**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**Lúcio Flávio Barboza**

**UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DO *TML* (LIMITE DE UMIDADE TRANSPORTÁVEL) DE EMBARQUES DE MINÉRIO DE FERRO**

São Luís

2024

**Lúcio Flávio Barboza**

**UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DO *TML* (LIMITE DE UMIDADE TRANSPORTÁVEL) DE EMBARQUES DE MINÉRIO DE FERRO**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

São Luís

2024 **LISTA DE ILUSTRAÇÕES**

[Figura 1 – Navio Hui Long (2005) naufrágio por liquefação de carga, fluorita. 7](#_Toc126258644)

[Figura 2 – Utilização da função merge. 13](#_Toc126258645)

[Figura 3 – Aparência do dataset “df\_final”. 13](#_Toc126258646)

[Figura 4 – Informações do dataset “df\_final”. 14](#_Toc126258647)

[Figura 5 – Histograma da variável TML. 15](#_Toc126258648)

[Figura 6 – Gráfico Boxplot da variável TML. 16](#_Toc126258649)

[Figura 7 – Identificação e eliminação de outliers. 16](#_Toc126258650)

[Figura 8 – Matriz de correlação com a variável TML. 18](#_Toc126258651)

[Figura 9 – Utilização do LazyRegressor. 19](#_Toc126258652)

[Figura 10 – Exibição do dataset e dos modelos sugeridos em (a). 19](#_Toc126258653)

[Figura 11 – Exibição do dataset e dos modelos sugeridos em (b). 21](#_Toc126258654)

[Figura 12 – Exibição do dataset e dos modelos sugeridos em (c) 21](#_Toc126258655)

[Figura 13 – Matriz reduzida de correlação com a variável TML . 22](#_Toc126258656)

[Figura 14 – Correlação das variáveis com o TML . 23](#_Toc126258657)

[Figura 15 – Exibição do dataset final - variáveis de maior correlação 2](#_Toc126258658)3

[Figura 16 – Melhores modelos obtidos a partir da redução de dimensionalidade. 24](#_Toc126258659)

[Figura 17 – Resultado das Métricas utilizando SVR 2](#_Toc126258660)5

[Figura 18 – Gráfico de valores reais x valores preditos (SVR) 2](#_Toc126258661)6

[Figura 19 – Cross Validation (SVR) 2](#_Toc126258662)6

[Figura 20 – Ajuste de hiper parâmetros (SVR) 2](#_Toc126258663)7

[Figura 21 – Influência das variáveis pela função SHAP (SVR). 2](#_Toc126258664)8

[Figura 22 – Resultado das Métricas utilizando LassoCV.](#_Toc126258665) 29

[Figura 23 – Gráfico de valores reais x valores preditos LassoCV .](#_Toc126258666) 29

[Figura 24 – Cross Validation (LassoCV) 30](#_Toc126258667)

[Figura 25 – Ajuste de hiper parâmetros (LassoCV) 30](#_Toc126258668)

[Figura 26 – Influência das variáveis pela função SHAP (LassoCV) 31](#_Toc126258669)

[Figura 27 – Resultado das métricas utilizando LinearRegression. 3](#_Toc126258670)2

[Figura 28 – Gráfico de valores reais x valores preditos (LinearRegression)](#_Toc126258671) 32

[Figura 29 – Cross Validation (LinearRegression). 3](#_Toc126258672)3

[Figura 30 – Ajuste de hiper parâmetros (LinearRegression). 3](#_Toc126258672)3

[Figura 31 – Influência das variáveis pela função SHAP (LinearRegression). 3](#_Toc126258672)4

**SUMÁRIO**

[1. Introdução 5](#_Toc74563991)

[1.1. Contextualização 5](#_Toc74563992)

[1.2. O problema proposto 7](#_Toc74563993)

[1.3. Objetivos 9](#_Toc74563994)

2[. Coleta de Dados 10](#_Toc74563994)

[3. Processamento/Tratamento de Dados 12](#_Toc74563995)

[3.1 Junção das bases de dados 12](#_Toc74563995)

[4. Análise e Exploração dos Dados 15](#_Toc74563996)

[5. Criação de Modelos de Machine Learning 19](#_Toc74563997)

[6. Apresentação dos Resultados 35](#_Toc74563999)

[7. Links 38](#_Toc74564000)

[REFERÊNCIAS 39](#_Toc74564001)

# 1. Introdução

## 1.1. Contextualização

No mercado transoceânico e na cabotagem, o minério de ferro é transportado em porões de navios destinados ao transporte de cargas sólidas a granel. Nos porões, durante a viagem marítima, esses minérios são submetidos a uma condição singular, chamada tecnicamente de carregamento cíclico. E é nesta condição que surge um risco à navegação que precisa ser gerenciado.

A atuação no setor marítimo conta com organizações que visam a orientar, regulamentar e proteger todas as partes envolvidas no transporte de carga. A principal é a IMO – Organização Marítima Internacional, que é a agência especializada das Nações Unidas responsável pela segurança do transporte marítimo e pela prevenção da poluição marinha e atmosférica dos próprios navios. O foco da IMO é criar uma estrutura regulamentar do transporte marítimo internacional para garantir que esse setor vital permaneça seguro, eficiente e ambientalmente correto.

Um importante instrumento da IMO é o IMSBC Code – International Maritime Solid Bulk Cargoes Code. Em português: Código Marítimo Internacional para Cargas Sólidas a Granel. O IMSBC Code é a ferramenta normativa da IMO, que determina as práticas obrigatórias para o transporte marítimo seguro dessas cargas. O código fornece informações sobre os riscos associados ao transporte de granéis sólidos e os procedimentos que devem ser adotados em cada caso para garantia da segurança.

Conforme o IMSBC Code, as cargas sólidas a granel são classificadas em três grandes grupos: Grupo A, Grupo B e Grupo C. O Grupo A é composto por cargas que podem se liquefazer se embarcadas com um conteúdo de umidade acima do seu limite de umidade transportável (conceito que será melhor explorado adiante). O Grupo B contém cargas que possuem um risco químico que pode levar o navio a uma situação de perigo. Por fim, o Grupo C é formado por cargas que não são sujeitas a liquefação e nem possuem riscos químicos. Vale ainda destacar que existem cargas que são classificadas simultaneamente nos Grupos A e B.

Uma vez conhecidos os três grupos em que estão distribuídas as cargas sólidas a granel no transporte marítimo, doravante nesse trabalho, trataremos somente das cargas do grupo A, de modo particular, os minérios de ferro fino sujeitos ao risco de liquefação.

Para facilitar o entendimento dos Aspectos Regulatórios no Transporte Marítimo é importante conhecer algumas definições:

**Carga sólida a granel**: qualquer carga que consista numa combinação de partículas, grânulos ou quaisquer fragmentos maiores de um material, geralmente de composição uniforme, que seja carregada diretamente nos porões de um navio, sem qualquer forma intermediária de acondicionamento.

**Conteúdo de umidade (MC – Moisture Content)**: quantidade de água presente no minério, expressa como percentagem da massa úmida total.

**Limite de Umidade Transportável (TML – Transportable Moisture Limit)**: o conteúdo de umidade máximo de uma carga que é considerado seguro para o seu transporte em navios.

**Carga que pode se liquefazer**: carga pertencente ao Grupo A que contenha uma certa proporção de partículas finas e uma certa quantidade de umidade. Ela pode se liquefazer caso seja embarcada além do Limite de Umidade Transportável (TML).

**Liquefação**: é um fenômeno perigoso porque o estado fluido viscoso da carga pode fazer com que ela se mova para um dos bordos do navio com um movimento de balanço, mas ela pode não retornar à posição inicial quando o navio se inclinar para o outro bordo. Consequentemente, o navio pode adernar progressivamente e chegar ao ponto de emborcar bruscamente.

