UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

GUILHERME RODRIGUES CABREIRA

DETECÇÃO DE ANOMALIAS POR MEIO DE MACHINE LEARNING: ESTUDO DE CASO DE SISTEMAS PNEUMÁTICOS EM CAMINHÕES

PATO BRANCO

2022

GUILHERME RODRIGUES CABREIRA

DETECÇÃO DE ANOMALIAS POR MEIO DE MACHINE LEARNING: ESTUDO DE CASO DE SISTEMAS PNEUMÁTICOS EM CAMINHÕES

Anomaly detection using Machine Learning in a case of anti-pressure system in trucks

Trabalho de conclusão de curso de graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia Mecânica da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador(a): Gilson Adamczuk Oliveira.

PATO BRANCO

2022

GUILHERME RODRIGUES CABREIRA

DETECÇÃO DE ANOMALIAS POR MEIO DE MACHINE LEARNING: ESTUDO DE CASO DE SISTEMAS PNEUMÁTICOS EM CAMINHÕES

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia Mecânica da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação:01/dezembro/2022

_____

Fábio Rodrigo Mandello Rodrigues
Doutorado
Universidade Tecnológica Federal do Paraná

_____

Paulo Cezar Adamczuk
Doutorado
Universidade Tecnológica Federal do Paraná

_____

Gilson Adamzuk Oliveira
Doutorado
Universidade Tecnológica Federal do Paraná

PATO BRANCO

2022

Dedico este trabalho à todas as pessoas com quem pude estar e conhecer durante a minha vida, aos amigos, família e principalmente ao meu pai que não se encontra mais nesse plano e sempre acreditou em meu potencial dando todo suporte possível.

AGRADECIMENTOS

Agradeço primeiramente a Deus que é responsável por tudo nos dando a dádiva da vida. À minha mãe responsável por todo ensinamento e valores em minha vida. Meu irmão Gabriel que sempre me proporcionou uma visão mais crítica de mundo. Ao meu pai Clóvis Mauá que não pode estar mais presente de corpo, mas estará sempre vivo em nossas memórias, orações e lembranças.

Aos meus amigos que estiveram sempre presente em momentos bons e ruins, sempre dando todo apoio, conselhos e broncas quando necessário sendo uma verdadeira família em minha vida durante a minha estadia no Paraná.

Agradeço também ao professor Gilson Adamczuk Oliveira que desde o início esteve aberto a novos desafios por proporcionar todo suporte necessário no trabalho.

RESUMO

A ideia principal do vigente trabalho é a aplicação de conceitos e técnicas de *Machine Learning*, para a predição de falhas em equipamentos relacionados a área de Engenharia Mecânica. Apesar da vasta possibilidade na escolha do software para desenvolvimento do trabalho, foi optado pela utilização do *Orange Data Mining* devido a sua interface amigável e intuitiva promovido pela programação visual existente no aplicativo. É utilizado no trabalho dados reais captados diariamente em caminhões Scânia. Para a garantia da segurança na condução de veículos pesados, se faz valioso um estudo na previsão de falhas em sistemas pneumáticos, visto que este é responsável por todo sistema de frenagem do veículo. No trabalho foram utilizadas métricas de medições como *Classification Acuracy,* Precisão, *Recall* e *Specificity.* No *Software* foram utilizados para medição algoritmos como: Regressão Logística, *Naive Bayes* e *Random Forest.* Após a realização de dois diferentes métodos no tratamento das 60000 instâncias existentes no banco de dados, foi obtido uma acurácia de 99,4% em *Random Forest*, 85,6% em *Naive Bayes* e 99,1% em Regressão Logística. A alta acurácia obtida nos algoritmos demonstra a real possibilidade na previsão de falhas em equipamentos pneumáticos de caminhões, existindo-se a possibilidade do desenvolvimento de equipamentos para a medição em tempo real buscando evitar possíveis acidentes.

Palavras-chave: *Machine Learning*; Estudo de Caso; *ex-post; Orange Data Mining*; Manutenção Preditiva; Sistema Pneumático de caminhões.

ABSTRACT

The main idea of the this paper is the application of Machine Learning concepts and techniques, for the prediction of equipment failure related to the Mechanical Engineering area. Despite the vast possibility in choosing a software for the development of the work, we chose to use Orange Data Mining due to its friendly and intuitive interface promoted by the visual programming of the software. Real data collected daily from Scania trucks are used in the work. The guarantee of safety while driving heavy vehicles, a study in the prediction of failures in pneumatic systems is valuable, since it is responsible for the entire braking system of the vehicle. Measurement metrics such as Classification Accuracy, Precision, Recall and Specificity were used in the work. Algorithms such as: Logistic Regression, Naive Bayes and Random Forest were used in the Software for measurement. After performing two different methods in the treatment of the 60000 existing instances in the database, an accuracy of 99.4% was obtained in Random Forest, 85.6% in Naive Bayes and 99.1% in Logistic Regression. The high accuracy obtained in the algorithms demonstrates the real possibility of predicting failures in pneumatic truck equipment, with the possibility of developing equipment for real-time measurement in order to avoid possible accidents.

Keywords: *Machine Learning*; case study ; *ex-post; Orange Data Mining*; Predictive Maintenance; Pneumatic System of trucks.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de anomalias em 2 regiões 18
Figura 2 – Exemplo anomalia de contexto 20
Figura 3 – Anomalia coletiva em eletrocardiograma humano 21
Figura 4 – Exemplo Classificação 23
Figura 5 – Exemplo Regressão 24
Figura 6 – Estrutura do Estudo de Caso 27
Figura 7 – *Widgets* 32
Figura 8 – Exemplo *Decision Tree* 33
Figura 9 – Exemplo de Regressão Logística 35
Figura 10 – Instâncias faltantes no banco de dados aplicado 38
Figura 11 – *Data Table* treinamento 39
Figura 12 – Separação dos dados em treinamento e teste utilizando *Bootstrap* 40
Figura 13 – Ilustração do processo de *bootstraping* 41
Figura 14 – Dados de teste gerado a partir do *Bootstrap* utilizando a média nos dados não existentes 41
Figura 15 – Dados de teste gerado a partir do *Bootstrap* utilizando -1 nos dados não existentes 42
Figura 16 – *Workflow* desenvolvido no *Orange* utilizando média nos dados não existentes 43
Figura 17 – *Workflow* desenvolvido no *Orange* utilizando -1 nos dados não existentes 43
Figura 18 – Acurácia medida em cada algoritmo utilizado no treinamento utilizando a média nos valores inexistentes 44
Figura 19 – Acurácia medida em cada algoritmo utilizado no treinamento utilizando -1 nos valores inexistentes 45
Figura 20 – Previsões e acurácia no conjunto de teste utilizando a média nos valores inexistentes 45
Figura 21 - Previsões e acurácia no conjunto de teste utilizando -1 nos valores inexistentes 46
Figura 22 – Demonstração de resultados obtidos na predição em regressão logística utilizando a média nos valores inexistentes 46
Figura 23 - Demonstração de resultados obtidos na predição em regressão logística utilizando -1 nos valores inexistentes 47
Figura 24 – Demonstração predição *Naive Bayes* utilizando a média nos valores inexistentes 47
Figura 25 – Demonstração predição *Naive Bayes* utilizando -1 nos valores inexistentes 48
Figura 26 – Demonstração predição *Random Forest* com 200 árvores utilizando a média nos valores inexistentes 48
Figura 27 – Demonstração predição *Random Forest* com 200 árvores utilizando -1 nos valores inexistentes 49
Figura 28 – Demonstração predição *Random Forest* com 500 árvores utilizando a média nos valores inexistentes 49
Figura 29 – Demonstração predição *Random Forest* com 500 árvores utilizando -1 nos valores inexistentes 50
Figura 30 - Custos obtidos na predição utilizando cada um dos algoritmos 51

1 INTRODUÇÃO 13
1.1 Objetivos 14
1.1.1 Objetivo geral 14
1.1.2 Objetivos específicos 15
1.2 Justificativa 15
1.3 Estrutura do trabalho 15
2 REVISÃO BIBLIOGRÁFICA 17
2.1 Manutenção preditiva 17
2.2 Detecção de anomalias 18
2.2.1 Tipos de anomalias 19
2.2.1.1 Anomalia de ponto 19
2.2.1.2 Anomalia de contexto 19
2.2.1.3 Anomalia coletiva 20
2.2.2 Dados de saída na detecção de anomalias 21
2.3 *Machine Learning* 21
2.4 Tipos de aprendizagem 22
2.4.1 Aprendizagem supervisionada 22
2.4.1.1 Classificação 23
2.4.1.2 Modelo de regressão 24
2.4.2 Aprendizagem não supervisionada 24
2.4.2.1 *Clustering* 25
3 MATERIAIS E MÉTODOS 26
3.1 Estudo de caso 26
3.1.1 Estrutura conceitual teórica 27
3.1.2 Planejamento do caso 27
3.1.3 Condução de um teste piloto 28
3.1.4 Análise de dados 28
3.1.5 Geração do Relatório da Pesquisa 29
3.2 Escolha de estudo de aplicação *ex-post* 29
3.3 *Orange Data Mining* 31
3.4 Algoritmos supervisionados utilizados e métricas de avaliação 32
3.4.1 Árvore de decisão (*decision tree*) 32
3.4.2 *Random Forest* 34
3.4.3 *Gradient Boosting* 34

3.4.4 Regressão logística 34
3.4.5 *Naive Bayes* 35
3.4.6 Métricas de avaliação 35
4 RESULTADOS E DISCUSSÕES 38
4.1.1 Detalhamento dos dados utilizados no estudo de caso 38
4.2 Análise e discussão 42
5 CONCLUSÃO 52
REFERÊNCIAS 53
ANEXO A - Lei n. 9.610, de 19 de fevereiro de 1998 56

1 INTRODUÇÃO

Com o desenvolvimento desenfreado de tecnologias e de computadores a medição de dados se tornou algo muito mais acessível, tornando a disponibilidade deste algo abundante nos dias atuais. Dessa forma a programação orientada por dados como o *Machine Learning* torna-se uma ferramenta de grande interesse de estudos, existindo a possibilidade de ser uma grande aliada da engenharia mecânica na detecção de anomalias para manutenção em equipamentos.

A utilização de métodos orientados por dados tem evoluído exponencialmente e se tornando um exemplo na fabricação e solução de mobilidade, desde em manutenção preditiva até em qualidade preditiva, incluindo análise de segurança, garantia e facilidades no monitoramento em plantas de fábricas (THEISSLER et al., 2021).

O desafio permanente de qualquer organização, no contexto de pressão competitiva verificado nas três últimas décadas, tem sido produzir produtos e serviços cada vez melhores, mais rápidos, com o menor custo e com a melhor aceitação possível por parte do mercado consumidor (CALLIGARO, 2003).

O emprego de *Machine Learning* em indústrias pode gerar grandes benefícios no mercado atual, visando o aumento do valor agregado de produtos assim como uma produção planejada e sob demanda, sendo essencial para sua sobrevivência devido à alta concorrência.

Somado a redução de custos para geração de produtos competitivos no mercado, temos a necessidade da redução de descartes, visando a diminuição do impacto global causado por industrias a partir do conhecimento prévio de que as matérias primas são bens esgotáveis.

A preservação da integridade física dos operadores de máquinas é outro ponto também importante na detecção de anomalias, buscando a partir do prognóstico de falhas garantir um ambiente de trabalho mais seguro e confiável.

O *Machine Learning* tem aplicabilidade em diversas áreas como administração, robótica, medicina dentre outras. Nesta ferramenta as simulações são realizadas a partir de erros e acertos visando o reconhecimento de padrões e buscando a construção de um modelo de aprendizado de maneira iterativa.

O *Machine Learning* pode ser dividido em aprendizagem supervisionada e aprendizagem não supervisionada. Diferindo-se esses dois tipos de aprendizagem na

maneira que são definidas as respostas na saída do sistema a partir da entrada de um conjunto de dados.

A otimização da gestão de tempo, a partir da não ocorrência de paradas inesperadas em setores específicos da linha de produção, faz com que a integridade de componentes de uma máquina seja algo essencial em indústrias, buscando-se ao máximo reduzir o investimento em componentes com critérios de urgência, em decorrência de por muitas vezes os prazos de entrega não serem cumpridos, podendo acarretar em uma parada por completo na linha de produção e consequentemente havendo um impacto negativo na produtividade.

Dentre os fatores a serem atendidos no segmento de atuação das Indústrias de Processamento Contínuo, destacam-se como normalmente presentes o preço baixo, qualidade alta e entrega confiável, o que determina automaticamente, os objetivos de desempenho a serem priorizados pela organização, quais sejam: Custo, Qualidade e Confiabilidade. Dada a estreita relação da atividade com estes objetivos de desempenho, os citados desafios permanentes da produção acabam sendo os próprios desafios da atividade de manutenção. Como consequência disso, temos a adoção de estratégias funcionais de manutenção adequadas alinhadas com a estratégia corporativa, buscando um custo efetivo sem comprometer a segurança e a confiabilidade da planta operacional, sendo de fundamental importância para a organização (CALLIGARO, 2003).

1.1 Objetivos

Todas as folhas do trabalho, a partir da folha de rosto, devem ser contadas sequencialmente, mas não numeradas. A numeração deve ser inserida à partir da primeira folha da parte textual (Introdução), em algarismos arábicos, no canto superior direito da folha. Havendo apêndices e anexos, as suas folhas devem ser paginadas de maneira contínua.

1.1.1 Objetivo geral

O objetivo geral desse trabalho é prolongar a vida útil do sistema hidráulico em caminhões, visando o aumento da segurança dos condutores e a redução de custo com trocas desnecessárias de equipamentos. Busca-se também realizar a

identificação do melhor método para o prognóstico de falhas em caminhões a partir da utilização de dados previamente captados por sensores. A ideia é conduzir um estudo de caso *ex-post*, ou seja, com dados secundários de um *case* real analisado à lúz de algoritmos de *Machine Learning* analisados nesse trabalho.

1.1.2 Objetivos específicos

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

- Estudar possíveis problemas de falhas em caminhões;
- Estudar métodos de *Machine Learning* com possíveis aplicação no estudo de caso;
- Dominar a utilização do aplicativo *open source;*
- Analisar os resultados obtidos nos algoritmos para verificação de uma possível aplicação com medição em tempo real.

