UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE MECÂNICA
CURSO DE ENGENHARIA MECÂNICA

MATHEUS VINÍCIUS DE CARVALHO

APLICAÇÕES DE *MACHINE LEARNING* NA ENGENHARIA MECÂNICA: UM ESTUDO DE CASO PARA DIAGNÓSTICO DA OPERABILIDADE DE SISTEMAS DE ABASTECIMENTO DE ÁGUA

TRABALHO DE CONCLUSÃO DE CURSO

PATO BRANCO
2021

MATHEUS VINÍCIUS DE CARVALHO

APLICAÇÕES DE MACHINE LEARNING NA ENGENHARIA MECÂNICA: UM ESTUDO DE CASO PARA DIAGNÓSTICO DA OPERABILIDADE DE SISTEMAS DE ABASTECIMENTO DE ÁGUA

Trabalho de Conclusão de Curso de graduação, apresentado à disciplina de Trabalho de Conclusão de Curso 2, do Curso de Engenharia Mecânica do Departamento Acadêmico de Mecânica – DAMEC – da Universidade Tecnológica Federal do Paraná – UTFPR, Câmpus Pato Branco, como requisito parcial para obtenção do título de Engenheiro Mecânico.

Orientador: Prof. Dr. Gilson Adamczuk Oliveira

PATO BRANCO

2021

FOLHA DE APROVAÇÃO

APLICAÇÕES DE MACHINE LEARNING NA ENGENHARIA MECÂNICA: UM ESTUDO DE CASO PARA DIAGNÓSTICO DA OPERABILIDADE DE SISTEMAS DE ABASTECIMENTO DE ÁGUA

Matheus Vinícius de Carvalho

Trabalho de Conclusão de Curso de Graduação apresentado no dia 19/08/2021 como requisito parcial para a obtenção do Título de Engenheiro Mecânico, do curso de Engenharia Mecânica do Departamento Acadêmico de Mecânica (DAMEC) da Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco (UTFPR-PB). O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora julgou o trabalho **APROVADO**.

_____
Prof. Dr. Giovanni Bratti
(UTFPR – Departamento de Mecânica)

_____
Prof. Dr. Paulo Rogerio Novak
(UTFPR – Departamento de Mecânica)

_____
Prof. Dr. Gilson Adamczuk Oliveira
(UTFPR – Departamento de Mecânica)
Orientador

_____
Prof. Dr. Bruno Bellini Medeiros
Responsável pelo TCC do Curso de Eng. Mecânica

*A Folha de Aprovação assinada encontra-se na Coordenação do Curso de Engenharia Mecânica

DEDICATÓRIA

Dedico esse trabalho a todos que me apoiaram no decorrer do curso, principalmente aos meus pais e minha irmã. Ao meu orientador Gilson Adamczuk Oliveira e meus aos amigos que estiveram presentes em todo o caminho.

AGRADECIMENTOS

Agradeço primeiramente aos meus familiares, que tornaram esse momento possível. À minha mãe por todo o suporte e ensinamento, que foram a base para suportar momentos difíceis e aproveitar os momentos bons. Ao meu pai por demonstrar toda a resiliência necessária para passar por momentos conturbados e tirar os melhores ensinamentos dos mesmos. À minha irmã por mostrar que certos elos jamais serão rompidos e podemos contar com a família em qualquer momento, seja para suportar as dificuldades ou para compartilhar as alegrias.

Agradeço também ao meu orientador, o professor Gilson Adamczuk Oliveira, que se dispôs a enfrentar a dificuldade de orientar um trabalho com tempo útil menor que o usual.

Aos amigos que, desde o começo, dedicaram tempo e paciência para transpor as mais diversas situações que foram vividas no decorrer dos anos, em especial ao Lucas, Ederson e William por todas as noites de distração, muita risada e aprendizado que certamente serão levados como base de conhecimento e vivência ao longo da vida.

EPÍGRAFE

“If you're committed enough, you can make any story work” (GOODMAN, Saul ; Breaking Bad, 2010)

“Se você se empenhar o suficiente pode fazer qualquer história resultar” (GOODMAN, Saul ; Breaking Bad, 2010)

RESUMO

CARVALHO, Matheus. Aplicações de machine learning na engenharia mecânica: um estudo de caso para diagnóstico da operabilidade de sistemas de abastecimento de água. 2021. 61 f. Trabalho de Conclusão de Curso – Curso de Engenharia Mecânica, Universidade Tecnológica Federal do Paraná. Pato Branco, 2021.

Este trabalho tem por objetivo aplicar conceitos e técnicas de *machine learning* na área de manutenção preditiva em engenharia mecânica. Diversos softwares têm sido desenvolvidos para auxiliar na extração de informação da grande quantidade de dados disponibilizada pelo *Big Data*. A escolha do *Orange Data Mining* foi feita visando aproveitar as características contidas no mesmo, principalmente por utilizar programação visual. O trabalho utiliza dados reais do governo da Tanzânia. Garantir água para a população da Tanzânia faz com que seja importante realizar uma previsão da operabilidade de bombas em postos de abastecimento de água. Dentre as técnicas de avaliação de acurácia do algoritmo de *machine learning* disponíveis, a utilizada foi a *classification accuracy* por possuir fácil entendimento e bons resultados. Tanto a limpeza quanto a classificação de variáveis foram feitas utilizando ferramentas presentes no *Orange Data Mining*. Os modelos de avaliação utilizados no *software,* decision *trees*, *random forests* e *gradient boosting*, modelos flexíveis que trabalham com diferentes tipos de variáveis e suportam não linearidade. Após ajustar um histórico de 47520 dados de treinamento e 11880 dados de teste, obteve-se uma acurácia de 78,8% utilizando o *random forests*, 77,6% utilizando o *gradiente boosting* e 75,7% com o método *decision trees*. O resultado final mostra que é possível fazer a previsão de operabilidade de bombas em postos de abastecimento com acurácia próxima de 77% em uma primeira análise, podendo alcançar melhores resultados com uma atualização de dados e otimização de variáveis inconsistentes.

Palavras-chave: *Machine Learning*; Estudo de Caso; Programas *Open Source* Para *Data Mining*; Bombas D'água na Tanzânia; Manutenção Preditiva.

ABSTRACT

CARVALHO, Matheus. Applications of machine learning in mechanical engineering: a case study for diagnosing the operability of water supply systems. 2021. 61 f. Course Completion Work – Mechanical Engineering Course, University Technological Federal of Paraná. Pato Branco, 2021.

This paper aims to apply concepts and techniques of machine learning *in* the area of predictive maintenance in mechanical engineering. Several software has been developed to assist in extracting information from the large amount of data made available by Big *Data.* The choice of *Orange Data Mining was* made in order to take advantage of the characteristics contained in it, mainly by using visual programming. The paper uses real data from the Tanzanian government. Ensuring water for tanzania's population makes it important to make a prediction of pump operability at water stations. Among the techniques of accuracy assessment of the machine learning algorithm *available,* the one used was the classification *accuracy* because it has easy understanding and good results. Both the cleaning and classification of variables were done using tools present in Orange Data *Mining.* The evaluation models used in *software,* decision *trees,* random *forests* and gradient *boosting,*flexible models that work with different types of variables and support nonlinearity. After adjusting a history of 47520 training data and 11880 test data, an accuracy of 78.8% was obtained using random *forests,*77.6% using the *boosting gradient* and 75.7% using the decision trees *method.* The final result shows that it is possible to forecast the operability of pumps in filling stations with accuracy close to 77% in a first analysis, being able to achieve better results with a data update and optimization of inconsistent variables.

Keywords: *Machine Learning;* Case Study; Open-Source *Programs* for Data *Mining*; Water pumps in Tanzania; Predictive maintenance.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTAS DE ABREVIATURAS E SIGLAS

ML	*Machine Learning*
DS	*Data Science*
DM	*Data Mining*
CA	*Classification Accuracy*
GB	*Gradient Boosting*
RF	*Random Forest*

SUMÁRIO

1 INTRODUÇÃO

A crescente inovação tecnológica fornece a possibilidade de acesso ao conhecimento praticamente ilimitado, mas o grande problema é fazer com que quantidades enormes de dados sejam processados e decifrados para virarem informações uteis para determinadas aplicações. Principalmente no presente século a evolução da quantidade de dados disponíveis para analise cresceu de forma rápida, com isso, os chamados cientistas de dados aproveitaram a abundância de dados, juntamente com a redução dos custos de armazenamento e processamento, para conseguir conhecimentos valiosos. *Data Science* poderia também ser uma soma de programação, matemática, estatística, *machine learning*, solução de problemas e a habilidade de preparar e alinhar os dados.

Machine Learning, ou aprendizado de máquina, é um termo que começou a ser utilizado no final dos anos 50, quando Arthur Samuel criou um sistema que conseguiu vencê-lo em um jogo de damas. A definição exata seria que *machine learning* é um campo de estudo que garante aos computadores a capacidade de aprendizado sem ter sido programado para determinada tarefa. Durante a evolução da capacidade de processamento dos computadores, a capacidade de aprendizado cresceu exponencialmente, pois o volume de dados disponível para análise também ganhou grande volume. Esse aprendizado foi inspirado no processo de aprendizado humano, onde elas aprendem de modo iterativo com os dados e isso permite que o sistema encontre informações escondidas nesses dados. Os algoritmos de aprendizagem podem ser divididos em três grandes categorias: aprendizagem supervisionada, não supervisionada e aprendizado de reforço, onde cada uma delas têm características que distinguem seus objetivos finais e modo de operação.

A obtenção desses dados é de grande importância, pois é nessa etapa que uma quantidade enorme de dados é separada por categorias, visando definir qual objetivo final cada tipo de dado pode conseguir. Esse tipo de detecção de dados é conhecido como *Data Mining*, onde algumas ferramentas e técnicas conseguem separar os dados em: agrupamentos, regras, árvores de decisão, hipóteses e outros. O grande objetivo com essa mineração de dados é conhecer melhor alguns padrões e motivações para que as decisões sejam tomadas de forma mais objetivas e assertivas de acordo com o objetivo final do processo.

Obviamente o mundo dos negócios tenta aproveitar ao máximo os pontos positivos que o Machine Learning pode trazer à empresa. Além das vantagens obvias que se espera de um sistema tão avançado tecnologicamente, como otimização do tempo e economia financeira, processos um pouco mais elaborados ajudam a tornar o sistema de produção cada vez mais eficaz e lucrativo. Automatização de tarefas, gerenciamento de dados, medição de risco, resolução de problemas, pesquisas, eficiência, prever tendências, segurança e previsão de mercado.

O grande desafio do ML no futuro será ter o envolvimento de pessoas que saibam como resolver problemas do mundo real, e é aí que a engenharia mecânica se encaixa perfeitamente, realizando modelos matemáticos que descrevem a física de cada problema. Por ter experiência teórica e prática com sistemas, sensores, materiais, térmica e fluidos, esses profissionais estão aptos a solucionar casos que necessitam de análises de escoamento, detecção de ineficiência e muito mais. Sendo assim, o DS reúne informações que são cruciais para que o profissional construa, atualize e otimize os modelos que serão posteriormente utilizados.

1.1 OBJETIVOS

1.1.1 Objetivo Principal

O objetivo principal desse trabalho é aplicar conceitos e técnicas de *Machine Learning* na área de manutenção preditiva no campo da Engenharia Mecânica.

1.1.2 Objetivo Específico

Para atingir o objetivo principal os seguintes objetivos específicos devem ser alcançados:

I. Aplicar um modelo de *Machine Learning* visando realizar manutenção preditiva em bombas de postos de abastecimento de água;
II. Apresentar um estudo de caso, mostrando e explicando o funcionamento de programas de programação;

III. Descrever e comparar os resultados provenientes do estudo de caso;
IV. Aplicar algoritmos de árvores (*trees*) para exemplificar aplicações de *Machine Learning.*

1.2 JUSTIFICATIVA

Nos últimos anos a procura por inovação tecnológica, juntamente com a otimização do tempo e melhoria no processo de produção fez com que empresas buscassem formas de melhorar os resultados e diminuíssem os desperdícios. A grande quantidade de dados gerados diariamente por todas as redes ao redor do mundo tornou-se uma ótima alternativa para buscar o conhecimento necessário para tal evolução. O *Machine Learning* é uma ferramenta que tem potencial para gerar níveis de aprendizado em larga escala, sem deixar de lado a rapidez e precisão que o mercado exige.

1.3 ESTRUTURA DO TRABALHO

No primeiro capítulo, são tratados a contextualização do problema, objetivos e justificativa do trabalho.

