Apresentação de TCC
MBA Data Science 2019

ANÁLISE DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE FALHA MECÂNICA DE EQUIPAMENTOS FLORESTAIS

Aluno: **Jean Carlos da Cruz**
Orientador: Denise Prado Kronbauer
Faculdade de Tecnologia SENAI Mato Grosso – Cuiabá-MT

FATECSENAI
Faculdade de Tecnologia SENAI Mato Grosso

AGENDA


INTRODUÇÃO
FUNDAMENTAÇÃO TEÓRICA
MATERIAIS E MÉTODOS
RESULTADOS E DISCUSSÃO
CONCLUSÕES
REFERÊNCIAS

INTRODUÇÃO

- Com a intensificação de uma economia globalizada, houve um crescimento na demanda por equipamentos e sistemas com melhor desempenho aliado ao baixo custo;

- Nesse contexto, entender a confiabilidade, se faz uma importante via de estudo e inovação, considerando que falhas podem levar a um aumento dos custos dos produtos ou até acidentes;

- Sendo assim, a confiabilidade pode ser definida como a probabilidade de um sistema operar de maneira satisfatória (sem falhas) em um período de tempo conhecido e condições definidas;

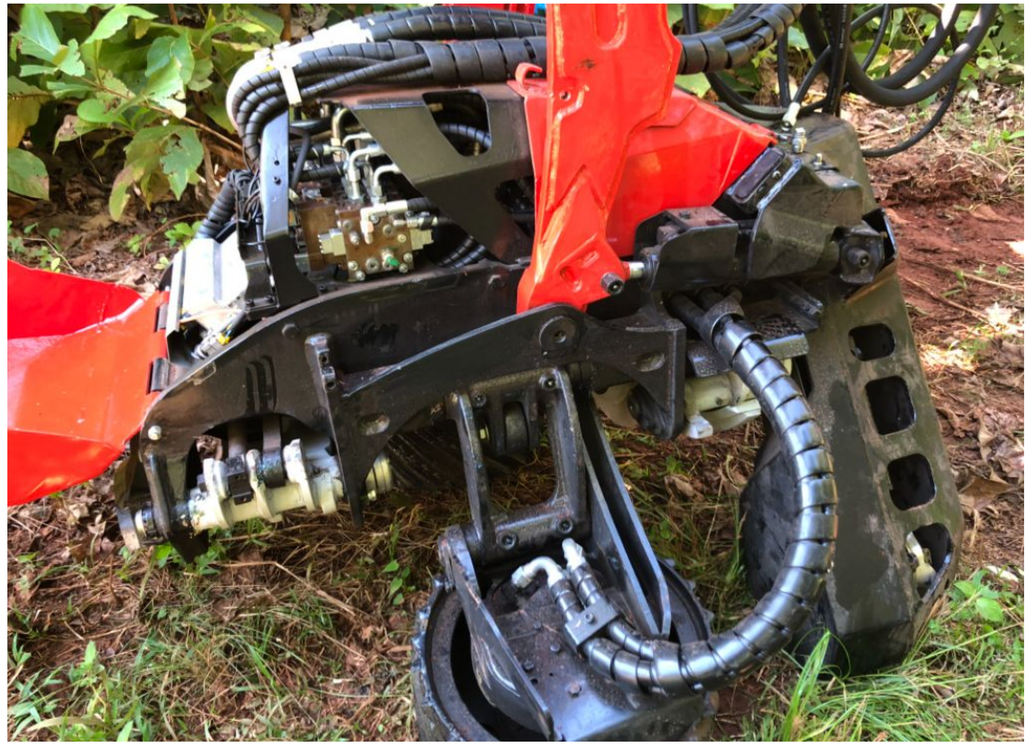
- Com base na confiabilidade dos equipamentos é possível definir com maior precisão o planejamento das manutenções preventivas, sendo essas as manutenções que ocorrem antes do equipamento falhar.

Fonte: Acervo Jean Cruz

INTRODUÇÃO

- Como saída para esse problema, podemos citar a aplicação de técnicas de aprendizado de máquina, descritas como técnicas que utilizam algoritmos capazes de aprender de acordo com as respostas esperadas por meio associações de diferentes dados. Estes dados podem ser números, imagens e tudo que possa ser identificado por essa tecnologia;

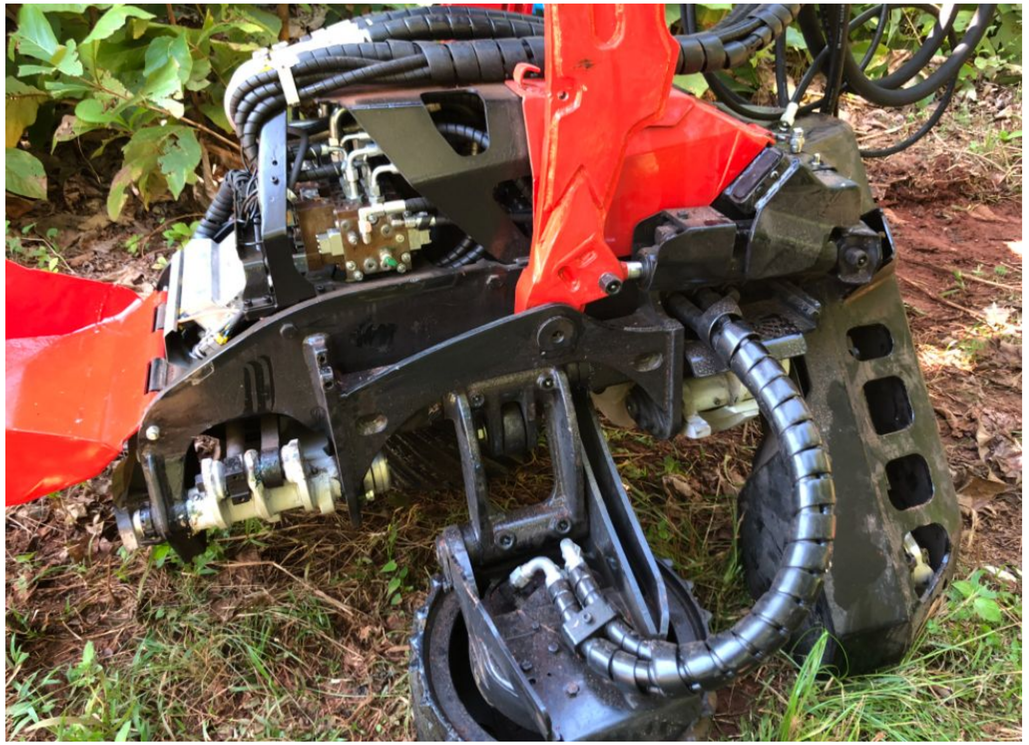
- Portanto, objetivou-se neste trabalho analisar modelos de aprendizado de máquina com o objetivo de predizer o tempo médio de falha. Para tanto, foi utilizado um dataset com 534 dados de manutenções realizadas com reparo de terminais e mangueiras hidráulicas, no período de janeiro de 2020 a janeiro de 2021, em uma empresa com 42 equipamentos, localizada no estado de Mato Grosso;



Fonte: Acervo Jean Cruz

INTRODUÇÃO

- Foram analisados os modelos de regressão, regressão linear múltipla, árvore de decisão, floresta aleatória, gradient boosting e regressão K vizinhos, onde concluiu-se que tais modelos não apresentam boa eficácia para predição do tempo médio a falha, em função da alta variabilidade das variáveis independentes e problemas estruturais nos dados obtidos.



Fonte: Acervo Jean Cruz

FUNDAMENTAÇÃO TEÓRICA

- **FALHA EM EQUIPAMENTOS MÓVEIS:**

  - A falha pode ser definida como uma inoperância de um produto, que não executa a função para a qual foi projetado (WUTTKE; SELLITO, 2008). Uma falha pode gerar uma situação indesejada como uma simples parada de máquina, prejuízos financeiros, e até algo pior como o risco de vidas humanas, logo, não devem ser poupados esforços para minimizar e evitar os riscos de uma falha (LAFRAIA, 2001);

  - Nesse contexto, novas técnicas foram desenvolvidas no setor de manutenção, dentre elas a manutenção preditiva, que tem como objetivo minimizar ou evitar a queda no desempenho seguindo um plano previamente elaborado, baseada nos intervalos definidos de tempos em tempos, sempre visando prolongar a vida útil das máquinas e equipamentos, garantindo assim o aumento da eficiência e da produtividade (KARDEC; NASCIF, 2009).

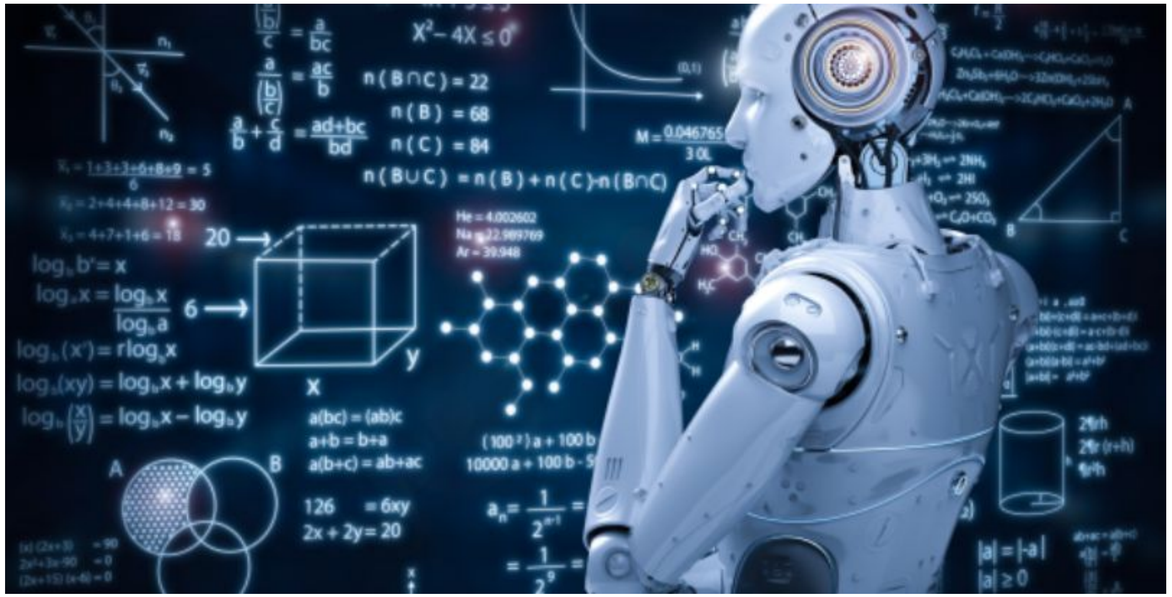
Fonte: Acervo Jean Cruz

FUNDAMENTAÇÃO TEÓRICA

- **APRENDIZADO DE MÁQUINA:**

  - Aprendizado de máquina (do inglês Machine Learning) é o nome que se dá em Ciência da Computação à técnica baseada nos princípios do aprendizado indutivo, onde algoritmos processam um conjunto de dados e extraem um modelo capaz de representar os intervalo de dados. Tais modelos podem ser usados também para representar um dado não amostrado (PERES; ROCHA; BISCARO, 2012);

  - As principais formas de aprendizado de máquinas são o aprendizado supervisionado e o não supervisionado. No aprendizado **supervisionado**, o algoritmo aprende a extrair informações de dados previamente conhecidos e classificados, sendo que após a execução do modelo testa-se a eficácia do aprendizado em dados desconhecidos.