1.2 Justificativa

O emprego do *Machine Learning* em indústrias na detecção de anomalias é de grande valia quando refletimos na ocorrência de paradas inesperadas em setores específicos da linha de produção. A redução de custos com manutenção corretiva de maquinários faz com que a manutenção da integridade de componentes mecânicos de uma máquina seja essencial, objetivando uma melhor gestão de tempo e sucessivamente uma maior produtividade, sendo algo vital para sobrevivência de indústrias no mercado atual. O papel da Manutenção ganha destaque especial quando se fala de Indústrias de Processamento Contínuo (IPC), nas quais as condições físicas de instalações e seus equipamentos tem um papel fundamental na continuidade da produção de seus bens e serviços, de modo estável e seguro (CALLIGARO, 2003).

1.3 Estrutura do trabalho

O primeiro capítulo é composto por introdução, definição do problema objetivos e justificativa do trabalho.

No segundo capítulo é realizada a revisão bibliográfica onde é enfatizado os principais conceitos utilizados, como manutenção preditiva, detecção de

anomalias, *Machine Learning* e tipos de aprendizagem, realizando-se o embasamento teórico para o desenvolvimento do trabalho.

No capítulo três é apresentada a metodologia para a realização do trabalho, demonstrando o aplicativo que será utilizado para programação e a apresentação dos dados utilizados.

O quarto capítulo apresenta os resultados obtidos a partir das simulações realizadas assim como a efetividade da aplicação de diferentes técnicas.

É apresentado no quinto capítulo as considerações finais do vigente trabalho.

2 REVISÃO BIBLIOGRÁFICA

2.1 Manutenção preditiva

De acordo com Paolanti et al. (2018) manutenção preditiva refere-se ao monitoramento inteligente dos equipamentos para evitar falhas futuras. A manutenção preditiva evoluiu do primeiro método que é a inspeção visual, para métodos automatizados que utilizam técnicas de processamento de sinal avançado com base no reconhecimento de padrões e *Machine Learning*.

A manutenção preditiva é um conjunto de atividades que detecta mudanças nas condições físicas dos equipamentos (sinais de falha), buscando realizar o trabalho de manutenção apropriada para maximizar a vida útil do equipamento sem aumentar seu risco de falha (WANG, 2017).

Em uma manutenção preditiva, primeiramente tenta-se prever o estado de integridade do sistema, após isso é apropriado planos de ações de acordo com o retorno das previsões. A atividade de prognóstico de falhas visa antecipar o tempo do defeito a partir da previsão do estado de saúde futuro de um dado componente, subsistema ou sistema e sua vida útil remanescente (TOBON-MEJIA et al., 2012).

Segundo Tobon Mejia et al. (2012) os métodos para manutenção preditiva em termos de sua condição de uso podem ser geralmente classificados em três categorias: método baseados em modelagem matemática, método baseado na experiência e método baseado em dados. Métodos baseados em modelagem matemática produzem bons resultados, porém requerem um vasto conhecimento nas funções do sistema, o que é visto como uma desvantagem e muitas vezes é de difícil obtenção. Métodos baseados na experiência tomam decisões baseados em eventos empíricos e frequentemente produzem resultados ruins de predição. Métodos baseados em dados investigam o problema a partir de uma perspectiva nos números (CHEN et al., 2021).

O método baseado em dados é também conhecido como mineração de dados ou *Machine Learning*, que usa dados históricos para aprender um modelo de comportamento do sistema. A abordagem baseada em modelo matemático tem a capacidade de incorporar a compreensão física do produto alvo, contando com o modelo analítico para representar o comportamento do sistema (PAOLANTI et al., 2018).

A manutenção preditiva utilizando métodos orientados por dados necessitam dos dados de falha como uma variável e utilizam destes para descrever os dados de entrada (GOURIVEAU; RAMASSO; ZERHOUNI, 2013).

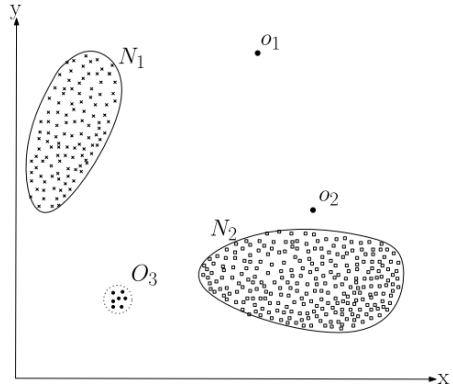
2.2 Detecção de anomalias

Segundo Ahmed; Naser Mahmood; Hu (2016), a detecção de anomalias é uma tarefa importante de análise de dados pois detecta dados anormais de um determinado conjunto. É uma área interessante de pesquisa de mineração de dados, pois envolve a descoberta de raros padrões.

A partir de Quatrini et al. (2020), as principais técnicas para detecção de anomalias são baseada na Classificação, *Clustering*, *Nearest Neighbor, Statistical* e na informação teórica e de espectros. Estudos sobre detecção de anomalia tem alcançado excelentes resultados utilizando os algoritmos de *Random Forest* (RFA), sendo o *Decisions Forest* o mais popular na predição de modelos.

De acordo com Chandola; Banerjee; Kumar (2009), anomalias são padrões nos dados que não se encontram em conformidade com uma noção de comportamento previamente definido. A Figura 1 ilustra anomalia em duas dimensões de conjunto de dados. Os dados tem duas regiões normais, N1 e N2, visto que a maior parte das observações se encontram nessas duas regiões. Pontos que se encontram suficientemente longe dessas regiões, isto é, pontos o1, o2 e O3 são denominadas anomalias.

Figura 1 – Exemplo de anomalias em 2 regiões



Fonte: Chandola; Banerjee; Kumar (2009)

2.2.1 Tipos de anomalias

Um importante aspecto da detecção de anomalias é a natureza da mesma. Anomalias podem ser classificadas nas seguintes 3 categorias apresentadas (CHANDOLA; BANERJEE; KUMAR, 2009).

2.2.1.1 Anomalia de ponto

Se a instância de um dado individual for considerada fora do domínio em relação ao resto dos dados, dessa forma a instância é determinada uma anomalia de ponto. Esse é o tipo mais simples de anomalia e é o foco da maioria das pesquisas de detecção de anomalia.

Tendo como exemplo a Figura 1, os pontos o1, o2 e O3 estão fora dos limite da região normal, sendo assim são pontos de anomalia por se localizarem fora do domínio N1 e N2 onde se encontram os pontos dos dados (CHANDOLA; BANERJEE; KUMAR, 2009).

2.2.1.2 Anomalia de contexto

A noção de contexto é induzida pela estrutura dos dados e tem que ser especificada como parte da formulação do problema. Cada instancia de dado é definida com base em dois atributos:

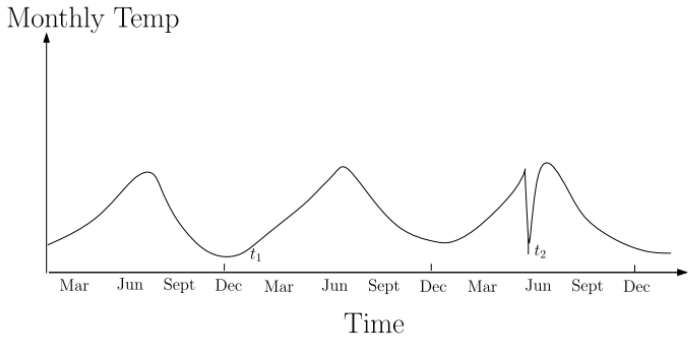
- Atributo contextual: é utilizado para determinar o contexto da instancia. Por exemplo em um conjunto de dados para determinação do espaço, a longitude e a latitude de uma localização são os atributos em contexto. Nos dados de uma série temporal, o tempo é o atributo em contexto que a determina a posição de uma instancia de uma sequência inteira.
- Atributo comportamental: defini as características não contextuais de uma instancia. Por exemplo, em um conjunto de dados espaciais que descreve o comportamento da precipitação média de todo o mundo, a quantidade de chuva em qualquer local específico é um atributo comportamental do mesmo.

O comportamento anômalo é determinado a partir da utilização de valores dos atributos comportamentais em um contexto específico. Uma instância de dados pode ser uma anomalia contextual em determinada ocasião (em termos de atributo

comportamental) e considerada normal em um contexto diferente (CHANDOLA; BANERJEE; KUMAR, 2009).

A figura 2, demonstra um exemplo de série temporal de temperatura, na qual é demonstrada a temperatura mensal de uma área em anos passados. A temperatura t1, é considerada normal visto que esta se situa no inverno norte americano, porém a temperatura t2 que têm o mesmo valor que t1 é considerada uma anomalia visto que está no verão.

Figura 2 – Exemplo anomalia de contexto

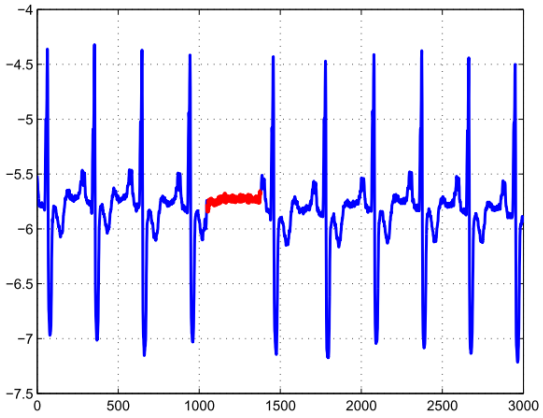


Fonte: Chandola; Banerjee; Kumar (2009)

2.2.1.3 Anomalia coletiva

Se uma coleção de instancia de dados relacionados for anômala no que diz a respeito a todo conjunto de dados, é denominada uma anomalia coletiva. As instancias de dados individuais em uma anomalia coletiva podem não ser anomalia por si mesmas, mas sua ocorrência em conjunto como uma coleção é anômala. A Figura 3 ilustra um exemplo de um eletrocardiograma humano. A região vermelha ilustra uma situação de anomalia coletiva devido o mesmo valor baixo existir por um grande período anormal (CHANDOLA; BANERJEE; KUMAR, 2009).

Figura 3 – Anomalia coletiva em eletrocardiograma humano



Fonte: Chandola; Banerjee; Kumar (2009)

2.2.2 Dados de saída na detecção de anomalias

Um aspecto importante na técnica de detecção de qualquer anomalia é a maneira de que as anomalias são reportadas. Normalmente o emprego de técnicas na detecção de anomalias é um dos dois tipos citados a seguir (CHANDOLA; BANERJEE; KUMAR, 2009).

Existem dois tipos de saída (*output*) que varia de acordo com a técnica empregada: pontuações (*scores*) e rotulações (*labels*). *Scores* é quando se atribui uma pontuação a cada instância. Essa pontuação representa o grau de anomalia daquele objeto a partir da utilização de um limite definido dentro de um domínio específico para seleção das anomalias mais relevantes. Por outro lado, uma *label* define uma rotulação categórica para cada instância como, por exemplo: normal ou anomalia (MATA, 2017).

2.3 *Machine Learning*

Machine Learning é um campo de estudo onde os computadores adquirem a habilidade de aprender sem ser explicitamente programados para isso (SAMUEL, 1959). Segundo Mitchell e Blum (1998), *Machine Learning* é um programa de computador que aprende a partir da variável (E) referente a experiência, a respeito de alguma tarefa definida pela variável (T) obtendo uma performance medida por (P), sendo quando a performance em P aumenta a partir da realização da tarefa em T, obtendo-se cada vez uma maior experiência E. Existindo dessa forma uma ciclo entre essas letras e tendo-se uma progressão da velocidade de processamento e capacidade de resolução de problemas com o passar do tempo.

A assertividade é maior quanto maior for o número de dados pré-existentes, alcançando-se resultados e previsões consideráveis para a tomada de decisões e identificação de anomalias. Estas previsões só são possíveis devido a teoria de *statistical learning* que foca principalmente em métodos estatísticos e na análise funcional. Se a nova observação não estiver representando a propriedade estatística dos dados, a observação é considerada uma anomalia. As técnicas estatísticas ajustam os dados de condição operacional esperados típicos e, em seguida, aplicam o teste de inferência estatística para determinar se a nova observação pertence ao modelo estatístico ajustado (SUTHARSSAN et al., 2015).

2.4 Tipos de aprendizagem

A partir de Paolanti et al. (2018) manutenção preditiva baseada em *Machine Learning* pode ser dividida em duas classes principais: supervisionado e não supervisionado.

2.4.1 Aprendizagem supervisionada

As informações da ocorrência de falhas estão presentes no conjuntos de dados havendo dessa forma dados de saídas previamente conhecidos (PAOLANTI et al., 2018). Se o algoritmo recebe as saídas rotuladas para um conjunto de entradas, o aprendizado é chamado de aprendizado supervisionado. Seu objetivo é prever uma saída correta para um novo dado de entrada. A maioria dos problema de manutenção preditiva podem ser tratados como problemas de aprendizagem supervisionada, onde um conjunto de dados de integridade e de falhas estão disponíveis (SUTHARSSAN et al., 2015). O aprendizado supervisionado consiste em duas classes de algoritmos: Classificação e Regressão.

Na classificação, é abrangido diversos algoritmos que permitem realizar a análise de uma variável dependente categórica, sendo utilizado como base, dados de entrada existente no banco de dados e a partir da análise desses dados, é gerado na saída um resultado qualitativo.

Assim como na classificação, na regressão são utilizados dos dados de entrada previamente analisados para se prever um resultado, tendo-se como distinção que a previsão é realizada numericamente e não categoricamente como na outra classe de algoritmo.

2.4.1.1 Classificação

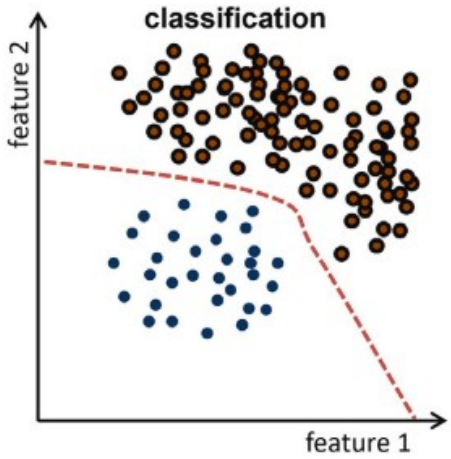
Segundo Ahmed; Naser Mahmood; Hu (2016) técnicas baseadas em classificação possuem importantes informações de dados na saída do sistema (*output*) para preservar ataques a rede. Quando dados anormais são fornecidos para a detecção de anomalia de um sistema, um ataque com um padrão conhecido pode ser notado evitando grandes prejuízos. Isso depende exclusivamente do banco de dados existente na saída de um sistema que só detecta a anomalia caso esta tenha sido previamente definida.