No capítulo dois, é detalhada a revisão bibliográfica, apresentando os principais pontos do trabalho, explicando termos como *Data Science*, *Big Data* e *Machine Learning*, assim como seu método de utilização e vantagens. Ainda, uma análise exploratória de aplicações de *Machine Learning* é apresentada.

O capítulo três exibe a metodologia utilizada para desenvolvimento do trabalho, desde a apresentação do problema proposto no estudo de caso até a limpeza de dados e escolha dos mesmos.

O quarto capítulo trata dos resultados obtidos, juntamente com uma avaliação e discussão sobre a precisão encontrada na simulação.

O quinto capítulo mostra as considerações finais deste trabalho, abordando os pontos mais significativos seguindo a metodologia prevista em capítulos anteriores.

2 REVISÃO BIBLIOGRÁFICA

2.1 *DATA SCIENCE*

De acordo com Filatro (2020) *Data Science*, ou Ciência de Dados, é a validação, análise e criação de significado para um conjunto de dados com o objetivo de extrair conhecimento de conjuntos que originalmente seriam grandes demais para análises tradicionais. Como forma de compreender os dados obtidos, o *data science* combina cinco componentes (propósito, pessoas, processos, plataformas e programabilidade) para criar um sexto componente (produto resultante). Dentre os 5 componentes iniciais, podemos destacar pessoas e propósito como itens mais direcionados para a área de aplicação, enquanto plataformas e programabilidade são mais voltados para a área tecnológica.

Para que esse grande conjunto de dados seja decifrado de forma correta e objetiva, o *data science* deve seguir algumas etapas que são: definição do problema, coleta de dados, preparação dos dados, exploração dos dados, modelagem dos dados, comunicação dos resultados e automatização da análise. Porém, normalmente essas etapas acabam não seguindo uma ordem e é necessário retornar à algumas etapas vistas anteriormente para otimizar o processo.

2.2 *BIG DATA*

De Dagi (2020), *Big Data* é definida como uma análise e interpretação de grandes volumes de dados que possuem grande variedade. Trabalhar com essa quantidade imensa de dados não é tão simples, e por isso, algumas soluções específicas são criadas para que os profissionais consigam trabalhar com informações não estruturadas a uma grande velocidade.

Para caracterizar os aspectos do *Big Data* utilizamos os cinco “Vs”: Volume, Variedade, Velocidade, Veracidade e Valor. Os três primeiros aspectos nos entregam informações relacionadas a grande quantidade de dados que devem ser analisados. A Veracidade é uma questão de confiabilidade e o Valor é relacionado aos benefícios que tais soluções conseguem trazer para uma determinada empresa.

2.3 *MACHINE LEARNING*

De acordo com Filatro (2020), da maneira mais direta e simples possível, *machine learning* é quando a máquina usa de algoritmos para coletar dados, aprender com parte deles e fazer alguma análise ou determinação específica que lhe foi dada. Em 1959, Arthur Samuel, criador do termo "*machine learning*" descreveu o conceito como sendo "um campo de estudo que dá aos computadores a habilidade de aprender sem terem sido programados para tal", Arthur foi também o pioneiro da inteligência artificial.

Com a imensa quantidade de dados obtidos pelo *Big Data*, as oportunidades que o *Machine Learning* entrega são muitas. Os algoritmos utilizados, que são cada vez mais evoluídos, fazem a varredura dessas informações buscando padrões que possam melhorar o sistema de atuação. Normalmente o método tradicional da programação de um software é feito com base em um conjunto de regras que geram respostas a partir do processamento dos dados introduzidos. Aqui no *Machine Learning* os algoritmos são criados a partir de dados que serão analisados e os resultados que esperamos dessa análise, no fim do processo o sistema cria as próprias regras. Com esse sistema, um aplicativo ou *software* com *Machine Learning* melhora automaticamente e gradualmente com o número de utilizações.

Os tipos de aprendizagem desses algoritmos são divididos em supervisionada e não supervisionada. A aprendizagem supervisionada começa com o algoritmo recebendo dados que contêm a resposta certa, sendo assim, é uma modalidade que é usada para treinar o algoritmo. Com a aprendizagem não supervisionada os dados não têm rótulos, o que faz com que os efeitos de cada variável não sejam previstos, e então, os resultados estão ligados com os padrões que são encontrados nos próprios dados. Porém, não somente com esses algoritmos que o ML alcança o resultado esperado e, para isso, alguns métodos também podem ser utilizados.

2.3.1 Classificação

Mencionado por Bianchi (2020), o processo de classificação tem por objetivo identificar algumas características presentes nos dados, para melhor aproximação de determinada classe. De forma um pouco mais simples, teríamos uma

classificação binária do tipo “sim” e “não”. Porém, em vários problemas de tomadas de decisões somente a previsão de classificação não é o suficiente, e precisaríamos também de uma pontuação que represente a probabilidade de um indivíduo pertencer a essa determinada classe.

2.3.2 Regressão Linear

A regressão linear é basicamente uma equação que utilizamos para determinar um valor esperado de uma variável ‘y’ em função de uma outra variável ‘x’, de acordo com Filatro (2020). É uma ferramenta utilizada para entender a relação entre alguns diferentes fenômenos com o objetivo de prever comportamentos de uma variável desconhecida tendo dados estatísticos de outra variável. Quando temos problemas envolvendo uma variável dependente e uma variável independente, chamamos esse caso de regressão linear simples. Caso o problema estudado envolva duas ou mais variáveis independentes, esse modelo passa a ser conhecido como regressão linear múltipla.

Para ser feita uma análise da relação de duas variáveis, é realizado um diagrama de dispersão, que consegue nos mostrar se o grau de relacionamento entre as variáveis é forte ou fraco. Se a relação for forte, o gráfico se aproximará de uma reta. Porém, existem casos em que uma forte correlação entre dois ou mais fenômenos não confirme que a existência de um deles garanta a ocorrência do outro.

2.3.3 Clusterização

De acordo com Honda (2017), clusterização é uma técnica que tem por objetivo identificar semelhanças em um conjunto de dados para que eles sejam analisados com uma possível característica em comum. A grande diferença dessa técnica para a Classificação fica pelo fato de que na clusterização buscamos associar similaridade. A classificação da qualidade de um cluster é feita avaliando a homogeneidade de cada grupo especifico e sua diferença para os demais grupos. Se um grupo possui grande homogeneidade e elevada diferença para os demais grupos, teremos então um bom cluster.

2.3.4 Combinação por Similaridade

Filatro (2020) ainda indica que esse tipo de método é talvez o mais popular para fazer recomendações a partir da identificação de semelhanças por produtos ou serviços que uma pessoa tenha adquirido. Funciona de forma bem simples, onde os dados conhecidos são analisados para identificar alguma semelhança entre os mesmos.

2.3.5 Detecção de Anomalias

Esse tipo de análise é mais utilizado em situações onde o número de variáveis não é tão elevado, pois é uma tarefa onde identificar ocorrências raras em um conjunto de dados não é feita de forma simples, como apresenta Schiezaro (2020). Uma técnica utilizada para evitar esforço de visualização e a criação de alarmes falsos juntamente com a perda de importantes eventos é realizar o aprendizado não supervisionado.

2.3.6 Mineração de Relacionamentos

Essa técnica tem por objetivo descobrir relações de uma variável desejada com outra variável em um grande número de dados. Filatro (2020) ainda informa que essa técnica também consegue medir a força desse relacionamento.

2.3.7 Destilação de Dados para o Julgamento Humano

Essa técnica tem por objetivo descobrir relações de uma variável desejada com outra variável em um grande número de dados. Essa técnica também consegue medir a força desse relacionamento. O principal objetivo aqui é determinar se um evento pode causar um outro evento estudando a característica dos mesmos em meio aos dados.

2.3.8 Redução de Dimensionalidade

De acordo com Clésio (2015), em termos computacionais nem é preciso dizer que o aumento no número de dados faz com que os algoritmos tenham que processar um volume de dados muito maior. Nessa técnica o objetivo é tentar substituir um conjunto maior de informações por um número menor, para que seja mais fácil trabalhar com os dados e, possivelmente, seja mais fácil revelar informações importantes. Esse processo é geralmente realizado em uma etapa de pós-processamento de dados já preparados. Por ser um processo inviável de se fazer manualmente, são utilizados modelos de redução. Essa técnica geralmente envolve a perda de informações, justamente por isso, escolher os dados que serão representados é de extrema importância.

2.3.9 Perfilamento

Também conhecido como descrição de comportamento, o perfilamento tem por objetivo caracterizar o comportamento do indivíduo. Porém, essa descrição deve ser feita de forma minuciosa, avaliando o comportamento por dias, semanas, meses e anos para que a confiabilidade seja garantida. Com essa análise feita, uma predição de vínculo também pode ser realizada, para mapear ligações baseadas nas preferências observadas.

2.3.10 Modelagem Casual

A modelagem casual é utilizada para descobrir os fatores que realmente influenciam as pessoas. Para ser feita essa análise, geralmente fazemos experiências controladas para levantar dados e comparar as variáveis obtidas.

2.3.11 Análise de Redes Sociais

Essa análise é de extrema importância para entender a estrutura dos relacionamentos existentes entre pessoas, seja somente entre duas pessoas ou entre a população mundial, levando em consideração todos os tipos de preferências existentes. A coleta de dados pode ser feita de forma quantitativa e qualitativa, onde, a forma qualitativa pode ser feita através de entrevistas e questionários e as coletas quantitativas geralmente focam em bases de dados preexistentes, como informa Rosa (2018). O objetivo principal dessa análise é entender a força dos relacionamentos entre entidades e detectar grupos mais conectados entre si.

2.4 APLICAÇÕES DE *MACHINE LEARNING* NA ENGENHARIA MECÂNICA

O grande problema do ML em todas as áreas de aplicação é se sua implementação será relevante para o processo e trará resultados expressivos para a mesma, como explica Vdma (2018). O ML oferece oportunidades sem precedentes para a Engenharia Mecânica para otimizar processos de produção, tanto para operadores, onde o sistema pode ser atualizado com o objetivo de facilitar a utilização, quanto para o próprio maquinário com relação a manutenções e outros serviços. Para que se faça valer o investimento, é importante que os benefícios econômicos sejam previamente analisados e quantificados.

2.4.1 Visão de máquina

Fazer a análise e o julgamento de texturas e superfícies é uma tarefa de extrema dificuldade onde os atuais sistemas de processamento chegaram ao seu limite, enquanto o olho humano tem a capacidade de reconhecer texturas, padrões objetos e estruturas e classificar os mesmos de forma precisa e confiável com pouco treinamento. De acordo com Vdma (2018), alguns tipos de sensores podem ser utilizados para que a máquina seja capaz de fazer uma análise tão precisa quanto a citada anteriormente, como por exemplo sistemas 2D, 3D, ultrassom, raios-x e forma de sombreamento.

Como mostra Homem; Cardoso; Lourenço (2019), algoritmos que utilizam conceitos de ML, quando expostos a um conjunto de dados, podem distinguir o “bom” do defeituoso.

2.4.2 Controle Adaptativo para Otimização do Processo

Sistemas técnicos, onde o comportamento é influenciado por inúmeras variáveis, são difíceis de se modelar com fórmulas físicas, o que pode inviabilizar o processo de otimização. O ML pode ser muito útil nesse caso, visto que o comportamento do sistema pode ser entendido e, consequentemente, previsões sobre o mesmo podem ser realizadas. De acordo com Vdma (2018), o tempo morto pode ser superado com base nos controles adaptativos, mas isso exige uma compreensão profunda de todo o processo de produção.

Homem; Cardoso; Lourenço (2019), ainda diz que o gerenciamento de recursos é uma outra força dos algoritmos baseados em ML, onde é citado um exemplo da otimização de consumo de energia realiza pela empresa Google, resultando em uma redução de até 40% em suas contas de energia elétrica.

2.4.3 Smart Tendering

Nos últimos anos, a personalização de produtos e máquinas fez com que a variedade e complexidade de versões dos mesmos aumentasse desproporcionalmente, como indica Vdma (2018). O grande problema é que essa variedade pode trazer uma confusão muito grande entre fabricantes e clientes, assim como uma dificuldade de registrar cada máquina e produto, o que pode acarretar demora na entrega e, consequentemente, perda de produtividade.

2.4.4 Inovação Baseada em Dados

Para que uma análise industrial seja feita, a eficácia geral do sistema é determinada, isso faz com que seja descoberto o potencial de otimização de um

sistema. De acordo com Vdma (2018), utilizando algoritmos de ML os problemas podem ser analisados e resolvidos, caso o problema não tenha solução, ele deve, ao menos, ser reconhecido com antecedência.