Fonte: Google Images

FUNDAMENTAÇÃO TEÓRICA

- **APRENDIZADO DE MÁQUINA:**

  - *Regressão Linear Múltipla*: Modelo matemático que associa a variável **dependente**, ou seja, aquela que se deseja conhecer ou estimar, com as **independentes** (também chamadas de regressoras). Nesse modelo, assumimos que existe uma relação linear entre a variável dependente e *n* variáveis independentes (RODRIGUES, 2012) ;

  - *Árvore de decisão*: Modelo onde uma árvore de decisão constitui uma estrutura de dados que pode ser definida recursivamente como um nó folha que corresponde a uma classe ou um nó de decisão que contém um teste sobre algum atributo de interesse. Em cada resultado do teste existe uma possível aresta para um subárvore. Cada subárvore produzida tem a mesma estrutura que a árvore principal (MONARD; BARANAUSKAS, 2003).

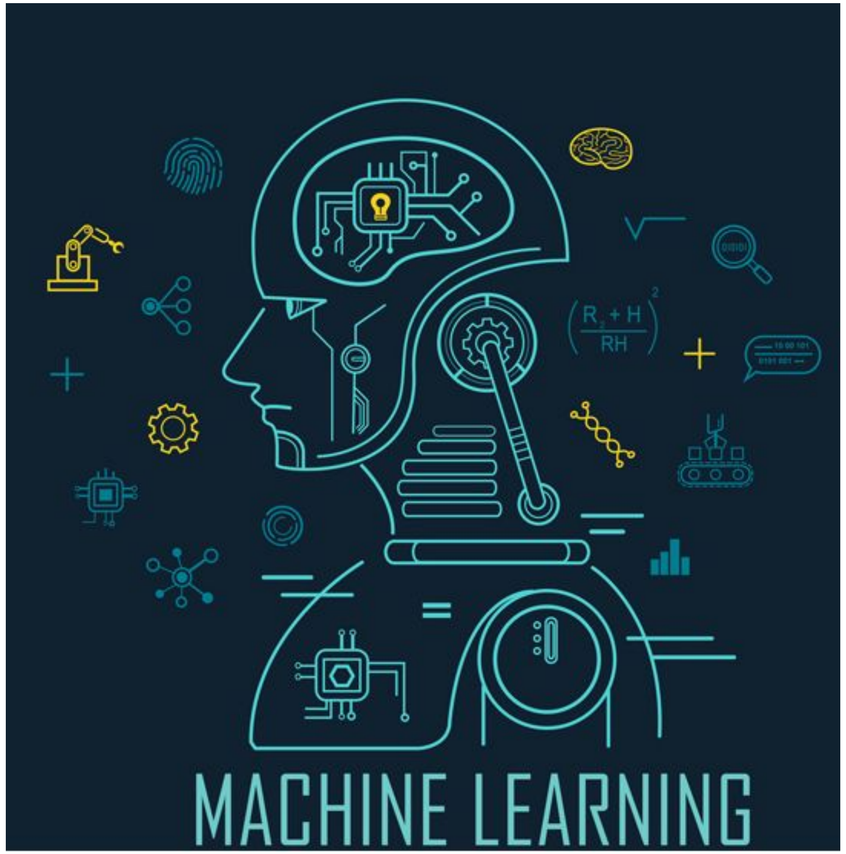
  - *Floresta Aleatória*: Constituem conjuntos de árvores de decisão, onde cada árvore é treinada com um subconjunto da base de dados disponível, sendo sua seleção realizada de forma aleatória (RIQUETI; RIBEIRO; ZÁRATE, 2018);

  - *Gradient Boosting*: Modelo de previsão onde se ajustam novos modelos de previsão fracos, geralmente árvores de decisão . Ela constrói o modelo em etapas, objetivando-se estimativas mais apuradas da variável dependente, onde temos novas bases de aprendizado para ser correlacionada ao máximo com o gradiente negativo da função de perda de todo o conjunto. Dessa forma, o gradient boosting proporciona muita liberdade durante a modelagem, tornando a escolha da função de perda mais apropriada uma questão de tentativa e erro (NATEKIN; KNOLL, 2013).

FUNDAMENTAÇÃO TEÓRICA

- **APRENDIZADO DE MÁQUINA:**

  - *Regressão K Vizinhos Mais Próximos*: Esse método é do tipo não paramétrico, uma vez que não há um modelo a ser ajustado. O princípio que molda a regressão K vizinhos é a procura por um número pré definido de amostras de treinamento que são as mais próximas em distância de um novo ponto e dali predizer e catalogar. Essa distância, em geral, pode ser qualquer métrica de medida, sendo a distância euclidiana a mais comum entre elas. Apesar de ser simples, k vizinhos obtém bons resultados em problemas de classificação e regressão (PEDREGOSA *et al.,* 2011).



Fonte: buffaloboy / Shutterstock.com

FUNDAMENTAÇÃO TEÓRICA

- **PYTHON:**
  - Criado por Guido van Rossum no final dos anos 80, como sucessor da linguagem ABC, o Python teve sua primeira publicação em 1991 na versão 0.9.0 (ROSSUM, 2009);

  - Python em comparação a outras linguagens não utiliza colchetes para delimitar blocos, apresenta também menos exceções sintáticas e casos especiais que linguagens como C ou Pascal (PYTHON SOFTWARE FOUNDATION, 2012);

  - Atualmente a grande quantidade de pacotes em bibliotecas disponíveis é considerada uma das maiores forças da linguagem Python (PIOTROWSKI, 2006). Em setembro de 2021 o repositório oficial de pacotes de terceiros continha mais de 300.000 pacotes com os mais diversos objetivos, tais quais: análise de dados, aprendizado de máquina, desenvolvimento de aplicativos mobile, processamento de texto, processamento de imagens, etc.

- **PYCARET:**
  - A biblioteca de código aberto em Python PyCaret, é uma ferramenta de baixo código que automatiza fluxos de trabalho de aprendizado de máquina. A biblioteca possui uma solução de ponta a ponta de gestão de modelos que acelera de forma exponencial o ciclo de experimentação e aumenta a produtividade dos projetos que a utilizam (MOEZ, 2020);

  - A biblioteca encapsula diversas outras bibliotecas de aprendizado de máquina, tais quais: scikit-learn, XGBoost, LightGBM, CatBoost, spaCy, Optuna, Hyperopt, Ray, etc. (MOEZ, 2020).

Fonte: https://pycaret.org/

MATERIAIS E MÉTODOS

- **LOCAL E PROCESSOS:**
  - O presente estudo foi realizado a partir de dados de manutenção mecânica realizados em equipamentos que atuam na etapa de colheita florestal de Tectona grandis L. F. (Teca) em três regiões do estado de Mato Grosso (Cáceres, Rosário Oeste e Tangará da Serra). A empresa dispõe de **42** equipamentos que atuam em três etapas no processo de colheita, sendo eles: Corte, Arraste, Baldeio e Carga;

  - A empresa possui departamento de manutenção próprio, composto por um time de mais de 10 pessoas (entre líderes, mecânicos e auxiliares), onde esse time realiza as manutenções de duas formas, sendo elas: Preventivas ou Corretivas;

  - A empresa dispõe de software de mercado que auxilia na coleta de dados de manutenção em campo. Uma vez inseridos, os dados são sincronizados com o banco de dados Microsoft SQL Server da companhia, onde as informações são armazenadas.

- **BASE DE DADOS:**
  - Ao total foram fornecidos 2.852 registros coletados entre os meses de janeiro de 2020 a janeiro de 2022, sendo fornecidos em formato .xlsx (Microsoft Excel). Os campos contidos no banco de dados são:

Campo	Tipo	Descrição
MAINTENANCE_ID	int	Código único utilizado para identificar cada manutenção
START_TIME	datetime	Início da Manutenção Realizada
END_TIME	datetime	Término da Manutenção Realizada
PART_TYPE	object	Tipo de peça utilizada na manutenção
SUB_SYSTEM	object	Subgrupo de sistema (Nomenclatura Interna)
MAINTENANCE_LOCAL	object	Local onde a Manutenção foi realizada
GROUP_NAME	object	Nome do grupo de manutenção
MAINTENANCE_TYPE	object	Tipo de manutenção realizada (Corretiva ou Preventiva)
SERVICE_NAME	object	O nome do serviço que foi realizado na manutenção
OBSERVATION	object	Observações que os mecânicos podem fazer em cada manutenção
PART_NUMBER	float	Número total de peças utilizado na manutenção
MACHINERY_ID	object	Código único utilizado para identificar cada equipamento
MACHINERY_HOURMETER	int	Horímetro do equipamento no momento do início da manutenção
MECHANIC	object	Nome do mecânico responsável pela manutenção
PURCHASE_DATE	datetime	Data que o equipamento foi adquirido
BUILT_YEAR	float	Ano que o equipamento foi montado
AGE	object	Idade dos equipamentos em meses quando a manutenção aconteceu
OPERATION	object	Operação Realizada pelo equipamento (Corte, Arraste, Baldeio ou Carga)

Quadro 1 - Tabela descritiva das variáveis disponíveis (Fonte: Teak Resources Company)

MATERIAIS E MÉTODOS

- **PROCESSAMENTO E MODELAGEM DOS DADOS:**
  - O processamento dos dados foi realizado utilizando a linguagem Python e as seguintes bibliotecas open source: Numpy, Pandas, csv, datetime, seaborn e PyCaret. O código utilizado foi construído utilizando o framework de notebook utilizando a ferramenta Jupyter em nuvem através do serviço grátis Google Colaboratory. Nessa etapa o trabalho foi dividido em três etapas, **limpeza e tratamento dos dados recebidos**, **análise exploratória** e **modelagem dos dados**;

  - *Limpeza e tratamento dos dados recebidos:* Foram removidas as colunas GROUP_NAME, MECHANIC, PURCHASE_DATE e BUILT_YEAR, uma vez que essas não têm utilização prática para o objetivo deste trabalho. Na base de dados havia informações de manutenções de equipamentos que não estão relacionados com a colheita florestal, dessa forma, esses registros foram removidos do conjunto.