As técnicas de detecção baseada em classificação operam de maneira bifásica, a fase de treinamento aprende um classificador a partir da utilização de dados de treinamento disponíveis. A fase de teste realiza a classificação de uma instância de teste como normal ou anomalia usando um classificador (CHANDOLA; BANERJEE; KUMAR, 2009).

O reconhecimento do tipo de uma cena capturada em uma fotografia pode ser lançado como uma tarefa de classificação, onde a saída desejada é um rótulo categórico discreto, por exemplo: uma cena de praia, uma paisagem urbana, uma imagem de interior (CRIMINISI; SHOTTON; KONUKOGLU, 2011).

Na figura 4 está demonstrado um exemplo de classificação, onde estão separados dados normais de dados com anomalias.

Figura 4 – Exemplo Classificação

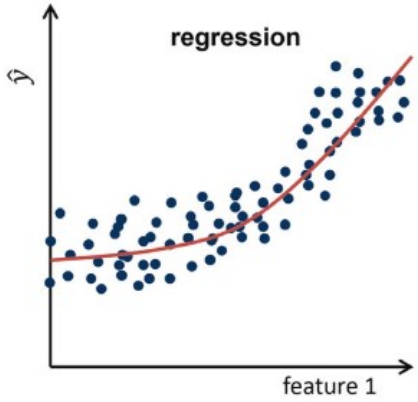


Fonte: Theissler et al. (2021)

2.4.1.2 Modelo de regressão

Modelos de regressão na manutenção preditiva são utilizados para calcular o tempo de vida útil restante de um equipamento, definindo a quantidade de tempo de operação restante antes da ocorrência da próxima falha (PAOLANTI et al., 2018). Algoritmos de regressão usam os recursos de entrada para prever os valores de saída dos dados escondidos no sistema (AL-AMRI et al., 2021). Prever o preço de uma casa em função da sua distância de uma boa escola pode ser considerado um problema de regressão (CRIMINISI; SHOTTON; KONUKOGLU, 2011). Um exemplo de regressão está ilustrado na figura 5 demonstrada a seguir:

Figura 5 – Exemplo Regressão



Fonte: Theissler et al. (2021)

2.4.2 Aprendizagem não supervisionada

As informações de processo estão disponíveis, mas não existem dados de saída definidos, havendo uma execução da tarefa com base em suas características e padrões semelhantes (PAOLANTI et al., 2018). A partir de Sutharssan et al. (2015) a aprendizagem não supervisionada é usada onde não há dados rotulados disponíveis. É usado para descobrir grupos semelhantes nos dados com base em técnicas de agrupamento. Tendo a disponibilidade na maioria dos novos sistemas apenas de dados operacionais normais, havendo o reconhecimento de um sistema saudável a partir da utilização destes e de maneira subsequente realizando a prevenção de possíveis falhas, mantendo a confiabilidade e realizando a estimativa do tempo de vida restante de um determinado sistema. Em sistemas não

supervisionados existem algumas técnicas que utilizam regras de *Clustering* e associação.

2.4.2.1 *Clustering*

De acordo com A.K. Jain; M.N. Murty; P.J. Flynn (1999), *clustering* é a classificação não supervisionada de padrões em grupos. Segundo Al-Amri et al. (2021), o objetivo da utilização de *clustering* é a identificação de dados normais na entrada, o mesmo ressalta que os dados de entrada possuem uma estrutura de certos padrões que ocorrem com uma maior frequência do que outros. Em estatística, esta etapa é chamada de estimativa da densidade. O agrupamento (*clustering*) é um método para estimativa de densidade.

O método de aprendizagem não supervisionada mais comum é a análise de agrupamento (*Clustering*), que é usada para análise exploratória de dados para encontrar padrões ou agrupamentos ocultos nos dados (THEISSLER et al., 2021).

3 MATERIAIS E MÉTODOS

3.1 Estudo de caso

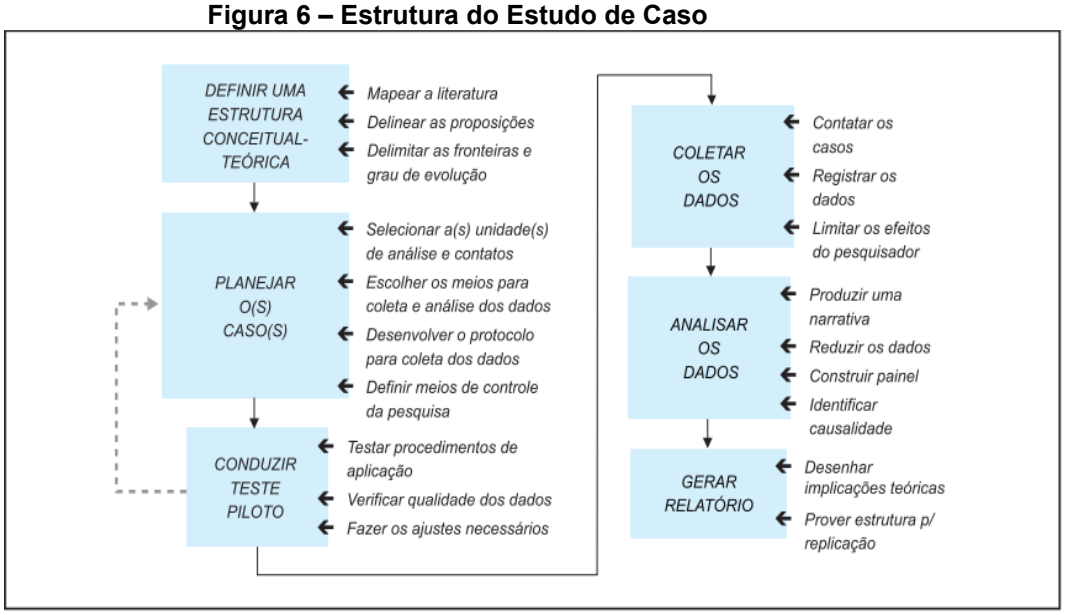
Cauchick Miguel (2007) define estudo de caso como um estudo de natureza empírica que investiga um determinado fenômeno, geralmente contemporâneo, dentro de um contexto real de vida, quando as fronteiras entre o fenômeno e o contexto em que ele se insere não são claramente definidas. Tendendo a esclarecer a tomada de determinadas decisões e escolhas, assim como quais resultados foram alcançados a partir das mesmas.

Na vigente pesquisa, o problema é tratado como um *ex-post facto* isto é, um estudo de um determinado acontecimento que já ocorreu no passado. Buscando-se averiguar o algoritmo de *Machine Learning* com a maior assertividade na manutenção preditiva de máquinas utilizando dados pré-existentes. As etapas para condução de um estudo de caso estão definidas na Figura 6.

A escolha do estudo *ex-post* foi realizada devido à dificuldade na obtenção de dados em produtos. A manutenção da propriedade intelectual e a garantia de autenticidade é uma estratégia de mercado que visa assegurar a exclusividade de produtos pelas marcas.

A etapa de teste é realizada antes do lançamento de equipamentos e produtos, sendo realizadas em laboratórios controlados, visando que haja a menor interferência possível e uma maior exatidão na captação dos dados. Dificilmente esses dados são divulgados pelas empresas detentoras dos equipamentos, e quando são realizadas algumas características são ocultas ou estes são exteriorizado em quantidades pequenas o que impossibilita a obtenção de resultados confiáveis na realização de um estudo.

Figura 6 – Estrutura do Estudo de Caso



Fonte: Cauchick Miguel (2007)

3.1.1 Estrutura conceitual teórica

Segundo Cauchick Miguel (2007) primeiramente deve-se definir um referencial conceitual-teórico para o trabalho, de forma a resultar em um mapeamento da literatura. Servindo o referencial teórico também para a delimitação das fronteiras do que será investigado no trabalho e suporte técnico para a realização da pesquisa.

3.1.2 Planejamento do caso

Primeiramente para a realização do planejamento de caso, é de enorme importância o entendimento do tema em pauta. Uma compreensão abrangente do problema, assim como a contextualização e reflexão do impacto causado pelo desenvolvimento do vigente estudo de caso, é garantia de um encaminhamento correto no estudo sendo de suma importância para a obtenção de bons resultados.

O planejamento do caso é realizado após a definição do referencial conceitual-teórico, onde tem-se como primeira tarefa a escolha das unidades que serão analisadas. Seguido da seleção dos casos, determina-se os métodos e técnicas que serão utilizados na coleta e na análise dos dados devendo ser empregadas múltiplos métodos e evidências. Visando uma validade construtiva a partir da mensuração do conceito que se tem como alvo.

A partir da escolha dos métodos e técnicas para a coleta de dados, é realizado um protocolo com regras e procedimentos para o prosseguimento da pesquisa. Sendo este protocolo mais que um mero roteiro e sim um instrumento para melhora da confiabilidade e validade na condução da pesquisa.

A escolha da aplicação do estudo de caso tendo como enfoque os caminhões, se deu a partir do entendimento de que predominantemente as rodovias são as maiores responsáveis pelo transporte de cargas em nosso país sendo responsáveis por 61,1% em 2009, segundo dados da Confederação Nacional de Transporte. Somado a esse fato, existe também que apesar da frota de caminhões representar apenas 5% da frota nacional no ano de 2021, esta foi responsável por 47% das mortes que ocorreram em rodovias federais neste mesmo ano.

3.1.3 Condução de um teste piloto

A condução de um teste piloto tem como foco o estudo da qualidade dos dados obtidos, a partir da avaliação da contribuição dos mesmos, realizando-se os ajustes necessários para que o objetivo da pesquisa seja alcançado e sejam obtidos resultados confiáveis.

O teste piloto no estudo de caso foi realizado a partir de um teste com um menor número de instância para a investigação se a partir do banco de dados utilizados seriam obtidos resultados expressivos e que justificassem a utilização do mesmo.

3.1.4 Análise de dados

Na análise de dados é feita uma avaliação do conjunto de dados coletados, visando que no conjunto permaneçam somente dados pertinentes para que o objetivo da pesquisa seja alcançado, sendo geralmente necessário realizar uma redução dos mesmos, permanecendo somente o essencial para a pesquisa. Podendo ser necessário nessa etapa remover ou complementar dados incompletos, selecionando as variáveis mais adequadas para a composição de modelos.

No estudo de caso, fez-se necessário complementar 4980 instâncias numéricas incompletas, sendo realizado duas tratativas diferentes para a comparação da geração de resultados com a maior precisão.

3.1.5 Geração do Relatório da Pesquisa

É realizada a sintetização de atividades das etapas anteriores, havendo considerações e relacionando os resultados alcançados com a teoria já apresentada no trabalho. Havendo a preocupação de que à teoria não seja ajustada com os resultados obtidos, e sim de que o inverso aconteça para que o estudo de caso tenha confiabilidade e validade.

3.2 Escolha de estudo de aplicação *ex-post*

O estudo de caso escolhido para aplicação *ex-post* é pertencente a um banco de dados constituídos por 60000 variáveis que foram coletados por sensores a partir da utilização diária de caminhões, buscando prevenir a integridade e prolongar a vida útil do sistema de ar pressurizado de caminhões Scania.

Buscando a conservação da integridade dos elementos das máquinas, a manutenção aliada com a tecnologia é um assunto de grande relevância nas indústrias, havendo uma grande expectativa no futuro de que custos desnecessários sejam evitados. A inserção de tecnologias no meio fabril e na manutenção de maquinários vem de encontro com a ideia da indústria 4.0, visando evitar paradas inesperadas a partir de dados previamente captados, buscando dessa forma uma produção ininterrupta e sem grandes surpresas.

O banco de dados foi desenvolvido a partir da leitura de sensores os quais tendo como foco indicar a necessidade de reparo de um caminhão antes de que haja a falha de seu sistema por completo.

Apesar dos veículos pesados representarem apenas 5% dos automóveis da frota brasileira, os acidentes envolvendo esse tipo de veículo possui um percentual expressivo de fatalidade. No ano de 2021 estas colisões foram responsáveis por 47% das fatalidades em acidentes nas rodovias federais brasileiras a partir de dados divulgados no anuário da polícia rodoviária federal.

Os acidentes de veículos pesado majoritariamente tem da falha humana como um de seus maiores responsáveis, segundo estudos. No entanto, o estudo na detecção precoce de possíveis falhas em caminhões é um assunto de grande relevância para que exista uma maior segurança no deslocamento de veículos em vias urbanas no mundo, buscando a manutenção da integridade física não só dos

condutores de veículos pesados, mas também dos utilitários que trafegam pelas vias urbanas e rodovias.

Dentre as falhas mecânicas mais comuns em caminhões, destacam-se:

- Motor enfraquecido : defeito acarretado pelo possível entupimento na mangueira de filtro de ar e diesel, utilização de combustíveis de má qualidade, vazamento em mangueiras do *intercooler*, desgaste interno de pistões e anéis, dentre outros;
- Falta de estabilidade: problema acarretado devido à falta de manutenção em itens da suspensão sendo acarretado diversas vezes devido a qualidade precária dos asfaltos das rodovias;
- Desgaste nos pneus: problema acarretado pela falta de verificação e demora para a realização da substituição;
- Problemas no sistema pneumático: sistema que possui maior demanda de funcionamento em um caminhão devido as altas cargas transportadas, podendo muitas vezes ser danificado devido a choques com obstáculos em estradas.

O sistema pneumático desempenha diversas funções em caminhões tendo como uma das principais delas a frenagem do veículo. Visando a manutenção da integridade e segurança do condutor do veículo, o estudo de caso tem como foco realizar a predição na falha do sistema afim de evitar possíveis acidentes e também poupar recurso em casos de trocas desnecessárias de equipamentos.

Os sistemas pneumáticos em caminhões possuem o seguinte funcionamento na maioria das vezes, o compressor comprime o ar e o envia para o regulador de pressão o qual realiza o controle da pressão de trabalho. Após isso a pressão regulada é distribuída igualmente em 4 cilindros independentes por uma válvula de proteção.