2.4.5 Manutenção Preditiva

Por meio de algoritmos computacionais, o ML permite o monitoramento preditivo, prevendo falhas de equipamentos antes que elas ocorram e agendando a manutenção em tempo hábil, como explica Homem; Cardoso; Lourenço (2019). Com isso, em alguns casos, já se observa o sistema de planejamento e manutenção com uma eficiência 20% maior com relação aos sistemas atuais, isso acontece pois os sistemas utilizados de ML conseguem uma precisão superior a 90%.

Alguns sistemas autônomos utilizados em carros já utilizam tal ferramenta para monitorar sua própria condição para decidir se precisam de manutenção.

2.5 PROGRAMAS *OPEN SOURCE* PARA *DATA MINING*

A mineração de dados pode ser tida como a parte central da descoberta do conhecimento no conjunto de dados, sendo assim, é importante que uma preparação seja feita para extrair o máximo de informação possível. Depois que essa preparação é feita, os modelos serão construídos dependendo do tipo da pesquisa a ser realizada. Para facilitar o processo de análise que é bastante complexo, diversas ferramentas têm sido desenvolvidas nos últimos anos, sendo que algumas têm por objetivo oferecer uma alternativa gratuita aos pesquisadores interessados.

2.5.1 RapidMiner

De acordo com Jovic; Brkić; Bogunović (2015), *RapidMiner* é uma plataforma de *software* desenvolvida pela empresa *RapidMiner* que fornece um ambiente para preparação dos dados, aprendizado de máquina, aprendizado profundo, mineração de texto e análise preditiva. Normalmente é utilizado para

aplicativos comerciais, pesquisa, educação, treinamento, prototipagem rápida e desenvolvimento de aplicativos. Dentre as licenças disponíveis, está a *RapidMiner Studio Free Edition*, que é limitada a um processador lógico e dez mil linhas de dados.

O RapidMiner é escrito na linguagem Java, cada operador executa uma única tarefa dentro do processo e a saída de cada operador forma a entrada do próximo. Ele fornece também esquemas de aprendizado, modelos e algoritmos e pode ser estendido usando scripts R e *Phyton*.

Como indica Belge Consultoria ([s.d.]), a plataforma é completa e integrada para a coleta e preparação de dados e possui um ambiente de programação visual com mais de 1500 funções de ML, além disso, possui funções como: correlação e regras associativas, agrupamento por significados K, análise discriminante, regressão, árvores decisórias e muitas outras.

2.5.2 R

Ainda de acordo com Jovic; Brkić; Bogunović (2015), a ferramenta de código de linguagem de programação, R, também é uma excelente escolha para DM. O código é escrito em C++, Fortran e no próprio R. Para os usuários de DM, R oferece uma implementação muito rápida de muitos algoritmos do ML, comparáveis em número ao *RapidMiner* e ao *Weka*. Possui também tipos de dados específicos para manipulação de *big data*, fluxos de dados, mineração na web, mineração gráfica, mineração espacial e outras tarefas avançadas, mas suas principais características estatísticas incluem modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação e agrupamento.

Segundo Didatica, (2019a), os *scripts* possuem as seguintes categorias: variáveis, onde o utilizador pode salvar informações, funções, onde um conjunto de instruções executam uma ou mais tarefas, operadores, onde são feitas operações matemáticas, tipos de dados, que podem ser numéricos, caracteres e lógicos, estrutura de dados, que podem ser vetores, listas, matrizes ou dados *frames* e condicionais, que são “*if*, *for* e *while*”.

2.5.3 *Weka*

O *Weka* começou a ser escrito em 1993 na Universidade de *Waikato*, na nova Zelândia, e utiliza linguagem Java. O software é gratuito para fins não comerciais e teve sua popularidade estável ao longo dos anos por possuir uma interface amigável ao utilizador, assim como um grande número de algoritmos de DM implementados. De acordo com Jovic; Brkić; Bogunović (2015), o software não é tão popular quanto *RapidMiner* ou R pois possui recursos mais lentos e mais exigentes. O *Weka* oferece quatro opções para DM, sendo elas: Interface de linha de comando, explorar, experimentar e fluxo de conhecimento. O suporte para big data e aprendizado semi supervisionado é limitado e o aprendizado profundo ainda não é suportado.

Clésio (2012), diz também que o *Weka* possui uma série de algoritmos que são desenvolvidos pela comunidade, assim como uma grande flexibilização na utilização de suas técnicas de mineração.

2.5.4 *Orange*

De acordo com Unis (2018) o Orange é uma ferramenta voltada para análise e visualização de dados de código aberto, onde é possível extrair dados via programação visual ou scripts *Phyton*, além de explorar estatísticas, realizar *box plots* ou *scatter plots*, agrupamento hierárquico, *heatmaps* e projeções lineares. Jovic; Brkić; Bogunović (2015) ainda fala que o *Orange* Canvas oferece uma interface estruturada em funcionalidades agrupadas em nove categorias, que são: operações de dados, visualização, classificação, regressão, avaliação, aprendizagem não supervisionada, associação, visualização usando *widgets* e protótipo de implementações.

2.5.5 *Knime*

Produzido e mantido pela empresa suíça *Knime*, o programa começou a ser desenvolvido em 2004 e foi lançado em 2006. De acordo com a própria empresa e com Jovic; Brkić; Bogunović (2015), o software é usado por mais de três mil

organizações em mais de sessenta países. Unis (2018) ainda informa que a plataforma ajuda na manipulação, análise e modelagem de dados por meio de programação.

Indoworld ([s.d.]), ainda informa que o *Knime* possui uma ampla aplicabilidade, muitas integrações e, por possuir mais de 2000 nós, entrega uma grande variedade de ferramentas para ajudar na descoberta de possíveis novos conhecimentos.

2.5.6 Scikit-Learn

De acordo com Jovic; Brkić; Bogunović (2015) o *Scikit-Learn* é um programa gratuito, feito em *Phyton*, que expande as funcionalidades dos pacotes NumPy e SciPy com grandes algoritmos de DM, assim como usar o matplotlib para gerar gráficos, como indica Bormann (2016). Um dos pontos fortes do software é sua documentação online de algoritmos implementados, pois acaba se tornando um requisito importante para outros usuários. Os colaboradores são solicitados a otimizar o código em diversos aspectos, isso faz com que o sistema seja bastante rápido, mesmo sendo escrito em uma linguagem interpretada. O grande problema é que seu utilizador deve ser um programador habilidoso em Phyton por causa de sua interface de linha de comando, fazendo com que outros softwares sejam mais atraentes nesse aspecto.

O diferencial do *Scikit-Learn* é ser uma biblioteca criada especificamente para aplicações práticas de ML. De acordo com Didatica (2020a), possui também ferramentas simples e eficientes para análise preditiva de dados. Suas principais aplicações são: pós processamento, classificação, regressão, *clusterização*, redução de dimensionalidade e ajuste de parâmetros.

3 MATERIAIS E MÉTODOS

3.1 PROCEDIMENTOS METODOLÓGICOS

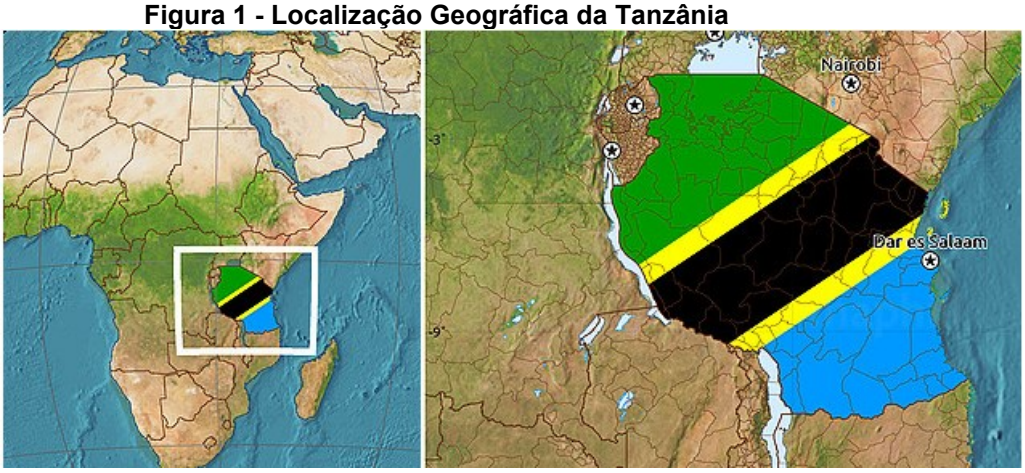
Uma das aplicações de *machine learning* na engenharia mecânica é quanto à manutenção preditiva. Dentre os métodos de manutenção utilizados, a manutenção preditiva se destaca pela confiabilidade e potencial retorno sobre o investimento, visto que há um acompanhamento de todo o equipamento através de dados coletados.

3.1.1 O problema: dados de poços na Tanzânia

A escassez de água é possivelmente o maior risco global em termos de impactos potenciais na próxima década de acordo com estudos recentes. O dado mais alarmante é que cerca de dois terços da população mundial vivem com grande escassez de água por pelo menos um mês por ano, sendo que meio bilhão de pessoas vivem com esse problema o ano todo.

“A República Unida da Tanzânia é um país da África Oriental. É um estado unitário composto por 26 regiões” Salles (2018).

Figura 1 - Localização Geográfica da Tanzânia



Fonte: Salles, 2018.

A Tanzânia é um país que possui um terço de suas regiões no clima árido ou semiárido, dificultando a tarefa de encontrar água potável. Desse modo, a utilização de águas subterrâneas é, provavelmente, a melhor alternativa para conseguir água. O grande problema é que o processo para conseguir acesso à essa água não é tão simples, o investimento inicial é relativamente alto e o sistema de manutenção deve ser eficiente. O sistema de manutenção eficiente pode garantir o abastecimento de água para milhares de pessoas. Existem diversos motivos para um ponto de abastecimento parar de funcionar, como: torneiras quebradas, bombas quebradas, canos quebrados, fonte de abastecimento danificada e até mesmo o roubo de bombas d'água.

Figura 2 - Ponto de Abastecimento de Água

Fonte: Salles, 2018.

Sendo assim, o grande objetivo é fazer com que o sistema de manutenção possa prever, com a utilização de dados reais, quando as bombas estão funcionando de forma correta, quando necessitam de manutenção e quando não estão em funcionamento.

3.1.2 Dados

Base para o presente trabalho, Salles (2018) apresenta os dados reais, obtidos por Taarifa e Tanzanian Ministry of Water, que foram utilizados nessa pesquisa. Os dados estão disponíveis em: (Salles, 2018b). Para que o sistema seja preciso e confiável os dados devem ser extremamente detalhados, garantindo que todos os problemas sejam levados em consideração na0020hora da decisão final. Sendo assim, o conjunto de dados disponibilizados é composto pelas seguintes informações:

- *id* – número de identificação do ponto de abastecimento;
- *amount_tsh* – quantidade de água disponível no ponto;
- *date_recorded* – data de gravação dos dados;
- *funder* – quem financiou o ponto de abastecimento;
- *gps_height* – altitude do ponto de abastecimento;
- *installer* – organização que instalou o ponto de abastecimento;
- *longitude* – coordenadas do GPS, longitude;
- *latitude* – coordenadas do GPS, latitude;
- *wpt_name* – Nome do ponto de abastecimento;
- *num_private* – sem informações sobre essa coluna;
- *basin* – localização geográfica;
- *subvillage* – localização geográfica;
- *region* – localização geográfica;
- *region_code* – localização geográfica (código);
- *District_code* – localização geográfica (código);
- *lga* – localização geográfica;
- *ward* – localização geográfica;
- *population* – população nos arredores do ponto de abastecimento;
- *public_meeting* – *true*/*false*;
- *recorded_by* – grupo que inseriu os dados;
- *scheme_management* – quem opera o ponto de abastecimento;
- *scheme_name* – quem opera o ponto de abastecimento;
- *permit* – se o ponto de abastecimento é permitido;

- *construction_year* – ano de construção do ponto de abastecimento;
- *extraction_type* – tipo de extração no ponto de abastecimento;
- *extraction_type_group* – tipo de extração no ponto de abastecimento;
- *extraction_type_class* – tipo de extração no ponto de abastecimento;
- *management* – como o ponto de abastecimento é gerenciado;
- *management_group* – como o ponto de abastecimento é gerenciado;
- *payment* – custos da água;
- *payment_type* – forma de pagamento;
- *water_quality* – qualidade da água;
- *quality_group* – qualidade da água;
- *quantity* – quantidade de água;
- *quantity_group* – quantidade de água;
- *source* – fonte do ponto de abastecimento;
- *source_type* – fonte do ponto de abastecimento;
- *source_class* – fonte do ponto de abastecimento;
- *waterpoint_type* – o tipo de ponto de abastecimento;
- *waterpoint_type_group* – o tipo de ponto de abastecimento.