**Figura 1 – Navio Hui Long (2005) naufrágio por liquefação de carga, fluorita**



Fonte: *Maritime Mutual Risk Bulletin No. 12*

Uma vez conhecidas as definições acima, é de fundamental importância compreender que, o requisito básico para carregamento de uma carga do Grupo A em navio é: **MC < TML**. Dito isso, cargas que podem liquefazer somente serão aceitas para carregamento quando o conteúdo de umidade da carga for menor que seu TML. Nota: há exceção a essa regra, o transporte em navios especialmente construídos, mas que não é objeto desse trabalho.

## 1.2. O problema proposto

Conforme descrito na sessão anterior, embarques de minério de ferro fino só podem ser transportados respeitando uma condição básica e fundamental: umidade de embarque < umidade limite transportável (MC < TML). Portanto, durante o processo de formação de carga para embarque, torna-se imperativo garantir que o TML não sofra variações significativas para baixo vindo a ameaçar a segurança da navegação em função da redução da margem entre a umidade da carga e seu TML. Essa condição se torna especialmente crítica na estação chuvosa por conta do aumento natural da umidade do minério provocado pelas chuvas.

Assim, o problema proposto consiste na abordagem de modelos de Machine Learning que possam apoiar o processo de gestão de risco das cargas através da predição desta importante variável que é o TML.

Visando proporcionar uma melhor visão do problema e da solução proposta, utilizou-se a técnica dos 5W (Why, Who, When, Where e Why):

Por quê? Uma queda brusca e não prevista no valor do TML pode determinar a redução da cadência dos embarques e, no pior cenário, ameaçar a segurança do transporte marítimo.

Quem? Dados de qualidade química e física de embarques de minério de ferro fino embarcados pelo Terminal Marítimo de Ponta da Madeira, incluindo seu TML e a composição percentual por mina de origem.

O que? Identificar algoritmos de Machine Learning que sejam capazes de prever, com boa aderência, o TML dos embarques de minério de ferro. A performance esperada para o modelo é conseguir um MAE (Erro Médio Absoluto) < 0,15. Não será aprofundada aqui a razão pela escolha desse valor, mas ela é baseada na margem de segurança existente entre a umidade real e o TML. Ou seja, um erro de até +0,15p.p. na predição do TML (valor predito 0,15p.p. > valor real) pode ser absorvido pelas camadas de gestão de segurança que acompanham a realização do embarque para garantir MC < TML (umidade < umidade limite).

Onde? Terminal Marítimo de Ponta da Madeira, São Luís-MA.

Quando? Período de 02/01/2021 a 01/08/2024

# **1.3. Objetivos**

Sem a pretensão de trazer uma solução inédita para a indústria, em cima de um tema de inegável relevância, nem de substituir ou confrontar soluções existentes, esse trabalho acadêmico objetiva a aplicação dos conhecimentos adquiridos no Curso de Ciência de Dados e Big Data para construção de modelos empíricos de predição do TML de embarques marítimos de minério de ferro fino a partir da sua qualidade química e física.

2. Coleta de Dados

Os dados utilizados nesse trabalho foram extraídos a partir de consultas a um sistema de informação interno utilizado no Terminal Marítimo de Ponta da Madeira. e exportados para o formato Excel. Foram utilizados três dataset.

O primeiro dataset, denominado “Base\_Embarques”, contendo informações gerais de 2.375 registros de embarques de finos de minérios de ferro, comercialmente chamado de sínter feed, exportados pelo porto de Ponta da Madeira entre os anos de 2021 e 2024.

|  |  |  |
| --- | --- | --- |
| Nome da coluna/campo | Descrição | Tipo |
| ID Embarque | Identificação do embarque | String |
| Data DTR | Data de desatracação | String |
| Produto | Produto embarcado | String |
| Ton. Emb. | Tonelagem embarcada | Inteiro |

Um segundo dataset, denominado “Base\_Qualidade”, contendo dados de qualidade química e física dos embarques, além da sua composição percentual por mina de origem.

|  |  |  |
| --- | --- | --- |
| Nome da coluna /campo | Descrição | Tipo |
| ID Embarque | Identificação do embarque | String |
| FE | Teor de ferro | Float |
| SIO2 | Teor de Sílica | Float |
| AL2O3 | Teor de alumina | Float |
| P | Teor de fósforo | Float |
| MN | Teor de manganês | Float |
| TIO2 | Teor de titânio | Float |
| PPC | Perda por calcinação | Float |
| H2O | Conteúdo de umidade | Float |
| +6,3 | Granulometria acumulada em 6,3mm | Float |
| +1 | Granulometria acumulada em 1mm | Float |
| -0,15 | Granulometria passante em -0,15mm | Float |
| 25 | Granulometria retida em 25mm | Float |
| 19 | Granulometria retida em 19mm | Float |
| 16 | Granulometria retida em 16mm | Float |
| 12,5 | Granulometria retida em 12,5mm | Float |
| 10 | Granulometria retida em 10mm | Float |
| 8 | Granulometria retida em 8mm | Float |
| 6,3 | Granulometria retida em 6,3mm | Float |
| 4 | Granulometria retida em 4mm | Float |
| 2 | Granulometria retida em 2mm | Float |
| 1 | Granulometria retida em 1mm | Float |
| 0,5 | Granulometria retida em 0,5mm | Float |
| 0,25 | Granulometria retida em 0,25mm | Float |
| 0,15 | Granulometria retida em 0,15mm | Float |
| %\_SFCK | Participação em % do protoproduto SFCK | Float |
| %\_SFLS | Participação em % do protoproduto SFLS | Float |
| %\_FCSS | Participação em % do protoproduto FCSS | Float |

E um terceiro dataset, denominado “Base\_TML”, contendo a identificação do embarque com seu respectivo TML analisado.

|  |  |  |
| --- | --- | --- |
| Nome da coluna/campo | Descrição | Tipo |
| ID Embarque | Identificação do embarque | String |
| TML LAB | Valor do TML analisado | Float |

Importante destacar que os três dataset utilizados se relacionam pela chave ID Embarque.

# 3. Processamento/Tratamento de Dados

O tratamento de dados em machine learning é uma etapa fundamental que pode determinar o sucesso ou o fracasso de um modelo. A qualidade dos dados usados para treinar um algoritmo influencia diretamente sua capacidade de fazer previsões confiáveis.

Dados brutos frequentemente contêm erros, valores ausentes e ruídos que podem distorcer os resultados. O tratamento de dados envolve limpar, transformar e organizar essas informações, garantindo que elas estejam em um formato adequado para análise. Isso inclui a remoção de duplicatas, o preenchimento ou exclusão de valores ausentes e a normalização de dados, que ajuda a manter a consistência.

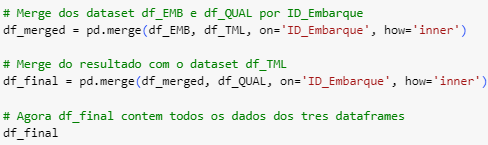
Antes de entrarmos na etapa de processamento de dados propriamente dita, vale destacar que todo o desenvolvimento desse trabalho foi realizado no ambiente Google Colab, o qual trouxe diferentes facilidades, tais como: execução diretamente no navegador (*browser*), acesso gratuito a GPUs e TPUs, Integração com Google Drive, e utilização de bibliotecas populares como Python, NumPy, Pandas, Scikit-learn, Seaborn, Lazypredict entre outras.

Conforme mencionado no item anterior, os três dataset utilizados nesse trabalho foram extraídos através de consulta a bancos de dados relacionais de produção garantindo a qualidade dos dados. Não obstante, como forma de confirmar a consistência e integridade dos dados durante a sua manipulação em planilhas, o processo de merge listado abaixo se encarregou de eliminar a eventual existência de registros duplicados. A normalização dos dados não foi necessária uma vez que todas as variáveis utilizadas se encontram na mesma escala, percentual (%).