  - *Análise exploratória*: Para guiar essa etapa foram definidas sete perguntas relevantes para o projeto, sendo elas:

    1- Qual equipamento e operação apresentou o maior número de manutenções?;

    2- Qual é o tempo médio para manutenção em horas por mês?;

    3- Qual é o tempo médio entre manutenções por máquina e grupo de operação?;

    4- Entre as manutenções realizadas, quantas são Corretivas? Preventivas?;

    5- Quais são as principais peças usadas nas manutenções?;

    6- Qual a estação operacional com mais manutenções registradas?;

    7- Qual é a idade média dos equipamentos por grupo de operação?

MATERIAIS E MÉTODOS

- **PROCESSAMENTO E MODELAGEM DOS DADOS:**
  - *Modelagem dos dados:* Nessa etapa de modelagem utilizou-se a biblioteca PyCaret. Os modelos selecionados foram: Regressão linear múltipla, Árvore de decisão, Floresta aleatória, Gradient boosting e regressão K vizinhos, por serem modelos de ampla utilização em nível acadêmico e privado;

  - Definiu-se o **erro absoluto médio** e o **erro quadrático médio** como as métricas de sucesso dos modelos a serem analisados. A primeira métrica é utilizada comumente na avaliação de modelos, pois mede o erro entre pares de observações e é de fácil explicação, uma vez que ela é construída na mesma unidade dos dados. A segunda métrica, o erro quadrático, também mede a magnitude do erro, porém ao obter o quadrado antes da média, leva a uma maior ponderação a altos erros.

  - Anteriormente à efetiva aplicação dos modelos de regressão, testou-se a distribuição das variáveis independentes que foram utilizadas bem como a correlação entre elas, utilizando o método de **Pearson**;

  - Após o primeiro teste dos modelos aplicados aos dados, obteve-se os resultados das métricas de sucesso. Com base nestes resultados selecionou-se o melhor modelo entre os analisados, e com base nesta seleção a ferramenta de "Tunning" do PyCaret foi utilizada com o objetivo de melhorar os resultados obtidos através do aumento do número de interações do modelo com os dados e alterando a métrica otimizadora para o erro médio absoluto. Após a etapa de tunning, realizou-se a análise de importância de cada variável independente utilizada no melhor modelo, bem como a análise de distribuição dos erros.

RESULTADOS E DISCUSSÃO

RESULTADOS BASEADOS NA ANÁLISE EXPLORATÓRIA DOS DADOS:

Por meio das perguntas formuladas na etapa de análise exploratória dos dados, observou-se padrões e tendências nos dados que serviram de embasamento para a seleção de variáveis independentes visando a predição do momento de falha.

1- QUAL EQUIPAMENTO E OPERAÇÃO APRESENTOU O MAIOR NÚMERO DE MANUTENÇÕES?

Foi possível evidenciar que o equipamento HV08 é o equipamento com o maior número de manutenções registradas. Em linhas gerais, é possível evidenciar também que equipamentos com o código TAC, que são usados na atividade de baldeio, se mostram fortemente presentes entre os dez primeiros registros visualizados, indicando que essa grupo de atividade tem a maior frequência de manutenções entre todas as analisadas, representando mais de 50% do total de manutenções realizadas

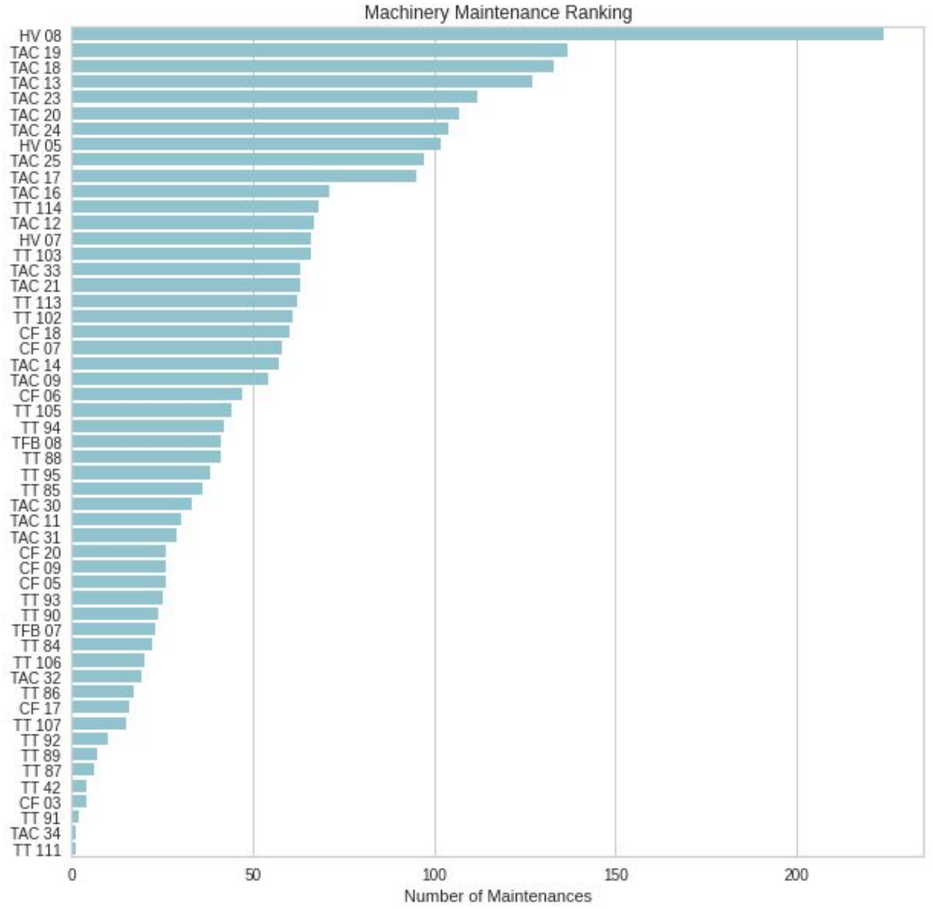


Figura 1 - Gráfico de barra ordenado para o número de manutenções por equipamento.

FATECSENAI
Faculdade de Tecnologia SENAI Mato Grosso

RESULTADOS E DISCUSSÃO

RESULTADOS BASEADOS NA ANÁLISE EXPLORATÓRIA DOS DADOS:

Por meio das perguntas formuladas na etapa de análise exploratória dos dados, observou-se padrões e tendências nos dados que serviram de embasamento para a seleção de variáveis independentes visando a predição do momento de falha.

2- QUAL É O TEMPO MÉDIO PARA MANUTENÇÃO EM HORAS POR MÊS?

O tempo médio para reparo nas manutenções dentro do conjunto de dados é de 43,96 horas. Analisando o resultado consolidado mês a mês são percebidos picos de aumento do tempo médio em relação a média, como exemplo, o mês de agosto de 2020 onde o tempo médio para manutenção foi superior a 300 horas (Figura 3). Este dado indica que algumas manutenções nesse período foram demasiado longas em relação ao habitual, provavelmente uma falha mais complexa que demandou maiores cuidados ou aquisição de peça de reposição que não estava disponível em estoque no momento que a falha ocorreu.

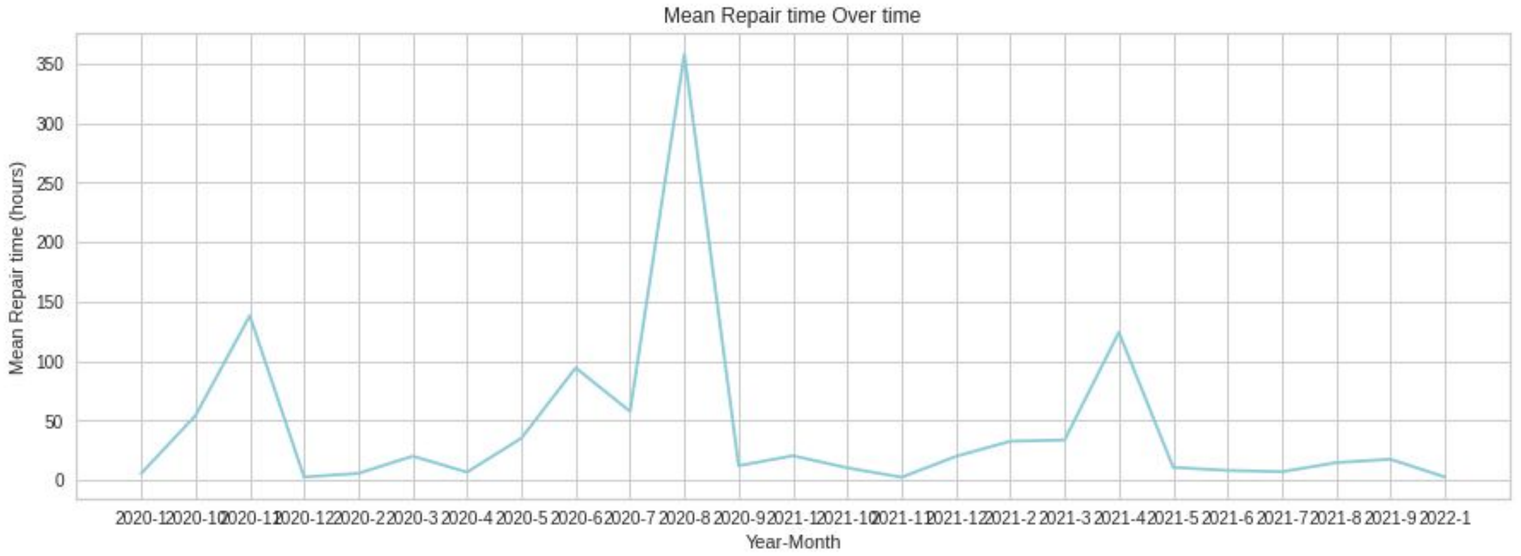


Figura 2 - Gráfico de linha ordenado por mês para o tempo médio para manutenção.