E assim, o estado do ponto de abastecimento é descrito como:

- *functional* – funcionando normalmente;
- *non functional* – completamente parado;
- *functional needs repair* – funcionando, mas precisando de manutenção.

3.1.3 Limpeza de Dados

Dentre os dados obtidos, alguns problemas são encontrados, onde os mais relevantes são: informações incompletas e informações que não agregam valor. Isso acontece, pois, algumas informações constam somente para atualização de cadastro das bombas. Para as informações incompletas que não foram descartadas, as

lacunas serão completadas utilizando algumas funcionalidades presentes no *Orange* que serão apresentadas posteriormente.

As informações categóricas são outro problema, pois em qualquer estratégia para tratamento haverá uma perda de dados, que podem ser ou não de grande importância para o resultado final.

3.1.4 Seleção das variáveis para Análise

Observando o conjunto de dados, duas colunas serão removidas pois não possuem informações que agregam valor final ao estudo, são elas: *num_private* e *recorded_by*, onde, a primeira, possui todos os valores iguais a zero “0” e a segunda é totalmente preenchida com o nome do grupo que inseriu os dados.

O *Orange Data Mining* separa as variáveis em três grupos: *numeric*, *categorial* e *text*. No grupo *numeric*, os dados são obrigatoriamente numéricos, os *categorials* contém dados com valores distintos entre si e os *text* são informações em texto. Os grupos de dados numéricos e categóricos podem ser classificados em *feature*, *target*, *meta* e *skip*. Os *features* são variáveis dependentes, *target* é o alvo principal para resolução do problema, *meta* são dados considerados somente para visualização, já que não constam na programação e *skip* são dados que não são considerados.

Sendo assim, a divisão de variáveis, juntamente com a explicação direcionada ao programa é apresentada nos quadros 1 e 2:

Quadro 1 - Utilização das Variáveis na Programação

Variável	Analisada? (Sim/Não)	Justificativa
id	Não	Identidade do ponto de abastecimento, mas não qualifica se o ponto está operável ou não.
amount_tsh	Sim	Informação indispensável
Funder	Sim	Pode dar informações importantes, as instalações podem ser qualidades divergentes entre si.
gps_height	Sim	Informação necessária para definir o trabalho que cada ponto de abastecimento terá
Installer	Sim	As instalações podem ser qualidades divergentes entre si.
Longitude	Não	Informação que agrega valor ao texto do projeto
Latitude	Não	Informação que agrega valor ao texto do projeto
wpt_name	Não	Não agrega valor funcional, somente à título de curiosidade
Basin	Sim	Pode dizer se existirá ou não água disponível no ponto de abastecimento
Subvillage	Sim	É possível saber a quantidade de água necessária para um bom funcionamento
Region	Sim	É possível saber a quantidade de água necessária para um bom funcionamento
region_code	Não	Não agrega valor funcional, somente à título de curiosidade
district_code	Não	Não agrega valor funcional, somente à título de curiosidade
Lga	Não	Não agrega valor funcional, somente à título de curiosidade
Ward	Não	Não agrega valor funcional, somente à título de curiosidade
Population	Sim	É possível saber a quantidade de água necessária para um bom funcionamento
public_meeting	Sim	Informa sobre possíveis mudanças no manuseio do ponto de operação
scheme_management	Sim	Pode trazer informações sobre a qualidade de operação no ponto de abastecimento
scheme_name	Sim	Pode trazer informações sobre a qualidade de operação no ponto de abastecimento

Fonte: Autoria Própria (2021)

Quadro 2 - Utilização das Variáveis na Programação (Continuação)

Variável	Analisada? (Sim/Não)	Justificativa
permit	Sim	Informação necessária para garantir a qualidade final do serviço prestado
construction_year	Sim	O tipo de manutenção e materiais necessários pode mudar
extraction_type	Sim	É possível fazer previsões do tempo de utilização e necessidade de manutenção
extraction_type_group	Não	Complemento de informações, pode trazer dados à título de curiosidade
extraction_type_class	Não	Complemento de informações, pode trazer dados à título de curiosidade
management	Sim	É possível fazer predições a respeito da necessidade de manutenção
management_group	Sim	É possível fazer predições a respeito da necessidade de manutenção
Payment	Sim	Pode explicar a necessidade de manutenção imediata ou um funcionamento prolongado
payment_type	Não	Complemento de informações, pode trazer dados à título de curiosidade
water_quality	Sim	Informação extremamente importante para garantir uma maior vida útil do equipamento
quality_group	Não	Complemento de informações, pode trazer dados à título de curiosidade
Quantity	Sim	Informação extremamente importante para garantir uma maior vida útil do equipamento
quantity_group	Não	Complemento de informações, pode trazer dados à título de curiosidade
Source	Sim	É possível prever o equipamento necessário, bem como sua manutenção
source_type	Sim	É possível prever o equipamento necessário, bem como sua manutenção
source_class	Sim	É possível prever o equipamento necessário, bem como sua manutenção
waterpoint_type	Sim	É possível prever o equipamento necessário, bem como sua manutenção
waterpoint_type_group	Não	Complemento de informações, pode trazer dados à título de curiosidade
status_group	Sim	Foco principal do trabalho

Fonte: Autoria Própria (2021)

3.1.5 Algoritmos de *Machine Learning* Utilizados – Métodos Baseados em Árvores de Decisão

Para que todos os dados sejam manipulados da melhor forma possível, alguns métodos serão utilizados, onde todos têm relações entre si. Os métodos baseados em árvores são bastante populares por apresentar fácil utilização e interpretação. Contudo, quando em comparação com outros métodos como regressão, esses métodos não apresentam resultados de acurácia tão elevados. Sendo assim, o método de *random forest* e o *boosting* serão utilizados para garantir boa acurácia, sacrificando parte da capacidade de interpretação.

O *random forest* aproveita de todas as qualidades encontradas no *decision trees* para entregar um melhor entendimento no algoritmo estudado. O *random forest* faz com que seja possível analisar uma quantidade de dados muito maior fazendo uma combinação de árvores de decisão utilizando um método conhecido como método *ensemble*. De acordo com James et al. (2013), “abordagens envolvendo múltiplas árvores (*bagging*, *random forests* e *boosting*) frequentemente aumentam consideravelmente a acurácia, perdendo, porém, um pouco da capacidade de interpretação”. Existem outras opções, sendo a ênfase em *trees* uma escolha metodológica adequada ao escopo de um trabalho de graduação.

3.1.5.1 *Decision Trees*

De acordo com Didatica, (2020b), *decision trees*, ou árvores de decisão, são algoritmos de *machine learning* largamente utilizados, possuindo fácil entendimento e bons resultados em suas previsões. São também a base do funcionamento do *random forests*, que será explicado posteriormente. A estrutura básica da árvore de decisões é composta por “nós”, que são pontos de decisão, onde em cada um desses pontos haverá alguns caminhos a serem seguidos, esses caminhos são “ramos”. Em cada um dos nós é feita uma pergunta, onde as respostas podem ser “sim” ou “não” e estas levarão a caminhos diferentes. No processo de criação de uma árvore de decisão é extremamente importante que os pontos de

decisão sejam específicos e pontuais, para que as respostas entreguem dados utilizáveis.

Para realizar a construção de uma árvore de regressão duas etapas são seguidas: É feita a divisão do espaço de previsão X1, X2, ..., Xp em J regiões R1, R2, ..., Rj. Após isso, para cada observação que pertence a Rj é feita a mesma previsão, que é a média das previsões das observações de treino em Rj. Nesse caso, o objetivo é encontrar as regiões que minimizem o RSS (*Residual Sum of Squares*), que é dado por:

$$\sum_{j=1}^{J} \epsilon \sum_{i \epsilon Rj} (yi - \hat{}yrj)^2 \qquad (1)$$

Para realizar a divisão binária, primeiro é selecionado o espaço de previsão 'Xj' e o ponto de corte 'S' de modo que dividir o espaço leva a maior redução possível do RSS. A divisão dos espaços é feita de forma que:

$$R_1(j,s) = \{X | X_j < s\} \qquad (2)$$

$$R_2(j,s) = \{X | X_j \geq s\} \qquad (3)$$

E assim, o objetivo é encontrar valores de 'j' e 's' que minimizem a equação (4):

$$\sum_{i:x_i \epsilon R1(j,s)} (yi - {}^{yR1})^2 + \sum_{i:x_i \epsilon R2(j,s)} (yi - {}^{yR2})^2 \qquad (4)$$

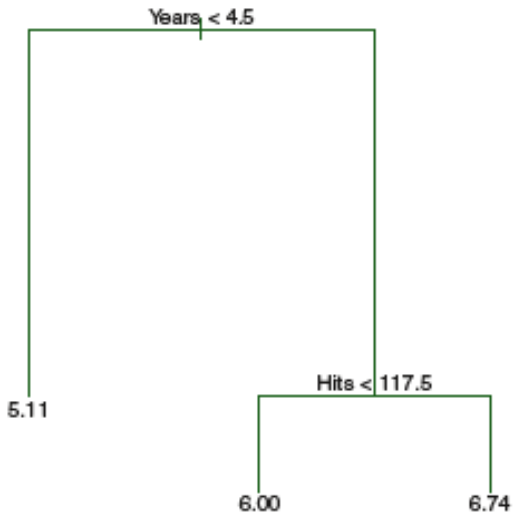
Em seguida, o processo é repetido, procurando melhorar o preditor e encontrar o melhor ponto de corte para diminuir o RSS. O processo continua até que o critério pré-definido seja alcançado.

Como exemplo para explicar melhor como funciona o método de *decision trees*, (JAMES et al., 2013) usa um estudo feito para determinar o salário de jogadores de um time de basebol, onde a primeira ramificação divide jogadores que tem mais de quatro anos e meio no time e jogadores que ainda não atingiram tal tempo. Dentre os jogadores que não possuem pelo menos esse tempo integrado ao time, o salário segue uma regra que já determina seu valor, já os outros jogadores passam por mais uma análise, que nesse caso é ver se o jogador atingiu pelo menos 117,5 rebatidas na temporada passada. O salário dos jogadores que possuem menos de 117,5

rebatidas é determinado por uma outra equação, enquanto os que possuem uma marca maior do que a informada passam por uma nova análise.

Uma grande vantagem da árvore de decisões é sua fácil visualização, como é visto na (Figura 3), apresentada por (James et al. 2013) em sua análise, aliada a uma possibilidade de “reengenharia” visto que podemos voltar o caminho feito para descobrir a fonte desejada.

Figura 3 - Esquematização realizada pelo método de Decision Trees para determinar o salário de jogadores de um time de Basebol



Fonte: JAMES et al., 2013

Para que a visualização fique ainda melhor, um esquema de cores será utilizado, onde quando o ponto de abastecimento estiver funcional, a cor será verde, quando estiver funcionando, mas necessitando de reparos, a cor será amarela e quando não estiver funcionando a cor será vermelha. Ainda de acordo com o sistema de cores, podemos observar que para valores com 100% de acerto por parte do programa, a cor terá maior intensidade, enquanto para valores menores essa intensidade vai reduzindo, isso acontece para que a visualização seja feita de forma mais simples e objetiva.

3.1.5.2 *Random Forests*

Didatica (2019b) diz que floresta aleatória, ou *random forest*, é um algoritmo de ML flexível e fácil de usar que entrega excelentes resultados. Como o próprio nome já diz, o algoritmo cria uma floresta de modo aleatório, essa floresta é uma combinação de árvores de decisão, essa combinação é conhecida como *ensemble*.

O método *ensemble* faz com que os algoritmos fiquem mais robustos e complexos, o que pode fazer com que o custo operacional se torne maior, mas também entrega um resultado mais preciso. Usualmente, quando se cria um modelo, o algoritmo que entrega o melhor desempenho é escolhido, sendo que alguns testes são feitos dentro deste algoritmo e, assim, alguns modelos diferentes são obtidos e, no fim do processo, apenas um modelo é escolhido. No método ensemble, quando esses modelos são criados a escolha não é de apenas um, mas sim de todos, fazendo com que o conjunto de dados seja muito mais amplo e preciso.

