interface
AR	Augmented Reality
BI	Business Intelligence
BPM	Business Process Management
C/O	Changeover Time
C/T	Cycle Time
COPICS	Construction Operation Building Information Exchange for Product and Systems Data
CPS	Cyber Physical System
CQL	Cassandra Query Language
CSRF	Cross-Site Request Forgery
CSS	Cascading Style Sheet
CSV	Comma-Separated Values
DA	Data Analytics
DM	Data Mining
DS	Data Science
DSR	Design Science Research
DSRM	Design Science Research Methodology
ECS	Elastic Container Service
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load
FP	Functional Programming
FPY	First Pass Yield

GKE	Google Container Engine
HTML	Hyper Text Markup Language
I4.0	Industry 4.0
ID	Identification
IOT	Internet of Things
JIT	Just in Time
JS	JavaScript
LM	Lean Manufacturing
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multilayer Perceptron
MMS	Manufacturing Message Specification
MQL	MongoDB Query Language
MQTT	Message Queuing Telemetry Transport
MSE	Mean Squared Error
MVT	Model-View-Template
OEE	Overall Equipment Effectiveness
OOP	Object-Oriented Programming
ORM	Object-Relational Mapping
PM	Process Mining
PPC	Process Planning and Control
Q&A	Questions and Answers
QOS	Quality of Service
RAM	Random Access Memory
RMSE	Root Mean Squared Error
SMED	Single-Minute Exchange of Die
SQL	Structured Query Language
SSL	Secure Sockets Layer
SVM	Support Vector Machines
SVR	Support Vector Regression

TCP/IP	Transmission Control Protocol/ Internet Protocol
TLS	Transport Sockets Layer
TPM	Total Productive Maintenance
URL	Uniform Resource Locator
VM	Virtual Machine
VS	Virtual Simulation
VSCODE	Visual Studio Code
VSM	Value Stream Mapping

TABLE OF CONTENTS

1 INTRODUCTION

Advancements in technology over the last decades have been followed by many challenges of industrial competitiveness at different levels, mainly when decision-making is needed (MEYER *et al.*, 2014). Challenges have increased since this rapid evolution has brought many radical changes in market requirements, and customer needs as a priority have a significant impact on industrial productivity (RAWAT; GUPTA; JUNEJA, 2018).

Reid & Sanders (2019) state that a manufacturing industry must manage the activities, resources, and materials used to create products or services. A transformation process consists of inputs, usually raw materials, that undergo a transformation process to create products (outputs).

A service organization, on the other hand, responds to the demands of consumers and embraces their needs through the service delivery process. A service organization only sometimes use inputs or produces something the consumer will use.

An *operating system* is one of many activities that transform inputs into helpful output using a transformation process. With this, it is possible to classify manufacturing and service systems.

Operations management deals with the activities, decisions, and responsibilities of managing the resources dedicated to producing and delivering products and services. As such, operations managers are the people responsible for overseeing and managing the resources that make up the operations functions. Operations management can significantly contribute to the success of an organization, using available resources effectively to produce products and services and thus obtain customer satisfaction (KUMAR, 2022).

The day-to-day management of operations and services is complex, as the function involves numerous processes. Among them are monitoring costs, integrating sectors, ensuring the services' excellence, worrying about the customer experience, and making data-based decisions.

In this scenario, it is common for entrepreneurs to face a series of challenges to accurately understand the variability of their production and achieve the expected results in their company. However, the advancement of digital technologies and the beginning of I4.0 (Industry 4.0), despite bringing many advantages, also brought many new challenges for companies that seek to innovate and reach new levels of competitiveness.

Martinuzzi et al. (2018) highlight two significant challenges in the industry. The first concerns the accelerating race for innovation to gain a competitive advantage in a global environment, and the second relates to public trust in the industry through innovations.

The I4.0 was brought about by the emergence, advancement, and convergence of a series of technologies that allowed computer science to standardize human problem-solving procedures and transfer them to computers to invent new and efficient solutions, as well as a course of action (PASCHEK; LUMINOSU; DRAGHICI, 2017). The term "Industry 4.0" originated from a German high-tech program following the first Industrial Revolution "Mechanization", the second "Mass Production", and the third "Automation" (SCHMIDT *et al.*, 2015).

Even companies that have not yet implemented digital technology to improve their processes and those that already apply digital resources are looking for metrics, methods, techniques, and tools to control their current processes and monitor new enhancements.

Various tools and methodologies promote various ways to achieve high productivity and innovation, differing in the technical and engineering domain from which they originated. Among existing ones, Lean is the most widely used in manufacturing engineering as it has been proven to be the most effective waste elimination (GHOLAMI *et al.*, 2021).

According to Garre et al. (2017), LM (Lean Manufacturing) has created many improvements in production efficiency for all industry sectors. Even so, many challenges will still be overcome in today's fast-changing market.

LM is "a set of principles, philosophies and business processes to enable its implementations, which is widely known and implemented since 1960" (ROSE *et al.*, 2011). LM aims to maximize efficiency and minimize waste in all stages of the production process. This approach is based on continuous improvement, elimination of waste, and involving employees in the search for problem solutions.

According to Permin et al. (2016), manufacturers always need to adjust and adapt at all levels of value creation. LM has successfully challenged mass production, where all not added value is waste. For greater flexibility of production systems, either a link or integration between lean production and I4.0 tools is required for better production and planning (MRUGALSKA; WYRWICKA, 2017).

Seifullina et al. (2018) revealed a need for more coherent and conceptual models to guide Lean implementation in the industry. Kolberg & Zühlke (2015) stated that there are existing combinations of Lean Production and Automation, also called Lean Automation. However, there are only a few studies for applying I4.0 solutions. Identified examples of possible combinations are:

Smart Operator: Equipped with smartwatches and AR (Augmented Reality) goggles, employees may support JIT (Just in Time) proceedings.

Smart Product: They could collect process data for analysis during and after production through Kanban information.

Smart Machine: CPS (Cyber Physical System), systems monitored by computers and can support Lean Poka Yoke.

Smart Planner: Workstation-integrated CPS can transform traditional Kanban systems into dynamic productions, automatically adapting to current production schedules.

I4.0 combines the strengths of traditional industries with cutting-edge internet technologies, which encompass the development of intelligent production and manufacturing processes, new capabilities through communication between the physical and virtual environment, and interoperability among computer systems (ZENG; YIN, 2017).

Increasingly, there are new global and multidisciplinary developments related to AI (Artificial Intelligence) and DS (Data Science). According to Neupane & Echaiz (2019), the number of AI articles in Scopus has increased seven times since 1996; Organizations that adopt AI provided the most value in their industry, and AI articles became 1.5x more positive from 2016 to 2018.

Thus, based on this context, the search for process improvements has been intensified both by the need to improve the competitive position in the market and by the opportunity to use new technologies to improve quality and productivity.

This research started with investigating unfilled gaps and opportunities for process improvement applications aided by specific AI and DS techniques. Process improvement often depends on the knowledge and experience of the field engineer.

Considering the massive amount of new research and publications on AI and DS, a preliminary and conceptual analysis was performed with the leading techniques related to process improvement in both areas to narrow down the choice of two specific techniques and focus on a deeper study of what has been developed on these topics.

In the field of AI, ML (Machine Learning) has been chosen as one of the fastest-growing technical fields that have progressed massively, according to Jordan and Mitchell (2015), and it has been increasingly used in different applications such as robotics, neuroscience research, and autonomous vehicle control.

In the field of DS, among all its applications, PM (Process Mining) has been chosen, despite being a relatively new research discipline which has started in 2005. It has been beneficial for the organization's processes. PM works with three classes: process discovery, conformance checking, and organizational mining, and its criteria are based on fitness, simplicity, precision, and generalization (POURMASOUMI; BAGHERI, 2017).

So, a bibliographic and systemic analysis of ML and PM applied to Process Optimization (FERNANDES *et al.*, 2019) was carried out. Among these studies, there were problems raised by the authors related to process management, complex models, and decision-making support. For the problems raised and their solutions, some unfilled gaps can be identified, such as (i) Additional Functionalities and Programming, in order to improve the previously tested and applied systems; (ii) Adaptations of Intelligent Methods, through various combinations, mainly using traditional methods such as LM Methods, Total Quality Management, among others; (iii) Environmental Adaptability, to evaluate complete intelligent systems for different areas; (iv) Smart Validation Data Set, due to lack of sufficient data to accurately evaluate a DS system. An innovative dataset can quickly generate data to simulate the real ones.

Therefore, in analyzing unfilled gaps (i) and (ii), new research was carried out to identify a method conducive to developing an intelligent adaptation that benefits manufacturing engineering. Lai et al. (2019) analyzed the impact of I4.0 in conjunction with LM and identified several positive aspects and examples of I4.0 with different types of waste targeted in LM.

Valamede & Akkari (2020) proposed integrating LM tools and 4.0 technologies through perspectives of the industrial field in the digital era. Among the analyses carried out, they identified 25 points of synergy between LM tools: JIT, Kanban, Poka-Yoke, VSM (Value Stream Mapping), Kaizen, and TPM (Total Productive Maintenance) with digital technologies: Big Data Analytics, AGV (Automated Guided Vehicle), cloud computing, AM (additive manufacturing), cybersecurity, VS (virtual simulation), and AR.

Although the importance of each Lean method is recognized, the present research primarily focuses on VSM because it is widely used and flexible to be used alone or in conjunction with other methods to reduce production lead time and make the manufacturing process more efficient, reducing waste in the manufacturing space, inventory, and cycle times. It is considered a critical process improvement method used to analyse, design, and manage the flow of materials and the information required to deliver a product to a customer. Furthermore, since VSM requires future planning mappings for its development, using prediction algorithms becomes appropriate and predominant for enhancing the method.

VSM was developed in Japan in the 1950s as part of the Toyota Production System, which sought to improve vehicle production with maximum efficiency and quality. In the 1990s, VSM began to gain popularity in other industrial sectors and countries, where the concept was adopted and incorporated into continuous improvement practices in companies (LUGERT; BATZ; WINKLER, 2018).

In addition, VSM is used to create the current and future states of the material and information flow of processes, mainly for identification, controlling, and improving value-added and non-value-added activities related to the customer's perceived rating of the value delivered (LACERDA; XAMBRE; ALVELOS, 2016; YUVAMITRA; LEE; DONG, 2017).

Some studies have related analyses and studies with the analysis of advantages of using VSM with digital technologies and analysis of alternatives on how to perform this combination. Horsthofer-Rauch et al. (2022) evaluated the potential of combining PM and VSM. They identified that PM contributes to VSM, especially in analyzing and modeling the production process, significantly reducing the efficiency challenges in conventional VSM. They did not effectively combine PM with VSM, indicating this development for future studies.

Valamede & Akkari (2020) identified that VSM is aligned with Big Data Analytics, Virtual Simulation, and Cloud Computing. Through sensors, or intelligent products and machines, equipment performance data and object location are automatically collected, allowing for constant of VSM 4.0, updates and accessing data that would feed the mappings by linking VSM to cloud computing or virtual machines.

Sultan & Khodabandehloo (2020) analyzed the improvements that VSM and an internal logistics system can obtain with digitization. In the study, they analyzed the challenges in VSM and the opportunities for improvement in VSM to increase productivity and then conducted an empirical study in a company. The study concludes that the combination of VSM with simulations and real-time data collection is vital to increase the flexibility and responsiveness of VSM.

Through related works and research gaps presented by the literature, it is identified that many studies relate VSM with digitalization, highlighting the potential that could be achieved with digital adaptations. However, there was a lack of an application for functional VSM for industrial production companies, both for the digital display of VSM and for developing VSM with intelligent predictions.

The traditional use of VSM is a process that requires a considerable amount of time for detailed observation, careful annotation of parameters, meticulous organization of data, and periodic meetings for decision-making. Additionally, it is necessary to align the schedule of all involved parties to ensure that the mapping development is done according to expectations. However, once bottlenecks and issues are identified, the challenge arises to develop a mapping capable of producing effective results. Testing predefined improvement alternatives in future state mapping is not considered efficient by disrupting process operations. Therefore, there are challenges in organizing and effectively analyzing data collected over long periods and challenges in testing new improvement alternatives and process changes without having a clear view of how they will affect the value stream, as there is a risk of compromising operating time during these tests, improper implementation, or undesired effects on results.

From the research opportunity and the challenges presented, the established research question is:

How can a flexible, digital, and dynamic application for Value Stream Mapping provide accurate representations of the current state of production and generate predictions for the future state to test alternatives, effectively provide insights for new process improvements, and contribute to decision-making?

The present work has been developed within the Product Life Cycle Research Group, which established two development programs for academic research: Intelligent Manufacturing and Sustainable Manufacturing. Such programs aim to develop projects that can contribute to manufacturing companies to improve their competitiveness in the market. This thesis is an outcome of the Intelligent Manufacturing program, which aims to contribute to the demand for the creation of planning and design steering factors based on knowledge.

1.1 Research Objectives

1.1.1 General Objective

The main objective of the research is to develop and apply a flexible and intelligent application for creating the current state mapping and predicting the future state mapping in a production flow scenario for an integrated fuel tank supply and assembly process.

1.1.2 Specific Objectives

The following specific goals can be identified as stepping stones for achieving the main goal:

- Identifying the application context of the solution in the manufacturing industry.
- Investigating different methods, algorithms, and platforms.
- Selecting a layered software architecture for the VSM application.
- Structuring the specifications and requirements of the application.
- Developing the back-end and front-end algorithms for Current State Mapping.
- Developing the back-end and front-end algorithms for Future State Mapping.
- Testing and debugging prediction functions for creating future mappings.
- Demonstrating the system application in a manufacturing company through a proof of concept.
- Evaluating the system for its efficiency, reliability, and applicability.

1.2 Justification

This research project aims to develop an intelligent web-based data control and prediction system to enable precise and dynamic control of VSM. Therefore, the project does not propose a completely new and improved VSM but rather an adaptation of VSM that can be used to predict and demonstrate better process alternatives.

This project does not intend to create a process redesign tool, meaning it does not involve a complete restructuring or redesign of the process. Instead, it aims to develop a decision support tool that provides valuable insights for identifying alternatives and improvements to the process.

In this context, this project seizes a research opportunity for the development of VSM through the use of digital and intelligent resources, which facilitates the organization of data, the manipulation between the connections of all information, the interpretation of the production and information flow, the identification of bottlenecks, and the future projection for identifying improvements in the analyzed processes.

Among the research motivation factors, the absence of digital developments to improve existing methods is highlighted since they are already consolidated methods in organizations and have been providing excellent results for many decades. Many organizations do not seek new solutions but rather the application of digital technologies to enhance the results that these methods can already achieve. According to Savastano et al. (2019), implementing digital technologies brings numerous benefits to manufacturing, especially its remarkable capacity to provoke paradigm shifts.

Many manufacturing companies are unaware of the potential, implications, and challenges of using AI and DS methods in conjunction with manufacturing methods.

The manufacturing industry faces numerous challenges with the vast amount of information and rapid-changing customer requirements. Process management and resource allocation issues (ZENG; YIN, 2017) have been growing and creating difficulties for companies not investing in intelligent production and manufacturing processes (LONG; ZEILER; BERTSCHE, 2016). Most organizations store their data in one or more database management systems but cannot analyze and make valuable insights from that data (SAVICKAS; VASILECAS, 2018). Failure to properly process or develop insufficient data mapping can cause time and costly resources to be wasted (ER et al., 2018; GENGA et al., 2018; LATORRE-BIEL et al., 2018; LI et al., 2017).

Decision strategies in dynamic environments do not consistently achieve desired outcomes or improvements, and this has been a considerable problem outlined by several authors (LATORRE-BIEL et al., 2014; LEE et al., 2018; MEYER et al., 2014; WYNN et al., 2017). Deviations from standard operating procedures, uncertainties, and delays (GERHARDT; VALIATI; CANTO DOS SANTOS, 2018) are just a few examples of unforeseen events that can occur without proper supervision of what is happening at a current time and what the likelihood is that any of these problems will occur in the future. Logistics (BECKER; INTOYOAD, 2017; WANG *et al.*, 2014), supply chain (BLACKHURST *et al.*, 2018) and distributed business processes (BORKOWSKI et al., 2019; EVERMANN; REHSE; FETTKE, 2017; KARRAY; CHEBEL-MORELLO; ZERHOUNI, 2014; LATORRE-BIEL; JIMÉNEZ-MACÍAS; PÉREZ-PARTE, 2014; LEE et al., 2018; UMER et al., 2017) were applications identified by authors that have been associated with decision support since the execution of actual processes can be highly unpredictable and vulnerable to disruptive events (SAVICKAS; VASILECAS, 2018), due to lack of maintenance (KARRAY; CHEBEL-MORELLO; ZERHOUNI, 2014) or human error (PARK *et al.*, 2018).

Thus, the combination of AI, DS, and VSM methods will provide many advantages, mainly due to the following:

1) There is a lack of integration between LM methods and I4.0 tools (DOMBROWSKI; RICHTER; KRENKEL, 2017; KOLBERG; ZÜHLKE, 2015; MAYR *et al.*, 2018; MRUGALSKA; WYRWICKA, 2017);

2) Disbelief in the implementation of lean thinking by AI self-learn processes (BAUER *et al.*, 2018; KOLBERG; ZÜHLKE, 2015; MAYR *et al.*, 2018; SARTAL *et al.*, 2017);

3) Automated information is not widespread for continuous improvement (KOLBERG; ZÜHLKE, 2015; MRUGALSKA; WYRWICKA, 2017);

4) Many manufacturing companies face many challenges and little flexibility daily (ALMANEI; SALONITIS; XU, 2017; RAUCH; DALLASEGA; MATT, 2015; SALONITIS; TSINOPOULOS, 2016; SUSILAWATI *et al.*, 2015);

5) Many companies have faced barriers to successful LM implementations (ALMANEI; SALONITIS; XU, 2017; RAUCH; DALLASEGA; MATT, 2015; SALONITIS; TSINOPOULOS, 2016; SUSILAWATI *et al.*, 2015).

Implementing any LM method is not a straightforward process, but unsuccessful implementation can significantly impact the organization's resources and affect employees and their confidence in the lean philosophy (SALONITIS; TSINOPOULOS, 2016).

Applying an intelligent system can be highly beneficial for manufacturing industries so that engineers cannot spend time doing routine activities to address more significant challenges.

The following section describes the research delimitation of the thesis project.

1.3 Delimitation

The data collection for the project is not carried out in real-time through IoT (Internet of Things) technology, i.e., the project development starts from an existing digital database.

1.4 Structure of the Thesis

The document is structured in five chapters, including the presented Introduction chapter. This chapter introduces the topics related to this research, how recent promising technologies adapt and enhance the traditional manufacturing system, how the research opportunity has been identified, their application perspectives, research objectives, and the justification for this study.

Chapter 2 presents the theoretical background of the main topics of interest for a complete understanding and development of the thesis project. It covers additional vital aspects of the manufacturing system, as well as its advances and challenges. A literature review on process planning and improvement and digitalization in manufacturing includes information on the leading related technologies and tools, flow-based modeling, MQTT (Message Queuing Telemetry Transport), web frameworks, BI (BI) tools, and virtualization systems.

Chapter 3 demonstrates the methodology aspects and methodological procedures applied to the development of this research, such as DSR (Design Science Research), with the steps taken to develop this project.

Chapter 4 covers the project development, with all the details about demonstrating the results and the evaluation.

Chapter 5 presents the conclusions about the research, summarizes the entire project’s development, details the main challenges encountered in achieving the research objective, and describes the primary suggestions and recommendations for future developments.

2 THEORETICAL BACKGROUND

In this chapter, the literature review aims to provide the main concepts about process planning and enhancement in industrial companies, production data analysis, LM methods, highlighting the VSM, reviews about DS and ML developments and techniques, web frameworks, flow-based programming, message brokers, BI, and container virtualization.

2.1 Process Planning and Enhancement

Planning is a ubiquitous human activity, and Process Planning is one of the most crucial manufacturing steps, which is a plan that determines all needed operations or processes to produce any tangible component. It decides how to manufacture a work part and finds all parameters and methods to transform raw material into a product (KARIM; TAI TIONG, 2019).

For several years, process planning was done manually, using route sheets and operations sheets to check and track the movement of parts produced through a factory. Route sheets provide all information to operators from one workstation to another, and operations sheets are a document that encompasses all the details of operations to complete a product.

Over time, many process management, control, and planning methodologies have been created to contribute to and improve industrial development, and they have been widely used in manufacturing.

LM is a management methodology oriented towards continuous improvement of processes and aims to eliminate waste, improve the efficiency of production processes, and consequently, unify the stages that add value to the product (ROHANI; ZAHRAEE, 2015); (PALANGE; DHATRAK, 2021). LM was formalized through five fundamental principles:

a) Identify the value: the first step is to identify what the customer considers value. This value derives from the customer's need, and companies must satisfy it by charging the price the customer is willing to pay;

b) Identify value chain: identify all stages of the production process and analyze how they relate;

c) Establish continuous flow: seek to eliminate waste and create continuous flow without interruptions. The immediate effect of this creation can be seen in the reduction of product conception times, order processing, and inventories;

d) Pull production: produce the product only when the customer demands it; and

e) Achieve perfection: focus all of the company's efforts on achieving perfection by eliminating waste and creating value, that is, by applying continuous improvement.

These fundamental principles can be fulfilled using tools and techniques associated with Lean. Among these tools and techniques are: VSM, JIT, Pull Production, Continuous Flow, Kanban System, Poka-Yoke, Standardized Work, OEE (Overall Equipment Effectiveness), TPM, SMED (Single-Minute Exchange of Die), Kaizen, 5S, Jidoka (Autonomation), Heijunka (Production Leveling), and Andon (Problem Signaling) (LEKSIC; STEFANIC; VEZA, 2020).

The basis of the lean thinking concept is mainly the elimination of waste, which refers to all production elements that increase costs without adding value. The production system has eight commonly accepted wastes: defects, overproduction, waiting, non-talented person, transport, inventory, motion, and extra processing. (RAMANI; KSD, 2021) .

All tools and techniques mentioned are used individually or in combination to eliminate these eight wastes. However, the focus of this thesis will be on VSM, as the development of this thesis is directed toward adapting this method.

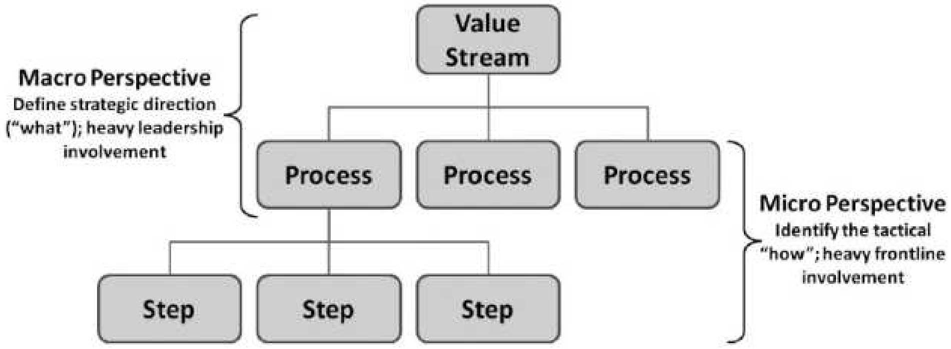
The VSM is a tool that allows the visualization of the entire production process, from receiving the customer order to delivering the final product, including support processes. This technique identifies activities that add value and those that do not add value to the process, thus helping to eliminate waste. In a broader sense, a value stream is the sequence of activities necessary to design, produce, and deliver a product or service to the customer, including information and material flows (DINIS-CARVALHO *et al.*, 2019; MARTIN; OSTERLING, 2013).

From a macro perspective, VSM provides effective means to establish strategic directions in the improvement process, while at a process-level mapping, it allows people to design tactical improvements (Figure 1).

VSM brings a highly visual projection of the complete cycle, as it demonstrates how work progresses from a request to the fulfillment of that request. This cyclical view provides a powerful means of visualizing the entire work system for delivering value to the customer. As shown in Figure 2, the VSM cycle includes three components: information flow, workflow or process, and the summarized timeline.

Thus, VSM explores an organizational understanding of work systems that deliver value and support the delivery of value to customers. This method significantly contributes to better decision-making in work design. Creating representations that assist different professionals at different levels in organizations, which all stakeholders can understand, brings excellent advantages to the visual establishment of a work system restructuring that can bring additional value, become faster, with lower cost, and in a safer work environment. VSM uses two mappings, the current state mapping and the future state mapping. The current state helps to capture the snapshot of how things are currently done and also to visualize potential areas for improvement. The future state consists of a lean transformation to improve the bottlenecks identified in the previous map (TYAGI *et al.*, 2015).

Figure 1 - Granularity of Work

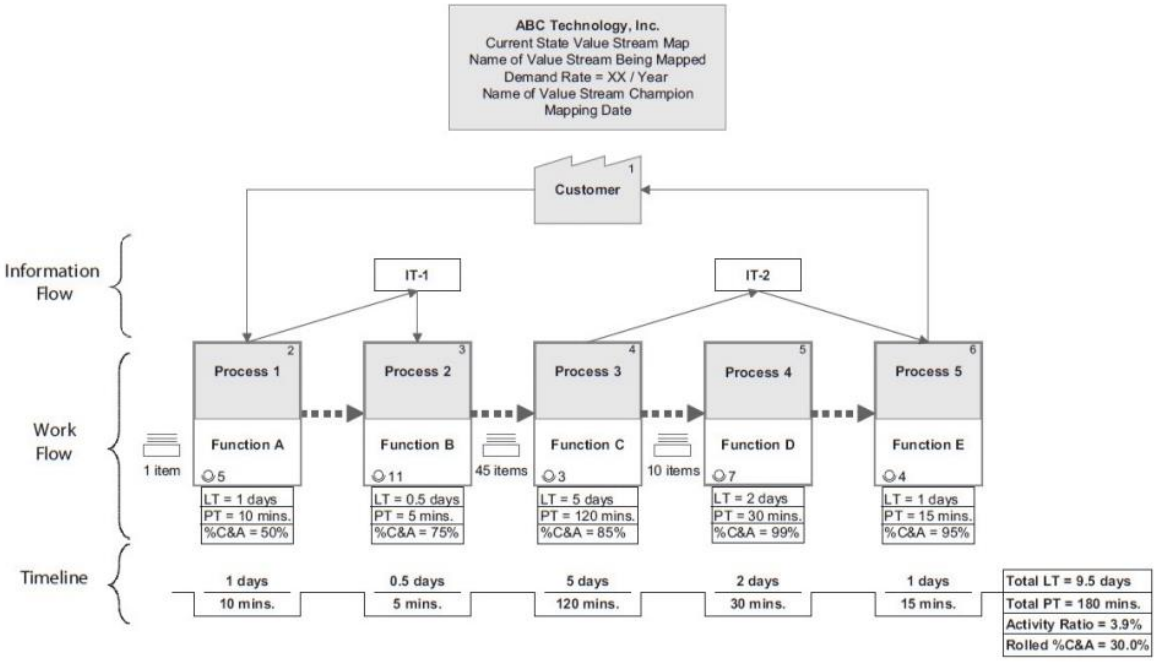

Value Stream
Process
Process
Process
Step
Step
Step
Macro Perspective
Define strategic direction ("what"); heavy leadership involvement
Micro Perspective
Identify the tactical "how"; heavy frontline involvement

Source: Martin & Osterling (2013)

VSM is a highly iterative tool that needs to be frequently consulted and updated with changes in the value flow. Traditionally, this method is physically applied with post-its in strategic locations where periodic meetings can occur to discuss value flow performance and how to make new process improvements (LANGSTRAND, 2016; MARTIN & OSTERLING, 2013).

The recommended development time for VSM is at least four weeks of planning before mapping. However, the development time for both mappings depends on the process's complexity and the amount of data that needs to be collected and analyzed. The present state mapping can take from a few weeks to several months, and the future state mapping can also last this same amount of time, depending on the complexity of the process and the number of proposed changes.

Figure 2 - Current state VSM example



Source: Martin & Osterling (2013)

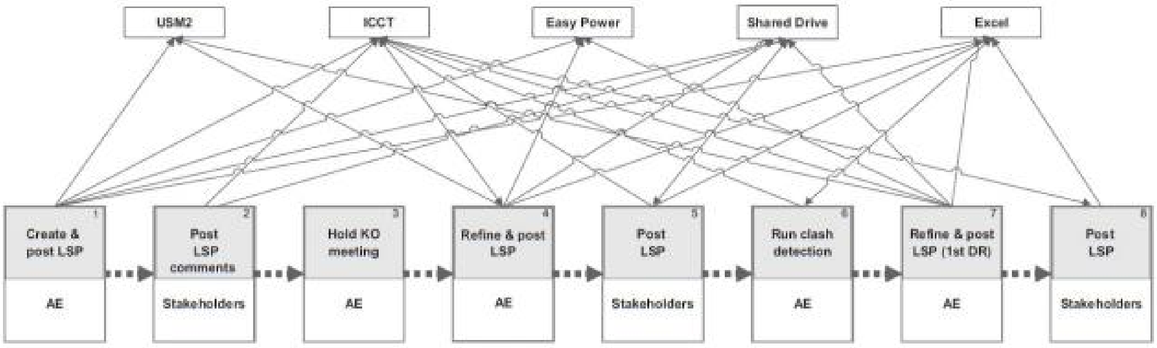
A critical step in developing the present state mapping is to physically walk the value stream, observe the work, talk to operators in their environment, and learn about some challenges and obstacles that may be occurring. It is recommended that the mapping team walks the value stream at least twice on the same day so that learning and identifying the values obtained are performed with attention and precision. This procedure can be done twice from start to finish, or it can be performed for the second time by analyzing the process from end to beginning (LACERDA; XAMBRE; ALVELOS, 2016; LIU; YANG; XIN, 2020).

Thus, this documentation of the current state mapping can be divided into five activities: walk the value stream, note everything on the mapping, walk the value stream for the second time, add more details, and organize and review this mapping. The variables included in the mapping will also depend on the complexity of the process. However, most VSM maps generally have variables such as cycle time, processing time, waiting time, inventory, rework, movement, lead time, and efficiency. These variables help identify bottlenecks, waiting times, inefficiencies, other process problems, and improvement opportunities (MARTIN; OSTERLING, 2013; RAMANI; KSD, 2021; YUVAMITRA; LEE; DONG, 2017).

Lead time is also an essential metric in the current VSM map, as it shows the total time from when work becomes available to an operator or work team until it is completed. Lead time can be indicated in hours, days, weeks, or months.

The information flow (Figure 3) is as vital in VSM as the process flow, as both need to be related. Understanding how information flow affects value flow is critical to decision-making for potential changes. The team needs to identify the systems or applications exist in the analyzed process and how they are used.

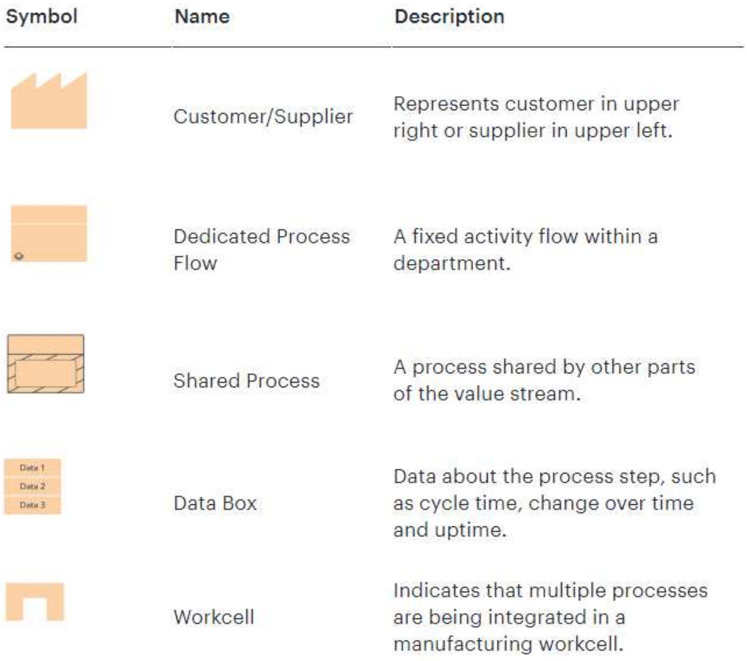
Figure 3 - Information flow along with process flow



Source: Martin & Osterling (2013)

The current state mapping should be created as a discovery activity, always seeking to represent the truth for both performance and significant barriers in the process. Value mapping is mainly used as a storyboard that clarifies how the work is done and reveals problems. In addition to the variables described in the mapping, some symbols can assist in organizing the mapping. These symbols are used in four categories: process, material, information, and general, which can be complicated for those who do not have adequate experience to interpret each one correctly, but some marks are very intuitive, such as the truck icon, commonly used for external shipments, and the glasses symbol for something that should be observed. In Figure 4, Figure 5, Figure 6, and Figure 7, the process, material, information, and general signs are shown, respectively.

Figure 4 - Process Symbols

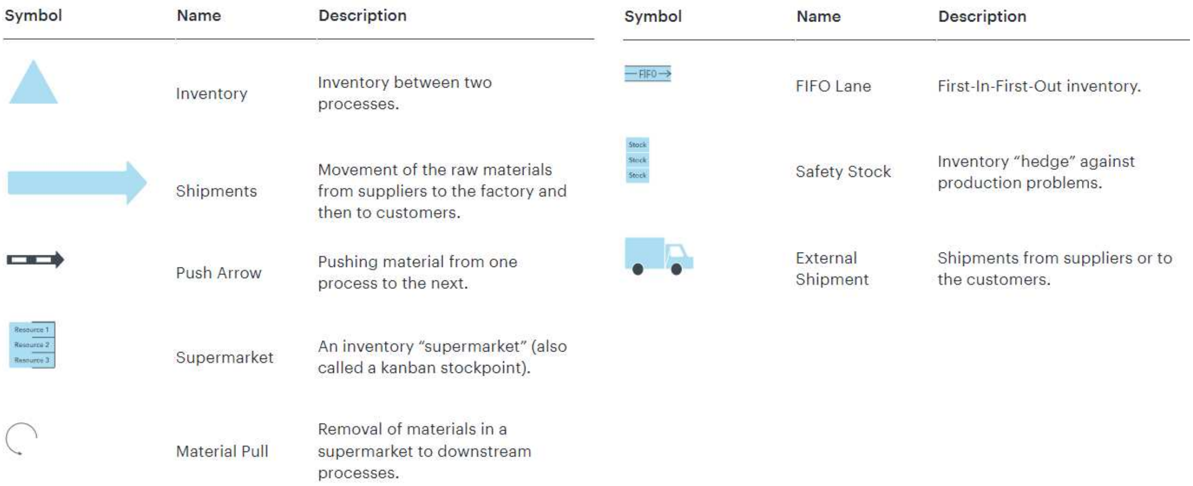


Symbol	Name	Description
	Customer/Supplier	Represents customer in upper right or supplier in upper left.
	Dedicated Process Flow	A fixed activity flow within a department.
	Shared Process	A process shared by other parts of the value stream.
Data 1 Data 2 Data 3	Data Box	Data about the process step, such as cycle time, change over time and uptime.
	Workcell	Indicates that multiple processes are being integrated in a manufacturing workcell.

Source: Lucidchart (2018)

Once the team thoroughly understands the current state, future state mapping should be performed. No correct mapping exists to be created, as developments are relative to the ideas and decisions made through team analysis. Therefore, this mapping may have multiple alternative constructions. Two teams can create ideas to improve the value stream, and both future mappings may work (LOBO; CALADO; CONCEIÇÃO, 2018).

Figure 5 - Material Symbols

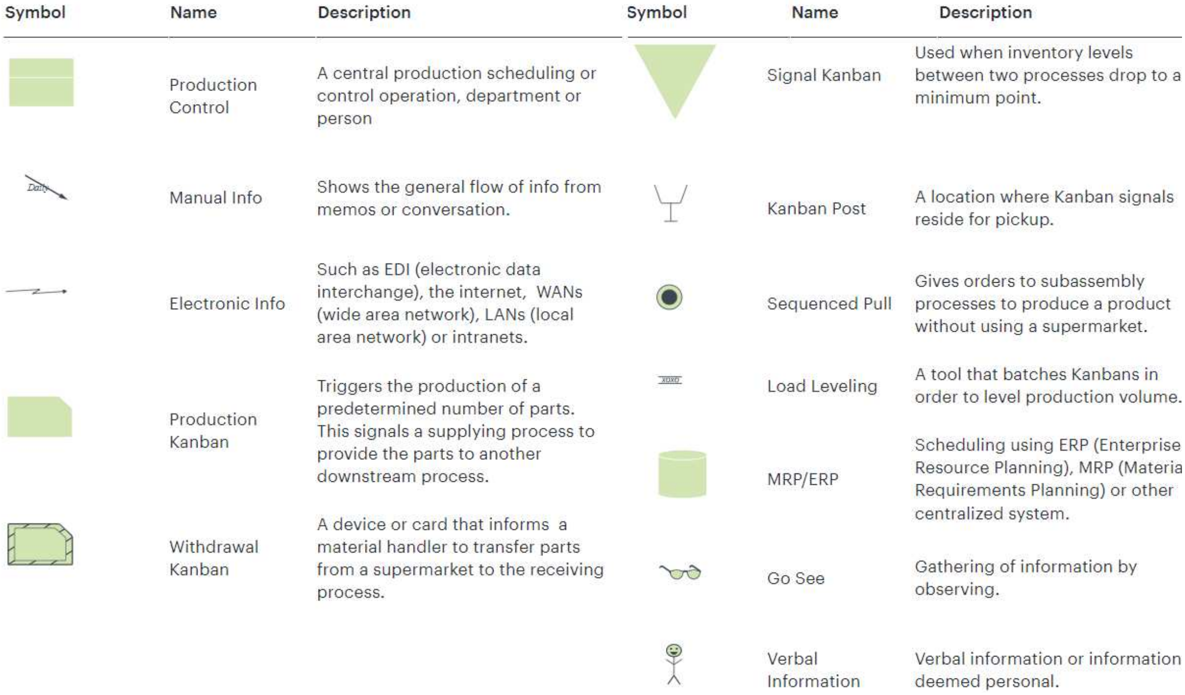


Symbol	Name	Description
▲	Inventory	Inventory between two processes.
	Shipments	Movement of the raw materials from suppliers to the factory and then to customers.
	Push Arrow	Pushing material from one process to the next.
Resource 1 Resource 2 Resource 3	Supermarket	An inventory "supermarket" (also called a kanban stockpoint).
	Material Pull	Removal of materials in a supermarket to downstream processes.

Symbol	Name	Description
FIFO	FIFO Lane	First-In-First-Out inventory.
Stock Stock Stock	Safety Stock	Inventory "hedge" against production problems.
	External Shipment	Shipments from suppliers or to the customers.

Source: Lucidchart (2018)

Figure 6 - Information Symbols

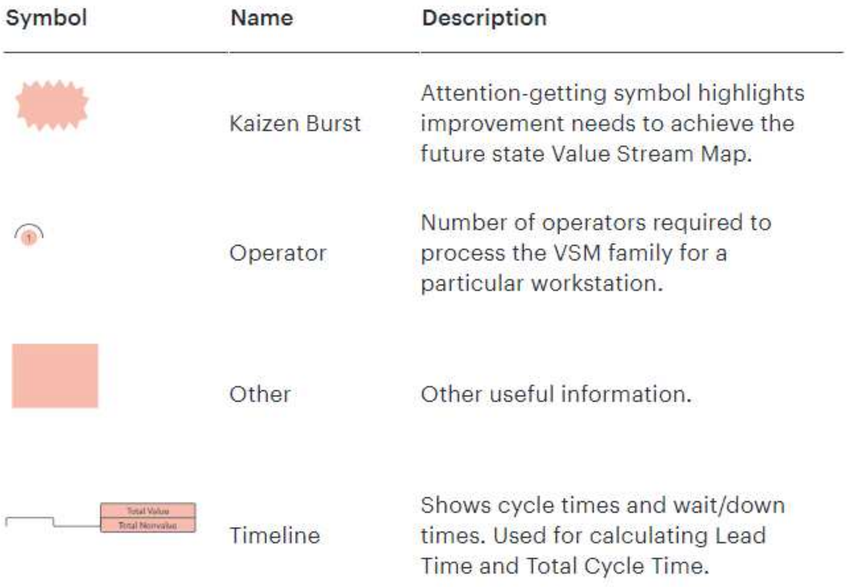


Symbol	Name	Description
🟩	Production Control	A central production scheduling or control operation, department or person
Daily	Manual Info	Shows the general flow of info from memos or conversation.
↯	Electronic Info	Such as EDI (electronic data interchange), the internet, WANs (wide area network), LANs (local area network) or intranets.
🟩	Production Kanban	Triggers the production of a predetermined number of parts. This signals a supplying process to provide the parts to another downstream process.
▨	Withdrawal Kanban	A device or card that informs a material handler to transfer parts from a supermarket to the receiving process.

Symbol	Name	Description
▼	Signal Kanban	Used when inventory levels between two processes drop to a minimum point.
Y	Kanban Post	A location where Kanban signals reside for pickup.
◉	Sequenced Pull	Gives orders to subassembly processes to produce a product without using a supermarket.
XOXO	Load Leveling	A tool that batches Kanbans in order to level production volume.
🟩	MRP/ERP	Scheduling using ERP (Enterprise Resource Planning), MRP (Materia Requirements Planning) or other centralized system.
👓	Go See	Gathering of information by observing.
☺	Verbal Information	Verbal information or information deemed personal.

Source: Lucidchart (2018)

Figure 7 - General Symbols



Symbol	Name	Description
✸	Kaizen Burst	Attention-getting symbol highlights improvement needs to achieve the future state Value Stream Map.
1	Operator	Number of operators required to process the VSM family for a particular workstation.
🟥	Other	Other useful information.
Total Value Total Nonvalue	Timeline	Shows cycle times and wait/down times. Used for calculating Lead Time and Total Cycle Time.

Source: Lucidchart (2018)

From a Lean perspective, improvements focus on adding value by eliminating waste (muda), unevenness (mura), and overburden (muri) (RICHTER *et al.*, 2022). In most value streams, there is always much waste to be eliminated, but it is essential to remember that two paths can be taken: eliminate and add work. Eliminating work does not always result in waste elimination. In many situations, adding work enables new divisions of activities or even more significant gains in process quality.

The development of VSM flow diagrams requires regular meetings and the expertise of individuals to achieve good results, as a large amount of data collected over one year or more from the production system can lead to erroneous conclusions.

The frequency of regular meetings for the development of flow diagrams can vary depending on the organization's needs and the analysis and interpretation capabilities of the team members involved. However, it is common for these meetings to occur periodically, which can be monthly, quarterly, or semi-annually.

Despite the proven efficiency of VSM, it is essential to recognize that it is prone to human errors due to the large amount of information that needs to be analysed and interpreted. However, the growing digitalization has brought new resources that can assist in applying VSM, and this area has been constantly explored in literature. Using digital tools and technologies such as real-time data analysis, process automation, and virtual simulation can reduce the margin of error and improve accuracy in visualizing and understanding the value stream. Integrating these digital solutions with VSM can bring significant benefits in process planning and improvement, enabling more precise analysis and facilitating the implementation of improvements based on more reliable data.

Lugert et al. (2018) evaluate the status of VSM to investigate how it will be suitable for digitalization. Thus, a survey of more than 92% of the participating experts demanding a VSM map using digitalization to compensate for its weaknesses is applied.

The study to improve manufacturing productivity has become an important area for many researchers (ABOLHASSANI *et al.*, 2018), and I4.0 has become a huge factor in increasing productivity. Digitalization enables organizations to collect, analyze, and use data more efficiently, thus improving production process quality, productivity, and efficiency.

Planning and process improvement are essential in the manufacturing industry and benefit from various tools that assist organizations. VSM is a critical method that provides valuable and significant resources, offering visual and qualitative results for process planning. Traditional methods with proven effectiveness over time can benefit from new digital technologies derived from Industry 4.0, allowing the development of complementary resources that enhance the quality and productivity of their applications. Integrating digital technologies such as automation, artificial intelligence, and the IoT provides new opportunities to improve process efficiency, optimize decision-making, and drive innovation in the manufacturing industry.

2.2 Digitalization in the Manufacturing Industry

Digitalization in industry refers to integrating digital technologies into all aspects of the industrial production process, from design and planning to production and maintenance. Smart systems have become very useful in increasing speed and productivity in the industry. According to Burggraf et al. (2018), AI has been mainly used in PPC (Process Planning and Control) in the last 20 years. Therefore, in the next 20 years, there will probably be a significant transformation in managerial decision-making.

Developing intelligent systems encompasses technologies such as IoT, cybersecurity, AR, AM, cloud computing, Big Data, Advanced Analytics, AI, and Digital Twin.

Dornelles et al. (2022) identified the positive and negative impacts of applying intelligent technologies in manufacturing. Among the main positive aspects are waste reduction, improved worker health and safety, time reduction, effectiveness in training, and ease of access to real-time information in activities. The main negative aspects include discomfort in equipment use, technology failure, high investment, and a high level of technical skill required. This information was obtained through a bibliographic review of various publications between 2014 and 2022 and in applications of different technologies in various industry activities. However, advances in search of improvements based on lessons learned have been progressing rapidly, especially when these developments involve the combination of several technologies, further improving manufacturing activities.

Among the technologies of intelligent systems in production and process planning, Cadavid et al. (2019) declared that Advanced Analytics, DS, AI, and ML technologies have been providing new opportunities to make intelligent decisions based on data in PPC. Both technologies are distinct concepts but are often used together to achieve the best results. Both technologies are different concepts but are often used together to achieve the best results.

AI is a branch of computer science that focuses on developing systems and algorithms to perform any task that requires human intelligence. It has been applied in various ways in various areas such as healthcare, finance, retail, transportation, education, marketing, and agriculture, and extensively in the industry in sectors such as manufacturing, industrial automation, predictive maintenance, quality, supply chain, energy, and many others (LEE et al., 2018). AI can solve many internal problems in the industry, including complexities in decision-making (MISRA *et al.*, 2022).

DS is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from mathematics, statistics, AI, and computer engineering to analyze large amounts of data. DS involves the analysis of complex and unstructured data to identify trends, patterns, and insights and aims to provide information and forecasts that can be used to make strategic decisions (SARKER, 2021). Regardless of the source, manufacturing companies need clean data to work. When dealing with real-world data, it is easy to find corrupt or inaccurate records and incomplete, incorrect, or irrelevant parts of the data. Clearing this data requires using one of the many processes of DS.

DS encompasses three skills: the skills of a statistician who knows how to model datasets, the skills of a computer scientist who effectively can use algorithms, and the domain expertise who needs to formulate questions and appropriately put their answers on them (VANDERPLAS, 2016).

DS is a superset for three other fields: DA (Data Analytics), DM (Data Mining), and PM, which can be used separately or together (VANDERPLAS, 2016).

Analytics is a common business word and describes a quantitative critical thinking method. DA is a subset of BA (Business Analytics). It is the process of examining data sets to conclude, enabling organizations to make more-informed business decisions and verify scientific models, theories, or hypotheses (ZHANG, 2017).

A taxonomy, according to (MOREIRA et al., 2018), that exists for DA is:

i) Descriptive analytics: summarize or condense data to extract patterns.

ii) Predictive Analytics: extract models from data for future predictions.

According to (DAI *et al.*, 2020), DA can extract vast business values from manufacturing data. There are many reasons to use DA in manufacturing, such as improving factory operations and production, reducing machine downtime, improving product quality, enhancing supply chain efficiency, and improving customer experience.

Nevertheless, due to the large volume of data and heterogeneous data types, there are also many challenges, such as difficulty in data representation, data transmission, data integration, data cleaning, data compression, redundancy reduction, reliability, data persistence, storage, scalability, efficiency, temporal and spatial correlation of data, efficient DM schemes and privacy and security.

At the same time, it has been doing many studies with enabling technologies to figure out new ways to overcome these challenges. Regarding DA, statistical modelling methods have been used for descriptive statistics, inferential statistics, stochastic methods, and several DM and ML algorithms.

Differences between DM and DA are related to their i) definition: DM is the process of discovering useful patterns and trends in large data sets, and DA is the process of extracting information from large data sets to make predictions about future outcomes; ii) importance: DM helps to better understand collected data (e.g. stock price timeline analysis, purchase pattern across geography or time), and DA predicts on top of DM result by applying domain knowledge; iii) scope: DM applies ML algorithm like regression, classification on collected data to find hidden patterns, and DA applies business knowledge on DM patterns with any additional data needed to get business valid predictions ; iv) outcome: DM outputs a pattern in data in form of a timeline varying distribution or clusters, and DA tries to find answers behind the pattern with applying business knowledge and thus making it more actionable piece of information ; and v) people involved: DM is mainly done by statisticians and ML engineers, and DA involves business analysts and other domain experts who can analyse and interpret the patterns discovered by the machines (KRISHNASWAMY *et al.*, 2019; POORNIMA; PUSHPALATHA, 2018).

The main aim of DM is to structure data by processing and categorizing the data using various techniques such as classification, and feature selection. The purpose is to understand the data with great insight and then to make predictions for new and unknown data (DUAN; XIONG, 2015).

Vazan et al. (2017) define DM as an interdisciplinary field aimed at predicting outcomes and uncovering relationships in data, employing sophisticated algorithms, and discovering hidden patterns, anomalies, and structures of vast amounts of data stored in information repositories. Prediction and description are two ultimate DM goals in a manufacturing context.

There are nine laws of DM, according to Hindle & Vidgen (2018), which are used as guiding principles for any data miner:

1) Business objectives are the origin of every DM solution;
2) Business knowledge is central to every step of the DM process;
3) Data preparation is more than half of every DM process;
4) The suitable model for a given application can only be discovered by experiment;
5) There are always patterns;
6) DM amplifies perception in the business domain;
7) Prediction increases information locally by generalization;
8) The value of DM results is not determined by the accuracy or stability of predictive models; and
9) All patterns are subject to change.

Like DM, PM techniques also provide a powerful method, but they can discover, analyze, checking the conformance of different processes and extend it with bottleneck identification and resource information. All recorded information is stored in event logs (Table 1). Event logs are created in data frames primarily to establish suitable relationships among data and analyze information on what was done, by whom, for whom, where, when, and so on. They are a set of events grouped into traces.

According to van der Aalst (2016) and Myers et al. (2018), PM is the missing link between model-based process analysis and data-oriented analysis techniques. A process model captures the control flow of a process event, and its order is explicitly expressed graphically, and it is used to compare existing process models with event logs to compare actual behaviors with expected behaviors.

Hence, from a more general perspective, both DS and ML methods are closely related to enhancement. Before these new technologies started to grow, LM methods were the most proven, advanced, and effective methods to implement in the industry.

Data analysis is one of the main stages of DS and involves collecting, cleaning, transforming, and exploring data to extract useful information, make decisions, and solve problems. Like data analysis, advanced analytics is also part of

DS. Advanced analytics refers to the technologies and techniques used to promote autonomous or semi-autonomous data analysis, which enables the emergence of insights, predictions, or suggestions.

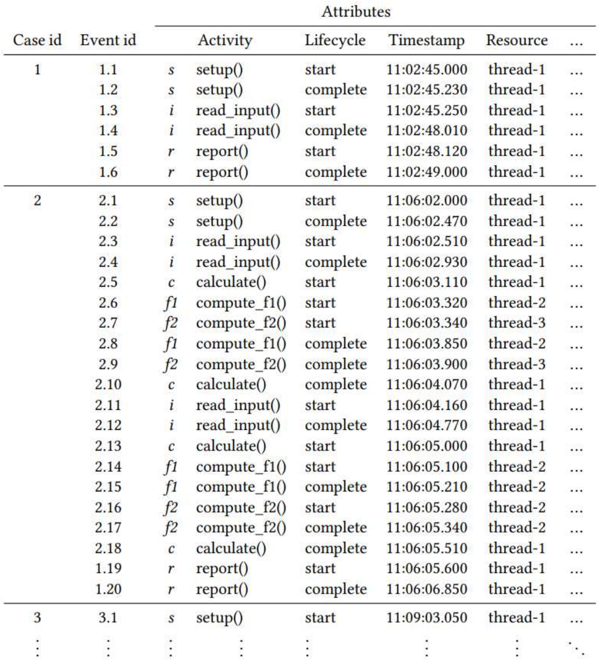
Advanced Analytics is a sub-area of DA that focuses on more advanced techniques such as predictive modeling, DM, network analysis, and other sophisticated methods. It seeks to predict what may happen in the future using statistical models and advanced algorithms, such as ML algorithms (VARSHA *et al.*, 2021).

So far, all the methods, techniques, and fields presented have been integrated into various multidisciplinary ways (Figure 8) to improve and create innovative solutions to many areas.

Particularly for the manufacturing industries, DS has taken the automation and enhancement of manufacturing processes to a new level since it triggered the beginning of I4.0.

Several programming languages are generally used in Advanced Analytics, such as Python, R, SQL (Structured Query Language), JS (JavaScript), Java, and Matlab. However, considering the development of this thesis, the highlighted languages are Python, SQL, and JS.

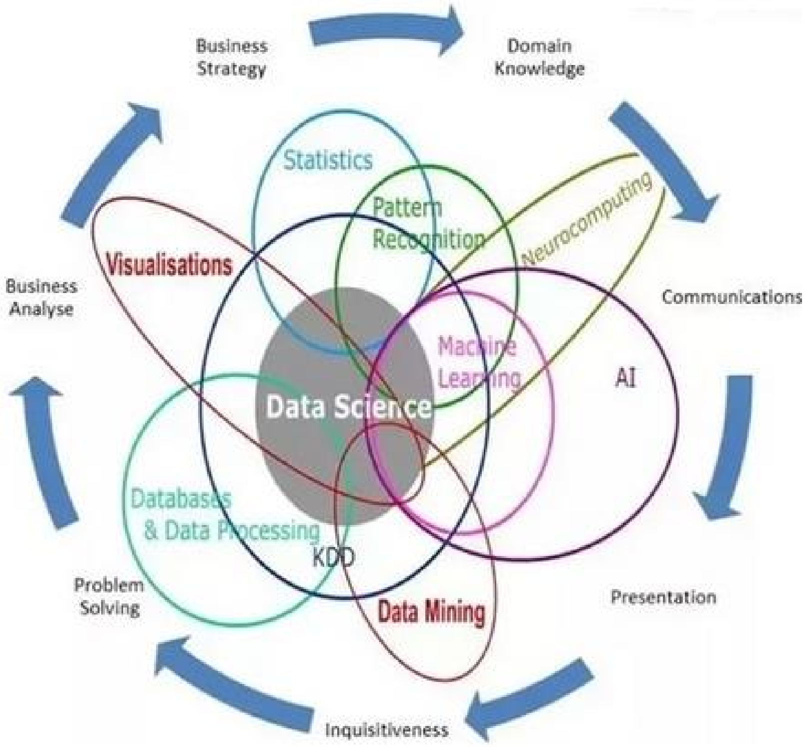
Table 1 - Event log for a software system



		Attributes					
Case id	Event id		Activity	Lifecycle	Timestamp	Resource	...
1	1.1	*s*	setup()	start	11:02:45.000	thread-1	...
	1.2	*s*	setup()	complete	11:02:45.230	thread-1	...
	1.3	*i*	read_input()	start	11:02:45.250	thread-1	...
	1.4	*i*	read_input()	complete	11:02:48.010	thread-1	...
	1.5	*r*	report()	start	11:02:48.120	thread-1	...
	1.6	*r*	report()	complete	11:02:49.000	thread-1	...
2	2.1	*s*	setup()	start	11:06:02.000	thread-1	...
	2.2	*s*	setup()	complete	11:06:02.470	thread-1	...
	2.3	*i*	read_input()	start	11:06:02.510	thread-1	...
	2.4	*i*	read_input()	complete	11:06:02.930	thread-1	...
	2.5	*c*	calculate()	start	11:06:03.110	thread-1	...
	2.6	*f1*	compute_f1()	start	11:06:03.320	thread-2	...
	2.7	*f2*	compute_f2()	start	11:06:03.340	thread-3	...
	2.8	*f1*	compute_f1()	complete	11:06:03.850	thread-2	...
	2.9	*f2*	compute_f2()	complete	11:06:03.900	thread-3	...
	2.10	*c*	calculate()	complete	11:06:04.070	thread-1	...
	2.11	*i*	read_input()	start	11:06:04.160	thread-1	...
	2.12	*i*	read_input()	complete	11:06:04.770	thread-1	...
	2.13	*c*	calculate()	start	11:06:05.000	thread-1	...
	2.14	*f1*	compute_f1()	start	11:06:05.100	thread-2	...
	2.15	*f1*	compute_f1()	complete	11:06:05.210	thread-2	...
	2.16	*f2*	compute_f2()	start	11:06:05.280	thread-2	...
	2.17	*f2*	compute_f2()	complete	11:06:05.340	thread-2	...
	2.18	*c*	calculate()	complete	11:06:05.510	thread-1	...
	1.19	*r*	report()	start	11:06:05.600	thread-1	...
	1.20	*r*	report()	complete	11:06:06.850	thread-1	...
3	3.1	*s*	setup()	start	11:09:03.050	thread-1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Source: Leemans et al. (2018)

Figure 8 - Multidisciplinary Set



Source: Tierney & Kelleher (2018)

2.2.1 Python

Python is known for its readability, simplicity, and ease of learning, with an intuitive syntax that makes code creation clear and efficient. Python is a language that supports OOP (Object-Oriented Programming) and FP (Functional Programming). Python allows developers to create classes, inheritance, encapsulation, and polymorphism, among others. On the other hand, Python also supports FP, which makes it a very flexible language allowing developers to use different programming paradigms depending on the project's needs.

Python has many valuable data loading, visualization, image processing, statistics, and more libraries. Some of the most popular libraries in Python include NumPy, SciPy, Pandas, and Matplotlib (MARTINUZZI *et al.*, 2018).

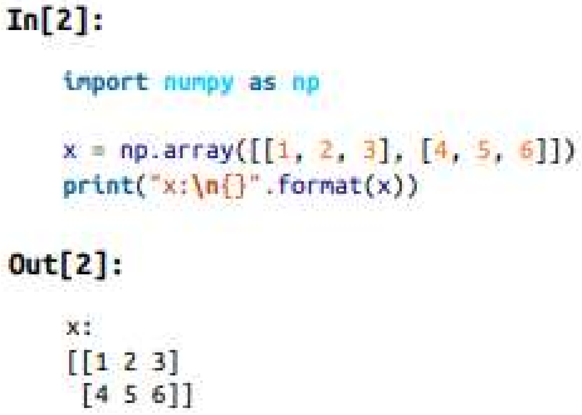
NumPy is a fundamental Python scientific computing package that uses high-level mathematical functions such as linear algebra operations. It is one of the most popular libraries due to its efficiency, ease of use, and array and matrix processing functions. Numpy is built around the ndarray (N-dimensional array) object, which allows data to be represented in various dimensions. It is possible to efficiently perform mathematical and statistical operations on large data sets (Figure 9).

SciPy is also a collection of functions for scientific computing, but it is used for advanced linear algebra routines, signal processing, special mathematical functions, sparse matrices, and statistical distributions. This library is divided into several submodules that provide functions for various tasks, such as signal processing (scipy.signal), image processing (scipy.ndimage), optimization (scipy.optimize), statistics (scipy.stats), among other diverse areas. Scipy also offers various tools for data interpolation, numerical integration, linear algebra, and Fourier transform, among other functions (Figure 10).

Pandas is a library for data wrangling and analysis. This library is built around two main objectives: Series and Dataframes. Series is a one-dimensional data structure that can store various data types, while Dataframe is a two-dimensional data structure that can store tabular data with columns and rows. There are many features for cleaning and manipulating data, merging, and joining data, filtering, and aggregation, among others. It is widespread to use the Pandas library in conjunction with data analysis and ML libraries, such as scikit-learn (Figure 11).

Finally, Matplotlib provides functions to make many graphical visualizations. This library offers many different types of charts, including line charts, histograms, scatter plots, area charts, and surface charts, among others, and it is possible to customize these visualizations with different colors, styles, sizes, annotations, and labels (Figure 12).

Figure 9 - Example of Numpy algorithms



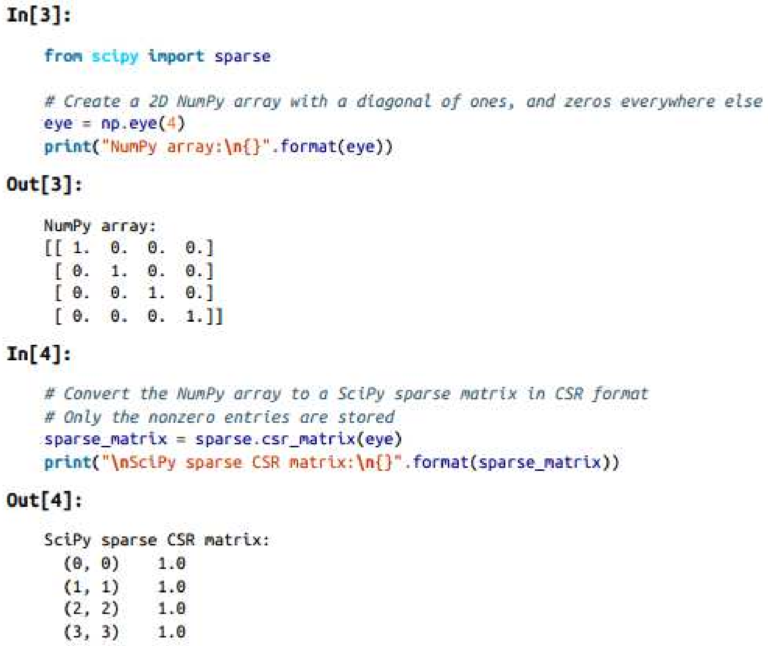
```
In[2]:

    import numpy as np

    x = np.array([[1, 2, 3], [4, 5, 6]])
    print("x:\n{}".format(x))

Out[2]:

    x:
    [[1 2 3]
     [4 5 6]]
```

Source: Müeller & Guido (2016)

There have been many developments in improvement in the manufacturing industry, using DS through functions with Python libraries and mainly through ML.

Sahoo et al. (2019) used exploratory data analysis using Pandas, Numpy, and Matplotlib libraries to analyze and identify patterns and improvements in an Amazon electronics dataset and achieved excellent results.

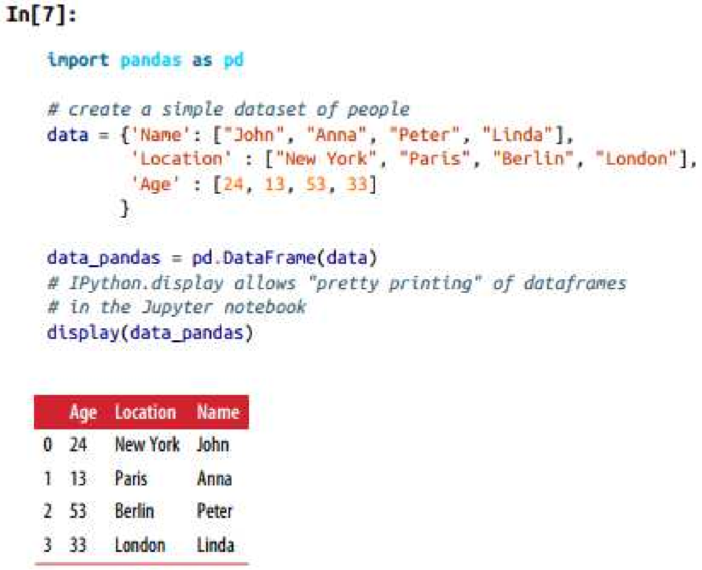
Figure 10 - Example of SciPy algorithms



```
In[3]:

    from scipy import sparse

    # Create a 2D NumPy array with a diagonal of ones, and zeros everywhere else
    eye = np.eye(4)
    print("NumPy array:\n{}".format(eye))

Out[3]:

    NumPy array:
    [[ 1.  0.  0.  0.]
     [ 0.  1.  0.  0.]
     [ 0.  0.  1.  0.]
     [ 0.  0.  0.  1.]]

In[4]:

    # Convert the NumPy array to a SciPy sparse matrix in CSR format
    # Only the nonzero entries are stored
    sparse_matrix = sparse.csr_matrix(eye)
    print("\nSciPy sparse CSR matrix:\n{}".format(sparse_matrix))

Out[4]:

    SciPy sparse CSR matrix:
      (0, 0)    1.0
      (1, 1)    1.0
      (2, 2)    1.0
      (3, 3)    1.0
```

Source: Müeller & Guido (2016)

Figure 11 - Example of Pandas algorithm

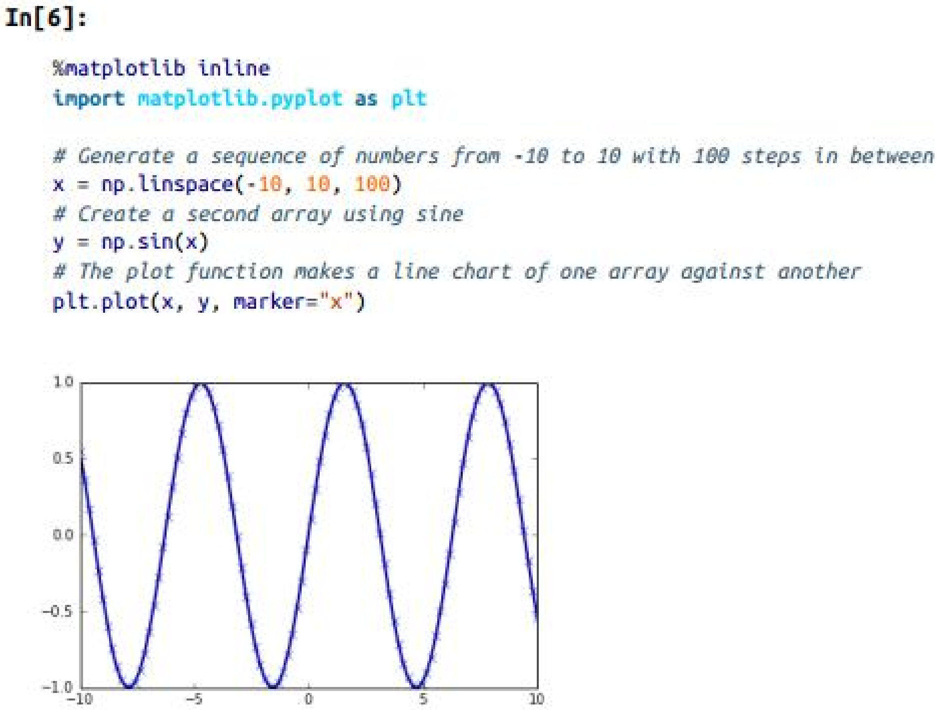


```
In[7]:

    import pandas as pd

    # create a simple dataset of people
    data = {'Name': ["John", "Anna", "Peter", "Linda"],
            'Location' : ["New York", "Paris", "Berlin", "London"],
            'Age' : [24, 13, 53, 33]
           }

    data_pandas = pd.DataFrame(data)
    # IPython.display allows "pretty printing" of dataframes
    # in the Jupyter notebook
    display(data_pandas)
```

	Age	Location	Name
0	24	New York	John
1	13	Paris	Anna
2	53	Berlin	Peter
3	33	London	Linda

Source: Müeller & Guido (2016)

Figure 12 - Example of Matplotlib algorithm



```
In[6]:

    %matplotlib inline
    import matplotlib.pyplot as plt

    # Generate a sequence of numbers from -10 to 10 with 100 steps in between
    x = np.linspace(-10, 10, 100)
    # Create a second array using sine
    y = np.sin(x)
    # The plot function makes a line chart of one array against another
    plt.plot(x, y, marker="x")
```

Source: Müeller & Guido (2016)

Morariu et al. (2020) developed a hybrid control solution through big data techniques and ML algorithms to process real-time flows of time information in large-scale manufacturing systems, focusing on energy consumption aggregated in various layers. This approach allowed for the accurate prediction of energy consumption patterns during production and the ability to detect anomalies based on predicted energy data.

ML is a field of AI that allows computers to learn and improve from provided data and use this data to make future predictions or decisions on new data (NAMDEV; AGRAWAL; SILKARI, 2015). Moreover, it is also known as statistical learning or predictive analytics, and it helps open new possibilities for improving decision-making and performance on several tasks. Martinuzzi et al. (2018b) emphasized that the most successful ML algorithms are those which automate decision-making processes by generalizing from known examples.

Although ML is a subfield of AI, they do not have the same goal because AI is intended to simulate natural intelligence to solve complex problems. AI is not a system, but it can be implemented within a system to drive computer programs that can work smartly. On the contrary, ML's goal is to learn from data for a specific task to maximize performance, then it is a system that can work and learn from datasets. ML algorithms are divided into three categories: supervised learning, unsupervised learning, and reinforcement learning (PAPER, 2020).

Supervised learning has pre-existing learning that allows the algorithm to make decisions or predictions based on the data it used for its training, i.e., the inputs and outputs are known. An example of use is fraud detection in credit cards because, in this context, supervised learning can be used to create a model that learns from transactions labeled as fraudulent or legitimate.

Supervised learning has three main categories: classification, regression, and pattern recognition.

Classification involves estimating a qualitative output of unobserved data based on input data, which covers observations with already defined categories. Predicting whether an email is spam or not or whether a patient has a disease are examples of classification learning.

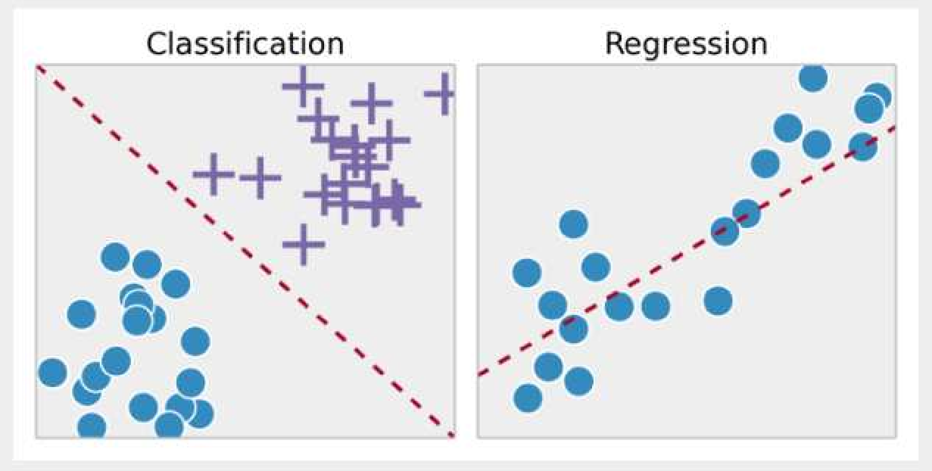
Regression, although using already observed input data, like classification, should estimate a numerical value. Regression mathematically models the unknown or dependent variable and the known or independent variable as a linear equation. Based on its financial data, the created model will predict a serial number, such as the price of a house or a company’s revenue, in the next year. In Figure 13, it is possible to visualize a representation of the prediction difference between classification and regression.

Pattern recognition involves identifying patterns in input data. Facial recognition in images and anomaly detection in sensor data are examples of this type of learning.

Unsupervised learning aims to find patterns and organize them without prior knowledge of the data. An example of use is customer segmentation, intending to segment customers and better understand their behaviors, preferences, and needs, thus offering better products and services. The unsupervised learning model can group customers into different segments based on their similarities.

Reinforcement learning uses algorithms that can learn from their own experience. This category makes the model capable of making the best decisions in different situations according to a trial-and-error process, in which there may be a reward for successful outcomes. An example of use would be temperature control in a heating system with a thermostat. If the thermostat is equipped with sensors, it can be adjusted with a learning model that keeps it within a desired temperature range by adjusting its heat output.

Figure 13 - Classification and Regression Prediction



Source: Edell (2015)

There are several popular models for Machine Learning in classification and regression. Some of the leading classification models are:

Logistic Regression is a classification model that estimates the probability of an instance belonging to a specific class. Decision Trees use a tree structure to make decisions based on data features. Random Forest is a set of decision trees that combines their predictions to perform a more robust classification. SVM (Support Vector Machines) is a model that maps the data into a higher-dimensional space to find a decision boundary that maximizes the separation between classes. Naive Bayes is a model based on Bayes' theorem that assumes independence among features and calculates the probability of an instance belonging to a class.

Some of the leading regression models widely used are:

Random Forest Regressor model combines multiple decision trees to perform regression. Each tree is trained on a random sample of the data and makes an individual prediction. The predictions of all the trees are combined to obtain a more robust and accurate final prediction.