O ar comprimido é distribuído de uma maneira separada, sendo duas vias destinadas para os freios do veículo (dianteiro e traseiro), a terceira para o estacionamento e a quarta com saída para os acessórios do caminhão. Caso haja vazamento de ar de uma das saídas as outras são automaticamente bloqueadas para que não se tenha perdas de ar no circuito de freio. Para ainda uma maior segurança do veículo as válvulas de freio priorizam os freios dianteiro e traseiro e dos acessórios, deixando por último o freio de estacionamento, visando evitar que o veículo saia sem ar comprimido nos circuitos citados anteriormente.

O circuito utilizado em freios é chamado de circuito duplo, onde o ar fica retido na válvula de comando e só é liberado quando o pedal de freio é acionado. Caso ocorra queda de pressão em um dos circuitos (traseiro ou dianteiro) o outro funciona de maneira independente. Após o pedal de freio ser desativado, o ar que foi utilizado para acionar a válvula é liberado para a atmosfera através da descarga e o compressor simultaneamente repõe o ar consumido pelo sistema.

O sistema de frenagem é composto por:

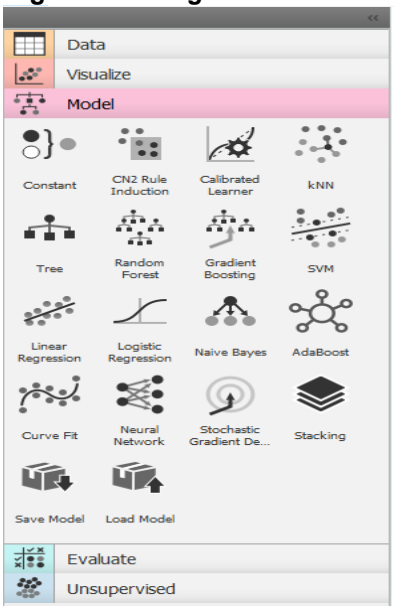
- Válvula reguladora de pressão;
- Elemento secador;
- Reservatório de regeneração;
- Válvula de proteção de 4 circuitos;
- Câmara de freio de estacionamento e serviço;
- Válvula moduladora do freio de estacionamento;
- Válvula relé;
- Válvula de descarga rápida;
- Válvula moduladora do freio de serviço.

Normalmente os problemas ocorrem devido a vazamentos em alguma válvula ou a partir da contaminação por óleo lubrificante do compressor ou água as quais fazem o desgaste das válvulas ser mais severo. Outros possíveis problemas existentes podem ser causados devido à perda de vedação das válvulas, entupimento do cano de saída e vazamentos. Podendo haver a contaminação do sistema por um todo, devendo-se inspecionar toda a tubulação de saída a qual deve ser limpa ou em casos mais severos substituídas.

3.3 *Orange Data Mining*

O *Orange Data Mining* é um programa criado no ano de 1996 que conta com programação na linguagem C++ e Python, contando em sua instalação padrão algoritmos de *Machine Learning*, com pré-processamento e visualizações de dados divididos em 5 conjuntos de *widgets* demonstrados na Figura 7, sendo eles: dados, visualização, modelo, avaliação e não supervisionado. Tendo também funcionalidades adicionais como análise de imagens, geografia, *network*, *single cel*, fusão de dados, mineração de texto, dentre outras.

Figura 7 – *Widgets*



Fonte: Autoria Própria (2021)

A programação utilizando o *Orange Data Mining* aparece como uma opção muito interessante devido ao visual atrativo e interativo do programa. Havendo a possibilidade de realizar uma exploração de dados com análise qualitativa de forma rápida e com visualização limpa, sem haver a necessidade de realizar a codificação e podendo dessa forma enfatizar a análise exploratória de dados.

Além de seu visual atrativo, o programa conta com a vantagem de ser um aplicativo completamente gratuito e capaz de retornar para o usuário uma perspectiva diferente de um determinado problema, podendo-se obter uma visão mais abrangente a partir da inserção de dados simples.

Dentre diversas ferramentas de análise existente no programa temos distribuições estatísticas, diagramas de caixa, gráficos de dispersão, agrupamento hierárquico, mapas de calor, projeções lineares etc.

3.4 Algoritmos supervisionados utilizados e métricas de avaliação

3.4.1 Árvore de decisão (*decision tree*)

Al-Amri et al. (2021) definiu *decision tree* como uma abordagem que constrói técnicas de regressão ou classificação em uma estrutura de árvore,

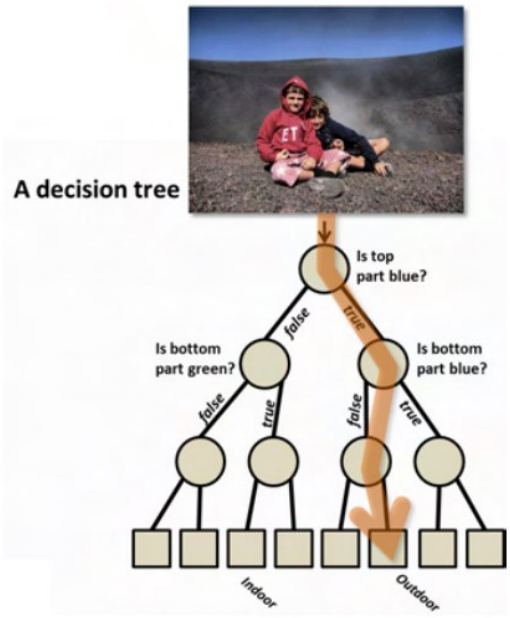
separando o conjunto de dados em pequenos grupos, enquanto ao mesmo tempo este conjunto de dados se torna cada vez maior.

O método *Decision Tree* possui fácil compreensão e podemos pensar neste como uma representação gráfica para um processo de decisão, onde são feitas determinadas perguntas para um dado e a partir dessas respostas são obtidos diferentes caminhos com diferentes resultados. As perguntas sobre o dado percorrem nós chegando a respostas, podendo ser feita uma analogia com a raiz de uma árvore que se ramifica até um galho no topo, respeitando-se uma hierarquia que se inicia a partir de um nó raiz e se ramifica em nós filhos até chegar em um determinado nó terminal, ou nó folha.

De acordo com Quatrini et al. (2020), em *machine learning*, uma árvore de decisão é um modelo preditivo onde cada nó interno é um nó de decisão e representa uma variável de predição. Onde uma decisão representa uma faixa de valor em um nó que vai se estreitando até um valor de predição na variável alvo. Cada nó interno divide os dados em duas ramificações que move a decisão para um consecutivo nó.

Na Figura 8 é demonstrado um exemplo de *Decision Tree*, detalhando-se a maneira que são feitas as perguntas e as respostas para o dado.

Figura 8 – Exemplo *Decision Tree*



Fonte: Criminisi; Shotton; Konuglu (2011)

3.4.2 *Random Forest*

Random forest é um método de conjuntos, combinando várias árvores de decisão. Tendo como objetivo o desenvolvimento de diversas árvores que são obtidas a partir da utilização de diferentes subconjuntos de dados de treinamento. Os dados de saída são determinados a partir da votação majoritária de todas as árvores. Florestas aleatórias podem ser utilizadas em problemas que envolvam classificação ou regressão podendo lidar com dados de alta dimensão (BREIMAN, 2001).

A ideia de *Random Forest* reside na junção de muitas árvores de decisão que, a partir de critérios de seleção, se ramificam e chegam cada uma a uma resposta para o problema. Em seguida a resposta que tiver mais votos é a solução geral. O modelo como um todo é formado por árvores de decisão que crescem a partir dos dados de entrada (LIMA; AMORIM, 2020).

3.4.3 *Gradient Boosting*

Segundo Lima; Amorim (2020), a lógica do modelo *Gradient Boosting*, é a junção de vários modelos fracos, onde o processo de aprendizagem se ajusta conforme novos modelos são adicionados tendo como objetivo obter uma estimativa mais precisa da variável de resposta.

Criminisi; Shotton; Konukoglu (2011), pondera que *Boosting* é a construção de classificadores fortes a partir da combinação linear de vários classificadores fracos. Um classificador é impulsionado a partir do treinamento de cada iteração dos exemplos de treinamentos, enfatizando os exemplos que possuem o pior desempenho e buscando-se a partir da impulsão aumentar o seu peso de treinamento.

3.4.4 Regressão logística

Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada diretamente. No caso da variável dependente Y assumir apenas dois possíveis estados (1 ou 0) e haver um conjunto de p variáveis independentes X1 X2 , ... , Xp, o modelo de regressão logística pode ser escrito da seguinte forma (MINUSSI; DAMACENA; NESS JR, 2002):

$$P(Y = 1) = \frac{1}{1+e^{-g(x)}} \quad (1)$$

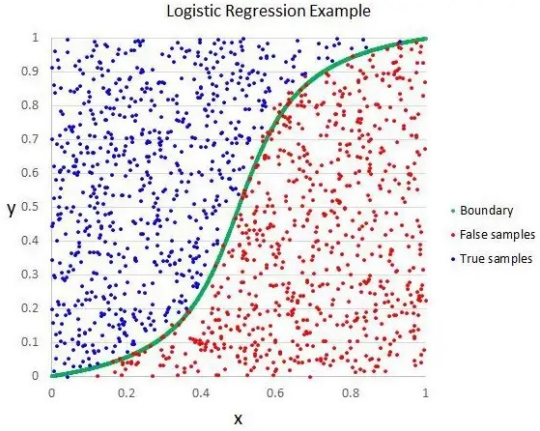
Onde,

$$g(x) = B_0 + B_1X_1 + \cdots + B_pX_p \quad (2)$$

Estimando-se os coeficientes a partir do conjunto pelo método da verossimilhança. Segundo Minussi; Damacena; Ness Jr (2002) considerando os valores de B0, B1, ..., Bp e variando os valores de X, nota-se que a curva logística tem comportamento probabilístico no mesmo formato da letra S.

Neste método é assumido que as variáveis podem ser linearmente separadas, tendo como objetivo fazer a melhor aproximação para encontrar os pontos de divisões positivos e negativos e sucessivamente a curva ser traçada separando os valores falsos e verdadeiros. Um exemplo de regressão logística está ilustrado na Figura 9 demonstrada a seguir:

Figura 9 – Exemplo de Regressão Logística



Fonte: Chauhan (2019)

3.4.5 *Naive Bayes*

Em termos simples, um classificador *Naive Bayes* assume que a presença ou abstenção de uma determinada característica não pode ser relacionada pela presença ou abstenção de outra característica. Por exemplo, uma fruta pode ser considerada uma maçã caso esta seja vermelha e tenha em torno de 4 polegadas de diâmetro. Mesmo que essas características sejam dependentes uma da outra, um classificador *Naive Bayes* considera as propriedades que contribuem para a probabilidade dessa fruta ser uma maçã como sendo independentes (ROCHA, 2006).

3.4.6 Métricas de avaliação

A avaliação da assertividade de cada algoritmo é representada pela métrica de avaliação, refletindo esta a qualidade de um modelo selecionado e sua precisão.

O entendimento do contexto em que os dados estão contidos é de suma importância e necessidade na escolha de uma boa métrica para a avaliação de modelos.

No vigente trabalho serão utilizados métricas como:

1. *Classification Acuracy*;
2. Precisão;
3. *Recall*;
4. *Specificity;*
5. *F1 Score.*

Na *Classification Acuracy* (CA) é indicado a razão de quantas classificações o modelo obteve de maneira correta dentre todos os exemplos analisados.

$$CA = \frac{Resultados\ corretos}{Total\ de\ dados} \quad (3)$$

A precisão é uma medida de quantas predições positivas foram feitas corretamente dividido pelo número total de predições positivas feitas.

$$Precisão = \frac{Número\ de\ predições\ positivas\ feitas\ corretamente}{Número\ total\ de\ predições\ positivas\ feitas} \quad (4)$$

Recall é a medida de quantas classificações positivas foram feitas da corretamente dividido pelo total de instâncias positivas existentes no banco de dados.

$$Recall = \frac{Número\ de\ predições\ positivas\ feitas\ corretamente}{Número\ total\ de\ predições\ positivas\ existentes} \quad (5)$$

No *Specificity*, é a medida de quantas predições negativas foram feita corretamente.

$$Specificity = \frac{Número\ de\ predições\ negativas\ feitas\ corretamente}{Número\ total\ de\ predições\ negativas\ existentes} \quad (6)$$

O *F1 Score,* é uma medição onde é realizada a combinação da precisão com a *Recall.* Sendo uma harmonia dessas duas métricas e uma outra maneira de se calcular uma média, sendo mais adequada em proporções do que uma média

aritmética tradicional. A ideia é proporcionar uma métrica que faça a abrangência dos dois índices.

$$F1_{Score} = 2 * \frac{Precisão * Recall}{Precisão + Recall} \quad (7)$$

Buscando facilitar a visualização dos resultados obtidos em cada algoritmo, será também utilizado a seguinte métrica:

- Para trocas desnecessárias realizadas por mecânicos (falso positivo) os valores serão multiplicados por 10.
- Para trocas necessárias não realizadas por mecânicos (falso negativo) que podem acarretar em uma futura falha no sistema, os valores serão multiplicados por 500.

Os valores determinados como falso negativos, são valores que na predição foram classificados como negativos e atualmente são pertencentes a classe positiva. Os valores definidos como falso positivo são valores do banco de dados que são pertencentes a classe positiva, porém na predição foram classificados como negativo.

$$CustoTotal = 10 * fp + 500 * fn \quad (8)$$

Segundo Solomon (2021) a meta do estudo, é minimizar o custo total, podendo-se notar que a longo prazo os custos das trocas que seriam necessárias e não foram realizadas por mecânicos excedem em muito o custo 1.

4 RESULTADOS E DISCUSSÕES

4.1.1 Detalhamento dos dados utilizados no estudo de caso

O estudo de caso é composto por dados adquirido por sensores obtidos a partir da leitura diária na utilização de caminhões Scania, tendo como foco minimizar a substituição de componentes que não se encontram no final de sua vida útil e, portanto, estão desempenhando seu papel sem apresentar falha. A partir do conhecimento dos componentes que não necessitam ser substituídos, busca-se não somente reduzir o gasto em manutenção com falsas predições como também aumentar a segurança dos condutores.

No banco de dados possuem duas classes de dados, sendo uma delas a classe negativa que é a que possui falha de componentes não relacionadas ao sistema pneumático de caminhões e a classe positiva que sucessivamente possui as falhas relacionadas ao sistema pneumático.