A seleção de dados é feita de forma diferente da árvore de decisões, aqui o algoritmo seleciona aleatoriamente algumas amostras dos dados, utilizando o *bootstrap*, que é um método de reamostragem onde as amostras selecionadas podem ser repetidas na seleção. Para a escolha da variável do próximo nó, outras variáveis são escolhidas, excluindo as já selecionadas anteriormente. Para a construção da próxima árvore, o processo se repete, mas provavelmente o resultado será outro, visto que as variáveis aleatoriamente escolhidas não devem ser as mesmas. A quantidade de árvores construídas fica por decisão do utilizador, lembrando sempre que quanto maior o número de árvores, melhores serão os resultados obtidos e maior será o tempo de criação do modelo.

Para construir a floresta, a cada divisão de árvore o algoritmo não pode considerar a maior parte do conjunto de dados. Normalmente, em cada divisão o número de preditores ‘m’ é aproximadamente a raiz de todos os preditores do conjunto ‘p’.

$$m \approx \sqrt{p} \qquad (5)$$

Essa técnica é utilizada para que, caso exista um preditor muito forte no conjunto de dados, esse preditor seja considerado em quase todas as árvores, fazendo com que as árvores sejam muito mais parecidas e assertivas.

3.1.5.3 *Boosting*

Tradicionalmente, construir um modelo de ML consiste em ensinar um único aluno para que ele se torne um indivíduo forte. Com os métodos de conjunto, a estratégia é treinar um grupo de alunos para que, juntos, se tornem mais fortes, essa estratégia é conhecida como grupo de alunos fracos. Boosting é uma das técnicas presentes nos métodos de conjunto. Essa técnica consiste em construir hipóteses sucessivas, para que exemplos classificados incorretamente sejam melhor classificados nas hipóteses seguintes, como indica Bianca (2010). Com isso, os "alunos fracos" aprendem a fazer previsões mais precisas em todos os tipos de dados, não somente os mais simples e fáceis. De um modo geral, as técnicas de ensemble ajudam a diminuir a variância, fazendo com que os dados sejam muito mais precisos e estáveis.

3.1.6 Métricas de Avaliação na Classificação

Ao desenvolver um projeto de ML é crucial a utilização de métricas apropriadas para cada situação. A precisão dos valores garante a qualidade do modelo, portanto, se forem mal escolhidas, será impossível atender aos requisitos necessários. Escolher uma métrica incorreta pode trazer prejuízos financeiros muito grandes, ou até pior, pode comprometer a saúde ou a vida de pessoas.

Nesse caso, a métrica utilizada será o *Classification Accuracy* (CA) (6), que diz quantos de nossos exemplos foram, de fato, classificados corretamente. A métrica é definida pela razão entre o que o modelo conseguiu acertar dentre todos os exemplos. O uso da CA é muito simples, de fácil uso e interpretável, e é dado por:

$$CA = \frac{Verdadeiros\ Positivos + Verdadeiros\ Negativos}{Total} \qquad (6)$$

3.2 PROGRAMAÇÃO VISUAL (*OPEN SOURCE*) *ORANGE DATA MINING*

3.2.1 Orange

Criado em 1996, o *Orange* é um pacote de software de programação visual baseada em *Phyton* e C++ voltada para análise visual de dados, aprendizado de máquina e mineração de dados. A interface do *software* consiste em uma tela em branco onde o usuário insere *widgets* para criar um fluxo de trabalho, esses *widgets* podem ler dados, mostrar tabelas e gráficos, comparar algoritmos, visualizar elementos e etc.

Dentre os *widgets* disponíveis, alguns se destacam por maior utilização e variedade de funções, como por exemplo os de visualização comum, que podem entregar gráficos de diferentes formas e histogramas. Os algoritmos usados para classificação de aprendizado de máquina supervisionado são chamados de *classify*. Os *widgets* de regressão são algoritmos de aprendizado supervisionado para, como o próprio nome diz, regressão. Já o avaliar é utilizado para fazer validações, estimativa de confiabilidade e pontuações de métodos de previsão. O *widget* não supervisionado é utilizado para algoritmos de aprendizagem não supervisionada e técnicas de projeção de dados. Além dessas funções principais, o *software* conta também com alguns *widgets* de complemento com algumas funções que podem ajudar em determinada tarefa, dentre eles destacam-se os *widgets* que podem fundir diferentes conjuntos de dados, ensinar conceitos de aprendizado de máquina, análise de imagens, análise de gráficos e rede, processamento de linguagem natural e modelagem de séries temporais.

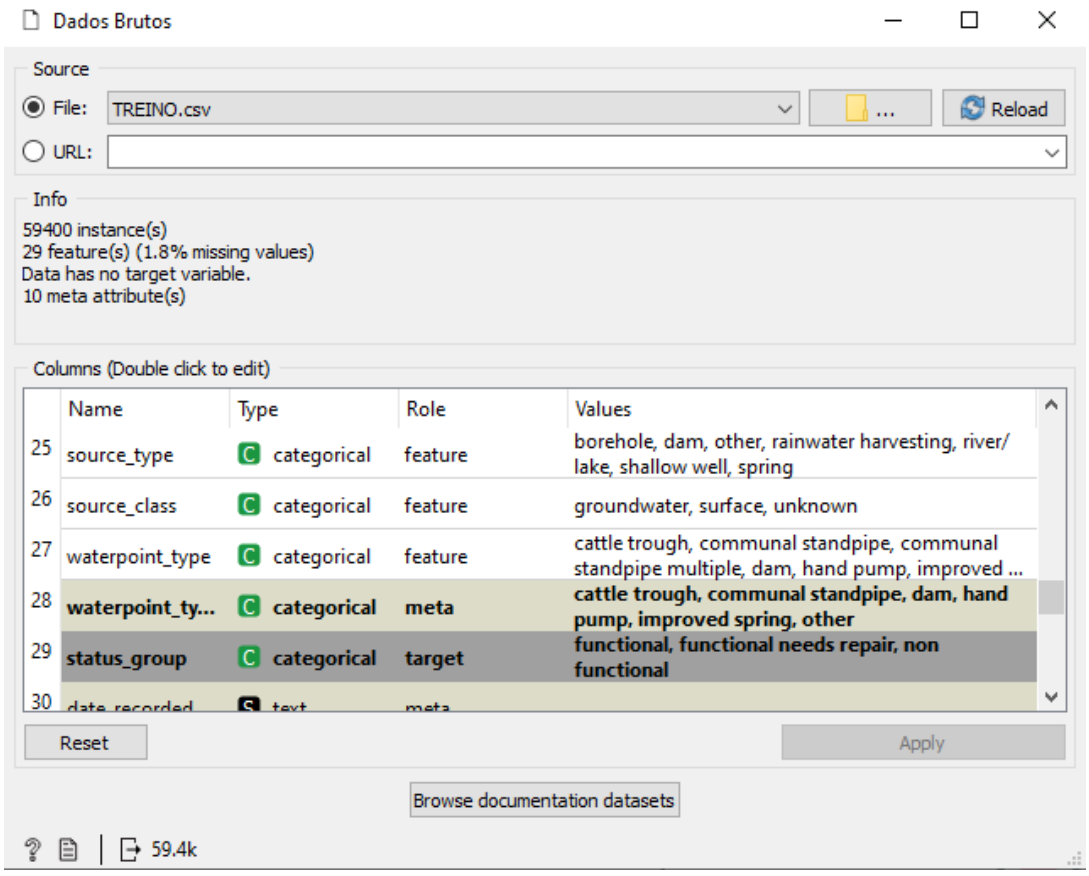
De acordo com Jovic; Brkić; Bogunović (2015), a interface de trabalho é visualmente atraente, o que oferece uma experiência agradável ao usuário. Por outro lado, o número de *widgets* é limitado quando comparado com outras ferramentas, mesmo assim, a cobertura de técnicas de mineração de dados é muito boa. Além disso, uma boa quantidade de *widgets* estão em desenvolvimento, significando que no futuro o conjunto de recursos deve ser expandido.

4 RESULTADOS E DISCUSSÕES

4.1 PRÉ-PROCESSAMENTO DE DADOS

A inserção do arquivo que possui 29 *features* (variáveis) e 59400 instâncias (pontos de abastecimento) é feita na ferramenta *Data Table* presente no *Orange*. É nesse momento que é feita a seleção de variáveis que serão levadas em conta nas análises seguintes. O grande problema nesse ponto é que o conjunto de dados disponibilizado possui 1,8% de valores sem informação, podendo fazer com que a análise perca precisão, como observado na Figura 4.

Figura 4 - Data Table – Dados Brutos

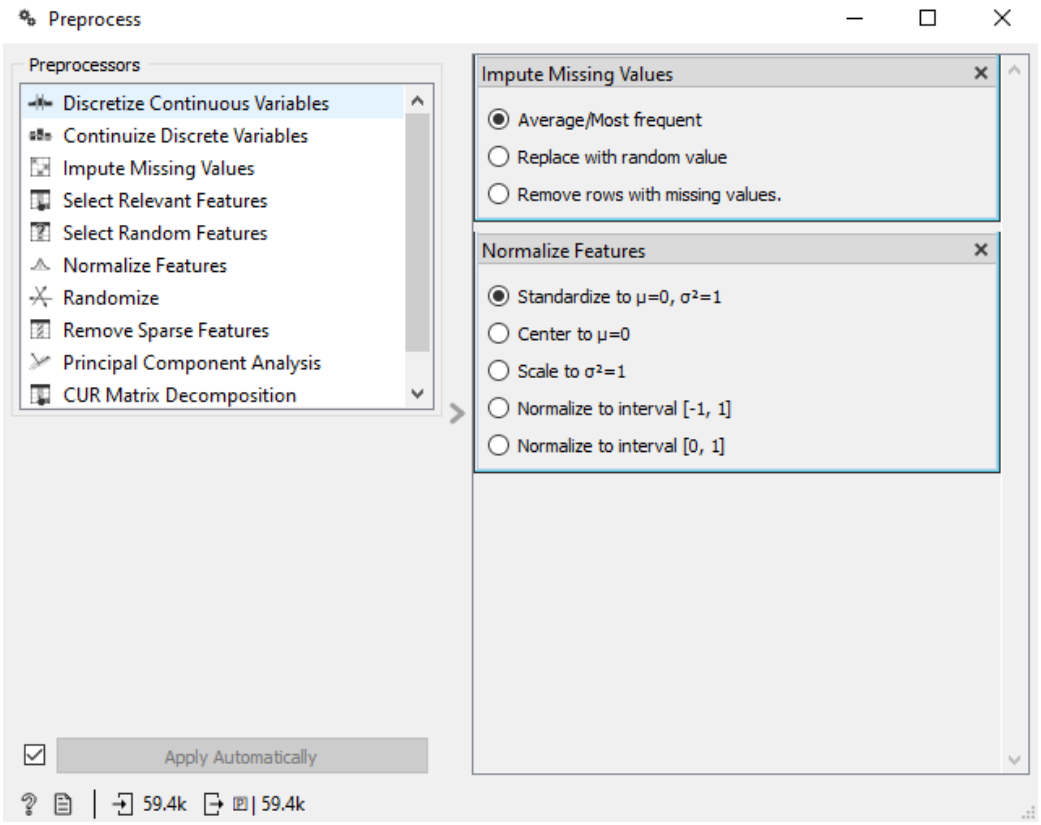


Fonte: Autoria Própria (2021)

A estratégia proposta foi utilizar a ferramenta de pré-processamento presente no *Orange*. Os valores sem informação serão completados com a média dos

valores contidos em cada coluna e, para melhorar a análise realizada, a normalização de variáveis também foi feita.