KNeighbors Regressor is an instance-based model that uses the "k" nearest neighbors to perform regression. It finds the "k" closest instances in the feature space and averages or weights them to make the prediction. Linear Regression is a classic model that estimates the linear relationship between input variables and the output variable. It finds the coefficients that best fit the straight line to the data and uses this line to make predictions.

SVR (Support Vector Regression) maps the data into a higher-dimensional space and finds a decision boundary that best fits the data. SVR aims to minimize the distance between the data points and the decision boundary.

MLP (Multilayer Perceptron) Regressor is a regression model based on artificial neural networks. It consists of multiple layers of interconnected artificial neurons, where each neuron receives weighted inputs and applies a non-linear activation function to generate an output. MLP Regressor can learn complex relationships in the data and make accurate predictions.

Gradient Boosting Regressor is a model that combines multiple regression trees in a sequence, where each tree attempts to correct the errors of the previous trees. It performs a boosting process, where each new tree is adjusted to the residuals of the previous model. The final prediction is obtained by summing the predictions of all the trees. These models offer different approaches and characteristics for classification and regression tasks, and the choice of the most suitable model depends on the data characteristics and specific objectives of the problem.

When developing applications and solutions using machine learning, programming using a database, and controlling the predicted responses' quality is essential. For this control to happen, it is necessary to use lines of code with functions that analyze the results with appropriate metrics (CARVALHO; PEREIRA; CARDOSO, 2019; ZHOU *et al.*, 2021). The validation of a ML model is critical for the accuracy and effectiveness of the models used. These validations allow developers to find problems in the model and improve it before putting it into production.

Among the metrics for classification include recall, F1 score, and precision, and for regression include MAE (Mean Absolute Error), MSE (Mean Squared Error), and R^2, which is also called the coefficient of determination (HANDELMAN *et al.*, 2019).

Recall measures the proportion of positive instances correctly identified by the model relative to the total number of positive instances in the test data. In other words, recall indicates the model's ability to retrieve relevant instances correctly. The higher the recall, the better the model correctly identifies positive instances.

Precision measures the proportion of positive instances correctly identified by the model relative to the total number of instances predicted as positive (i.e., the sum of true positives and false positives). Precision indicates the model's ability to make correct optimistic predictions. High precision means that the model's positive predictions are reliable and have a low rate of false positives.

F1 Score is a metric that combines both recall and precision into a single value, providing a balanced measure of the model's performance. It is calculated as the harmonic mean between recall and precision. The F1 Score is beneficial when there is an imbalance between the interest classes. It provides an overall measure of performance that considers both true and false positives.

MAE calculates the average absolute differences between the predictions and the actual values. It indicates the average magnitude of the prediction error.

MSE calculates the average of the squared differences between the predictions and the actual values. It is a metric that penalizes more significant errors.

R2 measures the proportion of the variance in the output variable that the independent variables can explain. A value of R2 close to 1 indicates a good fit of the model to the data.

RMSE is the square root of the MSE and represents the square root of the average prediction errors. It is a metric that estimates the standard deviation of the prediction errors.

SQL language is another programming language considered essential in projects developed in conjunction with data stored in databases.

2.2.2 SQL

SQL is a programming language used to manage and manipulate data in relational databases. SQL is widely used in data storage projects and is essential for projects that involve data analysis or software development that involves data.

Many programming libraries in other programming languages, such as Python and Java, support SQL. Several relational database management systems on the market use SQL as the standard for managing these databases, such as MySQL, Oracle, PostgreSQL, Microsoft SQL Server, IBM DB2, Amazon Aurora, and others. The SQL language is designed to be a query language, allowing for retrieving, updating, and manipulating data in relational databases.

This language is based on statements, which are commands sent to the database to perform specific operations, such as SELECT, INSERT, DELETE, and UPDATE. The SQL language is set-oriented, allowing operations to be quickly performed on large sets of data (MUKHERJEE, Shubham *et al.*, 2015).

SQL supports transactions, allowing operations to be handled safely and consistently, with functions grouped into a single unit, as well as supporting referential integrity and stored procedures, ensuring that the information stored in the database is accurate and reliable, and allowing for the creation of automated and customized workflows, respectively.

This language was specifically developed to manipulate data stored in a relational data model, i.e., where information is organized in tables. Non-relational databases, or NoSQL databases, use more flexible and scalable data models, as they are based on different structures, such as documents, graphs, key-value pairs, and columns (MUKHERJEE, Shubham *et al.*, 2015; NAYAK; PORIYA; POOJARY, 2013).

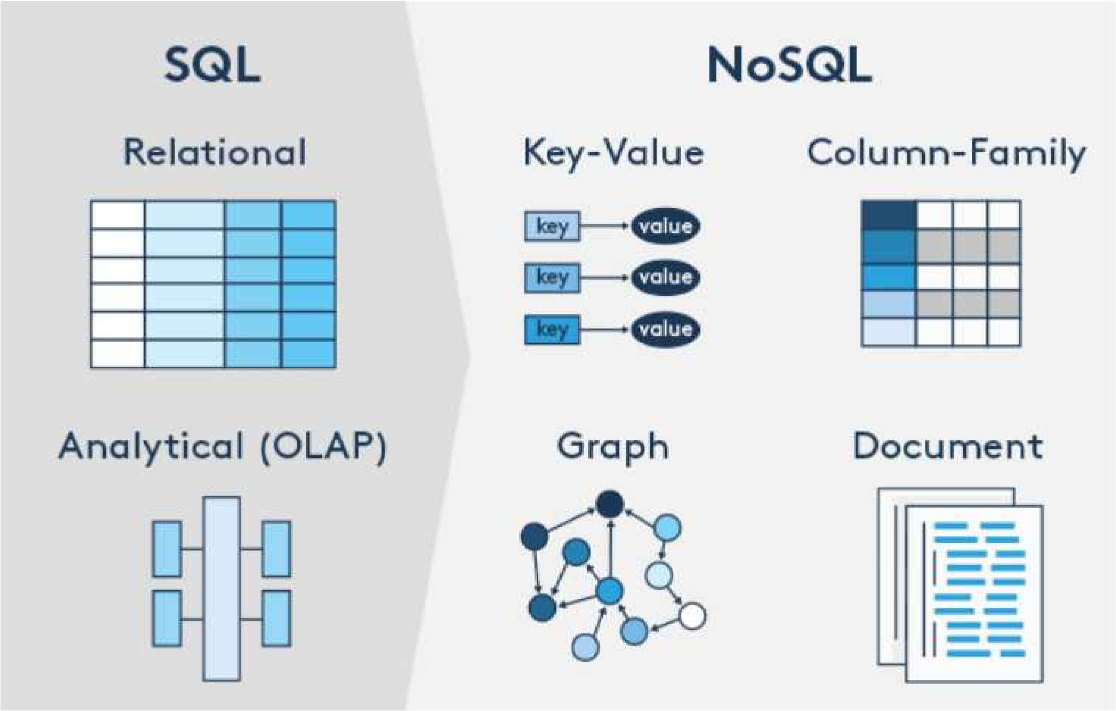
Non-relational databases typically use other query languages and several databases available on the market, such as MongoDB, Apache Cassandra, Redis, Couchbase, Amazon DynamoDB, and others. MongoDB uses its query language called MQL (Mongo DB Query Language), which is based on documents, while Cassandra uses CQL (Cassandra Query Language), which is based on columns (APACHE, 2023b; MONGODB, 2022)

The advantages of relational databases over non-relational ones include a highly organized structure, ACID (Atomicity, Consistency, Isolation, Durability) support transactions that ensure transactions are completed, and changes are rolled back in case of failures, vertical scalability, and reliability. On the other hand, the advantages of non-relational databases over relational ones include schema flexibility, horizontal scalability through the addition of more servers, better performance, and more cost-effective in terms of hardware and maintenance costs, especially in terms of hardware and maintenance costs (NAYAK; PORIYA; POOJARY, 2013).

Therefore, the choice of the database structure is relative to the project's needs, and some of these differences can still be seen in Figure 14.

In this project, it is essential to establish relationships between data and handle various relationships to establish a proper information order. Data consistency and accuracy are also of utmost importance. Additionally, relational MySQL has been identified as the ideal choice due to the need to analyze a large volume of data. The organized structure of tables and the ability to define complex relationships in MySQL offer the flexibility and integrity required for this project. The capability of MySQL to handle complex queries and its scalability are advantages that meet the requirements for analyzing large datasets. These considerations will be further addressed in the detailed explanation of the research methodology and the obtained results.

Figure 14 - Database Types



Source: DbaExperts (2002)

2.2.3 Javascript

JS is a programming language widely used to create interactive and dynamic web pages, including web applications, and is supported by many libraries and frameworks (FRISBIE, 2019). Although originally structured for web development, this language can also be used in data analysis, with the support of libraries such as Highcharts, D3.js, and Chart.js.

These libraries allow the creation of interactive and dynamic data visualizations and charts, with many features for exploring data. JS can also be used in conjunction with the Python language and SQL language, and there are many data analysis and BI tools, such as Tableau, PowerBI, and Metabase, support the incorporation of JS for dashboard development (NETEK; BRUS; TOMECKA, 2019).

JS is considered a front-end language used for developing the user interface on a website or application, which is what the user sees and interacts with. Python and SQL can be used for both front-end and back-end development, but Python is more frequently used as a back-end language, and SQL is mainly used to interact with the database (FRISBIE, 2019; STAICU *et al.*, 2019).

Back-end languages are used to develop server-side logic, which means they are used to process and manage data on the server side. These languages are used to manipulate data, perform database operations, authenticate users, manage sessions, protect against security threats, generate content dynamically, and provide quick responses to users. Both languages are essential for developing robust and scalable software solutions (LI, Nian; ZHANG, 2021).

JS has various libraries and considerable front-end development ability, with interactive and dynamic interfaces, animations, visual effects, and advanced form elements. JS is versatile and integrated with various other systems, such as Node-RED, a flow-based modeling tool.

2.3 Flow-Based Modelling

Flow-based modeling is a process modeling approach that uses visual representations of activity flows to describe a business process. These flows represent the sequence of activities occurring in a process and the interaction between them, allowing for a more precise and organized view of the process. They can represent various processes, from business processes to software processes and industrial processes (GIBELHAUS *et al.*, [*s. d.*]; PEDROSA; PUIG; NEJJARI, 2022)

BPM (Business Process Management) and flow-based modeling are directly related. BPM is a management discipline focused on improving business performance results based on automation, measurement, design, execution, and enhancement of processes for a company. Flow-based modeling is one of the main techniques BPM uses to visually represent business processes and thus aid in identifying bottlenecks and understanding the process as a whole (SLAATS, 2020).

However, flow-based modeling is often used with business process automation tools. Several models perform this combination, such as Apache NiFi, which is a data processing platform that uses the flow programming model to automate data movement between systems (APACHE, 2023a), LabView, which is a graphical development environment that uses a flow programming model to design control and data acquisition systems (NATIONAL INSTRUMENTS, 2023), Simulink, which is a flow-based tool used to develop control and communication systems (MATHWORKS, 2023), KNIME, which is a data analysis platform that uses a flow programming model to allow users to combine and process data from various sources (KNIME, 2023), Apache Camel, which is an integration framework with a flow-based programming model to enable the integration of heterogeneous systems (APACHE, 2023c), and Node-RED, which is a visual programming platform and flow-based modeling developed to connect devices and services quickly and easily.

These tools offer satisfactory results, but Node-RED has several advantages compared to Apache NiFi, LabVIEW, Simulink, KNIME, and Apache Camel. First and foremost, Node-RED is a visual development platform based on nodes, making it easy and fast to create and configure data flows intuitively. It provides a friendly and visual interface for creating and editing flows, allowing users to quickly build complex logic for data processing.

Node-RED has an extensive library of ready-to-use nodes covering many functionalities and integrations. This platform allows users to leverage various resources, from cloud service integration to communication with IoT devices, without the need to develop everything from scratch.

Another advantage of Node-RED is its scalability. It can run in various environments, from small devices to high-performance servers, enabling efficient processing of large volumes of data. Furthermore, Node-RED is highly modular and supports the creation of custom extensions, further enhancing its flexibility and adaptability to different requirements and use cases (NODE-RED, 2023).

Node-RED also benefits from an active community and continuous developer support, meaning a wide range of resources, tutorials, and examples are available. This support facilitates learning and solution development using Node-RED.

2.3.1 Node-RED

Node-RED was developed to assist with IoT project implementations. The open-source system comprises "nodes", which can be clicked, dragged, and dropped onto the screen. When the nodes are connected, they offer various options, from simple debugging to connection with a Raspberry Pi.

There are three types of nodes: input, output, and processing (Figure 15). Input nodes allow data to be entered into a particular application, output nodes enable data to be sent out of the application, and processing nodes allow data manipulation to provide new data to a new stage.

Data is obtained for use in Node-RED through its library's input nodes. These nodes allow the reception of data from various sources such as APIs, databases, web services, sensors, IoT devices, and others.

A specific input node can be added for the desired data source when creating a flow in Node-RED. For instance, an HTTP request node can be included to retrieve data from an API to send the request and receive the API's response. Specific nodes can be added to establish communication and receive the data they transmit if working with IoT devices.

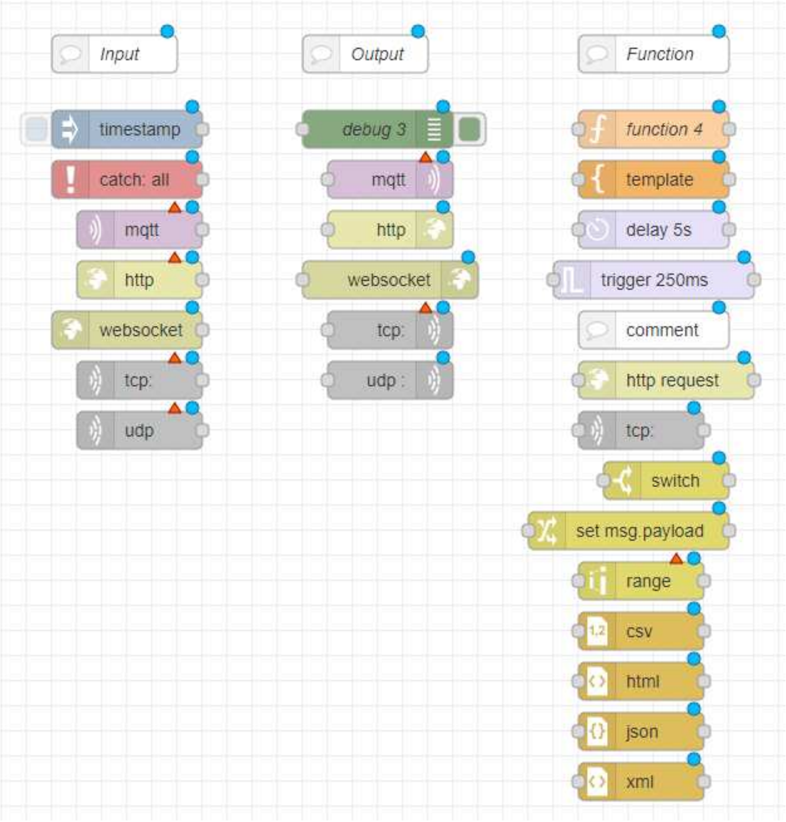
Each input node has its configurations, such as URL, authentication, and refresh interval, which depend on the specific data source. These configurations are defined according to the data source's requirements to ensure proper data retrieval by Node-RED.

Once the input nodes receive the data, it can be manipulated, processed, and sent to other nodes for additional actions. Node-RED provides a variety of processing and output nodes to assist in working with the data as desired.

Node-RED features a flow editor in the central area that facilitates configuration and connection of flows, a flow library on the left side of the main site that provides a wide variety of available flows, a debug panel which is a user interface section that allows for detailed information about the current flow execution to be visualized, and a properties panel which is a user interface section where the properties of a selected node are displayed and can be edited (Figure 16). Node-RED has a feature that allows users to share their flows with the community, and there are currently over 3000 nodes in this library. However, since many nodes have been created outside the community, the number of available nodes is much larger than this value (NODE-RED, 2023).

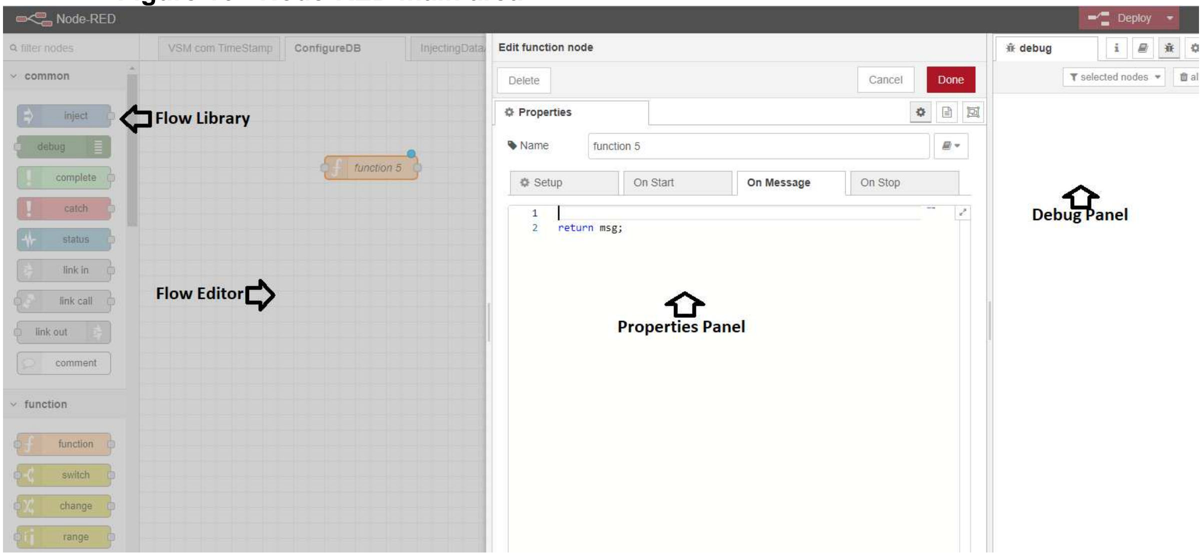
Node-RED’s lightweight runtime is built on Node.js, an open-source code runtime environment based on JS. Flows created in Node-RED are stored using JSON, a simple and lightweight data format for exchanging data between systems. They can be easily imported and exported for sharing with others.

Figure 15 - Node-RED nodes examples



Source: The Author (2023)

Figure 16 - Node-RED main area



Name: function 5

```
return msg;
```

Source: The Author (2023)

The wide variety of nodes and the ability to integrate different devices and services make it a powerful tool for building IoT, automation, and data analysis solutions. Node-RED can be run locally, with equipment such as Arduino, and Raspberry Pi, among others, and can also be run on a cloud computing service (PIANA; SILVA PINTO JUNIOR; GONÇALVES, 2022; SICARI; RIZZARDI; COEN-PORISINI, 2019).

Cloud computing is a computing model that allows on-demand access, via the internet, to a large set of shared resources, such as servers, databases, and storage, among other applications (SUNYAEV, 2020).

Sicari et al. (2019) developed a Node-RED application in a case study focused on intelligent transportation and logistics to demonstrate the tool's viability and usability in representing and testing IoT environments' behavior.

In IoT technology developments, it is common to use Node-RED with an MQTT broker to exchange messages with these technologies and thus create scalable and flexible IoT solutions integrating various devices.

2.4 Message Queuing Telemetry Transport

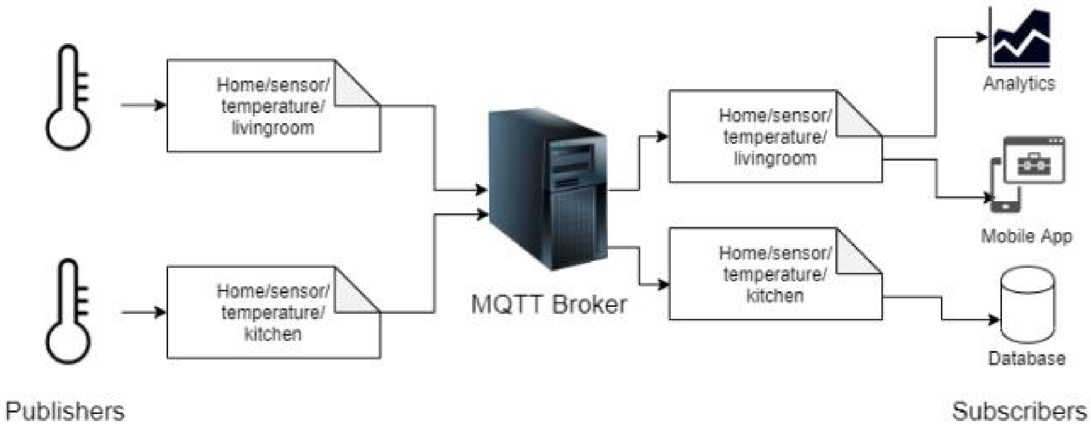
MQTT is a lightweight and efficient message communication protocol based on the Internet's foundation of TCP/IP (Transmission Control Protocol/Internet Protocol). According to Mishra et al. (2021), it can transmit telemetry data between constrained devices in unreliable networks. However, it is designed to operate with the TCP/IP protocol to provide reliable service and ensure that messages are delivered correctly. The MQTT protocol has three constituent components: (i) a publisher or producer (MQTT client), a broker (MQTT server), and a subscriber or consumer (MQTT client).

A client who is responsible for opening a network connection, creating, and sending messages to the server is called a publisher. The subscriber is a client that subscribes to a topic in advance so that it can receive messages. This subscriber can also unsubscribe from a topic to remove an application message request and thus close the connection to the server. The server acts as an intermediary between the producer and the consumer, receiving and forwarding messages to all subscribers (MACHESO *et al.*, 2021; MISHRA; MISHRA; KERTESZ, 2021; RATTANAPOKA *et al.*, 2019). Figure 17 demonstrates a basic model of MQTT functioning.

To increase reliability in unreliable networks, MQTT can be used as QoS (Quality of Service), i.e., a set of mechanisms used to manage the quality and reliability of data transmission for message delivery assurance with receipt confirmation, however, the use of QoS levels tends to increase network traffic and response time, thus bringing some disadvantages to this use (SONI; MAKWANA, 2017).

Many available brokers can be used, such as Mosquitto, ActiveMQ, CloudMQTT, RabbitMQ, VerneMQ, HiveMQ, and EMQX, among others. However, unlike others, Mosquitto is a popular and reliable option, especially for resource-limited environments. Among the main advantages of Mosquitto is its lightweight and low resource consumption with high performance, ease of installation and configuration, support for advanced features, supporting the three primary QoS levels, and use of SSL (Secure Sockets Layer) /TLS (Transport Layer Security) for communication encryption. In addition, it is open source, which means it can be modified and customized according to user needs (ECLIPSE, 2023).

Figure 17 - MQTT basic model


Home/sensor/ temperature/ livingroom
Home/sensor/ temperature/ kitchen
MQTT Broker
Home/sensor/ temperature/ livingroom
Home/sensor/ temperature/ kitchen
Analytics
Mobile App
Database
Publishers
Subscribers

Source: Mishra et al. (2021)

Mosquitto is often used with web frameworks for developing IoT applications, as many of the frameworks offer support for using MQTT, expanding the possibilities for digital developments with web interfaces.

A web framework is a set of libraries, tools, and standards to develop web applications and services. The main goal of using frameworks is to facilitate the web development process with a pre-prepared structure that includes primary resources. Frameworks allow developers to focus on more complex activities, as a web framework already has URL (Uniform Resource Locator) routing, session management, authentication, database management, form management, and other resources. With these resources, developers can already create web applications without starting the code from scratch (THOUTAM, 2021).

2.5 Web Framework

Many web frameworks are on the market, ready for various programming languages. Web frameworks for the Python language include Django, Flask, Pyramid, CherryPy, Tornado, and others. Django is a high-level web framework that enables rapid web development, as it is designed to be easy to use and has a wide range of features that can assist developers (GHIMIRE, 2020; GORE *et al.*, 2021). Although it may not be considered the best among all frameworks, several advantages and features make Django a convenient option for developers, reinforcing the choice of this framework for the development performed in this thesis.

2.5.1 Django

Django comes with a wide range of built-in features that can be used to develop complex web applications. This web framework includes ORM (Object-Relational Mapping), automatic admin panel administration, user authentication, and URL management (DUISEBEKOVA; KHABIROV; ZHOLZHAN, 2021).

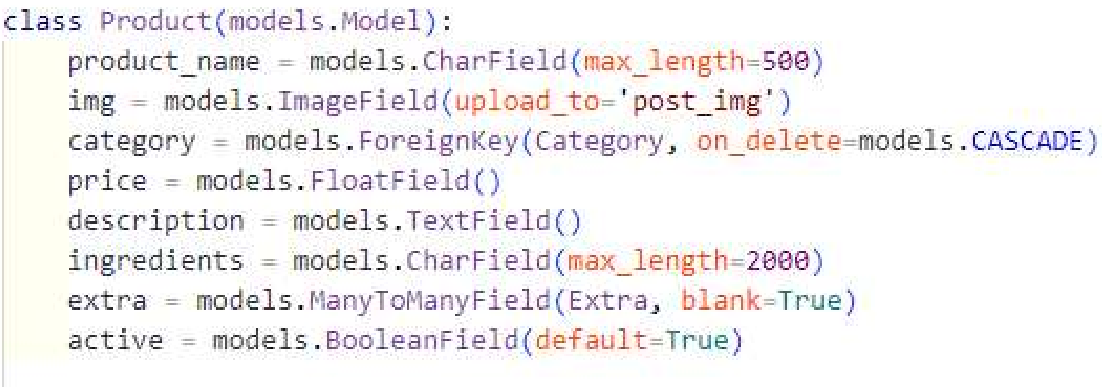
ORM is a software development standard that allows developers to use object-oriented language to interact with a relational database. In Django, the ORM is essential for developers to interact with the database more quickly and intuitively. The ORM is designed for developers do not have to manually write SQL language to insert, update, or retrieve data from the database (MELÉ, 2020). Instead, they can use Python objects and ORM methods for these tasks (Figure 18).

Django follows the MVT (Model-View-Template) architectural pattern, which is an approach that separates the responsibilities of the different parts of the system, making it more modular, easy to understand, and maintain (DUISEBEKOVA; KHABIROV; ZHOLZHAN, 2021; GORE *et al.*, 2021).

To start a new application in Django, it needs to create a "Project," which can have multiple "Applications" that serve different functions within the development to be carried out. An application is a set of Models, Views, Templates, and URLs.

Applications interact with the framework to provide specific functionalities and can be reused in various projects. Figure 19 shows the structure of the Django framework.

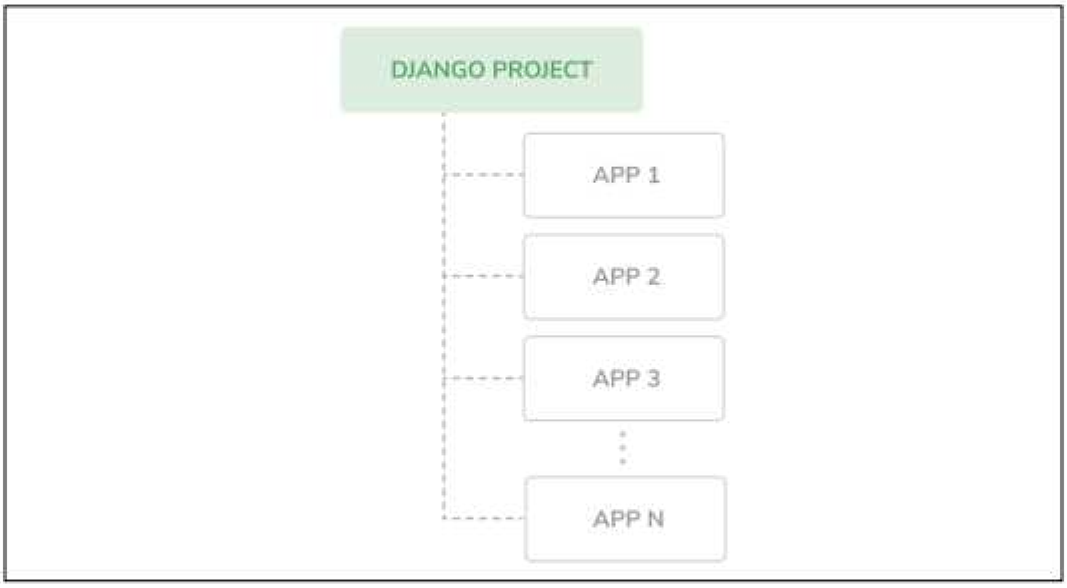
Figure 18 - Table creation with Django



```
class Product(models.Model):
    product_name = models.CharField(max_length=500)
    img = models.ImageField(upload_to='post_img')
    category = models.ForeignKey(Category, on_delete=models.CASCADE)
    price = models.FloatField()
    description = models.TextField()
    ingredients = models.CharField(max_length=2000)
    extra = models.ManyToManyField(Extra, blank=True)
    active = models.BooleanField(default=True)
```

Source: The Author (2023)

Figure 19 - Django project/ application structure



Source: Melé (2020)

Each created application has six files: admin.py, apps.py, migrations, models.py, tests.py, and views.py.

a) admin.py: location where models are registered for inclusion in the administration site;

b) apps.py: main configuration of the application;

c) migrations: directory with database migrations for the application;

d) models.py: information on the application’s data models, with the use of OOP in Python;

e) tests.py: directory to add tests to the application; and

f) views.py: all logical programming of the application will be created within this directory. Each views.py file receives an HTTP request, which will process and return a response to the URL.

Therefore, as stated in the previous paragraph, in the MVT pattern, the Model is responsible for accessing and manipulating the application’s data, and the View is responsible for the application’s logic, with Python functions and database access. The Template presents the data to the end user using HTML (Hyper Text Markup Language) and CSS (Cascading Style Sheet) code lines.

By using MVT, Django provides a clear structure that helps the developer organize the code, allowing the code to be understandable even in team projects, since a different application can be created for specific areas or pages of the web application (DUISEBEKOVA; KHABIROV; ZHOLZHAN, 2021).

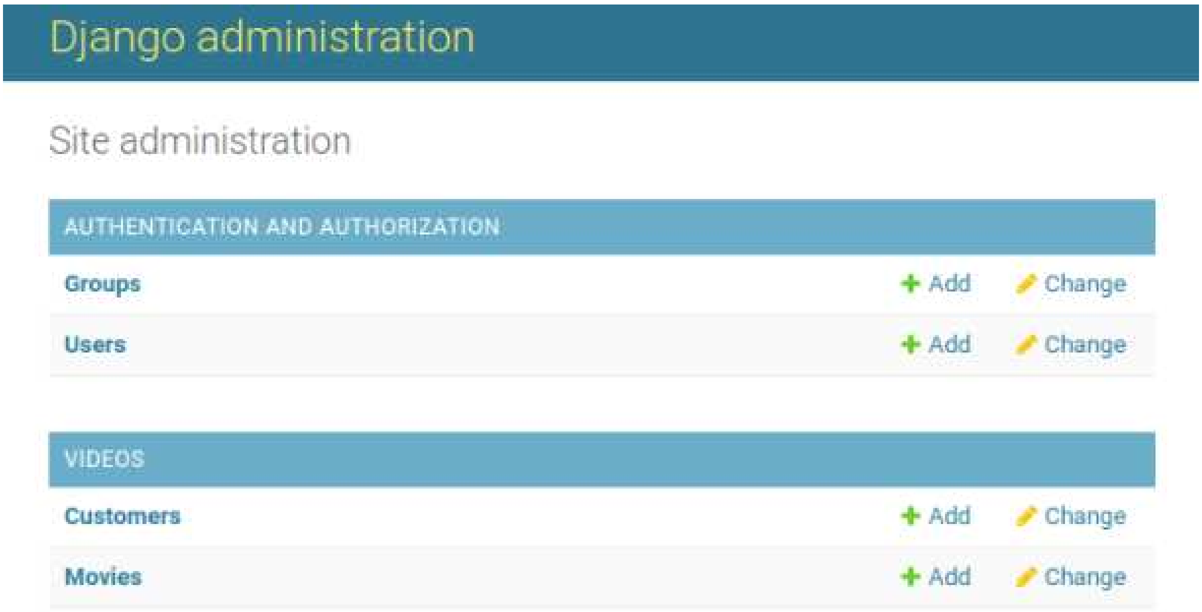
The automatic administration of Django is a pre-structured and native feature in Django for any project by which all application data can be managed. All tables and data can be managed through the Admin area (Figure 20). Access is made through the URL: localhost:8000/admin if the Django server runs on port 8000.

User authentication is an embedded feature of Django that provides an easy and secure way to manage user authentication in a web application. User authentication works through sessions and authentication tokens. When a user logs in, Django creates a session for that user and stores the session in the database. The authentication token is used to authenticate the user in each subsequent request. Several user models can be used to create users and manage permissions (MELÉ, 2020; SHAW *et al.*, 2021). The options for creating and changing new users and groups can also be seen in Figure 20.

The native URL management feature of Django allows developers to quickly and neatly manage the URLs of the application. This management is done through the "urls.py" file. This file contains a list of URL patterns that allows communication with the corresponding Python functions in the "views.py" file that will be called when the user accesses a particular URL.

Django has a very active and large community of developers, which provides high availability of support for framework users. This includes additional libraries and packages, detailed documentation, forums, and discussion lists. The framework is considered easy to use and learn, especially for developers who already program in Python. Django has clear syntax and well-written documentation that help and facilitate understanding of the entire functional structure (GORE *et al.*, 2021; MELÉ, 2020).

Figure 20 - Admin area



Django administration

Site administration

AUTHENTICATION AND AUTHORIZATION

Groups + Add ✎ Change

Users + Add ✎ Change

VIDEOS

Customers + Add ✎ Change

Movies + Add ✎ Change

Source: The Author (2023)

Django has several built-in security features, which make it easier for developers to create secure applications. These features include CSRF (Cross-Site Request Forgery) prevention, SQL injection prevention, and input validation.

CSRF is a security attack where an attacker can trick a legitimate user into sending malicious requests using the user's authentication. Django has a built-in security functionality that prevents this type of attack.

SQL injection prevention is also crucial to avoid attacks where attackers inject malicious SQL code into databases. Django has integrated features to prevent these attacks (MELÉ, 2020).

Django is scalable and designed to create large-scale web applications, which increases its usage by many large companies and organizations. This framework is very versatile and can also be used as a platform to build BI applications and create applications that collaborate with other BI applications.

2.6 Business Intelligence

According to Larson & Chang (2016), with the advent of Big Data and DA, the delivery of BI has been affected, as the increase in data speed has accelerated the need to transform all data into information.

BI is a set of technologies, tools, and practices that transform raw data into valuable, strategic information for companies. These transformations should bring insights about the organization, identify opportunities for improvement, and assist in anticipating market trends.

As Sharda et al. (2018) point out, the BI process can be divided into three pillars: data collection, organization and data analysis, and monitoring and control. Thus, the main objective of BI is to assist business decision-making. BI will not indicate the paths managers should follow but will generate reports and provide ways to evaluate information and bring insights to the company.

Applying BI to a company can bring several advantages and solutions to short and long-term problems, such as saving time and costs in identifying excessive expenses, discovering business opportunities based on market trends, greater accuracy in strategies and tactical plans when launching a new product, greater control over processes by visualizing end-to-end process characteristics.

Romero et al. (2021) conducted a study to analyze the current situation of BI technology in terms of positive impacts on the economic and commercial levels in terms of decision-making after the start of I4.0. The authors showed that digital technologies are fundamental development pillars for decision-making and prediction and also highlighted the importance that the integration of ERP (Enterprise Resource Planning), IoT, and BI technologies acquires, with high contribution in all organizational, operational, and managerial decision-making aspects.

The practical application of BI should be carried out through a BI tool, for which there are several possibilities, such as Tableau, Power BI, MicroStrategy, IBM Cognos Analytics, and Metabase, among others (MANUEL; NUNES, 2012).

The choice of the ideal tool depends on each organization's specific needs and the resources available. Some tools, such as Metabase, are open-source and valuable for small businesses or projects with limited resources.

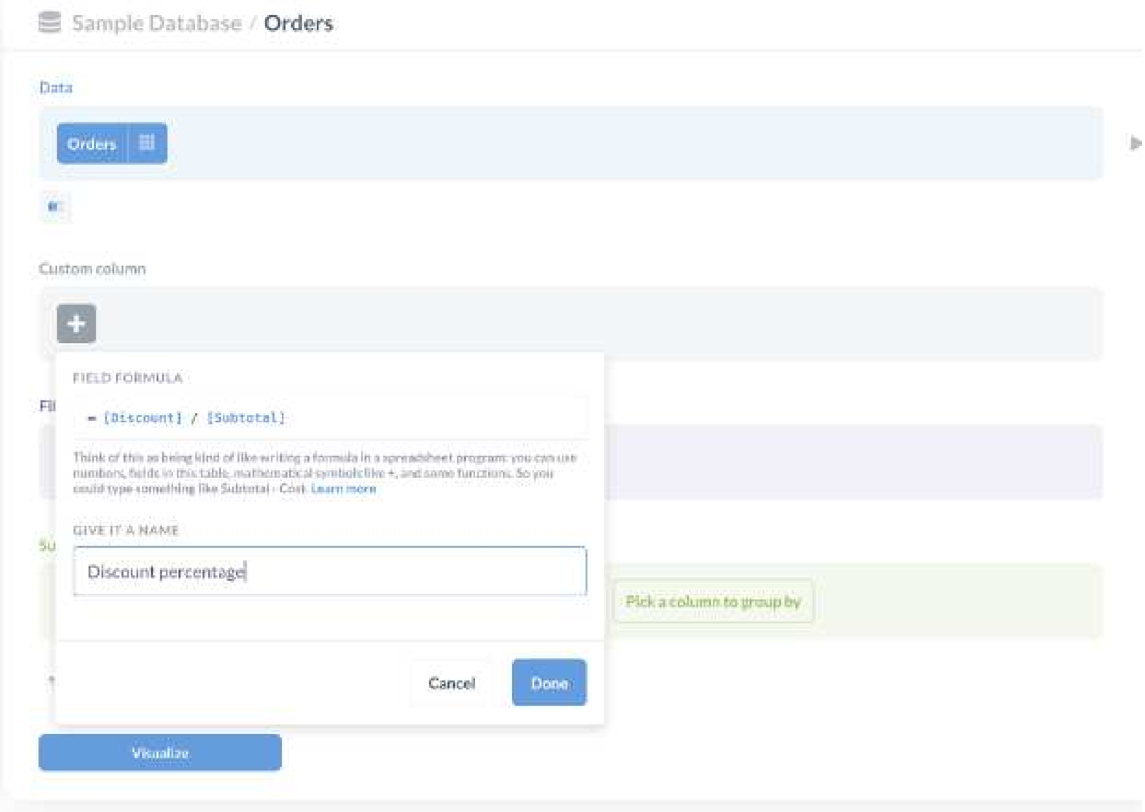
Although Metabase may be less sophisticated and complete than some proprietary tools, it can provide a basic BI solution for simple data analysis. BI tools can also be integrated into more extensive cloud computing or virtualization systems applications.

2.6.1 Metabase

Metabase offers a simple and intuitive interface for non-programmers to explore, visualize, and share data quickly. This tool also allows for creating Q&A (Questions and Answers) (Figure 21) with natural language on existing databases, enabling users to ask questions in English and receive graphical or tabular responses (JOVANOSKA; PETREVSKA NECHKOSKA; MANCHESKI, 2021). Several databases are compatible with Metabase, including MySQL, PostgreSQL, and SQL Server.

If users want to perform more complex or customized queries, in which it is impossible to obtain them with question tools, it is also possible to use code lines with SQL programming (Figure 22).

Figure 21 - Q&A in Metabase



Sample Database / Orders

Data

Orders

Custom column

FIELD FORMULA

= [Discount] / [Subtotal]

Think of this as being kind of like writing a formula in a spreadsheet program: you can use numbers, fields in this table, mathematical symbols like +, and some functions. So you could type something like Subtotal - Cost. Learn more

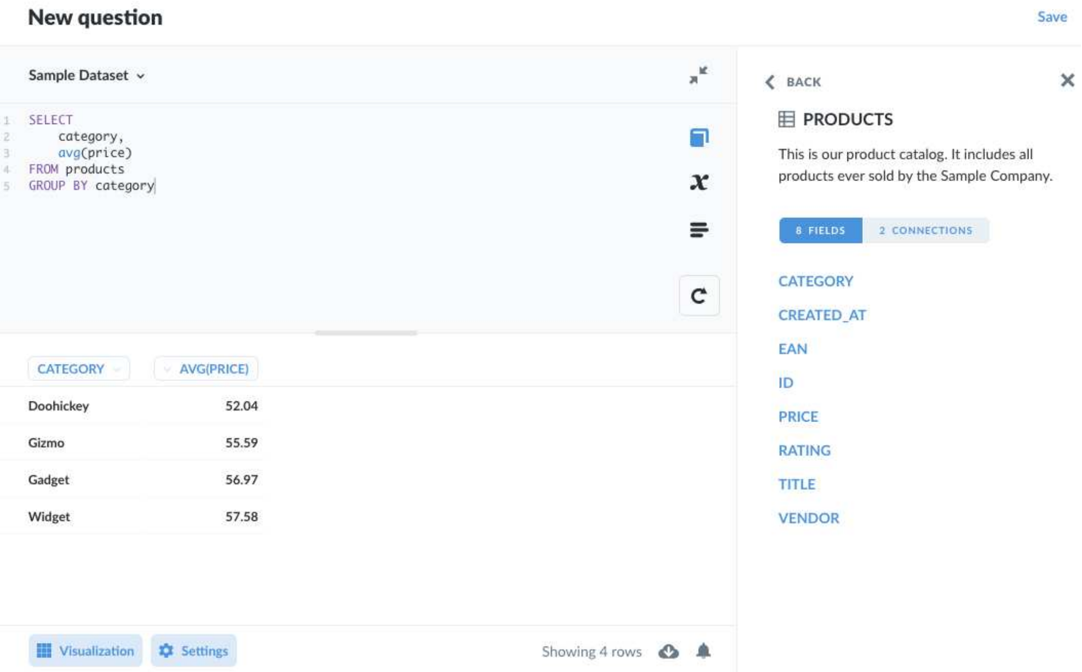
GIVE IT A NAME

Discount percentage

Source: Metabase (2023)

After using the question, filter, join, and subquery features, the user should select from various available chart types in the tool, such as line, bar, scatter plots, pie, tables, heatmaps, maps, and many others (Figure 23). Finally, after selecting and creating as many charts and queries as necessary, it is possible to gather them into a single dashboard (Figure 24).

Figure 22 - Writing SQL Queries



New question

Sample Dataset

```
SELECT
    category,
    avg(price)
FROM products
GROUP BY category
```

CATEGORY	AVG(PRICE)
Doohickey	52.04
Gizmo	55.59
Gadget	56.97
Widget	57.58

Showing 4 rows

PRODUCTS

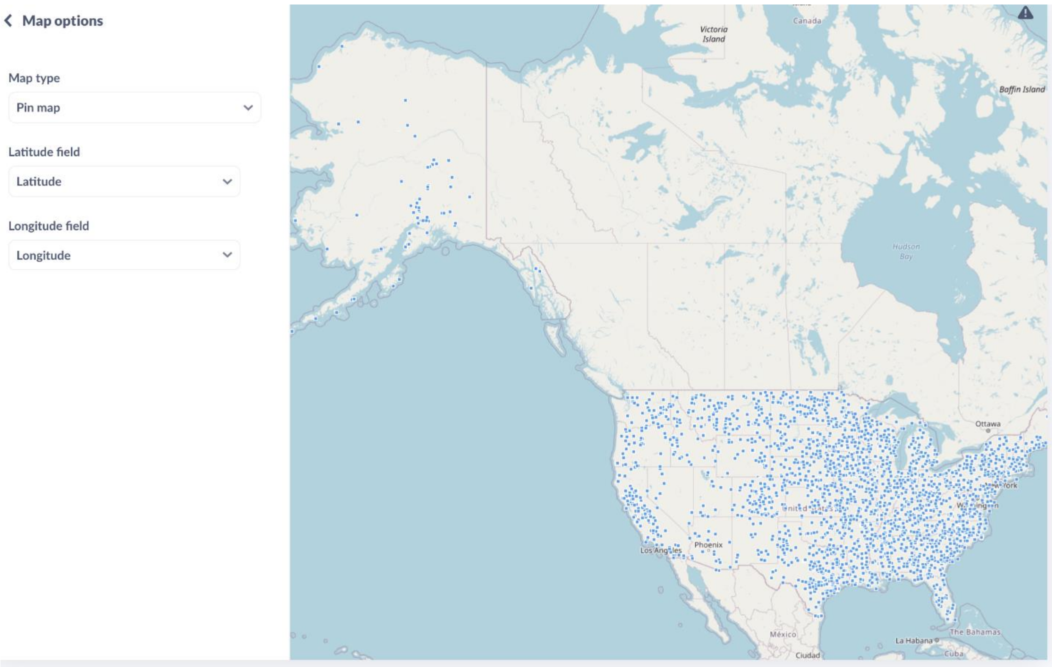
This is our product catalog. It includes all products ever sold by the Sample Company.

8 FIELDS | 2 CONNECTIONS

- CATEGORY
- CREATED_AT
- EAN
- ID
- PRICE
- RATING
- TITLE
- VENDOR

Source: Metabase (2023)

Figure 23 - Map Chart in Metabase



Map options

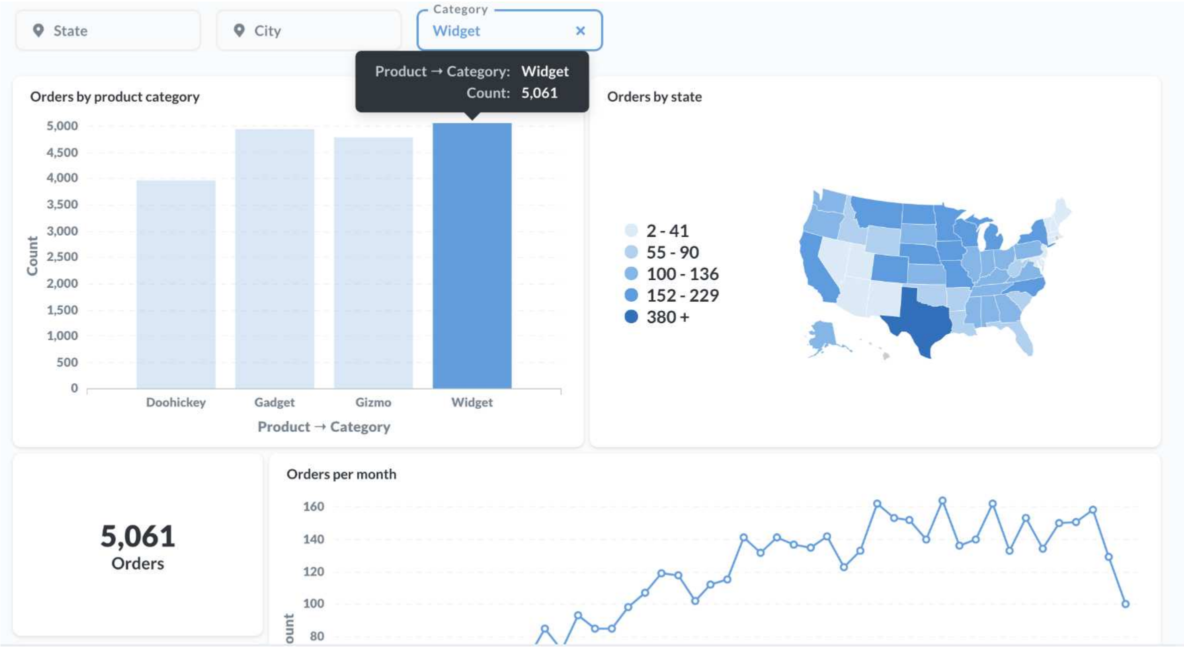
Map type: Pin map

Latitude field: Latitude

Longitude field: Longitude

Source: Metabase (2023)

Figure 24 - Dashboard example



Source: Metabase (2023)

Dashboards are graphical interfaces that visually and concisely present all relevant information for analysis. The Metabase dashboard is highly customizable, including charts, tables, maps, gauges, alerts, or other visual elements that make the results more precise and concise (ECKERSON, 2012; ELIAS, 2012).

Metabase can also be run in containers, such as Docker, as it is a tool that functions locally, greatly facilitating the deployment and management of the tool.

2.7 Virtualization and Containerization

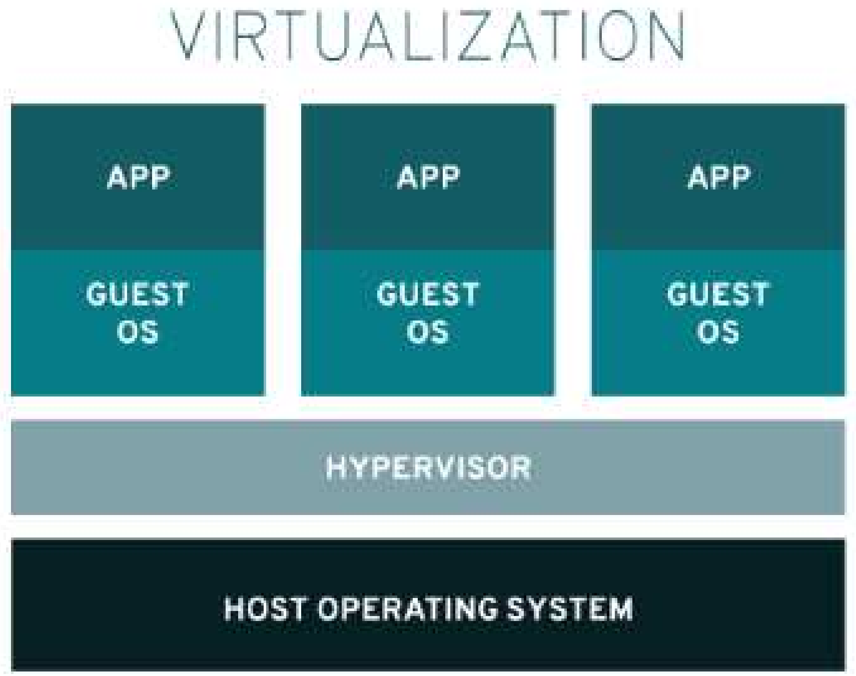
According to Xavier et al. (2013), virtualization technologies with various software solutions have become popular recently. The main benefits of virtualization include availability, isolation, hardware independence, and security. The virtualized resources are called VMs (virtual machines) and are seen as isolated execution contexts.

Virtualization can be achieved through virtualization software, such as VirtualBox, VMware, and Hyper-V, which create virtual machines, or through containers, such as Docker, which virtualize only the resources of the operating system, thus allowing multiple containers to share the same kernel of the primary operating system (SILVA, Vitor Goncalves da; KIRIKOVA; ALKSNIS, 2018). A kernel is the operating system's core, responsible for managing hardware resources and providing interfaces for programs to interact safely with hardware. Each operating system has its kernel, developed to work with the hardware and what the system needs (XAVIER *et al.*, 2013).

Virtualization is widely used to improve efficiency, scalability, and management of computing resources. Containers do not require an operating system, but this does not hinder them from running within a VM (CELESTI *et al.*, 2016). The main difference between virtualization and containers is how the operating system resources are shared between operations.

In virtualization, each virtual machine is an independent instance of a complete operating system with its resources and kernel. This instance allows applications to run on a single host with high hardware and storage consumption. For this virtualization, a hypervisor software layer enables multiple VMs to share a single set of physical resources. The hypervisor separates the physical machine's resources and distributes them to the virtual machines (VELTE; VELTE, 2009). Figure 25 demonstrates how the Virtualization structure with VMs works.

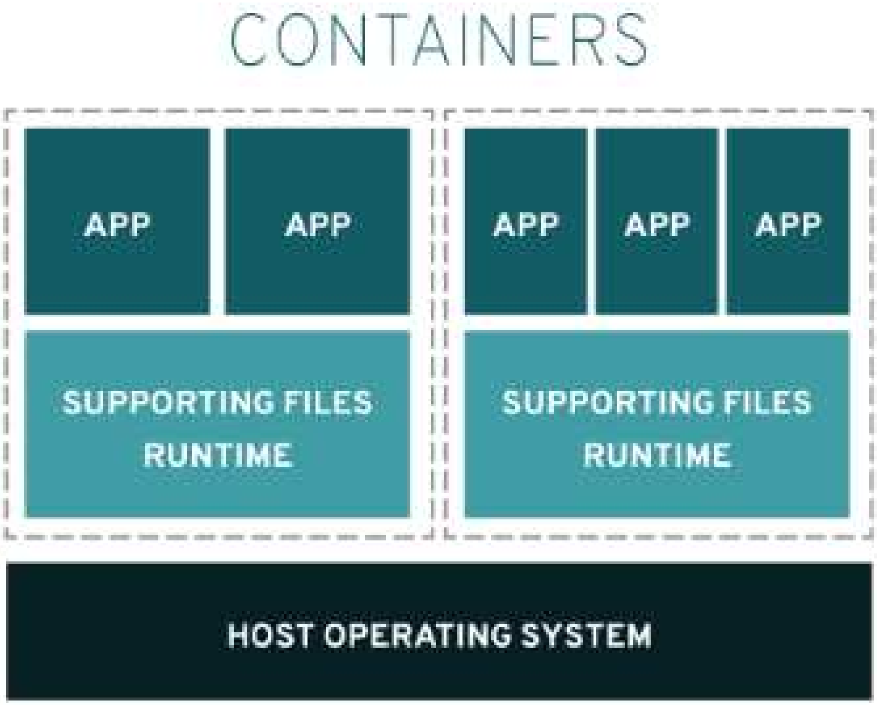
Figure 25 - Virtualization Structure



Source: Red Hat (2020)

In containers, applications share the same operating system kernel but have isolated environments where libraries and dependencies are installed locally. This isolation makes containers lighter and more efficient regarding system resources, allowing a single physical host to run multiple applications without needing various virtual machines (TURNBULL, 2014). Everything the container has is kept in an image, i.e., a code file that includes all libraries and dependencies. Figure 26 demonstrates how the container structure works.

Figure 26 - Container Structure



Source: Red Hat (2020)

In an application, there can be multiple containers, which is why container management platforms were created to allow for the creation, packaging, distribution, and management of applications in isolated environments. These platforms include Kubernetes, Apache Mesos, Red Hat Openshift, GKE (Google Container Engine), Amazon ECS (Elastic Container Service), and Docker (MCKENDRICK; GALLAGHER, 2017). Each platform has its characteristics, advantages, and disadvantages, and the choice will depend on the requirements and needs of the project.

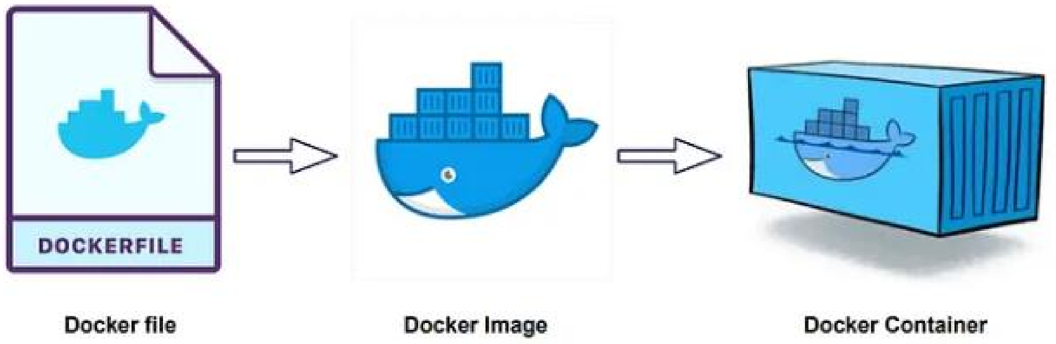
2.7.1 Docker

One of Docker's main strengths compared to other platforms is its portability. Docker allows it to be reinstalled with the same configurations on different machines or even within a cloud computing system (ACHARYA; SUTHAR, 2022)2).

The main components of a Docker architecture include Docker Image, Docker Container, Docker Hub, and Docker ComposDocker Image is a template, i.e., an image that contains all the necessary data and metadata to run containers. These images can be ready-made on the Docker Hub website or created from a file called Dockerfile, which describes all the necessary steps to build an image. Once created, the image is sent to be stored in the Docker hub repository, where it can be shared with other users (DOCKER, 2023; MCKENDRICK; GALLAGHER, 2017).

The Dockerfile is a text file that defines the details of a container's customized image (Figure 27). The Dockerfile can contain instructions detailing the necessary steps in creating the image for the container, including environment configuration and installation of dependencies and applications (TURNBULL, 2014).

Figure 27 - Container configuration



Source: The Author (2023)

The Docker Compose tool creates, defines, and runs Docker applications with multiple containers. Docker Compose allows defining a set of services and their dependencies in a YAML file, simplifying the process of creating all containers with just one command in the terminal that should execute this file (MCKENDRICK; GALLAGHER, 2017; TURNBULL, 2014).

The Docker Compose will create a unique network in an isolated environment so all containers can communicate. Among the configurations available for this file is Docker Volume, a feature that allows creating and managing data storage volumes for containers. A Docker volume is a directory or file that is stored outside the container and can be shared among multiple containers. This volume allows data to persist even if the container is removed or recreated. In addition, volumes allow containers to share data, which can be very useful (MORAVCIK; KONTSEK, 2020; ZEROUALI *et al.*, 2019).

For Docker to create an isolated network, the Docker network configurations must appear in the Docker Compose file, as it is a functionality that will allow the creation and configuration of isolated virtual networks and enable communication between the involved applications.

A bibliographic review of the essential concepts for understanding this research was presented in this section. Next, the methodological aspects for the elaboration of this research will be demonstrated, as well as the methodology used by the authors to meet the proposed objectives.

3 METHODOLOGICAL ASPECTS

This section discusses the research characterization (section 3.1), the methodological procedures (section 3.2), and the research limitations (section 3.3).

3.1 Research Characterization

According to (GIL, 2002; MORAIS; BOIKO, 2014; SILVA; MENEZES, 2005), this research is quantitative, applied, field research, bibliographic, and prescriptive. Research can be structured in Research Approach, Research Types, and Research Techniques.

From the Research Approach standpoint, this study can be characterized as quantitative since its environment is a direct source for data collection, interpretation of phenomena, and meaning attribution.

Research Types classification encompasses the primary research objective, the technical procedures, and the research design. In this context, the main research objective is characterized by its ends, and regarding the technical procedures and research design, it is classified by its means.

As for its ends, this study can be characterized as applied since its goals will generate practical knowledge and target specific problem-solution.

As for its means, this study can be characterized as bibliographic, field research, and prescriptive. It is bibliographic due to its study based on published papers and books. Field research is due to empirical investigations carried out at a place where a phenomenon occurs.

Prescriptive model research is designed to make solutions easier by leading problem solvers to the solution as efficiently as possible. In this study, models are developed to explore and predict variable behavior.

From the Research Techniques standpoint, it can be characterized as data analysis techniques, as once all the data has been captured and processed, it must be analyzed to bring meaning to the research.

3.2 Methodological Procedures

The current project is conducted based on DSR, fundamentally a problem-solving approach. Although DSR has been characterized as a paradigm for the Information Systems discipline, Engineering disciplines accept it as a valid and valuable research methodology since engineering research establishes value for practical problem-solving (HEVNER; CHATTERJEE, 2010; PEFFERS *et al.*, 2007)7). The DSR paradigm seeks to create innovations, or artifacts, in which information systems analysis, design, implementation, management, and use can be effectively and efficiently performed to solve real-world problems (HEVNER; CHATTERJEE, 2010). These artifacts, which define ideas, capabilities, techniques, practices, and products, can be complex and extend problem-solving boundaries by providing intellectual and computational tools.

In the engineering literature, many authors, such as (Eekels & Roozenburg (1991), have denoted the need for a common DSRM (Design Science Research Methodology). Bayazit (2004) believed the design could be coded and build system instantiations due to the research. Through his work, he defined six steps for DSR:

a) Programming;
b) Data collection and analysis;
c) Summary of the objectives and analysis results;
d) Development;
e) Prototyping; and
f) Documentation.

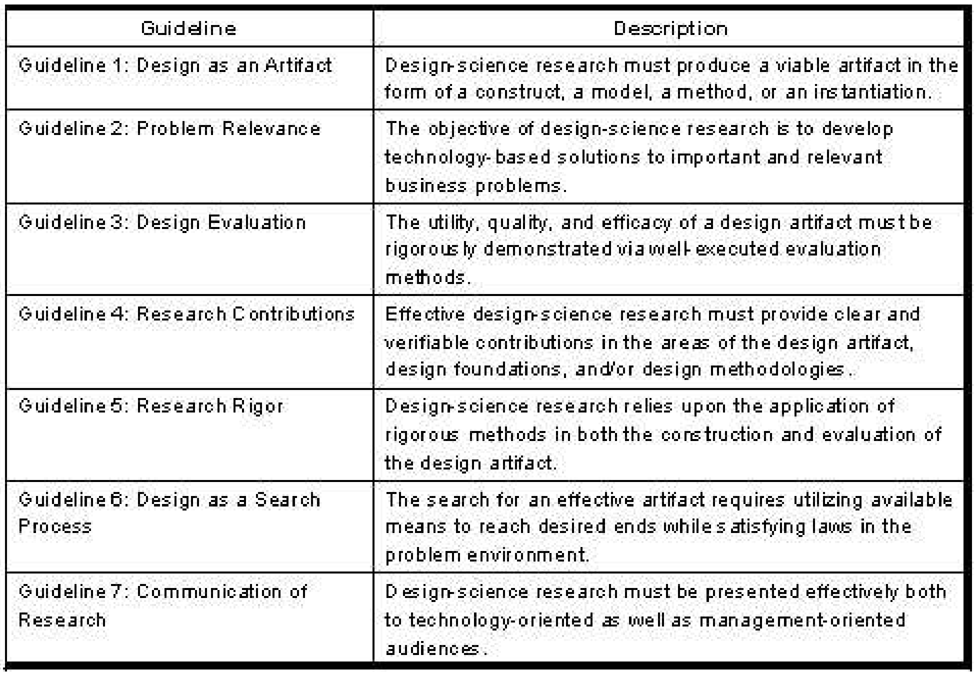
Thus, Bayazit (2004)stated that designers could systematically approach design problems to progress and achieve specific solutions.

Hevner & Chatterjee (2010) highlight seven guidelines (Table 2) derived from the building and application of an artifact over which knowledge and understanding of a design problem and its solution are achieved. The purpose of establishing these guidelines is to assist researchers, editors, and readers in understanding DSR requirements.

The present work proposes constructing a system with tools and instantiations to address and reach the main objective. Constructs provide the vocabulary and symbols used to define problems and solutions, significantly influencing how problems and tasks are produced (BOLAND JR., 2021). Instantiations may be intellectual or software tools designed for the information systems development process (HEVNER; CHATTERJEE, 2010).

In addition to these guidelines, the design research literature is rich in ideas on how to conduct research. However, Peffers et al. (2007), through important prior literature and based on DSR principles, sought to use a consensus-building approach to create the DSRM that would become a commonly accepted framework based on DSR. This framework, consisting of six activities, which are used for the development methodology of this project, is shown in Figure 28.

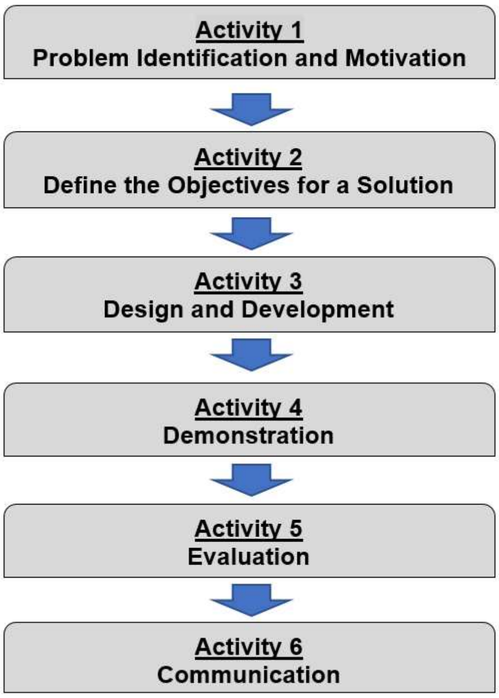
Table 2 - DSR Guidelines



Guideline	Description
Guideline 1: Design as an Artifact	Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.
Guideline 2: Problem Relevance	The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
Guideline 3: Design Evaluation	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
Guideline 4: Research Contributions	Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
Guideline 5: Research Rigor	Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
Guideline 6: Design as a Search Process	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
Guideline 7: Communication of Research	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

Source: Hevner & Chatterjee (2010)

Figure 28 - DSRM Activities



Source: Peffers et al. (2007)

Each of these six activities makes up an essential role for a DSR. Thus, each will be thoroughly explained according to its main concepts and integrated into each step of this study.

3.2.1 Problem Identification and Motivation

Define the research problem and validate the value of a solution. Resources for this activity contain knowledge of the state of the problem and the importance of its solution (PEFFERS *et al.*, 2007);

The creation of this artifact begins with identifying the application context of the solution in the manufacturing industry. Then, in the first activity of the DSR methodology some research demands were identified through a bibliometric and systemic analysis (FERNANDES *et al.*, 2019), in which opportunities for research related adapting PPC methods with digital technologies were identified. This analysis has brought together four unfilled gaps and the advances related to DS and ML techniques in improving manufacturing processes, as explained in Chapter 1.

The VSM method was chosen for the project development due to its potential and importance in identifying and optimizing value streams. However, it is essential to highlight that the application of VSM presents several challenges to be addressed. One of the main challenges is collecting and analysing data required to map the processes and identify value-added and non-value-added activities. The correct interpretation of the collected data and the identification of value-added and non-value-added activities require careful analysis. This interpretation can be complex and requires a deep understanding of the organization's processes and value streams.

The uncertainty regarding the effects of proposed changes in the processes also brings numerous difficulties to this development. However, using digital technologies, such as DS techniques and simulation based on ML, can help address these challenges. Data Science techniques enable the efficient and fast handling of large volumes of data, allowing for comprehensive data analysis and extraction of valuable information for value stream mapping. Through ML-based simulation, it is possible to create hypothetical scenarios and conduct virtual tests of the proposed changes. By providing relevant data to the simulation model and training it with ML algorithms, it is possible to anticipate the impact of these changes on the processes before implementing them. This simulation allows for a more precise and informed evaluation of the expected results.

Therefore, although VSM is a powerful tool, its application requires careful consideration and a continuous commitment to overcome challenges and achieve the expected benefits.

3.2.2 Define the objectives for a solution

Assume the aims of a solution from the problem definition and through what is feasible. Resources for this activity contain knowledge of the state of the problem and the solutions and effectiveness (PEFFERS *et al.*, 2007).

In the second activity, it is essential to elaborate on the expected objectives for problem-solving since it is related to investigating different methods, algorithms, and platforms compared to what has already been used.

The study contributes by developing a solution for two out of four identified gaps related to additional features, programming, and innovative method adaptation. Although VSM benefits organizations, its development still requires a significant amount of time since the necessary data is entered manually, which increases the possibility of errors in the analysis.

Daily visits were made to the partner company with analyzes of the last flow mapping carried out in the line of operations under study and observations in the field among all the activities carried out by the operators.

Several meetings were held with the engineers of the partner company. After many conversations, it was established that the artifact would be developed in a flexible web system for access on any device or location and that it would need to adapt to the provided scenario of a sequence of operations involving a integrated fuel tank supply and assembly process. The company had traditionally performed the mapping, and thus the solution should reconstruct the method's usability with dynamic functionalities and the aid of AI.

In this way, the artifact should bring advantages to the development of the mapping, with the capture of data in a digital environment, the transformation of data into the current state mapping, as well as creating and providing new insights with the mapping of the future state, with intelligent predictions that could simulate alternative solutions to be applied in the processes.

3.2.3 Design and Development

Create the artifact and determine the functionality of the artifact and its architecture (HEVNER; CHATTERJEE, 2010; PEFFERS et al., 2007).

The project and development of the artifacts which assist in solving the problem are developed in this step.

Therefore, the proposed system is based on virtualizing six applications using the container orchestrator tool Docker Compose. A container orchestrator tool assists in managing and automating the deployment, scaling, and management. Docker Desktop will manage multiple containers ensuring that both run reliably and efficiently. Docker has features that facilitate the installation of multiple applications, providing much more security to the user.

Node-RED, Mosquitto, MySQL, Django web Framework, Metabase, and Jupyter Notebook occupy the six containers (Figure 29).

MySQL is a relational database management system based on SQL (Structured Query Language), which stores data of varying sizes. MySQL will store the data tables used by all other applications.

Django is a web framework in the Python language, configured with the aid of the VSCode (Visual Studio Code) IDE for the significant development of the visual and FP of the current state mapping and future state mapping. The Python language is used as the back-end, but JS, HTML, and CSS will be used for the front-end development of the application. Django provides a component-based development framework that makes it easier to create complex web applications with less effort and code. According to (HOLOVATY; KAPLAN-MOSS, 2009), among the main advantages of using VSCode for Django programming assistance are:

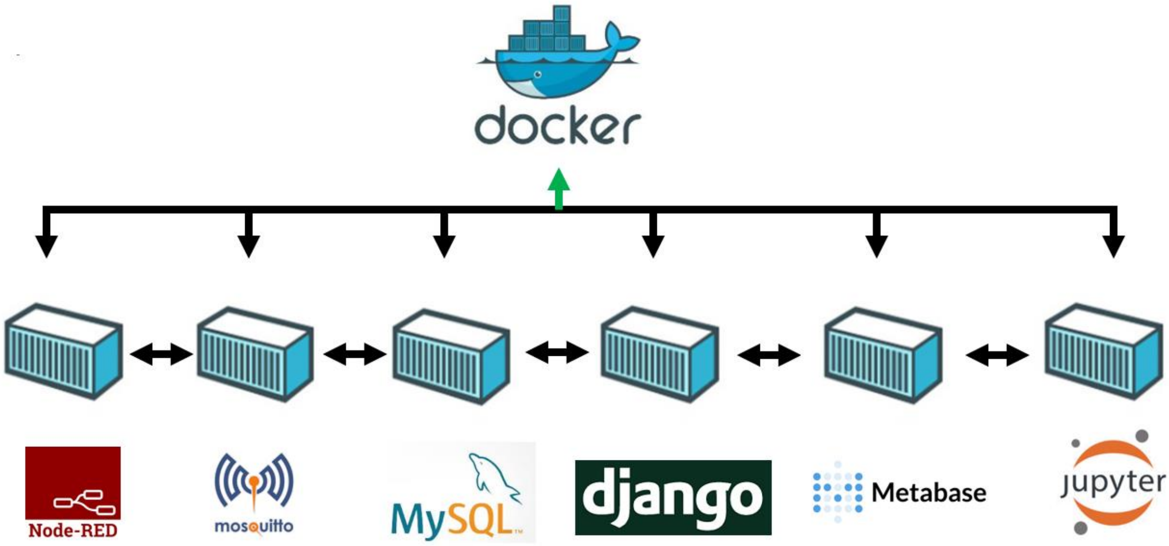
a) Python support: VSCode has excellent support for Python, allowing for more productive and efficient Django application development;

b) Integrated debugging: VSCode includes a built-in debugger for Python, allowing debugging of Django applications directly from the editor without the need to open a terminal or use external tools;

c) Extensibility: VSCode is highly extensible, allowing the installation of many useful plugins and extensions for web development in Python, such as Jupyter Notebook support, among others; and

d) Ease of use: VSCode is a modern and user-friendly code editor with an intuitive interface and many useful features to make development more productive and enjoyable.

Figure 29 - Isolated applications communicate with each other within Docker



Source: The Author (2023)

In this development, VSCode will access the data tables in MySQL, but the data must be manually entered for MySQL to have data tables. The visual framework Node-RED is used with the MQTT Broker Mosquitto to facilitate data entry.

Since Node-RED is a visual programming platform for IoT, it makes it easier to create automation flows. Among its extensive library of features, there is the ability to read CSV (Comma-Separated Values) files in local files on the user's device and then send them to an MQTT broker, which in this case is the Mosquitto broker already associated with a container inside Docker. Through the Node-RED nodes, the broker can receive the data and redirect it to MySQL; i.e., Node-RED is only an intermediary between the broker and MySQL.

Using Docker, Django, Metabase, and Jupyter, have access permissions to any available tables. Metabase complements the mappings created through Django. As a data analysis tool that can generate various visualizations and detailed reports, it is crucial to help organizations make strategic decisions.