**3.1 Junção das bases de dados**

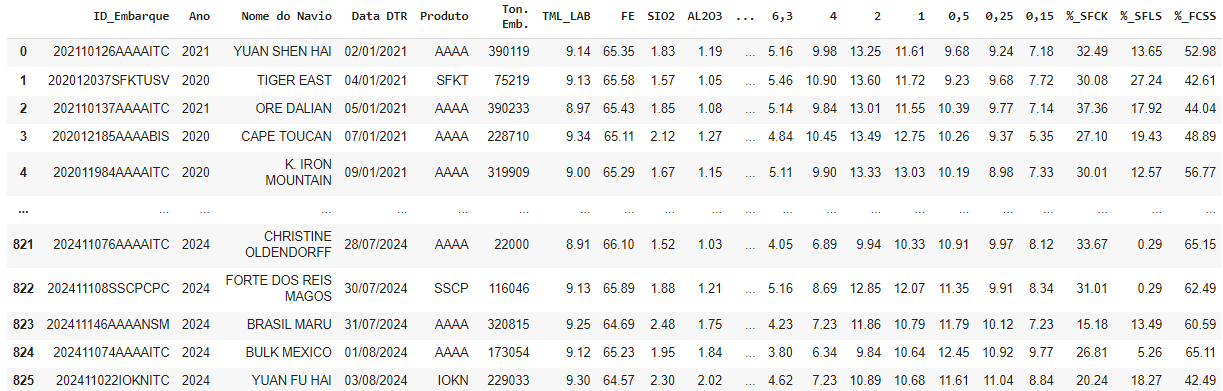
Uma etapa fundamental na construção de um modelo de machine learning é a junção de diversas bases em um único dataset. A figura 2 exibe o código em Python utilizando o método merge da biblioteca Pandas para consolidar os dataset de embarque (df\_EMB), TML (df\_TML) e qualidade (df\_QUAL). Como benefício adicional, o merge também se encarregou de eliminar 1.549 registros com valor de TML nulo.

**Figura 2: Utilização da função Merge**



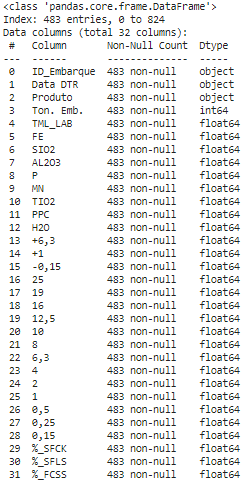
A figura 3, resultado do merge dos 3 dataset iniciais mostra como ficou o dataset final.

**Figura 3: Aparência do dataset “df\_final”**



Após a junção dos dados é importante checar se existe algum dado nulo, ocasião em que algum tratamento pode ser necessário. A seguir, com base na figura 4, verificamos que não existem dados nulos no novo dataset.

**Figura 4: Informações do dataset “df\_final”**



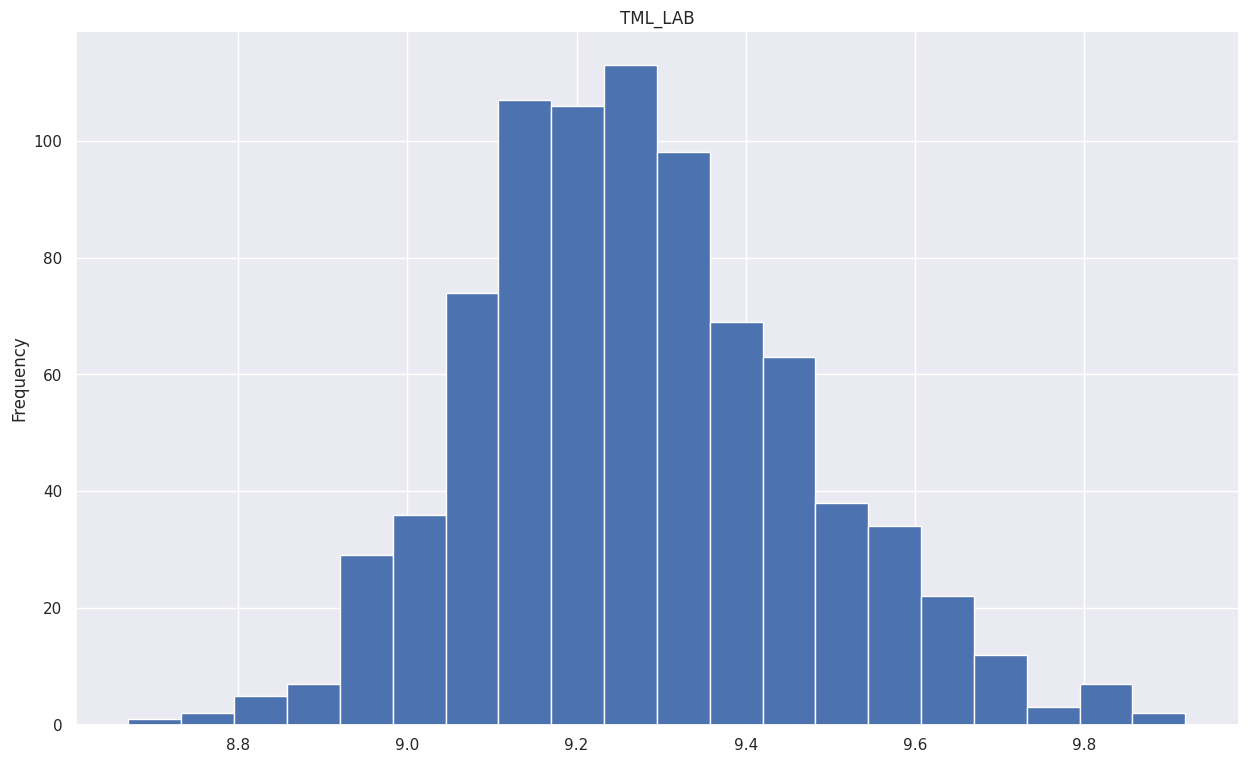
# 4. Análise e Exploração dos Dados

A análise e exploração de dados, frequentemente referida como EDA (*Exploratory Data Analysis*), desempenha um papel chave no desenvolvimento de projetos de machine learning. Este processo permite que os analistas e cientistas de dados compreendam melhor o conjunto de dados com o qual estão trabalhando.

A EDA oferece uma visão abrangente das características dos dados, incluindo suas distribuições, tipos e padrões subjacentes. Esse entendimento é vital para identificar as variáveis que influenciam o resultado e para formular estratégias apropriadas de modelagem.

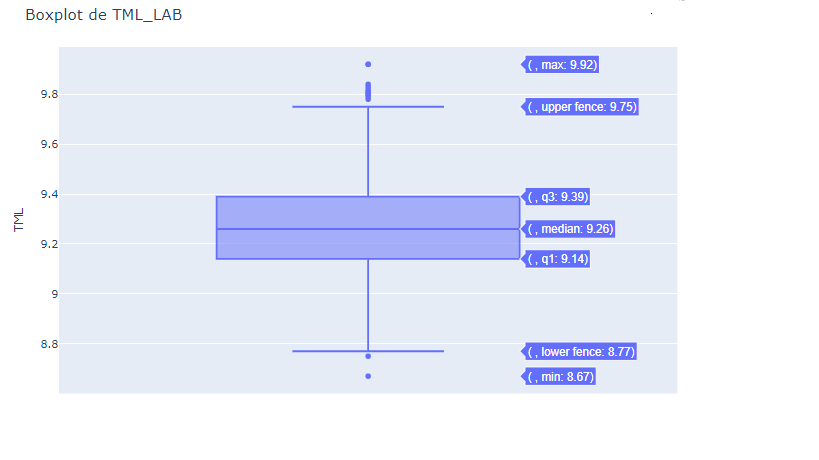
O histograma exibido na figura 5 mostra a distribuição de frequência da variável TML.