A realização do estudo de caso será feita em cima de 60000 instâncias, possuindo 170 características medidas para cada uma delas, conforme demonstrado na tabela demonstrada na Figura 10.

Figura 10 – Instâncias faltantes no banco de dados aplicado

Dados - Orange

Info
60000 instances
170 features (8.3 % missing data)
Target with 2 values
No meta attributes

Variables
☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

	class	aa_000	ab_000	cd_000	ch_000
1	neg	76698	?	1209600	0
2	neg	33058	?	1209600	?
3	neg	41040	?	1209600	0
4	neg	12	0	1209600	0
5	neg	60874	?	1209600	0
6	neg	38312	?	1209600	0
7	neg	14	0	1209600	?
8	neg	102960	?	1209600	0
9	neg	78696	?	1209600	?
10	pos	153204	0	1209600	?
11	neg	39196	?	1209600	0
12	neg	45912	?	1209600	0
13	neg	2104	?	1209600	0
14	neg	118950	?	1209600	0
15	neg	24416	?	1209600	?

Fonte: Autoria Própria (2022)

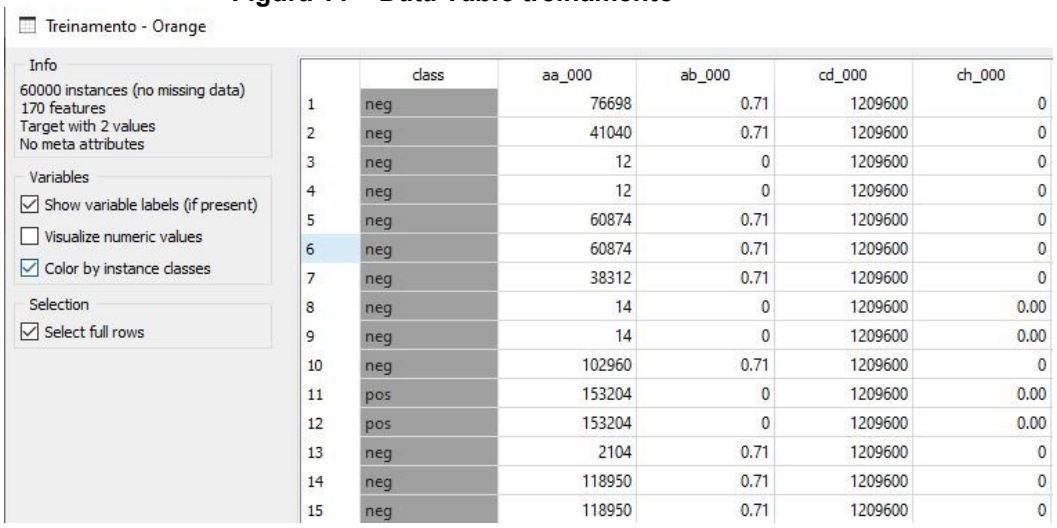
Das 60000 instâncias, tem-se 59000 pertencendo a classe negativa e 1000 pertencente a classe positiva. Dos 10200000 dados numéricos aferidos nessas 60000 instâncias, nota-se a inexistência de valores em 8,3% (846600 dados), os quais estão

definidas no conjunto como “na”, devido possivelmente ter acontecido algum erro durante a sua medição. Devido a motivos de propriedade intelectual e de preservação da marca, os nomes de cada característica e as unidades das mesmas não estão definidos.

A realização do tratamento desses valores inexistentes foi efetuado de duas formas. Primeiramente, em ambas as análises os dados em formato csv foram convertidos para o formato de excel (xlsx) para que a partir disso, os dados definidos como “na” fossem localizados e substituídos seguindo duas abordagens descritas a seguir:

Na primeira abordagem os valores definidos no banco de dados por “na” foram substituídos por um valor em branco. Após essa etapa foi realizado o *input* dos dados no *Orange* conforme demonstrado na Figura 10, para essas instancias inexistentes foi efetuado um pré-processamento atribuindo-se a média dos valores do banco de dados, conforme demonstrado na Figura 11, onde nota-se que todos os valores foram automaticamente preenchidos. Essa tratativa foi realizada buscando evitar de que o programa não fizesse a interpretação dos textos junto as instâncias numéricas como uma categoria e sim como um dado numérico.

Figura 11 – *Data Table* treinamento



	class	aa_000	ab_000	cd_000	ch_000
1	neg	76698	0.71	1209600	0
2	neg	41040	0.71	1209600	0
3	neg	12	0	1209600	0
4	neg	12	0	1209600	0
5	neg	60874	0.71	1209600	0
6	neg	60874	0.71	1209600	0
7	neg	38312	0.71	1209600	0
8	neg	14	0	1209600	0.00
9	neg	14	0	1209600	0.00
10	neg	102960	0.71	1209600	0
11	pos	153204	0	1209600	0.00
12	pos	153204	0	1209600	0.00
13	neg	2104	0.71	1209600	0
14	neg	118950	0.71	1209600	0
15	neg	118950	0.71	1209600	0

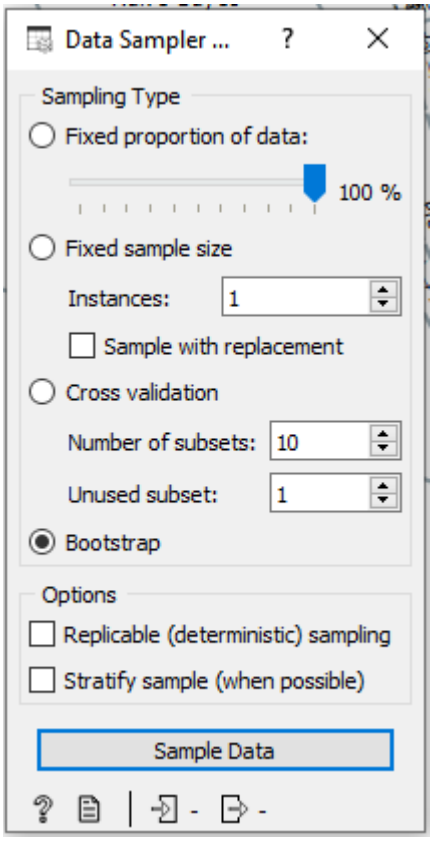
Fonte: Autoria Própria (2022)

A segunda abordagem utilizada, foi algo empírico e efetuada a partir da sugestão dos autores. Segundo, Schauhan (2019) diversas técnicas foram aplicadas

no banco de dados, porém foi obtido um melhor resultado substituindo-se os valores de “na” por -1.

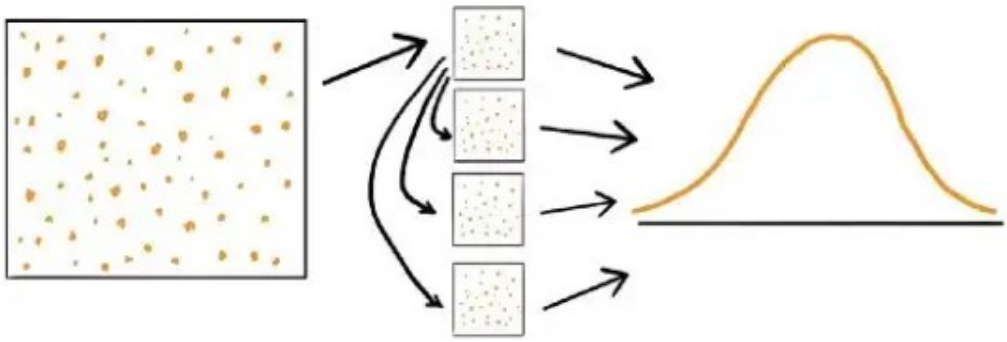
A estratégia utilizada para a divisão dos dados em treinamento e teste para que houvesse uma maior credibilidade no estudo de caso foi o *bootstrap* conforme demonstrado na Figura 12. Esse método consiste na reordenação dos dados existentes, fragmentando um dado aferido e simulando diversos outros dados a partir deste, conforme ilustrado na Figura 13. Das 60000 instâncias existente no banco de dados o *Bootstrap* gerou 21966 instâncias quando utilizado a média para serem definidos os dados inexistente e para uma futura realização de predição conforme demonstrado na Figura 14. Com o treinamento utilizando -1 nos dados inexistentes, o *Bootstrap* gerou 22066 conforme demonstrado na Figura 15.

Figura 12 – Separação dos dados em treinamento e teste utilizando *Bootstrap*



Fonte: Autoria Própria (2022)

Figura 13 – Ilustração do processo de *bootstraping*



Fonte: Joseph (2020)

Figura 14 – Dados de teste gerado a partir do *Bootstrap* utilizando a média nos dados não existentes

Teste - Orange

Info
21966 instances (no missing data)
170 features
Target with 2 values
No meta attributes

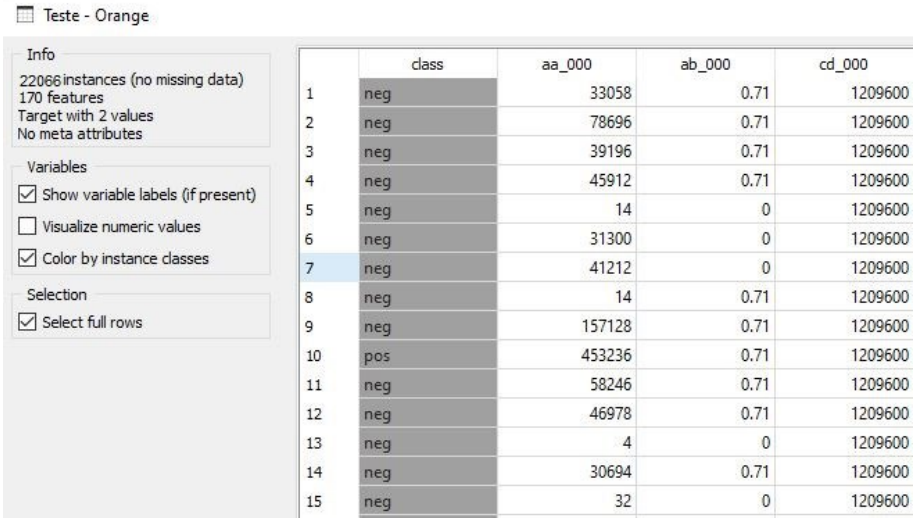
Variables
☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

	class	aa_000	ab_000	cd_000
1	neg	33058	0.71	1209600
2	neg	78696	0.71	1209600
3	neg	39196	0.71	1209600
4	neg	45912	0.71	1209600
5	neg	14	0	1209600
6	neg	31300	0	1209600
7	neg	41212	0	1209600
8	neg	14	0.71	1209600
9	neg	157128	0.71	1209600
10	pos	453236	0.71	1209600
11	neg	58246	0.71	1209600
12	neg	46978	0.71	1209600
13	neg	4	0	1209600
14	neg	30694	0.71	1209600
15	neg	32	0	1209600

Fonte: Autoria Própria (2022)

Figura 15 – Dados de teste gerado a partir do *Bootstrap* utilizando -1 nos dados não existentes



	class	aa_000	ab_000	cd_000
1	neg	33058	0.71	1209600
2	neg	78696	0.71	1209600
3	neg	39196	0.71	1209600
4	neg	45912	0.71	1209600
5	neg	14	0	1209600
6	neg	31300	0	1209600
7	neg	41212	0	1209600
8	neg	14	0.71	1209600
9	neg	157128	0.71	1209600
10	pos	453236	0.71	1209600
11	neg	58246	0.71	1209600
12	neg	46978	0.71	1209600
13	neg	4	0	1209600
14	neg	30694	0.71	1209600
15	neg	32	0	1209600

Fonte: Autoria Própria (2022)

4.2 Análise e discussão

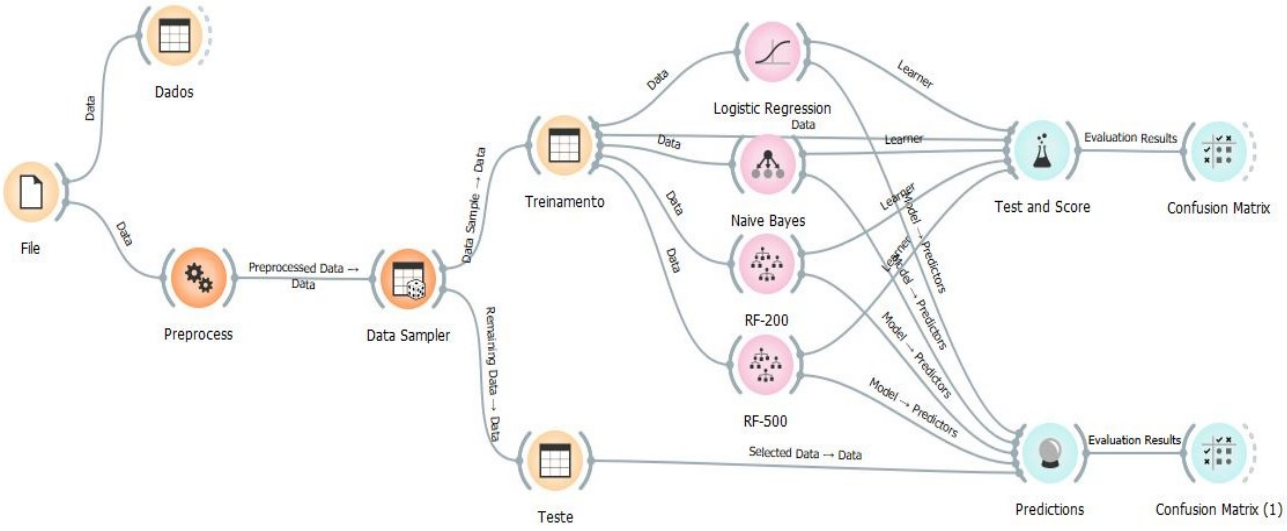
O conjunto de dados de treinamento demonstrado previamente na figura 11 será submetido a três métodos de aprendizagem, sendo eles: *Naive Bayes*, *Logistic Regression* e *Random Forest*.

Os dados de teste demonstrados nas figuras 14 e 15, serão processados pela ferramenta “Previsões” para futuramente compararmos a assertividade deste em relação aos dados de treinamento que foram submetidos aos métodos de aprendizagem.

A partir da utilização dessa métrica busca-se estimar o algoritmo com maior acurácia de acertos e que sucessivamente acarretará em um menor custo fazendo predições incorretas.