Django is the primary framework for developing the current and future state mapping, including the ML algorithms that will be essential to the development of future mapping. Nevertheless, since Jupyter Notebook is an interactive programming tool that allows for demonstrating the results line by line, the Jupyter container is included in Docker for two reasons. First, to verify line-by-line testing of the ML algorithms before effectively binding them to the Django web framework. Second, it will be used to analyze ML performance metrics, allowing for selecting the best learning model for each new table of data inserted. For testing and results of metric analysis, the interactivity and result visualization in Jupyter were considered more effective for the complete development of this application.

In structuring the project methodology, the ETL (Extract, Transform, Load) process was used to organize the importance of each application in achieving the overall project objective.

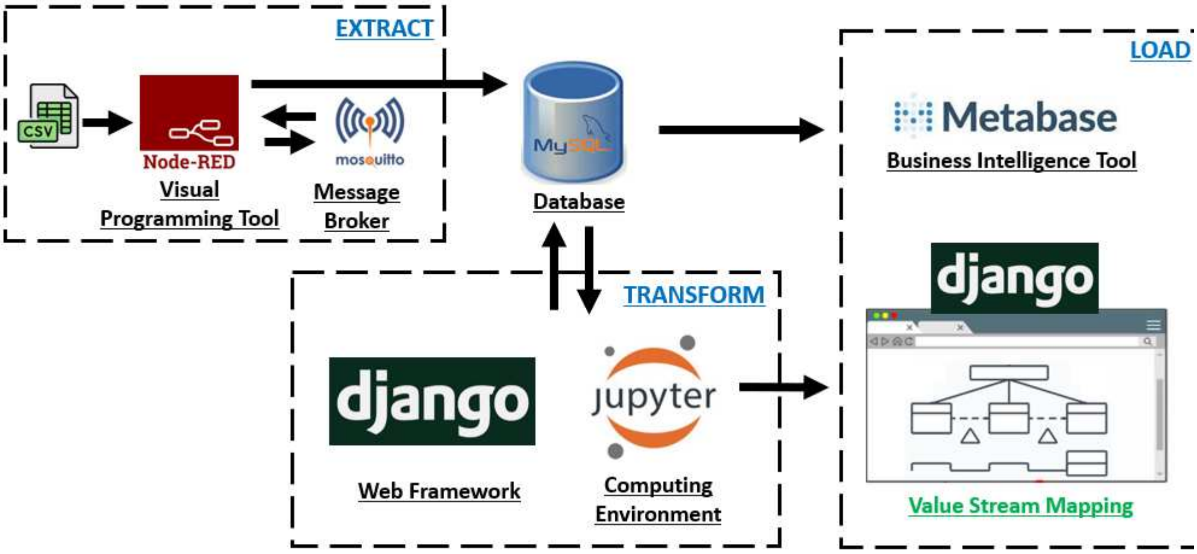
According to Diouf et al.(2018); Mukherjee & Kar (2017), ETL is essential because it helps improve the quality and consistency of data, as information is cleaned, standardized, and enriched before being stored. ETL allows analysis and business teams to work with accurate and reliable information essential for decision-making and achieving project goals.

The ETL process usually begins with extracting data from various sources such as databases, files, or web services, which are then transformed into a standard format, including data cleaning and processing, conversion, and aggregation for analysis. Finally, the data is loaded into a repository where users and applications can access and use it.

The classification of applications for this project within the ETL approach can be seen in Figure 30. The Extract comprises Node-RED and Mosquitto, as these applications access data in the CSV file and transfer it to the MySQL database. Transform is performed in the Jupyter Notebook and Django web framework as they explore DS and ML resources, capturing data stored in MySQL and allowing the development of functions in Python, JS, CSS, and HTML programming languages to extract new insights and patterns from the data. Load consists of the Metabase application and Django again, as Django renders web programming lines into HTML, enabling its visualization by any browser. Metabase accesses data directly in MySQL and allows the creation of different dashboards in various dynamic correlations for data visualization.

MySQL database is present in all stages of the ETL, as it connects all activities of all applications in this development.

Figure 30 - ETL for VSM Application Development



Source: The Author (2023)

A significant differential in this development is the adaptability of the mapping to different production lines, flexibility with an intuitive and easy-to-operate user interface, efficiency in providing predictions of future results already adapted to the mapping structure, ease of installation, and data updates.

3.2.4 Demonstration

Demonstrate the artifact used to solve one or more problem instances. Practical knowledge of using the artifact to solve the problem is necessary (HEVNER; CHATTERJEE, 2010; PEFFERS *et al.*, 2007).

The demo stage is the actual system test upon completion. Successful application of this artifact will be used as proof of concept. Therefore, the results of this test will be used to adjust any parameters to make it as appropriate as possible to ensure its effectiveness and efficiency.

The demonstration of this system will primarily take place via a web browser using the Docker Desktop virtualization platform. This approach offers the advantage of easy accessibility and compatibility across different devices and operating systems. By leveraging Docker's containerization technology, the system can be efficiently deployed and scaled, providing a seamless user experience. The target implementation of this system is within a multinational agricultural and construction equipment company located in Curitiba, Brazil. This context will enable real-world testing and evaluation of the artifact's effectiveness in addressing the company's specific needs and challenges.

3.2.5 Evaluation

Detect and evaluate how the artifact maintains a solution to the problem. This evaluation is made by comparing the objectives of a solution with the actual results detected by using the artifact in the demonstration (PEFFERS *et al.*, 2007).

This step develops metrics to evaluate the developed system’s efficiency, reliability, and adaptability. A satisfaction survey will be provided to the company’s engineers for the VSM Application, based on Vernadat (2014). This survey uses four options for the answer: (1) dissatisfied; (2) improvements are required; (3) satisfied; (4) very satisfied. This survey has questions related to mainly evaluating consistency, usability, accuracy, reliability, performance, and flexibility.

In addition to the questionnaire applied directly to the engineers of the partner company, a comprehensive analysis of the system in terms of generality, efficiency, and applicability is developed.

3.2.6 Communication

Communicate the problem’s importance, accuracy, usefulness, and effectiveness to appropriate researchers and practitioners (PEFFERS *et al.*, 2007).

The last activity is the communication of the analyses and results obtained. This communication occurs through a thesis elaboration and four articles published in International Journals.

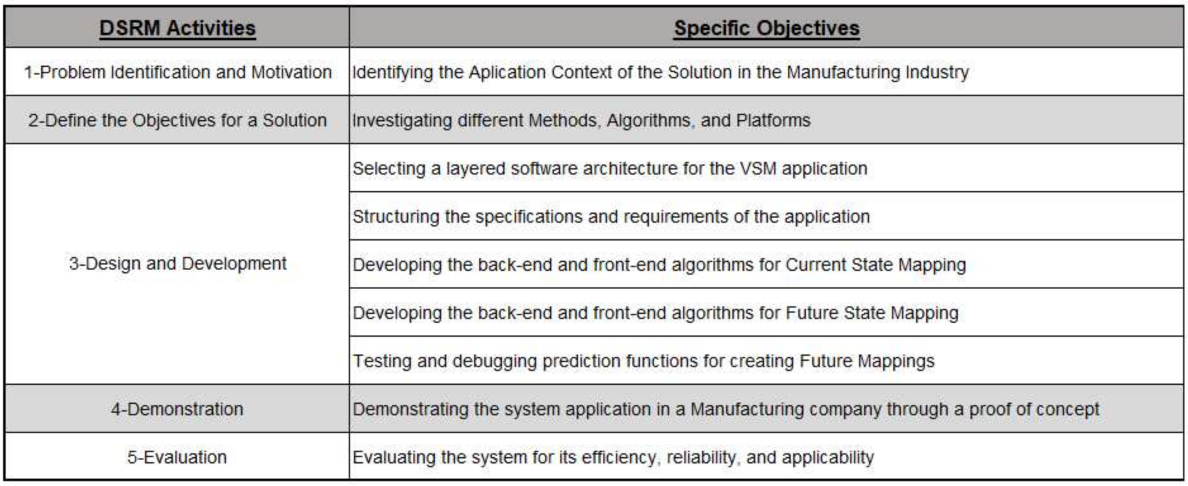
The methodology used for this research is developed according to the six DSRM activities. Thus, it is appropriate to correlate the specific objectives of this research (section 1.1) with the six activities from the DSRM (Table 3).

Problem Identification and Motivation and Define the Objectives for a Solution are activities related to the first two specific objectives: identifying the application context of the solution in the manufacturing industry and investigating different methods, algorithms, and platforms.

Design and Development is an activity related to five specific objectives: selecting a layered software architecture for the VSM application, structuring the specifications and requirements of the application, developing the back-end and front-end algorithms for Current State Mapping, developing the back-end and front-end algorithms for Future State Mapping, and testing and debugging prediction functions for creating Future Mappings.

A demonstration is an activity related to demonstrating the system application in a manufacturing company. Finally, Evaluation is an activity related to evaluating the system for its generality, efficiency, and applicability.

Table 3 - Correlation DSRM Activities and Specific Objectives



DSRM Activities	Specific Objectives
1-Problem Identification and Motivation	Identifying the Aplication Context of the Solution in the Manufacturing Industry
2-Define the Objectives for a Solution	Investigating different Methods, Algorithms, and Platforms
3-Design and Development	Selecting a layered software architecture for the VSM application
	Structuring the specifications and requirements of the application
	Developing the back-end and front-end algorithms for Current State Mapping
	Developing the back-end and front-end algorithms for Future State Mapping
	Testing and debugging prediction functions for creating Future Mappings
4-Demonstration	Demonstrating the system application in a Manufacturing company through a proof of concept
5-Evaluation	Evaluating the system for its efficiency, reliability, and applicability

Source: The Author (2023)

3.3 Limitations

Although the project follows all the guidelines of the DSR methodology, some limitations must be reported in the development of the model, demonstration, and evaluation.

In the development of the model, real-time data was not used as there are no sensors for this collection, and there are no data collection routines in the partner company. Therefore, since the amount of actual data available is insufficient for the development of the application, particularly for predicting future state mapping, the data used is only based on a sample of actual data obtained.

Data capturing can be done in real-time, but the predictions utilize a dataset that is not necessarily updated with each process execution.

4 RESULTS AND DISCUSSION

In this chapter, the main results of the development of this project are presented, with the creation and determination of the functionality and architecture of the artifact, the demonstration of the artifact in solving a problem, as well as the evaluation and communication of the applied solution.

4.1 Model Development

This section describes the complete process of building a dynamic web system for current and future state mapping by integrating various applications in a locally isolated network.

This integration became possible through the containerization of each application involved in this system and the creation of a development environment for the execution of each application. Thus, Docker was the software used to automate and manage each application's access, individually or in combination.

The containerization and integration of systems into a single network and installing Docker Desktop were first required. Since this is the initial development of this application and required various tests for its completion, Docker Desktop was selected instead of Docker Engine, the latter being exclusive to use on servers, i.e., ideal for use in production deployment. Docker enables the virtualization of each application with the help of complete images for each one of them, i.e., the containerization of the applications is only possible with access to their images or static representations of their configurations and dependencies.

The images can be created, but it is also possible to download images already pre-built by other developers, which can be found in the Docker Hub repository. In this project, it was possible to download the images of the MySQL database, the Node-Red visual programming platform, the Mosquitto message server, the Metabase BI platform, and the Jupyter Notebook web application.

In building an official Python 3.8 image linked to the initialization commands of the Django framework, a Dockerfile was used for this development. This file contains a base script linked to a file named "requirements.txt" in the working directory "/app/project" and thus installs Python and configures the environment with all necessary dependencies to the database. After installing these dependencies, it executes the installation of the MySQL client, which was used to connect to the database. Finally, it copies all files from the build directory to the working directory and exposes port 8000 to allow access to the web system from the network.

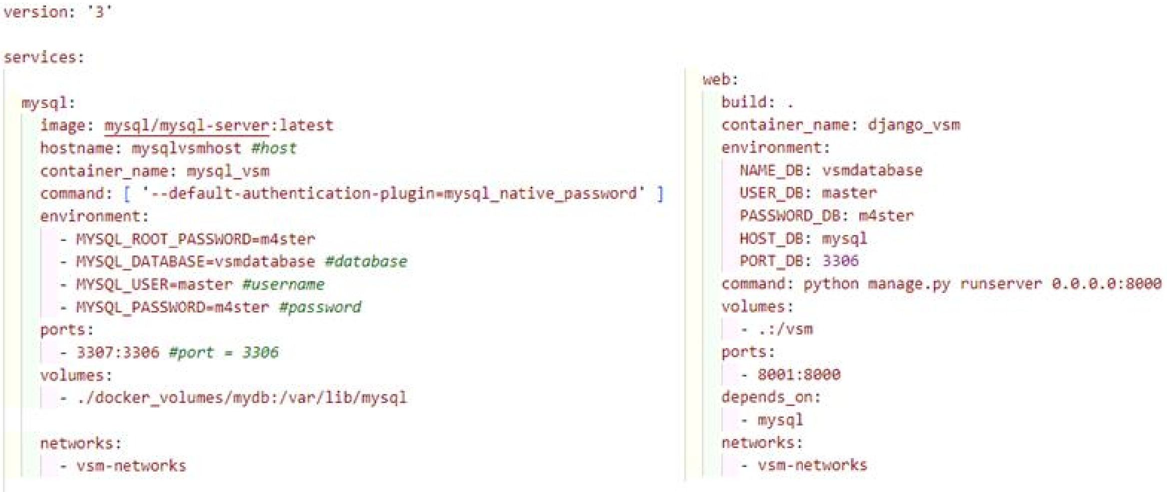
The Dockerfile used primarily to create the image linked to the execution of the Django web framework is complementary to the Docker Compose file since only the Docker Compose file allows for the complete configuration of all six containers with their respective ports and volumes (Figure 31, Figure 32, and Figure 33). The volumes allow for the sharing of files between the host and containers and allow data to persist even if the container is removed. In this development, the "mysql" container was linked to port 3007, the "web" container was linked to port 8001, the "mosquito" container was linked to port 1885, the "nodered" container was linked to port 1882, the "metabase" container was linked to port 3002, the "jupyter" container was linked to port 8890, and an isolated network called "vsm-networks" was created to allow communication between all of them.

Both files were gathered in a folder, and the "docker build" and "docker compose up" commands were executed through the command prompt. The "docker build" command was used to read the instructions in the Dockerfile, while the "docker compose up" command was used to start the services described in the dockercompose.yml file. This command created and started the containers, network, volumes, and all the configurations described in the file. Upon completing this process, all six containers were created and became manageable through the Docker Desktop platform (Figure 34).

Once all the containers became active on the platform, the next step was to configure each application for the web VSM with dynamic and intelligent features improved for the integration that Docker enabled to achieve the best possible result.

The Node-Red platform and the Mosquitto message server allowed for the operationalization of data input in the development environment. In this case study, the company provided data on an integrated fuel tank supply and assembly process, in a single Excel file. This process under study consists of 13 operations: Inbound Delivery, Typing, Unloading, Checking, Storing, Requesting, Order Picking, Loading (from inventory), Transshipment, Kitting, Loading (from factory), Partial Assembly, and Assembly. In each operation, the values of Changeover Time in minutes, Cycle Time in minutes, Availability in minutes, FPY (First Pass Yield) in units, number of operators in units, and Waiting time in minutes were measured.

Figure 31 – Docker Compose Programming Code for Mysql and Web

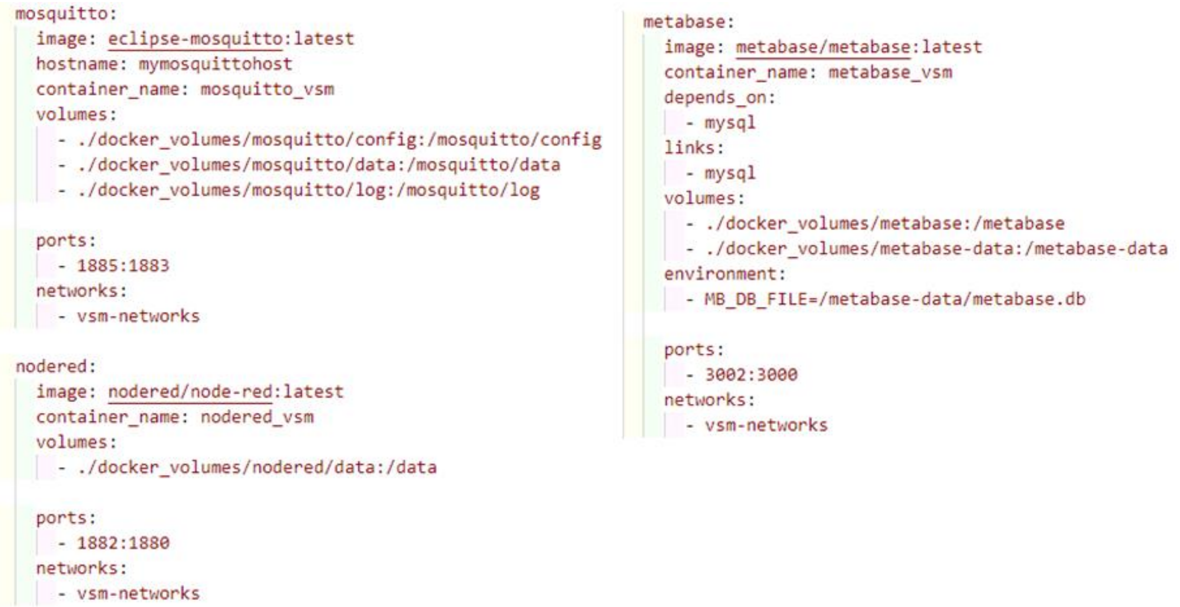


```
version: '3'

services:

  mysql:
    image: mysql/mysql-server:latest
    hostname: mysqlvsmhost #host
    container_name: mysql_vsm
    command: [ '--default-authentication-plugin=mysql_native_password' ]
    environment:
      - MYSQL_ROOT_PASSWORD=m4ster
      - MYSQL_DATABASE=vsmdatabase #database
      - MYSQL_USER=master #username
      - MYSQL_PASSWORD=m4ster #password
    ports:
      - 3307:3306 #port = 3306
    volumes:
      - ./docker_volumes/mydb:/var/lib/mysql

    networks:
      - vsm-networks
```

```
  web:
    build: .
    container_name: django_vsm
    environment:
      NAME_DB: vsmdatabase
      USER_DB: master
      PASSWORD_DB: m4ster
      HOST_DB: mysql
      PORT_DB: 3306
    command: python manage.py runserver 0.0.0.0:8000
    volumes:
      - .:/vsm
    ports:
      - 8001:8000
    depends_on:
      - mysql
    networks:
      - vsm-networks
```

Source: The Author (2023)

Figure 32 - Docker Compose Programming Code for Mosquitto, NodeRed, and Metabase

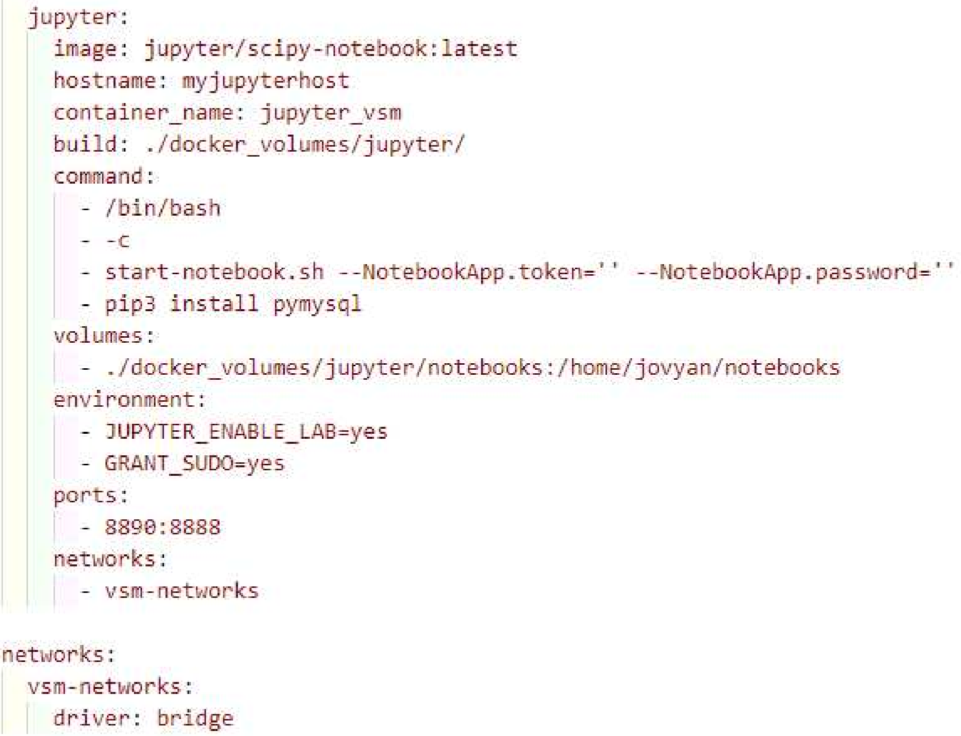


```
mosquitto:
  image: eclipse-mosquitto:latest
  hostname: mymosquittohost
  container_name: mosquitto_vsm
  volumes:
    - ./docker_volumes/mosquitto/config:/mosquitto/config
    - ./docker_volumes/mosquitto/data:/mosquitto/data
    - ./docker_volumes/mosquitto/log:/mosquitto/log

  ports:
    - 1885:1883
  networks:
    - vsm-networks

nodered:
  image: nodered/node-red:latest
  container_name: nodered_vsm
  volumes:
    - ./docker_volumes/nodered/data:/data

  ports:
    - 1882:1880
  networks:
    - vsm-networks
```

```
metabase:
  image: metabase/metabase:latest
  container_name: metabase_vsm
  depends_on:
    - mysql
  links:
    - mysql
  volumes:
    - ./docker_volumes/metabase:/metabase
    - ./docker_volumes/metabase-data:/metabase-data
  environment:
    - MB_DB_FILE=/metabase-data/metabase.db

  ports:
    - 3002:3000
  networks:
    - vsm-networks
```

Source: The Author (2023)

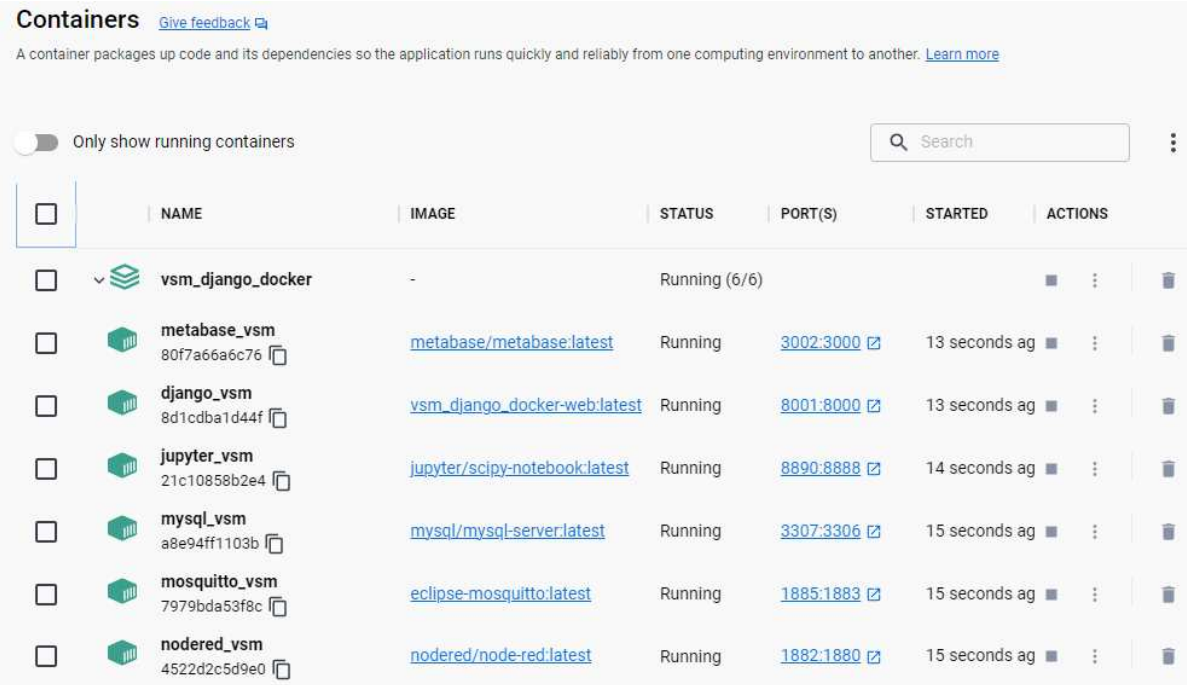
Figure 33 - Docker Compose Programming Code for Jupyter



```
  jupyter:
    image: jupyter/scipy-notebook:latest
    hostname: myjupyterhost
    container_name: jupyter_vsm
    build: ./docker_volumes/jupyter/
    command:
      - /bin/bash
      - -c
      - start-notebook.sh --NotebookApp.token='' --NotebookApp.password=''
      - pip3 install pymysql
    volumes:
      - ./docker_volumes/jupyter/notebooks:/home/jovyan/notebooks
    environment:
      - JUPYTER_ENABLE_LAB=yes
      - GRANT_SUDO=yes
    ports:
      - 8890:8888
    networks:
      - vsm-networks

networks:
  vsm-networks:
    driver: bridge
```

Source: The Author (2023)

Figure 34 – Docker Desktop



Containers Give feedback

A container packages up code and its dependencies so the application runs quickly and reliably from one computing environment to another. Learn more

Only show running containers

Search

□	NAME	IMAGE	STATUS	PORT(S)	STARTED	ACTIONS
□	vsm_django_docker	-	Running (6/6)			
□	metabase_vsm 80f7a66a6c76	metabase/metabase:latest	Running	3002:3000	13 seconds ag	
□	django_vsm 8d1cdba1d44f	vsm_django_docker-web:latest	Running	8001:8000	13 seconds ag	
□	jupyter_vsm 21c10858b2e4	jupyter/scipy-notebook:latest	Running	8890:8888	14 seconds ag	
□	mysql_vsm a8e94ff1103b	mysql/mysql-server:latest	Running	3307:3306	15 seconds ag	
□	mosquitto_vsm 7979bda53f8c	eclipse-mosquitto:latest	Running	1885:1883	15 seconds ag	
□	nodered_vsm 4522d2c5d9e0	nodered/node-red:latest	Running	1882:1880	15 seconds ag	

Source: The Author (2023)

The Changeover Time is the time needed to switch from producing one product to another, including the line production configuration and material preparation. Cycle time is the total time required to complete a single production cycle from start to finish. Availability refers to the capability of a machine, equipment, or resource to be available and functioning correctly during production time, i.e., it is a measure of the resource's availability to be used when necessary. FPY refers to the number of products that pass successfully in an initial inspection or process without the need for repair or correction, i.e., it is a measure that helps evaluate how many products were produced correctly on the first attempt. Waiting Time is when an item or work waits to be processed or inspected during production or operation.

As the partner company does not have a routine data collection process, the data obtained by the production line under study was based on the last VSM that had been carried out. Since the amount of data obtained in this mapping was not considered sufficient, especially for the quality of ML expected in the development of the future state mapping application, new data had to be generated from this obtained data. Thus, using a statistical function, the Excel software was used to create a larger amount of data.

Thus, 15,000 more data rows were created, considering only one operator in all operations. As mentioned in the previous paragraph, this is the number of operators currently used in these operations, but with a larger range for the other variables. In addition to these data, another 15,000 rows of additional data were added, but with a variation for two operators working in each operation. In order to obtain approximate values for actual operations with the addition of a new operator to each operation, it was established that in each operation, the cycle time would be reduced by 30%, and the waiting time would be reduced by 10%. The FPY would be increased by 5% to reflect the possible improvement in quality due to the involvement of an additional operator, and the changeover time and availability variables would remain in the same distribution of values used in operations with an operator.

The percentages used to represent variable changes were based on potential differences that could occur when adding a new operator. However, despite the reductions in Cycle Time by 30% and Waiting Time by 10%, as well as increases in FPY by 10%, being considered high values in a relatively stable process, the percentages were intentionally increased slightly above average in order to make the results in the software more perceptible and prevent them from becoming too close to each other. This approach was adopted to facilitate the comparison and validation of future prediction simulations, ensuring that the differences between the conditions with one operator and multiple operators were more distinct, thus proving the effectiveness of the learning algorithm used.

To create a column of 30,000 rows based on the data provided by the partner company, an average value for each variable in each operation was used as a basis for generating the other data. Therefore, a normal statistical distribution was used to generate random values that follow this distribution.

In Excel, the statistical function *=NORM.INV(probability, mean, standard_dev)* was used, replacing "probability" with the *RAND()* function, "mean" with the average value of the variable obtained in each operation, and "standard_dev" with 1.

NORM.INV function is commonly used in statistical analysis, simulations, and modeling to generate random values that follow a normal distribution. These values can create fictitious data or conduct experiments and simulations based on a known distribution. Approximately, most values generated by this function will be within the range of -3 and +3 from the value used in the "mean" parameter. This estimate is based on the property of the normal distribution, which states that about 68% of values are within one standard deviation of the mean, about 95% are within two standard deviations, and about 99.7% are within three standard deviations (FARRELL, 2018). However, it is essential to remember that the normal distribution is continuous and extends from negative infinity to positive infinity, and values outside this range can be generated, albeit with a lower probability.

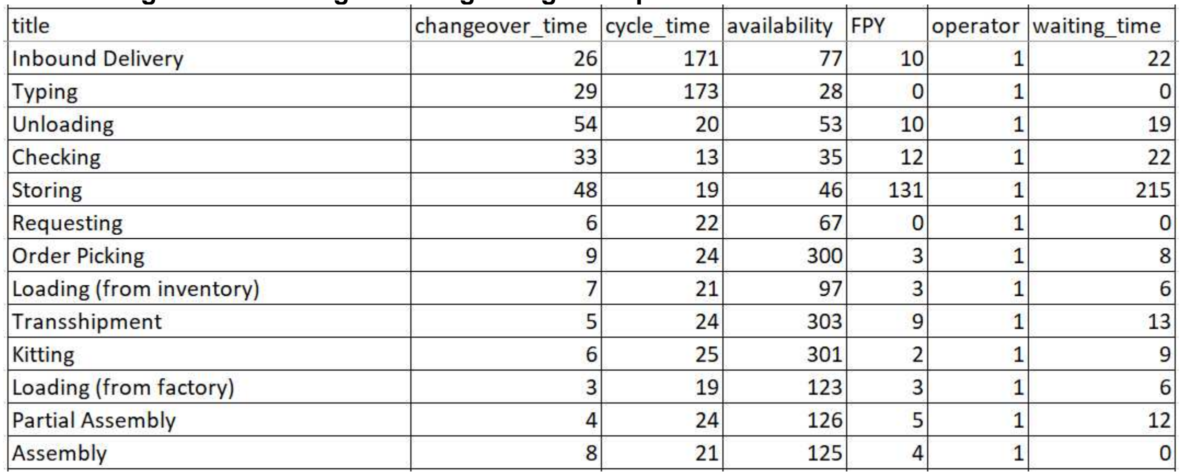
A row of each operation in the data table can be seen in Figure 35 as a partial example of the average values of each variable in each operation referring to a process with one operator.

A row of each operation in the data table can be seen in Figure 36 as a partial example of the average values of each variable in each operation referring to a process with two operators.

Although the changeover time is a more complex variable to achieve significant changes in a process, a distribution function was applied to it, as it can still vary according to different factors and circumstances.

While there is an estimated average time to perform the changeover time between operations or processes, according to Malindzakova; Malindzak; Garaj (2021), it can be influenced by the following:

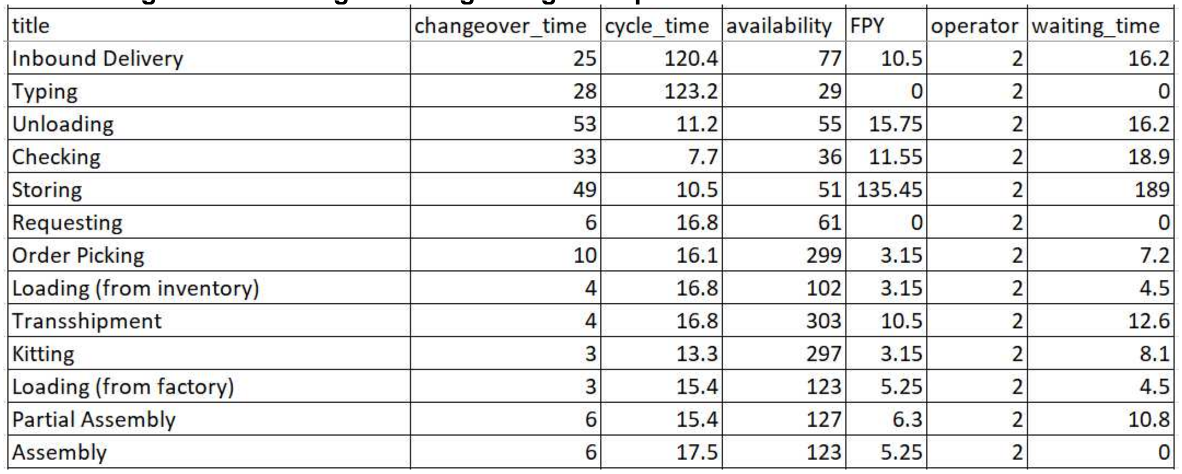
Figure 35 - Average data regarding one operator



title	changeover_time	cycle_time	availability	FPY	operator	waiting_time
Inbound Delivery	26	171	77	10	1	22
Typing	29	173	28	0	1	0
Unloading	54	20	53	10	1	19
Checking	33	13	35	12	1	22
Storing	48	19	46	131	1	215
Requesting	6	22	67	0	1	0
Order Picking	9	24	300	3	1	8
Loading (from inventory)	7	21	97	3	1	6
Transshipment	5	24	303	9	1	13
Kitting	6	25	301	2	1	9
Loading (from factory)	3	19	123	3	1	6
Partial Assembly	4	24	126	5	1	12
Assembly	8	21	125	4	1	0

Source: The Author (2023)

Figure 36 - Average data regarding two operators



title	changeover_time	cycle_time	availability	FPY	operator	waiting_time
Inbound Delivery	25	120.4	77	10.5	2	16.2
Typing	28	123.2	29	0	2	0
Unloading	53	11.2	55	15.75	2	16.2
Checking	33	7.7	36	11.55	2	18.9
Storing	49	10.5	51	135.45	2	189
Requesting	6	16.8	61	0	2	0
Order Picking	10	16.1	299	3.15	2	7.2
Loading (from inventory)	4	16.8	102	3.15	2	4.5
Transshipment	4	16.8	303	10.5	2	12.6
Kitting	3	13.3	297	3.15	2	8.1
Loading (from factory)	3	15.4	123	5.25	2	4.5
Partial Assembly	6	15.4	127	6.3	2	10.8
Assembly	6	17.5	123	5.25	2	0

Source: The Author (2023)

a) Complexity of the change: If the transition between operations requires complex adjustments, extensive configuration changes, or exchanging essential tools and equipment, the changeover time may be longer.

b) Experience and skill of the operators: More experienced and skilled operators can perform the changeover more efficiently, resulting in shorter times.

c) Availability of resources: If the resources needed for the changeover, such as tools, spare parts, or additional machinery, are readily available, the changeover time can be reduced.

d) Optimized processes and procedures: Having standardized processes and well-defined procedures for the changeover can help reduce the time required.

e) Continuous improvements: Implementing continuous improvements in the changeover process by identifying and eliminating bottlenecks or unnecessary steps can reduce the total time.

In many cases, 10,000 data lines could already be sufficient to obtain excellent result. However, 30,017 data lines were used to increase the probability of bringing good results from the start of the implementation tests.

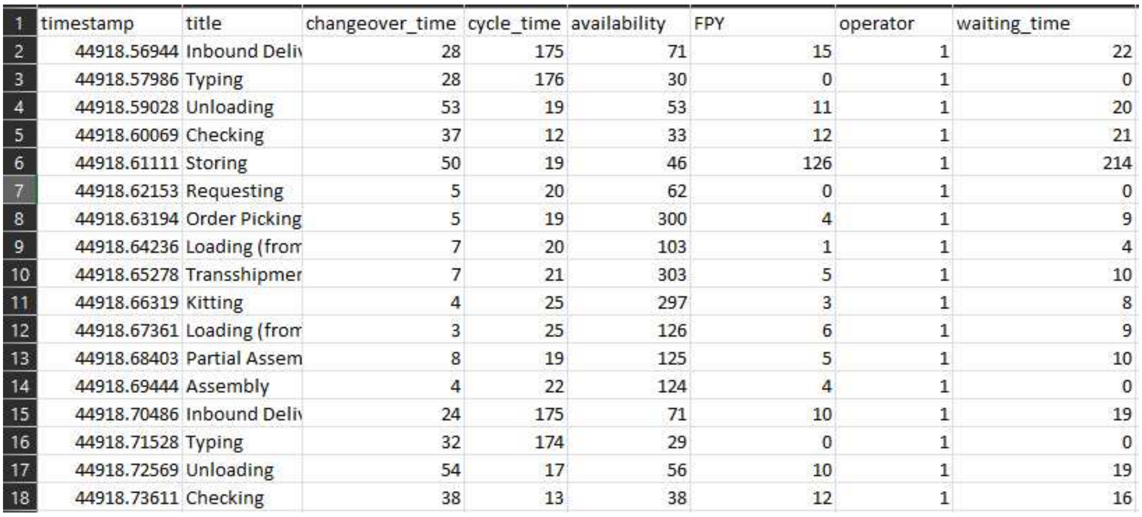
The ideal number of data rows for the machine learning algorithm used in future state mapping depends on various factors, such as the complexity of the problem, the number of variables, and the dimensionality of the data. Generally, the larger and more variable the amount of data, the better the machine learning model's performance will be.

This file comprises eight columns of data, with the first column for the timestamp record (i.e., the timeline of observation and record of each data row) to be used as a resource for sorting each operation in the algorithms developed. The timestamp was configured in Excel in the "General" format (i.e., in a decimal format) so that there would be no conflicts of interpretability of this information between the functions used, considering all the applications that will coordinate the data in this file. The following columns are related to the title of each operation, as well as the changeover time, cycle time, availability, FPY, the number of operators, and waiting time (Figure 37).

After creating these data rows in Excel, the .xlsx file was converted to a .csv file since this extension is compatible with data reading in the Node-Red platform. This file was named vsmF30t.csv and saved in a location where the Node-Red platform has permission to read this file, which was specified in the Docker Compose code line.

As previously explained, access to Node-Red is locally performed through port 1882 in any browser (i.e., localhost:1882). However, both Node-Red, Mosquitto, and the MySQL database need to be activated in Docker Desktop for the developed objective on this platform to be fulfilled (Figure 38).

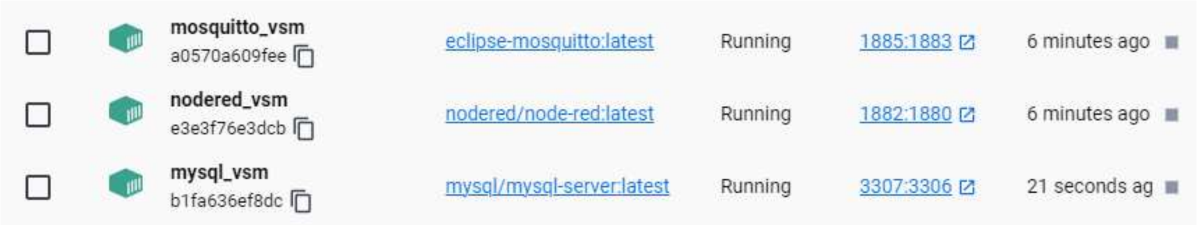
Figure 37 - Partial List of Data in CSV file



1	timestamp	title	changeover_time	cycle_time	availability	FPY	operator	waiting_time
2	44918.56944	Inbound Deliv	28	175	71	15	1	22
3	44918.57986	Typing	28	176	30	0	1	0
4	44918.59028	Unloading	53	19	53	11	1	20
5	44918.60069	Checking	37	12	33	12	1	21
6	44918.61111	Storing	50	19	46	126	1	214
7	44918.62153	Requesting	5	20	62	0	1	0
8	44918.63194	Order Picking	5	19	300	4	1	9
9	44918.64236	Loading (from	7	20	103	1	1	4
10	44918.65278	Transshipmen	7	21	303	5	1	10
11	44918.66319	Kitting	4	25	297	3	1	8
12	44918.67361	Loading (from	3	25	126	6	1	9
13	44918.68403	Partial Assem	8	19	125	5	1	10
14	44918.69444	Assembly	4	22	124	4	1	0
15	44918.70486	Inbound Deliv	24	175	71	10	1	19
16	44918.71528	Typing	32	174	29	0	1	0
17	44918.72569	Unloading	54	17	56	10	1	19
18	44918.73611	Checking	38	13	38	12	1	16

Source: The Author (2023)

Figure 38 – Activation of Mosquitto, Nodered, and MySQL on Docker Desktop



□	mosquitto_vsm a0570a609fee	eclipse-mosquitto:latest	Running	1885:1883	6 minutes ago
□	nodered_vsm e3e3f76e3dcb	nodered/node-red:latest	Running	1882:1880	6 minutes ago
□	mysql_vsm b1fa636ef8dc	mysql/mysql-server:latest	Running	3307:3306	21 seconds ag

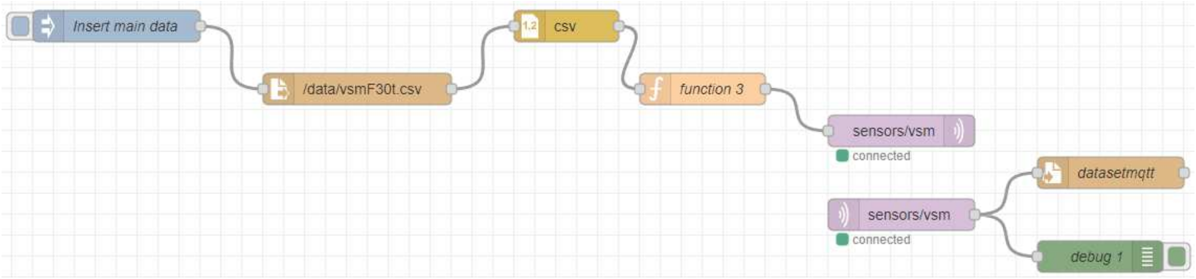
Source: The Author (2023)

In Node-Red, two data flows were developed in the workflow area of the platform. The first was responsible for publishing a topic in the MQTT protocol, as this protocol is an efficient, secure, and reliable way to transport data between different devices. The second data flow developed directs the data published on this topic to the MySQL database. MQTT is compatible with various database systems, allowing for adequate integration of this information in MySQL.

The first data flow developed has eight nodes: Inject, Read File, CSV, Function, MQTT Out, MQTT In, Write File, and Debug (Figure 39).

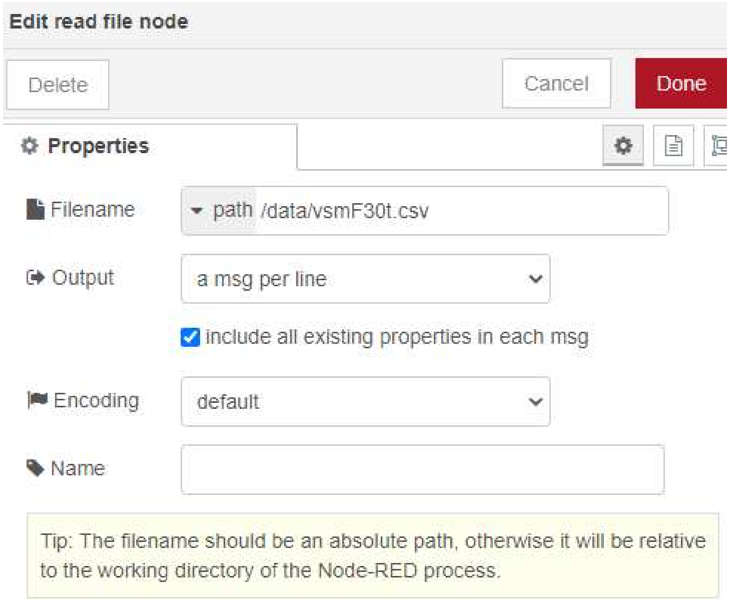
The Inject node, renamed to "Insert main data," allows for sending messages manually to the data flow, so it was used to initialize the execution of the flow. The Read File node was not renamed in this position. Still, the configuration of the path to the file's location to be read by the node (Figure 40) established this indication externally on the node displayed on the platform. It is important to note that this node only reads information from generic files and is not specific to CSV files, and for this reason, the CSV node had to be included in the sequence, which helps to separate the data into rows and columns.

Figure 39 - Node-Red flow (CSV to MQTT)



Source: The Author (2023)

Figure 40 - Edit Read File Node (CSV)



Filename: path /data/vsmF30t.csv

Output: a msg per line

☑ include all existing properties in each msg

Encoding: default

Name:

Tip: The filename should be an absolute path, otherwise it will be relative to the working directory of the Node-RED process.

Source: The Author (2023)

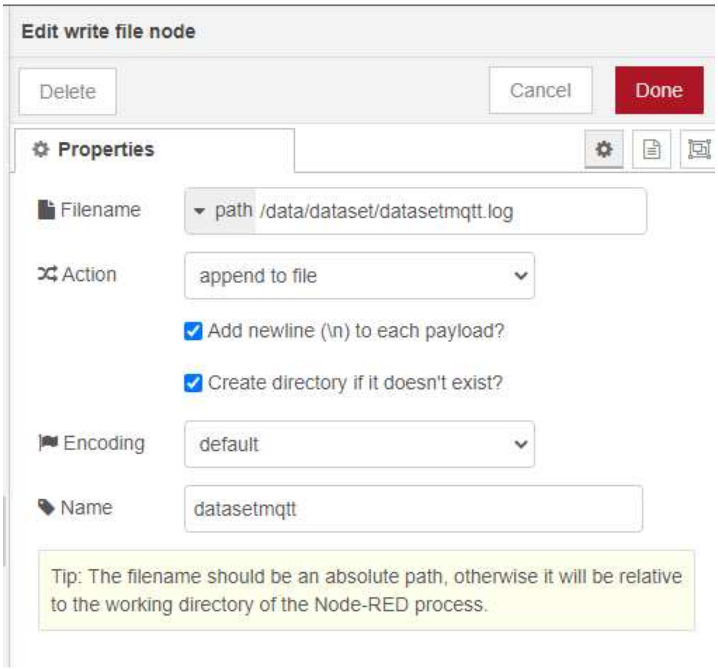
The Function node was used to transform the data received in CSV to JS, the platform's standard language, to format an MQTT input message appropriately. The message is processed, and the fields are rearranged, renamed, and assigned to a new "payload" object, which will be the content of the message that the flow will pass on. In addition to the data received from the CSV, the function adds another column, "Date," with the date these data were processed in creating this object.

The MQTT Out node will receive this message and publish it to a topic named sensor/vsm through the MQTT broker already installed in a Docker container. The broker is referenced in this node as test-client@mymosquittohost:1883. Port 1883 was established in the Docker Compose to access the Mosquitto container.

This topic published by the MQTT Out node is received by the MQTT In node through the same MQTT broker, as this intermediary controls the flow of messages and ensures that they are delivered correctly and in order. The MQTT In node processes this information and forwards it to the next node in the flow.

The Write File node collects the information forwarded by the MQTT In node and saves this information to a log file called datasetmqtt.log in a path specified in the node's configurations (Figure 41).

Figure 41 - Edit Write File Node



Edit write file node

Filename: path /data/dataset/datasetmqtt.log

Action: append to file

☑ Add newline (\n) to each payload?

☑ Create directory if it doesn't exist?

Encoding: default

Name: datasetmqtt

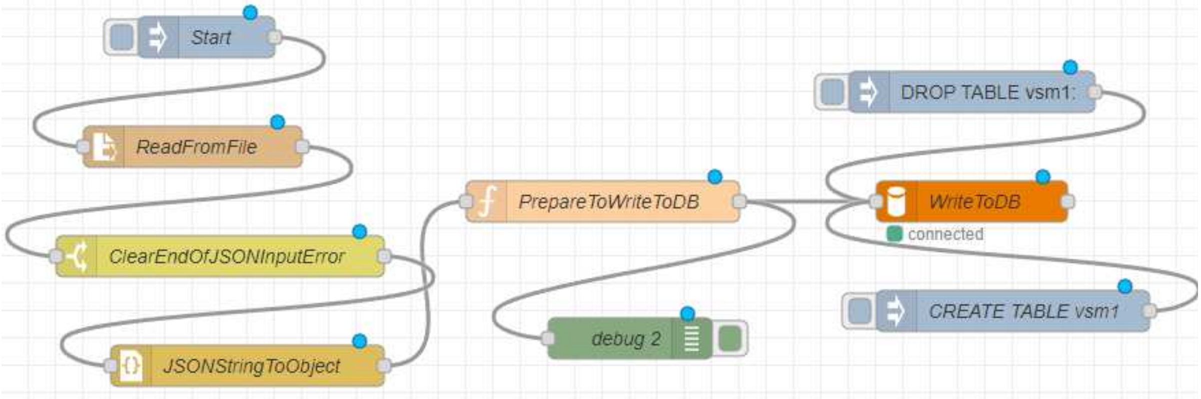
Tip: The filename should be an absolute path, otherwise it will be relative to the working directory of the Node-RED process.

Source: The Author (2023)

The debug node only confirms that the data is being appropriately received to create this log file.

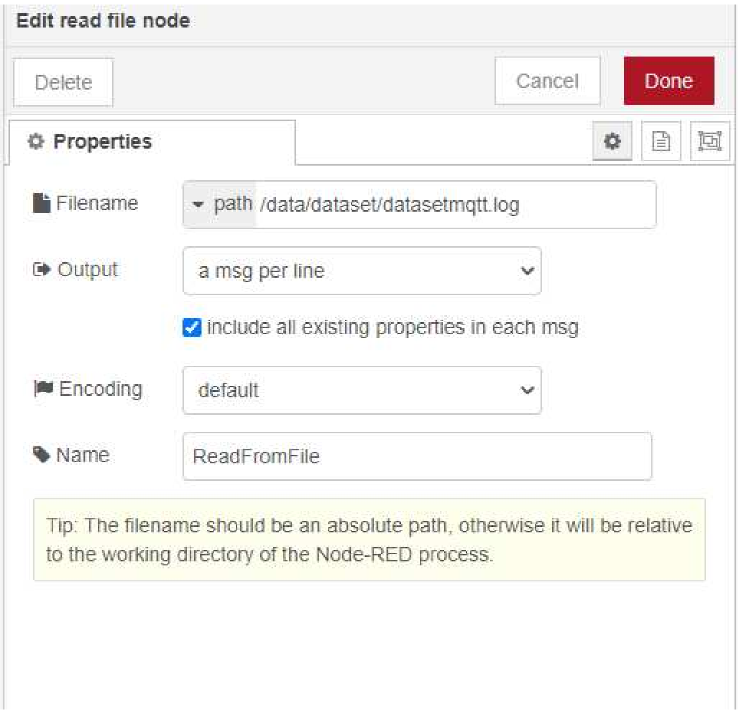
The second data flow consists of nine nodes: Inject, Read File, Switch, Json, Function, Debug, MySQL, and two more Injects (Figure 42). The first Inject is used to manually initialize the data flow, which starts with the Read File node for reading the log file created in the previous data flow (Figure 43).

Figure 42 - Node-Red flow (MQTT to MySQL)



Source: The Author (2023)

Figure 43 - Edit Read File node (Log)



Filename: path /data/dataset/datasetmqtt.log

Output: a msg per line

☑ include all existing properties in each msg

Encoding: default

Name: ReadFromFile

Tip: The filename should be an absolute path, otherwise it will be relative to the working directory of the Node-RED process.

Source: The Author (2023)

The Switch node, renamed ClearEndOfJSONINputError, directed the data flow according to specific rules. In this case, the rule "is not empty" was used only to allow the flow to continue if there is a value in the data input, and if the input is empty, the flow is interrupted by this node.

The JSON node, renamed to JSONStringToObject, was used to allow the conversion of JSON to String and thus obtain the necessary flexibility between the data sent as required. The Function node, renamed to PreparetoWritetoDB, is responsible for extracting data from the "payload" object and formatting it appropriately to insert it into a SQL table. In this way, ten variables (ID, Date, Timestamp, Title, Changeover Time, Cycle Time, Availability, FPY, Operator, and Waiting Time) that represent the columns of the SQL table were extracted from the object, grouped into an array, assigned to the object and finally set up an SQL statement. This statement has two commands: create a table named vsm1 (if it does not exist) and insert these values into the table. A Debug node was placed at the output of this node to check that this function formatted the data into a message object appropriately for insertion into an SQL table.

The MySQL node is used to establish a connection to a MySQL database and thus be able to develop the operations created in the previous node function. In this node, it is necessary to configure the information of the existing database, with the Host, Port, User, Password, and database name. These data must be equivalent to the configurations of the MySQL created by Docker Compose so that all applications linked to the vsm-networks can access and use the data from the created table.

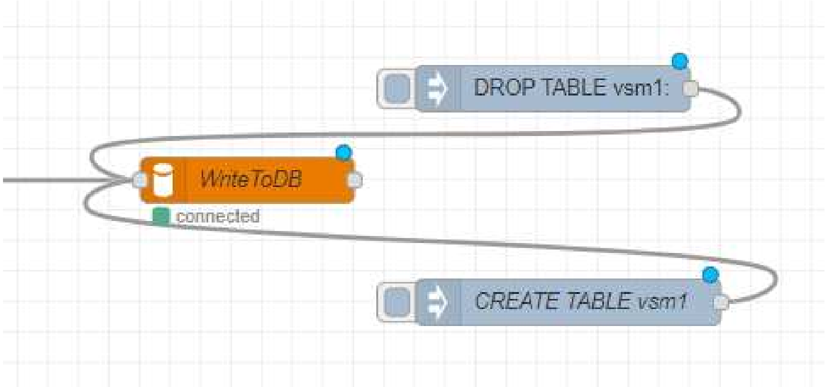
Two Inject nodes were connected to the MySQL node (Figure 44) so that it is possible to manually delete the vsm1 table completely (DROP TABLE vsm1) and then include a new empty table (CREATE TABLE vsm1) as an alternative in situations where it is necessary to update the database with new data, without risking injecting duplicate data. Every time the Inject button is clicked, the node-red will reinsert all the data in the CSV. If there are 30,000 rows in the CSV, with one click, the vsm1 table will have 30,000 rows. If there is a second click, the vsm1 table is duplicated, and it will have 60,000 rows, and so on. Duplicate tables are not recommended as they will not benefit the application usage and may unnecessarily increase the workload on MySQL.

There are two options available to update the data in the vsm1 table:

(i) delete the vsm1 table with the DROP TABLE vsm1 command, add new data rows to the same CSV file, delete the log in the folder where it was saved, inject the CSV data through the Node-RED by creating a new log, and then inject the data from the created log to a new vsm1 table.

(ii) use a new CSV file in the same specified path, with the same name, but only with new data, inject the latest data, overwrite the log with the new data, and then add the new data to the vsm1 table.

Figure 44 - MySQL node with two Inject nodes

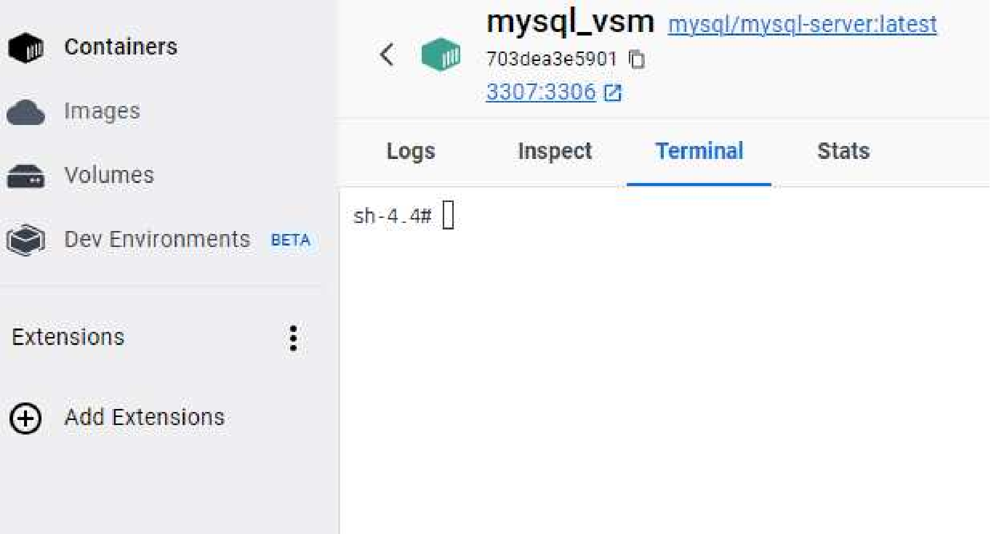


Source: The Author (2023)

Once the “vsmdatabase” database and the vsm1 table have been created and the data from the CSV file has been inserted into this table, any existing applications installed on the vsm-networks can access this data. The vsm1 table can be accessed through the terminal linked to the Docker desktop so that it is possible to query the existing data and make changes to the database, if necessary, through SQL language.

Access to this terminal (Figure 45) is achieved by clicking "Open in Terminal" next to the name of the mysql_vsm container in Docker Desktop.

Figure 45 - Container Terminal for MySQL access



```
sh-4.4# 
```

Source: The Author (2023)

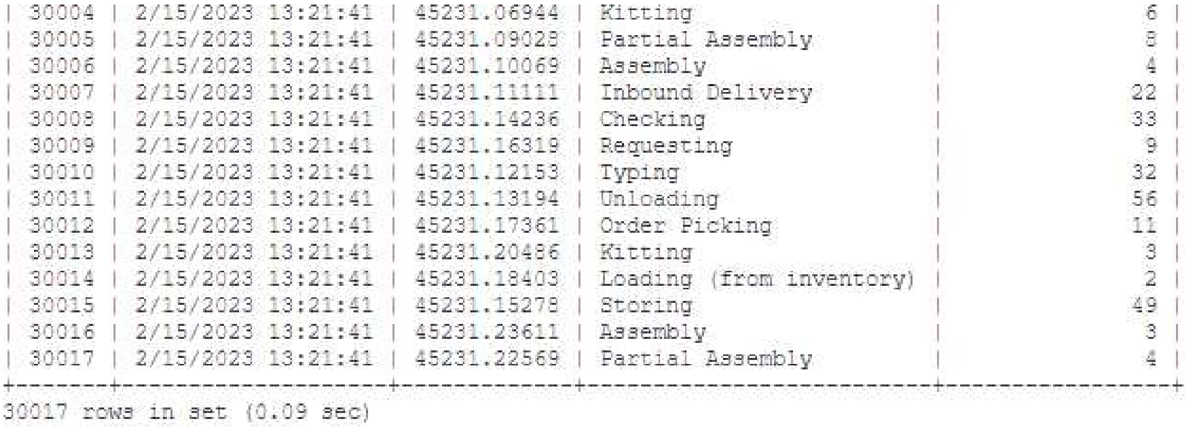
The MySQL access user and password have been configured in the Docker Compose code, and this access is allowed by typing the following codes in the terminal:

```
mysql --user=root --password=m4ster
use vsmdatabase;
select * from vsm1;
```

With these commands, all the data in the table can be viewed (Figure 46).

With all the data organized in MySQL, the web access programming lines were developed through the Django web framework. The programming was done with the aid of Microsoft VSCode IDE. Python was used in the back end, and JS, CSS, and HTML were used for the front end. The web system access is performed with the activation of the django_vsm container in Docker Desktop and with the execution of the code line "python manage.py runserver" via terminal (Figure 47), and port 8001 is used for web access in the browser (i.e., localhost:8001).

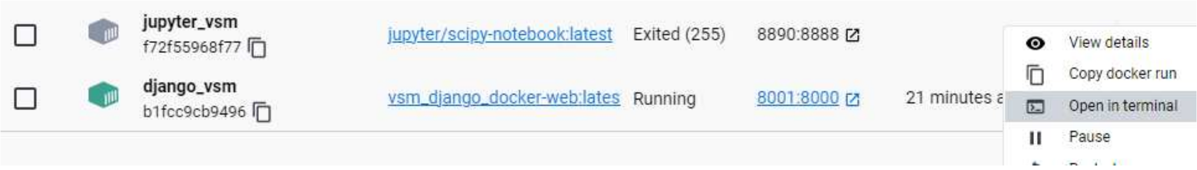
Figure 46 - Partial vsm1 table on MySQL



```
| 30004 | 2/15/2023 13:21:41 | 45231.06944 | Kitting                   |                 6 |
| 30005 | 2/15/2023 13:21:41 | 45231.09028 | Partial Assembly          |                 8 |
| 30006 | 2/15/2023 13:21:41 | 45231.10069 | Assembly                  |                 4 |
| 30007 | 2/15/2023 13:21:41 | 45231.11111 | Inbound Delivery          |                22 |
| 30008 | 2/15/2023 13:21:41 | 45231.14236 | Checking                  |                33 |
| 30009 | 2/15/2023 13:21:41 | 45231.16319 | Requesting                |                 9 |
| 30010 | 2/15/2023 13:21:41 | 45231.12153 | Typing                    |                32 |
| 30011 | 2/15/2023 13:21:41 | 45231.13194 | Unloading                 |                56 |
| 30012 | 2/15/2023 13:21:41 | 45231.17361 | Order Picking             |                11 |
| 30013 | 2/15/2023 13:21:41 | 45231.20486 | Kitting                   |                 3 |
| 30014 | 2/15/2023 13:21:41 | 45231.18403 | Loading (from inventory)  |                 2 |
| 30015 | 2/15/2023 13:21:41 | 45231.15278 | Storing                   |                49 |
| 30016 | 2/15/2023 13:21:41 | 45231.23611 | Assembly                  |                 3 |
| 30017 | 2/15/2023 13:21:41 | 45231.22569 | Partial Assembly          |                 4 |
+-------+--------------------+-------------+---------------------------+-------------------+
30017 rows in set (0.09 sec)
```

Source: The Author (2023)

Figure 47 - Activation of Django container on Docker Desktop



□	jupyter_vsm f72f55968f77	jupyter/scipy-notebook:latest	Exited (255)	8890:8888		
□	django_vsm b1fcc9cb9496	vsm_django_docker-web:lates	Running	8001:8000	21 minutes a	

Source: The Author (2023)

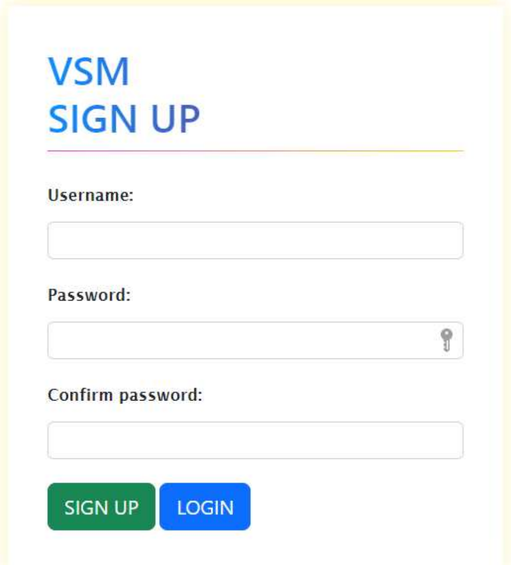
In web development, in the Django framework, a project named "vsm_django_docker" was created, which has four smaller applications: "app_infomapping," "app_login," "app_prediction," and "app_prodmapping." Creating smaller applications helped organize and structure the code, making it easier to manage and maintain the application. Both applications used the MVT architecture, which allows for the logic of data manipulation, user request manipulation, and data presentation to be separated for ease of maintenance and development of each functional feature.

The "app_login" application was organized only to manage the functions related to the signup and login screens of the application. Three functions in Python were created in this application: "signup," "login," and "exit." The "signup" function is responsible for creating new users in the system. When the user accesses the application's main page, this function checks if the user is already authenticated and, if so, redirects them to the main page. If the user is not authenticated, the "signup.html" page is rendered, and the user is redirected to a page to register the user and password (Figure 48). If the user has already registered, there is a button to switch to the login screen. During this registration, the function collects the entered user and password data, verifies that the passwords match, and verifies that all fields have been filled in. If any errors are found, an error message is added to the request, and the user is redirected back to the signup page. A new user is created and saved in the database if the information is valid.

The "login" function is responsible for authenticating users in the system. When the user accesses the system's main page, this function also checks if the user is already authenticated and, if so, redirects them to the main page. If the user is not authenticated, the "login.html" page is rendered, and the user is redirected to the login page (Figure 49). If the user still needs to register, there is a button to switch to the signup screen. During the login process, the function collects the entered user and password data and tries to authenticate them. If authentication fails, an error message is added to the request, and the user is redirected to the login page. If authentication is successful, the user is redirected to the main page. The "exit" function logs out the currently authenticated user and redirects them to the login page.

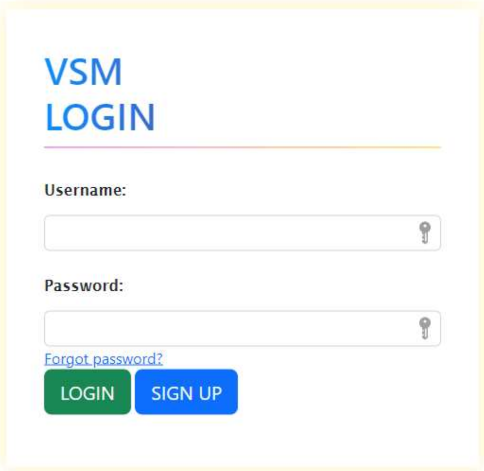
Django is a framework that provides an ORM API (Application Programming Interface) to work with data to manipulate the data using Python code, i.e., without having to write SQL directly. Models in Django are defined as Python classes, where attributes (fields) and behaviors (methods) of objects are specified and where an interface is provided to access and manipulate this data. This interface is accessed through the local URL localhost:8001/admin (Figure 50), and access is made using the same login and password created on the registration screen. The identification of all registered users, as well as the commands for adding new ones or deleting existing ones, can be performed through the "Users" option in this administrative area (Figure 51).

Figure 48 - Sign up page



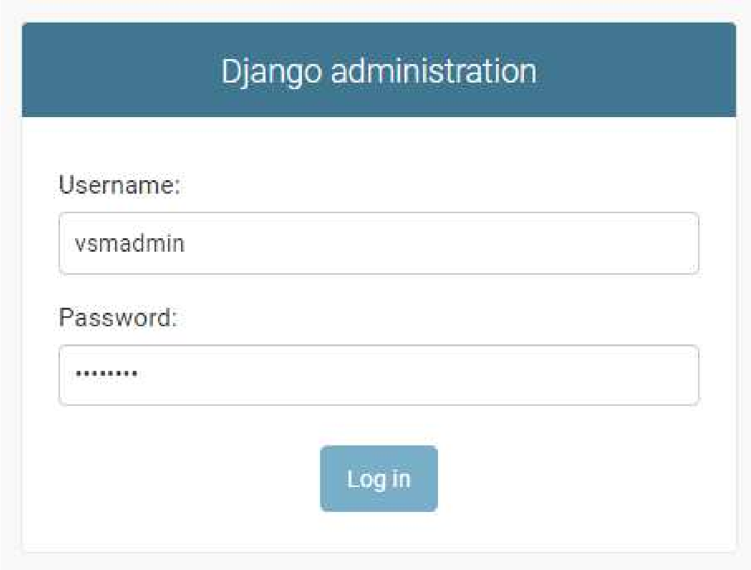
Source: The Author (2023)

Figure 49 - Login page



Source: The Author (2023)

Figure 50 - Django Administration Login

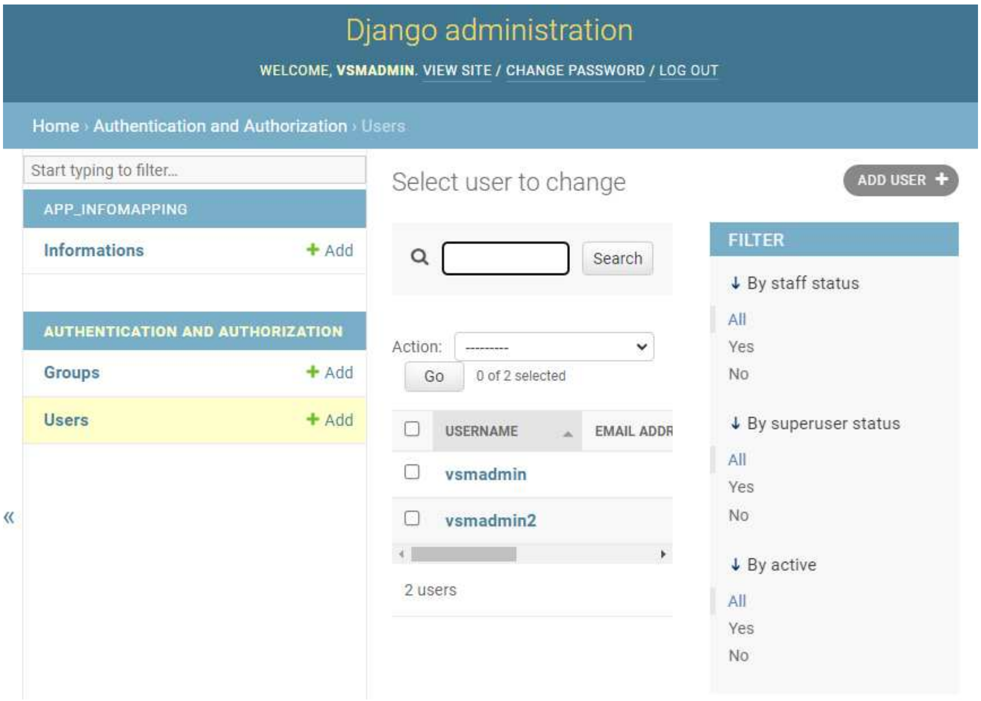


Username: vsmadmin

Password: ••••••••

Source: The Author (2023)

Figure 51 - Django Administration Main Interface



	USERNAME	EMAIL ADDR
□	vsmadmin	
□	vsmadmin2	

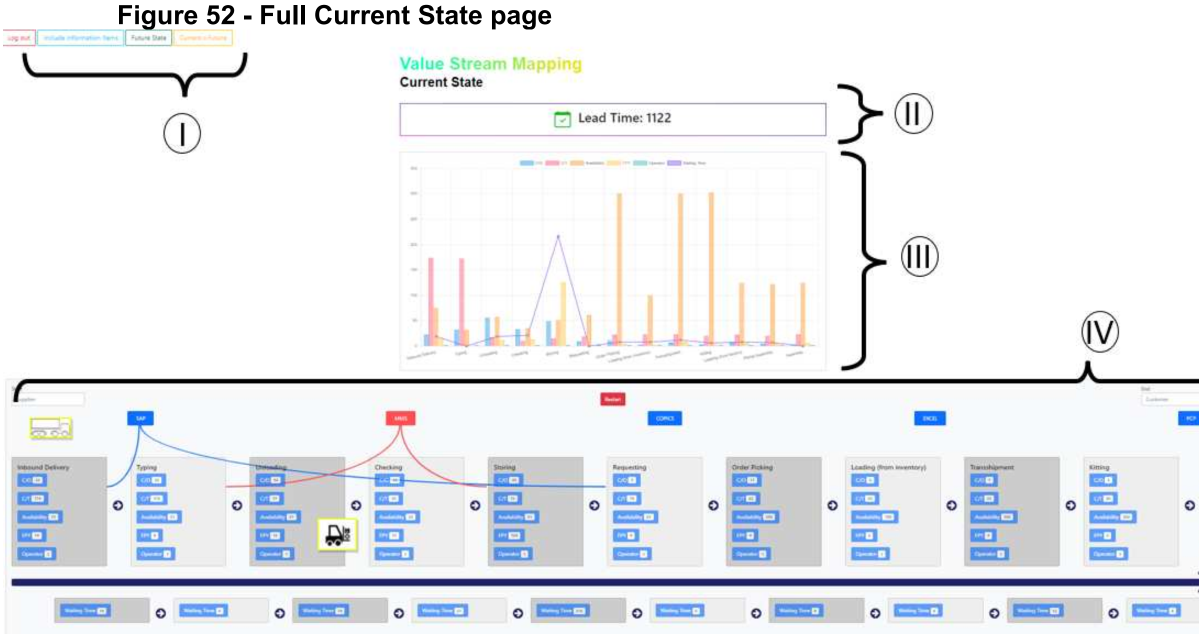
2 users

Source: The Author (2023)

In the app_infomapping, a class named Information was defined with a field called "Information_name" and type CharField, which means it is a limited text field. The definition of this class in the Django Models area defines the creation of a new table in the database with the same name as the class. This class allows the creation of as many objects as necessary through the Django administrative area. This table was created to store the items organized for the information item mapping in the VSM.