**Figura 5: Histograma da variável TML**



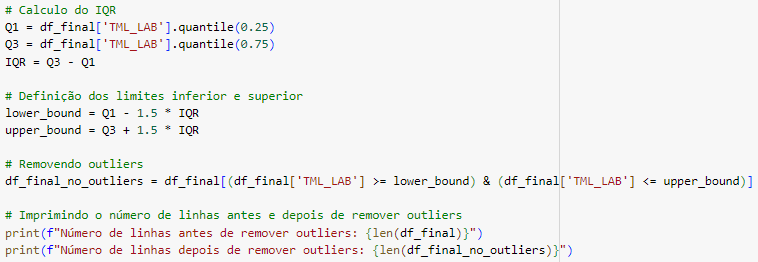
Problemas como outliers podem comprometer o desempenho do modelo. Ao realizar uma análise exploratória, o cientista de dados pode tomar medidas corretivas, como a remoção de outliers, antes de prosseguir para as etapas de modelagem. A figura 6 utiliza o gráfico Boxplot com a distribuição dos dados em quartis.

**Figura 6: Gráfico Boxplot da variável TML**



Não havendo possibilidade de proceder a uma análise de processo e investigar a origem dos outliers, optou-se pela sua exclusão:

**Figura 7: Identificação e eliminação de outliers**



Nesse ponto cabe destacar que, como forma de eliminar variações no modelo atribuídas pela variabilidade do TML em função do produto, optou-se por restringir a base de dados a um único produto comercial que teve sua identidade original alterada, passando a ser denominado “AAAA”.

Durante a exploração dos dados, é possível avaliar a importância de cada variável em relação ao problema em questão. A análise de correlações e a visualização das interações entre diferentes atributos permitem identificar quais recursos são mais relevantes para a predição. Essa etapa é crucial para reduzir a dimensionalidade do conjunto de dados, facilitando a modelagem e melhorando a performance do modelo.

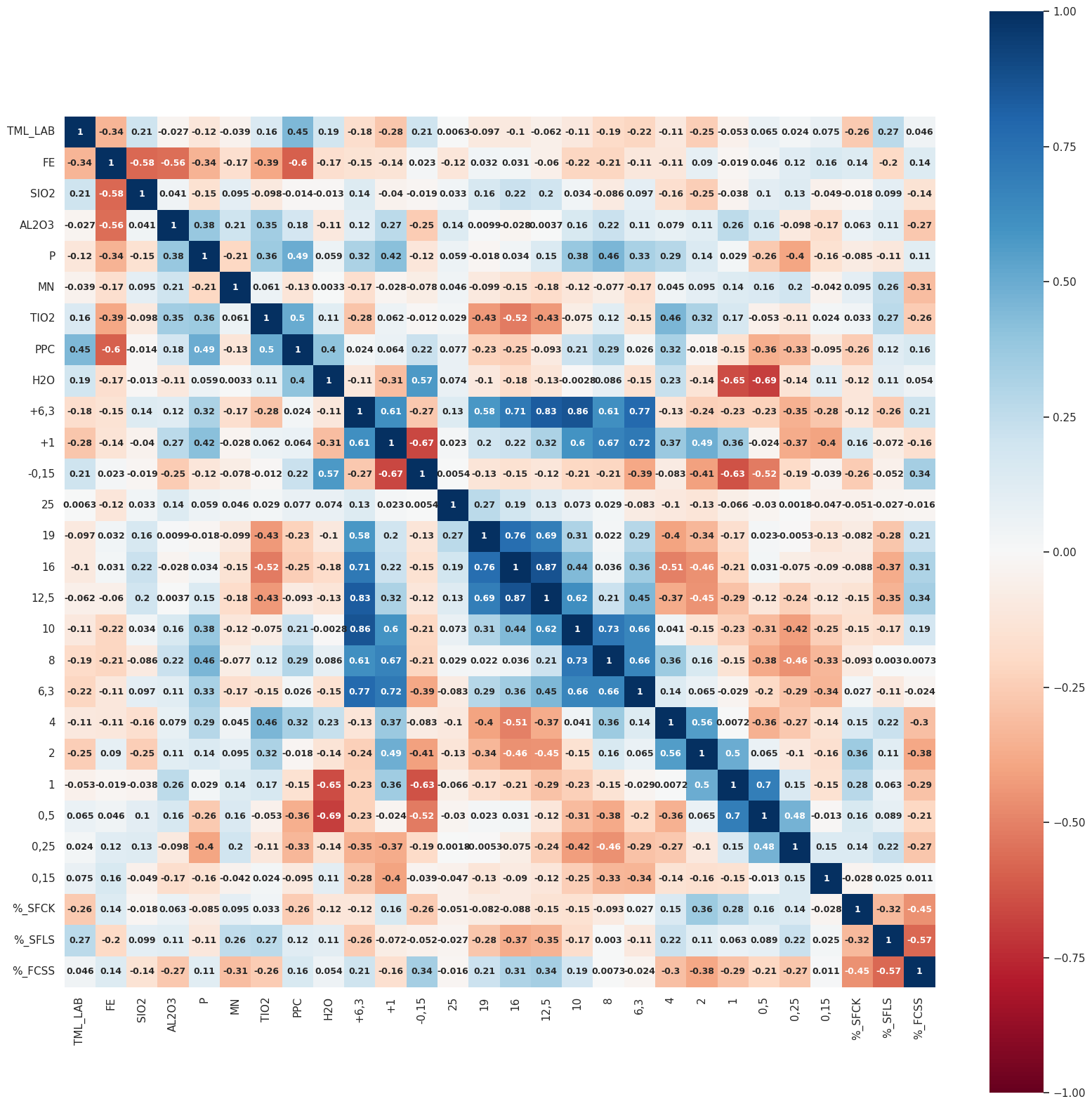
No trabalho de identificação das variáveis mais relevantes foi utilizada a função sns.heatmap da biblioteca Seaborn. A partir da matriz de correlação gerada (figura 8), é possível ver com clareza quais variáveis do modelo são mais relevantes e se estão positivamente ou negativamente correlacionadas com a variável de interesse, no caso, a variável TML\_LAB.

Interpretação da matriz de correlação:

* Correlação Positiva: Valores próximos de 1 indicam uma forte correlação positiva (quando uma variável aumenta, a outra também tende a aumentar).
* Correlação Negativa: Valores próximos de -1 indicam uma forte correlação negativa (quando uma variável aumenta, a outra tende a diminuir).
* Sem Correlação: Valores próximos de 0 indicam pouca ou nenhuma correlação
* Cores quentes (vermelho): Indicam correlações positivas fortes.
* Cores frias (azul): Indicam correlações negativas fortes.

Nessa primeira avaliação, apesar da baixa correlação entre as variáveis disponíveis no modelo, destaca-se o PPC como variável mais influente para o TML, correlação positiva.

**Figura 8: Matriz de correlação com a variável TML**



# 5. Criação de Modelos de Machine Learning

O avanço das técnicas de machine learning (ML) e a crescente disponibilidade de dados têm impulsionado a adoção dessas tecnologias em diversas áreas, como saúde, finanças, marketing e a indústria. No entanto, a seleção do modelo mais adequado para resolver um problema específico é um desafio significativo para profissionais e pesquisadores. O Lazy Predict é uma biblioteca desenvolvida em Python que busca simplificar essa tarefa, proporcionando uma maneira rápida e eficiente de avaliar diferentes algoritmos de ML.

O Lazy Predict automatiza a avaliação inicial de modelos de machine learning, permitindo a comparação rápida de diversos algoritmos sem a necessidade de ajustes complexos. O principal objetivo da ferramenta é facilitar o processo de experimentação, proporcionando resultados de forma intuitiva e visual.

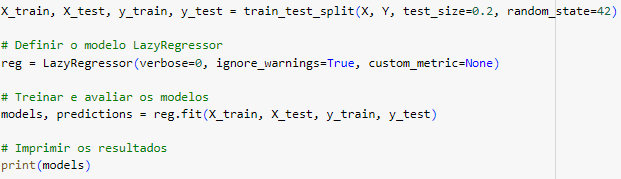
O conjunto de dados utilizado foi o dataset unificado que contém a qualidade ponderada das pilhas de minério de ferro juntamente com sua composição que, quando combinadas no plano de carregamento do navio, fornecerão a prévia do embarque.