FATECSENAI
Faculdade de Tecnologia SENAI Mato Grosso

RESULTADOS E DISCUSSÃO

RESULTADOS BASEADOS NA ANÁLISE EXPLORATÓRIA DOS DADOS:

Por meio das perguntas formuladas na etapa de análise exploratória dos dados, observou-se padrões e tendências nos dados que serviram de embasamento para a seleção de variáveis independentes visando a predição do momento de falha.

3- QUAL É O TEMPO MÉDIO ENTRE MANUTENÇÕES POR MÁQUINA E GRUPO DE OPERAÇÃO?

Para o cálculo do tempo médio entre manutenções um novo campo foi calculado, uma vez que essa informação não estava disponível de forma explícita no conjunto de dados. Essa variável foi obtida através da diferença entre os horímetros registrados da manutenção analisada e a anterior, gerando tal diferença em horas. O cálculo dessa variável gerou os primeiros indicadores onde foi possível evidenciar resultados incoerentes, como exemplo, tempos médios negativos. Esse fato indicou que existia algum problema nos dados de horímetro apontado em cada manutenção.

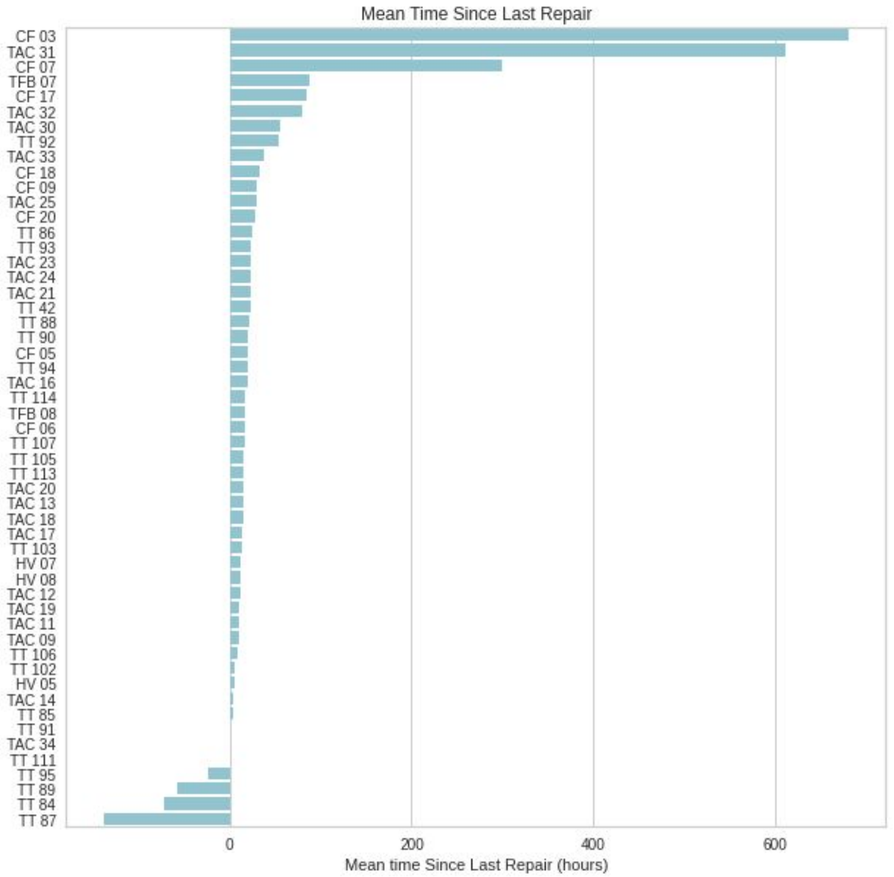


Figura 3 - Gráfico de linha ordenado por máquina para o tempo médio entre manutenções.

FATECSENAI
Faculdade de Tecnologia SENAI Mato Grosso

RESULTADOS E DISCUSSÃO

Sendo assim, não é possível utilizar tais números da maneira que foram disponibilizados na modelagem futura, sendo requerida uma análise mais detalhada sobre as possíveis correções. Para tal, foi plotada a dispersão do tempo médio entre manutenções por máquina (Figura 4).

Pelo gráfico de dispersão percebe-se que existem muitos outliers nos dados. Provavelmente esse problema tenha acontecido por falhas na digitação, erros de sincronização, erro entre o apontamento em papel e o aplicativo, falha da marcação do horímetro na máquina, etc. De qualquer maneira, para que essa variável pudesse ser utilizada posteriormente, foi necessário fazer a correção desses valores no conjunto de dados. Essa correção foi realizada em associação com os responsáveis pela manutenção da empresa, sendo que foram realizados 262 ajustes nas informações de horímetro e um novo gráfico de dispersão foi criado (Figura 5). Com as correções, os resultados agora refletem a realidade enfrentada em campo.

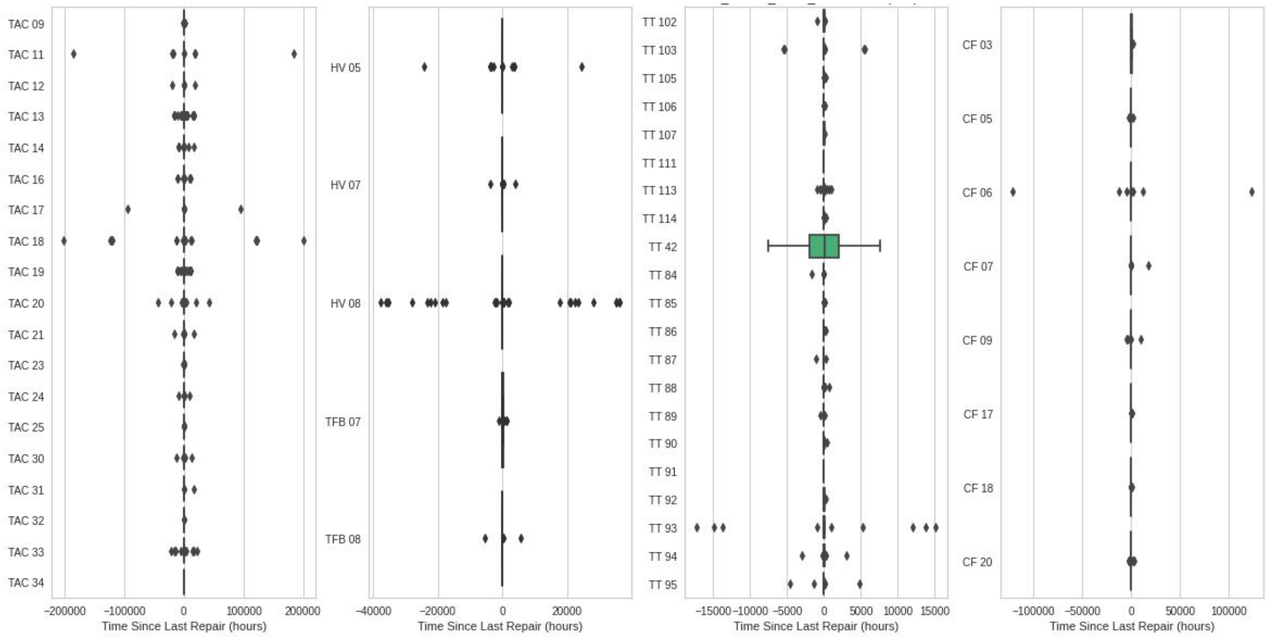


Figura 4 - Gráfico de dispersão por máquina para o tempo médio entre manutenções.

RESULTADOS E DISCUSSÃO

De acordo com os dados, é possível verificar que o tempo médio entre manutenções gerais é de 19 horas. Quando é analisado o tempo médio entre manutenções por grupo de operação, podemos ver que o maior tempo foi evidenciado na operação de Carga (26 horas), seguido por Baldeio (21 horas), Arraste (16 horas), Corte (13 horas), respectivamente.

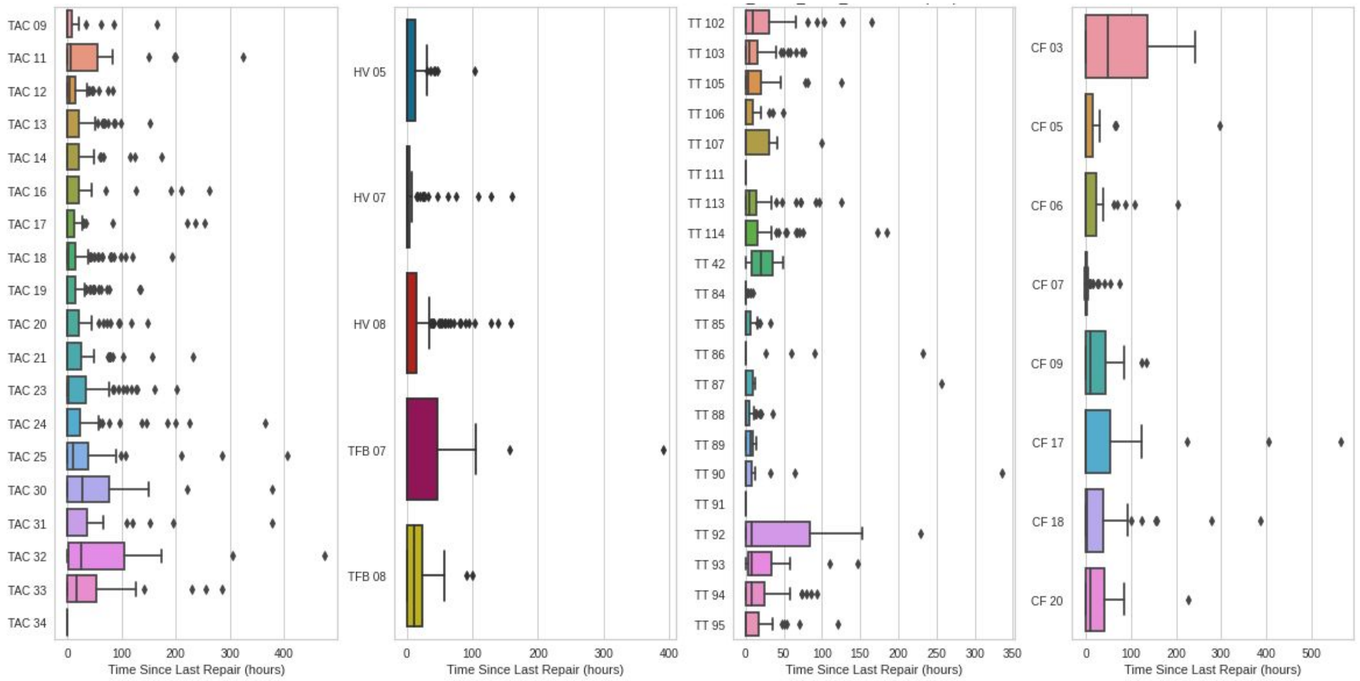


Figura 5 - Gráfico de dispersão ajustado por máquina para o tempo médio entre manutenções.