In the app_prodmapping, the main functions for developing the current state mapping were concentrated, which is the main access area after login. This page can be divided into four functional areas (Figure 52): (i) Buttons for transitioning to other pages (Figure 53), located at the top; (ii) calculation area for Lead Time; (iii) Dynamic chart for comparative data between each operation in the analyzed process (Figure 54); and (iv) Dynamic structure for the sequencing of operations in the current state mapping.

Figure 52 - Full Current State page

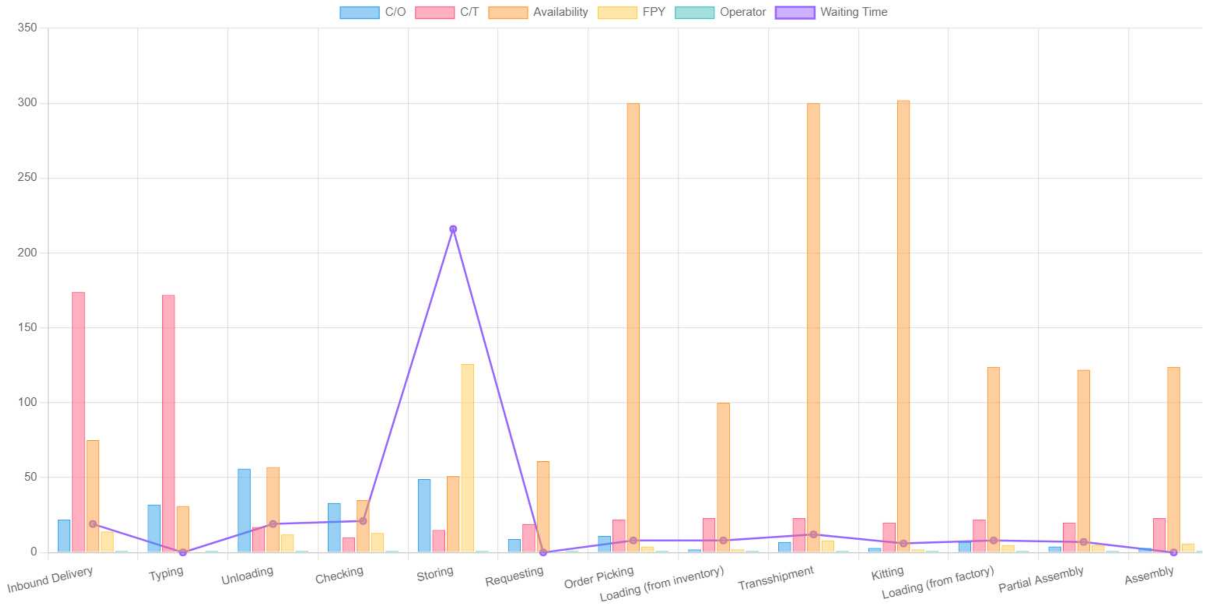


Source: The Author (2023)

Figure 53 – Current State page Buttons for transitioning to other pages

Source: The Author (2023)

Figure 54 – Current State Chart



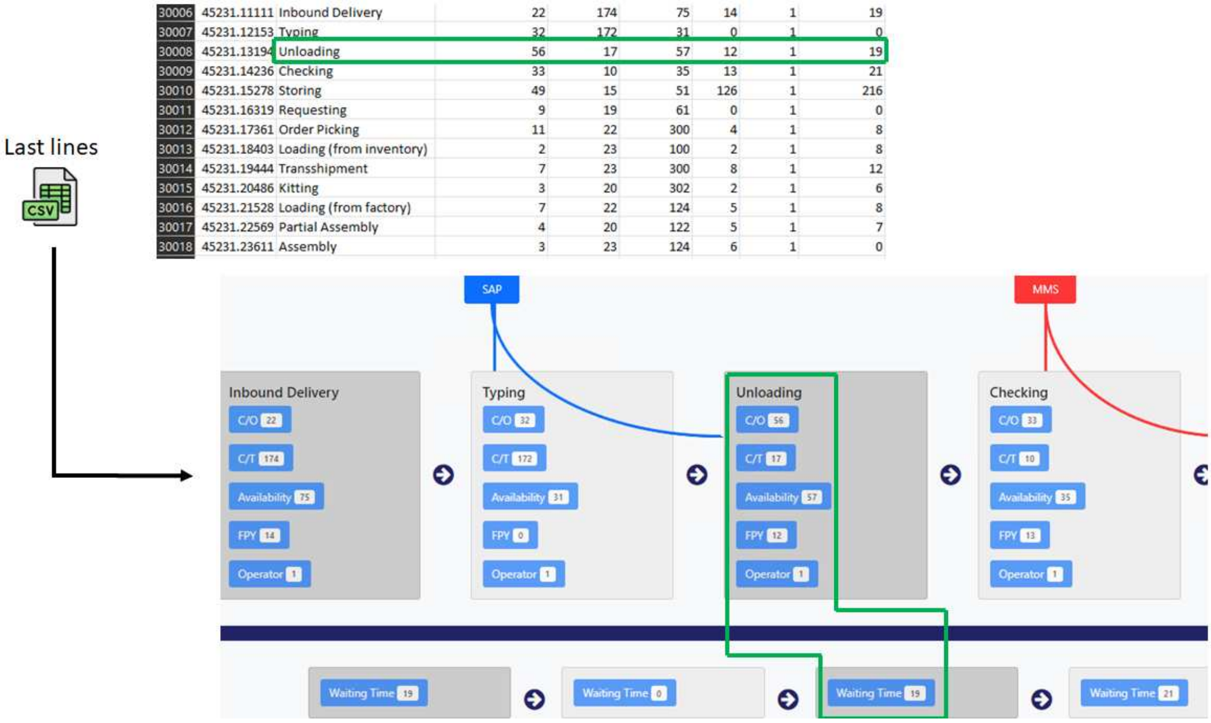
Source: The Author (2023)

In this system, four functions were created in Python that involve database connection, and several functions were created in JS for the dynamic front-end development of the system. Both functions interact with each other, and the display of these results is possible through the interaction of these functions with the HTML template, called in this code prod_mapping.html. All HTML templates used in this project have lines of code with imports of CSS and JS libraries.

In the View layer, the four Python functions are named: dynamic_mapping, dynamic_topbuttons, home, and dash_lead_time. The home function is used only to render the main HTML template of this system, but the buttons with the redirect links between pages were created in the HTML template itself by linking the URLs of the other templates of other pages. The dash_lead_time function, along with a JS function, returns JSON data with the total calculated time of the sum of the cycle time, changeover time, and waiting time fields of the last 13 operations of the process in the database. Even if the CSV file is updated, the function will always filter the data from the last operations, sum them up, and thus always display the current Lead Time on the page.

The dynamic_mapping function accesses all data from the vsm1 table and, just like the dash_lead_time function, a filter is used to access only the data belonging to the last operations in the database, and thus display them automatically as soon as the user accesses the main screen of the current mapping (Figure 55).

Figure 55 - Current Mapping connected to the last lines of the Database



30006	45231.11111	Inbound Delivery	22	174	75	14	1	19
30007	45231.12153	Typing	32	172	31	0	1	0
30008	45231.13194	Unloading	56	17	57	12	1	19
30009	45231.14236	Checking	33	10	35	13	1	21
30010	45231.15278	Storing	49	15	51	126	1	216
30011	45231.16319	Requesting	9	19	61	0	1	0
30012	45231.17361	Order Picking	11	22	300	4	1	8
30013	45231.18403	Loading (from inventory)	2	23	100	2	1	8
30014	45231.19444	Transshipment	7	23	300	8	1	12
30015	45231.20486	Kitting	3	20	302	2	1	6
30016	45231.21528	Loading (from factory)	7	22	124	5	1	8
30017	45231.22569	Partial Assembly	4	20	122	5	1	7
30018	45231.23611	Assembly	3	23	124	6	1	0

Source: The Author (2023)

The result is converted into a JSON format so that other functions in JS can use it and finally returned in an HTTP response and consequently linked to the code lines of the HTML template. Thanks to the TimeStamp column, each data line brought to the main page already has the pre-established order. Still, if a visual analysis is necessary, with the alternation between operations, a JS function was developed that allows the exchange between the display blocks with just a click on the arrows between these blocks (Figure 56).

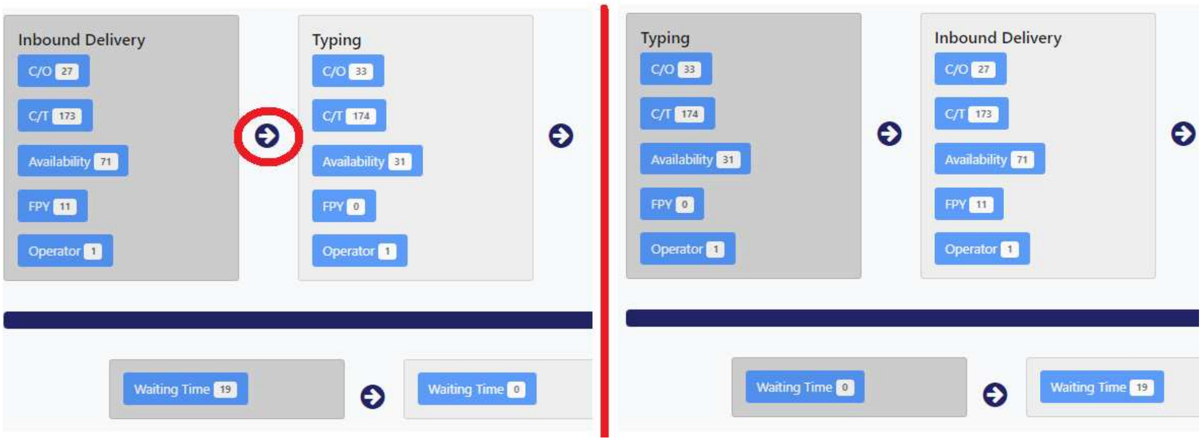
In order to make the mapping structure similar to the traditional VSM mapping, the Waiting Time data was allocated between operations in separate blocks from the main blocks. The same exchange functionality will be executed by clicking the arrows between the Waiting Time blocks. In the database, the Waiting Time values are always organized in the last column, i.e., the waiting time is related to the stock time after the operation that was executed, and therefore each waiting time value is always linked to an operation.

On accessing the main page, the latest data lines from the 13 operations are displayed on the block mapping, and simultaneously, the line and column chart, which is positioned just above the blocks, receives these same data, making it complementary to the quantitative data visualization that appears in the blocks. This chart displays the C/O (Changeover Time), C/T (Cycle Time), Availability, FPY, and Operator data in bars, but the Waiting Time data is displayed in Line, as already seen in Figure 54.

This chart has two primary dynamics, developed with the help of JS functions: (i) switching between the positions of each operation, which are triggered simultaneously with the change of blocks in the mapping (Figure 53), and (ii) dynamic filtering with the activation/deactivation of each bar/line, which is triggered by clicking on the respective color legend.

In this dynamic filter, deciding which values will be displayed and which will not be displayed at the analysis time is possible. In Figure 58, Figure 59, and Figure 60, it is possible to see examples of the same chart shown in Figure 54. Still, only the Cycle Time, FPY, and Waiting Time are displayed.

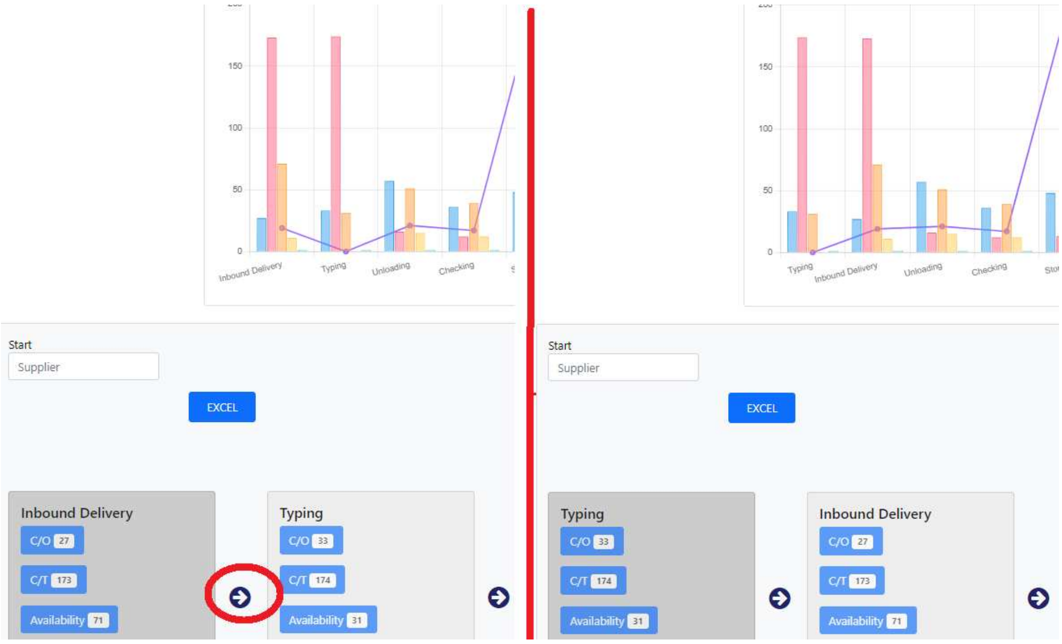
Figure 56– Arrow buttons for switching between blocks



Source: The Author (2023)

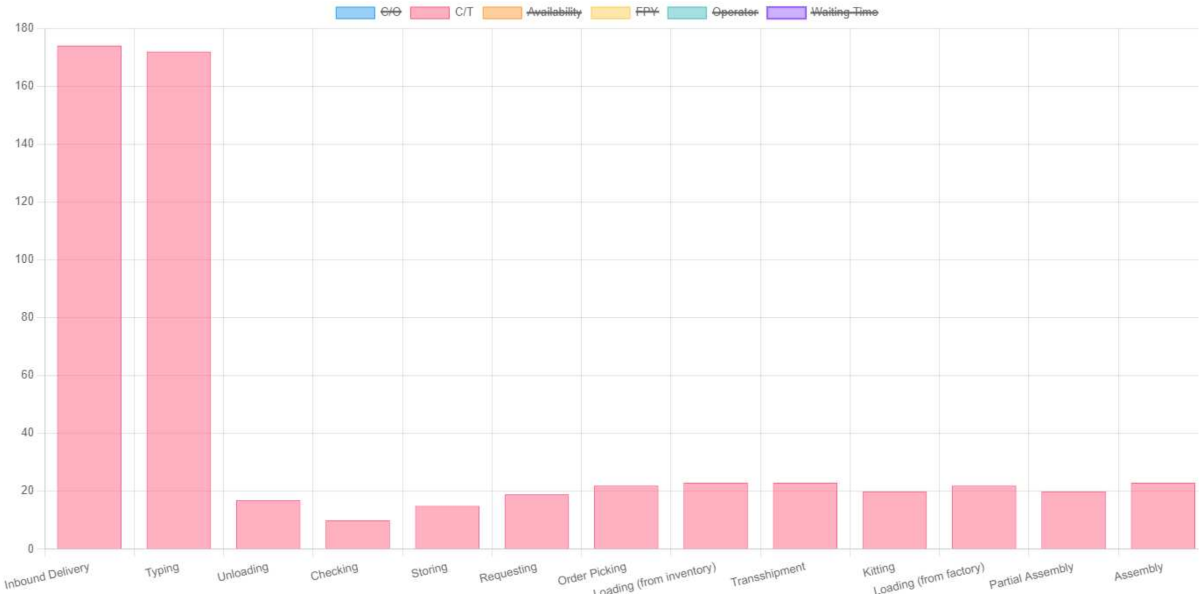
Moreover, the chart is also interactive, so the user can identify the value represented in each column by simply positioning the mouse arrow over one of these columns or a portion of the line (Figure 61).

Figure 57 – Before and after changing blocks



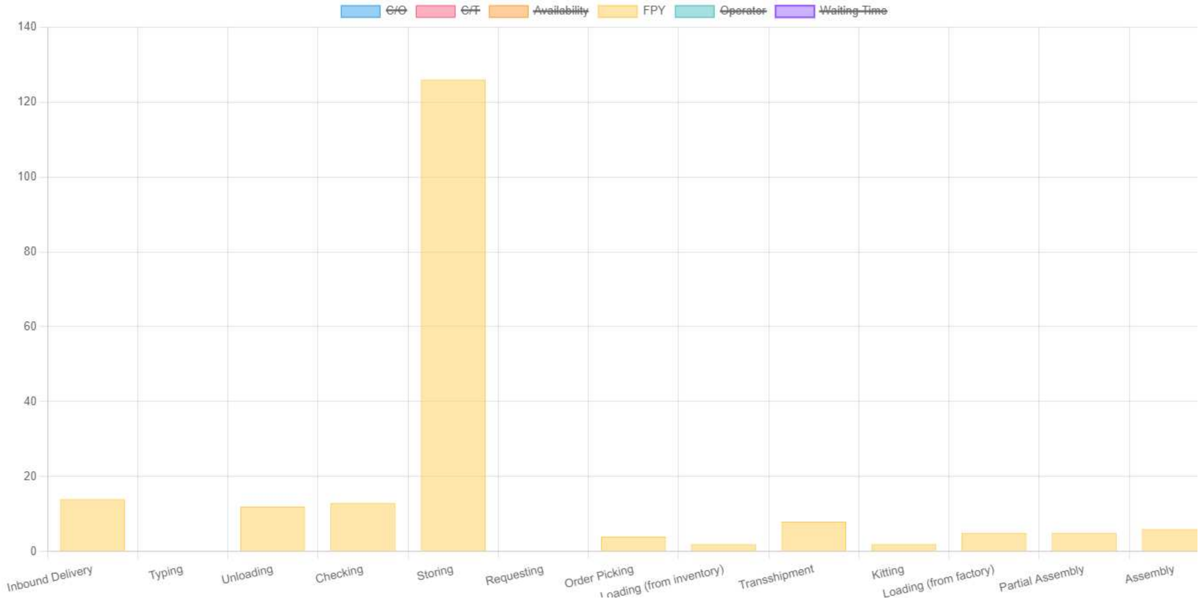
Source: The Author (2023)

Figure 58 - Cycle Time Bar filtering



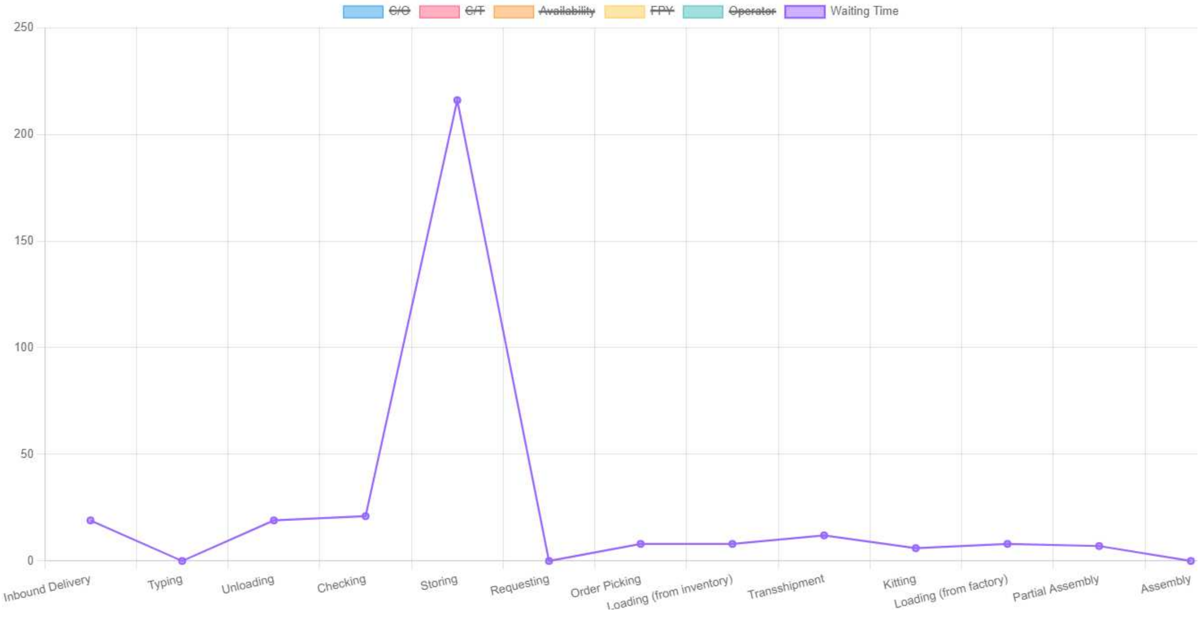
Source: The Author (2023)

Figure 59 - FPY Bar filtering



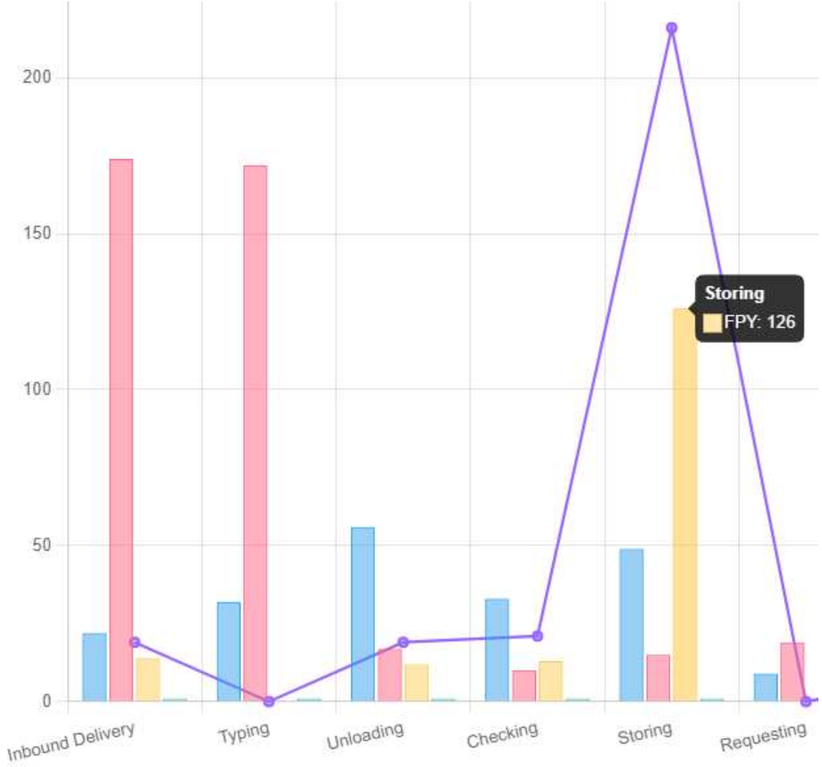
Source: The Author (2023)

Figure 60 - Waiting Time line filtering



Source: The Author (2023)

Figure 61 - Column value in Chart

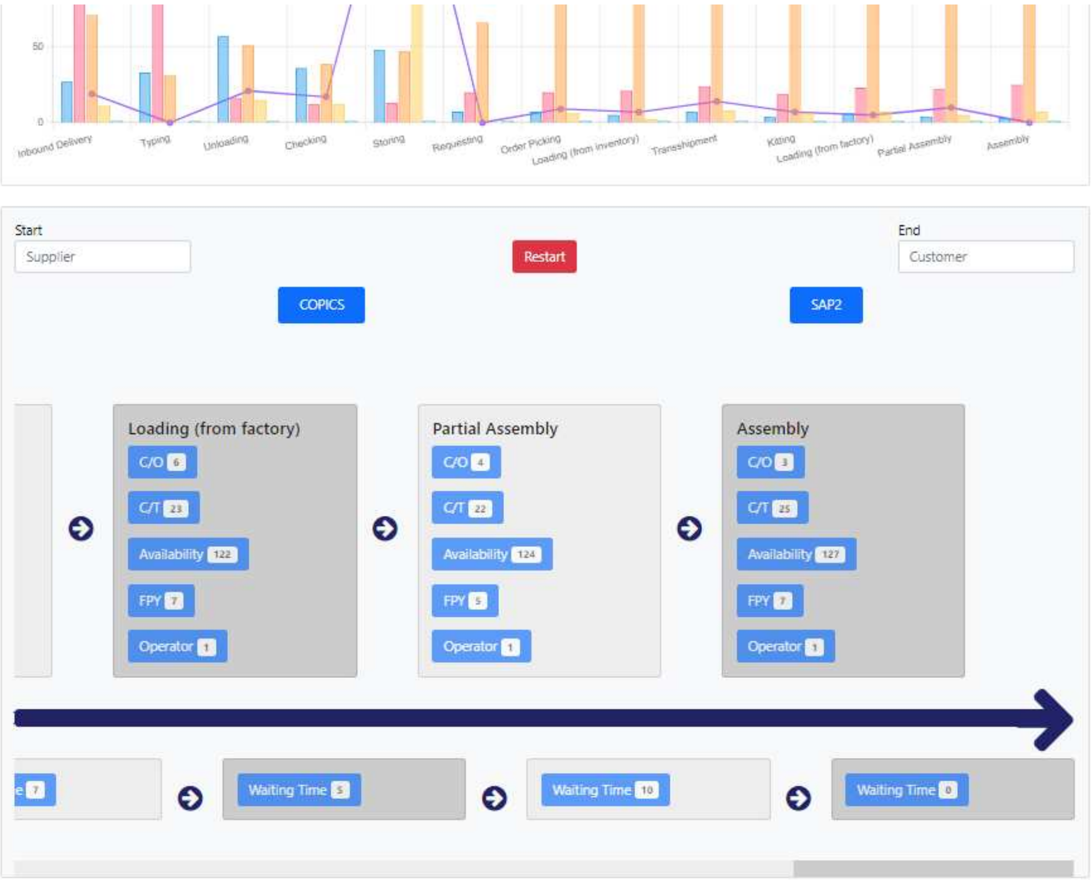


Source: The Author (2023)

Since the total width of the mapping between the 13 operations exceeds the maximum width required by the chart, and this can become larger or smaller depending on the information entered in the database, to make the user's visualization usability more dynamic and exciting, the open-source code design library Bootstrap was used to assist in the main structure of the interface for both pages in this project. Bootstrap provides a series of pre-built components, so the container layout element was used for this mapping to define a maximum width for the content, centering it horizontally on the page. In addition, a horizontal scrolling attribute was placed to view the entire extent of the mapping without this displacement moving away from the view of the chart (Figure 62).

Two input fields were also added for filling in the name of the Supplier, which is usually at the beginning of the flow of a process represented in a VSM, and another input field for filling in the name of the company, which is usually at the end of the process flow represented in a VSM. A placeholder was utilized in a web form's input field by default. This placeholder serves as descriptive text that appears within the field when empty. In this case, the input field is labeled "Supplier," while the output field is labeled "Customer." However, to provide flexibility in filling out the form, the field initially remains empty, allowing users to enter any name they desire. (Figure 62).

Figure 62 - Horizontal Scrolling demonstration



Loading (from factory)
C/O 6
C/T 23
Availability 122
FPY 7
Operator 1

Partial Assembly
C/O 4
C/T 22
Availability 124
FPY 5
Operator 1

Assembly
C/O 3
C/T 25
Availability 127
FPY 7
Operator 1

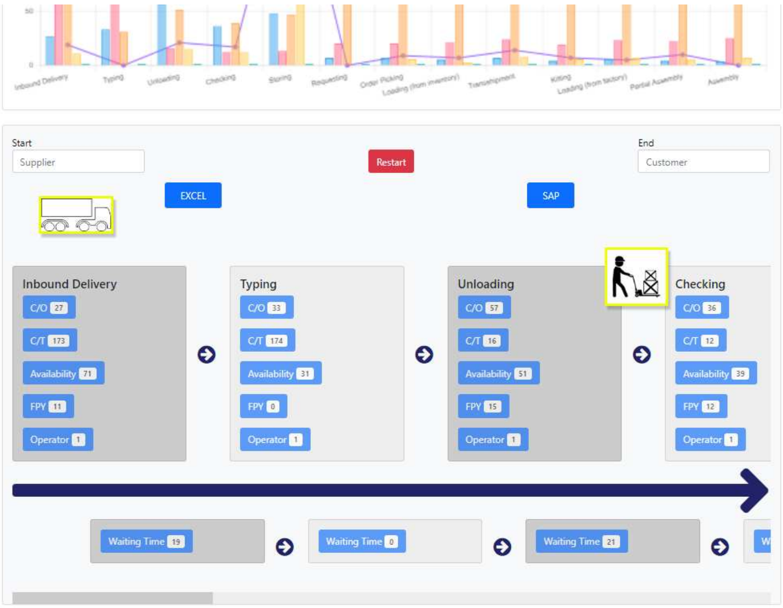
e 7
Waiting Time 5
Waiting Time 10
Waiting Time 0

Source: The Author (2023)

In case the user wants to view this mapping with a larger viewing area, a JS function has been linked to a button in the shape of a bidirectional arrow located in the upper right corner of the screen, which allows for the alternation of the width of the mapping area (Figure 63 and Figure 64).

A repository with the symbols used in VSM was configured in the main project folder in Django to allow the user to include them in current state mappings, such as glasses or exclamation point triangles, which help to visualize the process clearly and concisely and identify improvement points. This way, a JS function was determinant in allowing the user to include images and position them freely with just one click on the right mouse button at any position on the screen (Figure 65). Upon clicking, a frame is instantly displayed, showing all the symbols available for inclusion on the site. By selecting one of these images, it is possible to freely drag and drop them in the position the user considers appropriate for the analyzed process (Figure 66). To delete the images, clicking again with the right mouse button on the desired image will only be necessary.

Figure 63 – Current State Mapping standard visualization



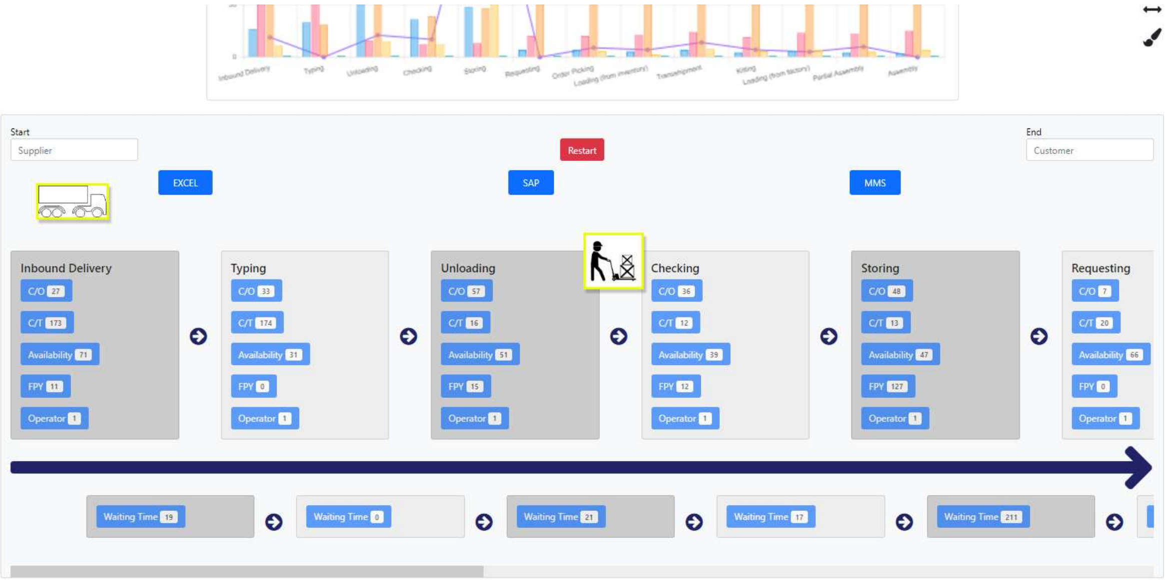
Start: Supplier
End: Customer

Inbound Delivery: C/O 27, C/T 173, Availability 71, FPY 11, Operator 1
Typing: C/O 33, C/T 174, Availability 31, FPY 0, Operator 1
Unloading: C/O 57, C/T 16, Availability 51, FPY 15, Operator 1
Checking: C/O 36, C/T 12, Availability 39, FPY 12, Operator 1

Waiting Time 19
Waiting Time 0
Waiting Time 21

Source: The Author (2023)

Figure 64 - Current State Mapping extended visualization



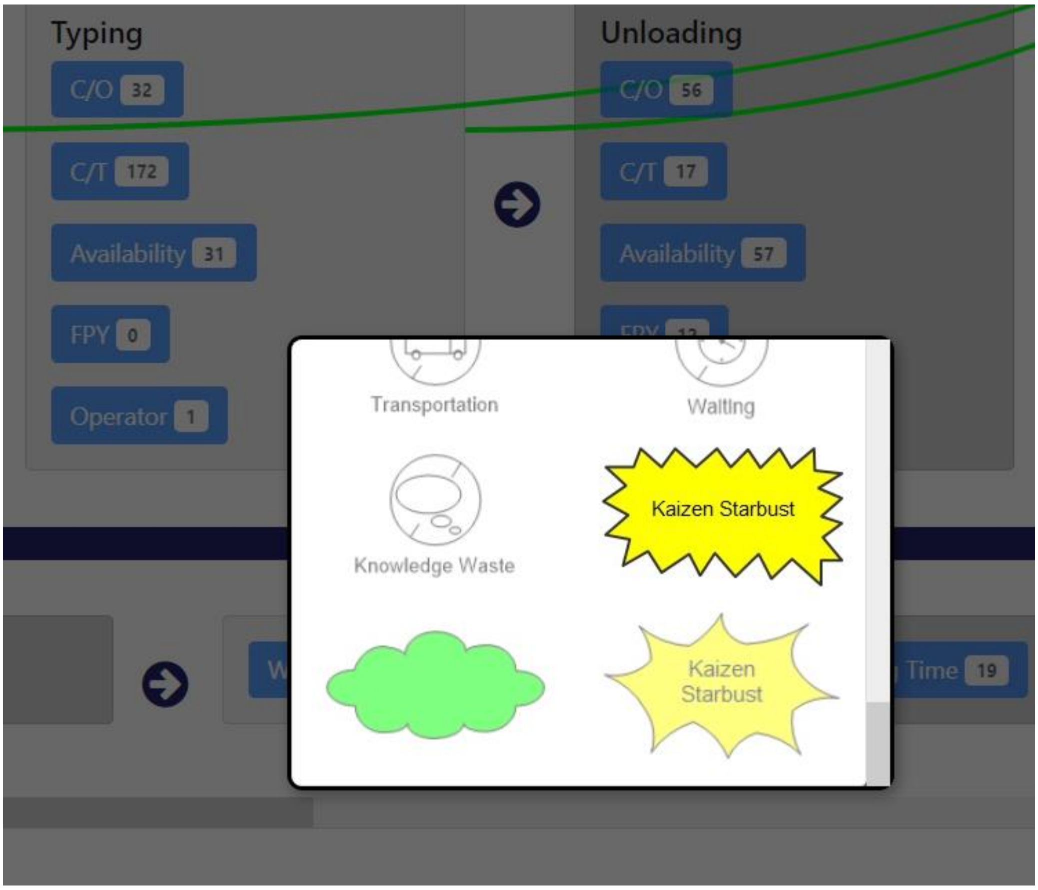
Start: Supplier
End: Customer

Inbound Delivery: C/O 27, C/T 173, Availability 71, FPY 11, Operator 1
Typing: C/O 33, C/T 174, Availability 31, FPY 0, Operator 1
Unloading: C/O 57, C/T 16, Availability 51, FPY 15, Operator 1
Checking: C/O 36, C/T 12, Availability 39, FPY 12, Operator 1
Storing: C/O 48, C/T 13, Availability 47, FPY 127, Operator 1
Requesting: C/O 7, C/T 20, Availability 66, FPY 0, Operator 1

Waiting Time 19
Waiting Time 0
Waiting Time 21
Waiting Time 17
Waiting Time 211

Source: The Author (2023)

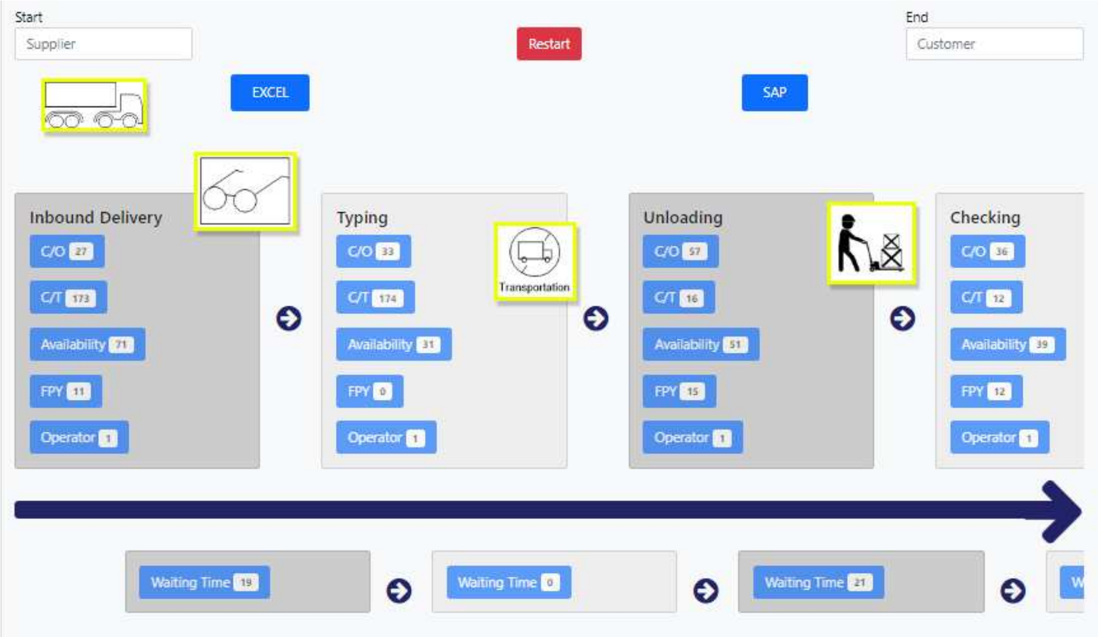
Figure 65 - Right Button Click Functionality



Source: The Author (2023)

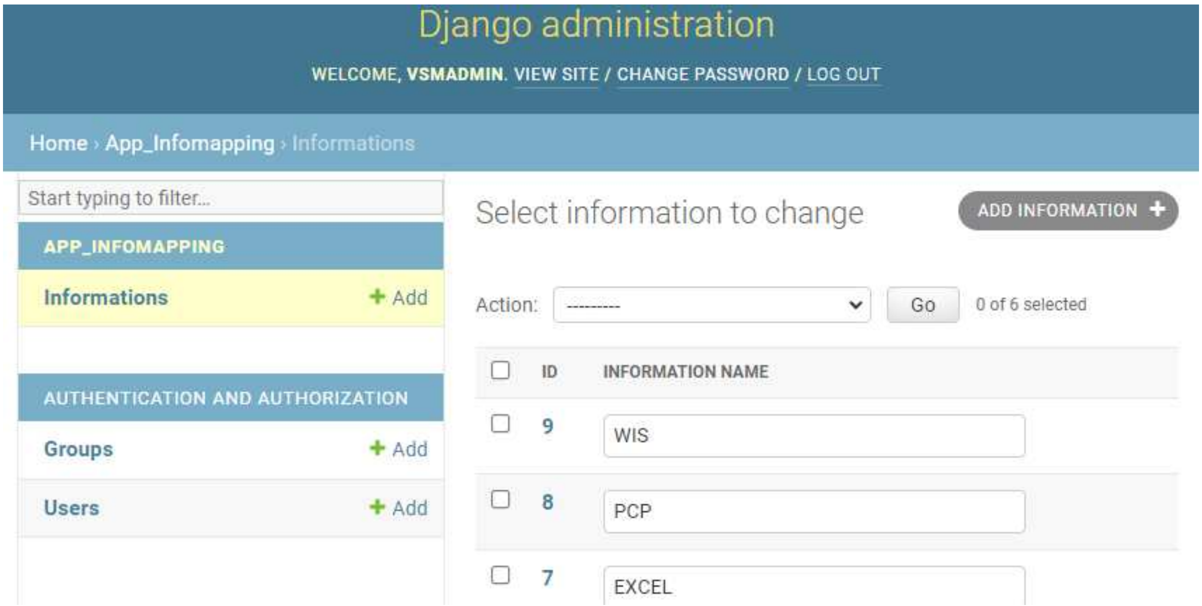
The dynamic_topbuttons function operates in the same prodmapping.html template but uses the data in the table called "Information" created by the app_infomapping application. Therefore, this functionality acts with the union of two applications. Like the dynamic_mapping function, the result is converted into a JSON format so that other functions in JS can use it and finally returned in an HTTP response and consequently linked to the HTML template code lines. The Django administrative area performs the insertion and modification of data related to the information mapping activities. This area can easily be accessed by clicking the "Include Information Items" button at the top of the page (Figure 53). Within the administrative area, by clicking on Information > Add Information, typing the name of the new item, and then clicking Save, new items will be included (Figure 67 and Figure 68). All item names can be edited with just a double-click on each name.

Figure 66 – VSM Symbols



Source: The Author (2023)

Figure 67 – Adding new Information items

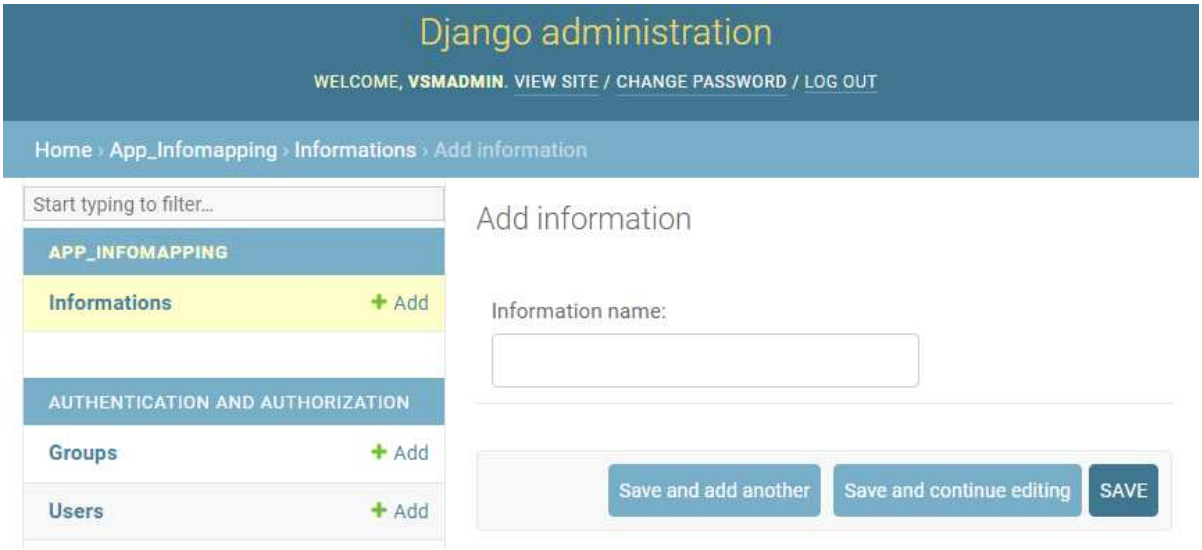

Django administration
WELCOME, VSMADMIN. VIEW SITE / CHANGE PASSWORD / LOG OUT
Home › App_Infomapping › Informations
Start typing to filter...
APP_INFOMAPPING
Informations + Add
AUTHENTICATION AND AUTHORIZATION
Groups + Add
Users + Add
ADD INFORMATION +
Action: ---------
Go
0 of 6 selected

Select information to change

□	ID	INFORMATION NAME
□	9	WIS
□	8	PCP
□	7	EXCEL

Source: The Author (2023)

Figure 68 – Information item name field


Django administration
WELCOME, VSMADMIN. VIEW SITE / CHANGE PASSWORD / LOG OUT
Home › App_Infomapping › Informations › Add information
Start typing to filter...
APP_INFOMAPPING
Informations
+ Add
AUTHENTICATION AND AUTHORIZATION
Groups
+ Add
Users
+ Add
Save and add another
Save and continue editing
SAVE

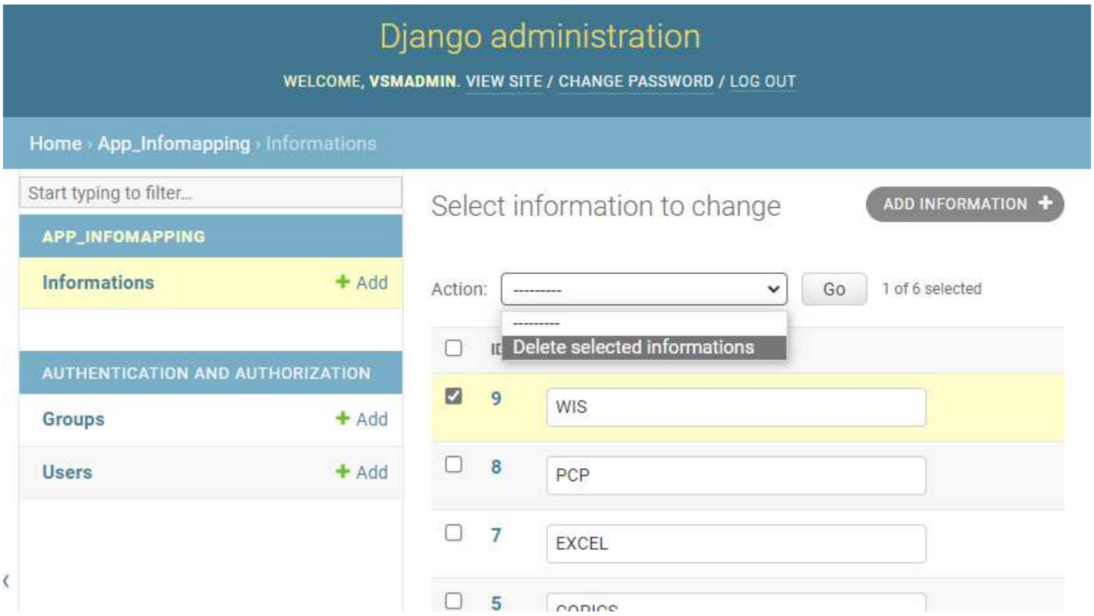
Add information

Information name:

Source: The Author (2023)

When deleting items, it will be necessary to select the desired items, change the "Action" field to "Delete selected information," and click on "Go" (Figure 69).

Figure 69 – Deleting an Information item


Django administration
WELCOME, VSMADMIN. VIEW SITE / CHANGE PASSWORD / LOG OUT
Home › App_Infomapping › Informations
Start typing to filter...
APP_INFOMAPPING
Informations
+ Add
AUTHENTICATION AND AUTHORIZATION
Groups
+ Add
Users
+ Add
ADD INFORMATION +
Go

Delete selected informations

Select information to change

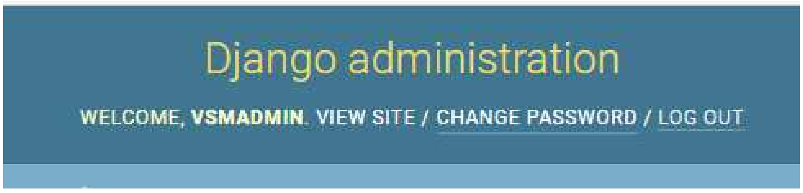
Action: --------- 1 of 6 selected

□	ID	
☑	9	WIS
□	8	PCP
□	7	EXCEL
□	5	COPICS

Source: The Author (2023)

When it is necessary to return to the main page, it is necessary to select "View Site" in the upper right corner (Figure 70).

Figure 70 – Django Administration exit



Django administration

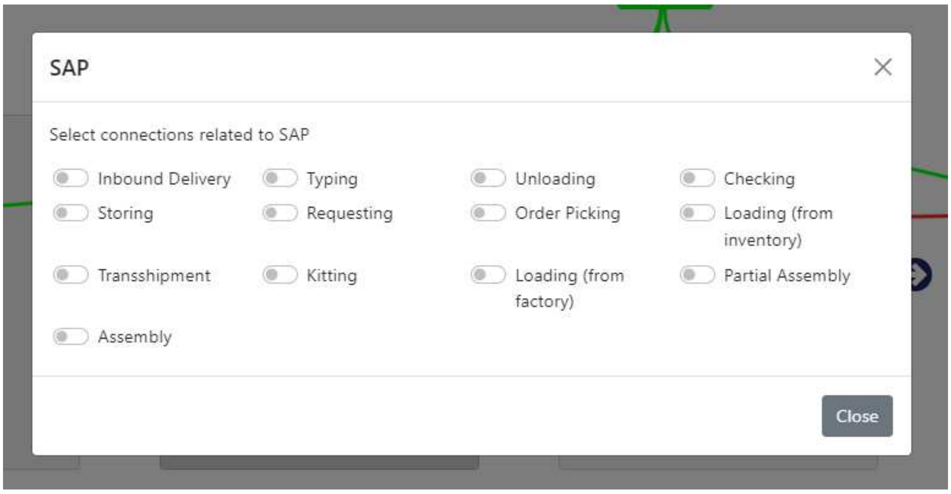
WELCOME, **VSMADMIN**. VIEW SITE / CHANGE PASSWORD / LOG OUT

Source: The Author (2023)

All items on the Django Administration page are displayed on the Current State Mapping page in the same order as the IDs (Identification). If it is necessary to change the order of any of these items, the user must edit the names so that they have different IDs, and thus the order is also changed on the main page.

In a VSM map, it is vital to identify the names of the information systems existing in a company's value chain and to identify which operations these information systems are connected to. Thus, some JS functions were implemented so that a modal appappears on the screen with just a click on any of the information frames, demonstrating th the alternatives for connecting existing operations (Figure 71). A modal is a UI element that appears as a pop-up window on the screen.

Figure 71 – Connection Selection Modal



SAP

Select connections related to SAP

○ Inbound Delivery ○ Typing ○ Unloading ○ Checking

○ Storing ○ Requesting ○ Order Picking ○ Loading (from inventory)

○ Transshipment ○ Kitting ○ Loading (from factory) ○ Partial Assembly

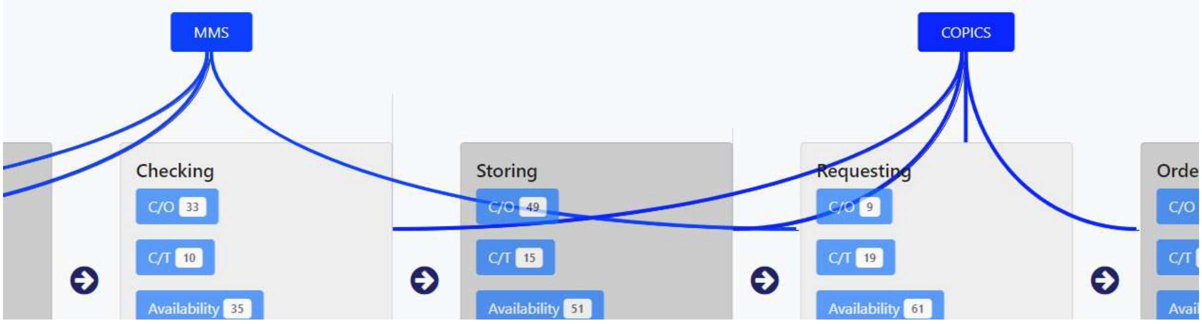
○ Assembly

Source: The Author (2023)

The selection options in the modal are related to the exclusive title options identified in the database, so if there are more or fewer titles, this will be dynamically changed in the modal.

In the example shown in Figure 72, three connections are made to the MMS (Manufacturing Message Specification) block to demonstrate this functionality, and four are made to the COPICS (Construction Operation Building Information Exchange for Product and Systems Data) block.

Figure 72 – Information items connecting to Operation blocks

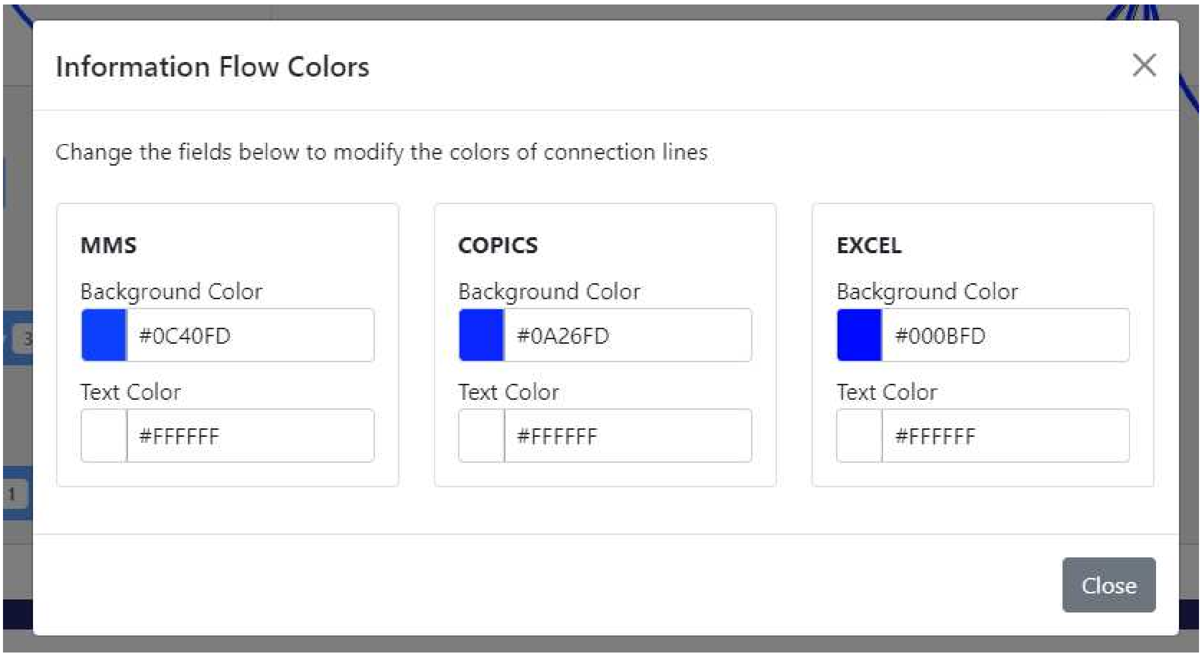


Source: The Author (2023)

In some situations, there may be an overlap between lines of the same color. Then, another JS function was applied to allow the change of connection line colors and, consequently, the information block and its text. This function is activated by clicking on a button in the upper right corner of the screen, which has a brush format (Figure 73, Figure 74, and Figure 75). Only the change of color for information blocks that already have a connection with an operation block will be allowed. By clicking on the "Background Color" or "Text Color" field, a color spectrum is displayed for the user to choose any of them freely.

Since the user can close the browser at any time or decide to analyze this mapping over several days, local storage was configured in the project's programming lines. Local storage is part of web storage that allows data to be stored on the client side persistently, even after the browser window is closed. Local storage is a more secure and efficient data storage option than cookies because it stores large amounts of data and is not sent to the server with each HTTP request. Local storage can be accessed through JS on the web page, making it a valuable and essential option for storing information. Therefore, all configurations developed on the web, such as the addition of symbols, connections, colors, and anything else configured, will always persist. When it is necessary to restart the entire process to make new changes, especially with additions and edits of information blocks, the only necessary action will be to click the red "Restart" button in the center of the screen.

Figure 73 – Information Flow Colors Modal after clicking on Brush button



Information Flow Colors

Change the fields below to modify the colors of connection lines

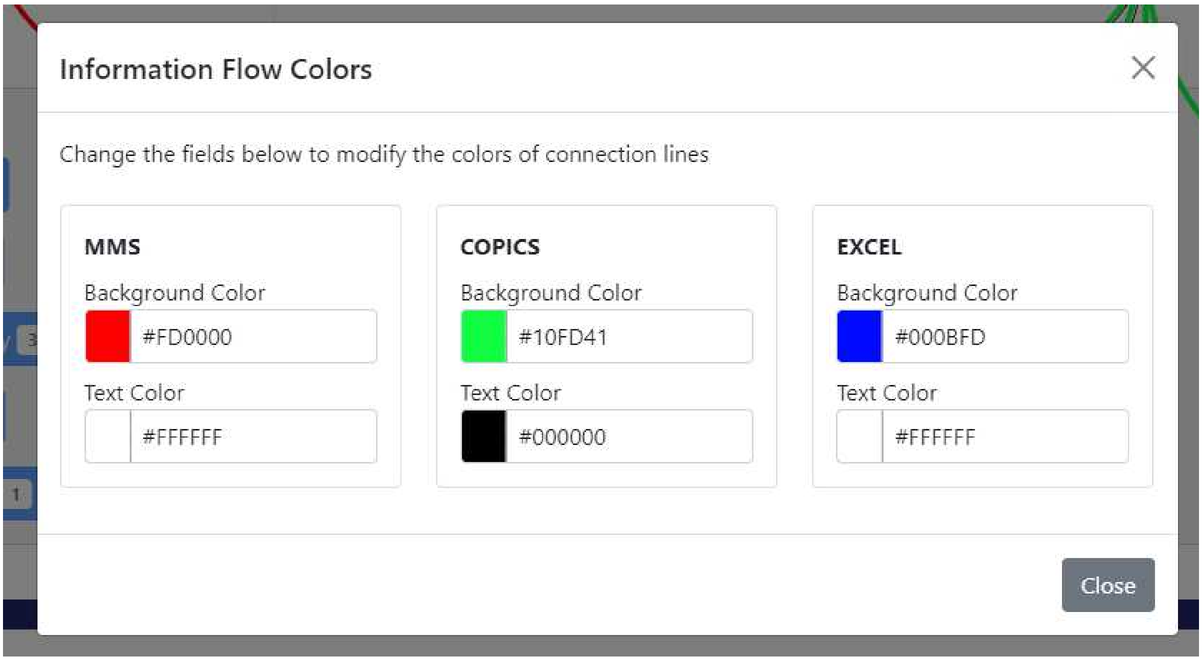
MMS
Background Color: #0C40FD
Text Color: #FFFFFF

COPICS
Background Color: #0A26FD
Text Color: #FFFFFF

EXCEL
Background Color: #000BFD
Text Color: #FFFFFF

Source: The Author (2023)

Figure 74 – Changing Background Colors and Text Colors for Information Items



Information Flow Colors

Change the fields below to modify the colors of connection lines

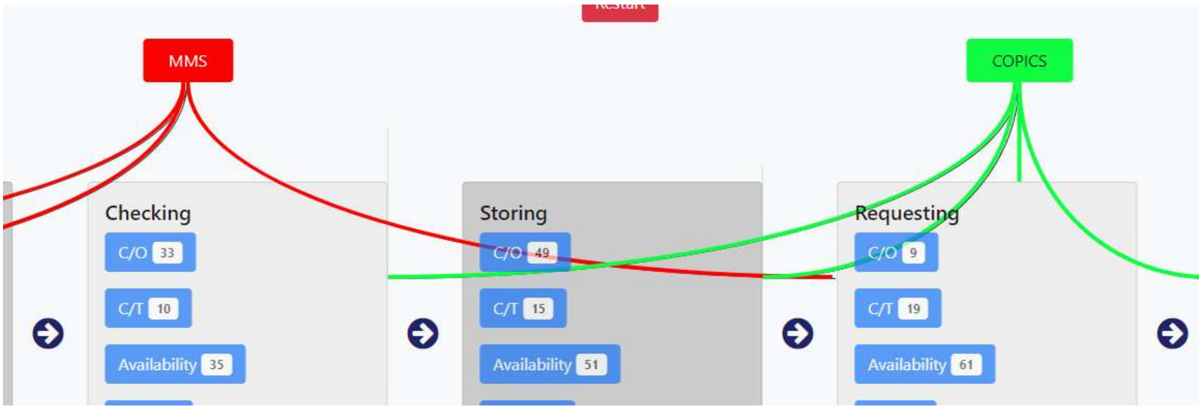
MMS
Background Color: #FD0000
Text Color: #FFFFFF

COPICS
Background Color: #10FD41
Text Color: #000000

EXCEL
Background Color: #000BFD
Text Color: #FFFFFF

Source: The Author (2023)

Figure 75 – Color Change Result



Source: The Author (2023)

The future state mapping was developed using the app_prediction application, linked to another HTML template called predict.html. In this application, seven functions were created in Python, which interacts with many other functions in JS.

The user must click the "Future State" button to access the future mapping page, as shown in Figure 53. The future state page has a different approach from the previous one, as it needs functions to predict the future using ML algorithms to develop future mapping. Therefore, a large amount of data existing in the database is essential.

This page can be divided into five functional areas (Figure 76): (i) buttons for transitioning to other pages (Figure 77), located at the top; (ii) chart and primary structure of the future state mapping; (iii) form for inserting new operation blocks (Figure 78), which will appear in the primary structure shown in Figure 59; (iv) form for operation prediction (

Figure **79**); (v) form for predicting parameters in a single operation (Figure 80).

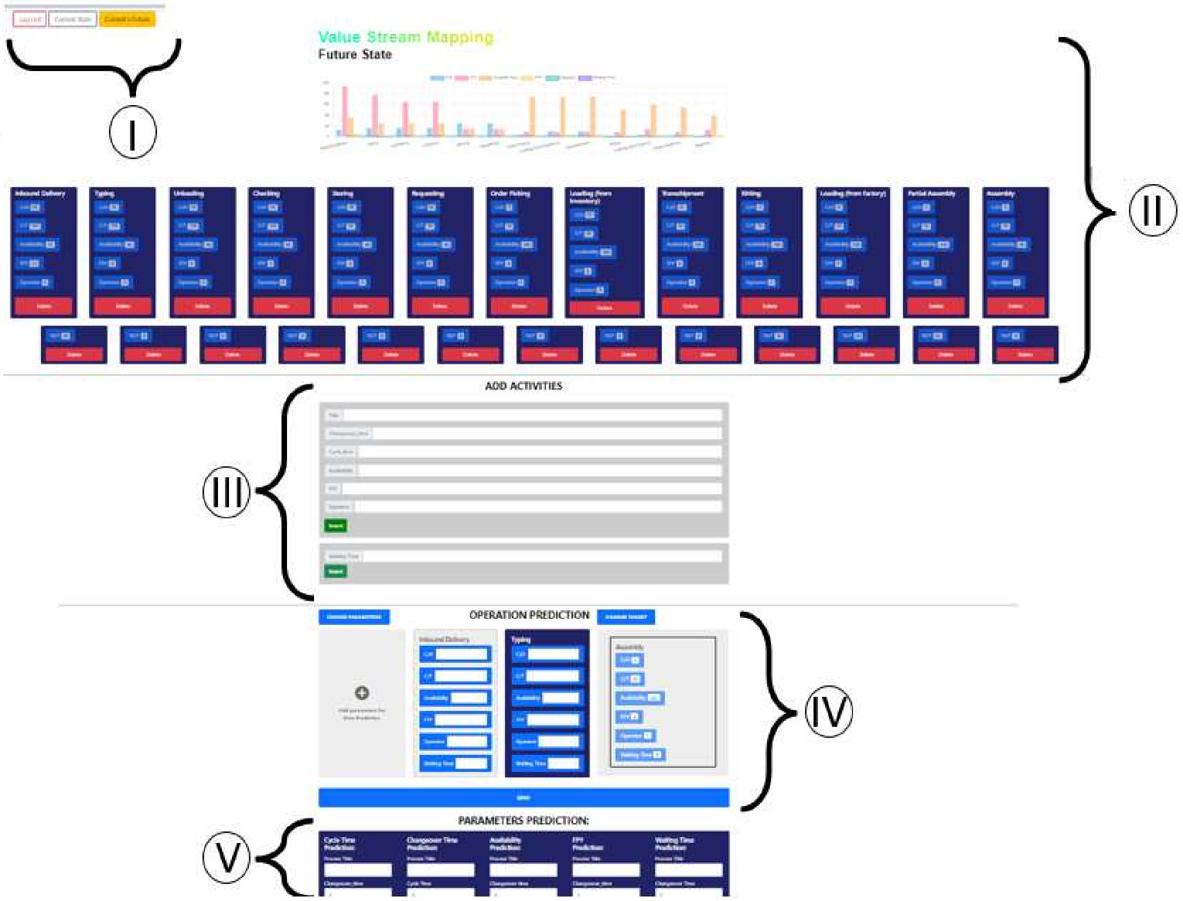
In the View layer, the seven Python functions were named: index, predict, result1, result2, result3, result4, and result5. The predict function was only used to render the main HTML template of this application. The index function uses Python to access the data in the database and perform predictions based on multiple regression models. Together with some JS functions, the configuration and response of these predictions are straightforward and dynamic.

This operation prediction allows the user to identify if changing one or some parameters in initial operations can positively or negatively influence the parameters of subsequent operations**.**

By clicking on "Add parameter for Data Prediction" (

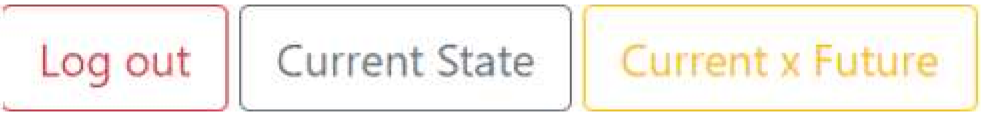
Figure **79**), the user can define how many operations they want to include and which parameters they will change concerning the current parameters presented in the current state mapping, and then by clicking on "Add the target for Data Prediction" the user will need to define which will be the target operation that they need the algorithm to perform this operation (Figure 81).

Figure 76 - Full Future State page



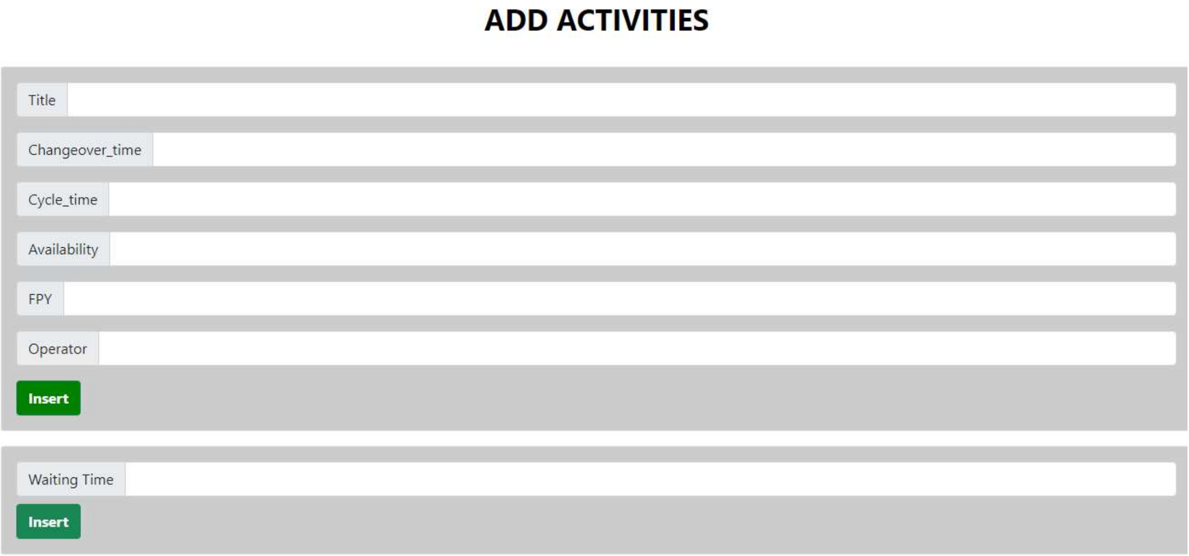
Source: The Author (2023)

Figure 77 - Future State page Buttons for transitioning to other pages



Source: The Author (2023)

Figure 78 – Form to add activities



ADD ACTIVITIES

Title

Changeover_time

Cycle_time

Availability

FPY

Operator

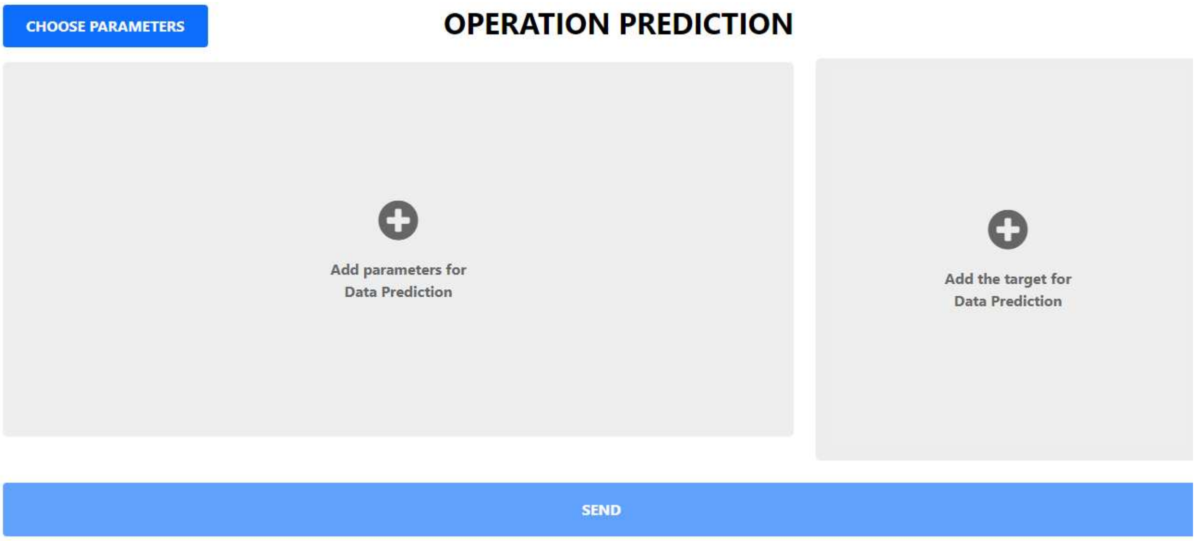
Insert

Waiting Time

Insert

Source: The Author (2023)

Figure 79 – Future Operation Prediction



CHOOSE PARAMETERS

OPERATION PREDICTION

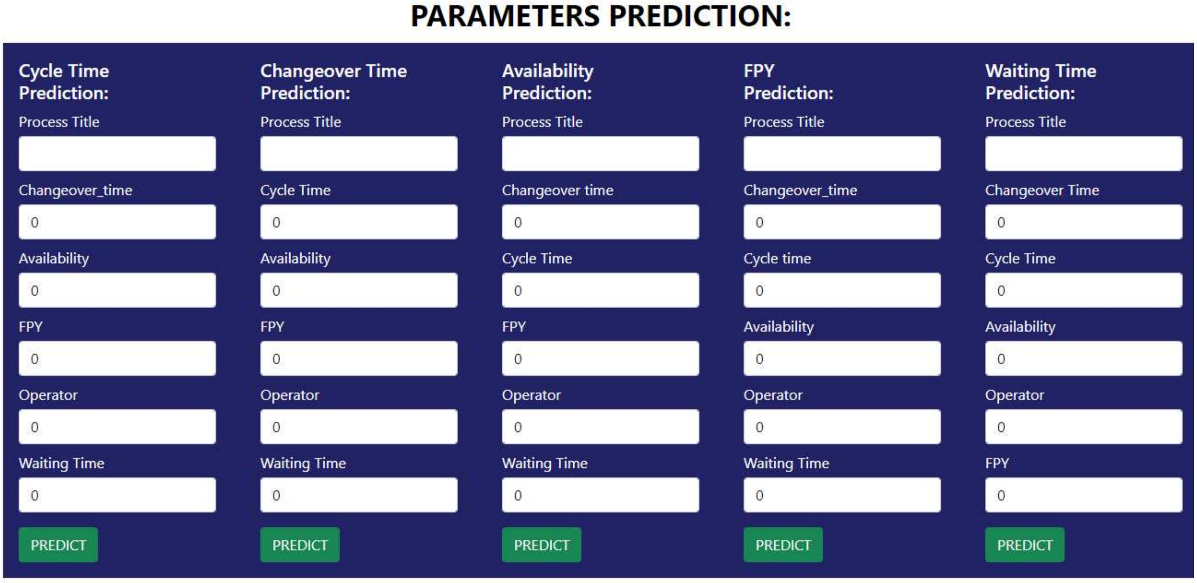
Add parameters for Data Prediction

Add the target for Data Prediction

SEND

Source: The Author (2023)

Figure 80 – Future parameters prediction



PARAMETERS PREDICTION:

Cycle Time Prediction:

Process Title:

Changeover_time: 0

Availability: 0

FPY: 0

Operator: 0

Waiting Time: 0

Changeover Time Prediction:

Process Title:

Cycle Time: 0

Availability: 0

FPY: 0

Operator: 0

Waiting Time: 0

Availability Prediction:

Process Title:

Changeover time: 0

Cycle Time: 0

FPY: 0

Operator: 0

Waiting Time: 0

FPY Prediction:

Process Title:

Changeover_time: 0

Cycle time: 0

Availability: 0

Operator: 0

Waiting Time: 0

Waiting Time Prediction:

Process Title:

Changeover Time: 0

Cycle Time: 0

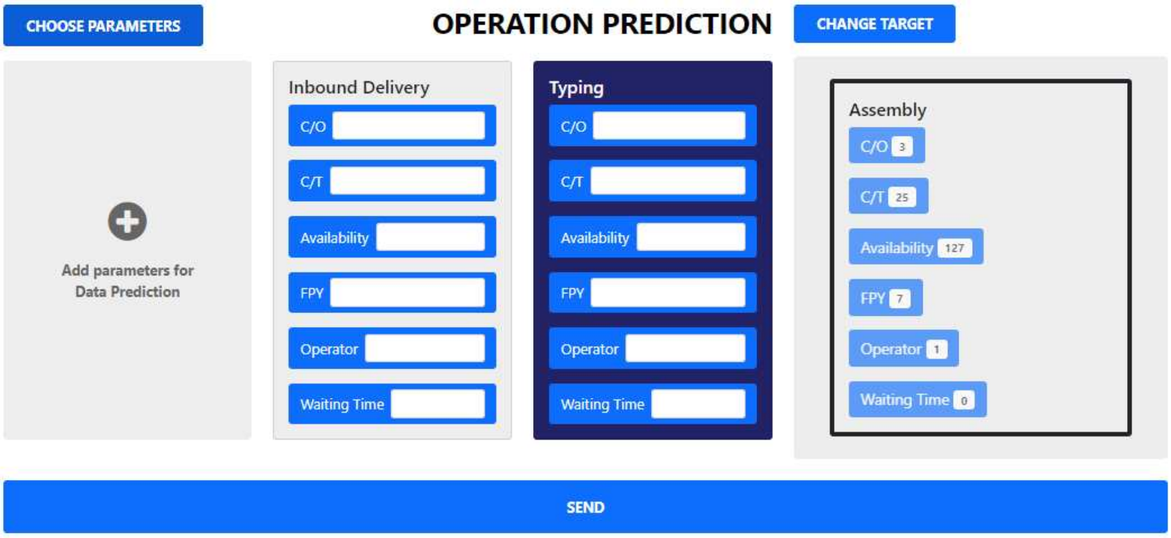
Availability: 0

Operator: 0

FPY: 0

Source: The Author (2023)

Figure 81 – Data Administration for Operation Prediction



OPERATION PREDICTION

Inbound Delivery

C/O:

C/T:

Availability:

FPY:

Operator:

Waiting Time:

Typing

C/O:

C/T:

Availability:

FPY:

Operator:

Waiting Time:

Assembly

C/O 3

C/T 25

Availability 127

FPY 7

Operator 1

Waiting Time 0

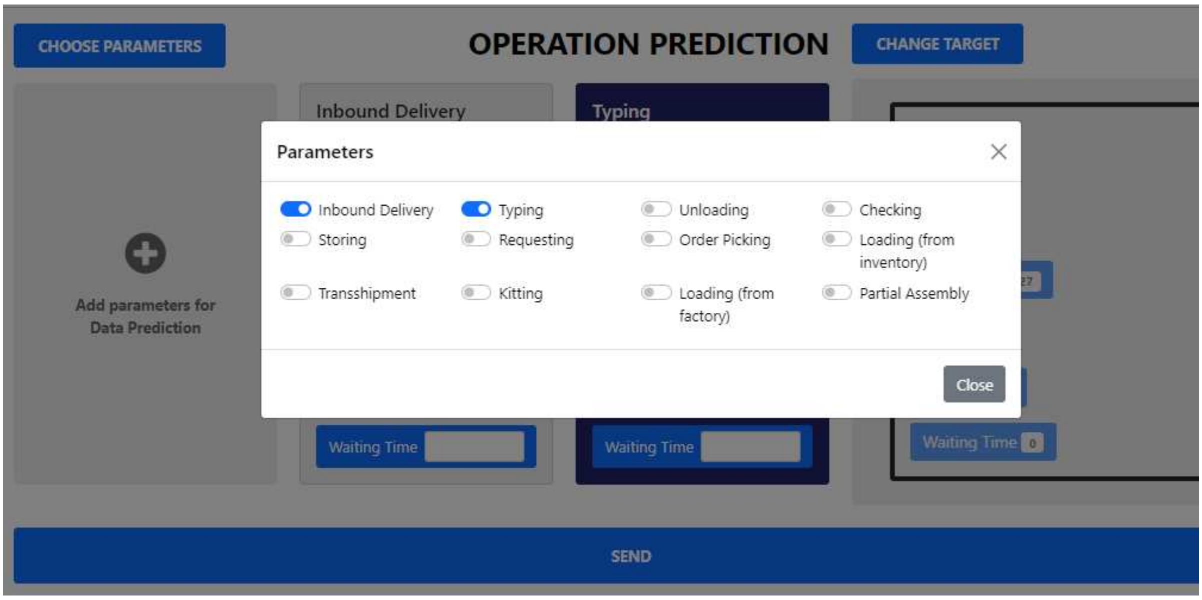
Source: The Author (2023)

Suppose the user needs to include more operations in the parameters area or even remove operations. In that case, the user should click the "Choose Parameters" option and make the desired selection changes (Figure 82).