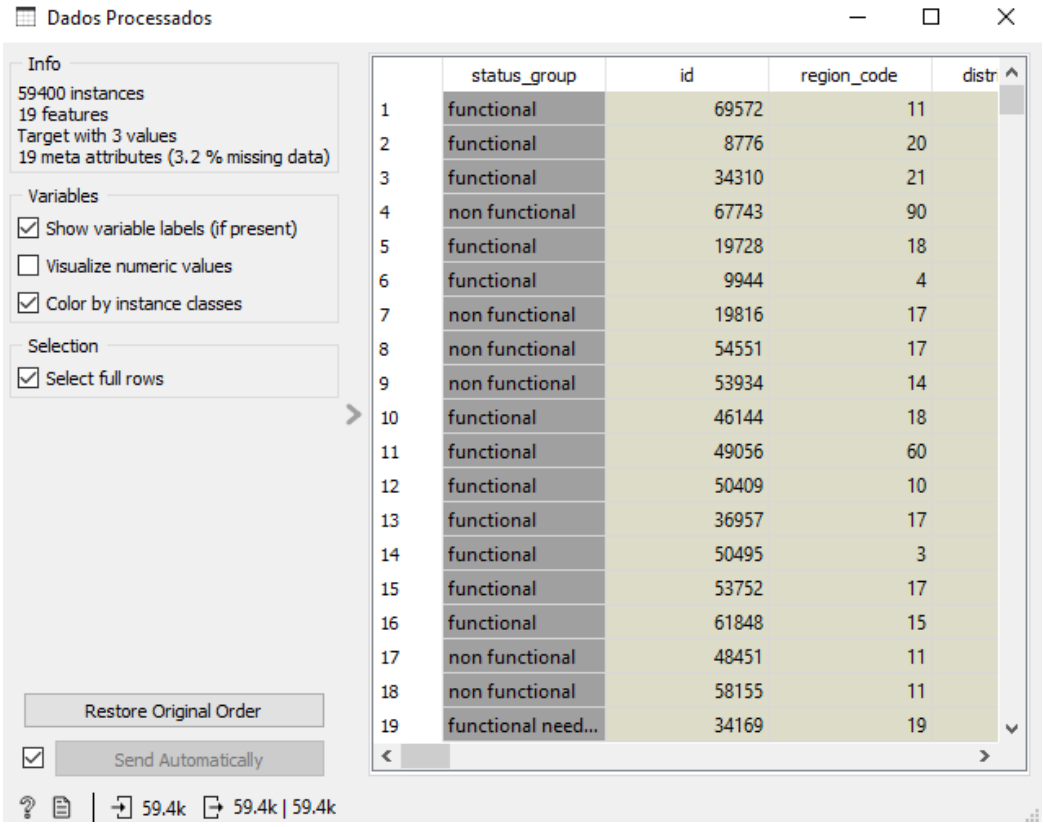
Figura 5 - Pré-processamento



Fonte: Autoria Própria (2021)

Nesse momento, das 29 variáveis iniciais, 19 serão levadas em consideração para a programação, com 0% de variáveis sem informação. As demais variáveis constam no software para agregar valor visual e à título de curiosidade.

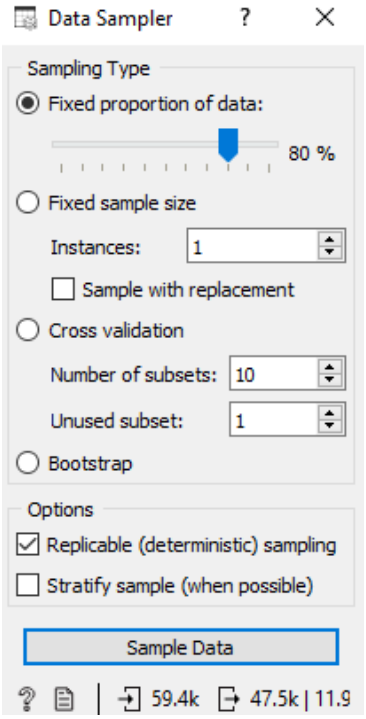
Figura 6 - Dados Processados



Fonte: Autoria Própria (2021)

Uma estratégia para garantir a veracidade da análise realizada é separar aleatoriamente o conjunto de dados em duas partes, treinamento e teste. O conjunto treinamento será submetido aos métodos de aprendizagem apresentados posteriormente, enquanto o conjunto teste passará pela ferramenta “Previsões”, onde será feita a comparação com os dados obtidos nos métodos de aprendizagem.

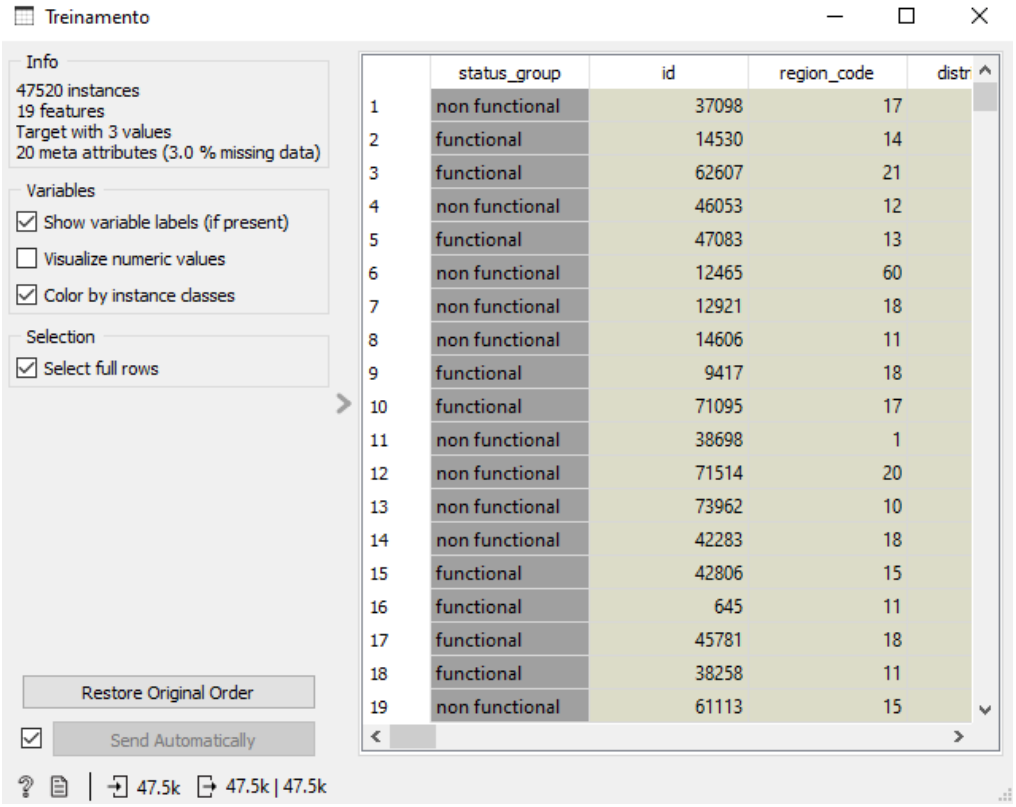
Figura 7 - Data Sampler



Fonte: Autoria Própria (2021)

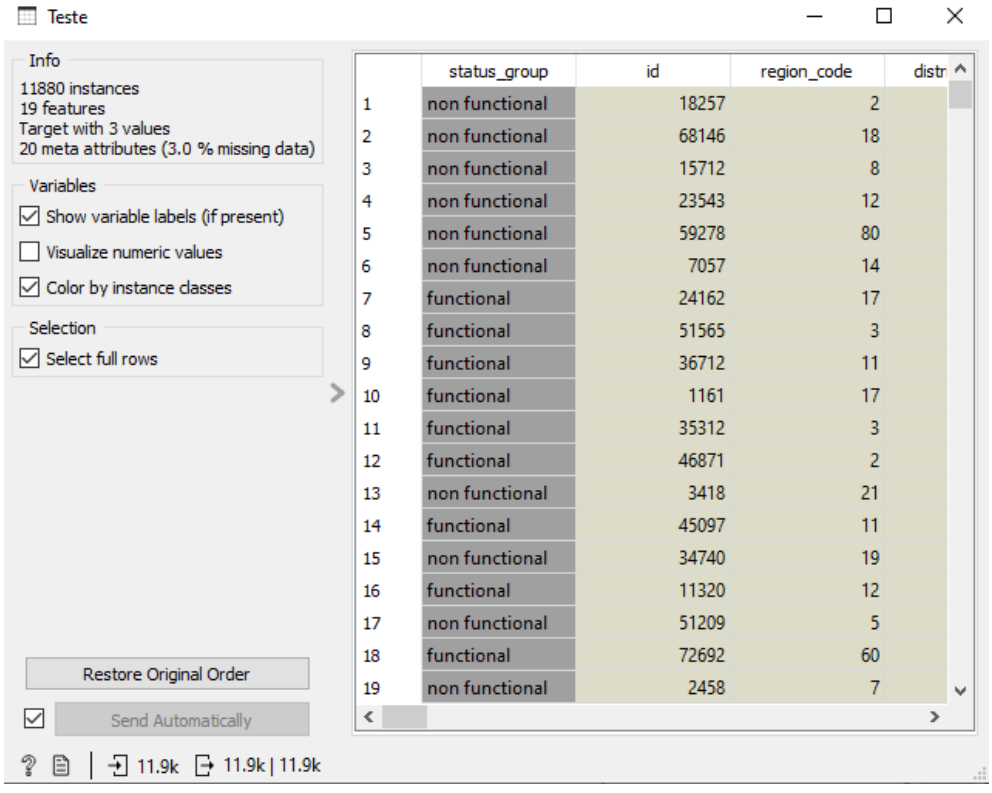
Como visto, a divisão escolhida foi de 80% dos dados para treinamento e 20% para teste, sendo assim, os dois conjuntos ainda possuem grande número de dados para que a análise final garanta boa precisão. O conjunto treinamento possui, nesse momento, 47520 instâncias e 19 variáveis, já o conjunto teste ficou com as 11880 instâncias restantes com as mesmas 19 variáveis.

Figura 8 - Data Table – Treinamento



Fonte: Autoria Própria (2021)

Figura 9 - Data Table - Teste



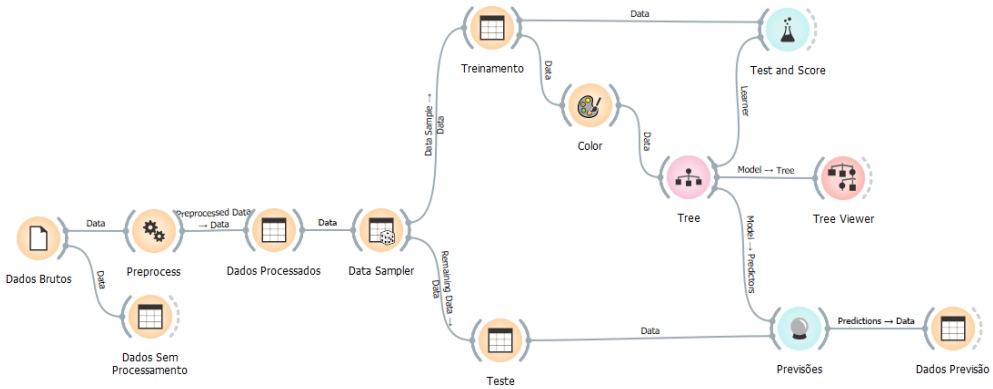
Fonte: Autoria Própria (2021)

4.2 MODELO NO ORANGE E RESULTADOS

O modelo final foi dividido em duas partes, uma com a árvore de decisões, visto na Figura 10, e outra com quatro métodos *random forests* e quatro *gradient boosting,* apresentado na Figura 13. Para que as métricas entre GB e RF sejam parecidas, a diferenciação ocorre pelo número de árvores em cada uma das análises, sendo B = 20, B = 40, B = 60 e B = 80. A decisão de não fazer a análise com um maior número de árvores ocorreu pela convergência de resultados, aliada a um tempo de processamento muito elevado por parte do programa.

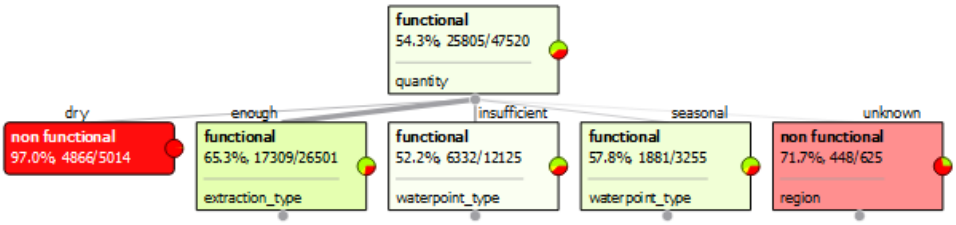
De acordo com a Figura 11, que é uma representação resumida da árvore obtida pelo modelo *tree*, consegue-se entender que em pontos de abastecimento que se encontram secos, quase a totalidade de bombas está sem funcionamento. Por outro lado, pouco mais de 65% dos pontos de abastecimento com água considerada suficiente estão funcionando, esse valor poderia ser maior, mas outras análises mais profundas foram avaliadas e não aparecem na representação simplificada.

Figura 10 - Layout Orange – Esquematização Realizada com o Decision Tree.