A implementação do Lazy Predict foi realizada em etapas:

1. Importação das Bibliotecas: Importadas as bibliotecas necessárias para manipulação de dados, visualização e aplicação do Lazy Predict.
2. Carregamento do Conjunto de Dados: separação das bases de treino e teste.
3. Aplicação do Lazy Predict: utilizada a função LazyRegressor para avaliar os modelos disponíveis na biblioteca, que foram aplicados ao conjunto de dados.
4. Análise dos Resultados: Os resultados foram coletados e analisados, considerando métricas como R2, RMSE, MAE, tempo de execução e visualização dos dados.

A figura 9 traz o trecho de código utilizado para avaliação inicial dos modelos através da ferramenta LazyRegressor da biblioteca Lazy Predict.

**Figura 9: Utilização do LazyRegressor**



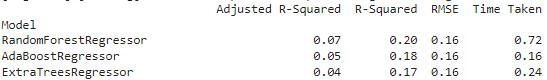
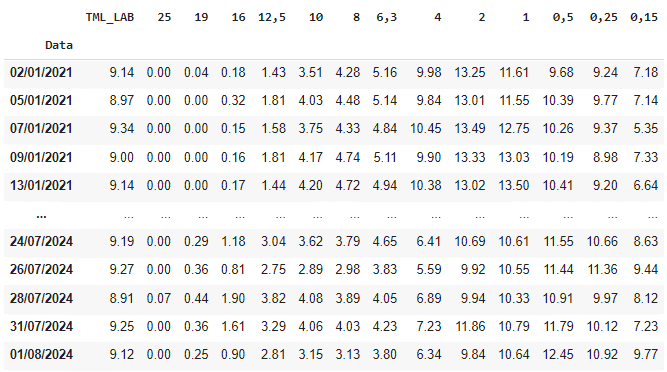
Com o objetivo de reduzir a dimensionalidade do modelo, antes de fazer uso da matriz de correlação apresentada no capítulo anterior, foram feitas algumas análises exploratórias para verificação dos resultados encontrados. Três abordagens iniciais foram elaboradas:

1. utilização apenas das variáveis físico granulométricas uma vez que o teste de TML também é um ensaio físico realizado a partir da compactação de uma amostra de minério, ou seja, à primeira vista, existiria uma lógica e uma expectativa positiva com essa abordagem;
2. acréscimo das variáveis químicas e;
3. acréscimo da composição do embarque por mina de origem

1. Somente variáveis granulométricas:

Avaliação dos resultados do Lazy Predict considerando apenas as variáves físico-granulométricas.

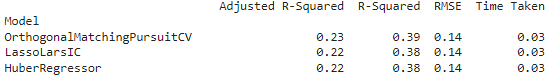
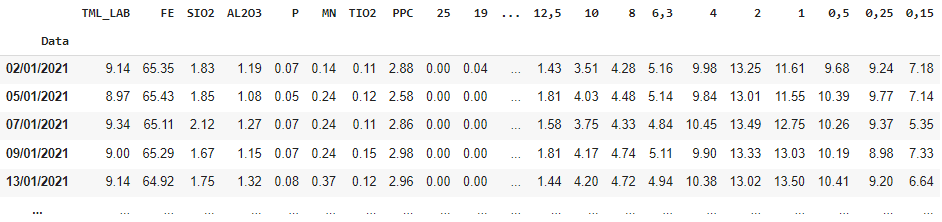
**Figura 10: Exibição do dataset e dos modelos sugeridos em (a)**



1. Variáveis granulométricas e químicas:

Avaliação dos resultados do Lazy Predict considerando as variáves físico-granulométricas juntamente com as variáveis químicas.

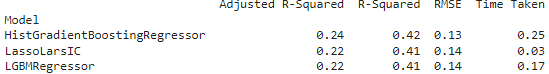
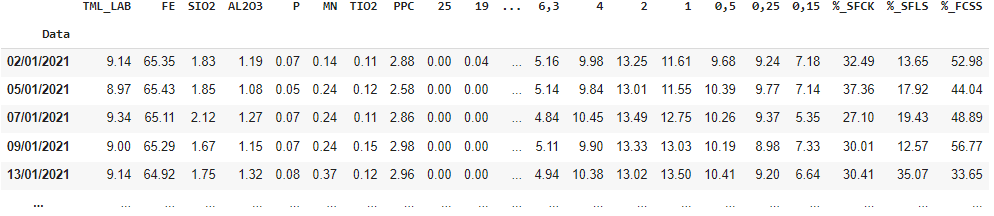
**Figura 11: Exibição do dataset e dos modelos sugeridos em (b)**



1. Variáveis granulométricas, química e composição por mina de origem:

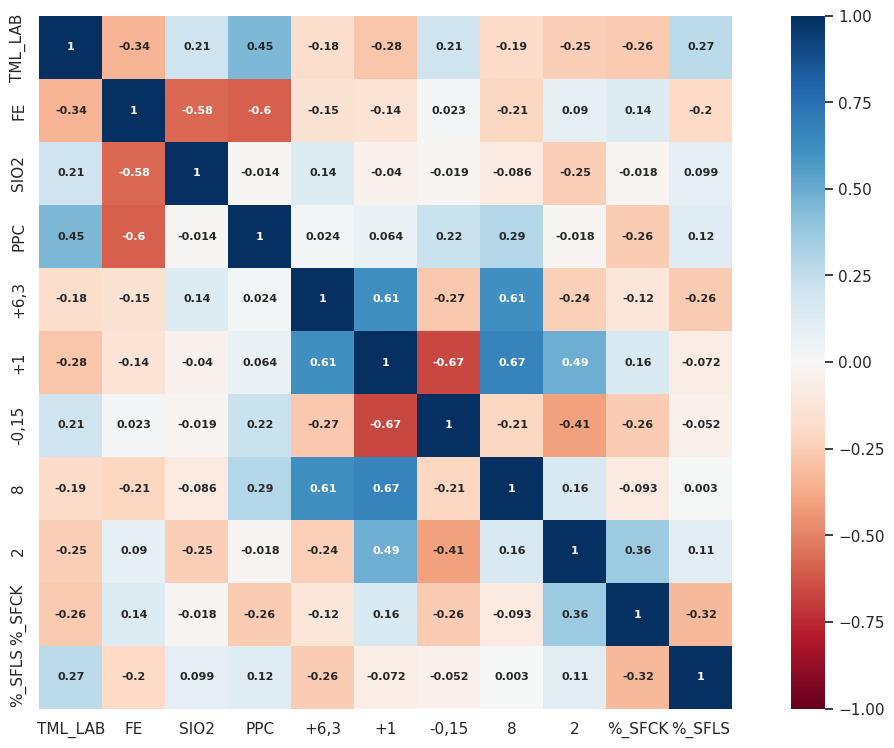
Avaliação dos resultados do Lazy Predict considerando as variáves físico-granulométricas juntamente com as variáveis químicas.