When including Operations in the parameters area, the fields of each of the variables need to be filled with the desired value, but when including the target operation, it displays the operation with the data it currently has, i.e., the same data that is displayed for this operation in the current state mapping (Figure 81).

After filling in all the data for the operations included in the parameters area and clicking the "Send" button, the data is sent to the ML algorithm. After a few seconds, the data displayed in the target area change, and to indicate to the user that this change has occurred, the target frame temporarily changes to yellow (Figure 83) as a warning. After a few more seconds, it returns to blue.

Figure 82 – Choosing parameters for Operation Prediction

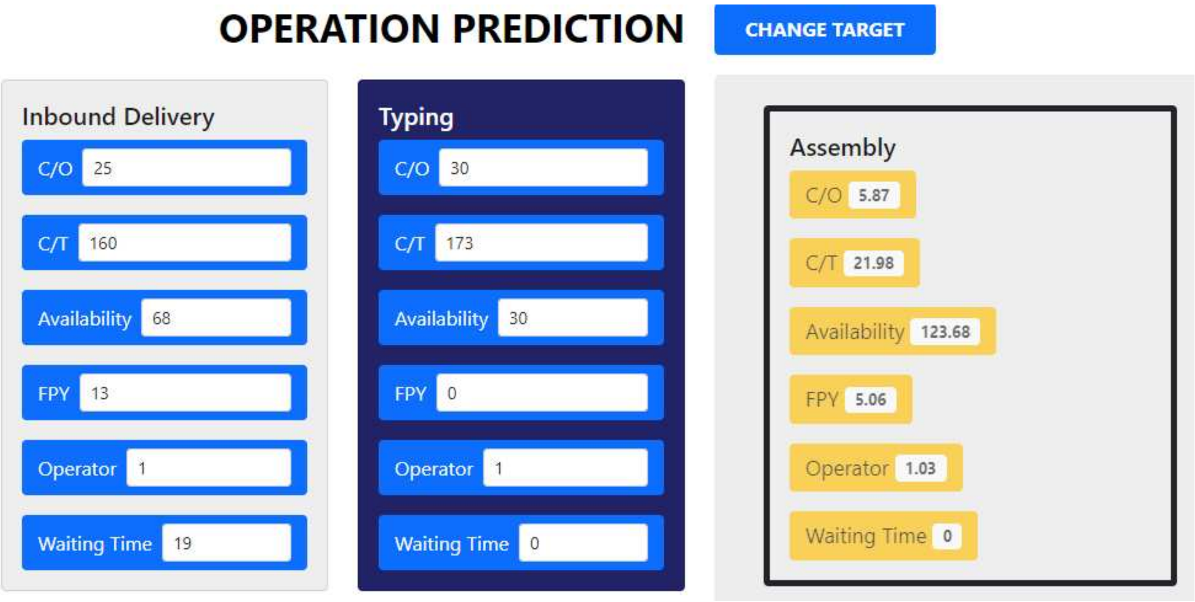


Parameters

- ● Inbound Delivery
- ● Typing
- ○ Unloading
- ○ Checking
- ○ Storing
- ○ Requesting
- ○ Order Picking
- ○ Loading (from inventory)
- ○ Transshipment
- ○ Kitting
- ○ Loading (from factory)
- ○ Partial Assembly

Source: The Author (2023)

Figure 83 – Results from Operation Prediction

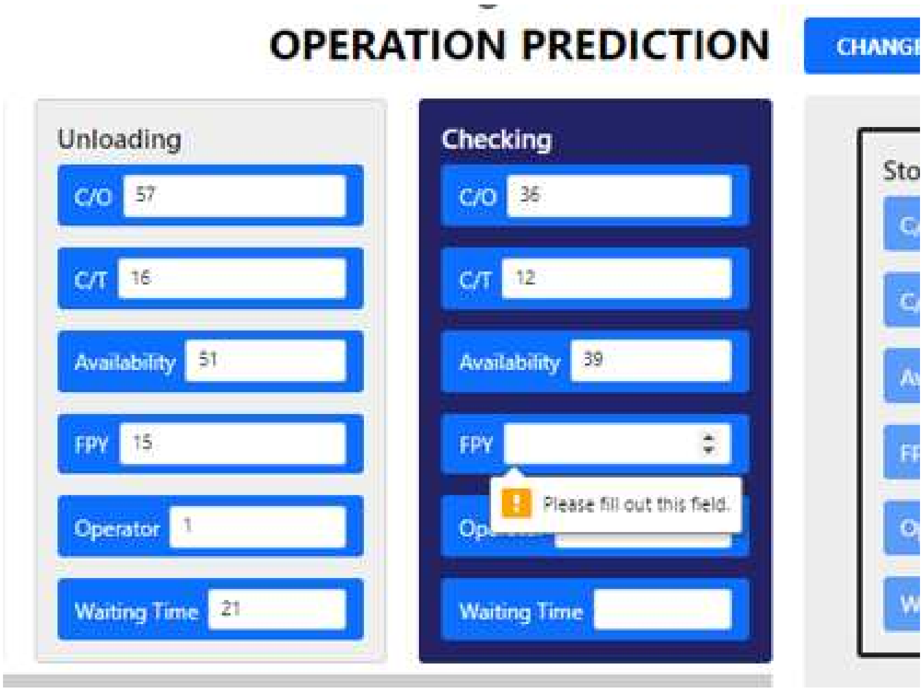


Inbound Delivery		Typing		Assembly	
C/O	25	C/O	30	C/O	5.87
C/T	160	C/T	173	C/T	21.98
Availability	68	Availability	30	Availability	123.68
FPY	13	FPY	0	FPY	5.06
Operator	1	Operator	1	Operator	1.03
Waiting Time	19	Waiting Time	0	Waiting Time	0

Source: The Author (2023)

If the user forgets to fill in any field and clicks “Send”, a message appears in the unfilled field with a warning “Please fill out this field” (Figure 84)

Figure 84 – Request for mandatory field completion



OPERATION PREDICTION

Unloading

C/O 57

C/T 16

Availability 51

FPY 15

Operator 1

Waiting Time 21

Checking

C/O 36

C/T 12

Availability 39

FPY

Please fill out this field.

Op

Waiting Time

Source: The Author (2023)

The data to be entered in the fields of the chosen operations will only be based on the current state mapping data, as the user will determine which data they would like to use as an experimental analysis to identify what the influence of this change would be, i.e., in search of alternatives for improving the process. The regression ML can always bring good results in its prediction if there is: (i) a good amount of relevant and clean training data, i.e., without outliers or missing values, (ii) an appropriate choice of parameters that have a strong correlation with the target, (iii) an appropriate choice of the regression model, taking into consideration the nature of the data and the relationship between the parameters and the target, (iv) cross-validation to evaluate the model's ability to generalize new data and accurately predict the target variation, and (v) a good evaluation of results, using appropriate metrics such as MSE and R^2.

All these items have been analyzed and evaluated with the help of the Jupyter Notebook platform, which has already been integrated with Docker Desktop, i.e., it has access to the database and can be accessed through port 8890 in the browser. In this development, 20% of the data was used for testing and 80% for training.

Thus, six ML models were used, and cross-validation was run on each of them, which is a technique for evaluating the performance of an ML model on training data, to assess the model's capability. The prediction results were also evaluated through four appropriate metrics for assessing the model's performance on test data.

The six models used were: Random Forest Regressor, Kneighbors Regressor, Linear Regression, SVR, MLP Regressor, and Gradient Boosting Regressor. The four metrics used were: MAE, MSE, R^2, and RMSE (Root Mean Squared Error).

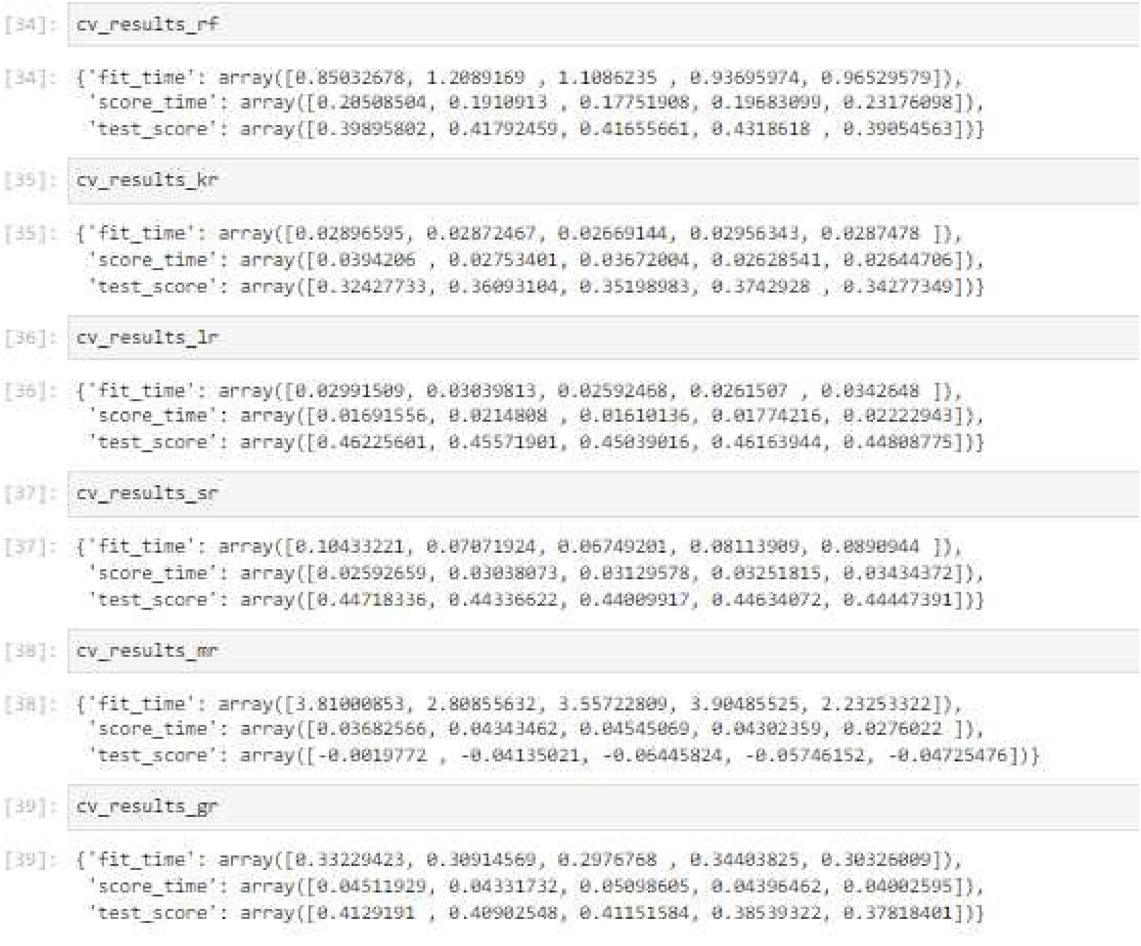
As operation predictions are predictions with multiple regression with six simultaneous predictions (i.e., Changeover Time, Cycle Time, Availability, FPY, Operator, and Waiting Time), each of the six models described in the previous paragraph is run in conjunction with the Multi-Output Regressor. The Multi Output Regressor is used for predicting multiple outputs but works as a wrapper for another regression model, which will be used as the basis for the prediction. This base model is indicated as an argument for the constructor of the MultiOutputRegressor class.

In search of the best model for future mapping, cross-validation was performed for all six chosen models, and the results can be viewed in Figure 85. All variables shown have the name "cv_results_XX", with the last two letters representing each model, i.e., rf for Random Forest Regressor, kr for Kneighbors Regressor, lr for Linear Regression, sr for SVR, mr for MLP Regressor, and gr for Gradient Boosting Regressor.

The fit_time metric represents the training time for each model, and the best model in this metric was Linear Regressor, followed by Kneighbors and Gradient Boosting, with similar and higher times than the other models. The score_time metric represents the time to evaluate the model score, and Linear is also the best model, followed by SVR and Kneighbors. The test_score metric represents the model's score for each of the five iterations performed by cross-validation, with the best model being Linear with 45%, followed by SVR with 44%, and Random Forest with 42%, while the MLP showed the worst results.

Then, the best mean score obtained by the Linear Regression in multiple regression was 45%. This percentage represents the average accuracy of the model concerning the validation data used during cross-validation. Usually, the higher the score, the better the model fits the training data and has a higher chance of performing well with new data. The value of 45% occurred because the model has six outputs, and the increase in these outputs can lead to additional complexity in the prediction problem. However, other factors can affect the test score, such as the size of the data sample, the complexity of the model, the quality of the training data, and the presence of outliers. In some cases, a value of 45% may be considered relatively low, while in others, it may be considered relatively high. For this reason, it is essential to compare this value with the performance of other models for the same problem and consider the user's expectations and needs for the model's performance.

Figure 85 – Cross-validation results for multiple regression



```
[34]: cv_results_rf

[34]: {'fit_time': array([0.85032678, 1.2089169 , 1.1086235 , 0.93695974, 0.96529579]),
       'score_time': array([0.20508504, 0.1910913 , 0.17751908, 0.19683099, 0.23176098]),
       'test_score': array([0.39895802, 0.41792459, 0.41655661, 0.4318618 , 0.39054563])}

[35]: cv_results_kr

[35]: {'fit_time': array([0.02896595, 0.02872467, 0.02669144, 0.02956343, 0.0287478 ]),
       'score_time': array([0.0394206 , 0.02753401, 0.03672004, 0.02628541, 0.02644706]),
       'test_score': array([0.32427733, 0.36093104, 0.35198983, 0.3742928 , 0.34277349])}

[36]: cv_results_lr

[36]: {'fit_time': array([0.02991509, 0.03039813, 0.02592468, 0.0261507 , 0.0342648 ]),
       'score_time': array([0.01691556, 0.0214808 , 0.01610136, 0.01774216, 0.02222943]),
       'test_score': array([0.46225601, 0.45571901, 0.45039016, 0.46163944, 0.44808775])}

[37]: cv_results_sr

[37]: {'fit_time': array([0.10433221, 0.07071924, 0.06749201, 0.08113909, 0.0890944 ]),
       'score_time': array([0.02592659, 0.03038073, 0.03129578, 0.03251815, 0.03434372]),
       'test_score': array([0.44718336, 0.44336622, 0.44009917, 0.44634072, 0.44447391])}

[38]: cv_results_mr

[38]: {'fit_time': array([3.81000853, 2.80855632, 3.55722809, 3.90485525, 2.23253322]),
       'score_time': array([0.03682566, 0.04343462, 0.04545069, 0.04302359, 0.0276022 ]),
       'test_score': array([-0.0019772 , -0.04135021, -0.06445824, -0.05746152, -0.04725476])}

[39]: cv_results_gr

[39]: {'fit_time': array([0.33229423, 0.30914569, 0.2976768 , 0.34403825, 0.30326009]),
       'score_time': array([0.04511929, 0.04331732, 0.05098605, 0.04396462, 0.04002595]),
       'test_score': array([0.4129191 , 0.40902548, 0.41151584, 0.38539322, 0.37818401])}
```

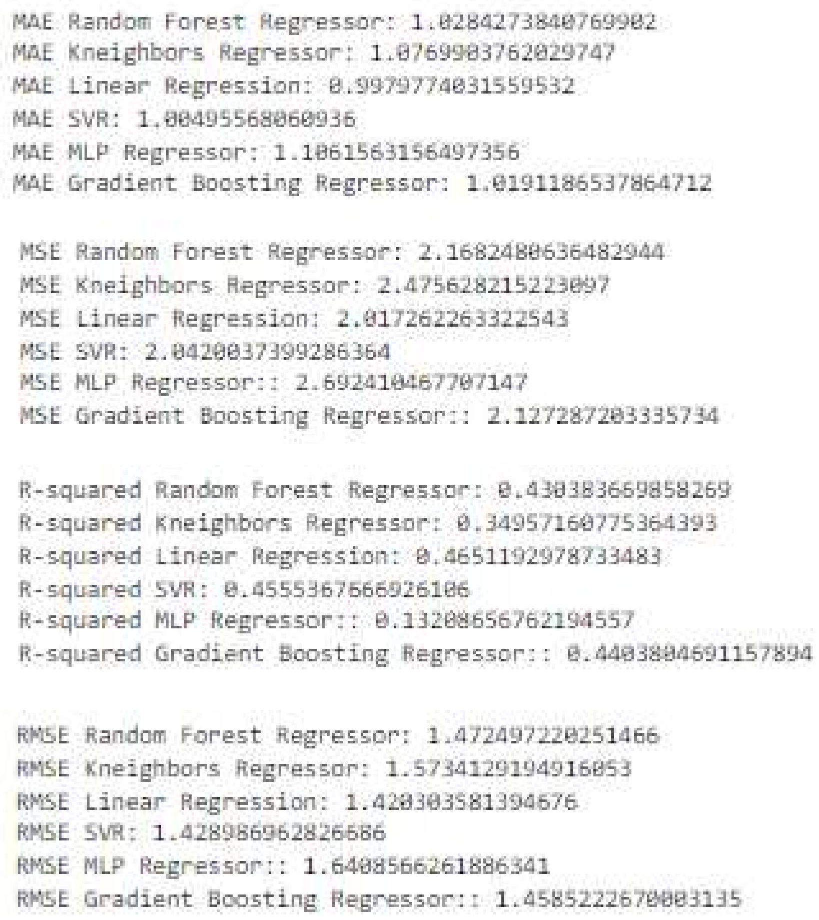
Source: The Author (2023)

At first, 30,000 rows of data were analyzed with data for one operator, and the best value obtained by linear regression was 32%. With a change to approximately 15,000 rows of data for one operator and 15,000 rows for two operators, the model's accuracy value significantly improved, confirming that data variability is as essential as data quantity. Detailed analysis of the problem and evaluation with other metrics are necessary to obtain a more meaningful conclusion about the model quality that can be used.

In addition to the cross-validation results, four metrics for regression evaluation in the test data were applied to the six ML models, and the results can be viewed in Figure 86.

From these metrics, which are appropriate for evaluating regression models, it can be concluded that the performance of the models is regular, with R^2 varying between 0.17 and 0.33, indicating that they explain only a tiny portion of the data variation. RMSE varies between 1.47 and 1.64, meaning that the average difference between the models' predictions and the valid values is around 1.5 units. However, the interpretation of these metrics depends on the context of the problem and business goals, so it is essential to evaluate whether these values are acceptable for the specific application.

Figure 86 – Metrics Results for Multiple Regression



```
MAE Random Forest Regressor: 1.0284273840769902
MAE Kneighbors Regressor: 1.0769903762029747
MAE Linear Regression: 0.9979774031559532
MAE SVR: 1.00495568060936
MAE MLP Regressor: 1.1061563156497356
MAE Gradient Boosting Regressor: 1.0191186537864712

 MSE Random Forest Regressor: 2.1682480636482944
 MSE Kneighbors Regressor: 2.475628215223097
 MSE Linear Regression: 2.017262263322543
 MSE SVR: 2.0420037399286364
 MSE MLP Regressor:: 2.692410467707147
 MSE Gradient Boosting Regressor:: 2.127287203335734

 R-squared Random Forest Regressor: 0.430383669858269
 R-squared Kneighbors Regressor: 0.34957160775364393
 R-squared Linear Regression: 0.4651192978733483
 R-squared SVR: 0.4555367666926106
 R-squared MLP Regressor:: 0.13208656762194557
 R-squared Gradient Boosting Regressor:: 0.4403804691157894

 RMSE Random Forest Regressor: 1.472497220251466
 RMSE Kneighbors Regressor: 1.5734129194916053
 RMSE Linear Regression: 1.420303581394676
 RMSE SVR: 1.428986962826686
 RMSE MLP Regressor:: 1.6408566261886341
 RMSE Gradient Boosting Regressor:: 1.4585222670003135
```

Source: The Author (2023)

Based on the presented metrics, the best-performing regression model in terms of error is Linear Regression, followed by Gradient Boosting Regressor and Random Forest. In terms of explaining the data, the best-performing model is Linear Regression, followed by Gradient Boosting Regressor and SVR. Thus, the currently configured model in the project's regression programming line is the Linear Regression model.

The result1, result2, result3, result4, and result5 functions use Python to access the data in the database and make predictions based on simple regression models. Together with some JS functions, the configuration and response of these predictions are straightforward and dynamic.

The lines of code related to simple regression ML are used for the parameter prediction area functionality. Figure 80 shows five columns with six input fields and a predict button in the last row. These five columns represent five parameter prediction systems. Each of these columns operates independently of the others, and the difference between them is the target parameter configured in the ML algorithm. The first column has Cycle time as the target prediction parameter, the second column has Changeover Time as the target prediction parameter, the third column has Availability as the target prediction parameter, the fourth column has FPY as the target prediction parameter, and the fifth and last column has Waiting Time as the target prediction parameter.

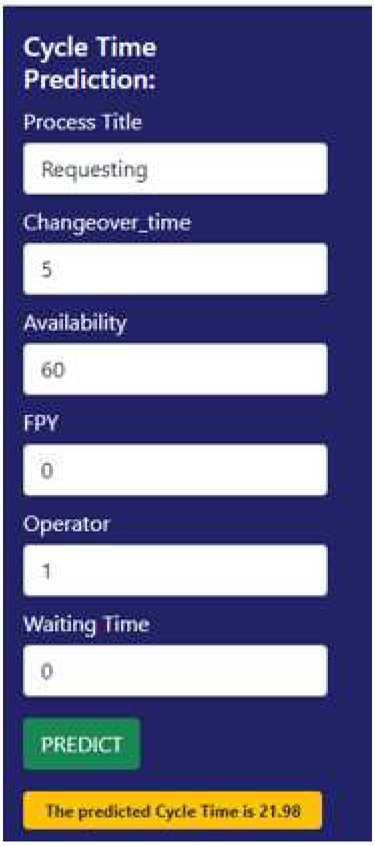
Unlike the multiple predictions, where it was necessary to fill in all the parameters of each selected operation to predict all the parameters of an operation chosen as the target, here, the focus of prediction will be only on a single parameter of a single operation. These five individual parameter predictors can complement the operation predictor to obtain the ideal future mapping for the entire process.

For example, supposing the user wants to analyze how changes in the Changeover Time, FPY, Availability, Operator, and Waiting Time parameters in the Requesting operation could increase or decrease Cycle Time. In that case, the user should take the first prediction system, enter all parameters, and click "predict". The result will be displayed just below the Predict button, as demonstrated in Figure 87.

In the current state mapping, it can be verified that in the Requesting operation, the Changeover Time is 7, the Availability is 66, and the Cycle Time is 20. Keeping all other data unchanged but changing the Changeover Time to 5 and the Availability to 60, the predictor indicated that the Cycle Time increased to 22.

The exact process can be carried out with the other four prediction systems, each with a different target. Just as capacity learning analyses were performed using regression algorithm performance metrics on the multiple regression prediction systems, the same analyses were performed on this simple regression prediction system.

Figure 87 – Result from Cycle Time Prediction



Cycle Time Prediction:

Process Title: Requesting

Changeover_time: 5

Availability: 60

FPY: 0

Operator: 1

Waiting Time: 0

The predicted Cycle Time is 21.98

Source: The Author (2023)

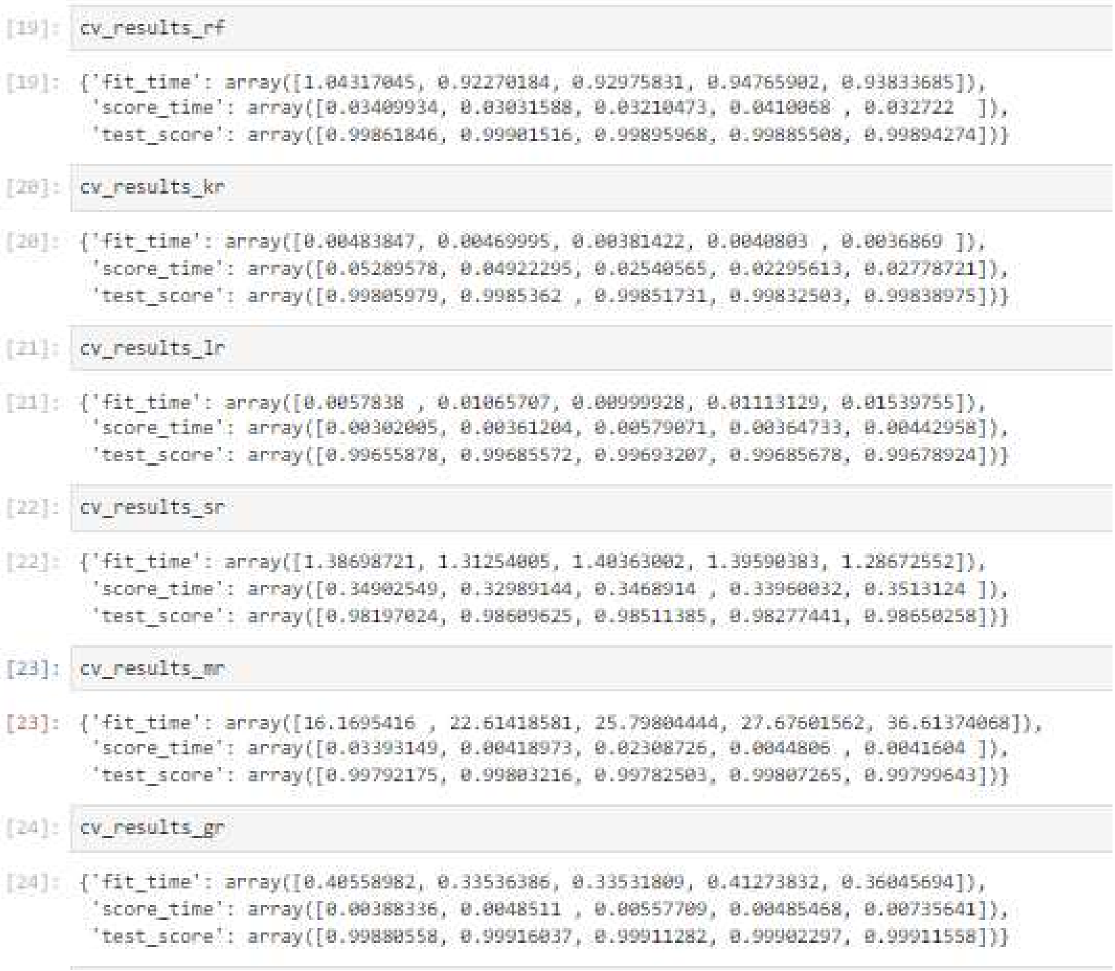
The same six models were used, with the difference that they are not placed as attributes in the MultiOutputRegressor model. The six models used are Random Forest Regressor, Kneighbors Regressor, Linear Regression, SVR, MLP Regressor, and Gradient Boosting Regressor. The results of the Cross Validation can be seen in Figure 88.

In the simple regression ML algorithm, the best model in terms of fit time was Linear Regressor, followed by Kneighbors and Gradient Boosting. The best model in terms of score time was Linear, followed by Kneighbors and Gradient Boosting, and the best model in terms of test score was Gradient Boosting, followed by Random Forest and Kneighbors. Although Linear Regression is not considered the best test score, it is very close to the other models, achieving a score of 99.7%.

The results of the four metrics for evaluating the model's performance on test data are shown in Figure 89.

Based on the obtained data, it can be concluded that the model with the lowest MAE is the Gradient Boosting with 1.234, the model with the lowest mean squared error is the Gradient Boosting with 2.623, the model with the highest coefficient of determination is the Gradient Boosting with 0.999. The model with the lowest root mean squared error is also Gradient Boosting with 1.62. Therefore, based on the obtained results, the model used in this project for the ML algorithms with simple regression was the Gradient Boosting Regressor.

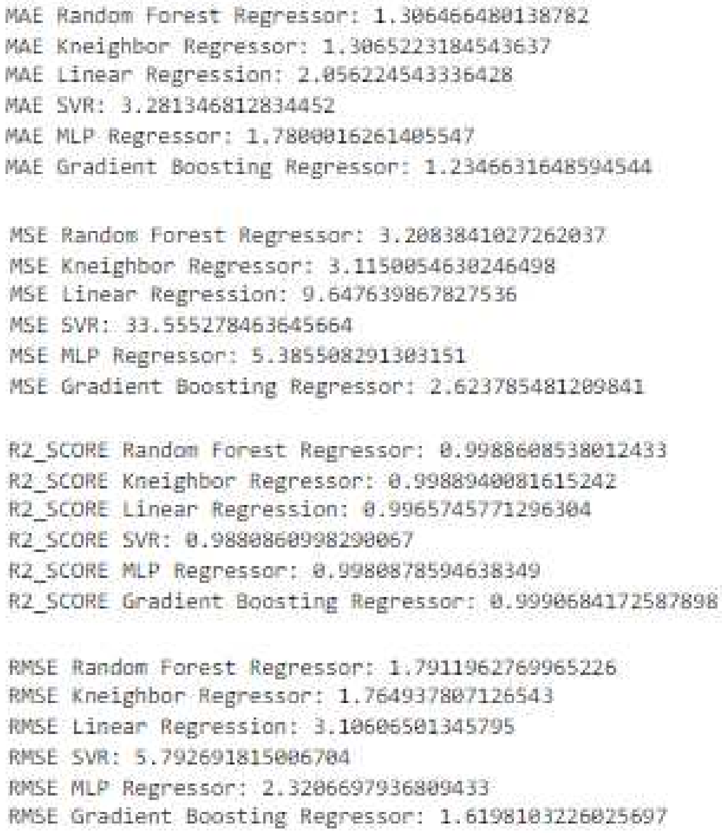
Figure 88 – Cross-Validation results for Simple Regression



```
[19]: cv_results_rf

[19]: {'fit_time': array([1.04317045, 0.92270184, 0.92975831, 0.94765902, 0.93833685]),
       'score_time': array([0.03409934, 0.03031588, 0.03210473, 0.0410068 , 0.032722  ]),
       'test_score': array([0.99861846, 0.99901516, 0.99895968, 0.99885508, 0.99894274])}

[20]: cv_results_kr

[20]: {'fit_time': array([0.00483847, 0.00469995, 0.00381422, 0.0040803 , 0.0036869 ]),
       'score_time': array([0.05289578, 0.04922295, 0.02540565, 0.02295613, 0.02778721]),
       'test_score': array([0.99805979, 0.9985362 , 0.99851731, 0.99832503, 0.99838975])}

[21]: cv_results_lr

[21]: {'fit_time': array([0.0057838 , 0.01065707, 0.00999928, 0.01113129, 0.01539755]),
       'score_time': array([0.00302005, 0.00361204, 0.00579071, 0.00364733, 0.00442958]),
       'test_score': array([0.99655878, 0.99685572, 0.99693207, 0.99685678, 0.99678924])}

[22]: cv_results_sr

[22]: {'fit_time': array([1.38698721, 1.31254005, 1.40363002, 1.39590383, 1.28672552]),
       'score_time': array([0.34902549, 0.32989144, 0.3468914 , 0.33960032, 0.3513124 ]),
       'test_score': array([0.98197024, 0.98609625, 0.98511385, 0.98277441, 0.98650258])}

[23]: cv_results_mr

[23]: {'fit_time': array([16.1695416 , 22.61418581, 25.79804444, 27.67601562, 36.61374068]),
       'score_time': array([0.03393149, 0.00418973, 0.02308726, 0.0044806 , 0.0041604 ]),
       'test_score': array([0.99792175, 0.99803216, 0.99782503, 0.99807265, 0.99799643])}

[24]: cv_results_gr

[24]: {'fit_time': array([0.40558982, 0.33536386, 0.33531809, 0.41273832, 0.36045694]),
       'score_time': array([0.00388336, 0.0048511 , 0.00557709, 0.00485468, 0.00735641]),
       'test_score': array([0.99880558, 0.99916037, 0.99911282, 0.99902297, 0.99911558])}
```

Source: The Author (2023)

Figure 89 – Metrics Results for Simple Regression



```
MAE Random Forest Regressor: 1.306466480138782
MAE Kneighbor Regressor: 1.3065223184543637
MAE Linear Regression: 2.056224543336428
MAE SVR: 3.281346812834452
MAE MLP Regressor: 1.7800016261405547
MAE Gradient Boosting Regressor: 1.2346631648594544

MSE Random Forest Regressor: 3.2083841027262037
MSE Kneighbor Regressor: 3.1150054630246498
MSE Linear Regression: 9.647639867827536
MSE SVR: 33.555278463645664
MSE MLP Regressor: 5.385508291303151
MSE Gradient Boosting Regressor: 2.623785481209841

R2_SCORE Random Forest Regressor: 0.9988608538012433
R2_SCORE Kneighbor Regressor: 0.9988940081615242
R2_SCORE Linear Regression: 0.9965745771296304
R2_SCORE SVR: 0.9880860998290067
R2_SCORE MLP Regressor: 0.9980878594638349
R2_SCORE Gradient Boosting Regressor: 0.9990684172587898

RMSE Random Forest Regressor: 1.7911962769965226
RMSE Kneighbor Regressor: 1.764937807126543
RMSE Linear Regression: 3.106065013457 95
RMSE SVR: 5.792691815006704
RMSE MLP Regressor: 2.3206697936809433
RMSE Gradient Boosting Regressor: 1.6198103226025697
```

Source: The Author (2023)

The development processing of the VSM application was performed on a computer with 16.0 GB of RAM and an i7 processor running at 2.60 GHz, and all results were obtained within a few seconds after each request. Since this system utilizes six containers in Docker, it is advisable to use a computer with a minimum of 16.0 GB RAM to achieve better performance when executing all of them simultaneously.

After analyzing the results of both prediction systems, it was identified that the table of 30,000 data lines used brought ideal accuracy for training and ideal accuracy for testing in the simple regression prediction system and regular accuracy for training and regular accuracy for testing in the multiple regression prediction system.

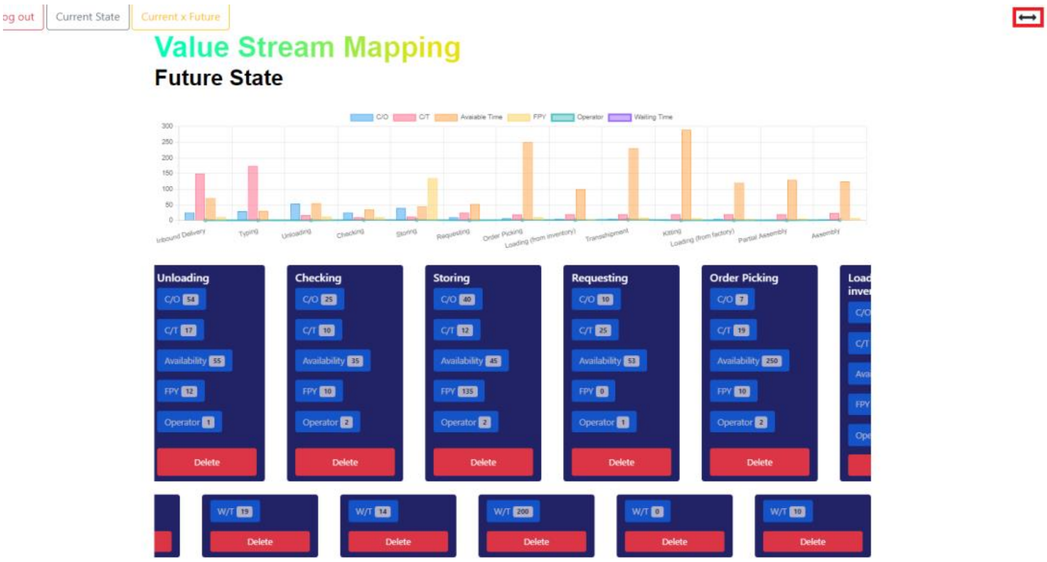
Although the accuracy of multiple regression is considered regular, the results obtained were identified as adequate to the reality of the source dataset. Increasing the amount of data alone does not guarantee that the multiple regression system can improve its accuracy, so it will be necessary to analyze beyond different quantities the variability and quality of the data with different approaches to obtain better performance and, consequently, better generalization to new data.

After conducting analyses to determine the best alternative for improving the process under study, the user must create the blocks for constructing the future state mapping. These blocks are created through the activity addition form shown in Figure 78. All fields must be filled in the correct order of the operations by clicking the two "Insert" buttons. The first includes the operation title, changeover time, cycle time, availability, FPY, and operator. In contrast, the second insert will consist of the waiting time. This separation in the insertion was performed due to the waiting time data being in separate blocks. If the user identifies any typing error after insertion, he may delete any operations he wants in this mapping, as seen in the red "Delete" button (Figure 90). This future state mapping has the same dynamics as the current state mapping, i.e., there is a maximum width centralized on the page at the same width as the chart above it, but when the user wants to expand this width, he can click the bidirectional arrow button located in the upper right corner of the page (Figure 90 and Figure 91).

At the same time each new operation is added, the entered data is also sent to the graphic area, which has precisely the same dynamics as the current state mapping chart.

Once the future state mapping has been completed with the same number of operations as the current state mapping, it will be possible to compare the charts on a new page (Figure 92) which can be accessed through the "Current x Future" button, located at the top of the page. This button is available on both the current state mapping page and the future state mapping page.

Figure 90 - Future State Mapping standard visualization



Value Stream Mapping

Future State

Unloading
C/O 54
C/T 17
Availability 55
FPY 12
Operator 1

Checking
C/O 25
C/T 10
Availability 35
FPY 10
Operator 2

Storing
C/O 40
C/T 12
Availability 45
FPY 135
Operator 2

Requesting
C/O 10
C/T 25
Availability 53
FPY 0
Operator 1

Order Picking
C/O 7
C/T 19
Availability 250
FPY 10
Operator 2

W/T 19

W/T 14

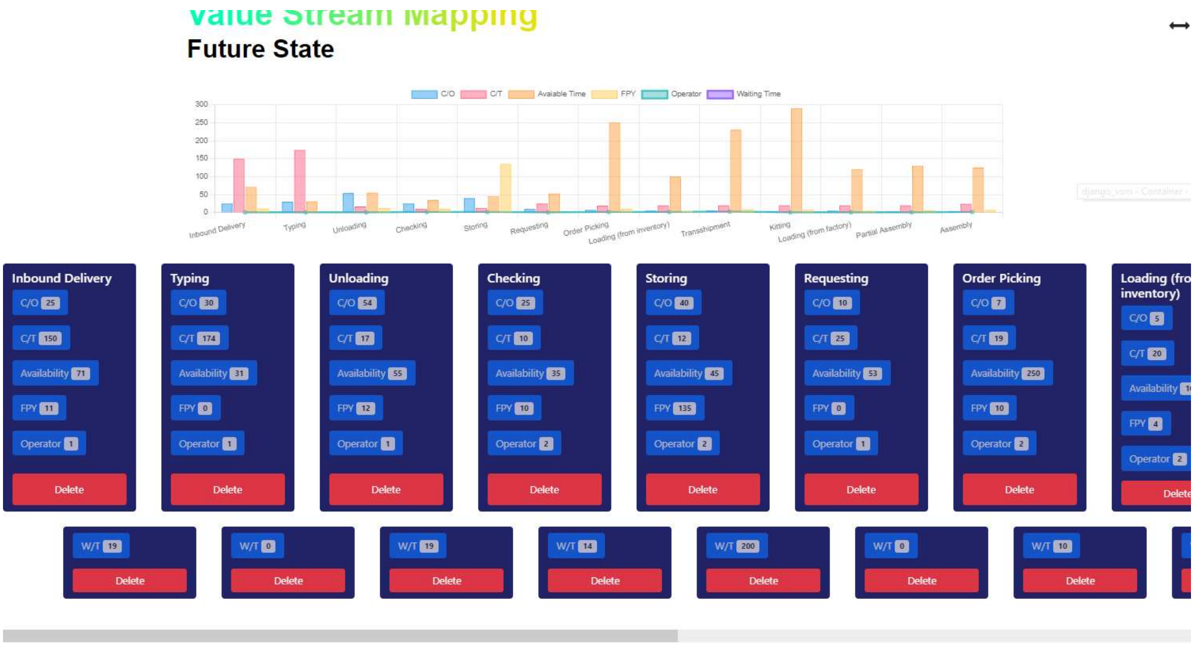
W/T 200

W/T 0

W/T 10

Source: The Author (2023)

Figure 91 - Future State Mapping extended visualization



Value Stream Mapping

Future State

Inbound Delivery
C/O 25
C/T 150
Availability 71
FPY 11
Operator 1
Delete

Typing
C/O 30
C/T 174
Availability 31
FPY 0
Operator 1
Delete

Unloading
C/O 54
C/T 17
Availability 55
FPY 12
Operator 1
Delete

Checking
C/O 25
C/T 10
Availability 35
FPY 10
Operator 2
Delete

Storing
C/O 40
C/T 12
Availability 45
FPY 135
Operator 2
Delete

Requesting
C/O 10
C/T 25
Availability 53
FPY 0
Operator 1
Delete

Order Picking
C/O 7
C/T 19
Availability 250
FPY 10
Operator 2
Delete

Loading (fro
inventory)
C/O 5
C/T 20
Availability 1
FPY 4
Operator 2
Delete

W/T 19
Delete

W/T 0
Delete

W/T 19
Delete

W/T 14
Delete

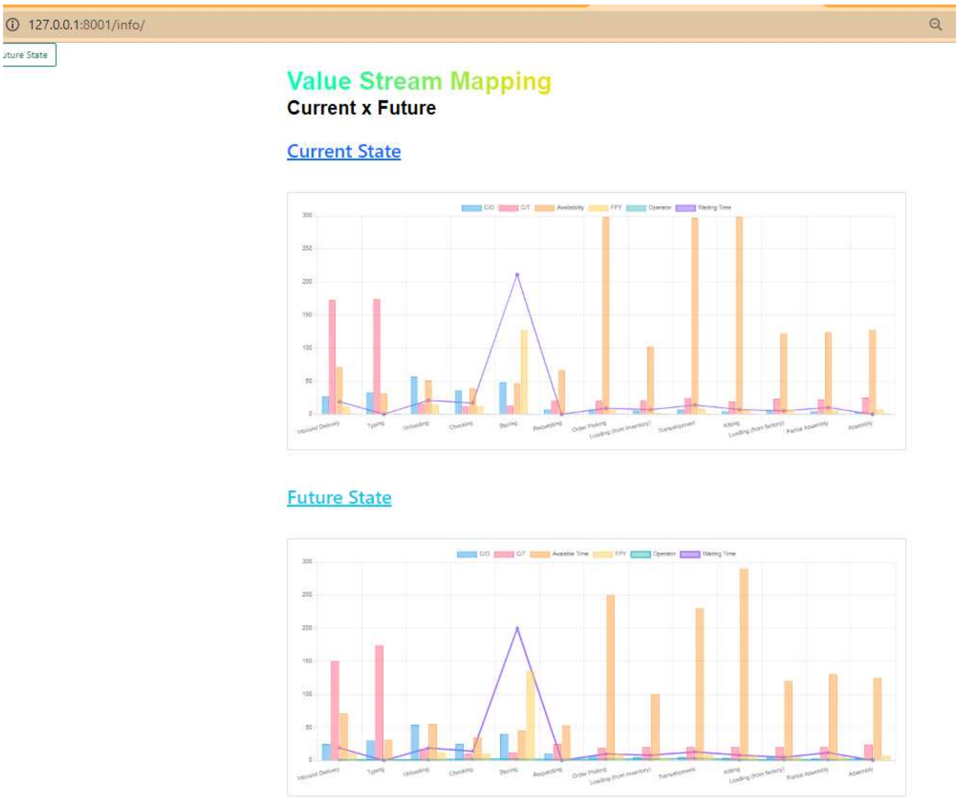
W/T 200
Delete

W/T 0
Delete

W/T 10
Delete

Source: The Author (2023)

Figure 92 Current Chart x Future Chart



127.0.0.1:8001/info/

uture State

Value Stream Mapping

Current x Future

Current State

Future State

Source: The Author (2023)

On this page, both charts have the same dynamics as other pages concerning the display filters of the columns and rows.

In the traditional VSM application, the mapping starts with the current mapping, which aims to obtain a complete and detailed understanding of how the process works, including cycle times, inventories, waiting time, unnecessary activities, and other waste. By identifying bottlenecks or opportunities during the current mapping, changes, and solutions are proposed to improve the value stream. Future mapping involves creating a new ideal state where waste is eliminated, processes are optimized, and flow is efficient.

Traditionally, this development is done through several meetings with multidisciplinary teams before practical analyzes that will confirm whether or not it can bring the expected results.

The development of the VSM application provides a different approach to the development of future mapping since it is carried out through prediction simulators using machine learning algorithms that learn from historical data from past mappings. These simulators help to get faster and more accurate answers based on trends and patterns learned from training data.

The simulators provide previous answers to practical analyses, saving time and costs since it will not be necessary to interrupt the production process to carry out the pre-established verifications in future mapping.

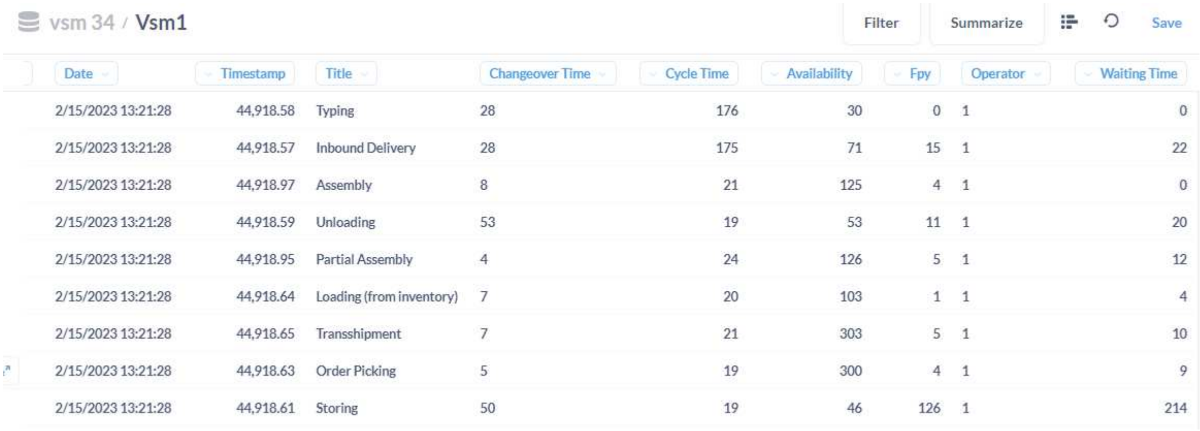
The interdisciplinary meetings will still take place, but the immediate responses to the predictions that will be made will enhance decision-making with the identification of the best alternatives.

The reliability and viability of the obtained predictions are directly related to the amount of data used in training and, consequently, the identification of the best machine-learning model through the previously explained metrics analyses.

In addition to the dynamic and intelligent features applied to the current and future state mappings through the web, with the graphical complement of each operation, this project still allows creating and using a complementary dashboard to the ones already demonstrated. This dashboard can be created through the Metabase BI system, which is also on the Docker Desktop and accessed through port 3002.

Metabase is connected to MySQL and therefore has access to both the vsm1 table (Figure 93) with all the data from the CSV file sent by Node-Red and the table created in Django for the information blocks, called App Infomapping Information (Figure 94).

Figure 93 – Vsm1 Table

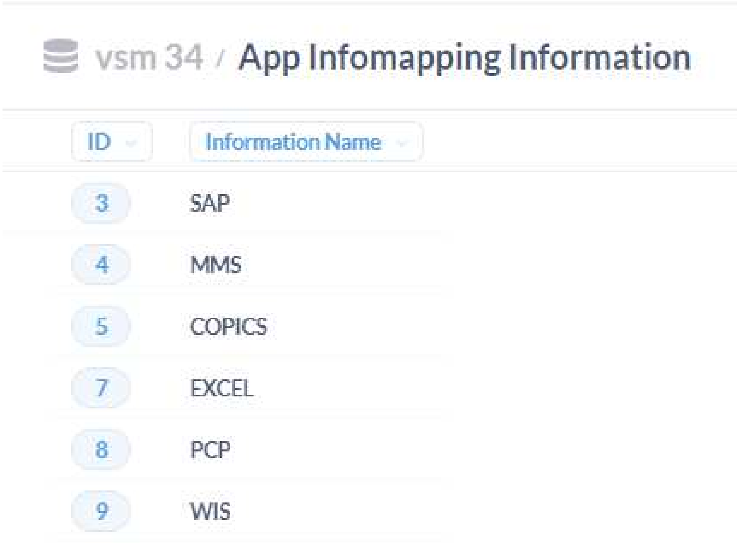


vsm 34 / Vsm1

Date	Timestamp	Title	Changeover Time	Cycle Time	Availability	Fpy	Operator	Waiting Time
2/15/2023 13:21:28	44,918.58	Typing	28	176	30	0	1	0
2/15/2023 13:21:28	44,918.57	Inbound Delivery	28	175	71	15	1	22
2/15/2023 13:21:28	44,918.97	Assembly	8	21	125	4	1	0
2/15/2023 13:21:28	44,918.59	Unloading	53	19	53	11	1	20
2/15/2023 13:21:28	44,918.95	Partial Assembly	4	24	126	5	1	12
2/15/2023 13:21:28	44,918.64	Loading (from inventory)	7	20	103	1	1	4
2/15/2023 13:21:28	44,918.65	Transshipment	7	21	303	5	1	10
2/15/2023 13:21:28	44,918.63	Order Picking	5	19	300	4	1	9
2/15/2023 13:21:28	44,918.61	Storing	50	19	46	126	1	214

Source: The Author (2023)

Figure 94 – App Infomapping Information Table



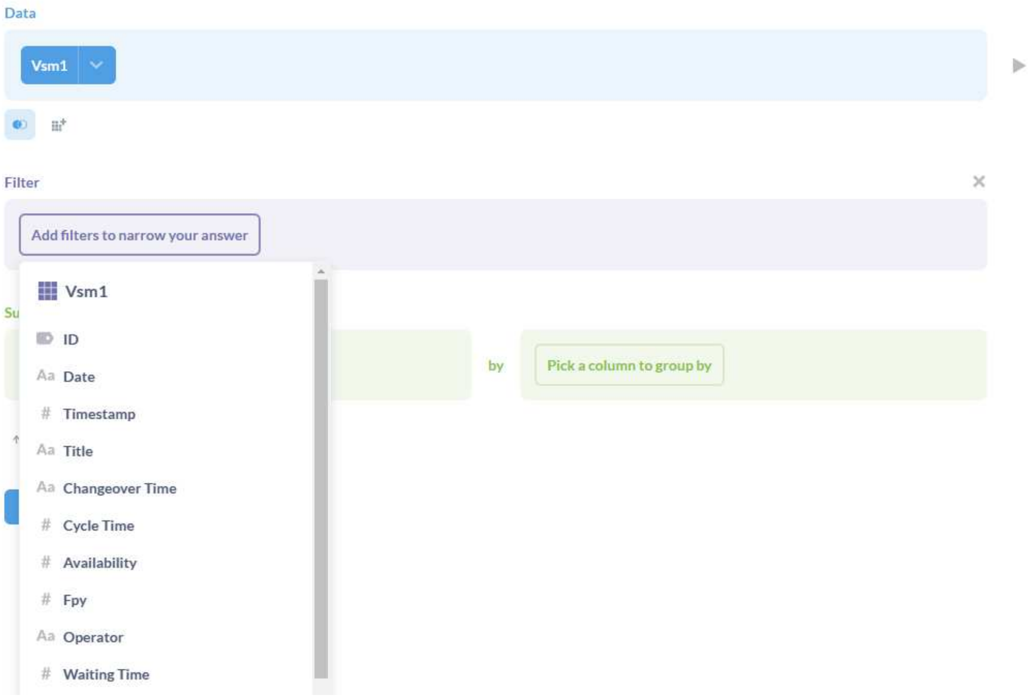
vsm 34 / App Infomapping Information

ID	Information Name
3	SAP
4	MMS
5	COPICS
7	EXCEL
8	PCP
9	WIS

Source: The Author (2023)

Users can create dashboards as necessary, with the number of charts and information they identify as relevant to their needs. The charts can be dynamically created using a question editor, which allows the user to select any of the two tables, filter this data, group and aggregate them into results, and apply them to various data transformations (Figure 95).

Figure 95 - Question Editor in Metabase



Source: The Author (2023)

All charts are dynamic and interactive, and by simply hovering over some areas of the charts, more specific information about the demonstrated data can be obtained.

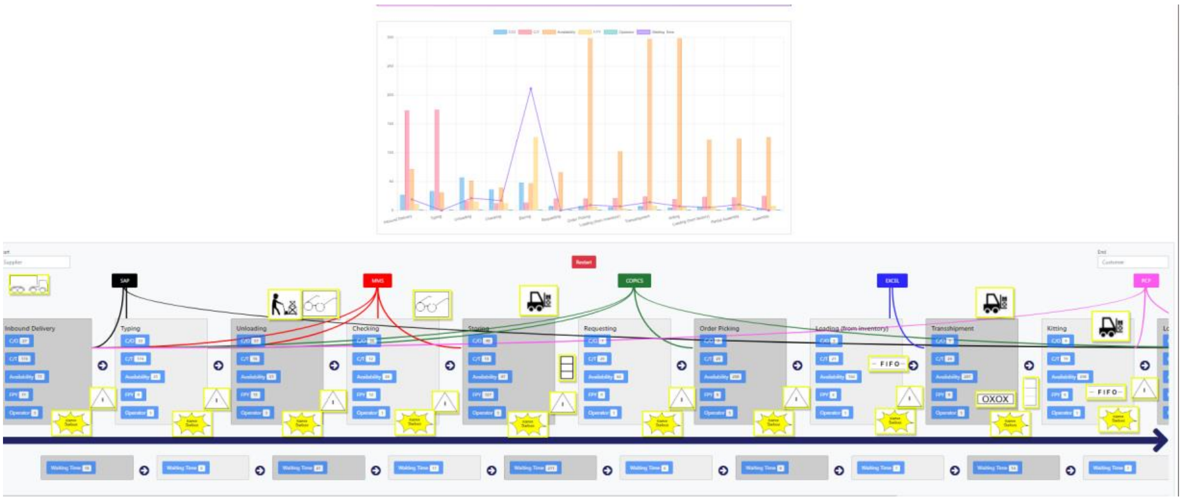
From the construction of this model, the third activity of DSR declared in Chapter 3 of the Methodology is concluded. The fourth activity will be developed in the next section, demonstrating the project developed in this thesis.

4.2 Demonstration

The fourth activity of the DSR methodology refers to the demonstration of the application, a real-life test through the direct usage of all functionalities by the engineers of the partner company.

There was a walking follow-up between operations to verify all activities and parameters used, and all data was reanalyzed through the CSV file to facilitate the comparison between what was created for the current mapping in the application and what is effectively practiced in the company. The system was used for practical application, in which it was effectively used to map the current situation with the full utilization of the available resources (Figure 96).

Figure 96 - Current state mapping demonstration



Source: The Author (2023)

After completing the current state mapping, a comprehensive analysis was performed to explore future state mapping development using the prediction algorithm. For this application demonstration, four prediction scenarios were analyzed. The first scenario was for predicting enhanced parameters for the Unloading operation from changes in the Inbound Delivery and Typing data. The second scenario was for predicting parameters for the Storing operation, from changes in the Inbound Delivery, Typing, Unloading, and Checking data. The third scenario was for the prediction of parameters for the Loading (from inventory) operation from changes in the Requesting and Order Picking data, and the fourth and final scenario was for the prediction of parameters for the Assembly operation from changes in the Unloading, Loading (from inventory), and Loading (from factory) data.

Since the data used in this project are not 100% actual, what was obtained was a projection of the trend that each variation in each parameter influenced to get other values. Still, there is no way to guarantee the accuracy of the obtained values, as 100% of the data used would need to be obtained through actual observations on the factory floor.

The positive or negative influence between two operations may occur, as both operations are interconnected and depend on each other to produce the final product. Suppose there is a reduction in the cycle time or changeover time of an operation. In that case, this can lead to a reduction in the total processing time and, therefore, an increase in machine availability. It can have a positive effect as it may reduce waiting time and improve overall performance. Similarly, a change in operator availability or the number of operators may affect the productivity and quality of the operations, but with the risk of negatively affecting the subsequent operations, as it may increase waiting time and reduce the quality of the products produced.

Therefore, it is vital to consider the interdependence of two or more operations when making changes in one of them and evaluate how these changes may affect the productivity and quality of the overall production.

In the first scenario, after several analyses of the first and second operations for the prediction of the third operation (Unloading), the unification of the Inbound Delivery and Typing operations was identified as an improvement alternative, but with the use of 2 operators. For this analysis, the Inbound Delivery data was set to zero, while the Typing data was changed proportionally as if two operators performed the two activities. The prediction brought an increase in Availability to Unloading and a reduction in cycle time with a single operator.

The values of the operations before and after the executed prediction can be visualized in Figure 97.

The comparison between the Unloading operation of the current mapping and the predicted Unloading operation for the future mapping (in yellow) is shown in Figure 97. In contrast, the individual comparison between the values of each operation, current and predicted, with the indication of increase or reduction, is shown in Figure 98.

In the second scenario, the proposal was to reduce 20% of the cycle time values for the four operations used as parameters for predicting the Storing operation. As this cycle time value will be directly reduced, all other values were obtained by the parameter prediction system.

The Cycle Time for Inbound Delivery was reduced from 174 to 139, the Cycle Time for Typing was reduced from 173 to 138, the Cycle Time for Unloading was reduced from 18 to 14, and the Cycle Time for Checking was reduced from 10 to 8.

In the simulations performed for the Inbound Delivery operation (Figure 99), the values obtained by the parameter simulator were: 25 for Changeover Time, 73 for Availability, 12 for FPY, and 17 for Waiting Time.

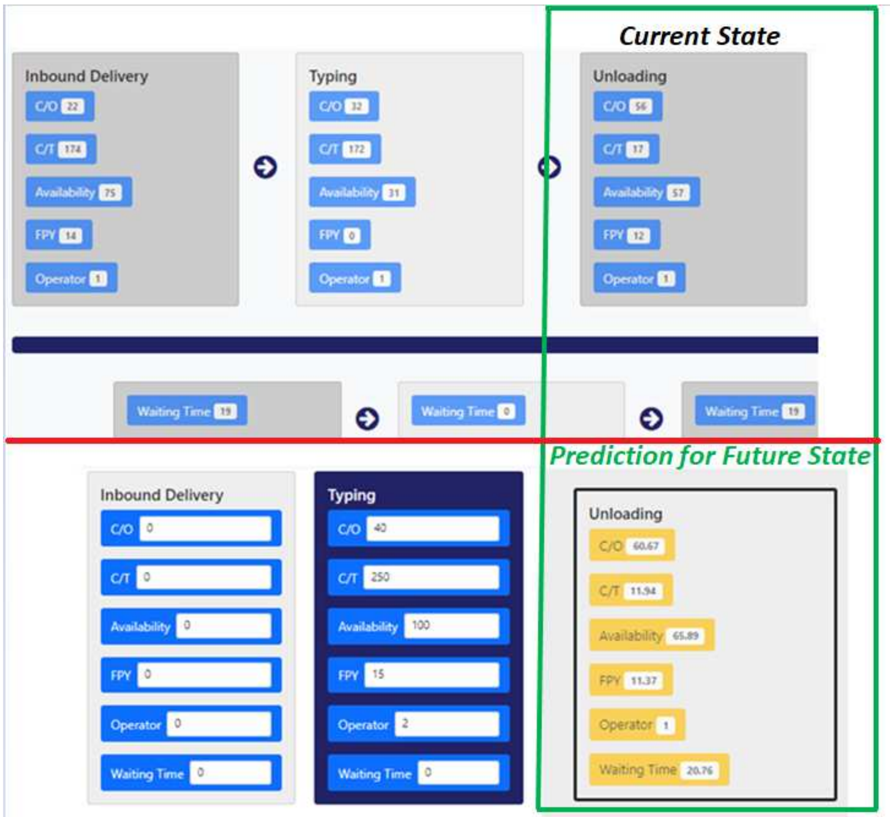
In the simulations performed for the Typing operation (Figure 100), the values obtained by the parameter simulator were: 30 for Changeover Time, 30 for Availability, 0 for FPY, and 0 for Waiting Time.

In the simulations performed for the Unloading operation (Figure 101), the values obtained by the parameter simulator were: 54 for Changeover Time, 54 for Availability, 13 for FPY, and 17 for Waiting Time.

In the simulations performed for the Checking operation (Figure 102), the values obtained by the parameter simulator were: 36 for Changeover Time, 36 for Availability, 13 for FPY, and 17 for Waiting Time.

In all operations, the number of operators was kept at 1.

Figure 97 - First scenario for the Future State Mapping Development



Current State

Inbound Delivery
C/O 22
C/T 174
Availability 75
FPY 14
Operator 1

Typing
C/O 32
C/T 172
Availability 31
FPY 0
Operator 1

Unloading
C/O 56
C/T 17
Availability 57
FPY 12
Operator 1

Waiting Time 19

Waiting Time 0

Waiting Time 19

Prediction for Future State

Inbound Delivery
C/O 0
C/T 0
Availability 0
FPY 0
Operator 0
Waiting Time 0

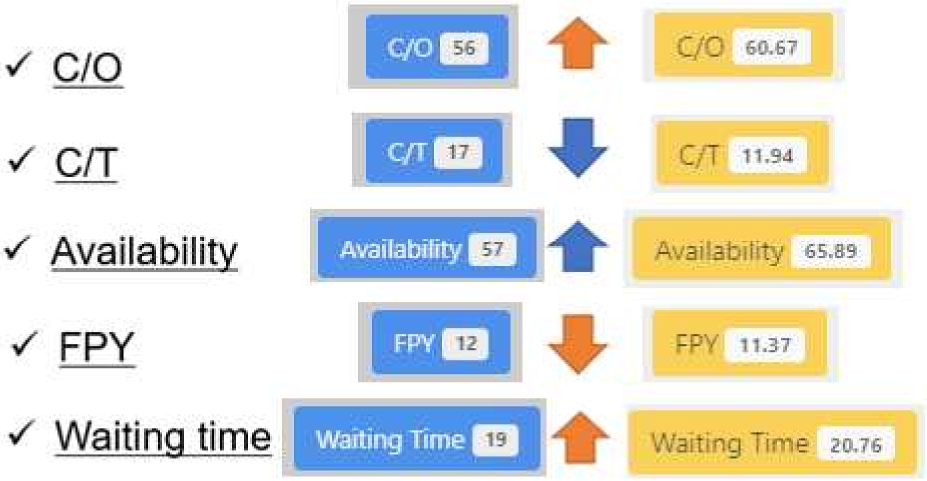
Typing
C/O 40
C/T 250
Availability 100
FPY 15
Operator 2
Waiting Time 0

Unloading
C/O 60.67
C/T 11.94
Availability 65.89
FPY 11.37
Operator 1
Waiting Time 20.76

Source: The Author (2023)

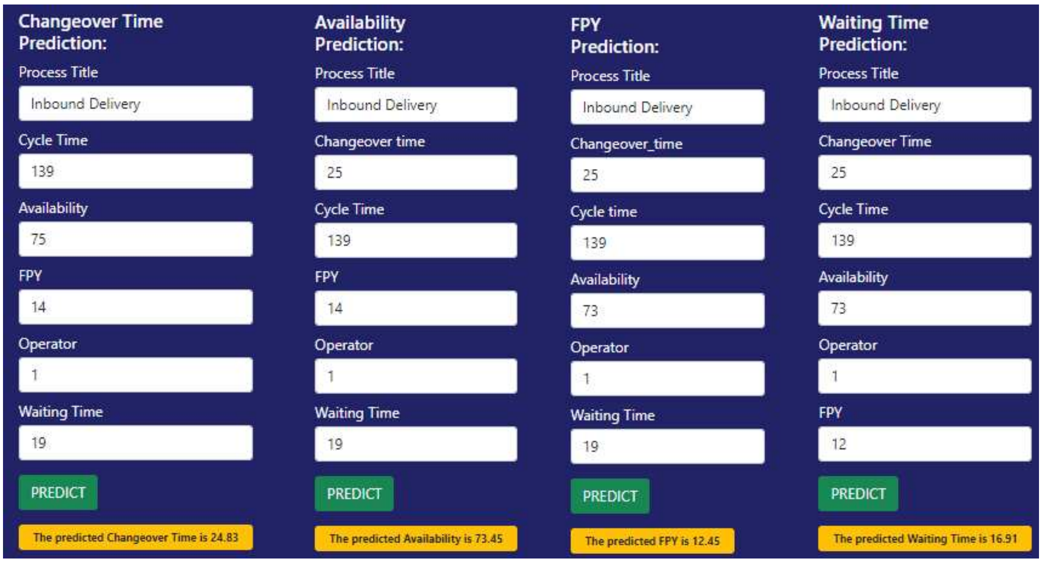
As previously explained in the adaptation of variable data for two operators, the 20% reduction for this demonstration scenario was also intentionally performed to make the simulations more perceptible in the comparative evaluation between the data from the current mapping and the data predicted for future mapping. If the values used were minimally altered, the responses generated by the machine learning algorithm could become very close to the original value, failing to provide evidence that the algorithms can consistently deliver results compatible with the data used during algorithm learning. Thus, based on the patterns and trends of the data obtained, the algorithm will always predict accurately within the established parameters.