FATECSENAI
Faculdade de Tecnologia SENAI Mato Grosso

RESULTADOS E DISCUSSÃO

RESULTADOS BASEADOS NA ANÁLISE EXPLORATÓRIA DOS DADOS:

Por meio das perguntas formuladas na etapa de análise exploratória dos dados, observou-se padrões e tendências nos dados que serviram de embasamento para a seleção de variáveis independentes visando a predição do momento de falha.

4- ENTRE AS MANUTENÇÕES REALIZADAS, QUANTAS SÃO CORRETIVAS? PREVENTIVAS?

Do total de manutenções realizadas no conjunto de dados, 68,4% do total são manutenções corretivas, 29,6% são manutenções preventivas e 2% delas manutenções periódicas. Quando visualizado em uma escala mensal de tempo, podemos ver uma grande heterogeneidade no número de manutenções corretivas e preventivas, e estabilidade no número de periódicas.

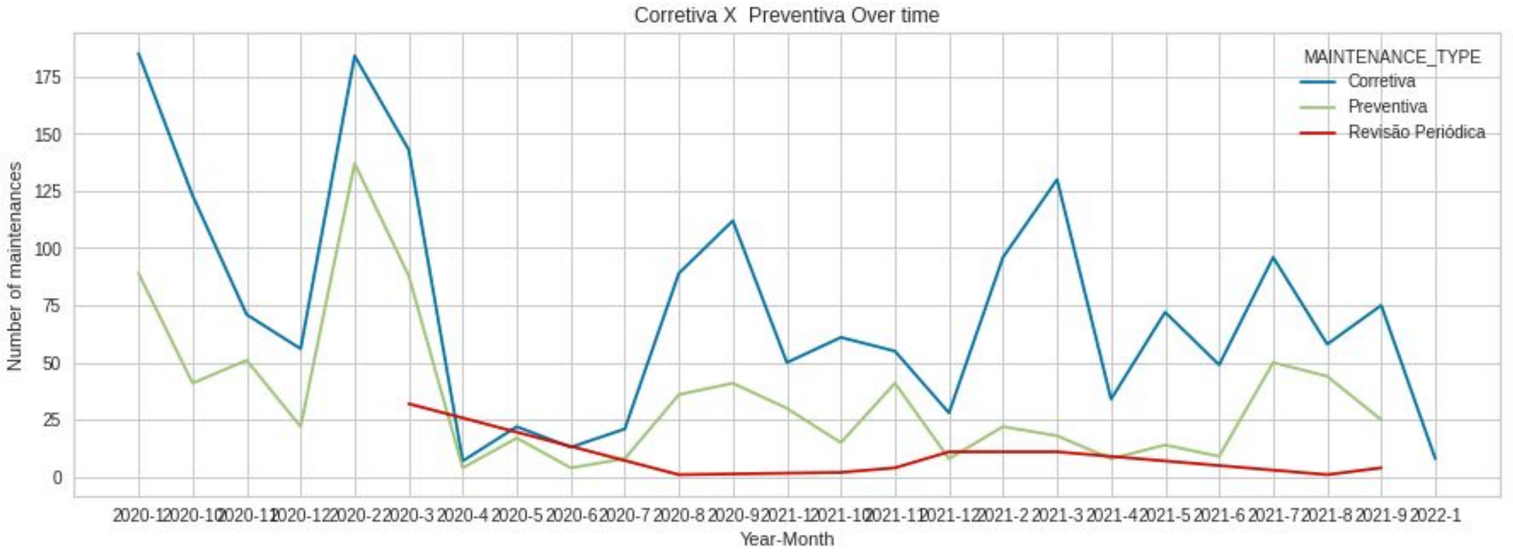


Figura 6 - Gráfico de linha ordenado por mês para o número de manutenções realizadas por tipo.

RESULTADOS E DISCUSSÃO

RESULTADOS BASEADOS NA ANÁLISE EXPLORATÓRIA DOS DADOS:

Por meio das perguntas formuladas na etapa de análise exploratória dos dados, observou-se padrões e tendências nos dados que serviram de embasamento para a seleção de variáveis independentes visando a predição do momento de falha.

5- QUAIS SÃO AS PRINCIPAIS PEÇAS USADAS NAS MANUTENÇÕES?

Terminal Hidráulico e Mangueira Hidráulica são as peças com maior frequência de utilização no conjunto de dados, aproximadamente 40% do total das manutenções, quando combinados. Como diversas peças são utilizadas nas manutenções, decidiu-se considerar para o futuro modelo de predição, apenas as manutenções desses dois tipos de peça, reduzindo a dispersão nos dados. Desta forma, foi possível evidenciar o tipo de manutenção mais frequente, ou seja, a maior dor da companhia.

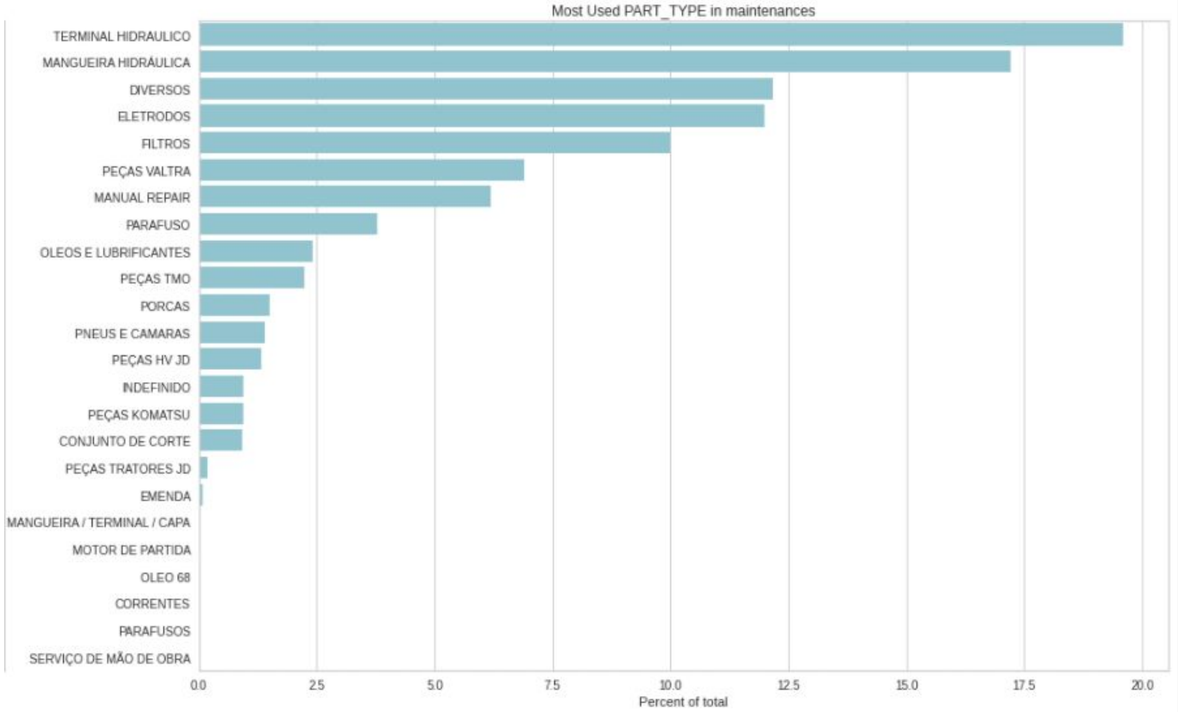


Figura 7 - Gráfico de barra ordenado por peça para o percentual do total de manutenções realizadas.

RESULTADOS E DISCUSSÃO

RESULTADOS BASEADOS NA ANÁLISE EXPLORATÓRIA DOS DADOS:

Por meio das perguntas formuladas na etapa de análise exploratória dos dados, observou-se padrões e tendências nos dados que serviram de embasamento para a seleção de variáveis independentes visando a predição do momento de falha.

6- QUAL A ESTAÇÃO OPERACIONAL COM MAIS MANUTENÇÕES REGISTRADAS?

A estação operacional é definida pela companhia como estação seca, que corresponde aos meses de abril, maio, junho, julho, agosto e setembro, sendo os meses mais secos do ano no estado do Mato Grosso. A estação chuvosa é definida pelos meses de outubro, novembro, dezembro, janeiro, fevereiro e março. Foi possível identificar que 65% do total de manutenções ocorreram na estação chuvosa. Esse resultado faz sentido, uma vez que, com o período chuvoso o terreno florestal e as próprias toras ficam mais pesadas em função da água acumulada, exigindo maior esforço dos equipamentos.