**Figura 12: Exibição do dataset e dos modelos sugeridos em (c)**



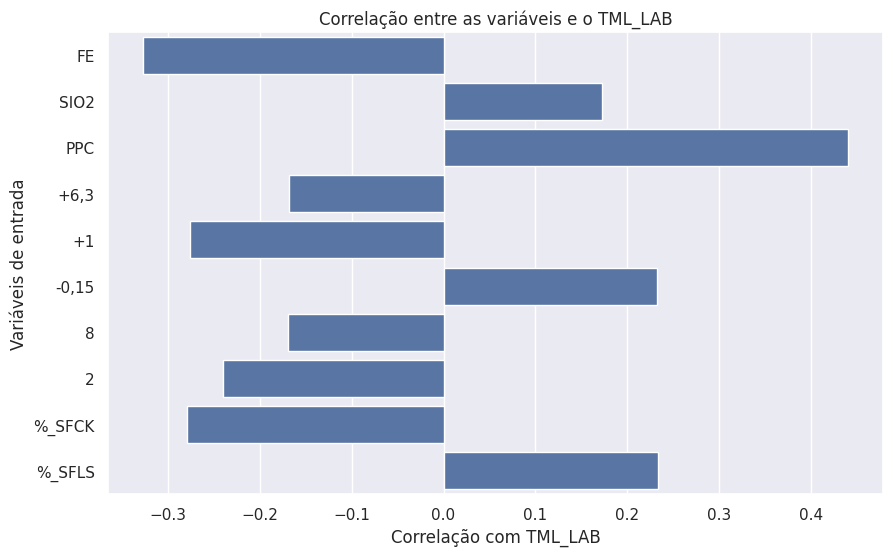
Concluída a análise exploratória, a qual ajudou a enfraquecer a hipótese formulada em (a), mostrando evolução nas métricas de avaliação à medida que novas variáveis foram acrescentadas aos modelos, retornou-se à matriz de correlação. Foi aplicado um corte eliminando todas as variáveis com correlação menor que 0.18, incluindo a variável H2O que, apesar de ter apresentado uma correlação de 0,19, sua utilização no modelo foi eliminada visto que é uma característica de qualidade analisada pelo laboratório, mas que não tem efeito no resultado do TML (a amostra passa por secagem antes do teste de compactação). A matriz de correlação na figura 13 mostra as variáveis selecionadas para alimentar os modelos.

**Figura 13: Matriz reduzida de correlação com a variável TML**



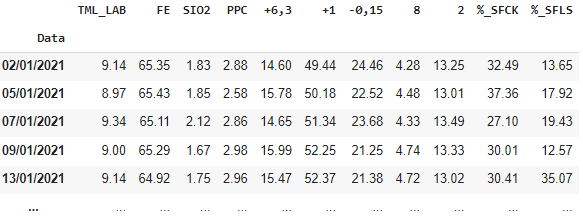
Uma outra forma interessante de visualização da correlação entre as variáveis de entrada e a variável de saída (TML), é mostrada no gráfico abaixo:

**Figura 14: Correlação das variáveis com o TML**



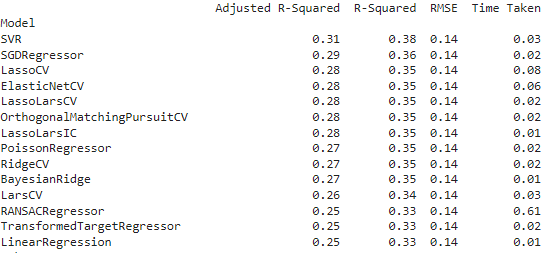
Concluída a etapa de análise exploratória das três abordagens formuladas em (a), (b) e (c), com base nos resultados obtidos, optou-se pela redução de dimensionalidade do modelo a partir da escolha das variáveis de maior correlação.

**Figura 15: Exibição do dataset final - variáveis de maior correlação**



A aplicação do Lazy Predict ao dataset mostrado na figura 14, produziu uma lista com 30 diferentes modelos testados. A figura 16 resume as métricas de desempenho dos quatorze primeiros modelos.

**Figura 16: Melhores modelos obtidos a partir da redução de dimensionalidade**



Os resultados demonstraram que todos os modelos apresentados na tabela acima alcançaram uma performance muito semelhante para as métricas R2 ajustado, R2 e RMSE. Vale ressaltar que essas métricas não foram escolhidas, mas foram aquelas disponibilizadas pelo Lazy Predict. De modo geral, o mesmo comportamento foi verificado para o tempo de execução, em que apenas o modelo RANSACRegressor consumiu quase oito vezes mais tempo que o penúltimo colocado, o modelo Lasso CV.

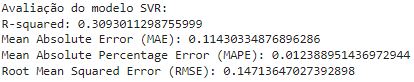
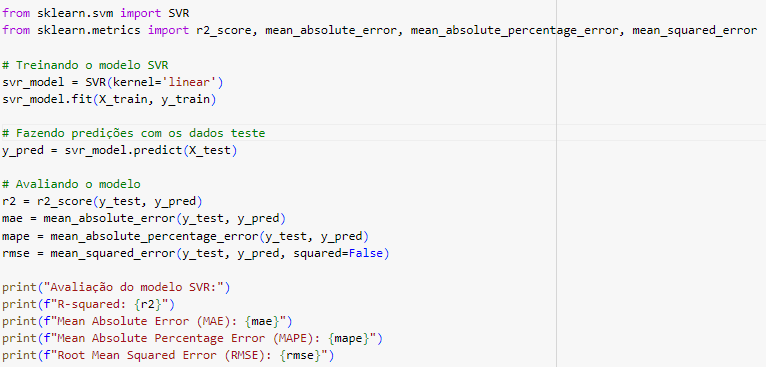
A partir do primeiro “*insight*” produzido com o uso do Lazy Predict, escolhemos três modelos de regressão com diferentes características para avaliar o seu desempenho e explorar seus hiper parâmetros: i) SVR (Regressão de Suporte Vetorial) ii) Regressão Lasso e iii) Regressão Linear.

**Modelo SVR**

A Support Vector Regression (SVR), ou Regressão de Vetores de Suporte, é uma técnica de aprendizado de máquina baseada nos princípios das Máquinas de Vetores de Suporte (SVM). Enquanto as SVMs são amplamente utilizadas para problemas de classificação, a SVR é projetada para prever valores contínuos.

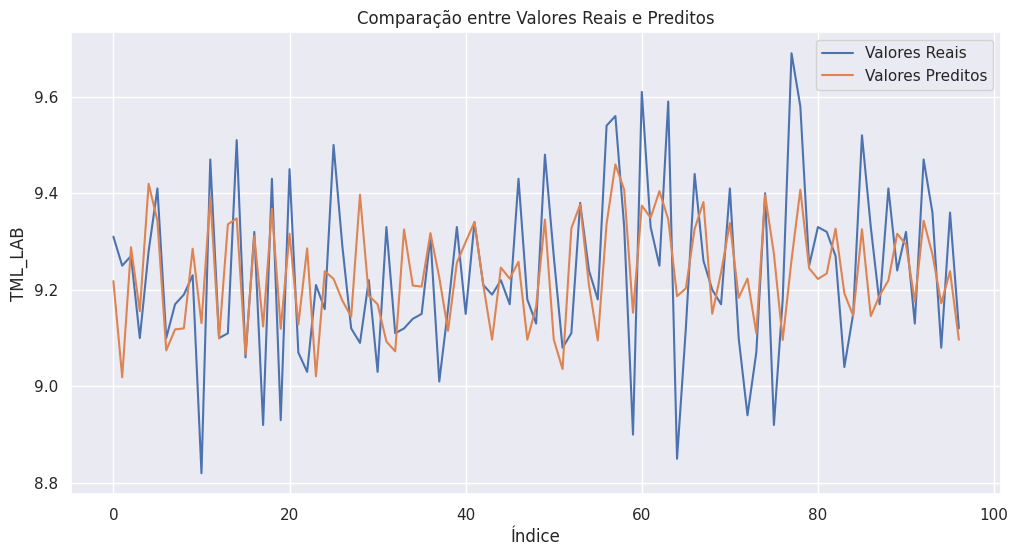
Abaixo é exibido o código utilizado com o resultado das métricas obtidas. Conforme já definido no item 1.2, é importante ressaltar que a métrica escolhida nesse trabalho para avaliar a performance do modelo será o MAE.

**Figura 17: Resultado das Métricas utilizando SVR**



O gráfico de linhas abaixo ajuda a ver o resultado obtido entre os valores reais e os valores preditos pelo modelo.

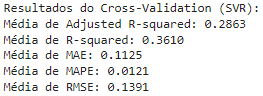
**Figura 18: Gráfico de valores reais x valores preditos (SVR)**



A validação cruzada (*cross validation*) é uma técnica estatística usada para avaliar a capacidade de generalização de um modelo preditivo. Ela é amplamente utilizada em Machine Learning para garantir que o modelo não esteja super ajustado (*overfitting*) aos dados de treinamento e possa generalizar bem com dados novos.