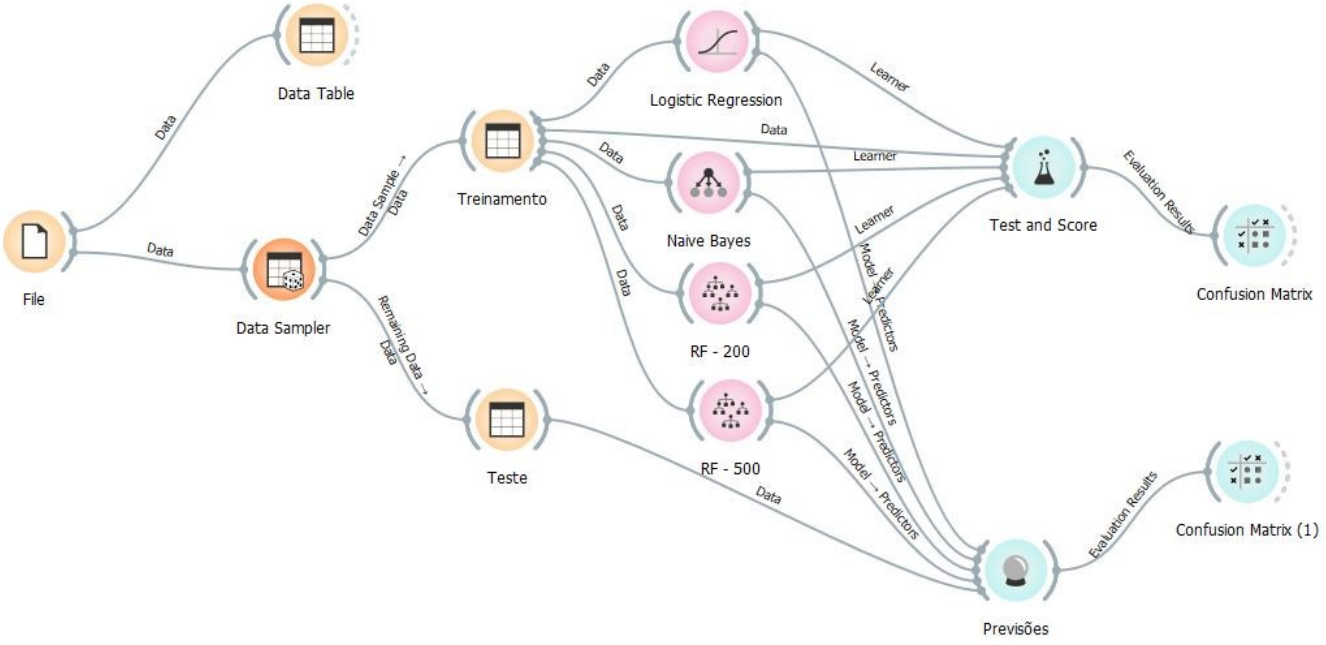
Nas figuras 16 e 17 demonstradas a seguir, temos a demonstração do *workflow* contendo os três algoritmos utilizados no *Orange* nas duas abordagens de estudo. No algoritmo de *Random Forest*, foi definido aleatoriamente o valor de 200 e 500 árvores para a geração dos resultados.

Figura 16 – *Workflow* desenvolvido no *Orange* utilizando média nos dados não existentes.



Fonte: Autoria Própria (2022)

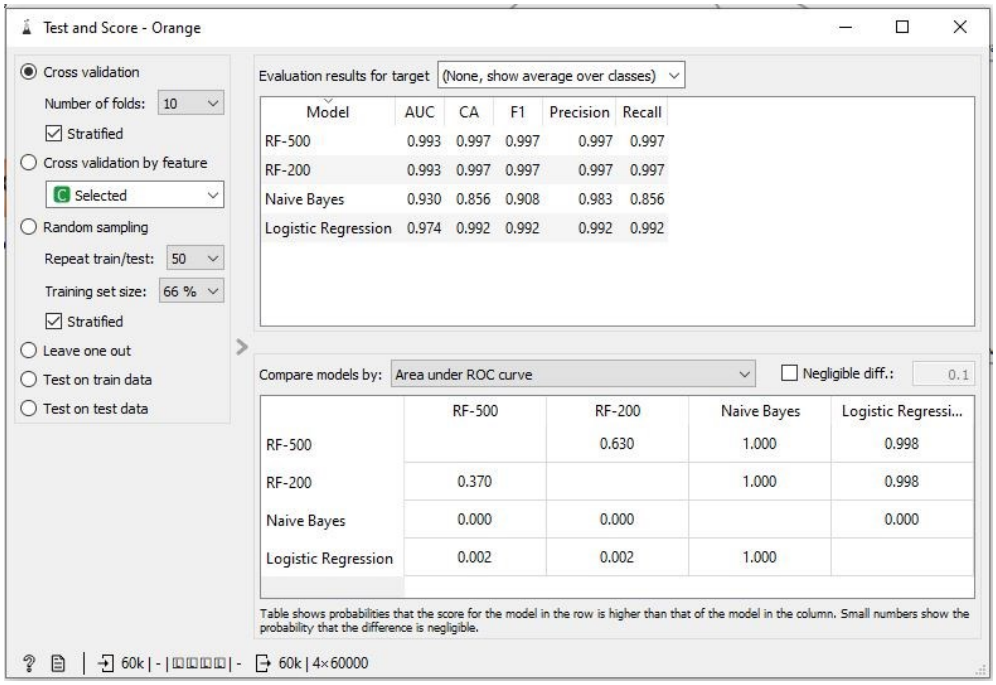
Figura 17 – *Workflow* desenvolvido no *Orange* utilizando -1 nos dados não existentes



Fonte: Autoria Própria (2022)

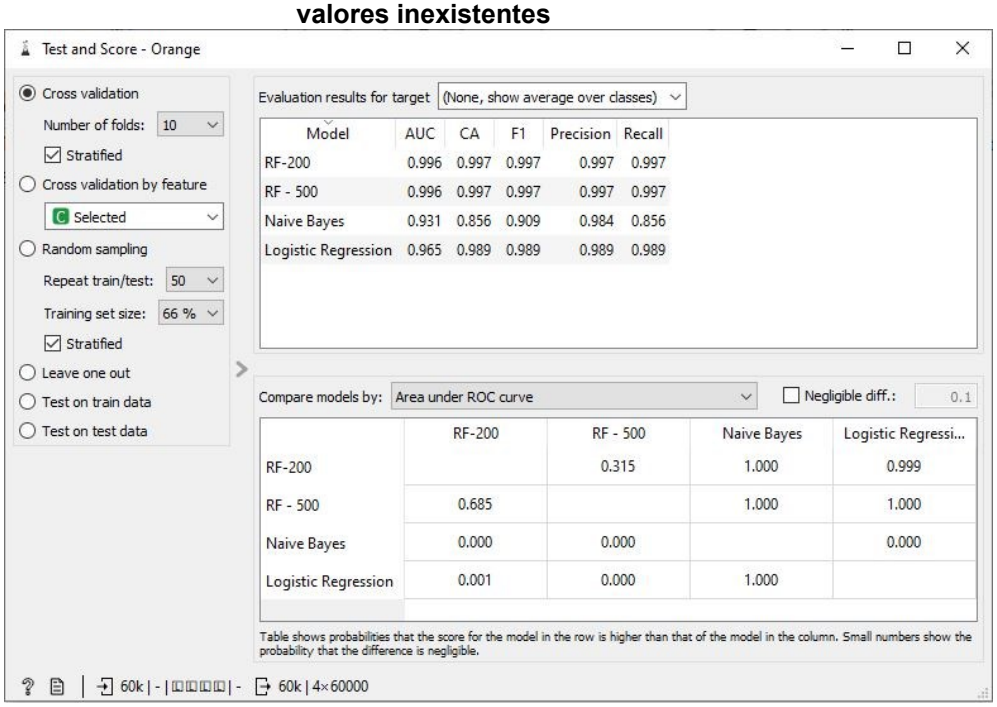
Após a realização de treinamentos e análises, o programa retornou os seguintes valores de *Classification Accuracy* (CA), *F1 Score,* Precisão e *Recall* no treinamento e nas predições para cada algoritmo utilizado, demonstrados nas figuras 18, 19, 20 e 21, sendo possível notar uma maior acurácia na utilização do algoritmo de *Random Forest* com maior número de árvores.

Figura 18 – Acurácia medida em cada algoritmo utilizado no treinamento utilizando a média nos valores inexistentes



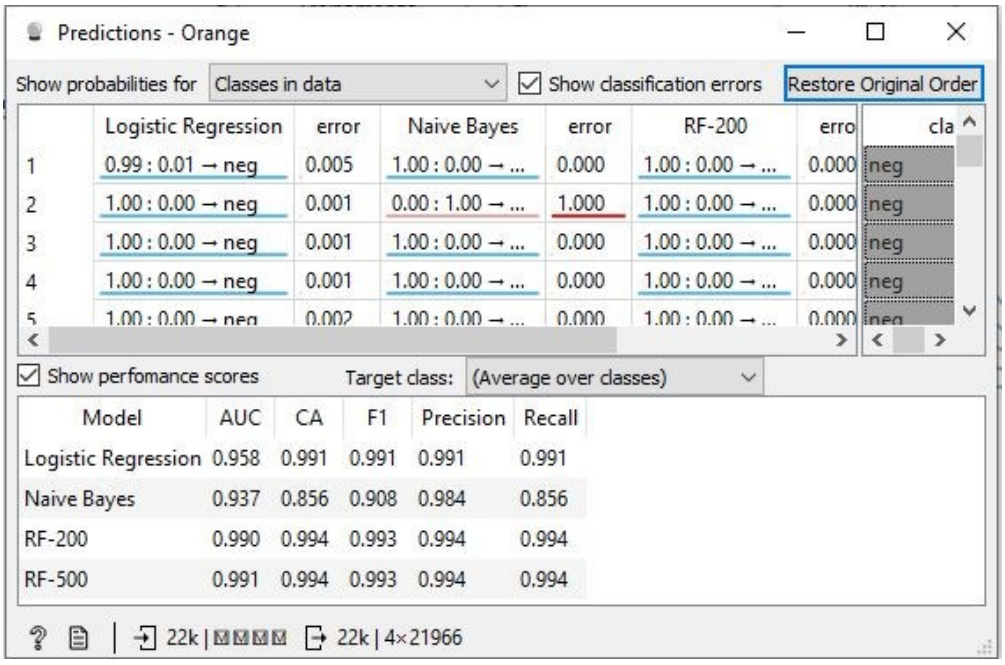
Fonte: Autoria Própria (2022).

Figura 19 – Acurácia medida em cada algoritmo utilizado no treinamento utilizando -1 nos valores inexistentes



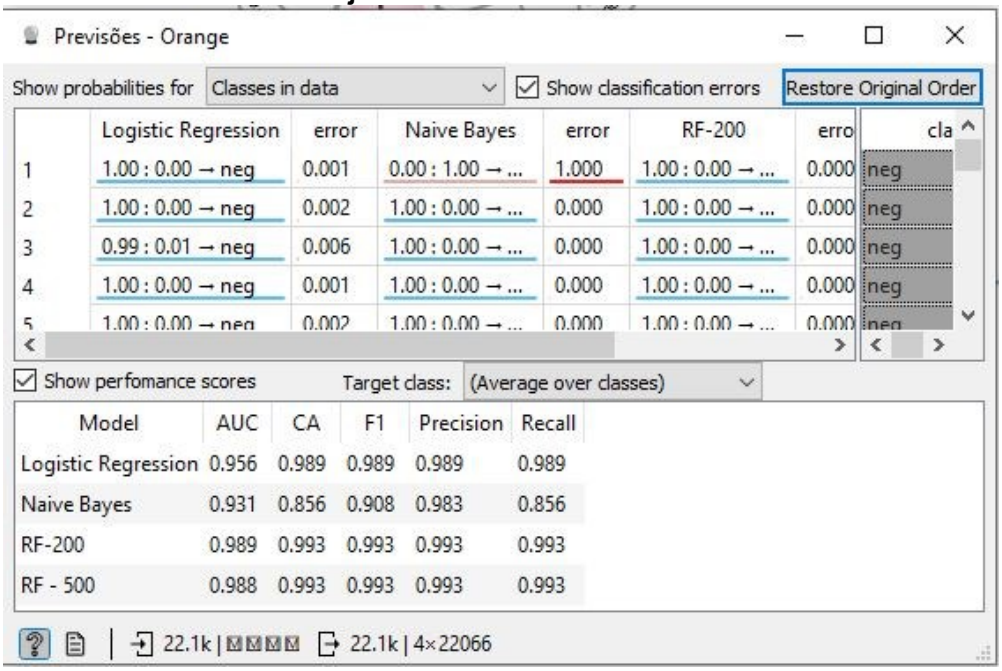
Fonte: Autoria Própria (2022)

Figura 20 – Previsões e acurácia no conjunto de teste utilizando a média nos valores inexistentes



Fonte: Autoria Própria (2022)

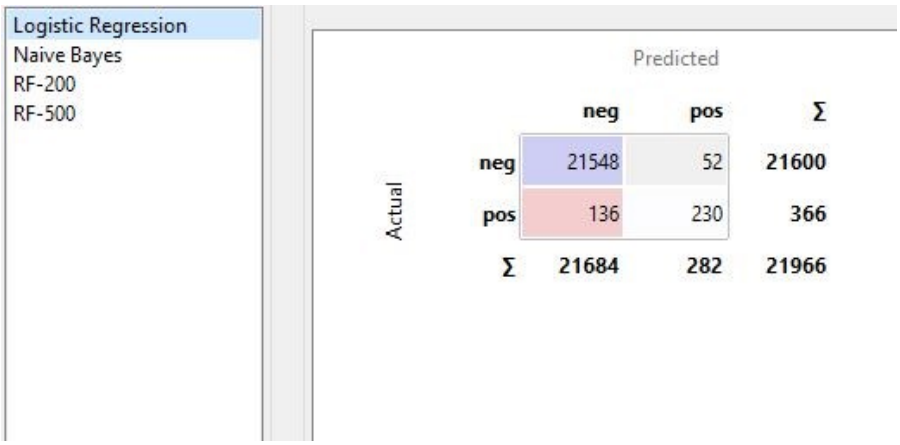
Figura 21 - Previsões e acurácia no conjunto de teste utilizando -1 nos valores inexistentes



Fonte: Autoria Própria (2022)

Conforme demonstrados nas figuras 20 e 21, podemos alcançar resultados bem satisfatórios na utilização desses algoritmos independente do tratamento utilizado nos dados inexistentes. O algoritmo que obteve uma maior quantidade de acertos na predição dos dados foi o *Random* Forest tendo uma acurácia de 99,4% com a utilização de 500 árvores e 200 árvores, seguido da Regressão logística com 98,9% e Naive Bayes com 85,6%.