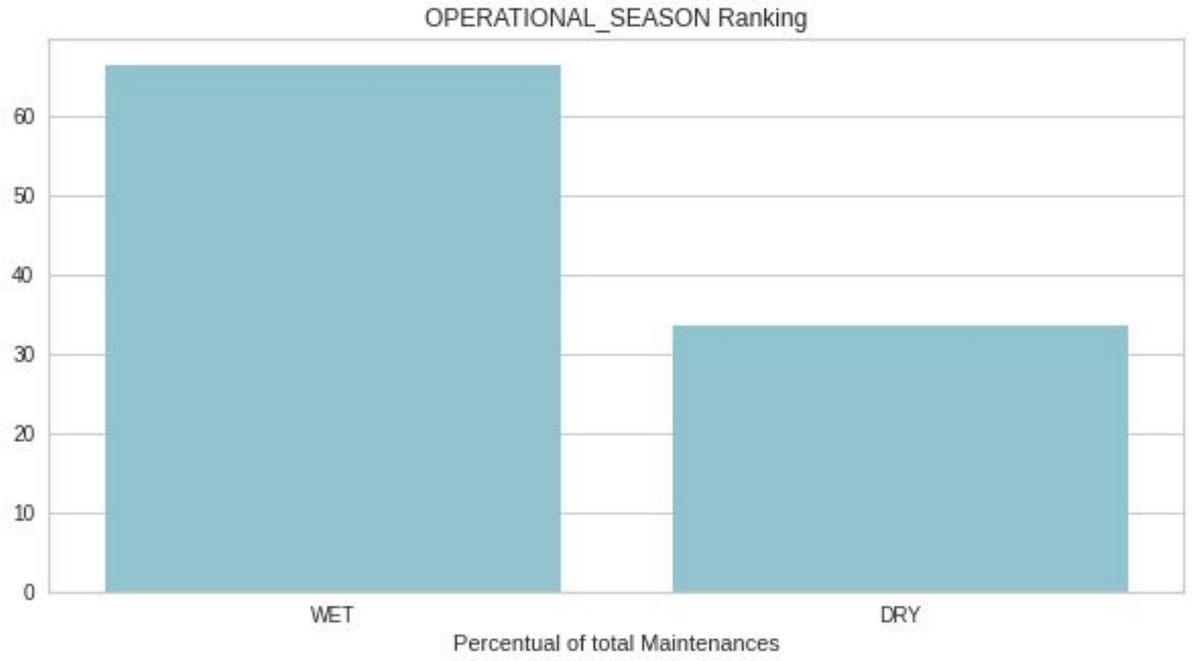


Figura 8 - Gráfico de barra ordenado por estação operacional.

RESULTADOS E DISCUSSÃO

RESULTADOS BASEADOS NA ANÁLISE EXPLORATÓRIA DOS DADOS:

Por meio das perguntas formuladas na etapa de análise exploratória dos dados, observou-se padrões e tendências nos dados que serviram de embasamento para a seleção de variáveis independentes visando a predição do momento de falha.

7- QUAL É A IDADE MÉDIA DOS EQUIPAMENTOS POR GRUPO DE OPERAÇÃO?

A idade média dos equipamentos é de 65 meses, resultado esse que indica que a maior parte dos equipamentos está próxima ou além do seu ciclo de vida estimado pelos fabricantes (em média 60 meses). Quando comparados os resultados por grupo de operação, a operação de arraste tem os equipamentos mais novos (27 meses) enquanto os equipamentos de carga são os mais velhos (100 meses). Devido a essa grande dispersão dentro da idade média, foi utilizada essa variável como independente na construção do modelo preditivo.

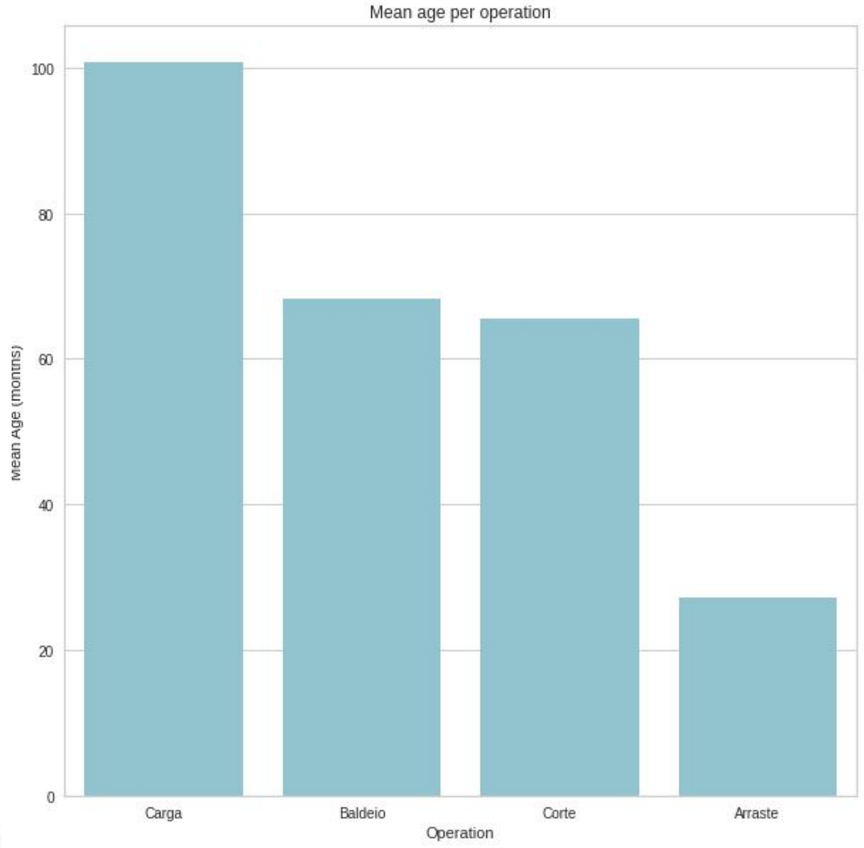


Figura 9 - Gráfico de barra ordenado por grupo de operação para a idade média em meses.

RESULTADOS E DISCUSSÃO

MODELAGEM:

- **ESTATÍSTICA DESCRITIVA**

Por meio da análise descritiva apresentada, observou-se que após todos os ajustes e definições realizadas, temos o total de 534 registros de manutenções disponíveis para a análise de modelos de preditivos. A média de horímetro é de 8473 horas, com desvio padrão de 9211 horas, ou seja, os dados de horímetro são bastante dispersos, fato esse que corrobora com o disposto no item 7 da etapa de análise exploratória, onde evidenciamos que a média de idade dos equipamentos é bastante dispersa.

	MAINTENANCE_ID	MACHINERY_HOURMETER	AGE	TIME_SINCE_LAST_REPAIR	TIME_TO_NEXT_REPAIR
count	534.000000	534.000000	534.000000	534.000000	534.000000
mean	81032.046816	8473.466292	63.172285	72.750936	72.750936
std	1376.985152	9211.408321	54.113704	128.150195	128.150195
min	78950.000000	19.000000	1.000000	0.000000	0.000000
25%	79567.250000	1427.000000	18.000000	0.000000	0.000000
50%	81205.500000	3077.000000	35.000000	26.000000	26.000000
75%	82409.750000	13206.750000	95.000000	83.500000	83.500000
max	83155.000000	31565.000000	181.000000	997.000000	997.000000

Figura 10- Estatística descritiva das variáveis utilizadas na análise de modelos.

RESULTADOS E DISCUSSÃO

MODELAGEM:

- **DISPERSÃO E HISTOGRAMAS**

Foram realizados gráficos de dispersão e histogramas, onde foi possível evidenciar que os dados não apresentam uma distribuição normal para o horímetro e idade. O tempo médio entre manutenções e o tempo médio para falha apresentam uma assimetria à esquerda. Pelo "Box-Plot", é possível visualizar como as variáveis estão distribuídas dentro dos quartis e suas medianas. Para o horímetro a mediana está abaixo das 5.000 horas, enquanto o valor máximo supera as 30.000 horas, corroborando com os resultados observados na estatística descritiva, o mesmo efeito é visível para as variáveis de tempo médio entre manutenções e tempo médio para falha, onde a mediana encontra-se abaixo de 20 horas e os valores máximos se aproximam das 100 horas.

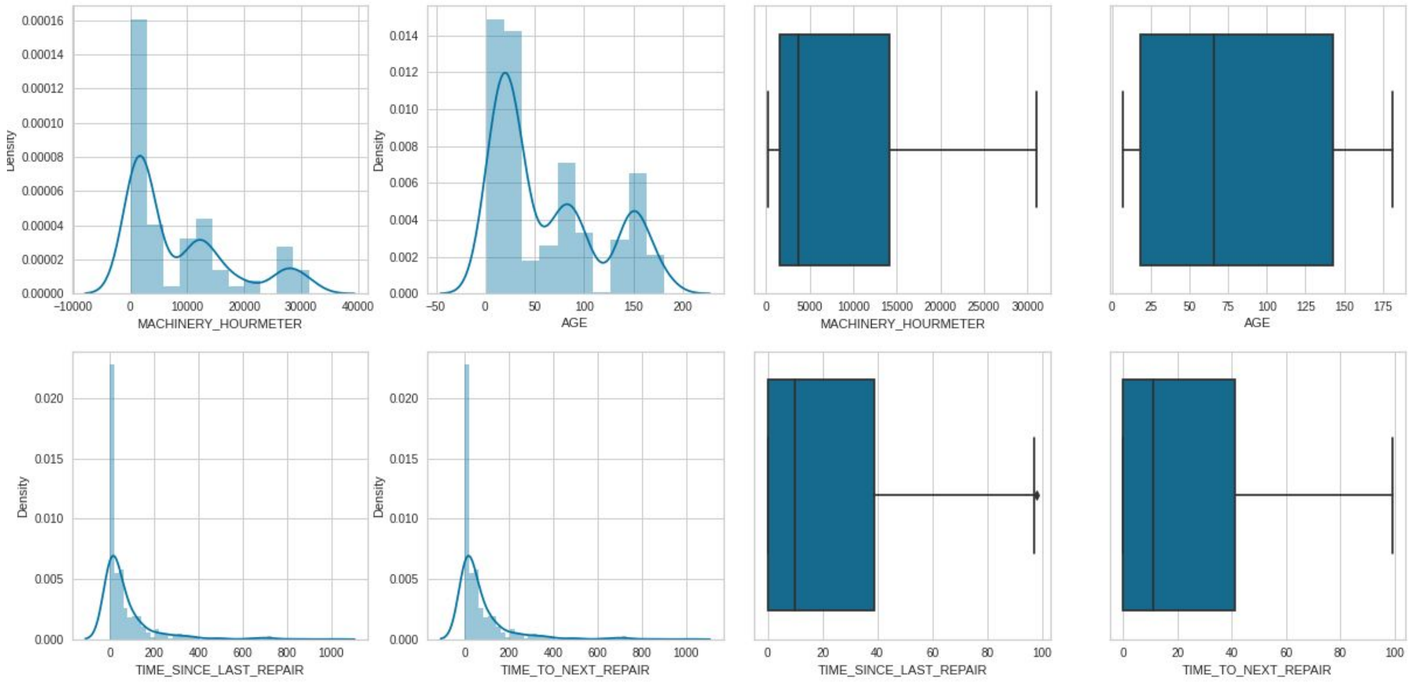


Figura 11- Histograma e "Box-Plot" das variáveis utilizadas na análise do modelo.