Figure 98 - Current Unloading variables x Future Unloading variables



Source: The Author (2023)

Figure 99 - Individual parameter predictions for the Inbound Delivery operation



Changeover Time Prediction:

Process Title: Inbound Delivery
Cycle Time: 139
Availability: 75
FPY: 14
Operator: 1
Waiting Time: 19

PREDICT

The predicted Changeover Time is 24.83

Availability Prediction:

Process Title: Inbound Delivery
Changeover time: 25
Cycle Time: 139
FPY: 14
Operator: 1
Waiting Time: 19

PREDICT

The predicted Availability is 73.45

FPY Prediction:

Process Title: Inbound Delivery
Changeover_time: 25
Cycle time: 139
Availability: 73
Operator: 1
Waiting Time: 19

PREDICT

The predicted FPY is 12.45

Waiting Time Prediction:

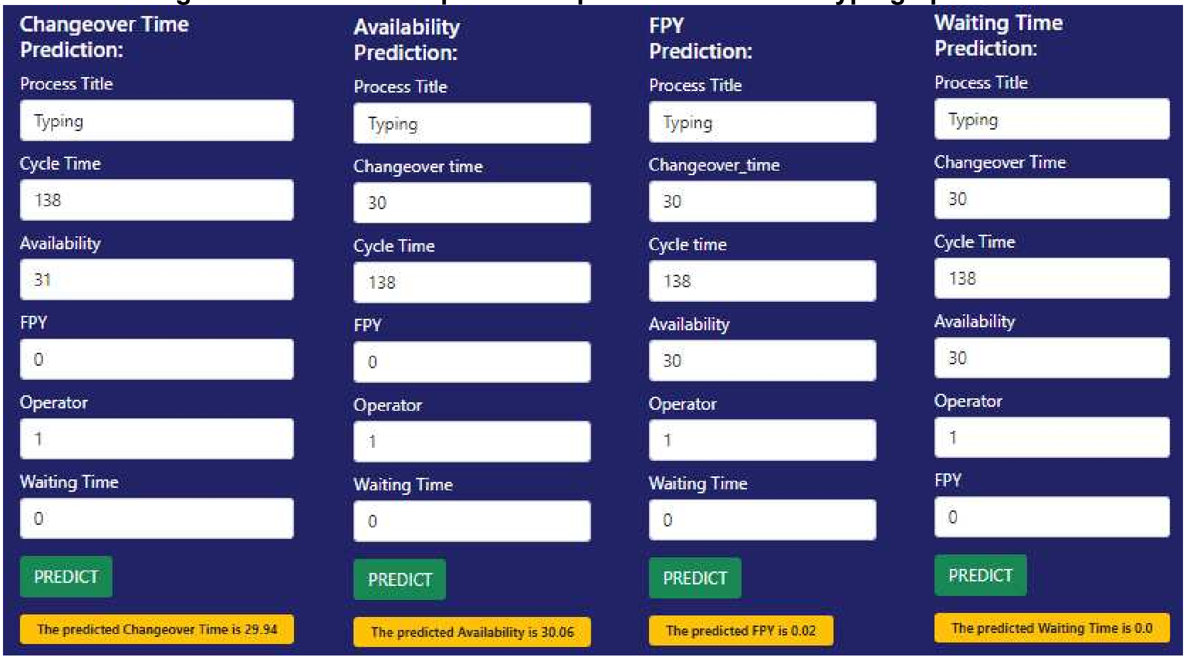
Process Title: Inbound Delivery
Changeover Time: 25
Cycle Time: 139
Availability: 73
Operator: 1
FPY: 12

PREDICT

The predicted Waiting Time is 16.91

Source: The Author (2023)

Figure 100 - Individual parameter predictions for the Typing operation



Changeover Time Prediction:

Process Title: Typing

Cycle Time: 138

Availability: 31

FPY: 0

Operator: 1

Waiting Time: 0

The predicted Changeover Time is 29.94

Availability Prediction:

Process Title: Typing

Changeover time: 30

Cycle Time: 138

FPY: 0

Operator: 1

Waiting Time: 0

The predicted Availability is 30.06

FPY Prediction:

Process Title: Typing

Changeover_time: 30

Cycle time: 138

Availability: 30

Operator: 1

Waiting Time: 0

The predicted FPY is 0.02

Waiting Time Prediction:

Process Title: Typing

Changeover Time: 30

Cycle Time: 138

Availability: 30

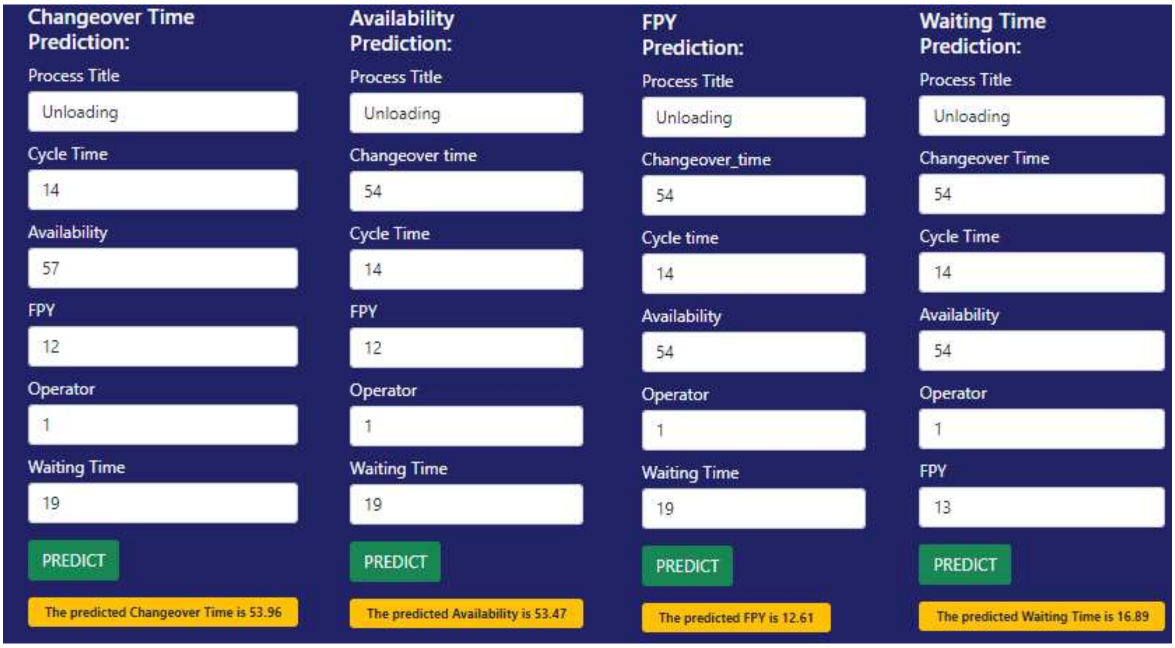
Operator: 1

FPY: 0

The predicted Waiting Time is 0.0

Source: The Author (2023)

Figure 101 - Individual parameter predictions for the Unloading operation



Changeover Time Prediction:

Process Title: Unloading

Cycle Time: 14

Availability: 57

FPY: 12

Operator: 1

Waiting Time: 19

The predicted Changeover Time is 53.96

Availability Prediction:

Process Title: Unloading

Changeover time: 54

Cycle Time: 14

FPY: 12

Operator: 1

Waiting Time: 19

The predicted Availability is 53.47

FPY Prediction:

Process Title: Unloading

Changeover_time: 54

Cycle time: 14

Availability: 54

Operator: 1

Waiting Time: 19

The predicted FPY is 12.61

Waiting Time Prediction:

Process Title: Unloading

Changeover Time: 54

Cycle Time: 14

Availability: 54

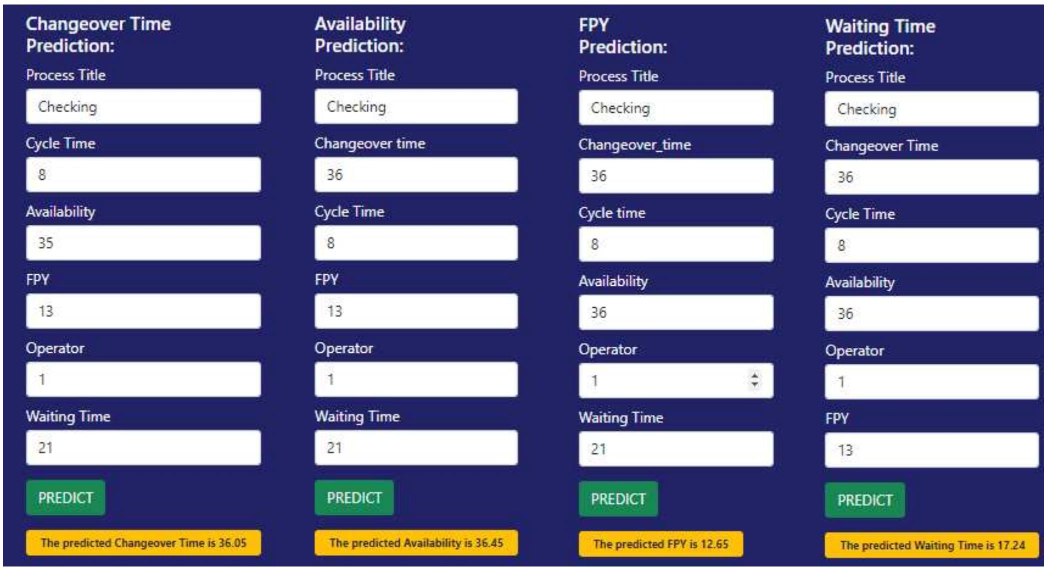
Operator: 1

FPY: 13

The predicted Waiting Time is 16.89

Source: The Author (2023)

Figure 102 - Individual parameter predictions for the Checking operation



Changeover Time Prediction:

Process Title: Checking

Cycle Time: 8

Availability: 35

FPY: 13

Operator: 1

Waiting Time: 21

PREDICT

The predicted Changeover Time is 36.05

Availability Prediction:

Process Title: Checking

Changeover time: 36

Cycle Time: 8

FPY: 13

Operator: 1

Waiting Time: 21

PREDICT

The predicted Availability is 36.45

FPY Prediction:

Process Title: Checking

Changeover_time: 36

Cycle time: 8

Availability: 36

Operator: 1

Waiting Time: 21

PREDICT

The predicted FPY is 12.65

Waiting Time Prediction:

Process Title: Checking

Changeover Time: 36

Cycle Time: 8

Availability: 36

Operator: 1

FPY: 13

PREDICT

The predicted Waiting Time is 17.24

Source: The Author (2023)

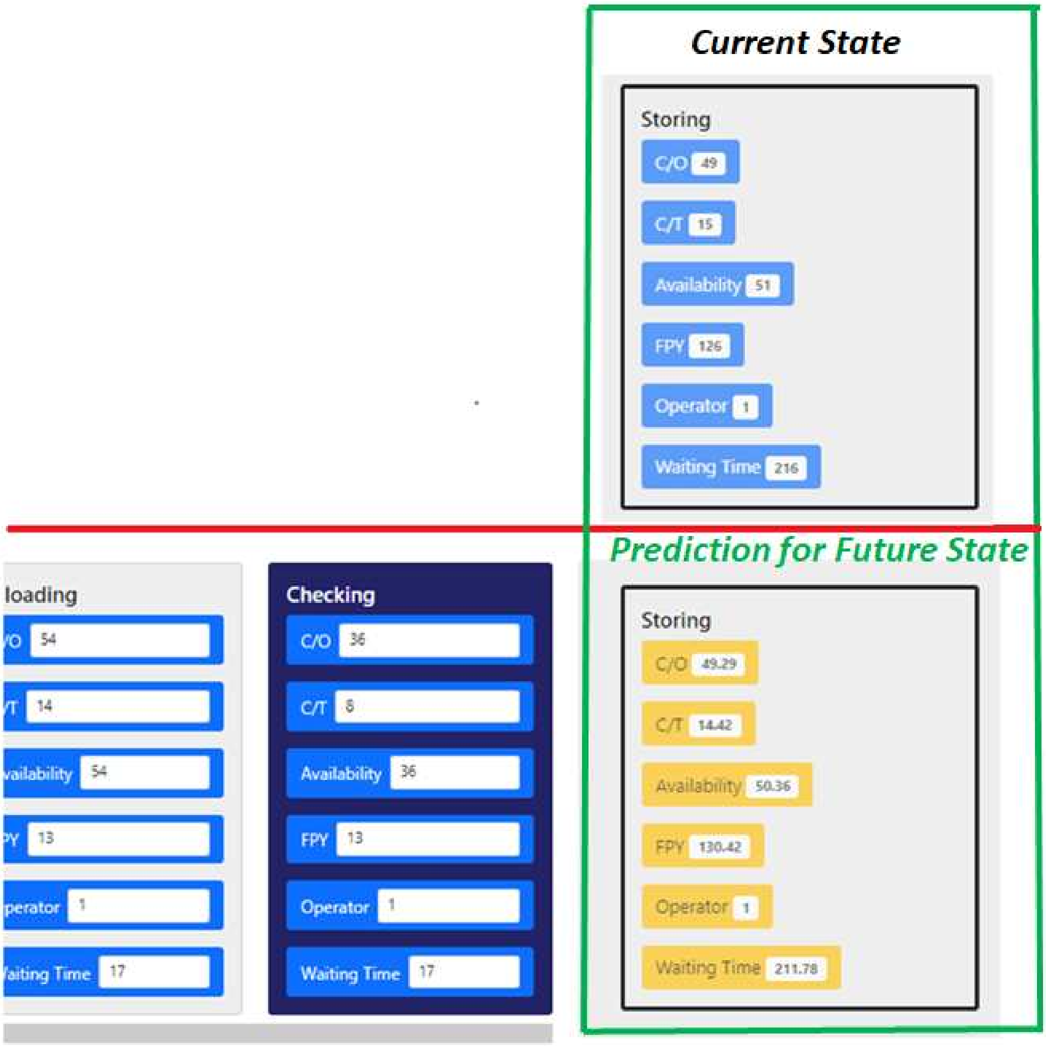
After the individual parameter prediction for each operation, all these parameters were filled into the operation prediction system to verify the prediction for the Storing operation.

The comparison between the Storing operation of the current mapping and the predicted Storing operation for the future mapping (in yellow) is shown in Figure 103. In contrast, the individual comparison between the most relevant values of each operation, current and predicted, with the indication of increase or reduction, is shown in Figure 104.

Thus, with the 20% reduction in all cycle times and predictions for the other parameters based on this change, the final result showed that if all these parameters are changed as predetermined, there is a high probability of increasing the FPY and reducing the waiting time.

In the third scenario, all parameters of the current state mapping were kept for the Requesting and Order Picking operations. Still, the number of operators was changed to two to verify the prediction made by the system for the Loading (from inventory) operation. The comparison between the Loading (from inventory) operation of the current mapping and the predicted Loading (from inventory) operation for the future mapping (in yellow) is shown in Figure 105. In contrast, the individual comparison between the values of each operation, current and predicted, with the indication of increase or reduction, is shown in Figure 106.

Figure 103 - Second scenario for the Future State Mapping Development



Current State

Storing

C/O 49

C/T 15

Availability 51

FPY 126

Operator 1

Waiting Time 216

Prediction for Future State

Storing

C/O 49.29

C/T 14.42

Availability 50.36

FPY 130.42

Operator 1

Waiting Time 211.78

loading

/O 54

/T 14

vailability 54

PY 13

perator 1

/aiting Time 17

Checking

C/O 36

C/T 8

Availability 36

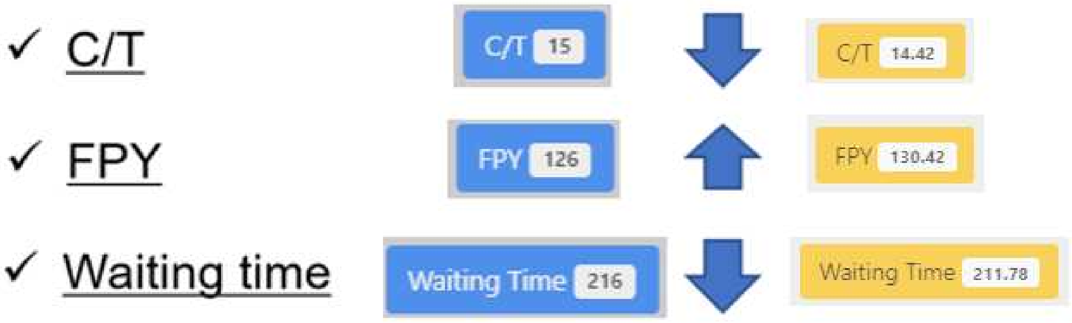
FPY 13

Operator 1

Waiting Time 17

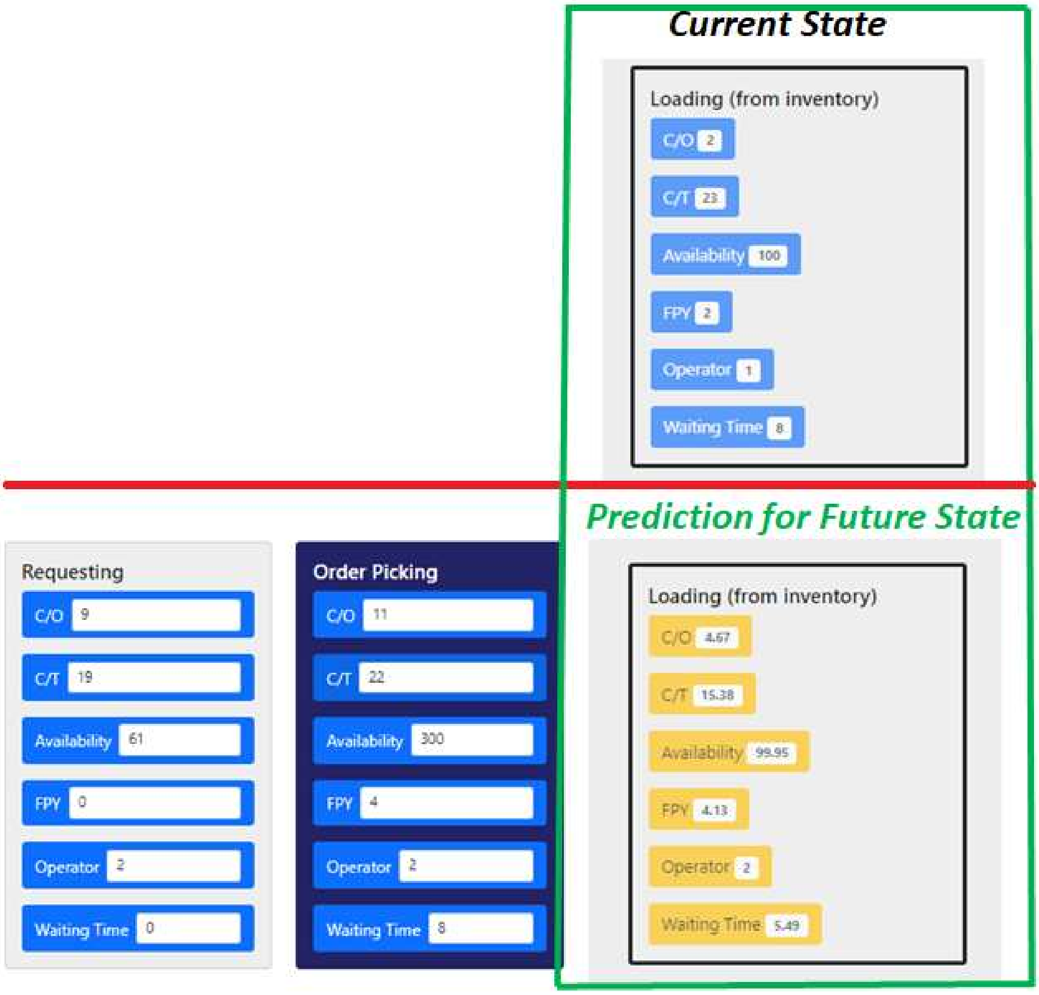
Source: The Author (2023)

Figure 104 - Current Storing variables x Future Storing variables



Source: The Author (2023)

Figure 105 - Third scenario for the Future State Mapping Development



Current State

Loading (from inventory)

C/O 2

C/T 23

Availability 100

FPY 2

Operator 1

Waiting Time 8

Prediction for Future State

Requesting

C/O 9

C/T 19

Availability 61

FPY 0

Operator 2

Waiting Time 0

Order Picking

C/O 11

C/T 22

Availability 300

FPY 4

Operator 2

Waiting Time 8

Loading (from inventory)

C/O 4.67

C/T 15.38

Availability 99.95

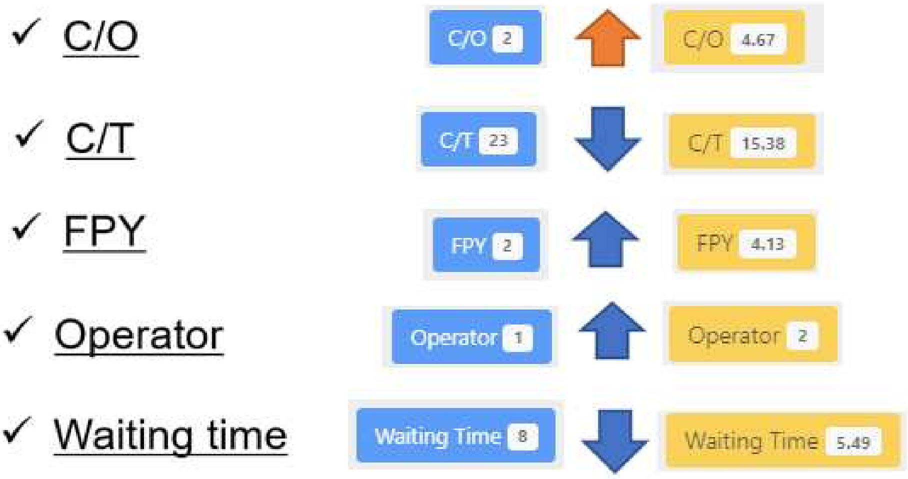
FPY 4.13

Operator 2

Waiting Time 5.49

Source: The Author (2023)

Figure 106 - Current Loading (from inventory) variables x Future Loading (from inventory) variables



Source: The Author (2023)

Although the idea for this analysis would be to adapt the parameters of Requesting and Order Picking to the reality of development with two operators, this scenario helped identify that the ML algorithm identified that the change of operators in the two previous operations induces the tendency to use two operators for the target operation. Thus, the algorithm has already adapted all the parameters in the target operation for this situation.

This scenario with two operators was not implemented in the company. However, it was used as a test to allow engineers to analyze the influence of data variability on the machine learning algorithm's performance. The algorithm exhibits a more robust learning capability as the data variability increases in the database. The greater the diversity and range of data the algorithm provides, the better results.

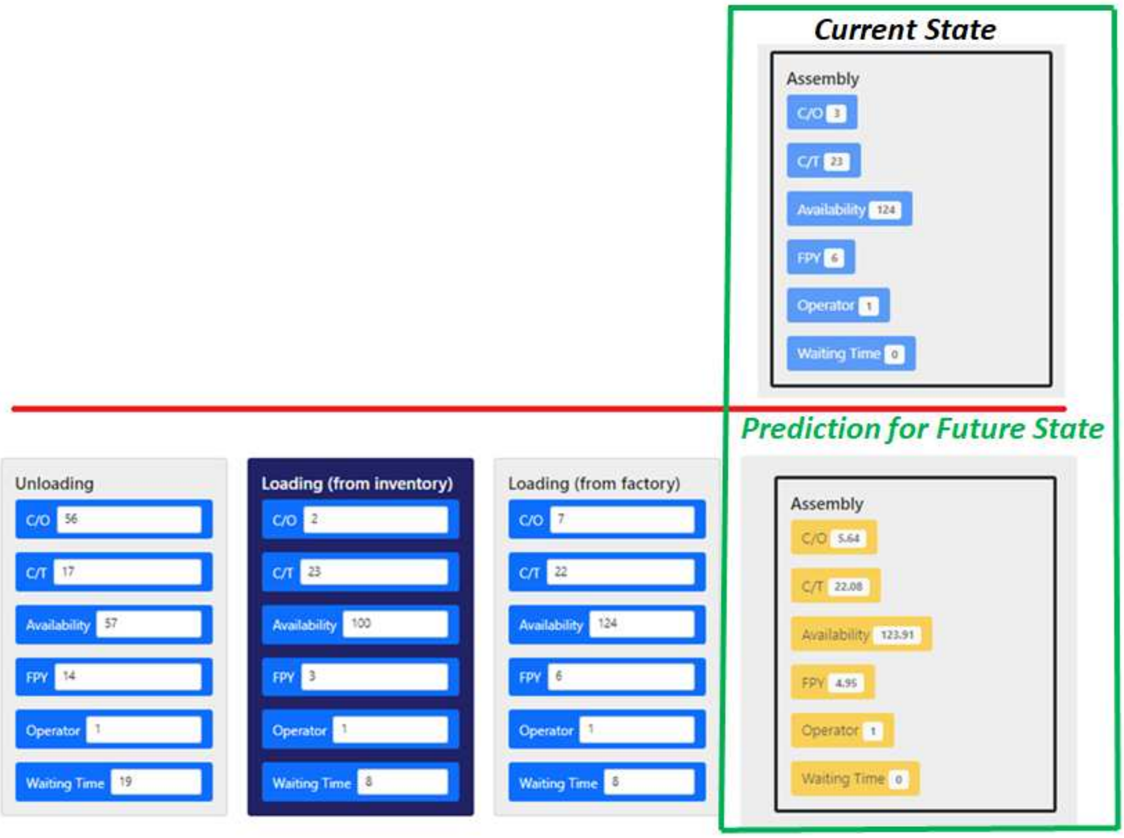
This scenario highlights that including a wide variety of data enables the algorithm to learn to generalize and make more accurate predictions in different situations. Furthermore, by avoiding overfitting and inconsistencies with changes in the process, the machine learning model becomes more reliable and effective.

Therefore, the test with two operators served as a practical demonstration of the principle that data variability plays a fundamental role in the success of machine learning, allowing the algorithm to learn more robust patterns and make more accurate predictions.

In the fourth scenario, three operations that are not in direct sequence were used as verification parameters, i.e., Unloading, Loading (from inventory), and Loading (from factory), precisely with the same parameters as in the current state mapping, except for FPY, which received an increase in its value to 15%. The target operation in this scenario was Assembly. The comparison between the Assembly operation of the current mapping and the predicted Assembly operation for the future mapping (in yellow) is shown in Figure 107. In contrast, the individual comparison between the most relevant values of each operation, current and predicted, with the indication of increase or reduction, is shown in Figure 108.

The result of the prediction obtained confirms that the parameters used for the prediction do not necessarily need to be positioned sequentially, as the ML algorithm is prepared to make these correlations, identifying both the order of operations and the trend of their parameters, even if the selected data is from separate operations. In this analysis, the third, eighth, and eleventh operations were included to predict the parameters of the thirteenth operation (Assembly).

Figure 107 - Fourth scenario for the Future State Mapping Development


Current State

Assembly
C/O 3
C/T 23
Availability 124
FPY 6
Operator 1
Waiting Time 0

Prediction for Future State

Unloading
C/O 56
C/T 17
Availability 57
FPY 14
Operator 1
Waiting Time 19

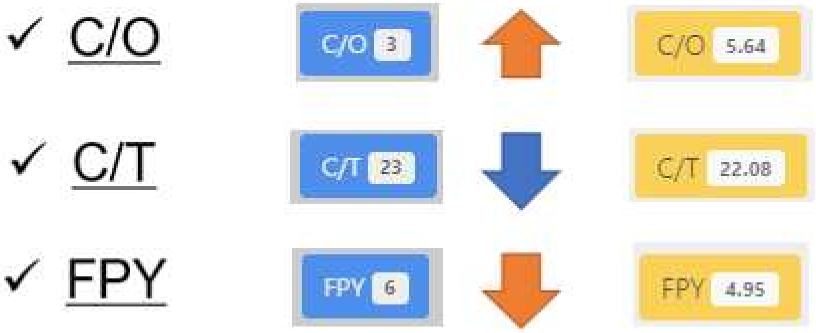
Loading (from inventory)
C/O 2
C/T 23
Availability 100
FPY 3
Operator 1
Waiting Time 8

Loading (from factory)
C/O 7
C/T 22
Availability 124
FPY 6
Operator 1
Waiting Time 8

Assembly
C/O 5.64
C/T 22.08
Availability 123.91
FPY 4.95
Operator 1
Waiting Time 0

Source: The Author (2023)

Figure 108 - Current Assembly variables x Future Assembly variables



Source: The Author (2023)

As a result of the prediction, it is identified that the increase in FPY in the three operations analyzed resulted in an increase in Changeover Time and a reduction in FPY for the Assembly operation.

In the first, third, and fourth scenarios, there was an increase in the C/O with the alteration of other variables in previous operations, and this result was expected since the dataset used included C/O data with variations in a random order within a normal distribution.

Although it is not usual for C/O to be influenced by changes in previous FPY values or even by increasing the number of operators, the prediction results are developed based on the quality of the dataset used to train the machine learning algorithm.

The functions developed for training and prediction allow the individual analysis of each variable in each operation, but always obeying the correct sequence of operations through the timestamp column.

If the C/T value is reduced, it does not necessarily mean the other values will also decrease. This change will depend on the overall variation that the algorithm has learned from the dataset used. If there is a global trend in the dataset, where reducing the C/T also reduces the C/O, that could be the result. However, the dataset has other possibilities where reducing the C/T does not reduce the C/O. In that case, the prediction result will be different and depend on the variability of the other data used.

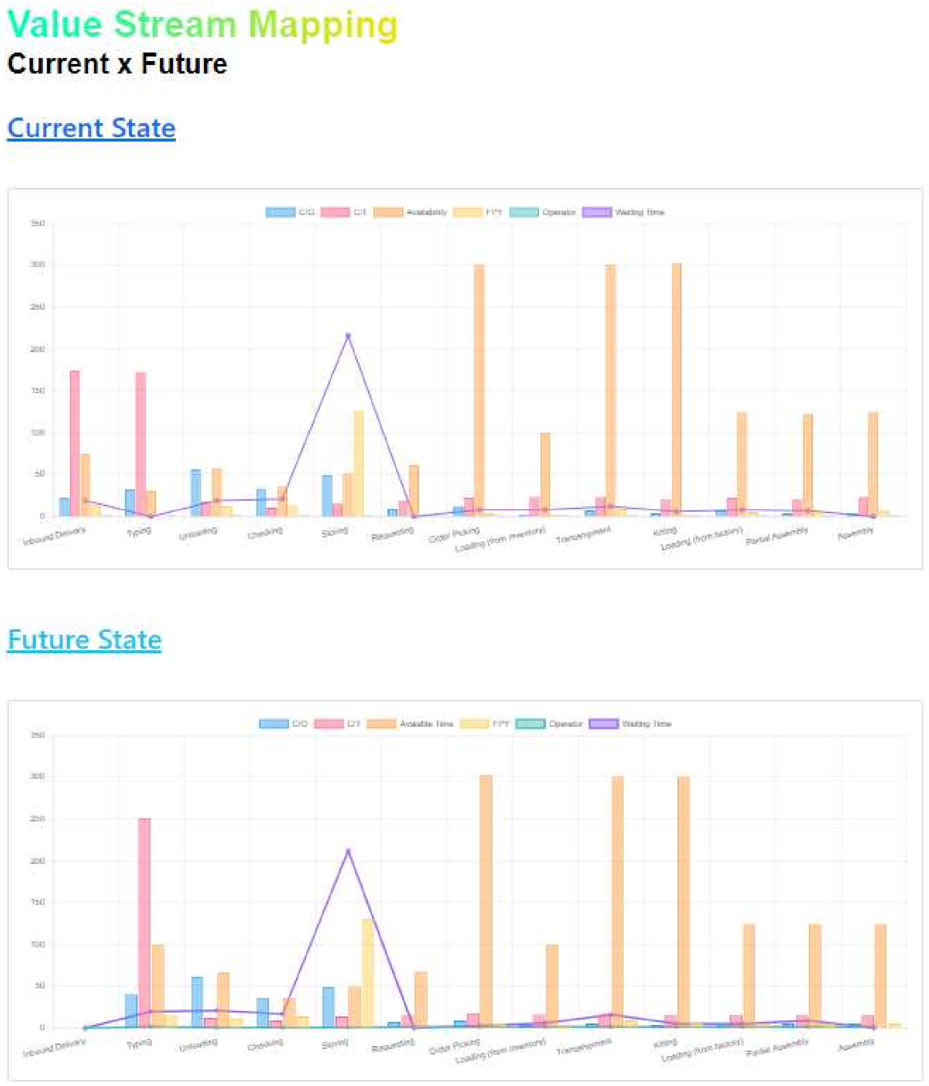
Both through the metrics used to evaluate the model's accuracy and through various analyses of the obtained results, the simulations were considered adequate and consistent with the dataset used for training. Therefore, the variations in the data used as parameters for the prediction simulators are consistent with the variations obtained as prediction results.

As analyzed by the metrics of the models used for these predictions, there is a margin of error in the obtained values, especially in the multiple regression prediction model, due to the complexity of the model, but this can be enhanced with appropriate data variability, accuracy, and precision, which can only be obtained with the application of actual data to this model. Despite the margin of error, this model presented values adequate to the trend of the data used in the database without divergences that exceed the reality of the process.

In addition to analyzing the created scenarios, a simulation of a complete future mapping was performed to compare the two charts (current and future) based on some settings created through the previously developed demonstration scenarios.

The first five operations were based on the first and second scenarios. In contrast, the fifth and sixth operations were taken from the third scenario but with variables adapted to the reality of two operators. From the sixth operation onwards, all parameters were predicted individually using the operational forecasting system. The resulting chart from the future mapping, together with the chart of the current mapping, can be seen in Figure 109.

Figure 109 – Current and Future Mapping Demonstration



Value Stream Mapping

Current x Future

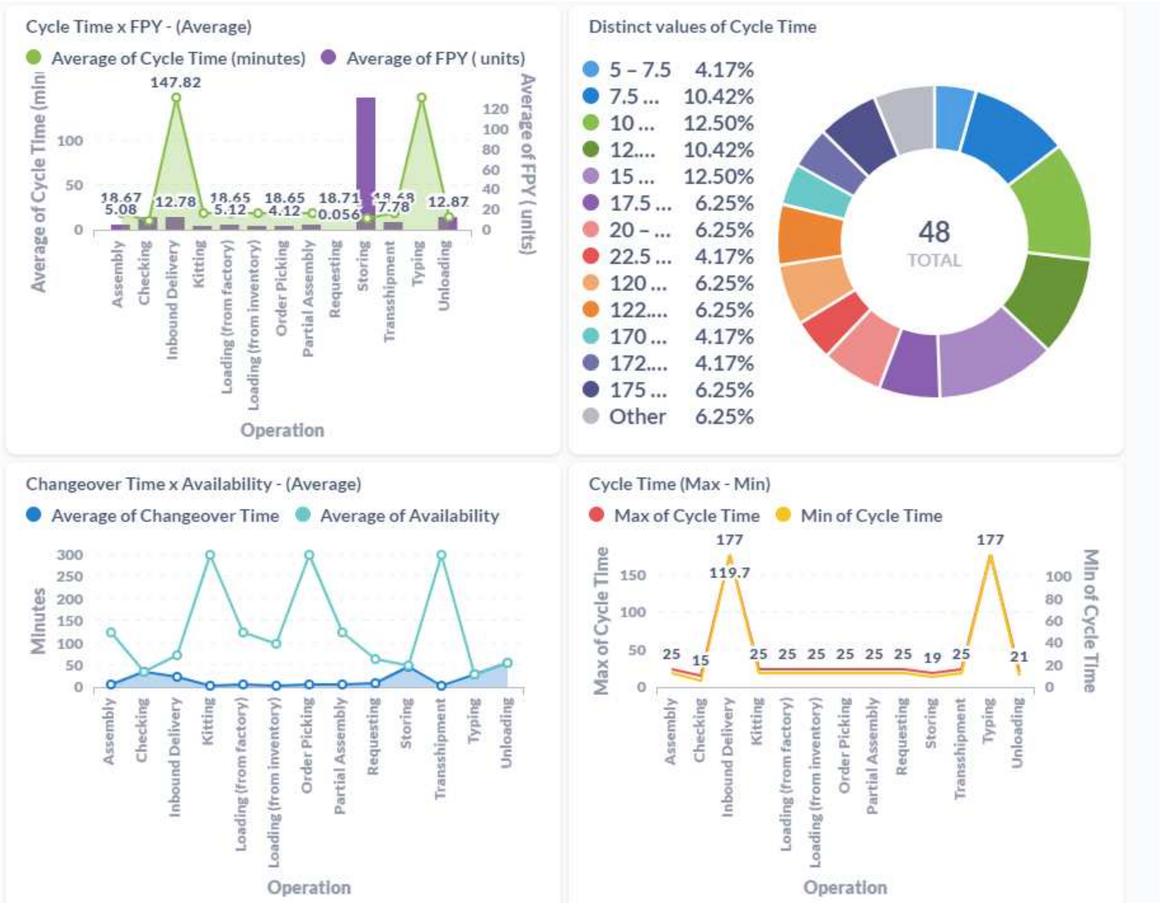
Current State

Future State

Source: The Author (2023)

After the complete mapping, when accessing port 3002 through the local URL localhost:3002, it was possible to create a dashboard with four charts to demonstrate its features, highlighting some alternatives of how the BI system integrated with Docker can bring even more insights to organizations along with all the features of the already demonstrated application. The created dashboard is shown in Figure 110. In this dashboard, the following charts were created: (i) a comparison between the averages of all cycle time values and the averages of all FPY values, both separated by titles and using the Combo visualization; (ii) a count of repeated or similar cycle time values using the Pie visualization; (iii) a comparison between the averages of all changeover time values and the averages of all Availability values, both separated by titles and using the Area visualization; and (iv) a comparison between the maximum cycle time values for each title and the minimum cycle time values for each title, using the Line visualization.

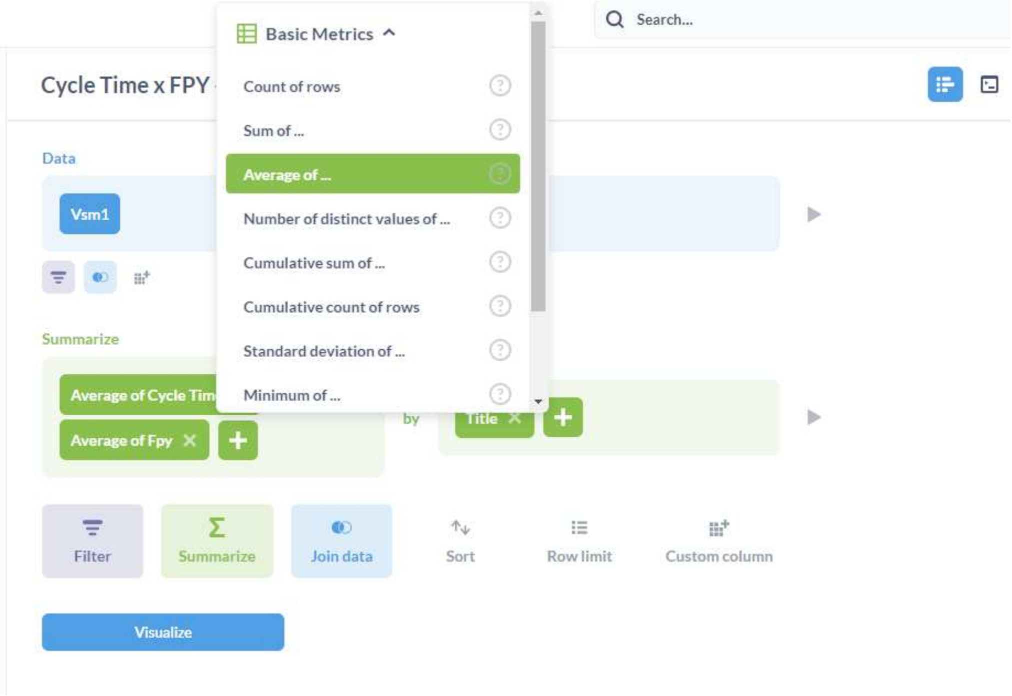
Figure 110 - Metabase Charts Demonstration



Source: The Author (2023)

When creating the first Chart, it was necessary to start the Question Editor, select the Vsm1 table, create the question with the average of both all Cycle Time values and create the question with the average of all FPY values, use the resource Summarize, group both averages by titles, and finally click on visualize (Figure 111).

Figure 111 – Questions for the development of the first chart



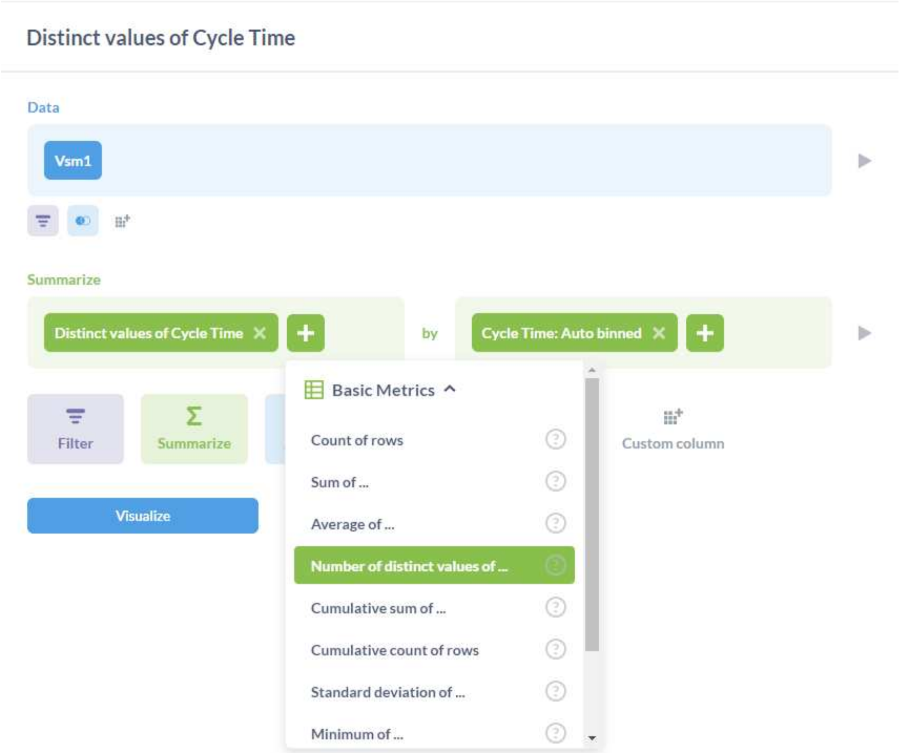
Source: The Author (2023)

When creating the second Chart, it was necessary to start the Question Editor, select the Vsm1 table, create the question with different values for the Cycle time, use the Summarize resource, select the auto binned option, and finally click on visualize (Figure 112).

When creating the third Chart, it was necessary to start the Question Editor, select the Vsm1 table, create the question with the average of all Changeover Time values, create the question with the average of all Availability values, group both averages by titles, and finally click on visualize (Figure 113).

When creating the fourth Chart, it was necessary to start the Question Editor, select the Vsm1 table, create the question with all minimum Cycle Time values, create the question with all maximum Cycle Time values, group both averages by titles, and finally click on visualize (Figure 114).

Figure 112 - Questions for the development of the second chart



Distinct values of Cycle Time

Data

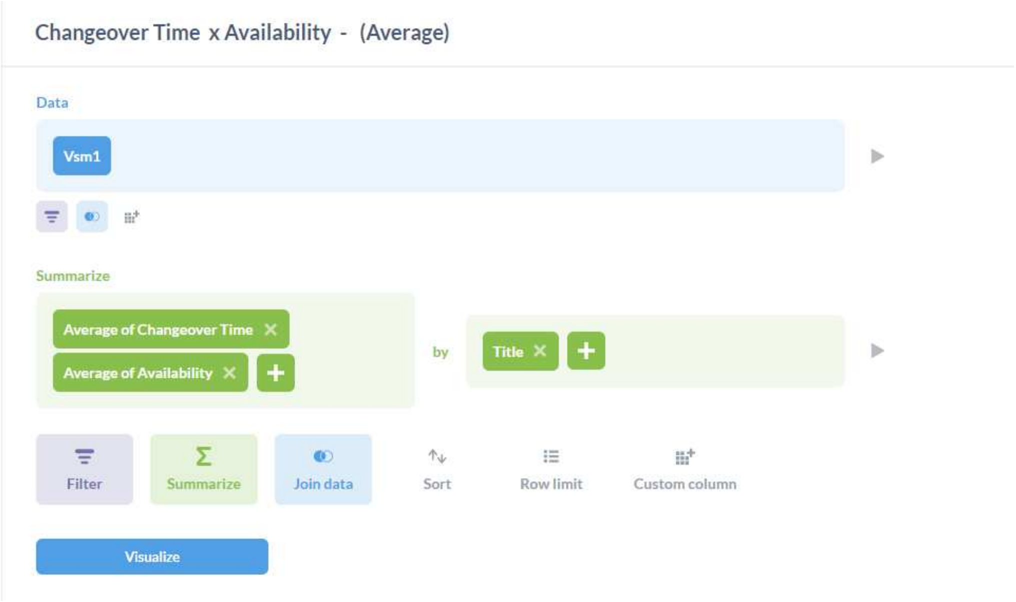
Vsm1

Summarize

Distinct values of Cycle Time by Cycle Time: Auto binned

Source: The Author (2023)

Figure 113 - Questions for the development of the third chart



Changeover Time x Availability - (Average)

Data

Vsm1

Summarize

Average of Changeover Time

Average of Availability

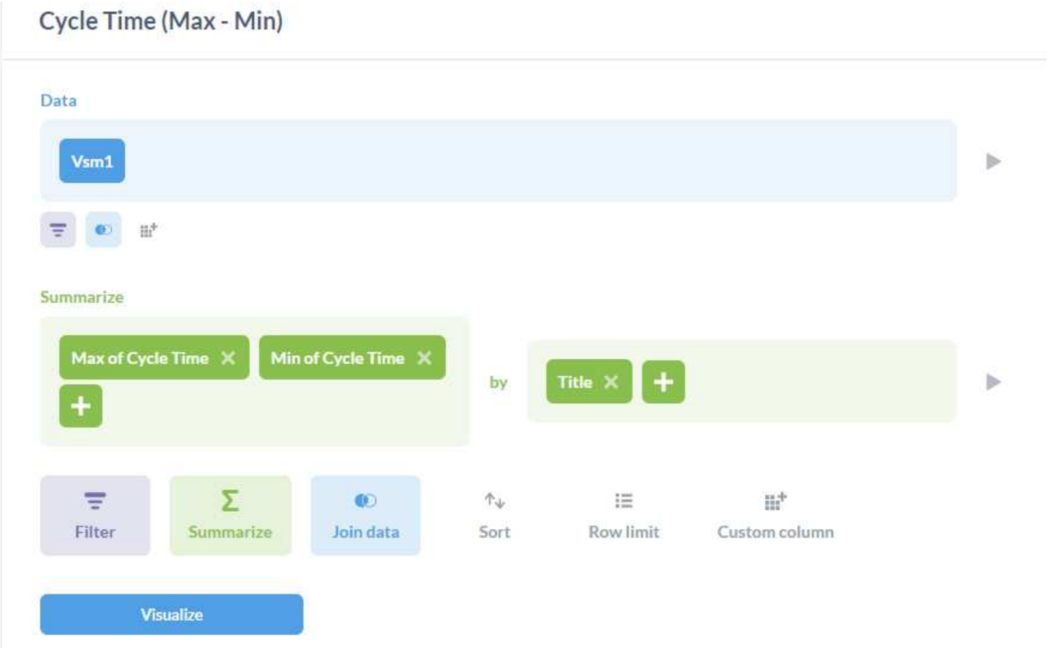
by Title

Source: The Author (2023)

After clicking on “Visualize”, the user will define which type of visualization is most suitable to obtain the best insights and settings for colors, legends, and axes. Then, both charts were made available on a single dashboard, as shown in Figure 110.

With the complete demonstration of the artifact, the fourth activity of the DSR methodology is concluded, as planned in section 3.2.4. Next, the processes of evaluation and communication of the artifact are described.

Figure 114 - Questions for the development of the fourth chart


Cycle Time (Max - Min)

Data

Vsm1

Summarize

Max of Cycle Time ✕ Min of Cycle Time ✕ + by Title ✕ +

Source: The Author (2023)

4.3 Model Evaluation and Communication

The DSR methodology's last activities, regarding the evaluation and communication of the artifact, will be described in this section. After the practical development of the use of the VSM system, an evaluative questionnaire (Appendix A) was sent to the engineering team of the partner company to obtain qualitative feedback on the applicability and usefulness of this system. This team comprises three engineers responsible for the operations lines used as a case study for this project. The questionnaire answered by them is in Appendix B.

Based on the results of the engineer's feedback, it was concluded that the engineering team was satisfied with all the functionalities and dynamics provided by the VSM application. They did not mark "very satisfied" with the accuracy due to not testing the tool with actual data, which could have produced results that could already be applied to the company's reality.

In addition to the results obtained from the evaluative questionnaire, as described in section 3.3.5, the results were evaluated according to their generality, efficiency, and applicability.

The system is compatible for use in any production line, with any amount of data, and in all applications created in the Docker network, as all are connected. The system mapping structure was developed to be flexible enough to adapt to any process in any company. However, the context of this project was for an integrated fuel tank supply and assembly process. The only limitation is the requirement to use the same process variables established in this development: cycle time, changeover time, availability, FPY, operator, and waiting time. Generality refers to the ability of an idea, concept, method, or theory to be applied to a wide range of situations or contexts rather than being limited to a single specific case (BRANCH; PENNYPACKER, 2013). In programming, generality can refer to the ability of a program or algorithm to work for different data sets or scenarios rather than being designed for only a single use case. Therefore, the system developed in this project is generalizable for use in other process areas and quantities of data but is not generalizable to the change of variables involved in mapping.

The efficiency of the project can be evaluated by the fulfillment of the second activity of the DSR methodology, which refers to the fulfillment of the research objectives. All objectives described in section 3.2.2 were effectively accomplished, and there was quality in the results presented in terms of response time, resource consumption, error quantity, portability, and ease of use.

The response time that the system uses is relatively low for changing pages and performing each task. The resource consumption required for virtualization in Docker Desktop is accessible, as there is no need for an internet connection since all processing of all applications can be done locally. The processor must support virtualization and have a minimum of 2 Gb of disk space and 8 Gb of RAM (Random Access Memory).

The number of errors relates to the quality of the database used. With the metrics used for the predictions, it is possible to identify that there is a way to keep control of these aspects even with changes in the data used. Once the system can handle different amounts of data and process them without losing performance, it is considered flexible and scalable. Finally, the system is easy to use due to the dynamic used in programming and being intuitive in mapping construction.

The evaluation of applicability was developed through analysis of its adaptability, flexibility, usability, reliability, security, cost-effectiveness, and update. According to the questionnaire results sent to the partner company, it was identified that the engineers who work at the partner company were satisfied. The engineers followed this development from the beginning. They identified that this system could be effectively used by the operators on the production line, using devices such as tablets to update and have real-time control of what is happening after each new observation.

The adaptability of this system in different scenarios or environments is confirmed by the adaptation to each new CSV file, regardless of whether the titles are different or the quantity of existing data lines. Moreover, there is the portability of Docker Desktop usage, which allows the installation of this system across different operating systems and cloud infrastructures without the need to adjust the code or application settings. Flexibility focuses on the users' needs, and since they can manipulate the construction of connections, include symbols, interact, and create charts, this brings a quality result to the system's applicability.

The system’s need for login and password access brings a positive result to its security. In terms of update aspects, as mentioned previously regarding adaptability aspects, whenever the user makes new observations on the production lines, they can update the CSV file and integrate these new updates to all applications that are connected to Docker.

The questionnaire evaluation and the generalizability, efficiency, and applicability evaluations identified in the system demonstration conclude the fifth activity of the Docker methodology.

The communication of the research, the sixth and final activity of the DSR methodology, is carried out for the publications developed based on this project. This communication is delimited to the final preparation of this document, along with the publication of five articles, three of which have already been published, one is in the process of publication, and the last one was recently submitted. The identification of the research opportunity and motivation was developed and published with the title

"Machine Learning and Process Mining applied to Process Optimization: Bibliometric and Systemic Analysis" (FERNANDES *et al.*, 2019) (Appendix C).

During the objective definition phase for a solution, an application for data generation was developed and published titled "Flexible Production Data Generator for Manufacturing Companies" (FERNANDES *et al.*, 2020) (Appendix D). Unfortunately, the integration of this application, which was developed using the Flask web framework, was incompatible with the data insertion using the flow-based modeling of the methodology used in this thesis. However, it was imperative to the knowledge acquired for defining the current methodology and essential to the programming skills learned throughout the project.

A proposal for a cloud computing framework was developed and published for this thesis (FERNANDES *et al.*, 2021) (Appendix E), but with a focus on another sector in the partner company involving improvements to worker ergonomics. Unfortunately, it could not be completed as the costs for this development's required cloud computing resources became too burdensome, making it unfeasible to continue.

The paper describing the initial development of the VSM application using the Django web framework has already been submitted and accepted (FERNANDES *et al.*, 2023) but only contains the current state mapping. The fifth article, which bears the same name as this thesis, was recently submitted and explained the full application. In addition, several presentations were made at conferences to demonstrate this project's advances and preliminary results.

5 CONCLUSIONS

Digital transformation in the manufacturing industry has brought various challenges and potential advantages in efficiency, quality, and flexibility, enabling greater automation and enhancement in multiple processes as digitally-enabled technologies lead to a paradigm shift in industrial production (SAVASTANO *et al.*, 2019). Digitalization has been critical for process management in various organizations, allowing for real-time data collection, analysis, and sharing, thus enabling more strategic decision-making and enhancing business processes. This approach is essential in innovation, driving companies' competitiveness and efficiency(AGOSTINI; GALATI; GASTALDI, 2019).

PPC provides efficient company management with strategies for organizing and coordinating manufacturing and production processes. Methods like VSM are helpful for this management, bringing a complete and integrated global view of the process. The digitalization of PPC methods aims to enhance process management. Thus, there was a search for how to adapt the construction and development of a method already consolidated and well-established in process management since the 1990s to a completely digital system, where it could bring the same essence of this method but with the support of technological resources that bring advantages to organizations.

Digital systems bring numerous benefits in precision, speed, storage, flexibility, costs, and communication. Therefore, the proposed solution for this research consisted of nine specific objectives for developing a single artifact: a dynamic and flexible web application with portability, easy deployment, environment isolation, scalability, ease of maintenance, resource management, and ease of sharing.

The development of this application brought many challenges in defining the best strategy to be used. This project underwent several changes in its methodology to identify the ideal path that would meet the established needs, which would be the virtual development of the VSM with the essence of how it is traditionally used but with a much more elaborate and intelligent dynamic.

All learning in back-end and front-end programming languages, database, DS, and AI were acquired without developing this project. In the first strategy for the project development, there was a study of the application of graph database tools. These tools are often used in social network analysis, fraud detection, and identity management applications. Some tools use query languages to access and manipulate data in the graph database. In this development, there was a study of alternatives using PM techniques, which would transform data into process networks, with representations between the relationships of activities, resources, decisions, and events recorded in the data. Due to incompatibilities and difficulties in developing a web application using HTML combined with PM techniques, the strategy had to be changed.

In the second development strategy, there was a study of creating a web application with hosting in cloud computing. This combination would bring portability and flexibility to the application, as cloud computing resources are scalable, secure, and perform better. There were several in-depth studies of existing cloud computing resources in different providers. However, the credits made available for free to students did not provide all the resources required to develop this application. The resources that needed to be obtained had a high acquisition cost, which made it impractical to carry out this project, which would require several tests until the complete development of the application.

Then, an alternative development strategy was created by creating a virtualized environment but installed on a local physical machine. This strategy was functional due to the applications and open-source virtualization environment. With the portability obtained by the virtualization environment, the development of the web application was enhanced with the initial use of the Flask web framework. However, it was later changed to the Django web framework based on identifying the robustness and scalability that the application would require.

Using the DSR methodology to delimit the activities of artifact construction was vital to the organization and successful completion of the project. The early definition of the solution and project development required an understanding of the importance of writing and reading CSV files with Node-Red, identifying and understanding the different approaches that can be achieved with the connection between programming languages, the scope, and flexibility of using BI systems, the applicability and reach that ML programs can achieve, the relevance of controlling and supervising data prediction through analysis of their metrics, and especially the ease, flexibility, and interactivity that containerization provides when integrating various applications into an isolated network. Understanding the structural base of each application was also very important for the functionalities of each one to complement each other for the best result.

The present research used the ETL process, which included extracting and compiling raw data, data transformation, and loading data into a target system. ETL was used to consolidate data from different sources and deliver them to a new environment. The applicability of this development occurred using this mapping in a case study in a multinational company of agricultural and construction equipment in Curitiba, Brazil.

The data obtained from the company were presented in the current state mapping, and improvement scenarios in the studied production line were developed to create and demonstrate future state mapping. Another challenge considered a limitation of this project was the absence of sufficient actual data to obtain more consistent and relevant predictions and therefore bring insights that could be applied directly to the company under study. However, the participation and monitoring of the partner company through feedback and suggestions for improvements in the meetings held were essential to achieve the best possible outcome.

The dynamic functionalities of information blocks, connections between blocks, and symbols were not included in the future mapping. While the main focus of the current state mapping was to explore the adaptation of diverse, dynamic functionalities that would provide the best usability for visualizing all the necessary data for the production flow, the focus of the future state mapping was set by its central differential, which was the use of ML algorithms to predict new information corresponding to the historical data used, seeking to bring accuracy in the identification of possible bottlenecks and the search for the best alternatives for an expected future flow.

Thus, with the demonstration and evaluation stages of the artifact, it was identified that the created solution had achieved the proposed objective. This project is considered original since no web-based VSM developments with traditional method functionalities combined with data analysis algorithms and ML, and all the dynamics provided by the JS programming language were identified in the literature. Among the main contributions of this project are the data-feeding strategy to the web system through the integration of Node-red, MySQL, and Django applications within the Docker virtualization environment and the development of a new tool for creating, editing, visually comparing, predicting, and assisting in decision-making for production planning and control.

5.1 Future Work

In future work, it is suggested to implement and test the system in various organizations and new contexts since the system can be updated through further feedback and evaluations to receive gradual improvements and become better over time.

Adapting the functionalities of connections with information blocks and including complementary symbols for future state mapping is also considered important as a complementary improvement of what has already been developed.

Additionally, adapting this system directly to cloud computing is considered highly relevant to allow ease of access on other devices, reduce operating costs, and provide greater agility in resource provisioning.

BIBLIOGRAPHY

ABOLHASSANI, Amir; HARNER, James; JARIDI, Majid; GOPALAKRISHNAN, Bhaskaran. Productivity enhancement strategies in North American automotive industry. **International Journal of Production Research**, vol. 56, no. 4, p. 1414–1431, 16 Feb. 2018. https://doi.org/10.1080/00207543.2017.1359700.

ACHARYA, Jigna N.; SUTHAR, Anil C. Docker Container Orchestration Management: A Review. **Springer**. [*S. l.*: *s. n.*], 2022. p. 140–153. https://doi.org/10.1007/978-3-030-97196-0_12.

AGOSTINI, Lara; GALATI, Francesco; GASTALDI, Luca. The digitalization of the innovation process. **European Journal of Innovation Management**, vol. 23, no. 1, p. 1–12, 20 Dec. 2019. https://doi.org/10.1108/EJIM-11-2019-0330.

ALMANEI, Mohammed; SALONITIS, Konstantinos; XU, Yuchun. Lean Implementation Frameworks: The Challenges for SMEs. **Procedia CIRP**, vol. 63, p. 750–755, 2017. https://doi.org/10.1016/j.procir.2017.03.170.

APACHE. **NiFi.** 2023a. Available at: https://nifi.apache.org/. Accessed on: 26 Feb. 2023.

APACHE. **The Cassandra Query Language (CQL).** 2023b. Available at: https://cassandra.apache.org/doc/latest/cassandra/cql/. Accessed on: 25 Feb. 2023.

APACHE. **Why Camel?** 2023c. Available at: https://camel.apache.org/. Accessed on: 26 Feb. 2023.

BAUER, Harald; BRANDL, Felix; LOCK, Christopher; REINHART, Gunther. Integration of Industrie 4.0 in Lean Manufacturing Learning Factories. **Procedia Manufacturing**, vol. 23, p. 147–152, 2018. https://doi.org/10.1016/j.promfg.2018.04.008.

BAYAZIT, Nigan. Investigating Design: A Review of Forty Years of Design Research. **Design Issues**, vol. 20, no. 1, p. 16–29, Jan. 2004. https://doi.org/10.1162/074793604772933739.

BECKER, Till; INTOYOAD, Wacharawan. Context Aware Process Mining in Logistics. **Procedia CIRP**, Procedia CIRP. Production Systems and Logistic Systems, Department of Production Engineering, University of Bremen, GermanyBremer Institut fur Produktion und Logistik, University of Bremen, Germany, vol. 63, p. 557–562, 2017. https://doi.org/10.1016/j.procir.2017.03.149.

BLACKHURST, Jennifer; RUNGTUSANATHAM, M. Johnny; SCHEIBE, Kevin; AMBULKAR, Saurabh. Supply chain vulnerability assessment: A network based visualization and clustering analysis approach. **Journal of Purchasing and Supply Management**, vol. 24, no. 1, p. 21–30, Jan. 2018. https://doi.org/10.1016/j.pursup.2017.10.004.

BOLAND JR., Richard J. 12 Design in the Punctuation of Management Action. **Managing as Designing**. [*S. l.*]: Stanford University Press, 2021. p. 106–112. https://doi.org/10.1515/9780804767439-014.

BORKOWSKI, Michael; FDHILA, Walid; NARDELLI, Matteo; RINDERLE-MA, Stefanie; SCHULTE, Stefan. Event-based failure prediction in distributed business processes. **Information Systems**, vol. 81, p. 220–235, Mar. 2019. https://doi.org/10.1016/j.is.2017.12.005.

BRANCH, Marc N.; PENNYPACKER, Henry S. Generality and generalization of research findings. **APA handbook of behavior analysis, Vol. 1: Methods and principles.** Washington: American Psychological Association, 2013. p. 151–175. https://doi.org/10.1037/13937-007.

BURGGRAF, Peter; WAGNER, Johannes; KOKE, Benjamin. Artificial intelligence in production management: A review of the current state of affairs and research trends in academia. Jan. 2018. **2018 International Conference on Information Management and Processing (ICIMP)** [...]. [*S. l.*]: IEEE, Jan. 2018. p. 82–88. DOI 10.1109/ICIMP1.2018.8325846. Available at: http://ieeexplore.ieee.org/document/8325846/. Accessed on: 18 Jan. 2020.

CADAVID, Juan Pablo Usuga; LAMOURI, Samir; GRABOT, Bernard; FORTIN, Arnaud. Machine Learning in Production Planning and Control: A Review of Empirical Literature. **IFAC-PapersOnLine**, vol. 52, no. 13, p. 385–390, 2019. https://doi.org/10.1016/j.ifacol.2019.11.155.

CARVALHO, Diogo v.; PEREIRA, Eduardo M.; CARDOSO, Jaime S. Machine Learning Interpretability: A Survey on Methods and Metrics. **Electronics**, vol. 8, no. 8, p. 832, 26 Jul. 2019. https://doi.org/10.3390/electronics8080832.

CELESTI, Antonio; MULFARI, Davide; FAZIO, Maria; VILLARI, Massimo; PULIAFITO, Antonio. Exploring Container Virtualization in IoT Clouds. May 2016. **2016 IEEE International Conference on Smart Computing (SMARTCOMP)** [...]. [*S. l.*]: IEEE, May 2016. p. 1–6. https://doi.org/10.1109/SMARTCOMP.2016.7501691.

DAI, Hong-Ning; WANG, Hao; XU, Guangquan; WAN, Jiafu; IMRAN, Muhammad. Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies. **Enterprise Information Systems**, vol. 14, no. 9–10, p. 1279–1303, 25 Nov. 2020. https://doi.org/10.1080/17517575.2019.1633689.

DBAEXPERTS. **Tipos de Bases de Datos NoSQL**. 2002. Available at: https://dbaexperts.tech/wp/bases-de-datos/tipos-de-bases-de-datos-nosql/. Accessed on: 25 Feb. 2023.

DINIS-CARVALHO, Jose; GUIMARAES, Levi; SOUSA, Rui M.; LEAO, Celina Pinto. Waste identification diagram and value stream mapping. **International Journal of Lean Six Sigma**, vol. 10, no. 3, p. 767–783, 16 Aug. 2019. https://doi.org/10.1108/IJLSS-04-2017-0030.

DIOUF, Papa Senghane; BOLY, Aliou; NDIAYE, Samba. Variety of data in the ETL processes in the cloud: State of the art. 8 May 2018. **2018 IEEE International Conference on Innovative Research and Development (ICIRD)** [...]. [*S. l.*]: IEEE, 8 May 2018. p. 1–5. https://doi.org/10.1109/ICIRD.2018.8376308.

DOCKER. **Docker Hub**. 2023. Available at: https://hub.docker.com/. Accessed on: 27 Feb. 2023.

DOMBROWSKI, Uwe; RICHTER, Thomas; KRENKEL, Philipp. Interdependencies of Industrie 4.0 & Lean Production Systems: A Use Cases Analysis. **Procedia Manufacturing**, vol. 11, p. 1061–1068, 2017. https://doi.org/10.1016/j.promfg.2017.07.217.

DORNELLES, Jéssica de Assis; AYALA, Néstor F.; FRANK, Alejandro G. Smart Working in Industry 4.0: How digital technologies enhance manufacturing workers' activities. **Computers & Industrial Engineering**, vol. 163, p. 107804, Jan. 2022. https://doi.org/10.1016/j.cie.2021.107804.

DUAN, Lian; XIONG, Ye. Big data analytics and business analytics. **Journal of Management Analytics**, vol. 2, no. 1, p. 1–21, 2 Jan. 2015. https://doi.org/10.1080/23270012.2015.1020891.

DUISEBEKOVA, Kulanda Seitbekovna; KHABIROV, Roman; ZHOLZHAN, Azamat. Django as Secure Web-Framework in Practice. **Вестник КазАТК**, vol. 116, no. 1, p. 275–281, 15 Mar. 2021. https://doi.org/10.52167/1609-1817-2021-116-1-275-281.

ECKERSON, Wayne. **Performance dashboards: Measuring, monitoring, and managing your business**. [*S. l.*]: Wiley, 2012. vol. 116, . https://doi.org/10.1002/9781119199984.

ECLIPSE**. Eclipse Mosquitto**. 2023. Available at: https://mosquitto.org/. Accessed on: 26 Feb. 2023.

EDELL, Laura. **Is Machine Learning the New EPM Black?** Available at: https://scorecardstreet.wordpress.com/2015/12/09/is-machine-learning-the-new-epm-black/. Accessed on: 9 Dec. 2015. .

EEKELS, J.; ROOZENBURG, N.F.M. A methodological comparison of the structures of scientific research and engineering design: their similarities and differences. **Design Studies**, vol. 12, no. 4, p. 197–203, Oct. 1991. https://doi.org/10.1016/0142-694X(91)90031-Q.

ELIAS, Micheline. **Enhancing User Interaction with Business Intelligence Dashboards**. 2012. 142 f. Ecole Centrale Paris, Paris, 2012.

ER, Mahendrawathi; ARSAD, Noval; ASTUTI, Hanim Maria; KUSUMAWARDANI, Renny Pradina; UTAMI, Rivia Atmajaningtyas. Analysis of production planning in a global manufacturing company with process mining. **Journal of Enterprise**

Information Management, vol. 31, no. 2, p. 317–337, 5 Mar. 2018. https://doi.org/10.1108/JEIM-01-2017-0003.

EVERMANN, Joerg; REHSE, Jana-Rebecca; FETTKE, Peter. Predicting process behaviour using deep learning. **Decision Support Systems**, vol. 100, p. 129–140, Aug. 2017. https://doi.org/10.1016/j.dss.2017.04.003.

FARRELL, Mary Beth. Statistics Refresher for Molecular Imaging Technologists, Part 2: Accuracy of Interpretation, Significance, and Variance. **Journal of Nuclear Medicine Technology**, vol. 46, no. 2, p. 76–80, Jun. 2018. https://doi.org/10.2967/jnmt.117.204719.

FERNANDES, Ederson Carvalhar; CAMATTI, Juliane Andressa; TABOR, Sandro Jessé Ferreira; ROMANEL, Luiz Gustavo; BROWN, Liam; BORSATO, Milton. Value Stream Mapping Application by using Django Web Framework. 2023. **Proceedings of 11th International Conference of Production Research - Americas - ICPR Americas 2022** [...]. Curitiba: Springer, 2023.

FERNANDES, Ederson Carvalhar; DOS SANTOS, Lucas Iuri; CAMATTI, Juliane Andressa; BROWN, Liam; BORSATO, Milton. Flexible Production Data Generator for Manufacturing Companies. **Procedia Manufacturing**, vol. 51, p. 1478–1484, 2020. https://doi.org/10.1016/j.promfg.2020.10.205.

FERNANDES, Ederson Carvalhar; FITZGERALD, Barry; BROWN, Liam; BORSATO, Milton. Machine Learning and Process Mining applied to Process Optimization: Bibliometric and Systemic Analysis. **Procedia Manufacturing**, vol. 38, p. 84–91, 2019. https://doi.org/10.1016/j.promfg.2020.01.012.

FERNANDES, Ederson Carvalhar; JESUS, Élcio Ricardo de; BROWN, Liam; BORSATO, Milton. Ergonomic and Production Data Analysis Framework for Cloud Computing. 2021. **Proceedings of 37th International Manufacturing Conference - IMC37** [...]. Available at: https://www.manufacturingcouncil.ie/imc-conference-archive. Athlone: Irish Manufacturing Council, 2021.

FRISBIE, Matt. **Professional JavaScript® for Web Developers**. [*S. l.*]: Wiley, 2019. https://doi.org/10.1002/9781119366560.

GARRE, Parthasarathy; NIKHIL BHARADWAJ, V.V.S.; SHIVA SHASHANK, P.; HARISH, Munigala; SAI DHEERAJ, M. Applying lean in aerospace manufacturing. **Materials Today: Proceedings**, vol. 4, no. 8, p. 8439–8446, 2017. https://doi.org/10.1016/j.matpr.2017.07.189.

GENGA, Laura; ALIZADEH, Mahdi; POTENA, Domenico; DIAMANTINI, Claudia; ZANNONE, Nicola. Discovering anomalous frequent patterns from partially ordered event logs. **Journal of Intelligent Information Systems**, Journal of Intelligent Information Systems. Eindhoven University of Technology, Eindhoven, NetherlandsUniversita Politecnica delle Marche, Ancona, Italy, vol. 51, no. 2, p. 257–300, 24 Oct. 2018. https://doi.org/10.1007/s10844-018-0501-z.

GERHARDT, Ricardo; VALIATI, João F.; CANTO DOS SANTOS, José Vicente. An Investigation to Identify Factors that Lead to Delay in Healthcare Reimbursement Process: A Brazilian case. **Big Data Research**, vol. 13, p. 11–20, Sep. 2018. https://doi.org/10.1016/j.bdr.2018.02.006.

GHIMIRE, Devndra. **Comparative study on Python web frameworks: Flask and Django**. 2020. 40 f. Metropolia University of Applied Sciences, Metropolia, 2020.

GHOLAMI, Hamed; JAMIL, Norhazrina; MAT SAMAN, Muhamad Zameri; STREIMIKIENE, Dalia; SHARIF, Safian; ZAKUAN, Norhayati. The application of Green Lean Six Sigma. **Business Strategy and the Environment**, vol. 30, no. 4, p. 1913–1931, 5 May 2021. https://doi.org/10.1002/bse.2724.

GIBELHAUS, Andrej ; POSTWEILER, Patrik ; SEILER, Jan ; BARDOW, André; GIBELHAUS, Andrej; PATRIK; SEILER, Jan; BARDOW; ANDRÉ. Computationally efficient, experimentally validated adsorption chiller model using a plug-flow-based modelling approach. **research-collection.ethz.ch**, [*s. d.*]. https://doi.org/10.3929/ethz-b-000459486.

GIL, Antonio Carlos. **Como Elaborar Projetos de Pesquisa**. 4th ed. São Paulo: Atlas, 2002.

GORE, Himanshu; SINGH, Rakesh Kumar; SINGH, Ashutosh; SINGH, Arnav Pratap; SHABAZ, Mohammad; SINGH, Bhupesh Kumar; JAGOTA, Vishal. Django: Web development simple & fast. **Annals of the Romanian Society for Cell Biology**, vol. 25, no. 6, p. 4576–4585, 2021. .

HANDELMAN, Guy S.; KOK, Hong Kuan; CHANDRA, Ronil v.; RAZAVI, Amir H.; HUANG, Shiwei; BROOKS, Mark; LEE, Michael J.; ASADI, Hamed. Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. **American Journal of Roentgenology**, vol. 212, no. 1, p. 38–43, 1 Jan. 2019. https://doi.org/10.2214/AJR.18.20224.

HEVNER, Alan; CHATTERJEE, Samir. **Design Research in Information Systems**. Boston, MA: Springer US, 2010. vol. 22, (Integrated Series in Information Systems). https://doi.org/10.1007/978-1-4419-5653-8.

HINDLE, Giles A.; VIDGEN, Richard. Developing a business analytics methodology: A case study in the foodbank sector. **European Journal of Operational Research**, vol. 268, no. 3, p. 836–851, 1 Aug. 2018. https://doi.org/10.1016/j.ejor.2017.06.031.

HOLOVATY, Adrian; KAPLAN-MOSS, Jacob. **The Definitive Guide to Django: Web Development Done Right**. 2nd ed. Berkeley: Apress, 2009.

HORSTHOFER-RAUCH, Julia; SCHUMANN, Marek; MILDE, Michael; VERNIM, Susanne; REINHART, Gunther. Digitalized value stream mapping: review and outlook. **Procedia CIRP**, vol. 112, p. 244–249, 2022. https://doi.org/10.1016/j.procir.2022.09.079.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, vol. 349, no. 6245, p. 255–260, 17 Jul. 2015. https://doi.org/10.1126/science.aaa8415.