Figura 22 – Demonstração de resultados obtidos na predição em regressão logística utilizando a média nos valores inexistentes

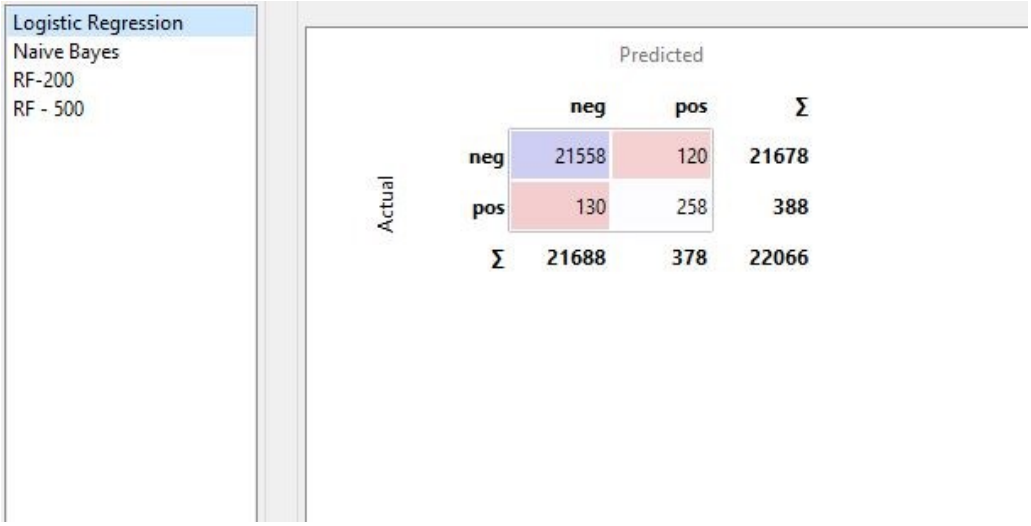


Fonte: Autoria Própria (2022)

O custo total na predição obtida, utilizando-se a média dos valores existentes para o tratamento dos dados inexistentes a partir da regressão logística, conforme demonstrado na Figura 22 foi:

$$CustoTotalRL_{média} = 52 * 10 + 136 * 500 = 68520 \quad (9)$$

Figura 23 - Demonstração de resultados obtidos na predição em regressão logística utilizando -1 nos valores inexistentes

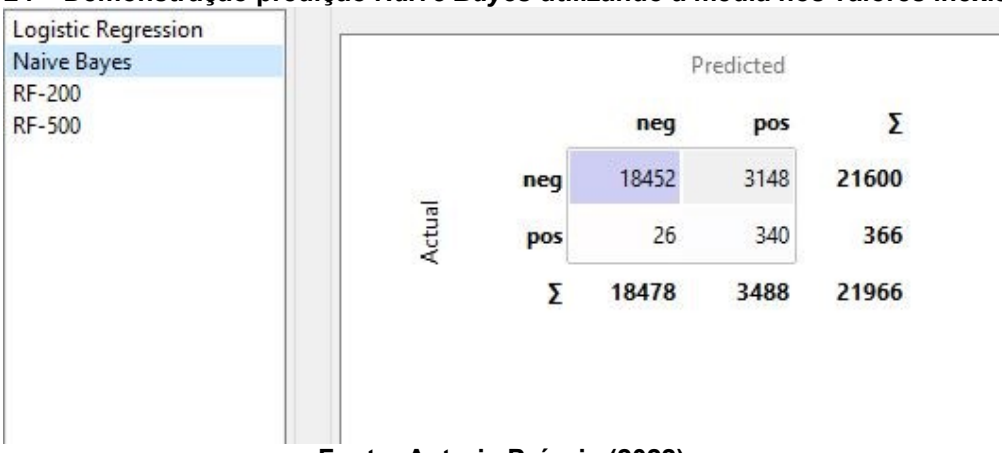


Fonte: Autoria Própria (2022)

O custo total na predição obtida, definindo como -1 no tratamento dos dados inexistentes a partir da regressão logística, conforme demonstrado na Figura 23 foi:

$$CustoTotalRL_{-1} = 120 * 10 + 130 * 500 = 66200 \quad (10)$$

Figura 24 – Demonstração predição *Naive Bayes* utilizando a média nos valores inexistentes

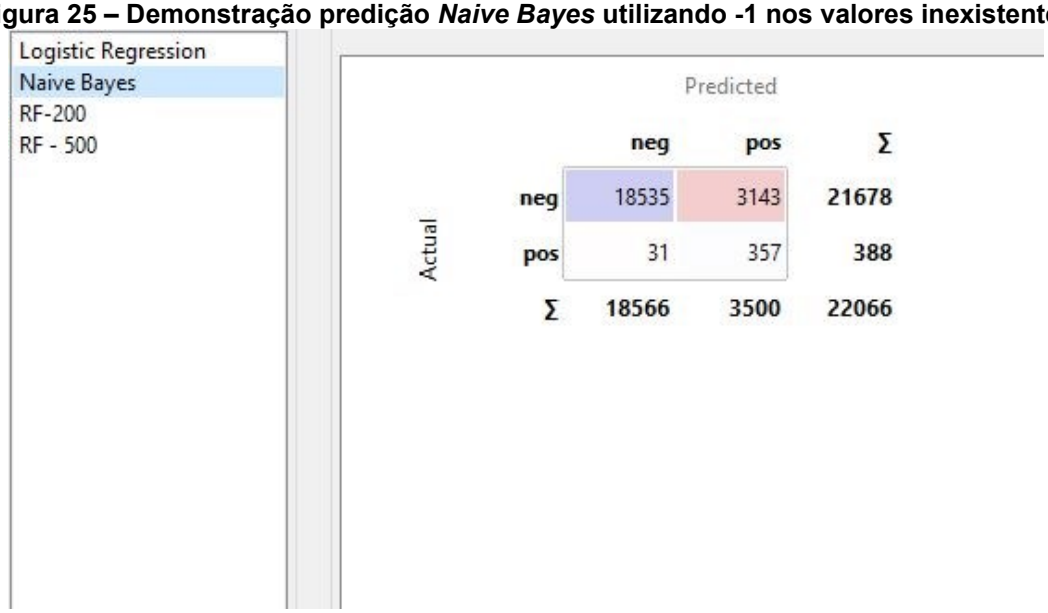


Fonte: Autoria Própria (2022)

O custo total na predição obtida a partir de *Naive Bayes* utilizando-se a média dos valores existentes para o tratamento dos dados inexistentes, conforme demonstrado na Figura 24 foi:

$$CustoTotalNB_{média} = 3148 * 10 + 26 * 500 = 44480 \quad (11)$$

Figura 25 – Demonstração predição *Naive Bayes* utilizando -1 nos valores inexistentes

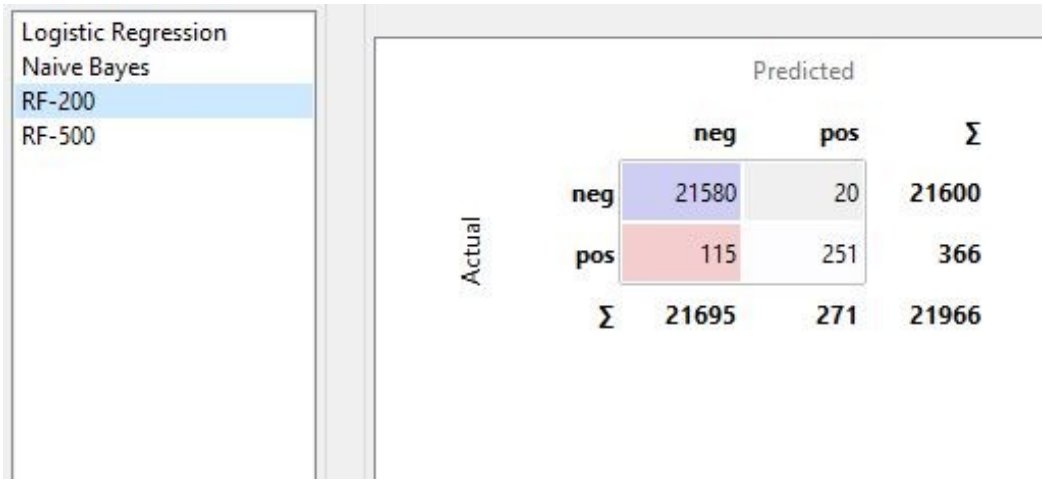


Fonte: Autoria Própria (2022)

O custo total na predição obtida, definindo como -1 no tratamento dos dados inexistentes a partir de *Naive Bayes*, conforme demonstrado na Figura 25 foi:

$$CustoTotalNB_{-1} = 3143 * 10 + 31 * 500 = 46930 \quad (12)$$

Figura 26 – Demonstração predição *Random Forest* com 200 árvores utilizando a média nos valores inexistentes

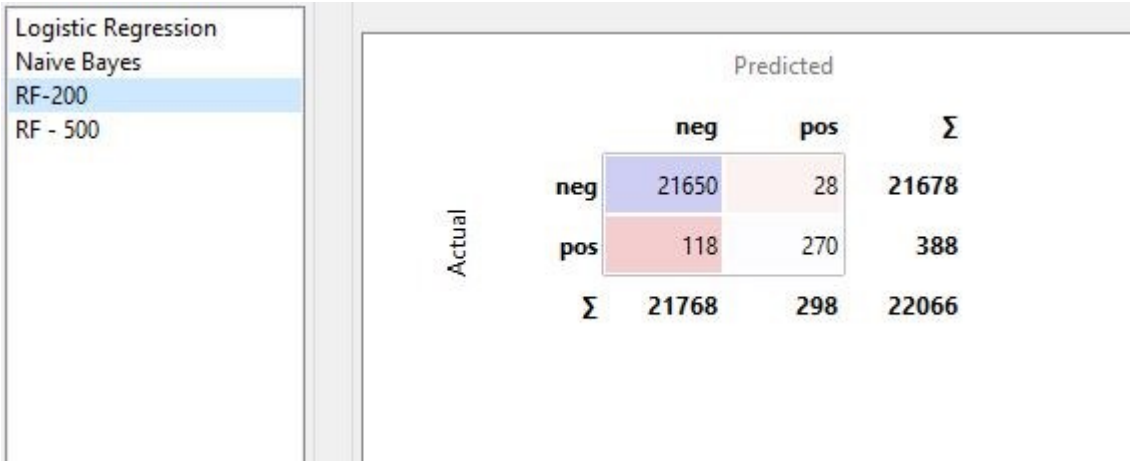


Fonte: Autoria Própria (2022)

O custo total na predição obtida a partir de *Random Forest* com 200 árvores, utilizando-se a média dos valores existentes para o tratamento dos dados inexistentes, conforme demonstrado na Figura 26 foi:

$$CustoTotalRF_{200_{média}} = 20 * 10 + 115 * 500 = 57700 \quad (13)$$

Figura 27 – Demonstração predição *Random Forest* com 200 árvores utilizando -1 nos valores inexistentes

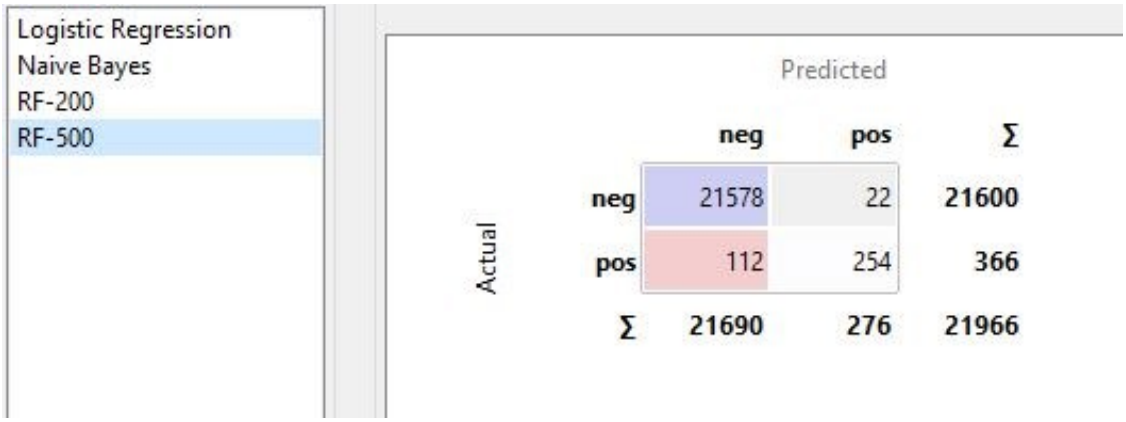


Fonte: Autoria Própria (2022)

O custo total na predição obtida, definindo como -1 no tratamento dos dados inexistentes a partir de *Random Forest* com 200 árvores, conforme demonstrado na Figura 27 foi:

$$CustoTotalRF_{200_{-1}} = 28 * 10 + 118 * 500 = 59280 \quad (14)$$

Figura 28 – Demonstração predição *Random Forest* com 500 árvores utilizando a média nos valores inexistentes