FATECSENAI
Faculdade de Tecnologia SENAI Mato Grosso

RESULTADOS E DISCUSSÃO

MODELAGEM:

- **CORRELAÇÃO**

Foi possível evidenciar uma forte correlação positiva entre as variáveis horímetro e idade dos equipamentos, resultado que apresenta-se bastante coerente, uma vez que, quanto mais velho o equipamento, provavelmente mais horas o mesmo trabalhou. As demais variáveis apresentam correlação negativa ou próximas a zero, resultados esses que podem indicar que a idade do equipamento de forma isolada não é um bom preditor para o tempo médio para a próxima falha. Possivelmente existem diversos outros fatores que não estão considerados no conjunto de dados disponível.

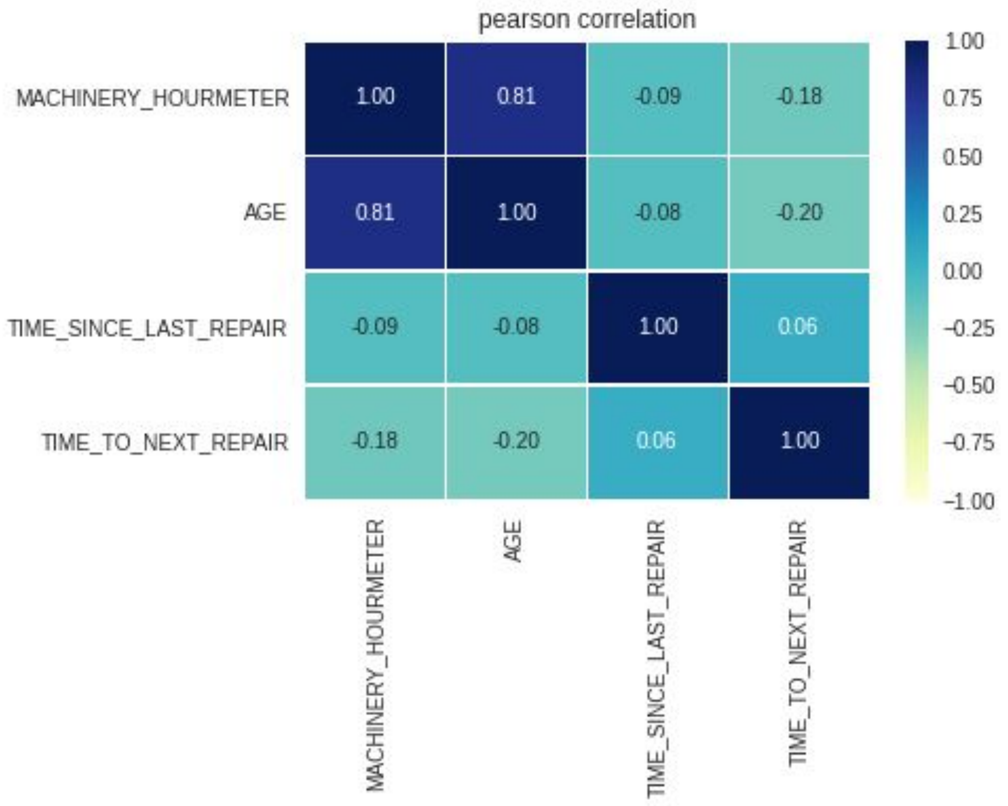


Figura 12 - Matriz de correlação entre as variáveis utilizadas na análise de modelos.

RESULTADOS E DISCUSSÃO

MODELAGEM:

- **CONFIGURAÇÕES DA ANÁLISE DE MODELOS**

Utilizando a biblioteca python PyCaret, definiu-se a variável tempo médio para falha como variável dependente, e as variáveis horímetro, idade e tempo médio entre manutenções como variáveis independentes. A biblioteca foi configurada para realizar a normalização dos dados, e com base nas configurações gerais desta biblioteca (onde é possível evidenciar que não existem dados faltantes, três variáveis numéricas, três variáveis categóricas), foi definida a proporção de 70%/30% entre conjunto de dados entre treinamento e testes. O número de interações padrão é de 10 e a normalização dos dados se dá por Z-Score.

	Description	Value		Description	Value		Description	Value		Description	Value
0	session_id	7412	16	Use GPU	False	32	PCA Method	None	48	Polynomial Threshold	None
1	Target	TIME_TO_NEXT_REPAIR	17	Log Experiment	False	33	PCA Components	None	49	Group Features	False
2	Original Data	(185, 8)	18	Experiment Name	reg-default-name	34	Ignore Low Variance	False	50	Feature Selection	False
3	Missing Values	False	19	USI	cea5	35	Combine Rare Levels	False	51	Feature Selection Method	classic
4	Numeric Features	3	20	Imputation Type	simple	36	Rare Level Threshold	None	52	Features Selection Threshold	None
5	Categorical Features	3	21	Iterative Imputation Iteration	None	37	Numeric Binning	False	53	Feature Interaction	False
6	Ordinal Features	False	22	Numeric Imputer	mean	38	Remove Outliers	False	54	Feature Ratio	False
7	High Cardinality Features	False	23	Iterative Imputation Numeric Model	None	39	Outliers Threshold	None	55	Interaction Threshold	None
8	High Cardinality Method	None	24	Categorical Imputer	constant	40	Remove Multicollinearity	False	56	Transform Target	False
9	Transformed Train Set	(129, 9)	25	Iterative Imputation Categorical Model	None	41	Multicollinearity Threshold	None	57	Transform Target Method	box-cox
10	Transformed Test Set	(56, 9)	26	Unknown Categoricals Handling	least_frequent	42	Remove Perfect Collinearity	True			
11	Shuffle Train-Test	True	27	Normalize	True	43	Clustering	False			
12	Stratify Train-Test	False	28	Normalize Method	zscore	44	Clustering Iteration	None			
13	Fold Generator	KFold	29	Transformation	False	45	Polynomial Features	False			
14	Fold Number	10	30	Transformation Method	None	46	Polynomial Degree	None			
15	CPU Jobs	-1	31	PCA	False	47	Trignometry Features	False			

Figura 13 - Configuração da análise de modelos utilizando PyCaret.

RESULTADOS E DISCUSSÃO

MODELAGEM:

- **RESULTADO DOS MODELOS ANALISADOS**

Entre os cinco modelos analisados, o modelo que apresentou o melhor resultado considerando as métricas de sucesso definidas anteriormente foi o modelo de regressão de floresta aleatória. O modelo de floresta aleatória obteve erro absoluto médio de 97,83 horas e o erro quadrático médio de 22723,76, evidenciando que o resultado é fraco para o objetivo de predizer o tempo médio para falha, uma vez que um erro médio de 97 horas entre estimado e real, não configura um erro aceitável para os padrões do negócio. O coeficiente de correlação neste trabalho (não definido como métrica de sucesso) mostra resultados negativos, indicando que os modelos não conseguem explicar de maneira satisfatória a variável dependente de acordo com as variáveis independentes disponíveis.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
knn	K Neighbors Regressor	101.1399	21846.2540	140.4576	-0.4541	1.5796	7.1654	0.080
lr	Linear Regression	98.9857	19155.2832	131.6107	-0.5063	1.6275	8.8708	0.025
rf	Random Forest Regressor	97.8355	22723.7639	144.5608	-0.8065	1.5811	9.8049	0.567
gbr	Gradient Boosting Regressor	101.1751	27753.8204	161.8613	-1.4531	1.5514	9.6592	0.068
dt	Decision Tree Regressor	128.0263	39296.8596	193.6178	-3.0360	1.9178	9.5279	0.040

Figura 14 - Resultado dos modelos analisados.

RESULTADOS E DISCUSSÃO

MODELAGEM:

- **OTIMIZAÇÃO DO MELHOR MODELO ANALISADO**

Com o objetivo de melhorar os resultados obtidos, realizou-se a separação do melhor modelo obtido, floresta aleatória, sendo otimizado os hyper parâmetros deste modelo, com o auxílio da biblioteca PyCaret a fim de buscar o melhor erro absoluto médio possível. A otimização melhorou o erro absoluto médio para 87,62 horas, uma diferença de aproximadamente 10 horas em comparação ao modelo não otimizado. Esse erro é inaceitável do ponto de vista do negócio, onde qualquer tomada de decisão baseada nos resultados desse modelo, poderá ocasionar em erros operacionais.

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	48.2692	3718.3654	60.9784	0.3863	0.9348	1.4283
1	76.7308	16414.0577	128.1174	0.5091	0.8071	0.9589
2	75.0000	10684.8462	103.3675	-1.0979	1.9772	13.9900
3	93.3846	35327.0769	187.9550	-0.1042	1.5197	3.7373
4	102.1923	30258.2885	173.9491	-0.0146	1.2294	1.7512
5	104.3077	25895.0769	160.9195	-0.5104	1.5384	4.1485
6	42.9231	3260.3077	57.0991	0.5896	1.2868	3.5919
7	70.6154	11624.6538	107.8177	-3.9355	1.8414	27.3811
8	120.3846	25433.7692	159.4797	-0.8188	1.7739	6.7616
9	142.4583	53227.8542	230.7116	-0.4766	1.6231	1.1963
Mean	87.6266	21584.4296	137.0395	-0.5473	1.4532	6.4945
SD	29.5265	14827.3467	52.9585	1.2499	0.3650	7.8798

Figura 15 - Resultado do modelo de floresta aleatória otimizado.