Fonte: Autoria Própria (2021)

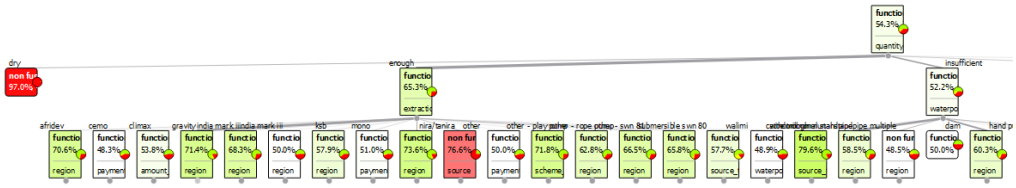
Figura 11 - Layout Orange –Tree Viewer (2 níveis)



Fonte: Autoria Própria (2021)

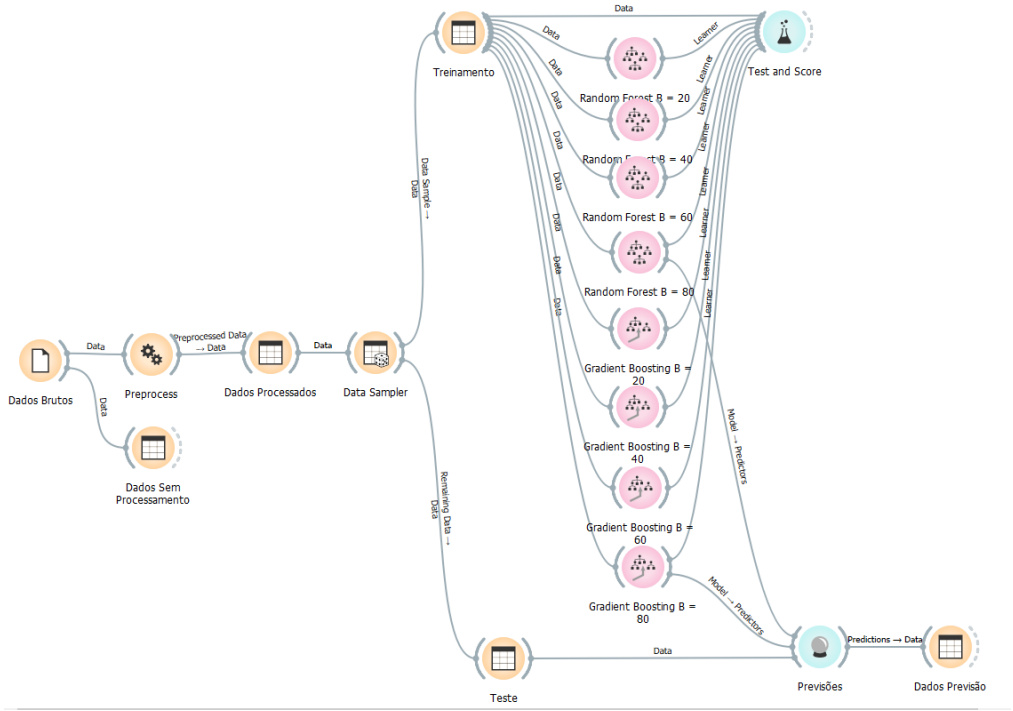
Na Figura 12 uma pequena parte da visualização em 3 níveis é apresentada, onde a localização da bomba já mostra regiões onde o abastecimento é comprometido por diferentes motivos. O modelo final conta com 10885 nós e 6156 folhas, isso faz com que seja inviável apresentar o modelo com imagens.

Figura 12 - Layout Orange –Tree Viewer (3 níveis)



Fonte: Autoria Própria (2021)

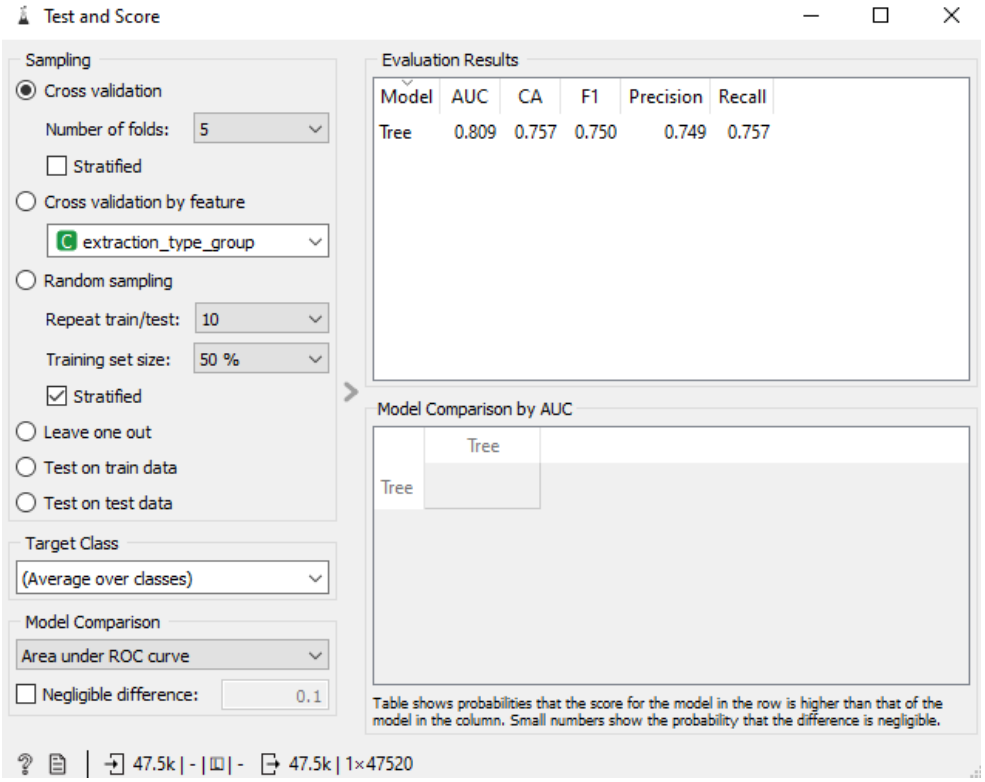
Figura 13 - Layout Orange – Esquematização Realizada com o Gradient Boosting e Random Forests



Fonte: Autoria Própria (2021)

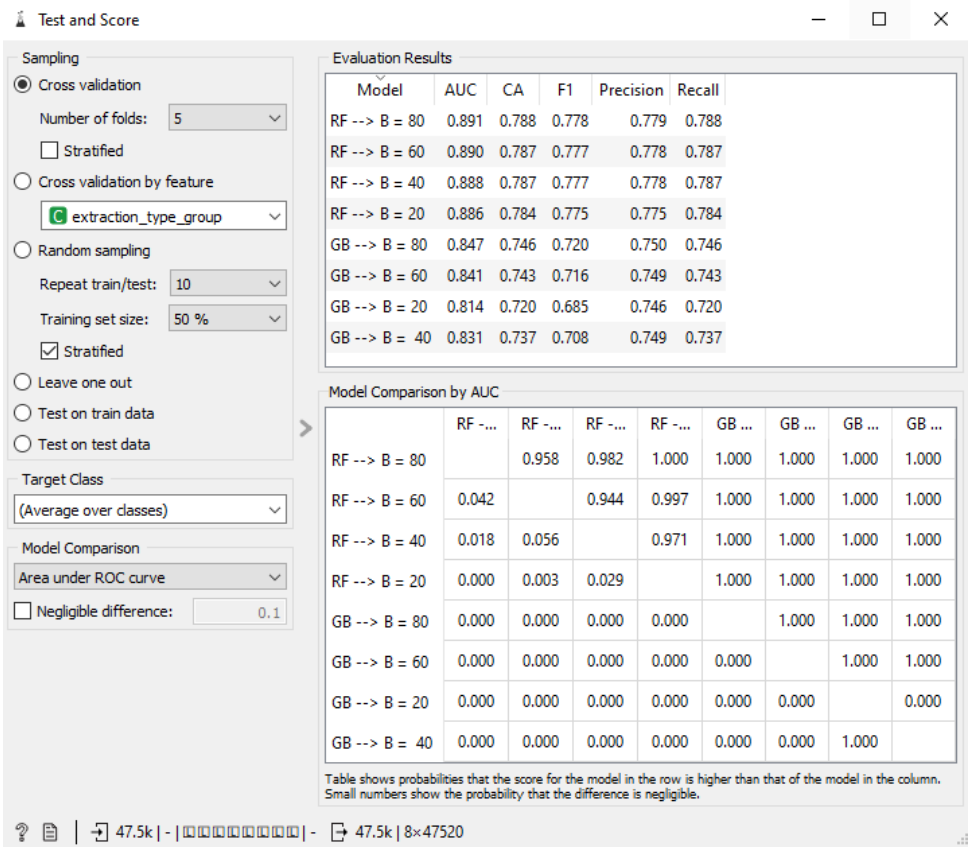
Após realizar a análise, o programa retornou os valores de *classification accuracy*. Nas Figuras 14 e 15 é possível ver que o modelo de decision trees apresenta boa acurácia e no modelo de *random forests* com 60 e 80 árvores o valor já foi o mesmo, portanto, não faria muito sentido aumentar drasticamente esse valor, visto que o tempo de processamento fica inviável para o estudo de caso.

Figura 14 - Teste Decision Tree Finalizado



Fonte: Autoria Própria (2021)

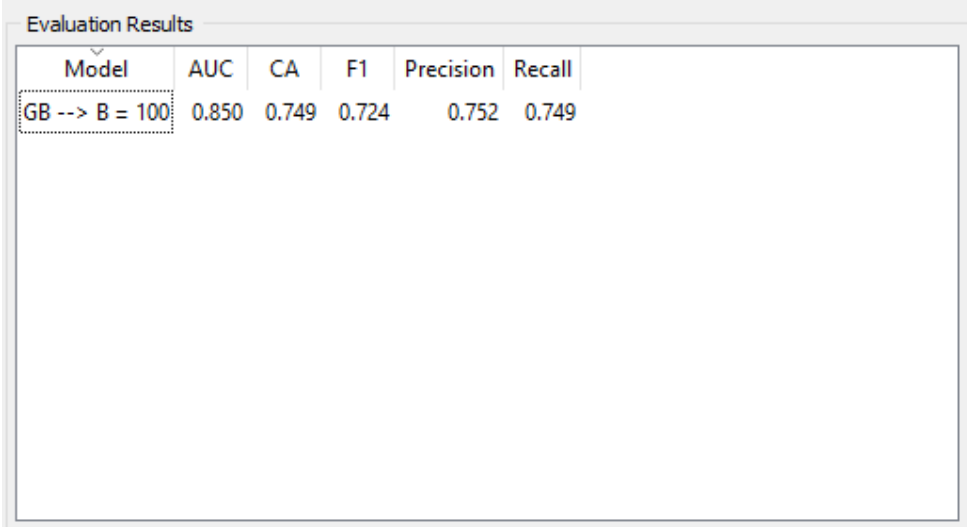
Figura 15 - Teste Gradient Boosting e Random Forests Finalizado



Fonte: Autoria Própria (2021)

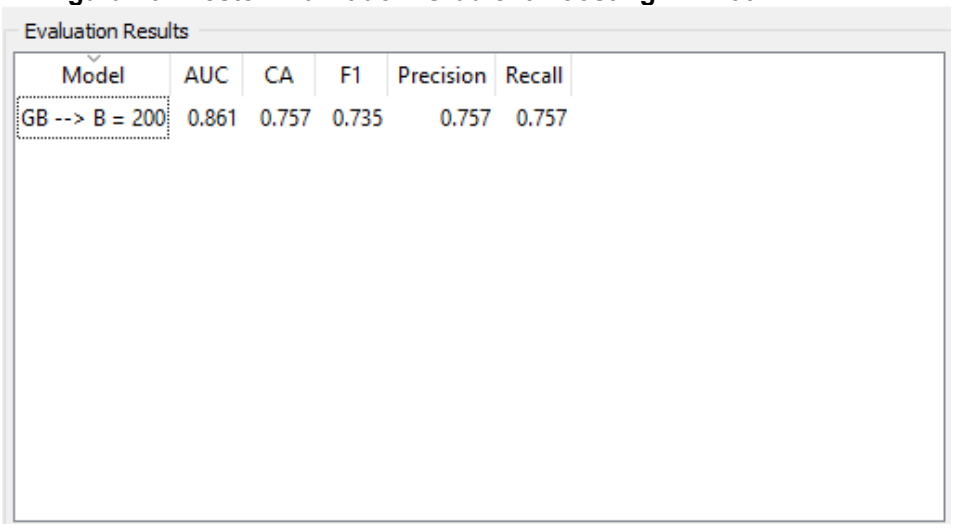
Como visto, no *gradient boosting* o valor de CA ainda aumenta com maior grandeza em comparação com o *random forest*, sendo assim, será feita uma nova análise, somente com o *gradient boosting* visando encontrar um valor mais preciso.