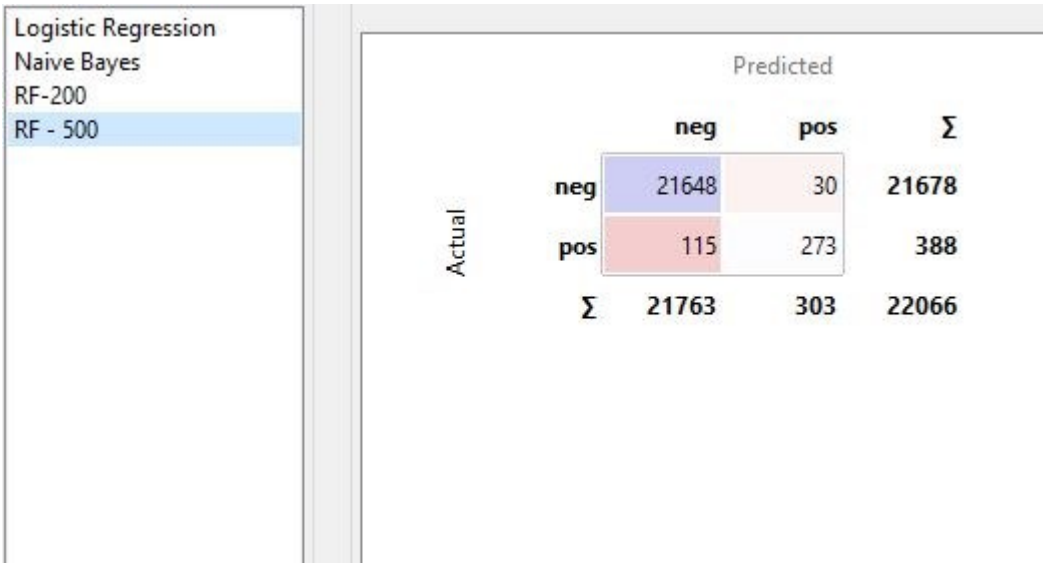


Fonte: Autoria Própria (2022)

O custo total na predição obtida a partir de *Random Forest* com 500 árvores, utilizando-se a média dos valores existentes para o tratamento dos dados inexistentes, conforme demonstrado na Figura 28 foi:

$$CustoTotalRF_{500_{média}} = 22 * 10 + 112 * 500 = 56220 \quad (15)$$

Figura 29 – Demonstração predição *Random Forest* com 500 árvores utilizando -1 nos valores inexistentes



Fonte: Autoria Própria (2022)

O custo total na predição obtida, definindo como -1 no tratamento dos dados inexistentes a partir de *Random Forest* com 500 árvores, conforme demonstrado na Figura 29 foi:

$$CustoTotalRF_{500_{-1}} = 30 * 10 + 115 * 500 = 57800 \quad (16)$$

Apesar da obtenção de resultados superiores nos algoritmos de *Random Forest* e Regressão logística em métricas como *Classification Acuracy, F1 Score,* Precisão e *Recall*, o algoritmo de *Naive Bayes* obteve uma maior exatidão na identificação das instâncias positivas dos dados, tendo esta classe um peso maior na definição da métrica de custos, sendo esta uma métrica na qual se buscava a obtenção do menor valor possível.

Notou-se também que os resultados gerados a partir do tratamento estabelecendo uma média nos dados ao invés da definição de todos os valores

inexistentes como -1, possuiu uma maior acurácia no treinamento e na previsão dos dados.

Dessa forma, concluímos que apesar de o algoritmo de *Random Forest* e Regressão logística apresentarem uma maior acurácia de acertos, para o foco do estudo que era a redução de custos a partir da predição antecipada de falhas em sistemas pneumáticos, os valores obtidos no algoritmo de *Naive Bayes* foram superiores conforme demonstrado na Figura 30.

Figura 30 - Custos obtidos na predição utilizando cada um dos algoritmos

Algoritmo	Custos Obtidos	
	Média	-1
Regressão Logística	68520	66200
Naive Bayes	44800	46930
Random Forest (200)	57700	59280
Random Forest (500)	56220	57800

Fonte: Autoria Própria (2022)

Nota-se também a partir da Figura 30, que com exceção do algoritmo de Regressão Logística, obteve-se melhores resultados na predição dos custos utilizando a média nos valores inexistentes no banco de dados.

Os valores de custos mais elevados quando aplicados aos algoritmos com maior *Classification Acuracy, F1 Score,* Precisão e *Recall*, podem ser compreendidos com uma maior facilidade a partir da utilização da métrica de *Specificity*.

$$Specificity_{NB} = \frac{340}{366} = 92{,}898\% \quad (17)$$

$$Specificity_{RL} = \frac{230}{366} = 62{,}84\% \quad (18)$$

$$Specificity_{RF200} = \frac{251}{366} = 68{,}579\% \quad (19)$$

$$Specificity_{RF500} = \frac{254}{366} = 69{,}398\% \quad (20)$$

5 CONCLUSÃO

Com a constante evolução da tecnologia e a possibilidade da leitura de bancos de dados cada vez maiores, a predição de falhas em caminhões e máquinas em geral, é essencial para o aumento da vida útil dos componentes e a redução de gastos indevidos.

Devido a se tratar de um banco de dados real e acurácia medida ser acima de 70% em todos os algoritmos estudados, é demonstrado nesse estudo a viabilidade na prevenção de falhas em caminhões. Os quais podem resultar em uma diminuição no número de acidentes e sucessivamente a segurança dos condutores nas vias públicas aumentar.

O aplicativo Orange Data Mining possibilita a realização de diversos temas de estudos tendo como vantagem não necessitar ter conhecimento prévio em programação, devido a se tratar a um aplicativo 100% visual.

Para futuros estudos, existe a possibilidade do desenvolvimento de equipamentos para aferir os dados em caminhões em tempo real, podendo ser evitados não só problemas relacionados ao sistema pneumático como também aos que não estão relacionados a esse sistema.

REFERÊNCIAS

A.K., JAIN; M.N., MURTY; P.J., F. Data Clustering: A Review. **ACM Computing Survey (CSUR)**, v. 31, p. 264–323, 1999.

AHMED, M.; NASER MAHMOOD, A.; HU, J. A survey of network anomaly detection techniques. **Journal of Network and Computer Applications**, v. 60, p. 19–31, 2016.

AL-AMRI, R. et al. A Review of Machine Learning and Deep Learning Techniques for Anomaly Detection in IoT Data. 2021.

BREIMAN, L. Random Forest. **Kluwer Academic Publishers.**, v. 12343 LNCS, p. 503–515, 2001.

CALLIGARO, C. Proposta de fundamentos habilitadores para a gestão da manutenção em indústrias de processamento contínuo baseada nos principios da manutenção de classe mundial. **UFRGS**, v. 18, n. 1, p. 22–27, 2003.

CAUCHICK MIGUEL, P. A. Estudo de caso na engenharia de produção: estruturação e recomendações para sua condução. **Production**, v. 17, n. 1, p. 216–229, 2007.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly Detection : A Survey. **ACM Computing Survey (CSUR)**, v. 41, n. 3, p. 1–72, 2009.

CHAUHAN, P. **ML for Mechanical Engineers!!!. APS failure detection in Scania Trucks… | by prashant chauhan | Medium**. Disponível em: <https://medium.com/@prashant23/ml-for-mechanical-engineers-5c5abea430a1>. Acesso em: 20 nov. 2022.

CHEN, X. et al. Application of data-driven models to predictive maintenance: Bearing wear prediction at TATA steel. **Expert Systems with Applications**, v. 186, n. October 2012, p. 115699, 2021.

CRIMINISI, A.; SHOTTON, J.; KONUKOGLU, E. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. **Foundations and Trends in Computer Graphics and Vision**, v. 7, n. 2–3, p. 81–227, 2011.

GOURIVEAU, R.; RAMASSO, E.; ZERHOUNI, N. Strategies to face imbalanced and unlabelled data in PHM applications. **Chemical Engineering Transactions**, v. 33, p. 115–120, 2013.

JOSEPH, T. **Bootstrapping Statistics. What it is and why it's used. | by Trist'n Joseph | Towards Data Science**. Disponível em: <https://towardsdatascience.com/bootstrapping-statistics-what-it-is-and-why-its-used-e2fa29577307>. Acesso em: 20 nov. 2022.

LIMA, M.; AMORIM, F. **Random Forest - LAMFO**. Disponível em: <https://lamfo-unb.github.io/2020/07/08/Random-Forest/>. Acesso em: 9 nov. 2021.

MATA, F. F. D. G. DA. Investigando métodos inteligentes para detecção de anomalias em comportamento de insetos sociais. p. 66, 2017.

MINUSSI, J. A.; DAMACENA, C.; NESS JR, W. L. Um modelo de previsão de solvência utilizando regressão logística. **Revista de Administração Contemporânea**, v. 6, n. 3, p. 109–128, 2002.

MITCHELL, T.; BLUM, A. Combining Labeled and Unlabeled Data with. **Acm**, p. 92–100, 1998.

PAOLANTI, M. et al. Machine Learning approach for Predictive Maintenance in Industry 4.0. **2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications, MESA 2018**, p. 1–6, 2018.

QUATRINI, E. et al. Machine learning for anomaly detection and process phase classification to improve safety and maintenance activities. **Journal of Manufacturing Systems**, v. 56, n. November 2019, p. 117–132, 2020.

ROCHA, A. DE R. Naive Bayes Classifier Teaching Material. p. 1–9, 2006.

SAMUEL, A. L. Eight-move opening utilizing generalization learning. (See Appendix B, Game G-43.1 Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal**, p. 210–229, 1959.

SOLOMON, S. **Scania Trucks Air Pressure System Failure Prediction | by Sharath Solomon | Analytics Vidhya | Medium**. Disponível em:

<https://medium.com/analytics-vidhya/scania-trucks-air-pressure-system-failure-prediction-ad6c43539d38>. Acesso em: 20 nov. 2022.

SUTHARSSAN, T. et al. Prognostic and health management for engineering systems: a review of the data- driven approach and algorithms. **The Journal of Engineering**, v. 2015, n. 7, p. 215–222, 2015.

THEISSLER, A. et al. Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. **Reliability Engineering and System Safety**, v. 215, 1 nov. 2021.

TOBON-MEJIA, D. A. et al. A data-driven failure prognostics method based on mixture of gaussians hidden markov models. **IEEE Transactions on Reliability**, v. 61, n. 2, p. 491–503, 2012.

WANG, K. Intelligent Predictive Maintenance (IPdM) system – Industry 4.0 scenario. **34th International Manufacturing Conference**, v. 113, n. August, 2017.

ANEXO A - Lei n. 9.610, de 19 de fevereiro de 1998

Presidência da República
Casa Civil
Subchefia para Assuntos Jurídicos

LEI Nº 9.610, DE 19 DE FEVEREIRO DE 1998[1].

Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências.

O PRESIDENTE DA REPÚBLICA Faço saber que o Congresso Nacional decreta e eu sanciono a seguinte Lei:

Título I - Disposições Preliminares

Art. 1º Esta Lei regula os direitos autorais, entendendo-se sob esta denominação os direitos de autor e os que lhes são conexos.

Art. 2º Os estrangeiros domiciliados no exterior gozarão da proteção assegurada nos acordos, convenções e tratados em vigor no Brasil.

Parágrafo único. Aplica-se o disposto nesta Lei aos nacionais ou pessoas domiciliadas em país que assegure aos brasileiros ou pessoas domiciliadas no Brasil a reciprocidade na proteção aos direitos autorais ou equivalentes.

Art. 3º Os direitos autorais reputam-se, para os efeitos legais, bens móveis.

Art. 4º Interpretam-se restritivamente os negócios jurídicos sobre os direitos autorais.

Art. 5º Para os efeitos desta Lei, considera-se:

I - publicação - o oferecimento de obra literária, artística ou científica ao conhecimento do público, com o consentimento do autor, ou de qualquer outro titular de direito de autor, por qualquer forma ou processo;

II - transmissão ou emissão - a difusão de sons ou de sons e imagens, por meio de ondas radioelétricas; sinais de satélite; fio, cabo ou outro condutor; meios óticos ou qualquer outro processo eletromagnético;

III - retransmissão - a emissão simultânea da transmissão de uma empresa por outra;

IV - distribuição - a colocação à disposição do público do original ou cópia de obras literárias, artísticas ou científicas, interpretações ou execuções fixadas e fonogramas, mediante a venda, locação ou qualquer outra forma de transferência de propriedade ou posse;

V - comunicação ao público - ato mediante o qual a obra é colocada ao alcance do público, por qualquer meio ou procedimento e que não consista na distribuição de exemplares;

VI - reprodução - a cópia de um ou vários exemplares de uma obra literária, artística ou científica ou de um fonograma, de qualquer forma tangível, incluindo qualquer armazenamento permanente ou temporário por meios eletrônicos ou qualquer outro meio de fixação que venha a ser desenvolvido;

VII - contrafação - a reprodução não autorizada;

VIII - obra:

a) em co-autoria - quando é criada em comum, por dois ou mais autores;

b) anônima - quando não se indica o nome do autor, por sua vontade ou por ser desconhecido;

c) pseudônima - quando o autor se oculta sob nome suposto;

d) inédita - a que não haja sido objeto de publicação;

e) póstuma - a que se publique após a morte do autor;

f) originária - a criação primígena;

g) derivada - a que, constituindo criação intelectual nova, resulta da transformação de obra originária;

h) coletiva - a criada por iniciativa, organização e responsabilidade de uma pessoa física ou jurídica, que a publica sob seu nome ou marca e que é constituída pela participação de diferentes autores, cujas contribuições se fundem numa criação autônoma;

i) audiovisual - a que resulta da fixação de imagens com ou sem som, que tenha a finalidade de criar, por meio de sua reprodução, a impressão de movimento, independentemente dos processos de sua captação, do suporte usado inicial ou posteriormente para fixá-lo, bem como dos meios utilizados para sua veiculação;

IX - fonograma - toda fixação de sons de uma execução ou interpretação ou de outros sons, ou de uma representação de sons que não seja uma fixação incluída em uma obra audiovisual;

X - editor - a pessoa física ou jurídica à qual se atribui o direito exclusivo de reprodução da obra e o dever de divulgá-la, nos limites previstos no contrato de edição;

XI - produtor - a pessoa física ou jurídica que toma a iniciativa e tem a responsabilidade econômica da primeira fixação do fonograma ou da obra audiovisual, qualquer que seja a natureza do suporte utilizado;

XII - radiodifusão - a transmissão sem fio, inclusive por satélites, de sons ou imagens e sons ou das representações desses, para recepção ao público e a transmissão de sinais codificados, quando os meios de decodificação sejam oferecidos ao público pelo organismo de radiodifusão ou com seu consentimento;

XIII - artistas intérpretes ou executantes - todos os atores, cantores, músicos, bailarinos ou outras pessoas que representem um papel, cantem, recitem, declamem, interpretem ou executem em qualquer forma obras literárias ou artísticas ou expressões do folclore.

Art. 6º Não serão de domínio da União, dos Estados, do Distrito Federal ou dos Municípios as obras por eles simplesmente subvencionadas.

---

[1] Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9610.htm.