RESULTADOS E DISCUSSÃO

MODELAGEM:

- **NÍVEL DE IMPORTÂNCIA DAS VARIÁVEIS INDEPENDENTES**

Com base nos dados disponíveis e nos modelos analisados, o horímetro da máquina é a principal variável de importância na predição do tempo médio para falha, seguido pela idade do equipamento em meses e do tempo desde o último reparo, as demais variáveis não apresentam importância estatística. Esse resultado corrobora com os resultados obtidos na modelagem, que identificam que a utilização de modelos de aprendizado de máquina não se mostrou satisfatória para predizer o tempo médio para falha. Na etapa de análise exploratória dos dados, notou-se que existe diferença nos dados entre o tipo de operação e a estação operacional, entretanto, tal diferença não foi expressa no modelo de melhor resultado.

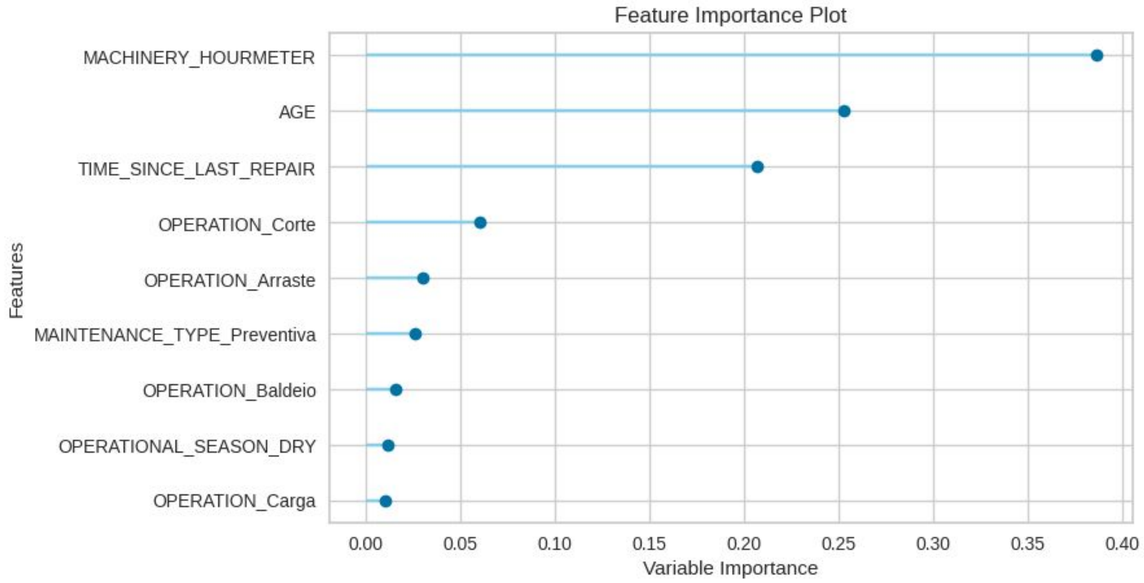


Figura 16 - Resultado de importância das variáveis utilizadas no modelo.

CONCLUSÕES

- A atividade de baldeio representa mais da metade do total de registros de manutenção, sendo maior que o dobro da segunda operação que é a de arraste;

- O tempo médio para reparo é de 43,96 horas, sendo um número considerado alto, pois algumas manutenções no conjunto de dados foram mais longas que a média (em alguns casos superior a 300 horas), devido a falhas complexas ou dificuldades de aquisição de peças de reposição;

- O conjunto de dados tem problemas em relação aos apontamentos de horímetro. Foram necessárias 262 interações de correção com auxílio do time de negócios da companhia para ajuste dos números;

- Após as correções foi possível evidenciar um tempo médio entre reparos de aproximadamente 19 horas, onde a atividade de corte apresenta o menor resultado (13 horas) indicando que essa operação apresenta mais necessidade de manutenção que as demais.

Fonte: Acervo Jean Cruz

CONCLUSÕES

- Evidenciou-se que 68,4% do total de manutenções foram corretivas e 29,6%, preventivas, indicando que na sua maioria as manutenções acontecem após a falha do equipamento, incorrendo em perdas de produção, por conta de equipamento parado de forma não programada;

- As peças terminal hidráulico e mangueira hidráulica são as peças mais frequentes nas manutenções, somando 40% do total quando combinadas;

- Mais de 65% das manutenções no conjunto de dados aconteceram na temporada chuvosa, entretanto, o resultado aqui obtido, pode conter algum viés em função da pandemia de COVID, onde algumas operações foram paralisadas temporariamente;

- A idade média dos equipamentos é de 65 meses, indicando que a maior parte dos equipamentos está no final do seu ciclo de vida (geralmente 60 meses). A atividade de arraste tem os equipamentos mais novos e a de carga os mais antigos.

Fonte: Acervo Jean Cruz

CONCLUSÕES

- O modelo com o melhor resultado foi o modelo de floresta aleatória com erro médio absoluto após otimização de 87,62 horas e erro quadrático de 21584,43, tais resultados mostram que a aplicação de modelos de aprendizado de máquina para predição do tempo médio para falha no conjunto de dados analisados **não têm resultados satisfatórios**, não podendo ser utilizado em escala de produção para auxílio na tomada de decisão da companhia;

- Sugere-se que o negócio invista em aumento da governança de dados, objetivando maior qualidade nos dados obtidos, uma vez que esses são a chave para o sucesso de qualquer análise exploratória ou modelagem preditiva;

- Recomenda-se também a continuidade do trabalho, através do incremento da engenharia de variáveis, com o objetivo de entender melhor a variabilidade dos dados. É aconselhável investigar a estratificação dos modelos e gerar resultados baseados em grupos de operação ou idade dos equipamentos, as quais podem apresentar erros menores que os aqui obtidos.

Fonte: Acervo Jean Cruz

REFERÊNCIAS

- ALI, M. PyCaret: An open source, low-code machine learning library in Python. **PyCartet version 1.0**, 2020. Disponível em: https://www.pycaret.org. Acesso em: 27 abr. 2022.
- ANZANELLO, M. J.; SILVA, P. R. S. da; RIBEIRO, J. L. D.; FOGLIATTO, F. S. Proposição de modelo de degradação para capacitores submetidos a ensaios acelerados. *In*: ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO, 23., 2003, Ouro Preto. **Anais** […]. Ouro Preto: ABEPRO, 2003.
- EZRA, O. Achieving Manufacturing Excellence with Predictive Maintenance and Machine Learning. **Industry 4.0 Insights**, [*s. l.*], 2018.
- FERNÁNDEZ-DELGADO, M. *et al*. Do we need hundreds of classifiers to solve real world classification problems. **Journal of Machine Learning Research**, [*s. l.*], v. 15, n. (1), p. 3133-3181, 2014.
- FOGLIATTO, F. S.; RIBEIRO, J. L. D. **Confiabilidade e manutenção industrial**. Rio de Janeiro: Elsevier, 2009. 265 p.
- KARDEC, A.; NASCIF, J. de A. **Manutenção**: Função estratégica. Rio de Janeiro: Qualitymark, 2009.
- KUHLMAN, D. Introductions Etc. *In*: KUHLMAN, D. **A Python Book**: Beginning Python, Advanced Python, and Python Exercises. [*S. l.*: *s. n.*], 2013.
- LAFRAIA, J. R. B. **Manual de confiabilidade, mantenabilidade e disponibilidade**. 2. ed. Rio de Janeiro: Qualitymark, 2001.
- LUCIAN, L. M. **Fundamentos de Aprendizagem de Máquina**. Porto Alegre: SAGAH, 2020.
- MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997.
- MONARD, M. C.; BARANAUSKAS, J. A. **Conceitos Sobre Aprendizado de Máquina. Sistemas Inteligentes Fundamentos e Aplicações**. Barueri, SP: Manole Ltda, 2003. p. 89-114.

REFERÊNCIAS

- NATEKIN, A.; KNOLL, A. **Gradient boosting machines, a tutorial**. Munich, Alemanha: Department of Informatics, Technical University Munich, Garching, 2013.
- PACCOLA, J. E. **Manutenção e Operação de Equipamentos Móveis**. São José dos Campos: JAC, 2017.
- PEDREGOSA, F. *et al*. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011.
- PERES, S. M. *et al*. Tutorial sobre fuzzy-c-means e fuzzy learning vector quantizations: Abordagens híbridas para tarefas de agrupamento e classificação. **Revista de Informática Teórica e Aplicada**, v. 19, n. 1, p. 120-163, 2012.
- PIOTROWSKI, P. Build a Rapid Web Development Environment for Python Server Pages and Oracle. **Oracle Technology Network**, 2006. Retirado do original em 2 abr. 2019. Acessado novamente em 12 mar. 2012.
- PYTHON SOFTWARE FOUNDATION. Is Python a good language for beginning programmers? **General Python FAQ**. Retirado do original em 24 out. 2012. Acessado novamente em jan. 2022.
- RICHTER, I. Achieving Zero Unplanned Downtime with Predictive Maintenance Analytics. **Industry 4.0 Insights**, [*s. l.*], 2019.
- RIQUETI, G. A.; RIBEIRO, C. E.; ZÁRATE, L. E. **Classificando perfis de longevidade de bases de dados longitudinais usando Floresta Aleatória**. [*S. l.*]: Symposium on Knowledge Discovery, Mining and Learning, KDMILE, 2018.
- ROSSUM, G. van. **The History of Python**: A Brief Timeline of Python. [*S. l.*: *s. n.*], 2009. Retirado do original em 2022.
- RODRIGUES, S. C. **Modelo de Regressão Linear e suas Aplicações**. Covilhã: [*s. n.*], 2012.
- WUTTKE, R. A.; SELLITO, M. A. Cálculo da disponibilidade e da posição na curva da banheira de uma válvula de processo petroquímico. **Revista Produção Online**, v. 8, n. 4, p. 1-23, 2008.

REFERÊNCIAS

OBRIGADO!

FATECSENAI
Faculdade de Tecnologia SENAI Mato Grosso

Apresentação de TCC
MBA Data Science 2019

ANÁLISE DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE FALHA MECÂNICA DE EQUIPAMENTOS FLORESTAIS

Aluno: **Jean Carlos da Cruz**
Orientador: Denise Prado Kronbauer
Faculdade de Tecnologia SENAI Mato Grosso – Cuiabá-MT

FATECSENAI
Faculdade de Tecnologia SENAI Mato Grosso