JOVANOSKA, Dijana; PETREVSKA NECHKOSKA, Renata; MANCHESKI, Gjorgji. Metabase Cockpits as a Base for BI in Strategic Management. 2021. **Proceedings of the 26th International Scientific Conference Strategic Management and Decision Support Systems in Strategic Management** [...]. [*S. l.*]: University of Novi Sad, Faculty of Economics in Subotica, 2021. https://doi.org/10.46541/978-86-7233-397-8_116.

KARIM, M.S. Abdul; TAI TIONG, Chan. Development of a Simple and Affordable Computer Aided Process Planning (CAPP). Mar. 2019. **2019 Advances in Science and Engineering Technology International Conferences (ASET)** [...]. [*S. l.*]: IEEE, Mar. 2019. p. 1–4. https://doi.org/10.1109/ICASET.2019.8714443.

KARRAY, Mohamed-Hedi; CHEBEL-MORELLO, Brigitte; ZERHOUNI, Noureddine. PETRA: Process Evolution using a TRAce-based system on a maintenance platform. **Knowledge-Based Systems**, vol. 68, p. 21–39, Sep. 2014. https://doi.org/10.1016/j.knosys.2014.03.010.

KNIME. **Knime Analytics Platform**. 2023. Available at: https://www.knime.com/. Accessed on: 26 Feb. 2023.

KOLBERG, Dennis; ZÜHLKE, Detlef. Lean Automation enabled by Industry 4.0 Technologies. **IFAC-PapersOnLine**, vol. 48, no. 3, p. 1870–1875, 2015. https://doi.org/10.1016/j.ifacol.2015.06.359.

KRISHNASWAMY, Pavitra; RAMASAMY, Savitha; AL-DUJAILI, Abdullah; SRIDHARAN, Srinath; GOH, Geraldine; CHUEN, Tong Shao; AUNG, Khin Chaw Yu; TOH, Gerard Leong Kui; MACDONALD, Michael Ross; GUANG, Sheldon Lee Shao; YAN, Cao; SUNDARAM, Suresh; LEN, Chow Wai. A Predictive Analytics Methodology to Assess and Optimize Readmission Risk in Heart Failure Patients. **Journal of Medical Systems**, vol. 43, p. 293, 5 Mar. 2019. .

KUMAR, R. **Operations Management**. 1st ed. Kolenchery: Jyothis Publishers, 2022.

LACERDA, António Pedro; XAMBRE, Ana Raquel; ALVELOS, Helena Maria. Applying Value Stream Mapping to eliminate waste: a case study of an original equipment manufacturer for the automotive industry. **International Journal of Production Research**, vol. 54, no. 6, p. 1708–1720, 18 Mar. 2016. https://doi.org/10.1080/00207543.2015.1055349.

LAI, Nai Yeen Gavin; WONG, Kok Hoong; HALIM, Dunant; LU, Jiawa; KANG, Hooi Siang. Industry 4.0 Enhanced Lean Manufacturing. 9 Mar. 2019. **2019 8th International Conference on Industrial Technology and Management (ICITM)** [...]. [*S. l.*]: IEEE, 9 Mar. 2019. p. 206–211. https://doi.org/10.1109/ICITM.2019.8710669.

LARSON, Deanne; CHANG, Victor. A review and future direction of agile, business intelligence, analytics and data science. **International Journal of Information Management**, vol. 36, no. 5, p. 700–710, Oct. 2016. https://doi.org/10.1016/j.ijinfomgt.2016.04.013.

LATORRE-BIEL, Juan-Ignacio; FAULÍN, Javier; JUAN, Angel A.; JIMÉNEZ-MACÍAS, Emilio. Petri Net Model of a Smart Factory in the Frame of Industry 4.0. **IFAC-PapersOnLine**, Institute of Smart Cities, Public University of Navarre, SpainInternet Interdisciplinary Institute, Dept. of Computer Science, Open University of Catalonia, SpainDept. of Electrical Engineering, University of La Rioja, Spain, vol. 51, no. 2, p. 266–271, 2018. https://doi.org/10.1016/j.ifacol.2018.03.046.

LATORRE-BIEL, Juan-Ignacio; JIMÉNEZ-MACÍAS, Emilio; PÉREZ DE LA PARTE, Mercedes; BLANCO-FERNÁNDEZ, Julio; MARTÍNEZ-CÁMARA, Eduardo. Control of Discrete Event Systems by Means of Discrete Optimization and Disjunctive Colored PNs: Application to Manufacturing Facilities. **Abstract and Applied Analysis**, vol. 2014, p. 1–16, Jan. 2014. https://doi.org/10.1155/2014/821707.

LATORRE-BIEL, Juan-Ignacio; JIMÉNEZ-MACÍAS, Emilio; PÉREZ-PARTE, Mercedes. Sequence of decisions on discrete event systems modeled by Petri nets with structural alternative configurations. **Journal of Computational Science**, vol. 5, no. 3, p. 387–394, May 2014. https://doi.org/10.1016/j.jocs.2013.09.001.

LEE, Jay; DAVARI, Hossein; SINGH, Jaskaran; PANDHARE, Vibhor. Industrial Artificial Intelligence for industry 4.0-based manufacturing systems. **Manufacturing Letters**, vol. 18, p. 20–23, Oct. 2018. https://doi.org/10.1016/j.mfglet.2018.09.002.

LEE, Wai Lam Jonathan; PARRA, Denis; MUNOZ-GAMA, Jorge; SEPÚLVEDA, Marcos. Predicting process behavior meets factorization machines. **Expert Systems with Applications**, vol. 112, p. 87–98, Dec. 2018. https://doi.org/10.1016/j.eswa.2018.05.035.

LEEMANS, Maikel; VAN DER AALST, Wil M. P.; VAN DEN BRAND, Mark G. J. Hierarchical performance analysis for process mining. 26 May 2018. **Proceedings of the 2018 International Conference on Software and System Process** [...]. New York, NY, USA: ACM, 26 May 2018. p. 96–105. https://doi.org/10.1145/3202710.3203151.

LEKSIC, I.; STEFANIC, N.; VEZA, I. The impact of using different lean manufacturing tools on waste reduction. **Advances in Production Engineering & Management**, vol. 15, no. 1, p. 81–92, 31 Mar. 2020. https://doi.org/10.14743/apem2020.1.351.

LI, Jun; MENG, Xianghu; ZHOU, MengChu; DAI, Xianzhong. A Two-Stage Approach to Path Planning and Collision Avoidance of Multibridge Machining Systems. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, vol. 47, no. 7, p. 1039–1049, Jul. 2017. https://doi.org/10.1109/TSMC.2016.2531648.

LI, Nian; ZHANG, Bo. The Research on Single Page Application Front-end development Based on Vue. **Journal of Physics: Conference Series**, vol. 1883, no. 1, p. 012030, 1 Apr. 2021. https://doi.org/10.1088/1742-6596/1883/1/012030.

LIU, Qingqi; YANG, Hualong; XIN, Yuchen. Applying value stream mapping in an unbalanced production line: A case study of a Chinese food processing enterprise. **Quality Engineering**, vol. 32, no. 1, p. 111–123, 2 Jan. 2020. https://doi.org/10.1080/08982112.2019.1637526.

LOBO, Cicero Vasconcelos Ferreira; CALADO, Robisom Damasceno; CONCEIÇÃO, Roberta Dalvo Pereira da. Evaluation of value stream mapping (VSM) applicability to the oil and gas chain processes. **International Journal of Lean Six Sigma**, vol. 11, no. 2, p. 309–330, 16 Nov. 2018. https://doi.org/10.1108/IJLSS-05-2018-0049.

LONG, F.; ZEILER, P.; BERTSCHE, B. Modelling the production systems in industry 4.0 and their availability with high-level Petri nets. **IFAC-PapersOnLine**, vol. 49, no. 12, p. 145–150, 2016. https://doi.org/10.1016/j.ifacol.2016.07.565.

LUCIDCHART. **What is Value Stream Mapping**. 2018. Available at: https://www.lucidchart.com/pages/value-stream-mapping. Accessed on: 25 Feb. 2023.

LUGERT, Andreas; BATZ, Aglaya; WINKLER, Herwig. Empirical assessment of the future adequacy of value stream mapping in manufacturing industries. **Journal of Manufacturing Technology Management**, vol. 29, no. 5, p. 886–906, 23 May 2018. https://doi.org/10.1108/JMTM-11-2017-0236.

MACHESO, Paul S.B.; MANDA, Tiwonge D.; MEELA, Angel G.; MLATHO, Justice S.; TAULO, Gracian T.; PHIRI, Joseph C. Industrial Temperature Monitor Based on NodeMCU ESP8266, MQTT and Node-RED. 17 Dec. 2021. **2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)** [...]. [*S. l.*]: IEEE, 17 Dec. 2021. p. 740–743. https://doi.org/10.1109/ICAC3N53548.2021.9725469.

MALINDZAKOVA, Marcela; MALINDZAK, Dusan; GARAJ, Patrik. Implementation of the Single Minute Exchange of Dies method for reducing changeover time in a hygiene production company. **International Journal of Industrial Engineering and Management**, vol. 12, no. 4, p. 243–252, 30 Dec. 2021. https://doi.org/10.24867/IJIEM-2021-4-291.

MANUEL, Filipe; NUNES, Ferreirinho. **Open source business intelligence**. 2012. 148 f. Instituto Universitário de Lisboa, Lisboa, 2012.

MARTIN, Karen; OSTERLING, Mike. **Value Stream Mapping - How to Visualize Work and Align Leadership for Organizational Transformation**. 1st ed. New York: McGraw-Hill Education, 2013.

MARTINUZZI, André; BLOK, Vincent; BREM, Alexander; STAHL, Bernd; SCHÖNHERR, Norma. Responsible Research and Innovation in Industry—

Challenges, Insights and Perspectives. **Sustainability**, Sebastapol, CA, vol. 10, no. 3 ed. First Edit, p. 702, 5 Mar. 2018. https://doi.org/10.3390/su10030702.

MATHWORKS. **Simulink is for Model-Based Design**. 2023. Available at: https://www.mathworks.com/products/simulink.html. Accessed on: 26 Feb. 2023.

MAYR, A.; WEIGELT, M.; KÜHL, A.; GRIMM, S.; ERLL, A.; POTZEL, M.; FRANKE, J. Lean 4.0 - A conceptual conjunction of lean management and Industry 4.0. **Procedia CIRP**, vol. 72, p. 622–628, 2018. https://doi.org/10.1016/j.procir.2018.03.292.

MCKENDRICK, Russ; GALLAGHER, Scott. **Mastering Docker.: Master this widely used containerization tool**. 2nd ed. Birminghan: Packt Publishing, 2017.

MELÉ, Antonio. **Django 3 by example : build powerful and reliable Python web applications from scratch**. 3rd ed. Birminghan: Packt Publishing, 2020.

METABASE. **Metabase**. 2023. Available at: https://www.metabase.com/product/. Accessed on: 26 Feb. 2023.

MEYER, Georg; ADOMAVICIUS, Gediminas; JOHNSON, Paul E.; ELIDRISI, Mohamed; RUSH, William A.; SPERL-HILLEN, JoAnn M.; O'CONNOR, Patrick J. A Machine Learning Approach to Improving Dynamic Decision Making. **Information Systems Research**, vol. 25, no. 2, p. 239–263, Jun. 2014. https://doi.org/10.1287/isre.2014.0513.

MISHRA, Biswajeeban; MISHRA, Biswaranjan; KERTESZ, Attila. Stress-Testing MQTT Brokers: A Comparative Analysis of Performance Measurements. **Energies**, vol. 14, no. 18, p. 5817, 14 Sep. 2021. https://doi.org/10.3390/en14185817.

MISRA, N. N.; DIXIT, Yash; AL-MALLAHI, Ahmad; BHULLAR, Manreet Singh; UPADHYAY, Rohit; MARTYNENKO, Alex. IoT, Big Data, and Artificial Intelligence in Agriculture and Food Industry. **IEEE Internet of Things Journal**, vol. 9, no. 9, p. 6305–6324, 1 May 2022. https://doi.org/10.1109/JIOT.2020.2998584.

MONGODB. **Getting Started with Atlas and the MongoDB Query API**. 2022. Available at: https://www.mongodb.com/developer/products/atlas/getting-started-atlas-mongodb-query-language-mql/. Accessed on: 25 Feb. 2023.

MORAIS, Márcia de Fátima; BOIKO, Thays J. Perassoli. Metodologia de Pesquisa : uma proposta de estrutura para pesquisas técnico-científicas em Engenharia de Produção. 1., 2014. **VIIIE Encontro de Engenharia de Produção Agroindustrial** [...]. [*S. l.*: *s. n.*], 2014. vol. 1, p. 12.

MORARIU, Cristina; MORARIU, Octavian; RĂILEANU, Silviu; BORANGIU, Theodor. Machine learning for predictive scheduling and resource allocation in large scale manufacturing systems. **Computers in Industry**, vol. 120, p. 103244, 1 Sep. 2020. https://doi.org/10.1016/j.compind.2020.103244.

MORAVCIK, Marek; KONTSEK, Martin. Overview of Docker container orchestration tools. 2020. **18th International Conference on Emerging eLearning Technologies and Applications (ICETA)** [...]. [*S. l.*: *s. n.*], 2020. p. 475–480.

MOREIRA, João Mendes; DE CARVALHO, André C. P. L. F.; HORVÁTH, Tomáš. **A General Introduction to Data Analytics**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2018. https://doi.org/10.1002/9781119296294.

MRUGALSKA, Beata; WYRWICKA, Magdalena K. Towards Lean Production in Industry 4.0. **Procedia Engineering**, vol. 182, p. 466–473, 2017. https://doi.org/10.1016/j.proeng.2017.03.135.

MÜELLER, Andreas C.; GUIDO, Sarah. **Introduction to Machine Learning with Python: A guide for Data Scientists**. 1st ed. Sebastapol: O'Reilly Media, 2016.

MUKHERJEE, Rajendrani; KAR, Pragma. A Comparative Review of Data Warehousing ETL Tools with New Trends and Industry Insight. 12 Jan. 2017. **2017 IEEE 7th International Advance Computing Conference (IACC)** [...]. [*S. l.*]: IEEE, 12 Jan. 2017. p. 943–948. https://doi.org/10.1109/IACC.2017.0192.

MUKHERJEE, Shubham; SEN, Pritam; BORA, Sudeshna; PRADHAN, Chittaranjan. SQL Injection: A sample review. 8., 1 Jul. 2015. **2015 6th International Conference on Computing, Communication and Networking Technologies (ICCCNT)** [...]. [*S. l.*]: IEEE, 1 Jul. 2015. vol. 8, p. 1–7. https://doi.org/10.1109/ICCCNT.2015.7395166.

MYERS, David; SURIADI, Suriadi; RADKE, Kenneth; FOO, Ernest. Anomaly detection for industrial control systems using process mining. **Computers & Security**, vol. 78, p. 103–125, Sep. 2018. https://doi.org/10.1016/j.cose.2018.06.002.

NAMDEV, Neeraj; AGRAWAL, Shikha; SILKARI, Sanjay. Recent Advancement in Machine Learning Based Internet Traffic Classification. **Procedia Computer Science**, vol. 60, no. 1, p. 784–791, 2015. https://doi.org/10.1016/j.procs.2015.08.238.

NATIONAL INSTRUMENTS. **What is LabVIEW**. 2023. Available at: https://www.ni.com/en-us/shop/labview.html. Accessed on: 26 Feb. 2023.

NAYAK, Ameya; PORIYA, Anil; POOJARY, Dikshay. Type of NOSQL databases and its comparison with relational databases. **International Journal of Applied Information Systems**, vol. 5, no. 4, p. 16–19, 2013. .

NETEK, Rostislav; BRUS, Jan; TOMECKA, Ondrej. Performance Testing on Marker Clustering and Heatmap Visualization Techniques: A Comparative Study on JavaScript Mapping Libraries. **ISPRS International Journal of Geo-Information**, vol. 8, no. 8, p. 348, 1 Aug. 2019. https://doi.org/10.3390/ijgi8080348.

NEUPANE, Bhanu; ECHAIZ, Lucia Flores. **Steering AI and Advanced ICTs for Knowledge Societies**. Paris: UNESCO Publishing, 2019. Available at: https://en.unesco.org/unesco-series-on-internet-freedom.

NODE-RED. Node-RED. 2023. Available at: https://nodered.org/. Accessed on: 26 Feb. 2023.

PALANGE, Atul; DHATRAK, Pankaj. Lean manufacturing a vital tool to enhance productivity in manufacturing. **Materials Today: Proceedings**, vol. 46, p. 729–736, 1 Jan. 2021. https://doi.org/10.1016/j.matpr.2020.12.193.

PAPER, David. **Hands-on Scikit-Learn for Machine Learning Applications**. Berkeley, CA: Apress, 2020. DOI 10.1007/978-1-4842-5373-1. Available at: http://link.springer.com/10.1007/978-1-4842-5373-1.

PARK, Jinkyun; JUNG, Jae-Yoon; HEO, Gyunyoung; KIM, Yochan; KIM, Jaewhan; CHO, Jaehyun. Application of a process mining technique to identifying information navigation characteristics of human operators working in a digital main control room – feasibility study. **Reliability Engineering & System Safety**, vol. 175, p. 38–50, Jul. 2018. https://doi.org/10.1016/j.ress.2018.03.003.

PASCHEK, Daniel; LUMINOSU, Caius Tudor; DRAGHICI, Anca. Automated business process management – in times of digital transformation using machine learning or artificial intelligence. **MATEC Web of Conferences**, vol. 121, p. 04007, 9 Aug. 2017. DOI 10.1051/matecconf/201712104007. Available at: http://www.matec-conferences.org/10.1051/matecconf/201712104007.

PEDROSA, Javier; PUIG, Vicenç; NEJJARI, Fatiha. Health-Aware Economic MPC for Operational Management of Flow-Based Networks Using Bayesian Networks. **Water**, vol. 14, no. 10, p. 1538, 11 May 2022. DOI 10.3390/w14101538. Available at: https://www.mdpi.com/2073-4441/14/10/1538.

PEFFERS, Ken; TUUNANEN, Tuure; ROTHENBERGER, Marcus A.; CHATTERJEE, Samir. A Design Science Research Methodology for Information Systems Research. **Journal of Management Information Systems**, vol. 24, no. 3, p. 45–77, 8 Dec. 2007. DOI 10.2753/MIS0742-1222240302. Available at: https://www.tandfonline.com/doi/full/10.2753/MIS0742-1222240302.

PERMIN, Eike; BERTELSMEIER, Felix; BLUM, Matthias; BÜTZLER, Jennifer; HAAG, Sebastian; KUZ, Sinem; ÖZDEMIR, Denis; STEMMLER, Sebastian; THOMBANSEN, Ulrich; SCHMITT, Robert; BRECHER, Christian; SCHLICK, Christopher; ABEL, Dirk; POPRAWE, Reinhart; LOOSEN, Peter; SCHULZ, Wolfgang; SCHUH, Günther. Self-optimizing Production Systems. **Procedia CIRP**, vol. 41, p. 417–422, 2016. https://doi.org/10.1016/j.procir.2015.12.114.

PIANA, Valerio Junior; SILVA PINTO JUNIOR, Joelias; GONÇALVES, Alexandre Leopoldo. Accessibility of the node-red iot framework: an integrative review. **Revista e-TECH: Tecnologias para Competitividade Industrial - ISSN - 1983-1838**, vol. 15, no. 4, 30 Dec. 2022. https://doi.org/10.18624/etech.v15i4.1231.

POORNIMA, S.; PUSHPALATHA, M. A survey of predictive analytics using big data with data mining. **International Journal of Bioinformatics Research and**

Applications, vol. 14, no. 3, p. 269, 2018. https://doi.org/10.1504/IJBRA.2018.092697.

POURMASOUMI, Asef; BAGHERI, Ebrahim. Business process mining. **Encyclopedia with Semantic Computing and Robotic Intelligence**, vol. 01, no. 01, p. 1630004, 23 Mar. 2017. https://doi.org/10.1142/S2425038416300044.

RAMANI, Prasanna Venkatesan; KSD, Laxmana Kumara Lingan. Application of lean in construction using value stream mapping. **Engineering, Construction and Architectural Management**, vol. 28, no. 1, p. 216–228, 3 Feb. 2021. https://doi.org/10.1108/ECAM-12-2018-0572.

RATTANAPOKA, Choopan; CHANTHAKIT, Somphop; CHIMCHAI, Apatsaraporn; SOOKKEAW, Amorntip. An MQTT-based IoT Cloud Platform with Flow Design by Node-RED. Dec. 2019. **2019 Research, Invention, and Innovation Congress (RI2C)** [...]. [*S. l.*]: IEEE, Dec. 2019. p. 1–6. https://doi.org/10.1109/RI2C48728.2019.8999942.

RAUCH, Erwin; DALLASEGA, Patrick; MATT, Dominik T. Synchronization of Engineering, Manufacturing and on-site Installation in Lean ETO-Enterprises. **Procedia CIRP**, vol. 37, p. 128–133, 2015. https://doi.org/10.1016/j.procir.2015.08.047.

RAWAT, Govind Singh; GUPTA, Ashutosh; JUNEJA, Chandan. Productivity Measurement of Manufacturing System. **Materials Today: Proceedings**, vol. 5, no. 1, p. 1483–1489, 2018. https://doi.org/10.1016/j.matpr.2017.11.237.

RED HAT. **Containers and Virtual Machines (VMs).** 2020. Available at: https://www.redhat.com/pt-br/topics/containers/containers-vs-vms. Accessed on: 27 Feb. 2023.

REID, R. Dan; SANDERS, Nada R. **Operations Management - An Integrated Approach**. 7th ed. [*S. l.*]: Wiley, 2019.

RICHTER, Ralph; SYBERG, Marius; DEUSE, Jochen; WILLATS, Peter; LENZE, David. Creating lean value streams through proactive variability management. **International Journal of Production Research**, , p. 1–12, 18 Aug. 2022. https://doi.org/10.1080/00207543.2022.2111614.

ROMERO, Carlos Andrés Tavera; ORTIZ, Jesús Hamilton; KHALAF, Osamah Ibrahim; PRADO, Andrea Ríos. Business Intelligence: Business Evolution after Industry 4.0. **Sustainability**, vol. 13, no. 18, p. 10026, 7 Sep. 2021. https://doi.org/10.3390/su131810026.

ROSE, Ahmad Nasser Mohd; DEROS, B.M.; RAHMAN, Mohd Nizam Ab; NORANI, Nordin. Lean manufacturing best practices in SMEs. 1., Jan. 2011. **International Conference on Industrial Engineering and Operation Management** [...]. Kuala Lumpur: [*s. n.*], Jan. 2011. vol. 1, p. 872–877.

SAHOO, Kabita; SAMAL, Abhaya Kumar; PRAMANIK, Jitendra; PANI, Subhendu Kumar. Exploratory Data Analysis using Python. **International Journal of Innovative Technology and Exploring Engineering**, vol. 8, no. 12, p. 4727–4735, 30 Oct. 2019. https://doi.org/10.35940/ijitee.L3591.1081219.

SALONITIS, Konstantinos; TSINOPOULOS, Christos. Drivers and Barriers of Lean Implementation in the Greek Manufacturing Sector. **Procedia CIRP**, vol. 57, p. 189–194, 2016. https://doi.org/10.1016/j.procir.2016.11.033.

SARKER, Iqbal H. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. **SN Computer Science**, vol. 2, no. 5, p. 377, 12 Sep. 2021. https://doi.org/10.1007/s42979-021-00765-8.

SARTAL, Antonio; LLACH, Josep; VÁZQUEZ, Xosé H.; DE CASTRO, Rodolfo. How much does Lean Manufacturing need environmental and information technologies? **Journal of Manufacturing Systems**, vol. 45, p. 260–272, Oct. 2017. https://doi.org/10.1016/j.jmsy.2017.10.005.

SAVASTANO, Marco; AMENDOLA, Carlo; BELLINI, Francesco; D'ASCENZO, Fabrizio. Contextual Impacts on Industrial Processes Brought by the Digital Transformation of Manufacturing: A Systematic Review. **Sustainability**, vol. 11, no. 3, p. 891, 9 Feb. 2019. https://doi.org/10.3390/su11030891.

SAVICKAS, Titas; VASILECAS, Olegas. Belief network discovery from event logs for business process analysis. **Computers in Industry**, vol. 100, p. 258–266, Sep. 2018. https://doi.org/10.1016/j.compind.2018.04.020.

SCHMIDT, Rainer; MÖHRING, Michael; HÄRTING, Ralf-Christian; REICHSTEIN, Christopher; NEUMAIER, Pascal; JOZINOVIĆ, Philip. Industry 4.0 - Potentials for Creating Smart Products: Empirical Research Results. **Lecture Notes in Business Information Processing**. [*S. l.*: *s. n.*], 2015. vol. 208, p. 16–27. https://doi.org/10.1007/978-3-319-19027-3_2.

SEIFULLINA, Aziza; ER, Ahmet; NADEEM, Simon Peter; GARZA-REYES, Jose Arturo; KUMAR, Vikas. A Lean Implementation Framework for the Mining Industry. **IFAC-PapersOnLine**, vol. 51, no. 11, p. 1149–1154, 2018. https://doi.org/10.1016/j.ifacol.2018.08.435.

SHARDA, Ramesh; DELEN, Dursun; TURBAN, Efraim. **Business intelligence, analytics, and data science: a managerial perspective**. 1st ed. London: Pearson, 2018.

SHAW, Ben; BADHWAR, Saurabh; BIRD, Andrew; KS, Bharath Chandra; GUEST, Chris. **Web Development with Django: Learn to build modern web applications with a Python-based framework**. Birmingham: Packt Publishing, 2021.

SICARI, Sabrina; RIZZARDI, Alessandra; COEN-PORISINI, Alberto. Smart transport and logistics: A Node-RED implementation. **Internet Technology Letters**, vol. 2, no. 2, p. e88, Mar. 2019. https://doi.org/10.1002/itl2.88.

SILVA, Edna Lúcia da; MENEZES, Estera Muszkat. **Metodologia de Pesquisa e Elaboração de Dissertação**. 4th ed. Florianopolis: UFSC, 2005.

SILVA, Vitor Goncalves da; KIRIKOVA, Marite; ALKSNIS, Gundars. Containers for Virtualization: An Overview. **Applied Computer Systems**, vol. 23, no. 1, p. 21–27, 1 May 2018. https://doi.org/10.2478/acss-2018-0003.

SLAATS, Tijs. Declarative and Hybrid Process Discovery: Recent Advances and Open Challenges. **Journal on Data Semantics**, vol. 9, no. 1, p. 3–20, 19 Mar. 2020. https://doi.org/10.1007/s13740-020-00112-9.

SONI, Dipa; MAKWANA, Ashwin. A survey on MQTT: a protocol of internet of things (iot). 2017. **International Conference on Telecommunication, Power Analysis and Computing Techniqes (ICTPACT - 2017)** [...]. Chennai: [*s. n.*], 2017.

STAICU, Cristian-Alexandru; SCHOEPE, Daniel; BALLIU, Musard; PRADEL, Michael; SABELFELD, Andrei. An Empirical Study of Information Flows in Real-World JavaScript. 15 Nov. 2019. **Proceedings of the 14th ACM SIGSAC Workshop on Programming Languages and Analysis for Security** [...]. New York, NY, USA: ACM, 15 Nov. 2019. p. 45–59. https://doi.org/10.1145/3338504.3357339.

SULTAN, Sahira; KHODABANDEHLOO, Aida. **Improvement of Value Stream Mapping and Internal Logistics through Digitalization: A study in the context of Industry 4.0**. 2020. Mälardalen University, Västerås, 2020.

SUNYAEV, Ali. Cloud Computing. **Internet Computing**. Cham: Springer International Publishing, 2020. p. 195–236. https://doi.org/10.1007/978-3-030-34957-8_7.

SUSILAWATI, Anita; TAN, John; BELL, David; SARWAR, Mohammed. Fuzzy logic based method to measure degree of lean activity in manufacturing industry. **Journal of Manufacturing Systems**, vol. 34, no. C, p. 1–11, Jan. 2015. https://doi.org/10.1016/j.jmsy.2014.09.007.

THOUTAM, Vivek. A Study On Python Web Application Framework. **Journal of Electronics,Computer Networking and Applied Mathematics**, no. 11, p. 48–55, 2 Sep. 2021. https://doi.org/10.55529/jecnam.11.48.55.

TIERNEY, Brendan; KELLEHER, John D. **Data Science**. 1st ed. Cambridge: The MIT Press, 2018.

TURNBULL, James. **The Docker Book: Containerization is the new virtualization**. Melbourne: James Turnbull, 2014.

TYAGI, Satish; CHOUDHARY, Alok; CAI, Xianming; YANG, Kai. Value stream mapping to reduce the lead-time of a product development process. **International Journal of Production Economics**, vol. 160, p. 202–212, Feb. 2015. https://doi.org/10.1016/j.ijpe.2014.11.002.

UMER, Rahila; SUSNJAK, Teo; MATHRANI, Anuradha; SURIADI, Suriadi. On predicting academic performance with process mining in learning analytics. **Journal of Research in Innovative Teaching & Learning**, vol. 10, no. 2, p. 160–176, 3 Jul. 2017. https://doi.org/10.1108/JRIT-09-2017-0022.

VALAMEDE, Luana Sposito; AKKARI, Alessandra Cristina Santos. Lean 4.0: A New Holistic Approach for the Integration of Lean Manufacturing Tools and Digital Technologies. **International Journal of Mathematical, Engineering and Management Sciences**, vol. 5, no. 5, p. 851–868, 1 Oct. 2020. https://doi.org/10.33889/IJMEMS.2020.5.5.066.

VAN DER AALST, Wil. **Process Mining - Data Science in Action**. 2nd ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016. https://doi.org/10.1007/978-3-662-49851-4.

VANDERPLAS, Jake. **Python Data Science Handbook: Essential Tools for Working**. Sebastopol: O’Reilly , 2016.

VARSHA, Ramesh; NAIR, Siddharth M.; TYAGI, Amit Kumar; ASWATHY, Su; RADHAKRISHNAN, R. The Future with Advanced Analytics: A Sequential Analysis of the Disruptive Technology’s Scope. **Advances in Intelligent Systems and Computing**. [*S. l.*]: Springer Science and Business Media Deutschland GmbH, 2021. vol. 1375 AIST, p. 565–579. https://doi.org/10.1007/978-3-030-73050-5_56.

VAZAN, Pavel; JANIKOVA, D.; TANUSKA, Pavol; KEBISEK, Michal; ČERVEŇANSKÁ, Zuzana. Using data mining methods for manufacturing process control. **IFAC-PapersOnLine**, vol. 50, no. 1, p. 6178–6183, Jul. 2017. https://doi.org/10.1016/j.ifacol.2017.08.986.

VELTE, A; VELTE, T. **Microsoft virtualization with Hyper-V**. 2009. Available at: https://dl.acm.org/doi/abs/10.5555/1593830. Accessed on: 27 Feb. 2023.

VERNADAT, François. Enterprise Modeling in the context of Enterprise Engineering: State of the art and outlook. **International Journal of Production Management and Engineering**, vol. 2, no. 2, p. 57, 9 Jul. 2014. https://doi.org/10.4995/ijpme.2014.2326.

WANG, Ying; CARON, Filip; VANTHIENEN, Jan; HUANG, Lei; GUO, Yi. Acquiring logistics process intelligence: Methodology and an application for a Chinese bulk port. **Expert Systems with Applications**, vol. 41, no. 1, p. 195–209, Jan. 2014. https://doi.org/10.1016/j.eswa.2013.07.021.

WYNN, Moe Thandar; POPPE, Erik; XU, J.; HOFSTEDE, A.H.M. ter; BROWN, R.; PINI, Azzurra; VAN DER AALST, Wil. ProcessProfiler3D: A visualisation framework for log-based process performance comparison. **Decision Support Systems**, vol. 100, p. 93–108, Aug. 2017. https://doi.org/10.1016/j.dss.2017.04.004.

XAVIER, Miguel G.; NEVES, Marcelo v.; ROSSI, Fabio. D.; FERRETO, Tiago C.; LANGE, Timoteo; DE ROSE, Cesar A. F. Performance Evaluation of Container-Based Virtualization for High Performance Computing Environments. Feb. 2013. **2013 21st**

Euromicro International Conference on Parallel, Distributed, and Network-Based Processing [...]. [*S. l.*]: IEEE, Feb. 2013. p. 233–240. https://doi.org/10.1109/PDP.2013.41.

YUVAMITRA, Korakot; LEE, Jim; DONG, Kanjicai. Value Stream Mapping of Rope Manufacturing: A Case Study. **International Journal of Manufacturing Engineering**, vol. 2017, p. 1–11, 16 Feb. 2017. https://doi.org/10.1155/2017/8674187.

ZENG, Yan; YIN, Yuehong. Virtual and Physical Systems Intra-referenced Modelling for Smart Factory. **Procedia CIRP**, vol. 63, p. 378–383, 2017. https://doi.org/10.1016/j.procir.2017.03.105.

ZEROUALI, Ahmed; MENS, Tom; ROBLES, Gregorio; GONZALEZ-BARAHONA, Jesus M. On the Relation between Outdated Docker Containers, Severity Vulnerabilities, and Bugs. Feb. 2019. **2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER)** [...]. [*S. l.*]: IEEE, Feb. 2019. p. 491–501. https://doi.org/10.1109/SANER.2019.8668013.

ZHANG, Arthur. **Data Analytics: Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life**. 1st ed. Scotts Valley: CreateSpace , 2017.

ZHOU, Jianlong; GANDOMI, Amir H.; CHEN, Fang; HOLZINGER, Andreas. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. **Electronics**, vol. 10, no. 5, p. 593, 4 Mar. 2021. https://doi.org/10.3390/electronics10050593.

APPENDIX A – Preliminary Evaluation Questionnaire for Current and Future Value Stream Mapping Application

PRELIMINARY EVALUATION QUESTIONNAIRE FOR CURRENT AND FUTURE VALUE STREAM MAPPING APPLICATION

1- *Consistency:* How do you evaluate the consistency of the VSM application for Manufacturing Companies? Is the application consistent in relation to the workflow?

( ) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions:

2- *Usability*: How do you evaluate the usability of the results presented by the VSM application? Is the application interface easy to use and understand?
( ) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions:

3- *Accuracy:* How do you evaluate the accuracy of the results presented by the VSM application for Manufacturing Companies?
( ) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions:

4- *Reliability:* How do you evaluate the reliability of the results presented by the VSM application? Do you trust the results presented by the application?
( ) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions:

5- *Performance:* How do you evaluate the performance of the VSM application for Manufacturing Companies? Is the application fast and responsive?
( ) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions:

6- *Flexibility:* How do you evaluate the flexibility of the activities evaluated by the VSM application? Does the application allow for customization and adaptation to the needs of the company?

( ) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions:

APPENDIX B – Preliminary Evaluation Questionnaire answered by the Partner Company

PRELIMINARY EVALUATION QUESTIONNAIRE ANSWERED BY THE PARTNER COMPANY

1- *Consistency:* How do you evaluate the consistency of the VSM application for Manufacturing Companies? Is the application consistent in relation to the workflow?

(X) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions: It is a good idea to use this methodology to identify the real problems and impacts generated inside the factory.

2- *Usability*: How do you evaluate the usability of the results presented by the VSM application? Is the application interface easy to use and understand?
(x) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions:

3- *Accuracy:* How do you evaluate the accuracy of the results presented by the VSM application for Manufacturing Companies?
( ) Very satisfied
(X) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions: I believe the next step is to use real data to get better understanding of the outcomes.

4- *Reliability:* How do you evaluate the reliability of the results presented by the VSM application? Do you trust the results presented by the application?
(X) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions: Yes, the results are grounded in statistics concept.

5- *Performance:* How do you evaluate the performance of the VSM application for Manufacturing Companies? Is the application fast and responsive?
(X) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied
Comments and suggestions: The application is fast and can provide different scenarios according to the inputs.

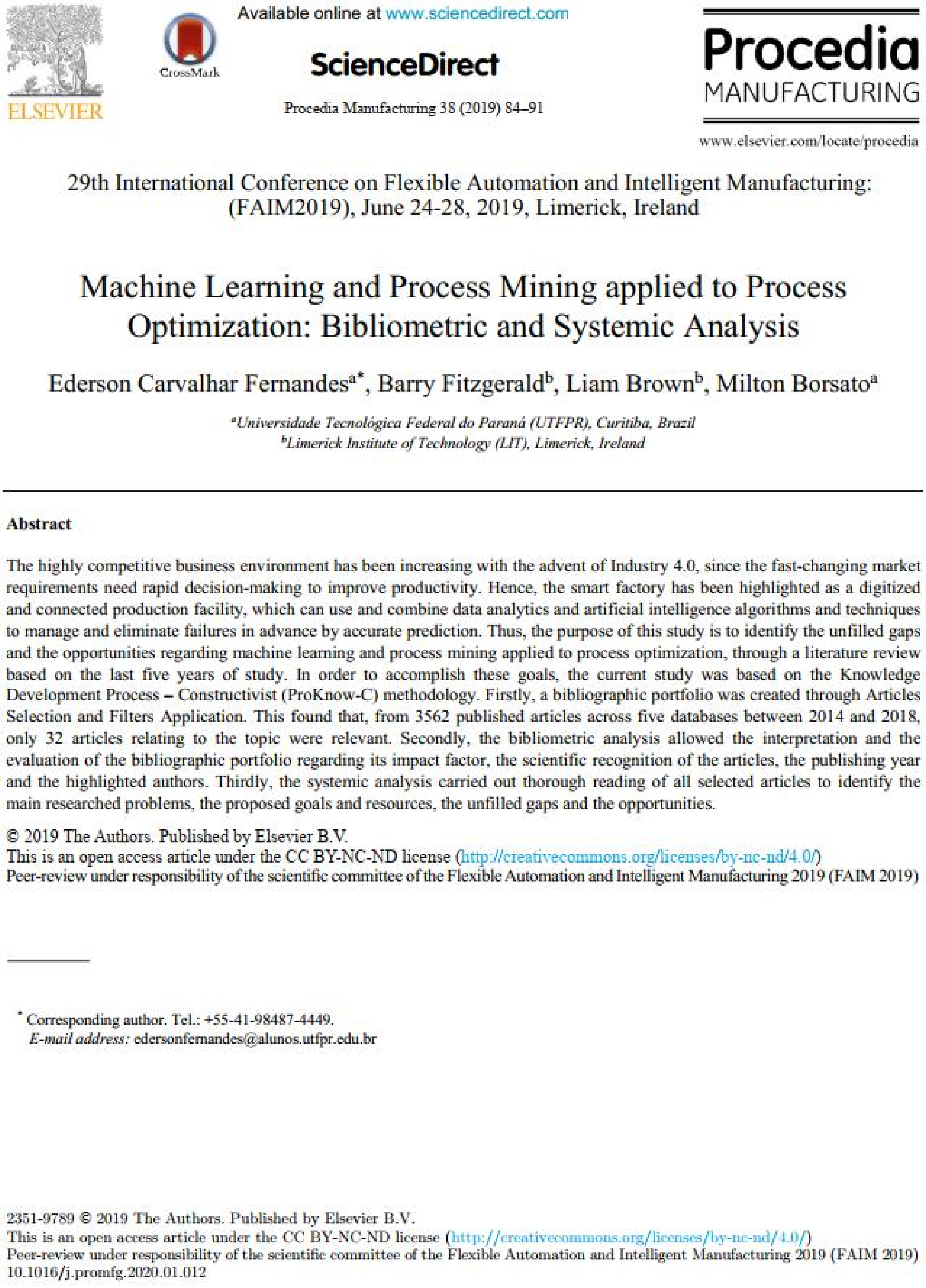
6- *Flexibility:* How do you evaluate the flexibility of the activities evaluated by the VSM application? Does the application allow for customization and adaptation to the needs of the company?

(X) Very satisfied
( ) Satisfied
( ) Improvements are necessary
( ) Dissatisfied

Comments and suggestions: The application is very flexible which is possible to add or remove steps of analysis according to the company reality.

APPENDIX C – Machine Learning and Process Mining applied to Process Optimization: Bibliometric and Systemic Analysis (Published Paper)

MACHINE LEARNING AND PROCESS MINING APPLIED TO PROCESS OPTIMIZATION: BIBLIOMETRIC AND SYSTEMIC ANALYSIS (PUBLISHED PAPER)



Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 38 (2019) 84–91

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

29th International Conference on Flexible Automation and Intelligent Manufacturing: (FAIM2019), June 24-28, 2019, Limerick, Ireland

Machine Learning and Process Mining applied to Process Optimization: Bibliometric and Systemic Analysis

Ederson Carvalhar Fernandes[a*], Barry Fitzgerald[b], Liam Brown[b], Milton Borsato[a]

[a]*Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Brazil*
[b]*Limerick Institute of Technology (LIT), Limerick, Ireland*

Abstract

The highly competitive business environment has been increasing with the advent of Industry 4.0, since the fast-changing market requirements need rapid decision-making to improve productivity. Hence, the smart factory has been highlighted as a digitized and connected production facility, which can use and combine data analytics and artificial intelligence algorithms and techniques to manage and eliminate failures in advance by accurate prediction. Thus, the purpose of this study is to identify the unfilled gaps and the opportunities regarding machine learning and process mining applied to process optimization, through a literature review based on the last five years of study. In order to accomplish these goals, the current study was based on the Knowledge Development Process – Constructivist (ProKnow-C) methodology. Firstly, a bibliographic portfolio was created through Articles Selection and Filters Application. This found that, from 3562 published articles across five databases between 2014 and 2018, only 32 articles relating to the topic were relevant. Secondly, the bibliometric analysis allowed the interpretation and the evaluation of the bibliographic portfolio regarding its impact factor, the scientific recognition of the articles, the publishing year and the highlighted authors. Thirdly, the systemic analysis carried out thorough reading of all selected articles to identify the main researched problems, the proposed goals and resources, the unfilled gaps and the opportunities.

[*] Corresponding author. Tel.: +55-41-98487-4449.
E-mail address: edersonfernandes@alunos.utfpr.edu.br

10.1016/j.promfg.2020.01.012

Keywords: Machine Learning; Process Mining; Process Optimization; Bibliometric Analysis; Systemic Analysis.

1. Introduction

Industrial competitiveness challenges have been increasing with the advent of the Industry 4.0. The fourth industrial revolution puts great emphasis on the smart factory, which encompasses the development of intelligent production and manufacturing processes, the new capabilities through communication between the physical and virtual environment, and the interoperability among computer systems [1].

Process management is an important component of the smart factory and has been deployed in several ways to help productivity improvement. Unforeseen events can cause deviations from procedures, resulting in excessive downtime, overproduction, defects and inventory issues. In this way, process management methods can provide valuable insights for organizations, such as the identification of bottlenecks and time lags [2].

Process optimization depends on the field engineer's knowledge and expertise. Yet, many companies still have been managing processes for one single process or workflow at a time, comparing results through regular meetings, mappings and manual documents [3]. However, the advent of Artificial Intelligence (AI) has enabled computer science to invent efficient and new solutions to human problem-solving tasks [4]. One of the main branches of AI, that gives computers the ability to learn without being explicitly programmed, is called Machine Learning (ML).

Along with ML, the use of Process Mining (PM) has increasingly grown in the manufacturing industry. [5] declare that PM is a family of techniques used to discover and improve real business processes by extracting knowledge from event logs available in process-aware information systems. Event logs are basic resources that help provide information about network traffic, usage and other conditions.

The focus of this paper, therefore, is to present a literature review for the methods of ML and PM applied to process optimization. The main goal of this research is the identification of the trends and proposed solutions from the last five years relating to the different approaches and strategies developed for process optimization, as well as finding unfilled gaps.

The sequence and explanations of how the activities carried out will be shown more detailed in section 3. In section 2 the methodological aspects will be shown. Section 4 includes the final considerations.

2. Methodological Aspects

According to [6], the research classification can be goals-based and technical procedures-based. The goals-based classification contains three main groups: exploratory, descriptive and explanatory. Technical procedures-based classification splits the research into the following categories: bibliographic, documentary, experimental, ex-post facto, survey and case-study. This study was classified as exploratory and bibliographic, respectively.

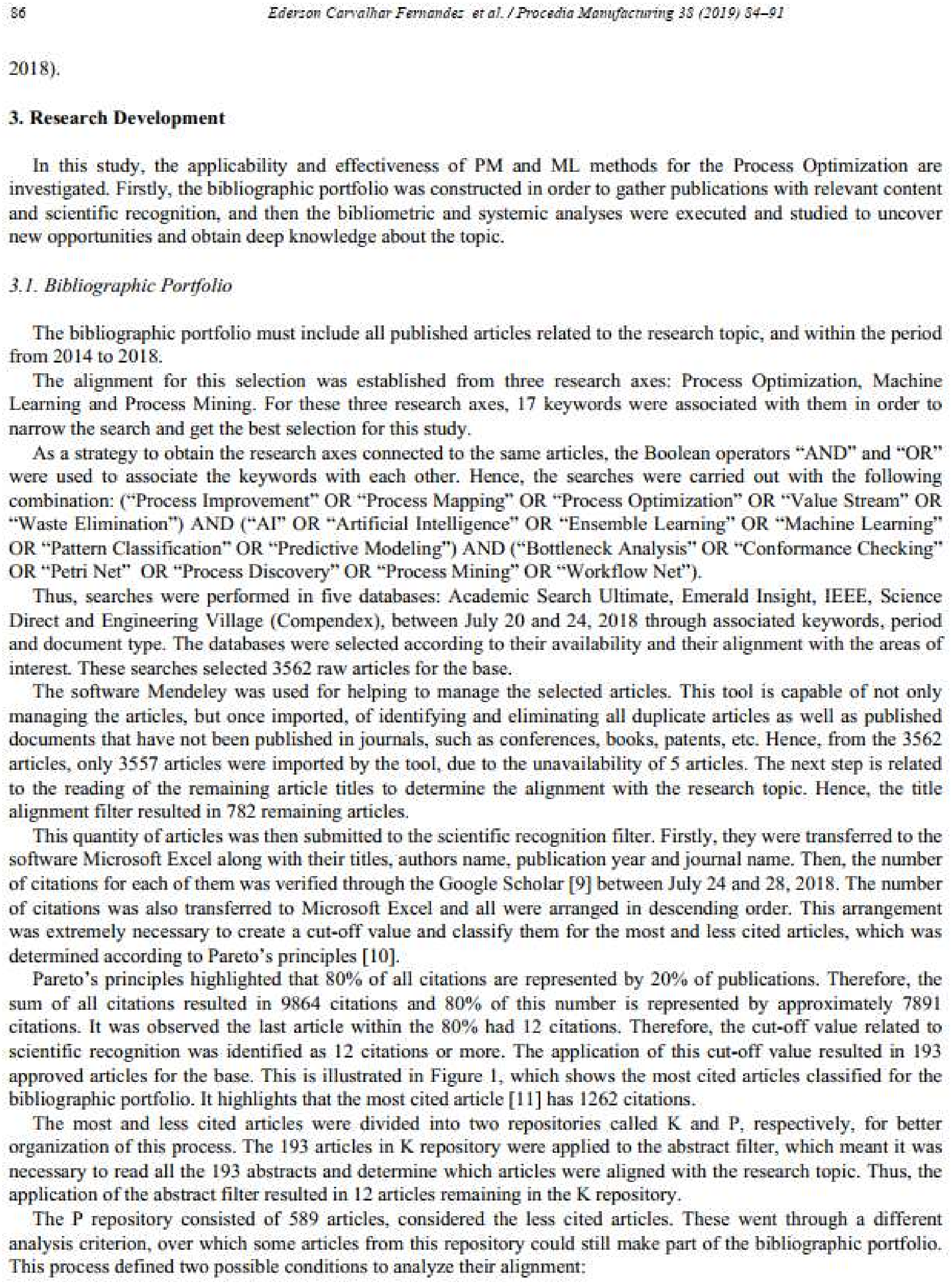
Exploratory research is research conducted for a problem that has not been studied more clearly and helps to determine the best research design, selection of subjects and data collection. Bibliographic research is based on published material, basically constituting books and scientific articles.

The goal of the presented work was obtained using the Knowledge Development Process – Constructivist (ProKnow-C). This methodology was proposed by [7] and [8] and developed to help the researchers to find the most relevant content for the research in an easier way than they were used to.

This methodology is divided into five stages: *(i)* article selection that constitutes the bibliographic portfolio, *(ii)* bibliometric analysis, *(iii)* systemic analysis, *(iv)* research question definition, and *(v)* research goal definition.

The current study was conducted only with the first three stages from the Proknow-C since the goal of this study is the identification of the trends and opportunities relating to the research. Mendeley and Microsoft Excel software were applied to this methodology for helping with the bibliographic management and data tabulation.

While executing this methodology two restrictions were applied to this process. Firstly, it was decided to check only articles from journals, and secondly, a 5-year period was selected to limit the research (i.e., between 2014 and



2018).

3. Research Development

In this study, the applicability and effectiveness of PM and ML methods for the Process Optimization are investigated. Firstly, the bibliographic portfolio was constructed in order to gather publications with relevant content and scientific recognition, and then the bibliometric and systemic analyses were executed and studied to uncover new opportunities and obtain deep knowledge about the topic.

3.1. Bibliographic Portfolio

The bibliographic portfolio must include all published articles related to the research topic, and within the period from 2014 to 2018.

The alignment for this selection was established from three research axes: Process Optimization, Machine Learning and Process Mining. For these three research axes, 17 keywords were associated with them in order to narrow the search and get the best selection for this study.

As a strategy to obtain the research axes connected to the same articles, the Boolean operators "AND" and "OR" were used to associate the keywords with each other. Hence, the searches were carried out with the following combination: ("Process Improvement" OR "Process Mapping" OR "Process Optimization" OR "Value Stream" OR "Waste Elimination") AND ("AI" OR "Artificial Intelligence" OR "Ensemble Learning" OR "Machine Learning" OR "Pattern Classification" OR "Predictive Modeling") AND ("Bottleneck Analysis" OR "Conformance Checking" OR "Petri Net" OR "Process Discovery" OR "Process Mining" OR "Workflow Net").

Thus, searches were performed in five databases: Academic Search Ultimate, Emerald Insight, IEEE, Science Direct and Engineering Village (Compendex), between July 20 and 24, 2018 through associated keywords, period and document type. The databases were selected according to their availability and their alignment with the areas of interest. These searches selected 3562 raw articles for the base.

The software Mendeley was used for helping to manage the selected articles. This tool is capable of not only managing the articles, but once imported, of identifying and eliminating all duplicate articles as well as published documents that have not been published in journals, such as conferences, books, patents, etc. Hence, from the 3562 articles, only 3557 articles were imported by the tool, due to the unavailability of 5 articles. The next step is related to the reading of the remaining article titles to determine the alignment with the research topic. Hence, the title alignment filter resulted in 782 remaining articles.

This quantity of articles was then submitted to the scientific recognition filter. Firstly, they were transferred to the software Microsoft Excel along with their titles, authors name, publication year and journal name. Then, the number of citations for each of them was verified through the Google Scholar [9] between July 24 and 28, 2018. The number of citations was also transferred to Microsoft Excel and all were arranged in descending order. This arrangement was extremely necessary to create a cut-off value and classify them for the most and less cited articles, which was determined according to Pareto's principles [10].

Pareto's principles highlighted that 80% of all citations are represented by 20% of publications. Therefore, the sum of all citations resulted in 9864 citations and 80% of this number is represented by approximately 7891 citations. It was observed the last article within the 80% had 12 citations. Therefore, the cut-off value related to scientific recognition was identified as 12 citations or more. The application of this cut-off value resulted in 193 approved articles for the base. This is illustrated in Figure 1, which shows the most cited articles classified for the bibliographic portfolio. It highlights that the most cited article [11] has 1262 citations.

The most and less cited articles were divided into two repositories called K and P, respectively, for better organization of this process. The 193 articles in K repository were applied to the abstract filter, which meant it was necessary to read all the 193 abstracts and determine which articles were aligned with the research topic. Thus, the application of the abstract filter resulted in 12 articles remaining in the K repository.

The P repository consisted of 589 articles, considered the less cited articles. These went through a different analysis criterion, over which some articles from this repository could still make part of the bibliographic portfolio. This process defined two possible conditions to analyze their alignment:

- •A) Published articles from the last 2 years of this analysis (i.e., 2017 and 2018) mainly in order to not eliminate good works which could be relevant, but haven't had the proper acknowledgement because they were published too recently;
- •B) Published articles more than 2 years from this analysis (i.e., 2014 to 2016) for which their authors are the same authors of the articles belonging to K repository.

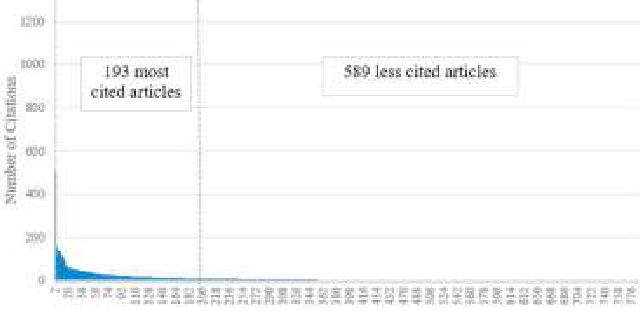


Fig. 1. Cut-off Value according to Citations.

After the condition's application, the found articles will only be approved after the abstract filter application.

Thus, 305 articles from the P repository were analyzed according to the A condition, since the publishing year was between 2017 and 2018. As with the most cited articles that went through the abstract filter, the same occurred with these articles. Hence, the application of the abstract filter resulted in 20 approved remaining articles for the A condition in P repository.

Before the analysis of the 284 articles belonging to the 2014 to 2016 publishing years, it was necessary to define which authors from the K repository would be considered at this stage. So, from the 12 articles selected in K repository, an author base was created with 57 authors. Using these 57 authors as a parameter to select new articles based on the B condition, 7 articles with the same authorship were found. However, none of these have been approved after the abstract filter application.

In conclusion for the Filters Application, 12 articles remained in the K repository and 20 in the P repository. Thus, the bibliographic portfolio ended with 32 articles for the bibliometric analysis.

3.2 Bibliometric Analysis

The bibliometric analysis allows the interpretation and the evaluation of the bibliographic portfolio. It establishes the journals' relevance regarding its impact factor, the scientific recognition of the articles, the publishing year, the highlighted authors and keywords.

For the 32 selected papers in the bibliographic portfolio, the most cited article [12] has 36 citations. These articles encompass 57 authors, and four of these authors have more than one publication. These authors are Juan-Ignacio Latorre-Biel, Emilio Jimenez-Macias, Mengshu Zhou and Hernan Ponce de Leon.

The analysis results of the Journals Relevance show that the journals: Expert Systems with Applications, and Procedia CIRP, are the most relevant journals for the research topic, with each one having three articles included in the bibliographic portfolio. Also, it is concluded that the two most relevant journals are not considered the best according to the scientific journal rankings (SJR), which are the science evaluation resources to assess worldwide universities and research-focused institutions.

An increase in the publishing rates in the last five years relating to Machine Learning and Process Mining applied to Process Optimization was detect through analysis of the publishing trends per year relating to the research topic. This is evidence that this topic has been well highlighted in academia.

Since the bibliographic portfolio was completed and the bibliometric analysis was concluded, the systemic analysis could be started as the final step of the methodology.

3.3 Systemic Analysis

The systemic analysis focused on analysing the articles' content from the bibliographic portfolio. This step works as a representativity test through a detailed evaluation to analyze their adherence to the research topic alignment and to look for knowledge opportunities for new and relevant developments.

The systemic analysis process was mainly structured according to [13]. All the articles were searched in their full version and this step was completely successful because all 32 articles from the bibliographic portfolio were available. During the full readings, unfortunately, three articles were eliminated since they were considered misaligned with the research topic. Thus, each remained article was fully read in order to address: *(i)* problem area, *(ii)* goal, *(iii)* methodology, *(iv)* main results, *(v)* future recommendations, *(vi)* research opportunities, and *(vii)* existing gaps.

3.3.1. Identified Research Problems

After completely reading the articles, the problems raised by the authors could be identified. The majority are described as follows.

(i)Process Management: Problems of process management and resource allocation [1] have been growing and creating several hardships for the companies that do not invest in the development of intelligent production and manufacturing process [14]. Nowadays most organizations store their data in one or more database management systems, but they do not have the capability to analyze and gain useful insights from this data [15]. The absence of adequate process planning, or the development of insufficient data mapping, can cause waste of time and costly resources [16]–[19].

(ii)Complex models: A few systems, methods and methodologies have already been used for process optimization, but due to the technology evolution, these strategies have become less effective for factories [12]. [20] described that up-to-date models and systems are essential for production control but challenging to maintain. Continuous improvement is very important [21], since companies may reach new levels of competitiveness, reliability and accuracy [22], [23]. The use of a new system or even the combination of more systems may overcome many weaknesses from past studies [24], [25].

(iii)Decision-making support: Decision strategies in dynamic environments do not always get either desired outcomes or optimizations, and this has been a huge problem outlined by several authors [2], [3], [26], [27]. Deviations from the standard operating procedures, uncertainties, and delays [28] are examples of unforeseen events on current and future states without proper supervision. Logistics [29], [30], supply chain [31] and distributed business processes [2], [32]–[36] were identified as applications of decision support by associated authors, since the execution of real processes can be highly unpredictable and vulnerable to disruptive events [15], mainly due to either lack of maintenance [36] or human error [37].

The next section presents the proposed goals defined by the authors, encompassing these problems, as well as resource used for reaching them.

3.3.2 Proposed Goals and Resources

Resources are either actions or strategies which may be adopted by several applications for getting better outcomes. All the resources, such as tools, methods, approaches, systems and frameworks will be presented along with the proposed goals.

(i)Process Management and Modelling: AI methodologies have been constantly applied to improve strategies and eliminate failures. [1], [14], [16], [20] used the modelling language Petri Net (PN) for modelling production systems in order to support analysis and availability optimization as well as supporting resources and performance evaluation. [14] built three PN models and simulated them for modelling interactions and self-organization.

[17] used PM techniques for modelling the production planning process of a manufacturing company. They discovered the process model with both heuristic miner and analysis.

[18] extracted anomalous frequent patterns from historical logging data. Consequently, it was possible to get more accurate analysis to exhibit parallel behaviors and correlate recurrent deviations which occurred in different portions of the process.

[19] used an PN model for both path planning and collision avoidance of a machine system. [30] took advantage of a methodological approach which can be used by managers for easily mapping, visualizing unforeseen and undesired events, supply chain and understanding potential weaknesses. Similarly [28] used PM and data mining techniques for investigating and identifying possible factors that cause delay in the production system.

[29] and [30] developed comprehensible methodologies for applying both PM and ML to add context awareness for unstructured event data in logistics (i.e., the execution of exploratory, performance and conformance analysis to get the potential to improve the results).

[33] used PN to solve complex problems in an efficient way (i.e., integrating different decisions into the same optimization problem).

(ii)New approaches, frameworks and combinations: [15] presented an approach for discovering probabilistic belief networks from practical real-life application event logs.

[12] used a hybrid PN by embedding a neural network algorithm which can model a runtime environment and collaborate in making adaptation decisions while the system is running.

Similar to these proposals, [24] integrated a five-stage framework of data mining, process improvement and process ontology for managing and improving processes with high volume of data, and [25] designed and implemented a PN based Generic Genetic Algorithm (GGA) framework which can be used for optimizing any given business processes modeled in Color Petri Nets (CPN) and exploiting simulation outputs.

[22] incorporated negative information in process discovery of complex systems for deriving less complex, fitting and precise process models, as well as being very good of generalizing the right behavior for an event log. [23] modelled a new architecture using production flow schemas (PFS) and their dynamic behaviors which are validated by PN models. This system was used to improve the communication between machines and products in a modular production system.

[3] filled a gap for the lack of tools which enable an intuitive and direct comparison of multiple management cohorts (i.e., a coherent group of process instances with one or more shared characteristics). The framework supports the cohort's selection and a tridimensional visualization for comparing their performance metrics. Among their resources, there is a set of plug-ins which includes a PM open source framework called PROM with two real-life datasets.

Due to the limited time required to obtain a solution, either the manufacturing control or decision-making problems may require unaffordable computations resources. Thus, [27] proposed an efficiency improvement throughout simulation-based optimization, PN models and genetic algorithm metaheuristics.

(iii)Predictive systems: [34] used a combination of ML techniques and PM features for predicting performance, and [35] applied both deep learning and recurrent neural networks for predicting future events in a business process.

[36] used a new system for tracking maintenance production along with a domain maintenance ontology. [2] used recommendation systems for incorporating factorization machines in event prediction tasks. [38] created a new predictive modelling technique based on both previous weaker biases, PM and grammatical inference which accurately predicts and provides comprehensible results. [26], [32] used a methodology based on ML techniques which may detect errors and predict failures with high accuracy. [37] used PM techniques to estimate human error probabilities when required tasks are conducted. For the presented solutions and results, a few opportunities were identified, which will be discussed in the next section.

3.3.3. Unfilled Gaps Identification

For the raised problems and their solutions, a few unfilled gaps could be identified.

(i)Additional Features and Programming: There are many improvement opportunities from proposed solutions which the researchers could not completely implement, and suggested should be the focus of additional research. [26] declared big potential for their ML approach improvement, PROCEDO, along with data mining. Similarly, [2], [3], [20], [23] and [31] suggested additional features for reaching better performance, response time, service availability and richer understanding of their process vulnerabilities. PM applications for logistician intelligence

support are considered to have several opportunities for enhancing logistics process transparency, strengthening the internal control of logistics firms and improving performance [29], [31].

(ii)Smart method adaptation: Several combinations of ML workflows, PM algorithms and PN language improvement may have potential for full execution of the activities [17], [21], [25], [27]. One of the most interesting developments of smart systems has been applied for logistics and supply chain [29]–[31], and this has opened possibilities for new improved manufacturing methods (e.g., Lean Manufacturing Method and Total Quality Management).

(iii)Environmental Adaptability: Smart systems applications which have had good outcomes in one specific environment may be advantageous in other environments [1]. Methodologies such as [28] and [34], through suitable adaptations may be tried out in different environments and processes for getting new accurate outcomes.

(iv)Smart Validation Dataset: Difficulties and limitations could be identified regarding the method validation process (i.e., in real applications there was not enough collected data for accurately simulating and validating them). Also, the study could be subject to insufficient conclusions since relevant knowledge and information may not have been considered. [14], [34], [35]. There are other developments which have obtained great results, however they still demonstrated dependency on manual experts' intervention in either validating or making final decisions [15], [36].

Therefore, the study conducted highlights the scope for continued research and further improvement based on the unfilled gaps.

4. Conclusions

The study has aimed a structured methodology called ProKnow-C in a bibliographic and systemic analysis of Machine learning and process mining applied to process optimization.

Firstly, the bibliographic portfolio was created, whereby from 3562 published articles found in five databases between 2014 and 2018, only 32 articles relating to the topic were approved.

Secondly, the bibliometric analysis allowed the interpretation of the bibliographic portfolio and the evaluation, through comparative charts, of the journals' relevance regarding its impact factor, scientific recognition of the articles, publishing year, highlighted authors and keywords.

Thirdly, the systemic analysis carried out thorough reading of all selected articles in order to identify the main research problems, proposed goals and resources, and the unfilled gaps. The research opportunities identified were: *(i)* smart method adaptation; *(ii)* environmental adaptability; *(iii)* additional features and programming; and *(iv)* smart validation dataset.

ProKnow-C methodology has been very effective for the understanding of current problems, goals, gaps and opportunities in several areas. However, it demands a lot of time and dedication from the researcher, due to the huge number of articles found in the databases and the application of filters, which caused many challenges during the analyses. In future studies deeper research is recommended into references from each article in this bibliographic portfolio. This will ensure more relevant articles and identify not only the recent authors but the ones who coined the concepts relating to this research topic.

References

[1] Y. Zeng and Y. Yin, Virtual and Physical Systems Intra-referenced Modelling for Smart Factory, Procedia CIRP, vol. 63, pp. 378–383, 2017.

[2] W. L. J. Lee, D. Parra, J. Munoz-Gama, and M. Sepúlveda, Predicting Process Behavior Meets Factorization Machines, Expert Syst. Appl., 2018.

[3] M. T. Wynn *et al.*, ProcessProfiler3D: A visualisation framework for log-based process performance comparison, Decis. Support Syst., vol. 100, pp. 93–108, 2017.

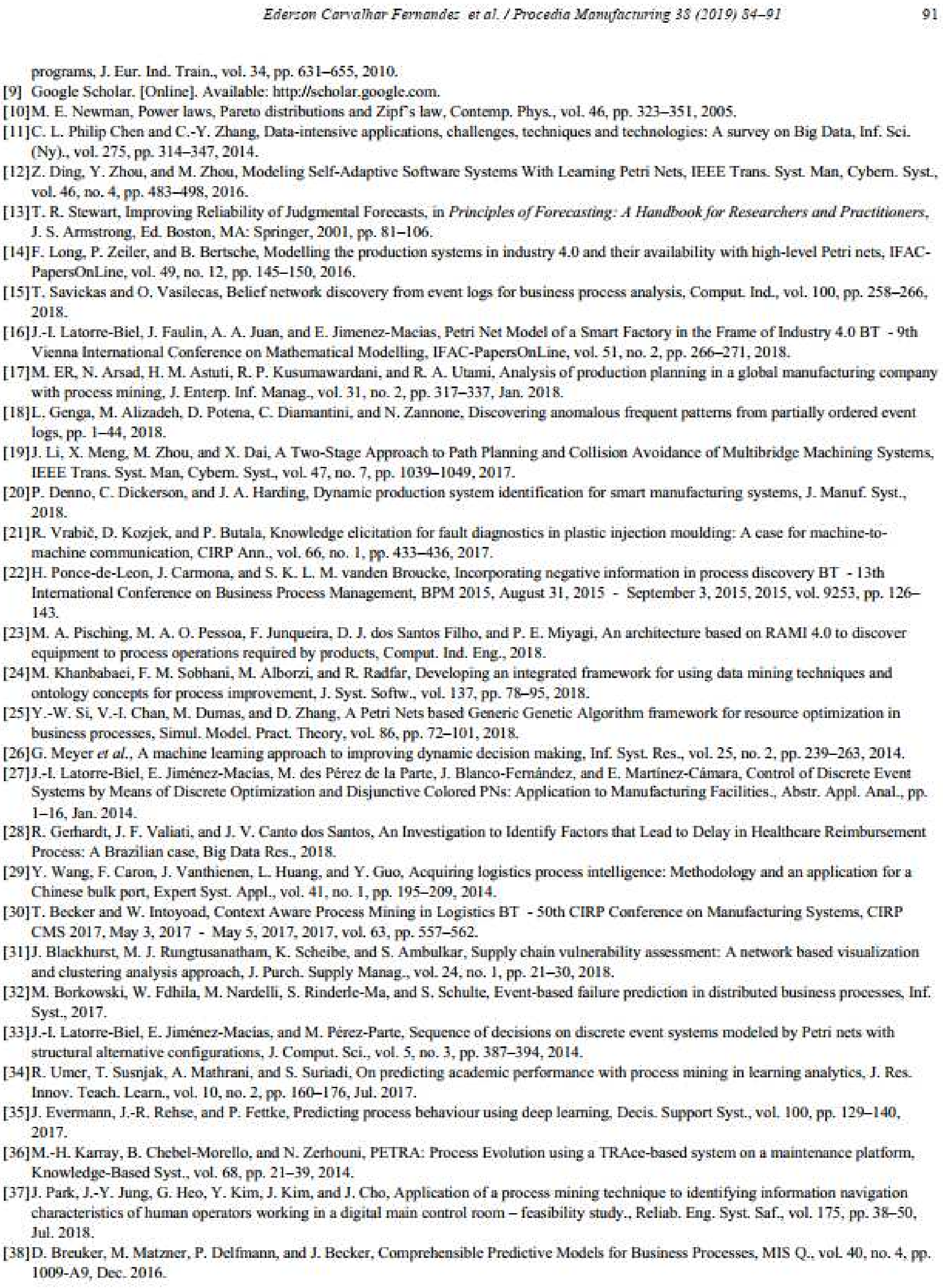
[4] D. Paschek, C. T. Luminosu, and A. Draghici, Automated business process management – in times of digital transformation using machine learning or artificial intelligence, MATEC Web Conf., vol. 121, p. 04007, 2017.

[5] A. R. C. Maita, L. C. Martins, C. R. López Paz, M. Fantinato, and S. M. Peres, Process mining through artificial neural networks and support vector machines: A systematic literature review, Bus. Process Manag. J., vol. 21, no. 6, pp. 1391–1415, Oct. 2015.

[6] A. C. Gil, Como Elaborar Projetos de Pesquisa, 3rd ed. Belo Horizonte: Atlas, 1996.

[7] L. Ensslin, O Design Na Pesquisa Quali-Quantitativa Em Engenharia De Produção – Questões Epistemológicas the Design in the Quali-Quantitative Research in the Production Engineering – Epistemological Issues, Rev. Produçao line, vol. 8, no. 48, p. 16, 2008.

[8] M. B. M. A. J. Eduardo Tasca, L. Ensslin, S. Rolim Ensslin, An approach for selecting a theoretical framework for the evaluation of training



programs, J. Eur. Ind. Train., vol. 34, pp. 631–655, 2010.

[9] Google Scholar. [Online]. Available: http://scholar.google.com.

[10] M. E. Newman, Power laws, Pareto distributions and Zipf's law, Contemp. Phys., vol. 46, pp. 323–351, 2005.

[11] C. L. Philip Chen and C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Inf. Sci. (Ny)., vol. 275, pp. 314–347, 2014.

[12] Z. Ding, Y. Zhou, and M. Zhou, Modeling Self-Adaptive Software Systems With Learning Petri Nets, IEEE Trans. Syst. Man, Cybern. Syst., vol. 46, no. 4, pp. 483–498, 2016.

[13] T. R. Stewart, Improving Reliability of Judgmental Forecasts, in *Principles of Forecasting: A Handbook for Researchers and Practitioners*, J. S. Armstrong, Ed. Boston, MA: Springer, 2001, pp. 81–106.

[14] F. Long, P. Zeiler, and B. Bertsche, Modelling the production systems in industry 4.0 and their availability with high-level Petri nets, IFAC-PapersOnLine, vol. 49, no. 12, pp. 145–150, 2016.

[15] T. Savickas and O. Vasilecas, Belief network discovery from event logs for business process analysis, Comput. Ind., vol. 100, pp. 258–266, 2018.

[16] J.-I. Latorre-Biel, J. Faulin, A. A. Juan, and E. Jimenez-Macias, Petri Net Model of a Smart Factory in the Frame of Industry 4.0 BT - 9th Vienna International Conference on Mathematical Modelling, IFAC-PapersOnLine, vol. 51, no. 2, pp. 266–271, 2018.

[17] M. ER, N. Arsad, H. M. Astuti, R. P. Kusumawardani, and R. A. Utami, Analysis of production planning in a global manufacturing company with process mining, J. Enterp. Inf. Manag., vol. 31, no. 2, pp. 317–337, Jan. 2018.

[18] L. Genga, M. Alizadeh, D. Potena, C. Diamantini, and N. Zannone, Discovering anomalous frequent patterns from partially ordered event logs, pp. 1–44, 2018.

[19] J. Li, X. Meng, M. Zhou, and X. Dai, A Two-Stage Approach to Path Planning and Collision Avoidance of Multibridge Machining Systems, IEEE Trans. Syst. Man, Cybern. Syst., vol. 47, no. 7, pp. 1039–1049, 2017.

[20] P. Denno, C. Dickerson, and J. A. Harding, Dynamic production system identification for smart manufacturing systems, J. Manuf. Syst., 2018.

[21] R. Vrabič, D. Kozjek, and P. Butala, Knowledge elicitation for fault diagnostics in plastic injection moulding: A case for machine-to-machine communication, CIRP Ann., vol. 66, no. 1, pp. 433–436, 2017.

[22] H. Ponce-de-Leon, J. Carmona, and S. K. L. M. vanden Broucke, Incorporating negative information in process discovery BT - 13th International Conference on Business Process Management, BPM 2015, August 31, 2015 - September 3, 2015, 2015, vol. 9253, pp. 126–143.

[23] M. A. Pisching, M. A. O. Pessoa, F. Junqueira, D. J. dos Santos Filho, and P. E. Miyagi, An architecture based on RAMI 4.0 to discover equipment to process operations required by products, Comput. Ind. Eng., 2018.