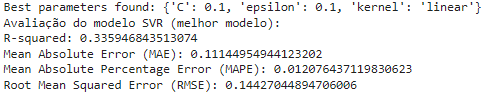
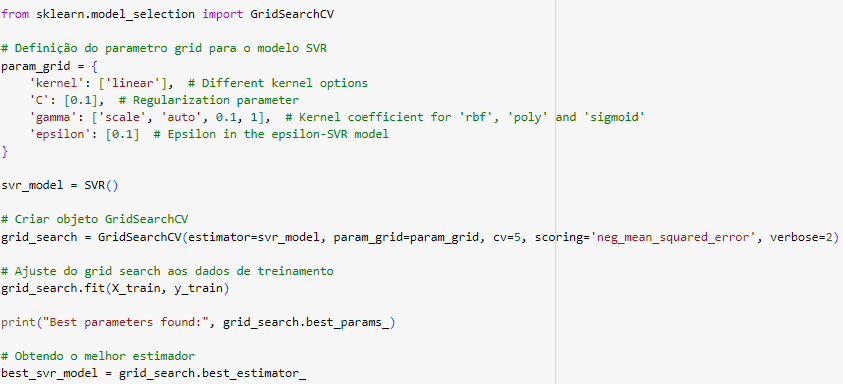
Os resultados obtidos na etapa de cross validation, figura 19, utilizando 5 folds e kernel linear evidenciaram a boa capacidade de generalização do modelo gerando um MAE médio = 0,11. Considerando o arredondamento em duas casas decimais, é o mesmo valor obtido anteriormente para todo o conjunto de dados.

**Figura 19: Cross Validation (SVR)**



A seguir, partiu-se para o ajuste de hiper parâmetros, processo crucial em Machine Learning que envolve a seleção dos melhores valores para os hiper parâmetros de um modelo, a fim de otimizar seu desempenho. Diferente dos parâmetros do modelo, que são aprendidos durante o treinamento, os hiper parâmetros são definidos antes do início do processo de aprendizagem e controlam o comportamento do algoritmo.

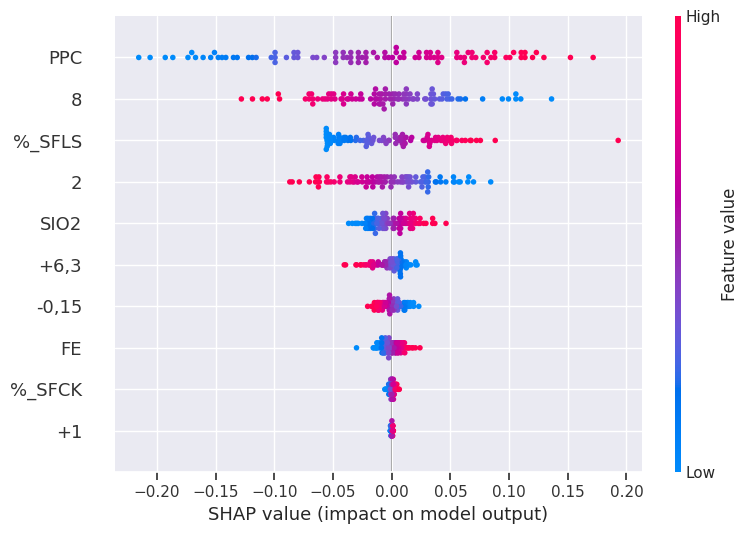
**Figura 20: Ajuste de hiper parâmetros (SVR)**



O ajuste de hiper parâmetros utilizando o objeto grid\_search, figura 20, permitiu testar várias combinações de valores de hiper parâmetros. Todavia, em função do elevado tempo de processamento, optou-se por testar individualmente os quatro valores individuais de ‘C’ e ‘epsilon’. O MAE de 0,11 obtido mostrou que não houve melhora do modelo com o ajuste realizado.

Como última etapa na avaliação do modelo SVR é apresentado abaixo o gráfico resumo gerado pela função SHAP, ele ajuda a entender como cada característica (variável) contribui para as previsões do modelo.

**Figura 21: Influência das variáveis pela função SHAP (SVR)**

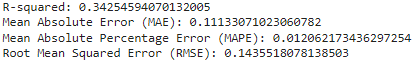
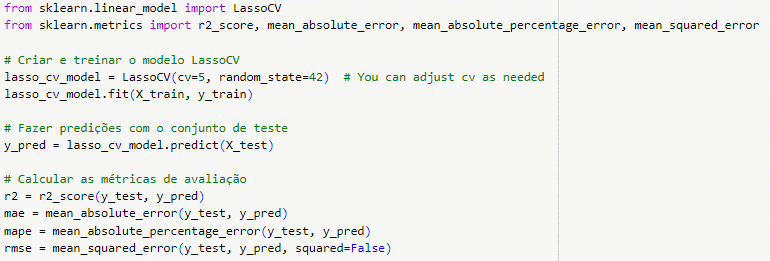


**Modelo LassoCV**

O LassoCV é uma variante da regressão Lasso que utiliza validação cruzada para encontrar automaticamente o melhor valor do parâmetro de regularização (\alpha). A regressão Lasso (Least Absolute Shrinkage and Selection Operator) é uma técnica de regularização que pode ser usada tanto para reduzir a complexidade do modelo quanto para selecionar características importantes, ao penalizar a soma dos valores absolutos dos coeficientes.

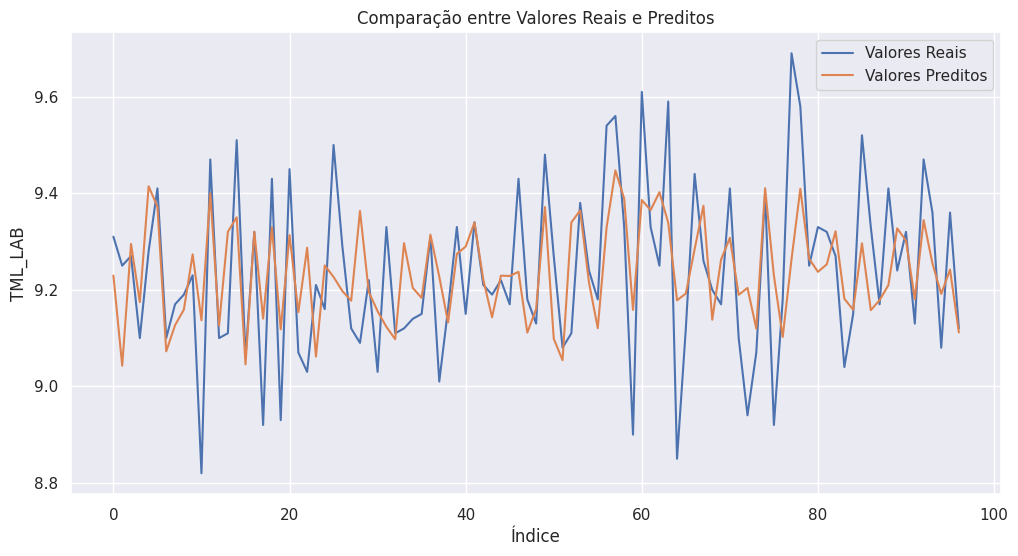
Abaixo é exibido o código utilizado com o resultado das métricas obtidas para o modelo LassoCV.

**Figura 22: Métricas utilizando LassoCV**



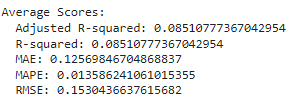
O gráfico de linhas abaixo ajuda a ver o resultado obtido entre os valores reais e os valores preditos pelo modelo.