Figura 16 - Teste Finalizado – Gradient Boosting B = 100



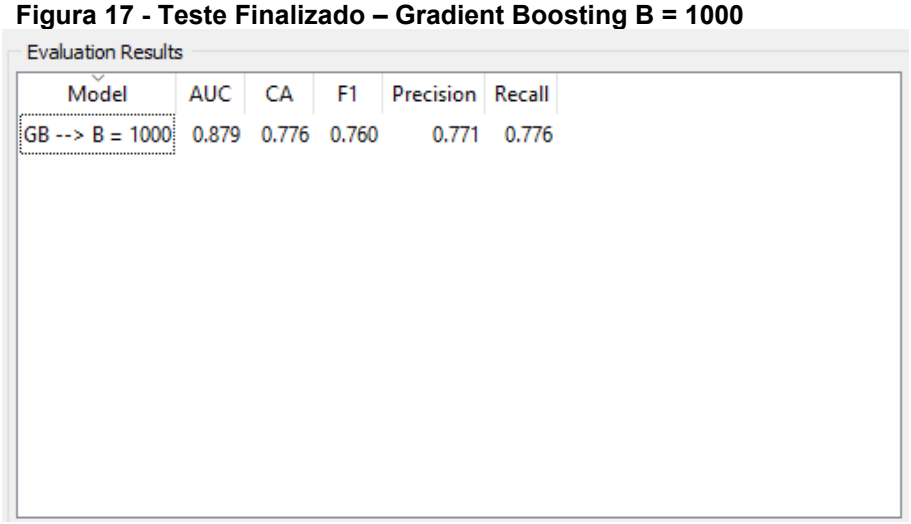
Fonte: Autoria Própria (2021)

Figura 16 - Teste Finalizado – Gradient Boosting B = 200



Fonte: Autoria Própria (2021)

Figura 17 - Teste Finalizado – Gradient Boosting B = 1000



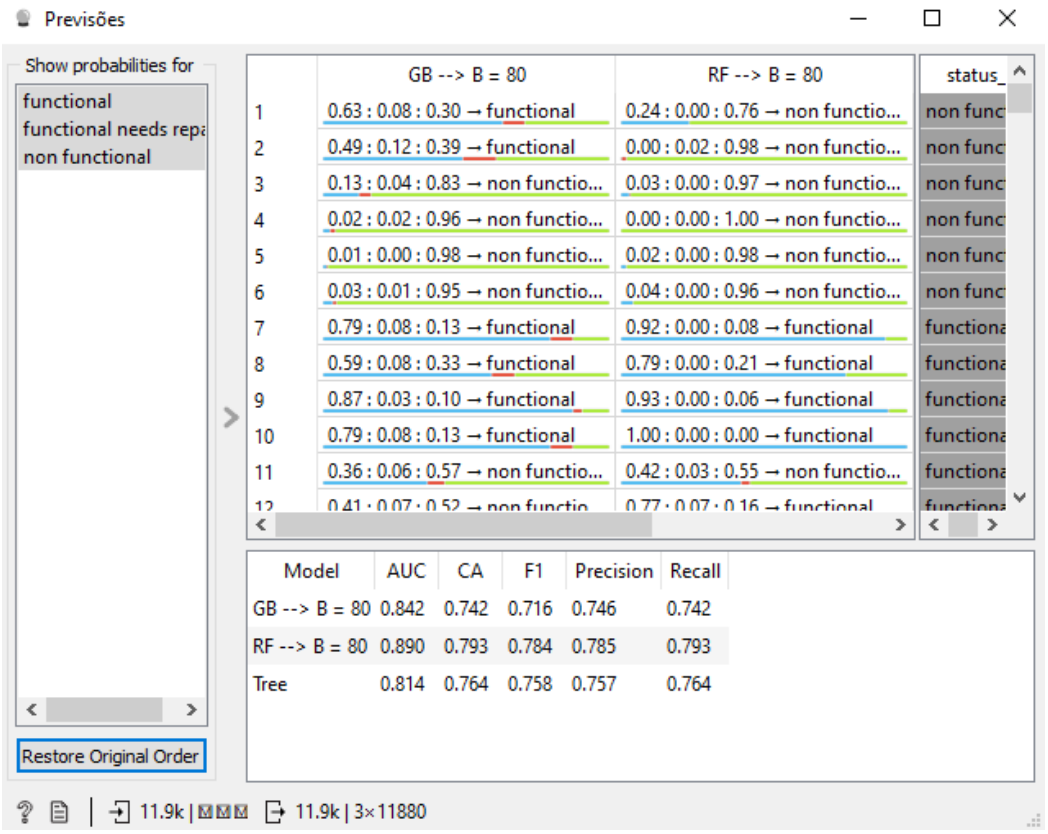
Fonte: Autoria Própria (2021)

As Figuras 16, 17 e 18 apresentam o *Classification Accuracy* do teste de *Gradient Boosting* para 100, 200 e 1000 árvores, respectivamente.

Com isso, é possível ver na Figura 17 que para 1000 árvores o valor de CA chega a 0,776, valor maior do que o encontrado no teste de 80 árvores (0,746), o grande problema é que no estudo em questão o equipamento demorou cerca de 55 minutos para realizar a análise, ficando inviável para a análise no presente estudo.

Para que a análise seja verificada, os valores encontrados na previsão devem ser próximos, para isso, somente as análises de *random forests* e *gradient boosting* com 80 árvores foram levadas em consideração, além da análise de *decision tree*.

Figura 18 – Previsões



Fonte: Autoria Própria (2021)

Como apresentado pelo programa, os valores de *classification accuracy* encontrados aumentam com um refino de precisão feito no mesmo, aumentando o tempo de processamento. Com CA média próxima à 77%, os valores são satisfatórios para uma primeira análise. Utilizando a ferramenta de previsões (Figura 19), conseguimos observar um valor de CA muito próximo aos encontrados nos modelos de avaliação, todos com uma diferença menor que 1%, com destaque para o método de *gradient boosting*, onde a diferença fica na casa de 0,5%.

5 CONCLUSÕES

Com evolução em larga escala nos últimos anos, o *machine learning* promove grande número de aplicações práticas em áreas distintas. A ideia de realizar um estudo de caso relacionando abastecimento de água e manutenção preditiva é de extrema importância. A realidade de boa parte da população mundial é diferente da observada na Tanzânia, onde a falta de água é uma preocupação iminente, sendo assim, a manutenção de bombas d'água visa garantir uma estabilidade de fornecimento de água para a maior parte da população.

Por tratar de dados reais e apresentar acurácia de quase 80%, os resultados mostram que é possível prever a operabilidade de cada ponto de abastecimento, sendo possível adequar as decisões de acordo com a necessidade vista em cada ponto.

A limpeza dos dados possibilita melhorar a acurácia do modelo, por esse motivo, é de extrema importância garantir dados precisos e coesos. No presente estudo, algumas variáveis poderiam ter sido ignoradas e uma atualização de informações é de grande necessidade para que os dados sejam utilizáveis no mundo real.

A programação com o *Orange Data Mining* é feita de forma objetiva e não é necessário ter conhecimento em linguagens de programação, o tempo de processamento varia de acordo com a configuração de cada equipamento utilizado, mas para análises com quantidades reduzidas de valores esse tempo se torna consideravelmente menor.

Uma grande vantagem da utilização de *machine learning* é a sua capacidade de aprendizado. Com o tempo, o número de dados vai se tornando cada vez maior, fazendo com que a previsão se torne cada vez mais exata e confiável.

5.1 SUGESTÕES PARA TRABALHOS FUTUROS

Uma atualização constante de novos locais de abastecimento e refino de informações potencialmente incoerentes farão com que a acurácia atinja valores mais confiáveis, sendo assim, para uma base de estudos futuros, é recomendável que os

dados sejam atualizados para anos mais recentes e o utilizador encontre ferramentas dentro do software escolhido para eliminar dados inconsistentes.

REFERÊNCIAS

BELGE CONSULTORIA. **Big Data, Analytics, Machine Learning, Ciência de dados, RapidMiner - Belge**. Disponível em: <https://www.belge.com.br/rapidminer.php>. Acesso em: 30 jun. 2021.

BIANCHI, A. **As classificações dos algoritmos de Machine Learning**. Disponível em: <https://www.viceri.com.br/insights/as-classificacoes-dos-algoritmos-de-machine-learning>. Acesso em: 30 abr. 2021.

BORMANN, F. **Aprendendo Scikit-Learn**. Disponível em: <https://medium.com/@felipebormann/aprendendo-scikit-learn-e-um-pouco-mais-de-python-6b27025f9d5b>. Acesso em: 26 abr. 2021.

CLESIO, F. **A utilização do WEKA como Minerador de Dados**. Disponível em: <https://flavioclesio.com/a-utilizacao-do-weka-como-minerador-de-dados>. Acesso em: 30 jun. 2021.

CLÉSIO, F. **7 técnicas para redução da dimensionalidade – Data Mining / Machine Learning / Data Analysis**. Disponível em: <https://mineracaodedados.wordpress.com/2015/06/13/7-tecnicas-para-reducao-da-dimensionalidade/>. Acesso em: 30 abr. 2021.

DAGI. **BIG DATA | Diretório Acadêmico de Gestão da Informação (DAGI) UFPE**. Disponível em: <https://sites.ufpe.br/dagi/2020/10/07/big-data/>. Acesso em: 30 abr. 2021.

DIDATICA. **Tudo sobre a Linguagem R: O que é, vantagens e como aprender**. Disponível em: <https://didatica.tech/a-linguagem-r/>. Acesso em: 30 jun. 2021a.

DIDATICA. **Entenda como funciona o Random Forest (Machine Learning)**. Disponível em: <https://didatica.tech/o-que-e-e-como-funciona-o-algoritmo-randomforest/>. Acesso em: 30 jun. 2021b.

DIDATICA. **A biblioteca scikit-learn – Pyhton para machine learning**. Disponível em: <https://didatica.tech/a-biblioteca-scikit-learn-pyhton-para-machine-learning/>. Acesso em: 30 jun. 2021a.

DIDATICA. **Como funciona o algoritmo de Árvore de Decisão (Decision Tree)**. Disponível em: <https://didatica.tech/como-funciona-o-algoritmo-arvore-de-decisao/>. Acesso em: 30 jun. 2021b.

FILATRO, A. **Data Science na educação: presencial, a distância e corporativa**. [s.l: s.n.].

HOMEM, W. L.; CARDOSO, B.; LOURENÇO, G. **JORNAL PET: BIG DATA E MACHINE LEARNING NA ENGENHARIA MECÂNICA**. [s.l: s.n.].

HONDA, H. **Introdução Básica à Clusterização - LAMFO**. Disponível em: <https://lamfo-unb.github.io/2017/10/05/Introducao_basica_a_clusterizacao/>. Acesso em: 30 abr. 2021.

INFOWORLD. **Como usar o Knime para a ciência de dados**. Disponível em: <https://por.small-business-tracker.com/how-use-knime-data-science-211481>. Acesso em: 30 jun. 2021.

JAMES, G. et al. **An Introduction to Statistical Learning**. 1st. ed. [s.l: s.n.].

JOVIC, A.; BRKIĆ, K.; BOGUNOVIĆ, N. **An overview of free software tools for general data mining**, 2015.

ROSA, C. S. **Estudo sobre as técnicas e métodos de análise de dados no contexto de Big Data**. [s.l: s.n.].

SALLES, R. **Data Science aplicado à manutenção preditiva: bombas d'água na Tanzânia.** Disponível em: <https://medium.com/ensina-ai/data-science-aplicado-à-manutenção-preditiva-bombas-dágua-na-tanzânia-32a85de8b41>. Acesso em: 10 maio. 2021a.

SALLES, R. **GitHub - RodrigoSalles/Tanzania_water_Pump: Manutenção preditiva de bombas d'água na Tanzânia**. Disponível em: <https://github.com/RodrigoSalles/Tanzania_water_Pump>. Acesso em: 11 ago. 2021b.

SCHIEZARO, M. **Como detectar anomalias utilizando Inteligência Artificial | Venturus**. Disponível em: <https://www.venturus.org.br/como-detectar-anomalias-utilizando-inteligencia-artificial/>. Acesso em: 30 abr. 2021.

UNIS. **Conheça 7 ferramentas que vão ajudar a aprimorar o seu processo de mineração de dados**. Disponível em: <https://blog.unis.edu.br/conheca-7-ferramentas-que-vao-ajudar-aprimorar-o-seu-processo-de-mineracao-de-dados>. Acesso em: 24 abr. 2021.

VDMA. **Machine Learning in Mechanical and Plant Engineering**Frankfurt, 2018.