[24] M. Khanbabaei, F. M. Sobhani, M. Alborzi, and R. Radfar, Developing an integrated framework for using data mining techniques and ontology concepts for process improvement, J. Syst. Softw., vol. 137, pp. 78–95, 2018.

[25] Y.-W. Si, V.-I. Chan, M. Dumas, and D. Zhang, A Petri Nets based Generic Genetic Algorithm framework for resource optimization in business processes, Simul. Model. Pract. Theory, vol. 86, pp. 72–101, 2018.

[26] G. Meyer *et al.*, A machine learning approach to improving dynamic decision making, Inf. Syst. Res., vol. 25, no. 2, pp. 239–263, 2014.

[27] J.-I. Latorre-Biel, E. Jiménez-Macías, M. des Pérez de la Parte, J. Blanco-Fernández, and E. Martínez-Cámara, Control of Discrete Event Systems by Means of Discrete Optimization and Disjunctive Colored PNs: Application to Manufacturing Facilities., Abstr. Appl. Anal., pp. 1–16, Jan. 2014.

[28] R. Gerhardt, J. F. Valiati, and J. V. Canto dos Santos, An Investigation to Identify Factors that Lead to Delay in Healthcare Reimbursement Process: A Brazilian case, Big Data Res., 2018.

[29] Y. Wang, F. Caron, J. Vanthienen, L. Huang, and Y. Guo, Acquiring logistics process intelligence: Methodology and an application for a Chinese bulk port, Expert Syst. Appl., vol. 41, no. 1, pp. 195–209, 2014.

[30] T. Becker and W. Intoyoad, Context Aware Process Mining in Logistics BT - 50th CIRP Conference on Manufacturing Systems, CIRP CMS 2017, May 3, 2017 - May 5, 2017, 2017, vol. 63, pp. 557–562.

[31] J. Blackhurst, M. J. Rungtusanatham, K. Scheibe, and S. Ambulkar, Supply chain vulnerability assessment: A network based visualization and clustering analysis approach, J. Purch. Supply Manag., vol. 24, no. 1, pp. 21–30, 2018.

[32] M. Borkowski, W. Fdhila, M. Nardelli, S. Rinderle-Ma, and S. Schulte, Event-based failure prediction in distributed business processes, Inf. Syst., 2017.

[33] J.-I. Latorre-Biel, E. Jiménez-Macías, and M. Pérez-Parte, Sequence of decisions on discrete event systems modeled by Petri nets with structural alternative configurations, J. Comput. Sci., vol. 5, no. 3, pp. 387–394, 2014.

[34] R. Umer, T. Susnjak, A. Mathrani, and S. Suriadi, On predicting academic performance with process mining in learning analytics, J. Res. Innov. Teach. Learn., vol. 10, no. 2, pp. 160–176, Jul. 2017.

[35] J. Evermann, J.-R. Rehse, and P. Fettke, Predicting process behaviour using deep learning, Decis. Support Syst., vol. 100, pp. 129–140, 2017.

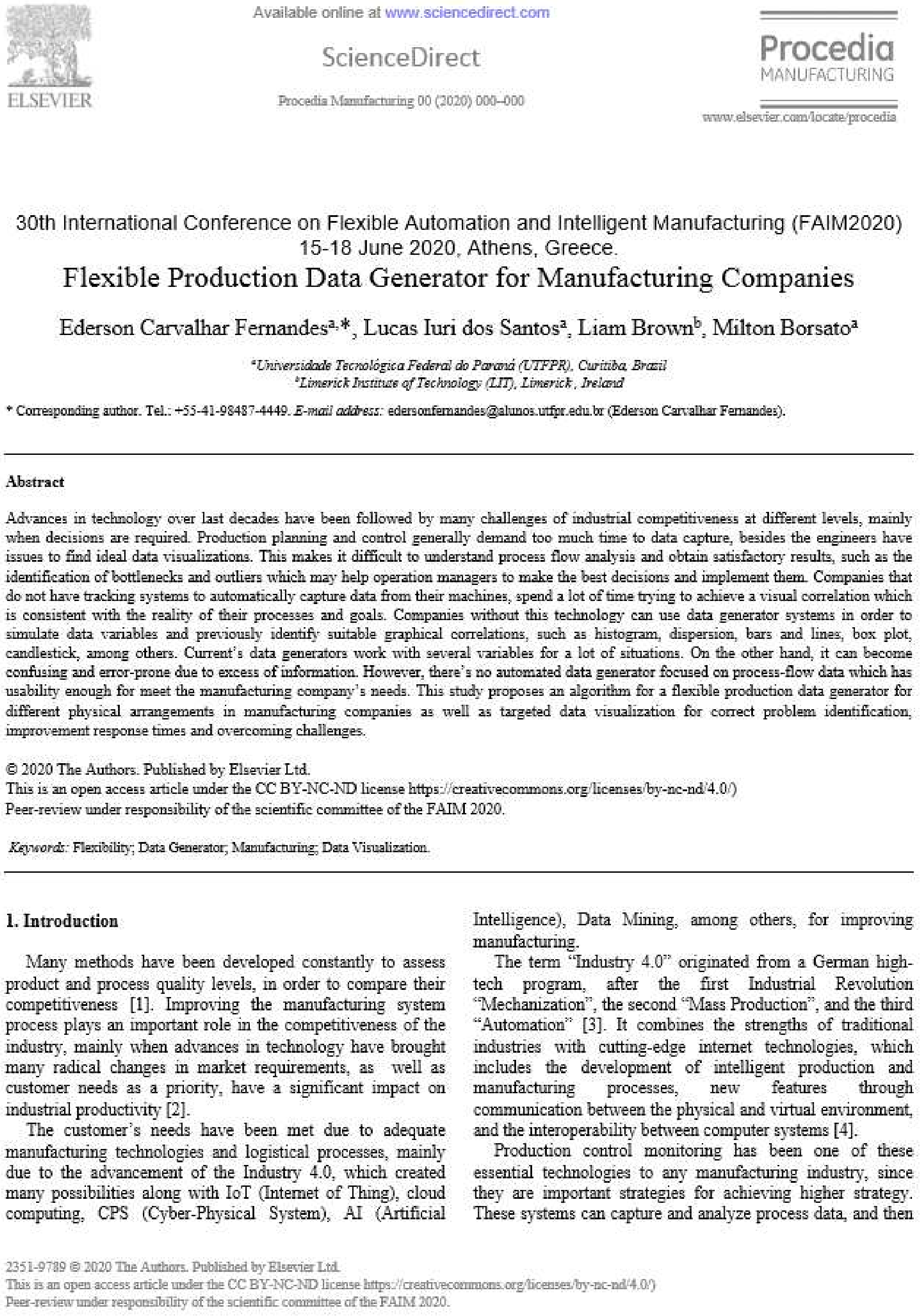
[36] M.-H. Karray, B. Chebel-Morello, and N. Zerhouni, PETRA: Process Evolution using a TRAce-based system on a maintenance platform, Knowledge-Based Syst., vol. 68, pp. 21–39, 2014.

[37] J. Park, J.-Y. Jung, G. Heo, Y. Kim, J. Kim, and J. Cho, Application of a process mining technique to identifying information navigation characteristics of human operators working in a digital main control room – feasibility study., Reliab. Eng. Syst. Saf., vol. 175, pp. 38–50, Jul. 2018.

[38] D. Breuker, M. Matzner, P. Delfmann, and J. Becker, Comprehensible Predictive Models for Business Processes, MIS Q., vol. 40, no. 4, pp. 1009-A9, Dec. 2016.

APPENDIX D – Flexible Production Data Generator for Manufacturing Companies (Published Paper)

FLEXIBLE PRODUCTION DATA GENERATOR FOR MANUFACTURING COMPANIES (PUBLISHED PAPER)



Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 00 (2020) 000–000

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

30th International Conference on Flexible Automation and Intelligent Manufacturing (FAIM2020)
15-18 June 2020, Athens, Greece.

Flexible Production Data Generator for Manufacturing Companies

Ederson Carvalhar Fernandes[a,*], Lucas Iuri dos Santos[a], Liam Brown[b], Milton Borsato[a]

[a]*Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Brasil*
[b]*Limerick Institute of Technology (LIT), Limerick, Ireland*

* Corresponding author. Tel.: +55-41-98487-4449. *E-mail address:* edersonfernandes@alunos.utfpr.edu.br (Ederson Carvalhar Fernandes).

Abstract

Advances in technology over last decades have been followed by many challenges of industrial competitiveness at different levels, mainly when decisions are required. Production planning and control generally demand too much time to data capture, besides the engineers have issues to find ideal data visualizations. This makes it difficult to understand process flow analysis and obtain satisfactory results, such as the identification of bottlenecks and outliers which may help operation managers to make the best decisions and implement them. Companies that do not have tracking systems to automatically capture data from their machines, spend a lot of time trying to achieve a visual correlation which is consistent with the reality of their processes and goals. Companies without this technology can use data generator systems in order to simulate data variables and previously identify suitable graphical correlations, such as histogram, dispersion, bars and lines, box plot, candlestick, among others. Current's data generators work with several variables for a lot of situations. On the other hand, it can become confusing and error-prone due to excess of information. However, there's no automated data generator focused on process-flow data which has usability enough for meet the manufacturing company's needs. This study proposes an algorithm for a flexible production data generator for different physical arrangements in manufacturing companies as well as targeted data visualization for correct problem identification, improvement response times and overcoming challenges.

Keywords: Flexibility; Data Generator; Manufacturing; Data Visualization.

1. Introduction

Many methods have been developed constantly to assess product and process quality levels, in order to compare their competitiveness [1]. Improving the manufacturing system process plays an important role in the competitiveness of the industry, mainly when advances in technology have brought many radical changes in market requirements, as well as customer needs as a priority, have a significant impact on industrial productivity [2].

The customer's needs have been met due to adequate manufacturing technologies and logistical processes, mainly due to the advancement of the Industry 4.0, which created many possibilities along with IoT (Internet of Thing), cloud computing, CPS (Cyber-Physical System), AI (Artificial Intelligence), Data Mining, among others, for improving manufacturing.

The term "Industry 4.0" originated from a German high-tech program, after the first Industrial Revolution "Mechanization", the second "Mass Production", and the third "Automation" [3]. It combines the strengths of traditional industries with cutting-edge internet technologies, which includes the development of intelligent production and manufacturing processes, new features through communication between the physical and virtual environment, and the interoperability between computer systems [4].

Production control monitoring has been one of these essential technologies to any manufacturing industry, since they are important strategies for achieving higher strategy. These systems can capture and analyze process data, and then correlate them in graphical charts for a continuous improvement.

Although the Industry 4.0 has many advantages, mainly for SMEs (Small and Medium-sized Enterprises), some manufacturing companies have difficulties and limitations to either implement new technologies or even transfer their data digitally, mainly due to the lack of tracking systems for their components and technology to collect this data, or there is not enough reliable and collected data to simulate and analyze it properly [5].

Many companies which control their processes manually tend not to get enough data to correlate their product variables. In addition, manual correlation is error-prone and a lack or excess of information might become very confusing and difficult to prepare and interpret. Others can access their data digitally, but do not know the suitable way to view and understand their relationships and get the best results and predictions.

Data visualization is a technique for converting raw data into an interactive way of presentation [6], including scientific visualization, information visualization, visual analytics, and statistical graphics.

Visualization process data is very important as it can help practitioners comprehend and analyze the relationship of all variables related to their production flow.

Thus, two issues are mainly identified in the manufacturing industries. Firstly, there are companies without a practical and confident control of their data, as they gather and analyze it by using manual and visual resources. Secondly, there are companies that, despite storing and using a digital resource to work with their data, do not use a system capable of interacting with that data and easily creating different charts to connect several of their production variables.

Many companies underestimate the potential of data visualization to transmit and communicate the best results.

Thus, the focus of this paper is to create a random and flexible data generator system for the production flow, based on real variables and offering different distributions and sizes of data, which can help researchers and practitioners to understand and configure random data in its own parameters.

This paper is organized as follows. Section 2 describes the methodological aspects. The programming and framework development, as well as the evaluation and results will be shown in Section 3. Conclusions and future studies are given in Section 4.

2. Methodological Aspects

Data generators are systems which they are capable of producing artificial datasets that does not contain the original records, and they offer possibilities to understand the environment of application [7].

Creating a data generator specifically targeted for production is very relevant and useful for the industry, as generality may reduce the accuracy in the results, hamper the usability of manufacturing practitioners and do not create the correct technical data correlation to meet the companies' needs [8].

A data generator can be implemented in different programming languages, but the author uses Python for the logical structure, along with the Microsoft Visual Studio IDE (Integrated Development Environment) to help create the graphical interface.

The choice of using Python is because this programming language is considered flexible, stable, and has many tools available. [9] highlight Python has many useful libraries for data loading, visualization, image processing, statistics, and more.

In the construction of this data generator, the creation of standard production parameters is considered vital for this development and, therefore, the user must fill in this information. These variables will be BS (Batch Size), AT (Available Time), LT (Lead Time), PT (Process Time), WT (Waiting Time), CD (Customer Demand), COT (Changeover Time), NW (Number of Workers), NS (Number of Shifts), NM (Number of Machines), I (Inventory), TO (Total Output) and DPU (Defects per Unit).

The user will fill in all this information and will be able to add any specific information related to his company that he deems necessary. After this completion, the generator will automatically calculate new variables based on standard parameters, such as CT (Cycle-Time), PE (Process Efficiency), TAKT (Takt Time), NVA (Non-Value-Adding Time), and DR (Defects Rate). In Fig. 1 is possible to identify (i) the input data reading process, (ii) the parameters, and, (iii) the output data process.

These parameters are very significant to a production process, since both are constantly applicable during a VSM (Value Stream Mapping) method, which is a Lean Manufacturing method to illustrate and analyze the logic of a production, in addition to providing an overview of the end-to-end business process [10].

Unlike VSM, a data generator will not be able to analyze a future state of a production, but it is a proof to the importance that graphical visualization of how activities and operations are connected helps to obtain new insights for improvements during production planning meetings.

The completion of this Data frame will be created in an intuitive and easy way, along with objective questions about each parameter, maximum and minimum values of each one, each distribution that can be normal, exponential or fixed, in addition to obtain the number of decimal digits of precision for each item included.

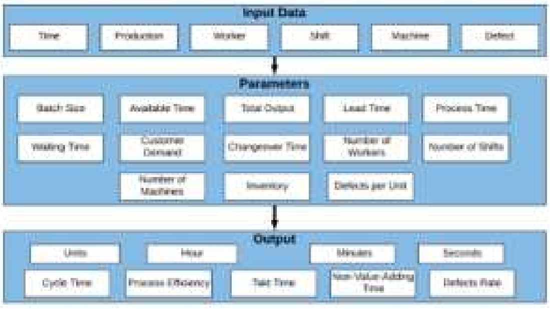


Fig. 1. Standard Data Generator Architecture.

All resulting parameters will be shown in columns within a data frame using Python Pandas, which is a software library created to manipulate and analyze data, specifically offering data structures and operations to manipulate numerical tables and time series [9].

The last step for the user it is to examine all available graphs and visualize any of his choice between the columns of parameters he wants. This will always be dynamically convenient for the user identify how the graph will best help him analyze bottlenecks and prevent himself from allowing outliers to affect production performance and productivity.

3. Framework Development

Firstly, the logic structure was created using Python programming language and the Visual Studio IDE.

Several functions were developed to create an easy, dynamic, flexible and comprehensively framework for any manufacturing practitioner and to provide a better understanding of its processes and parameters.

There are four questions presented to the first procedures of its development, regarding to the standard variables. However, as there are 13 standard variables, this results in 52 questions for each user, only in relation to these parameters.

The structure encompasses the following questions:

1. "How many rows you wish for the generated dataset?"
2. "What is the minimum value for the Batch Size (BS)?"
3. "What is the maximum value for the Batch Size (BS)?"
4. "How many decimal digits of precision do you want for the Batch Size (BS)?"
5. "What is the minimum value for the Total Output (TO)?"
6. "What is the maximum value for the Total Output (TO)?"
7. "How many decimal digits of precision do you want for the Total Output (TO)?"
8. "What is the minimum value for the Available Time (AT)?"
9. "What is the maximum value for the Available Time (AT)?"
10. "How many decimal digits of precision do you want for the Available Time (AT)?"
11. "What is the minimum value for the Lead Time (LT)?"
12. "What is the maximum value for the Lead Time (LT)?"
13. "How many decimal digits of precision do you want for the Lead Time (LT)?"
14. "What is the minimum value for the Process Time (PT)?"
15. "What is the maximum value for the Process Time (PT)?"
16. "How many decimal digits of precision do you want for the Process Time (PT)?"
17. "What is the minimum value for the Waiting Time (WT)?"
18. "What is the maximum value for the Waiting Time (WT)?"
19. "How many decimal digits of precision do you want for the Waiting Time (WT)?"
20. "What is the minimum value for the Customer Demand (CT)?"
21. "What is the maximum value for the Customer Demand (CT)?"
22. "How many decimal digits of precision do you want for the Customer Demand (CT)?"
23. "What is the minimum value for the Changeover Time (COT)?"
24. "What is the maximum value for the Changeover Time (COT)?"
25. "How many decimal digits of precision do you want for the Changeover Time (COT)?"
26. "What is the minimum value for the Number of Workers (NW)?"
27. "What is the maximum value for the Number of Workers (NW)?"
28. "How many decimal digits of precision do you want for the Number of Workers (NW)?"
29. "What is the minimum value for the Number of Shifts (NS)?"
30. "What is the maximum value for the Number of Shifts (NS)?"
31. "How many decimal digits of precision do you want for the Number of Shifts (NS)?"
32. "What is the minimum value for the Number of Machines (NM)?"
33. "What is the maximum value for the Number of Machines (NM)?"
34. "How many decimal digits of precision do you want for the Number of Machines (NM)?"
35. "What is the minimum value for the Inventory (I)?"
36. "What is the maximum value for the Inventory (I)?"
37. "How many decimal digits of precision do you want for the Inventory (I)?"
38. "What is the minimum value for the Defects per Unit (DPU)?"
39. "What is the maximum value for the Defects per Unit (DPU)?"
40. "How many decimal digits of precision do you want for the Defects per Unit (DPU)?"
41. "Do you wish to add more columns?"

Each of these 13 parameters, as well as the new specific parameters that the user can add from question 41, can be classified as [1] Random Generation, [2] Normal Distribution, [3] Exponential Distribution, or [4] Textual (i.e., words as year no). This classification will be free to the user's choice, as the data will be generated according to that choice.

In consideration of the length of this questionnaire, which is not practical for users, it was converted to a dynamic and friendly interface in Microsoft Visual Studio, allowing the compact and simultaneous viewing of these questions on the same screen, with free decision of which parameters will definitely use in their analysis (Fig.2).

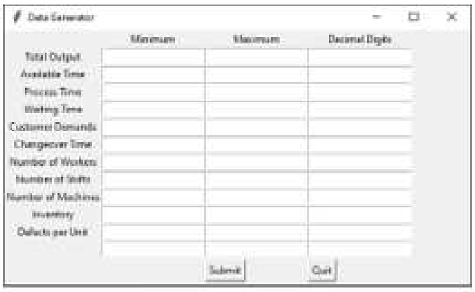
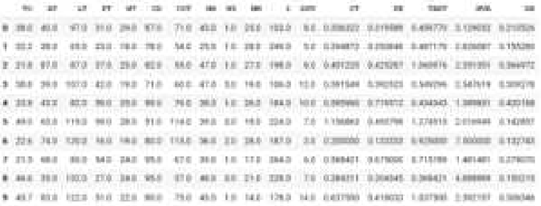


Fig. 2. Standard Parameters Data Collection.

After completing the standard data, they will automatically be shown all in a column structure, along with the variables calculated from this data (Fig.3). If any data is not included in the previous screen, the user will be asked that it is not possible to calculate any of the calculated variables (e.g., takt time, cycle time, non-value-adding time, defect rate, and process efficiency) and then a confirmation for this decision will be made.



	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Fig. 3. Example of a generated Dataframe.

As soon as this data is made available, a series of graphs will be demonstrated (Fig. 4), as well as the dataframe created, allowing the user to choose the graph and the variables that will be used on the X and Y axes, in order to obtain dynamically the development of these graphs with the data developed.

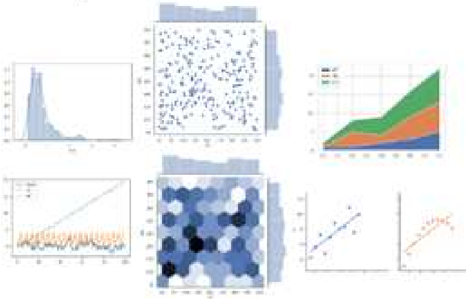


Fig. 4. Graphical Visualization from Generated Data.

In general histogram, dispersion, bars and lines, box plots, stack plots, joint plots, scatterplot and pair plots graphs can be created.

Then, from structured data and with the precise identification of various graphical perspectives, it is possible that professionals can identify in a much more dynamic way how to identify bottlenecks and outliers, being an excellent alternative to help the company overcome challenges in its production with a vision far more comprehensive and intelligent of your data.

4. Conclusions

The present work developed a data generator, focusing on developing a system that could help companies better understand their data and be able to previously identify any errors or deviations in their production, as well as obtain opportunities to create new insights of intelligent strategies for company's continuous improvement.

There are no data generators targeted for production, available on the market, hence this system will fulfill this need for companies to gain better control over their production planning and achieve process optimization.

As future studies, it is possible to highlight the integration with the Microsoft Azure cloud computing system along with its database, in order to provide access to the cloud on any computer or even be able to store the data and analyze it in real-time. The data generator creates the data randomly or by a specific distribution, but as a result of this development, it is identified that it is very relevant to upload the data, in any extension (e.g., .csv, .json, etc.) already developed to obtain data in real time, which the algorithm can transform and set up all data for a new comprehensive data structure which it can obtain better results for the industry.

References

[1] I. Petrov, V. Kharitonov, and M. Polyakova, "Procedure for evaluating competitiveness of production processes," in *Materials Science Forum*, 2019, vol. 946 MSF, pp. 726–731.

[2] G. S. Rawat, A. Gupta, and C. Juneja, "Productivity Measurement of Manufacturing System," in *Materials Today: Proceedings*, 2018, vol. 5, no. 1, pp. 1483–1489.

[3] R. Schmidt, M. Möhring, R. C. Härting, C. Reichstein, P. Neumaier, and P. Jozinović, "Industry 4.0 - Potentials for creating smart products: Empirical research results," in *Lecture Notes in Business Information Processing*, 2015, vol. 208, pp. 16–27.

[4] Y. Zeng and Y. Yin, "Virtual and Physical Systems Intra-referenced Modelling for Smart Factory," in *Procedia CIRP*, 2017, vol. 63, pp. 378–383.

[5] R. Umer, T. Susnjak, A. Mathrani, and S. Suriadi, "On predicting academic performance with process mining in learning analytics," *J. Res. Innov. Teach. Learn.*, vol. 10, no. 2, pp. 160–176, Jul. 2017.

[6] F. Z. Fezarudin, M. I. Illyas Tan, and F. A. Qasem Saeed, "Data visualization for human capital and halal training in halal industry using tableau desktop," in *Communications in Computer and Information Science*, 2017, vol. 752, pp. 593–604.

[7] Y. Chen, J. Taub, and M. Elliot, "CONFERENCE OF EUROPEAN STATISTICIANS Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Optimising the Trade-off between Information

Utility and Disclosure Risk in a GA Synthetic Data Generator Trade-off between Information Utility and," 2019.

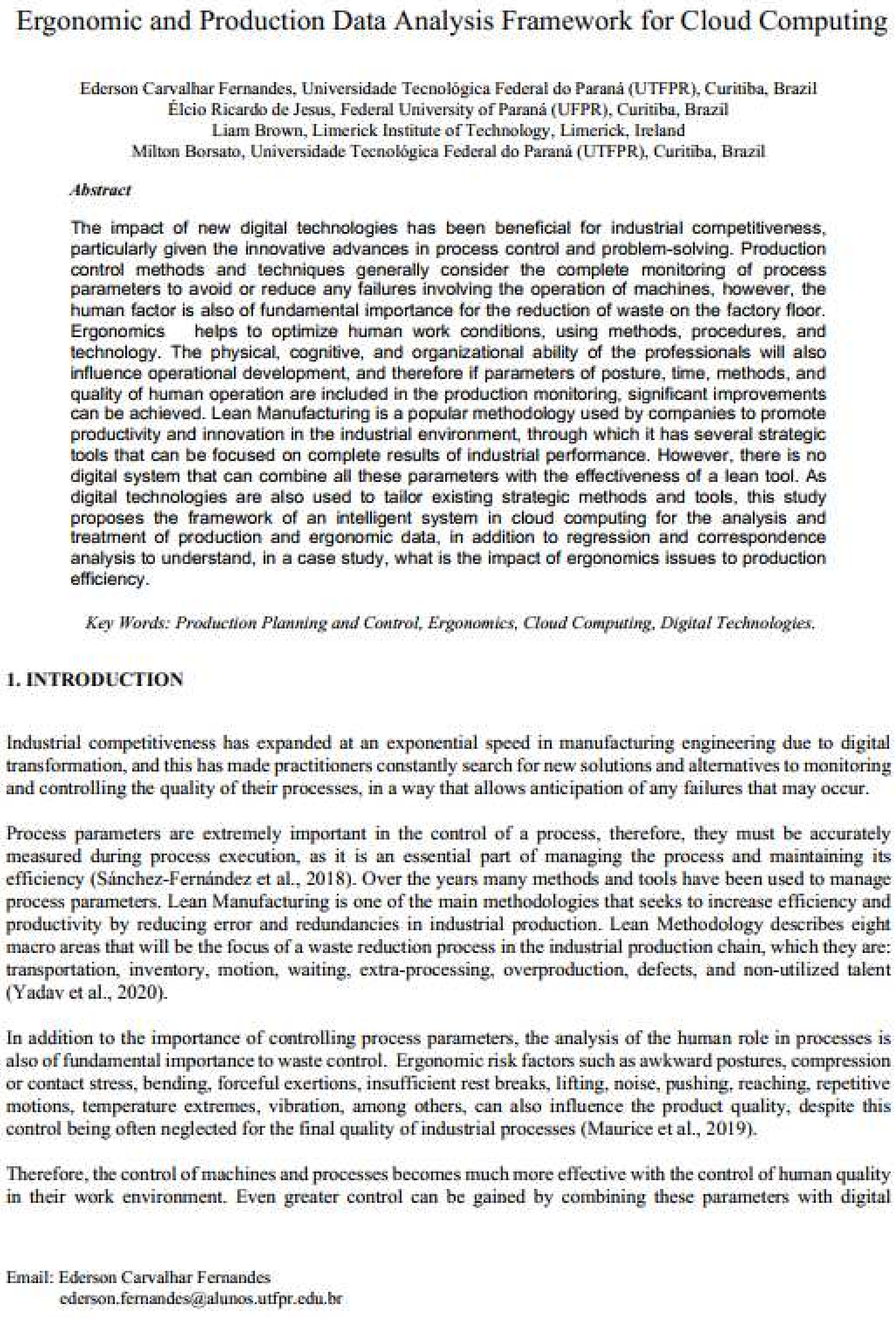
[8] V. Ayala-Rivera, A. O. Portillo-Dominguez, L. Murphy, and C. Thorpe, "COCOA: A synthetic data generator for testing anonymization techniques," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9867 LNCS, pp. 163–177, 2016.

[9] A. C. Muller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*, First Edit., vol. 1, no. 3. Sebastapol, CA: O'Reilly Media, 2017.

[10] J. Langstrand, "An introduction to value stream mapping and analysis," *Sweden*, no. November, pp. 1–25, 2016.

APPENDIX E – Ergonomic and Production Data Analysis Framework for Cloud Computing (Published Paper)

ERGONOMIC AND PRODUCTION DATA ANALYSIS FRAMEWORK FOR CLOUD COMPUTING (PUBLISHED PAPER)



Ergonomic and Production Data Analysis Framework for Cloud Computing

Ederson Carvalhar Fernandes, Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Brazil
Élcio Ricardo de Jesus, Federal University of Paraná (UFPR), Curitiba, Brazil
Liam Brown, Limerick Institute of Technology, Limerick, Ireland
Milton Borsato, Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Brazil

Abstract

The impact of new digital technologies has been beneficial for industrial competitiveness, particularly given the innovative advances in process control and problem-solving. Production control methods and techniques generally consider the complete monitoring of process parameters to avoid or reduce any failures involving the operation of machines, however, the human factor is also of fundamental importance for the reduction of waste on the factory floor. Ergonomics helps to optimize human work conditions, using methods, procedures, and technology. The physical, cognitive, and organizational ability of the professionals will also influence operational development, and therefore if parameters of posture, time, methods, and quality of human operation are included in the production monitoring, significant improvements can be achieved. Lean Manufacturing is a popular methodology used by companies to promote productivity and innovation in the industrial environment, through which it has several strategic tools that can be focused on complete results of industrial performance. However, there is no digital system that can combine all these parameters with the effectiveness of a lean tool. As digital technologies are also used to tailor existing strategic methods and tools, this study proposes the framework of an intelligent system in cloud computing for the analysis and treatment of production and ergonomic data, in addition to regression and correspondence analysis to understand, in a case study, what is the impact of ergonomics issues to production efficiency.

Key Words: Production Planning and Control, Ergonomics, Cloud Computing, Digital Technologies.

1. INTRODUCTION

Industrial competitiveness has expanded at an exponential speed in manufacturing engineering due to digital transformation, and this has made practitioners constantly search for new solutions and alternatives to monitoring and controlling the quality of their processes, in a way that allows anticipation of any failures that may occur.

Process parameters are extremely important in the control of a process, therefore, they must be accurately measured during process execution, as it is an essential part of managing the process and maintaining its efficiency (Sánchez-Fernández et al., 2018). Over the years many methods and tools have been used to manage process parameters. Lean Manufacturing is one of the main methodologies that seeks to increase efficiency and productivity by reducing error and redundancies in industrial production. Lean Methodology describes eight macro areas that will be the focus of a waste reduction process in the industrial production chain, which they are: transportation, inventory, motion, waiting, extra-processing, overproduction, defects, and non-utilized talent (Yadav et al., 2020).

In addition to the importance of controlling process parameters, the analysis of the human role in processes is also of fundamental importance to waste control. Ergonomic risk factors such as awkward postures, compression or contact stress, bending, forceful exertions, insufficient rest breaks, lifting, noise, pushing, reaching, repetitive motions, temperature extremes, vibration, among others, can also influence the product quality, despite this control being often neglected for the final quality of industrial processes (Maurice et al., 2019).

Therefore, the control of machines and processes becomes much more effective with the control of human quality in their work environment. Even greater control can be gained by combining these parameters with digital

Email: Ederson Carvalhar Fernandes
ederson.fernandes@alunos.utfpr.edu.br

technology tools. Digital technologies provide an opportunity to greatly increase productivity, reduce costs, gains deeper insights, adaptability, among others (Cadavid et al., 2020).

Despite the help of digital technologies, most cognitive work in the digital age will remain dependent on human work, as this type of work requires non-programmable activities, and for that it also requires human creative work (Akaev & Sadovnichii, 2021).

In regards of many combinations and developed smart method adaptations, as well as applications of smart systems that have been applied in areas such as logistics and supply chain, there are many unfilled gaps for new smart developments in order to improve the results from traditional manufacturing methods (Becker & Intoyoad, 2017; Blackhurst et al., 2018; ER et al., 2018; Wang et al., 2014).

Digital transformation is the process in which companies use digital technologies to solve traditional problems, such as declines in performance, productivity, agility, and effectiveness. This transformation must come from a structural change in organizations.

Currently, digital technologies are being used as key resources to support the transformation and productivity trajectories of organizations. The introduction of digital technologies in industrial processes has as its basic principle connecting machines, systems, products, and people, creating intelligent networks throughout the value chain, in addition to automating the production line (Núñez-Merino et al., 2020). These technologies change the way companies currently operate, transforming them into smarter factories capable of commanding all processes in the production chain through interconnected systems and devices that communicate to each other and to humans in real time.

Thus, all processes can be integrated, as well as being monitored in real time, increasing the company's productivity and effectiveness, and improving its position in the market and its consumers. Among the main technologies used, it stands out: IoT (Internet of Things), IoS (Internet of Services), Machine Learning, Artificial Intelligence, Big Data, Cloud Computing, Virtual Reality, Augmented Reality, Autonomous Robots, and 3D Printing.

Each technology has its own peculiarity with more or less application benefits depending on the context used, however, cloud computing is an evolution of information technology and a dominant business model for delivering IT (Information Technology) resources.

Cloud computing refers to the provision of memory and storage, plus the processing of data on computers and servers interconnected through the internet. By having these characteristics, the cloud can provide computing services such as servers, network, databases, software, among others, differing from physical infrastructures for its flexibility, high data availability and a processing capacity that can be unlimited. This allows optimization of resources and time, and provides more and more storage and data processing capacity for companies (Ramachandra et al., 2017; Sunyaev, 2020).

Focusing on improving the level of competitiveness of the production process and reducing ergonomic problems in a manufacturing company, this paper aims to develop a case study in a large construction and agricultural equipment company located in Brazil, which has been facing some issues regarding the ergonomics of its employees, and this is directly affecting its results. Thus, a cloud computing framework will be developed for a smart application of an ergonomic and production data analysis of an assembly line of this company.

Knowledge about the ergonomic and production parameters and variables that involve the analysed production line is essential for the applicability of the system processing. Thus, an exploratory study was previously developed in the company for the primary structuring of a sufficient dataset for use in cloud computing.

This project aims to contribute to the PPC (Production Planning and Control) with the development of an intelligent system in cloud computing for the analysis, treatment, query, and learning of ergonomic datasets.

The cloud computing framework encompasses a set of IT resources in which they are programmed to act automatically from a predetermined trigger, which will be the inclusion of a new dataset into the system.

It will assist engineers in decision-making and planning process improvements, due to predictive linear regression analysis, in addition to obtaining processing responses in seconds, to allow that practitioners can optimize their time for other activities.

This paper is organized as follows. The Introduction already presented in Section 1 of this project. Section 2 describes a brief literature review about PPC, ergonomics, machine learning, and cloud computing. The proposed framework, as well as the services and parameters will be shown in section 3. Some conclusions and future recommendations are given in Section 4.

2. LITERATURE REVIEW

Production planning is a part of PPC, which it helps to identify what to produce, when to produce, how much to produce, as it involves a long-term view of the overall production planning (Bueno et al., 2020).

Production planning has two main objectives: ensure the quantity and quality of the items that are available during the production and ensure the capacity utilization is in line with forecast demand. Production planning takes care of product and process planning, and it is done in long-term, medium-term, and short-term planning.

Production control seeks to use different types of control techniques to achieve optimal production system performance, and overall production planning goals. Ensuring a smooth flow of all production processes, ensuring production cost savings, controlling waste of resources, and maintaining the quality standard throughout the product lifecycle are the main advantages of robust production control (Agostino et al., 2020).

Just as it is important to obtain analysis of production planning, it is essential to obtain accurate diagnoses about the quality of the work of the practitioners involved. Ergonomics consists of a set of disciplines that study the organization of work in which interactions between human beings and machines exist. The main objective of ergonomics is to develop and apply techniques for adapting elements of the work environments to human beings, along with the aim of generating worker well-being and consequently increasing their productivity.

Ergonomics concept applies to the quality of adaptation of a machine to its operator, providing an effective handling and avoiding an extreme effort by the worker in carrying out the work. Using ergonomic solutions in the workplace is an initiative that can significantly increase worker satisfaction, effectiveness, and efficiency (Valamede & Lima, 2019).

Process data such as waiting time, inventory, customer demand, performance, lead time, total parts, among others, can usually be collected from observation, checklist, or even through recordings of industrial machines themselves. Ergonomic data, on the other hand, are obtained through observation, monitoring, or through information obtained from the operators themselves.

After the company periodically collects all this data, it is necessary to correlate and map them to acquire a broad reading of current productivity and calculate what can be optimized, in addition to identifying the production bottleneck or even the ergonomic situation.

Because these analyzes are too time-consuming, and often repetitive, the use of digital technologies can streamline and optimize this process, bringing dynamic and effective results. Human beings have always learned by observing patterns, formulating hypotheses, and testing them to discover rules. What is new in this computer age is the enormous amount of data that can no longer be scrutinized for patterns within a reasonable amount of time. Data Mining arises for this purpose and can be applied both to scientific research and to boost the profitability of a mature, innovative, and competitive company.

Data Mining is the process of exploring large amounts of data, looking for consistent patterns, as association rules or temporal sequences, to detect systematic relationships between variables, thus detecting new subsets of data. Data Mining consists of a set of tools and techniques that use learning or classification algorithms based on neural networks and statistics. These can explore a set of data, extracting or helping to evidence patterns in these data and assisting in knowledge discovery. Knowledge in Data Mining can be presented by these tools in several ways, such as clusters, hypotheses, rules, decision tress, graphs, or dendrograms (Accorsi et al., 2017; Alsrehin et al., 2019; Chowdhury et al., 2018).

In addition to Data Mining, Machine Learning (ML) can leverage established data patterns to analyze their correlations, learn these patterns, and then forecast future trends on company-defined objectives.

ML is the study of algorithms and mathematical models that gives computers the ability to learn without being explicitly programmed. Moreover, it is also known as statistical learning or predictive analytics and it helps open new possibilities up for improving decision-making and performance on several tasks. Muller and Guido, (2017) emphasized the most successful ML algorithms are those which automate decision-making processes by generalizing from knows examples.

Usually, ML uses two techniques: supervised and unsupervised learning (Namdev et al., 2015). The former provides the algorithm with pairs of inputs and desired outputs, then it learns with the input and seeks to obtain all desired outputs. The best algorithm will be when it is able to create an output for an input which has never seen before without any help from a human (e.g., detecting fraudulent activity in credit card transactions, determining whether a tumour is benign based on a medical image, identifying the zip code from handwritten digits on an envelope).

On the other hand, the latter is only able to know the input data, and, by itself, it needs to find hidden patterns or intrinsic structures (e.g., detecting abnormal access patterns to a website, segmenting customers into groups with similar preferences, identifying topics in a set of blog posts).

Both processes can be developed in various software and applications, as well as through a set of cloud computing platforms. Cloud Computing (CC) platforms are services where customers can use it to run applications and store data on Internet-accessible machines owned by other companies, in addition to providing a valuable application-specific service in many application domains (Ray, 2016).

Using CC, organizations can use shared computing and storage resources rather than building, operating, and improving infrastructure on their own. It is a model that allows users to provision and release resources on demand, resources to be scaled up or down automatically depending on load, resources accessible over a network with suitable security, and cloud services from service providers that can allow pay-per-use model, where customers are billed based on the type of features and per usage.

Cloud computing models are being used more and more by manufacturing organizations and in other areas, as a cloud computing model, with dynamic, secure, and private access, that can develop agile analysis and effective results after each new data entry, can become very useful for the company's competitiveness.

Therefore, the thorough analysis of process and ergonomics parameters can bring valuable insights into continuous improvement in an organization.

3. ERGONOMIC AND PRODUCTION DATA ANALYSIS FRAMEWORK

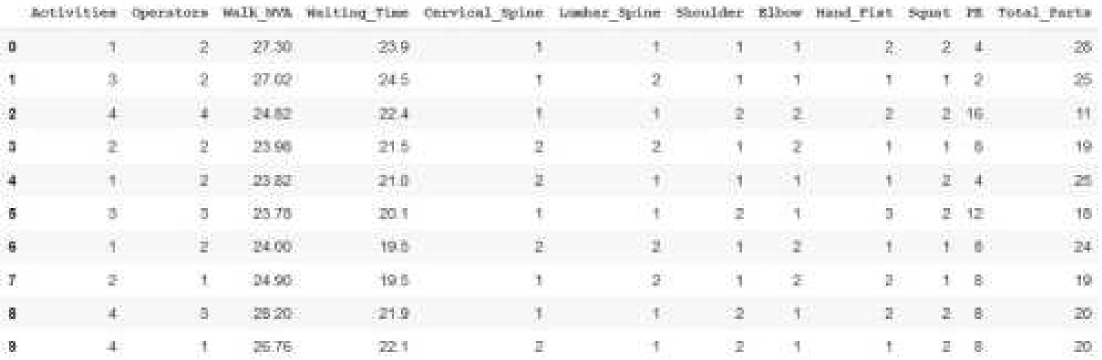
Create a smart system allocated on a cloud computing platform, where it will always have availability, especially in critical ones, is very relevant and useful for the industry, as it is a paradigm shift in the way hardware and software resources are managed and utilized (Bello et al., 2021).

The methodology used for the study of the application of CC is mainly through the secondary method, which includes data already gathered by the company. This study does not include the data capture method, as it is the entire responsibility of the organization to do so (e.g., computer vision, tracking systems, machine connectivity, etc.). The development of the framework is built from the inclusion of the data already obtained.

Data were obtained in an assembly line with four different activities, where four operators were observed at different times (i.e., each operator performs each of these four activities at different times). In total 365 readings make up the dataset, where operators are classified by numbers, as well as activities (i.e., 1,2,3, and 4). Three process parameters have been included, such as: walking time, waiting time, and total parts

Ergonomics data uses a numerical rating between 1 and 3 for Work Postures, such as Cervical Spine, Lumbar Spine, Shoulder, Elbow, Hand Fist, and Squat , which was adapted from Jarebrant (2016). Work postures mean the position of each member when a work task is carried out, where 1 corresponds to a favorable posture, and 3 corresponds to an unfavorable posture (Figure 1).

In a complete evaluation of all physical ergonomics parameters, all numerical values from work postures were multiplied together to obtain a single value to be included in the data correlation analysis.



	Activities	Operators	Walk_NVA	Waiting_Time	Cervical_Spine	Lumbar_Spine	Shoulder	Elbow	Hand_Fist	Squat	PE	Total_Parts
0	1	2	27.30	23.9	1	1	1	1	2	2	4	26
1	3	2	27.02	24.5	1	2	1	1	1	1	2	25
2	4	4	24.82	22.4	1	1	2	2	2	2	16	11
3	2	2	23.96	21.5	2	2	1	2	1	1	8	19
4	1	2	23.82	21.0	2	1	1	1	1	2	4	25
5	3	3	23.78	20.1	1	1	2	1	3	2	12	18
6	1	2	24.00	19.5	2	2	1	2	1	1	8	24
7	2	1	24.90	19.5	1	2	1	2	2	1	8	19
8	4	3	28.20	21.9	1	1	2	1	2	2	8	20
9	4	1	25.76	22.1	2	1	2	1	1	2	8	20

Figure 1. Dataset Sample

The cloud computing framework can be reproduced on any platform, however for this case study the AWS (Amazon Web Services) platform was used, and this choice was solely due to the exclusive access used by the institution affiliated to this research. Access to this system will be web-based, which allows access on any device, with no restriction on which browser to use.

The purpose of the applicability of this system will be to streamline visual responses to the user (i.e., practitioners of the company), so that only the inclusion of new data will get answers to what is going on in the current time and what may occur in a future on the organization's production lines, where they just need to update the system with new real data available on the shop floor, every day or week.

Thus, the AWS services used for the development of this research were: S3 (Simple Storage Service), Glue, Lambda, Athena, and Sagemaker (Figure 2). Security applications will not be covered in this paper, just to keep the focus on getting correlation and insights into the results, but security and monitoring applications are highly recommended to complement this development.

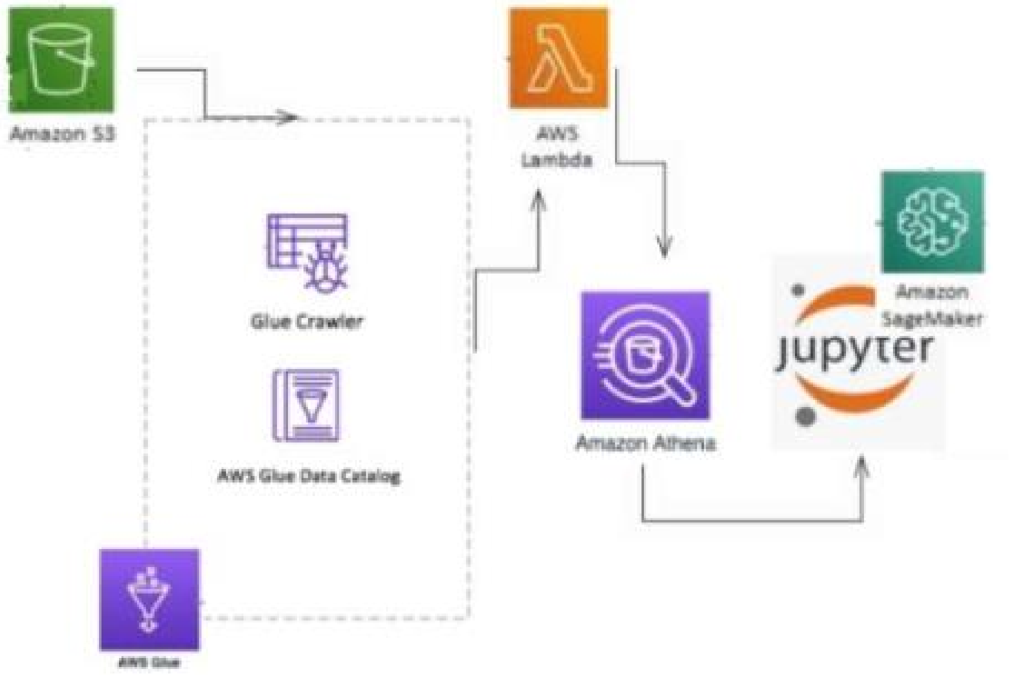


Figure 2. Data Analysis Framework

S3 is a file storage service, also called objects or blobs, focused on scalability, availability, security and performance, all at an extremely affordable cost (Amazon, 2021b). Glue is a fully managed extract, transform, load (ETL) service that makes it easier and more cost-effective to categorize data, cleanse it, enhance it, and reliably move it across multiple data stores and streams (Amazon, 2021d). Lambda is a serverless computing service that allows to run code without provisioning or managing servers. Lambda can run code for virtually any type of application or back-end service without administration (Amazon, 2021e). Athena is an interactive query service that maker it easy to analyse data using standard Structured Query Language (SQL) (Amazon, 2021a). Sage maker is a service that makes it easy to quickly and reliably create, train, and implement ML models. It is fully managed by AWS, which removes the heavy lifting from each step to develop high-quality ML models (Amazon, 2021c).

S3 is used to store data files, which can be collected daily, as the cloud space can be expanded if necessary. These files must be saved exclusively in the CSV (Comma-separated Value) format, by which they are the files configured for processing, and easily converted to structured datasets.

Glue is an intermediate service between S3 and Athena, which will transfer the CSV data to its own database that Athena will be able to access more easily. Using Lambda in this process will be vital to the development of the system, as each application must be updated individually, and Lambda creates triggers in the cloud so that when placing new files in S3, Glue automatically is executed to transfer these files to Athena, without the need for any isolated commands.

Once the dataset is transferred to Athena, it is possible to work with the SQL language, which is a programming language to deal with relational database (based on tables). It was created so that multiple developers could access and modify a company's data simultaneously, in an uncomplicated and unified way (Chandra et al., 2019).

The use of SQL queries is very important as it is an excellent way to remove outliers in the dataset. However, the use of SQL will not be done within Athena application, but in Sagemaker. The inclusion in Athena was strategically programmed to connect the dataset to this application and then enable the use of the SQL language. The ability to use ML with SQL queries makes it possible to perform complex tasks.

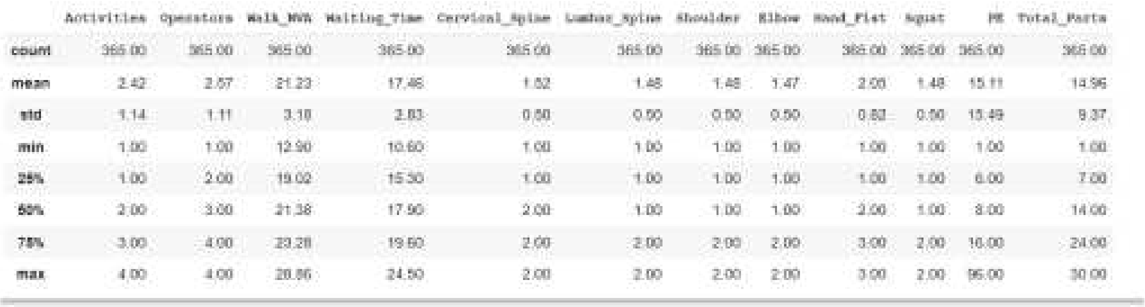
The next step will be to connect the dataset path between Athena and Sagemaker, whereby it will be possible by creating a new trigger by Lambda, where every time the dataset accesses Athena, it will be connected to Sagemaker.

Sagemaker provides access to Jupyter Notebook, which is a open-source web application, that exists to develop open source software, open standards and services for interactive computing in dozens of programming languages (Jupyter, 2021). The programming language used in this project was Python, since it is a flexible, stable language, and has several tools and functionalities available for free (Python, 2021).

When creating a code with SQL language accessibility it was necessary to install a library called PyAthena, in addition to other libraries to enable data loading, scientific calculations, graphical visualizations, analysis and statistics, such as Pandas, Matplotlib, and Numpy

PyAthena is a library for accessing Athena, which contains several methods for using SQL commands. Pandas is a library for data manipulation and analysis, and it offers data structures and operations for manipulating numerical tables and time series. Matplotlib is a plotting library for explicitly create figures and axes, and Numpy is an open-source numerical library, which can be utilized to perform several mathematical operations on arrays such as trigonometric, statistical, and algebraic routines.

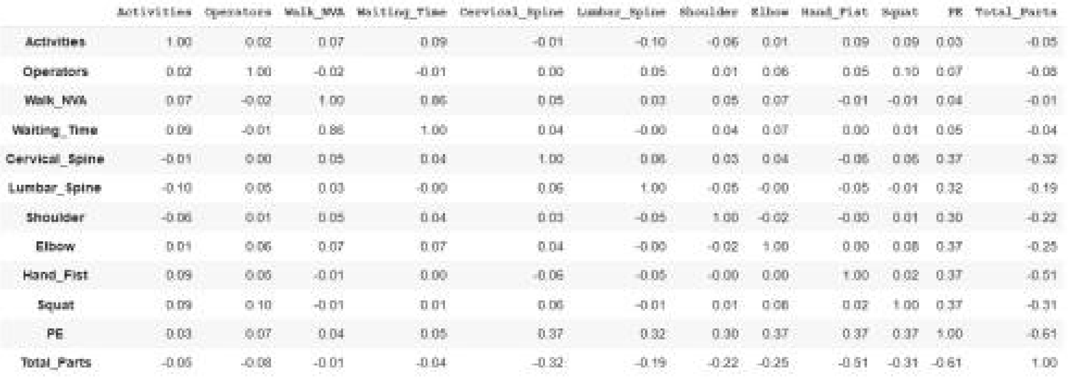
Firstly, descriptive statistics were analyzed, which summarized the central tendency, dispersion and distribution configuration of the data (Figure 3).



	Activities	Operators	Walk_MVA	Waiting_Time	Cervical_Spine	Lumbar_Spine	Shoulder	Elbow	Hand_Fist	Squat	PR	Total_Parts
count	365.00	365.00	365.00	365.00	365.00	365.00	365.00	365.00	365.00	365.00	365.00	365.00
mean	2.42	2.57	21.23	17.46	1.52	1.48	1.48	1.47	2.05	1.48	15.11	14.96
std	1.14	1.11	3.18	2.83	0.50	0.50	0.50	0.50	0.82	0.50	15.49	9.37
min	1.00	1.00	12.90	10.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
25%	1.00	2.00	19.02	15.30	1.00	1.00	1.00	1.00	1.00	1.00	6.00	7.00
50%	2.00	3.00	21.38	17.90	2.00	1.00	1.00	1.00	2.00	1.00	8.00	14.00
75%	3.00	4.00	23.28	19.60	2.00	2.00	2.00	2.00	3.00	2.00	16.00	24.00
max	4.00	4.00	28.86	24.50	2.00	2.00	2.00	2.00	3.00	2.00	96.00	30.00

Figure 3. Descriptive Statistics

Then, pairwise correlation of all columns in the dataset was performed, with the number 1 representing full correlation between the data (Figure 4)



	Activities	Operators	Walk_NVA	Waiting_Time	Cervical_Spine	Lumbar_Spine	Shoulder	Elbow	Hand_Fist	Squat	PE	Total_Parts
Activities	1.00	0.02	0.07	0.09	-0.01	-0.10	-0.06	0.01	0.09	0.09	0.03	-0.05
Operators	0.02	1.00	-0.02	-0.01	0.00	0.05	0.01	0.06	0.05	0.10	0.07	-0.08
Walk_NVA	0.07	-0.02	1.00	0.86	0.05	0.03	0.05	0.07	-0.01	-0.01	0.04	-0.01
Waiting_Time	0.09	-0.01	0.86	1.00	0.04	-0.00	0.04	0.07	0.00	0.01	0.05	-0.04
Cervical_Spine	-0.01	0.00	0.05	0.04	1.00	0.06	0.03	0.04	-0.06	0.06	0.37	-0.32
Lumbar_Spine	-0.10	0.05	0.03	-0.00	0.06	1.00	-0.05	-0.00	-0.05	-0.01	0.32	-0.19
Shoulder	-0.06	0.01	0.05	0.04	0.03	-0.05	1.00	-0.02	-0.00	0.01	0.30	-0.22
Elbow	0.01	0.06	0.07	0.07	0.04	-0.00	-0.02	1.00	0.00	0.08	0.37	-0.25
Hand_Fist	0.09	0.05	-0.01	0.00	-0.06	-0.05	-0.00	0.00	1.00	0.02	0.37	-0.51
Squat	0.09	0.10	-0.01	0.01	0.06	-0.01	0.01	0.08	0.02	1.00	0.37	-0.31
PE	0.03	0.07	0.04	0.05	0.37	0.32	0.30	0.37	0.37	0.37	1.00	-0.61
Total_Parts	-0.05	-0.08	-0.01	-0.04	-0.32	-0.19	-0.22	-0.25	-0.51	-0.31	-0.61	1.00

Figure 4. Pairwise Correlation

It is noteworthy that all these settings and correlation algorithms become fixed and automatically updated in the cloud after new data insertions by S3.

Graphical visualizations such as the analysis of total parts for each operator (Figure 5), full analysis of productivity of total parts (Figure 6), Joint Plot (Figure 7), Total Dispersion (Figure 8) and divided by each operator (Figure 9), brought in conclusion, in relation to the data used, the highest rate of ergonomic issues occurs when operators complete more than 20 executions in the analysed cycle time.

As with the box plot, the frequency distribution (figure 6) can also analyze the productivity of parts, but in a temporal aspect through days or weeks.

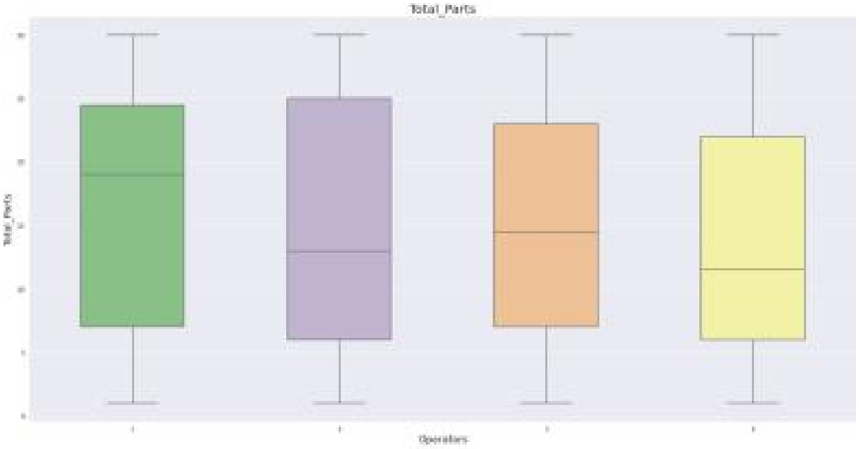


Figure 5. Box Plot

In Figure 5, it is possible to identify that there are variations in the performance of each operator, which helps to identify low or high productivity, then analyse how much this impacts on ergonomic indexes.

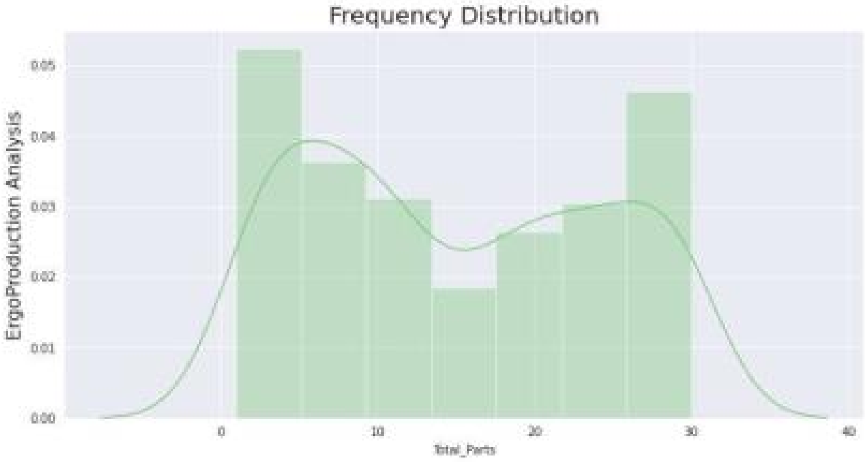


Figure 6. Frequency Distribution

As with the box plot, the frequency distribution (figure 6) can also analyze the productivity of parts, but in a temporal aspect through days or weeks. A low frequency at the center of this distribution can be identified, which may mean an issue in the process, which could be through a maintenance problem that affected productivity, or even an ergonomic problem that would hinder the amount of this production.

Joint plot is a visualization that together with the previous analysis, can help in several insights through the relationship of two variables of greater relevance to what is expected to be monitored. In this case, by joining the ergonomics variable with the total produced parts variable, the regression of this correlation can be verified. Low production is visible with the increase of ergonomic problems in the company under study.

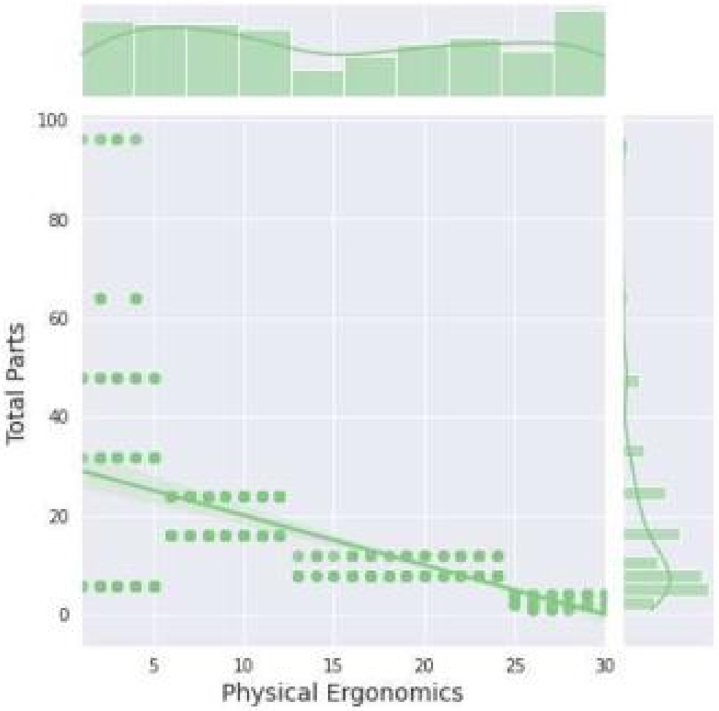


Figure 7. Joint Plot

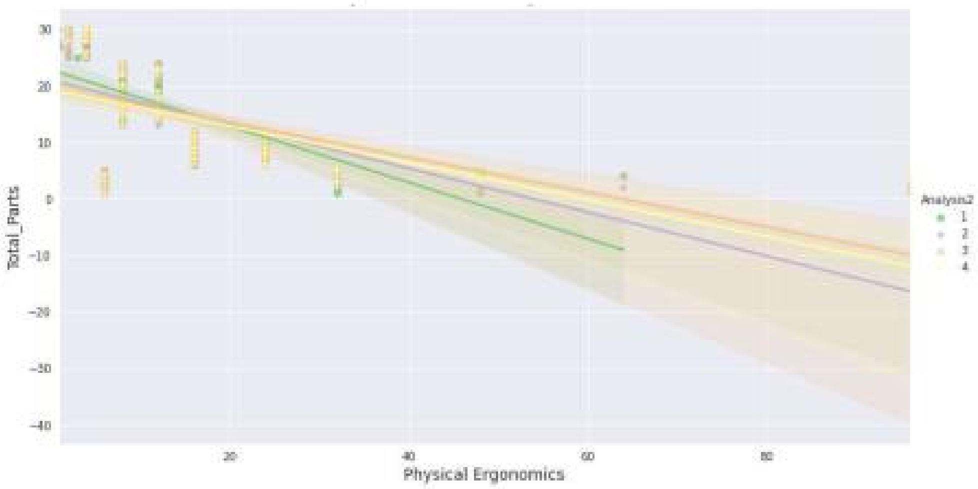


Figure 8. Total Dispersion

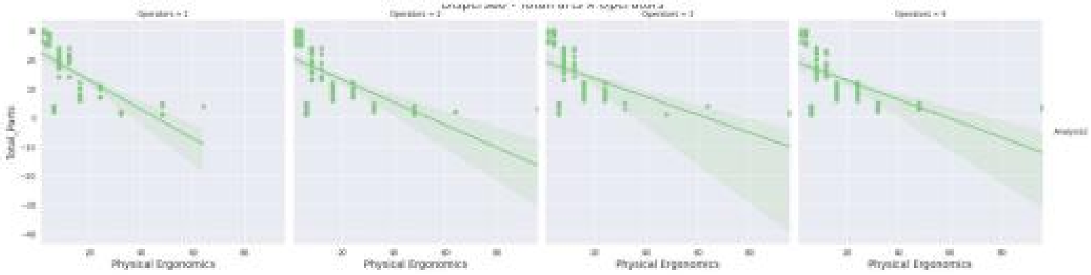


Figure 9. Dispersion by each operator

In a more specific analysis by the visualizations represented in Figures 8 and 9, each operator can be monitored individually to check their productivity. This provides a streamline of the investigation of causes that may affect the production cycle time. This further exploratory investigation will be able to determine if it is a problem with operator training, or even a specific issue to a single operator.

After this data analysis, a machine learning linear regression algorithm was performed to identify the future trend that could occur within this scenario, which for this case study, revealed an increase in the inversely proportional correlation between the number of completed activities and the number of unfavourable postures to work activities. This linear regression obtained an accuracy of 80%, due to the amount of data used, which identifies the need to increase the data used.

Although the accuracy still does not reach values above 80%, the practicality of responses explored by the framework was considered extremely relevant to the identification of causes of irregularities that may already occur or even to prevent future problems.

The issues that the company has been facing need to be prevented in advance. The use of camera monitoring could be an alternative, but it would still require someone to carefully analyze all the details that may occur during production activities, and this could incur even greater costs, as well as considering assumptions or misinterpretations of the real causes. This virtual system can analyze all data uninterruptedly, for greater learning with each new feed, thus obtaining new and effective insights to the company.

4. CONCLUSIONS AND OUTLOOK

The present work developed a case study in a agricultural and construction equipment company. The structure of cloud computing, as well the results, were evaluated by three practitioners, who identified that the predicted data is largely consistent with reality, as it presented residuals (i.e., outliers) in regression analysis.

Despite recent advances in cloud computing technologies, the existing literature shows that there is no analysis of the use of cloud computing with correlation of ergonomics data with process data. The study highlights the boundaries, and challenges of development and application in this area.

This system is stored in a virtual cloud instance, with access on any device with internet access, and with high scalability for the development of machine learning, which through large-scale data feeding can increase the accuracy rate and the efficiency of investigation of current and future analyzes of an industrial process.

SWOT (Strengths, Weaknesses, Opportunities and Threats analysis) was used to determine how well this system is working according to what was pre-established or expected by the organization (Gürel & Tat, 2017).

Strengths: The system provides a graphical view of the data, using different methods of analysis, in addition to the application of machine learning for the development of future analyses.

Weaknesses: it still does not have a direct communication to the data from the monitoring cameras, requiring the user to feed in new data.

Opportunities: Optimize the system to use machine learning results to generate graph visualizations with data relationship (i.e., create a value stream mapping to the data provided).

Threats: The system is already easy to use and brings useful information to improve processes, but a possible threat would be the union of this system with IoT applications.

In future studies, it is planned to investigate the possibility of creating graphs to map the relationships between each variable, enabling a visual reading and ease in obtaining new insights in current and future analyses. Furthermore, it can be tested with different process and ergonomic parameters, and analyse with more data, so that machine learning will be able to obtain greater accuracy in the results.

5. ACKNOWLEDGEMENTS

Special thanks to CNPq (nº443854/2018-1) for encouraging research.

6. REFERENCES

Accorsi, R., Manzini, R., Pascarella, P., Patella, M., & Sassi, S. (2017). Data Mining and Machine Learning for Condition-based Maintenance. *Procedia Manufacturing*, *11*, 1153–1161. https://doi.org/10.1016/j.promfg.2017.07.239

Agostino, Í. R. S., Broda, E., Frazzon, E. M., & Freitag, M. (2020). Using a digital twin for production planning and control in industry 4.0. *International Series in Operations Research and Management Science*, *289*, 39–60. https://doi.org/10.1007/978-3-030-43177-8_3

Akaev, A. A., & Sadovnichii, V. A. (2021). The Human Component as a Determining Factor of Labor Productivity in the Digital Economy. *Studies on Russian Economic Development*, *32*(1), 29–36. https://doi.org/10.1134/S1075700721010020

Alsrehin, N. O., Klaib, A. F., & Magableh, A. (2019). Intelligent Transportation and Control Systems Using Data Mining and Machine Learning Techniques: A Comprehensive Study. *IEEE Access*, *7*, 49830–49857. https://doi.org/10.1109/ACCESS.2019.2909114

Amazon. (2021a). *Amazon Athena*. https://aws.amazon.com/athena/

Amazon. (2021b). *Amazon S3*. https://aws.amazon.com/s3/

Amazon. (2021c). *Amazon Sagemaker*. https://aws.amazon.com/sagemaker/

Amazon. (2021d). *AWS Glue*. https://aws.amazon.com/glue/

Amazon. (2021e). *AWS Lambda*. https://aws.amazon.com/lambda/

Becker, T., & Intoyoad, W. (2017). *Context Aware Process Mining in Logistics BT - 50th CIRP Conference on Manufacturing Systems, CIRP CMS 2017, May 3, 2017 - May 5, 2017*. *63*, 557–562. https://doi.org/10.1016/j.procir.2017.03.149

Bello, S. A., Oyedele, L. O., Akinade, O. O., Bilal, M., Davila Delgado, J. M., Akanbi, L. A., Ajayi, A. O., & Owolabi, H. A. (2021). Cloud computing in construction industry: Use cases, benefits and challenges. *Automation in Construction*, *122*, 103441. https://doi.org/10.1016/J.AUTCON.2020.103441

Blackhurst, J., Rungtusanatham, M. J., Scheibe, K., & Ambulkar, S. (2018). Supply chain vulnerability assessment: A network based visualization and clustering analysis approach. *Journal of Purchasing and Supply Management*, *24*(1), 21–30. https://doi.org/https://doi.org/10.1016/j.pursup.2017.10.004

Bueno, A., Godinho Filho, M., & Frank, A. G. (2020). Smart production planning and control in the Industry 4.0 context: A systematic literature review. *Computers and Industrial Engineering*, *149*. https://doi.org/10.1016/j.cie.2020.106774

Cadavid, J., Lamouri, S., Grabot, B., … R. P.-J. of I., & 2020, undefined. (2020). Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0. *Springer*, *31*. https://doi.org/10.1007/s10845-019-01531-7

Chandra, S., Varde, A. S., & Wang, J. (2019). A Hive and SQL Case Study in Cloud Data Analytics. *2019 IEEE 10th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2019*, 0112–0118. https://doi.org/10.1109/UEMCON47517.2019.8992925

Chowdhury, M., Rahman, A., & Islam, R. (2018). Malware analysis and detection using data mining and machine learning classification. *Advances in Intelligent Systems and Computing*, *580*, 266–274. https://doi.org/10.1007/978-3-319-67071-3_33

ER, M., Arsad, N., Astuti, H. M., Kusumawardani, R. P., & Utami, R. A. (2018). Analysis of production planning in a global manufacturing company with process mining. *Journal of Enterprise Information Management*, *31*(2), 317–337. https://doi.org/10.1108/JEIM-01-2017-0003

Gürel, E., & Tat, M. (2017). SWOT ANALYSIS: A THEORETICAL REVIEW. *Uluslararası Sosyal Araştırmalar Dergisi The Journal of International Social Research*. https://doi.org/10.17719/jisr.2017.1832

Jarebrant, C. (2016). *Ergonomic Value Stream Mapping:(ErgoVSM): Tool and User Guide*. https://books.google.com/books?hl=pt-BR&lr=&id=LaLDDQAAQBAJ&oi=fnd&pg=PA5&dq=ergo+value+stream+mapping&ots=22-Fkthxc4&sig=3NflgGE8VaySxHOG2PI95XaTNWo

Jupyter. (2021). *Jupyter Notebook*. https://jupyter.org/

Maurice, P., Malaisé, A., Amiot, C., Paris, N., Richard, G.-J., Rochel, O., & Ivaldi, S. (2019). Human movement and ergonomics: An industry-oriented dataset for collaborative robotics: *Https://Doi.Org/10.1177/0278364919882089*, *38*(14), 1529–1537. https://doi.org/10.1177/0278364919882089

Muller, A. C., & Guido, S. (2017). Introduction to Machine Learning with Python: A Guide for Data Scientists. In D. Schanafelt (Ed.), *O'Reilly* (First Edit, Vol. 1, Issue 3). O'Reilly Media. https://doi.org/10.3390/su10030702

Namdev, N., Agrawal, S., & Silkari, S. (2015). Recent advancement in machine learning based internet traffic classification. *Procedia Computer Science*, *60*(1), 784–791. https://doi.org/10.1016/j.procs.2015.08.238

Núñez-Merino, M., Maqueira-Marín, J. M., Moyano-Fuentes, J., & Martínez-Jurado, P. J. (2020). Information and digital technologies of Industry 4.0 and Lean supply chain management: a systematic literature review. *Https://Doi.Org/10.1080/00207543.2020.1743896*, *58*(16), 5034–5061. https://doi.org/10.1080/00207543.2020.1743896

Python. (2021). *Python*. https://www.python.org/

Ramachandra, G., Iftikhar, M., & Khan, F. A. (2017). A Comprehensive Survey on Security in Cloud Computing. *Procedia Computer Science*, *110*, 465–472. https://doi.org/10.1016/J.PROCS.2017.06.124

Ray, P. P. (2016). A survey of IoT cloud platforms. *Future Computing and Informatics Journal*, *1*(1–2), 35–46. https://doi.org/10.1016/J.FCIJ.2017.02.001

Sánchez-Fernández, A., Baldán, F. J., Sainz-Palmero, G. I., Benítez, J. M., & Fuente, M. J. (2018). Fault detection based on time series modeling and multivariate statistical process control. *Chemometrics and Intelligent Laboratory Systems*, *182*, 57–69. https://doi.org/10.1016/J.CHEMOLAB.2018.08.003

Sunyaev, A. (2020). Cloud Computing. *Internet Computing*, 195–236. https://doi.org/10.1007/978-3-030-34957-8_7

Valamede, L. S., & Lima, M. Z. T. de. (2019). Technological ergonomic innovations applied at the final sector of an automotive industry assembly line. *Revista Dos Trabalhos de Iniciação Científica Da UNICAMP*, *26*. https://doi.org/10.20396/revpibic262018864

Wang, Y., Caron, F., Vanthienen, J., Huang, L., & Guo, Y. (2014). Acquiring logistics process intelligence: Methodology and an application for a Chinese bulk port. *Expert Systems with Applications*, *41*(1), 195–209. https://doi.org/https://doi.org/10.1016/j.eswa.2013.07.021

Yadav, G., Luthra, S., Huisingh, D., Mangla, S. K., Narkhede, B. E., & Liu, Y. (2020). Development of a lean manufacturing framework to enhance its adoption within manufacturing companies in developing economies. *Journal of Cleaner Production*, *245*. https://doi.org/10.1016/j.jclepro.2019.118726