**Figura 23: Gráfico de valores reais x valores preditos LassoCV**



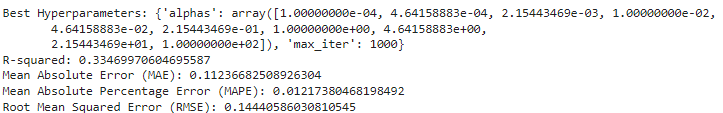
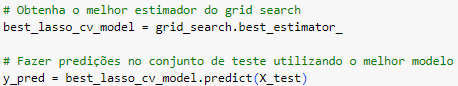
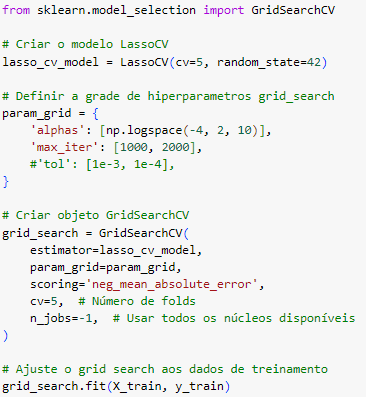
Os resultados obtidos na etapa de cross validation, figura 24, utilizando 5 folds mostraram a capacidade de generalização do modelo gerando um MAE médio igual a 0,13 contra um valor de 0,11 obtido anteriormente para todo o conjunto de dados.

**Figura 24: Cross Validation (LassoCV)**



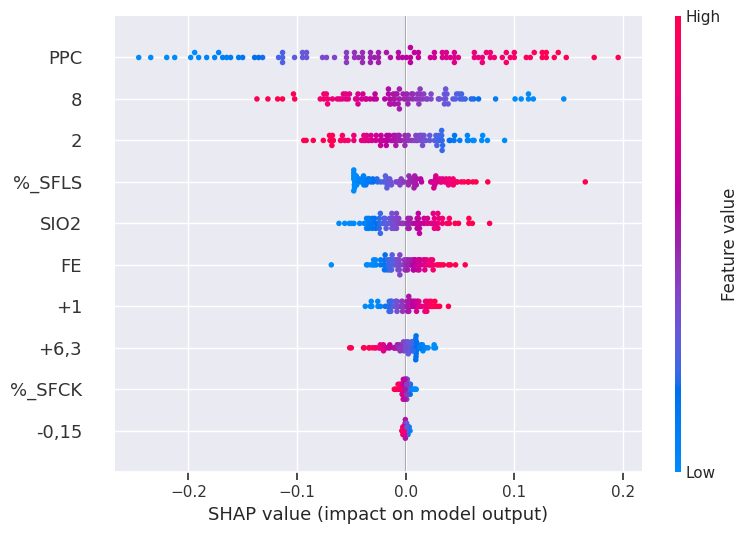
A seguir, partiu-se para o ajuste de hiper parâmetros. A figura 25 mostra o código utilizado e o resultado obtido.

**Figura 25: Ajuste de hiper parâmetros (LassoCV)**



Como última etapa na avaliação do modelo SVR é apresentado abaixo o gráfico resumo gerado pela função SHAP, ele ajuda a entender como cada característica (variável) contribui para as previsões nesse modelo.

**Figura 26: Influência das variáveis pela função SHAP (LassoCV)**

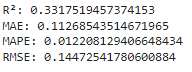
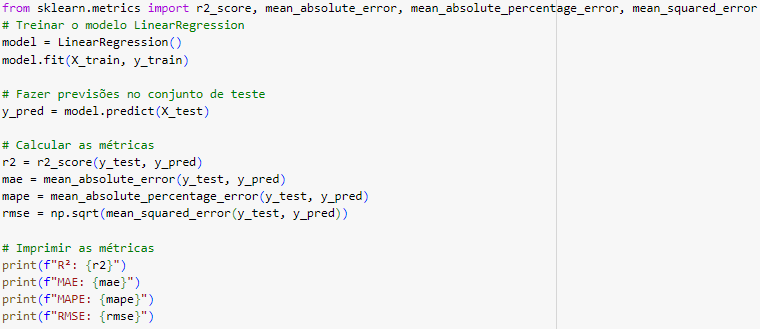


**Modelo Linear Regression**

O modelo LinearRegression é um dos mais simples e amplamente utilizados em aprendizado de máquina. Ele é usado para prever o valor de uma variável dependente (ou alvo) com base em uma ou mais variáveis independentes (ou características).

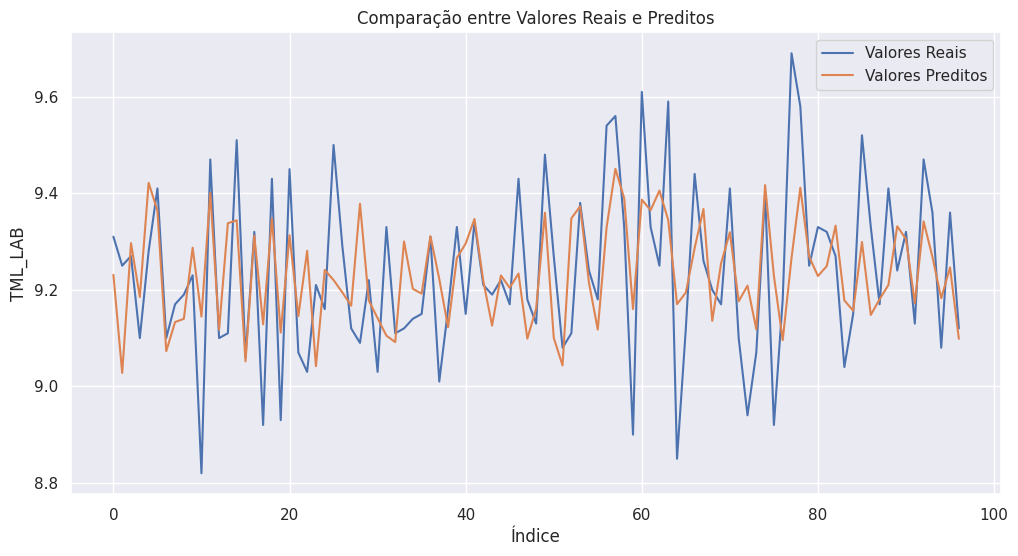
Abaixo é exibido o código utilizado com o resultado das métricas obtidas para o modelo LinearRegression.

**Figura 27: Resultado das métricas utilizando LinearRegression**



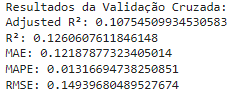
O gráfico de linhas abaixo ajuda a ver o resultado obtido entre os valores reais e os valores preditos pelo modelo.

**Figura 28: Gráfico de valores reais x valores preditos (LinearRegression)**



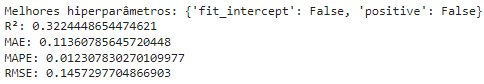
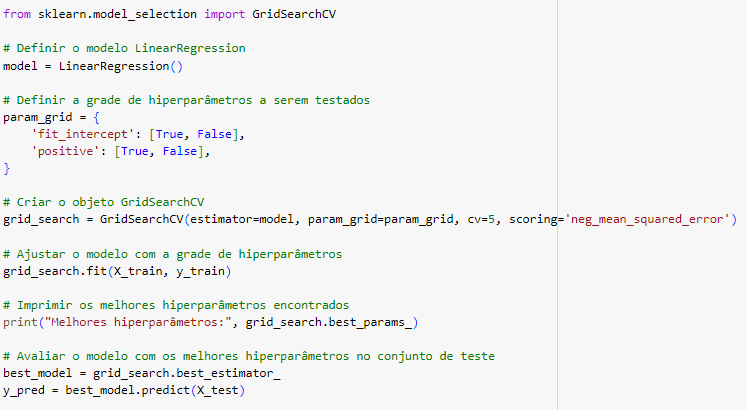
Os resultados obtidos na etapa de cross validation, figura 29, utilizando 5 folds mostraram a capacidade de generalização do modelo gerando um MAE médio igual a 0,13 contra um valor de 0,11 obtido anteriormente para todo o conjunto de dados.

**Figura 29: Cross Validation (LinearRegression)**



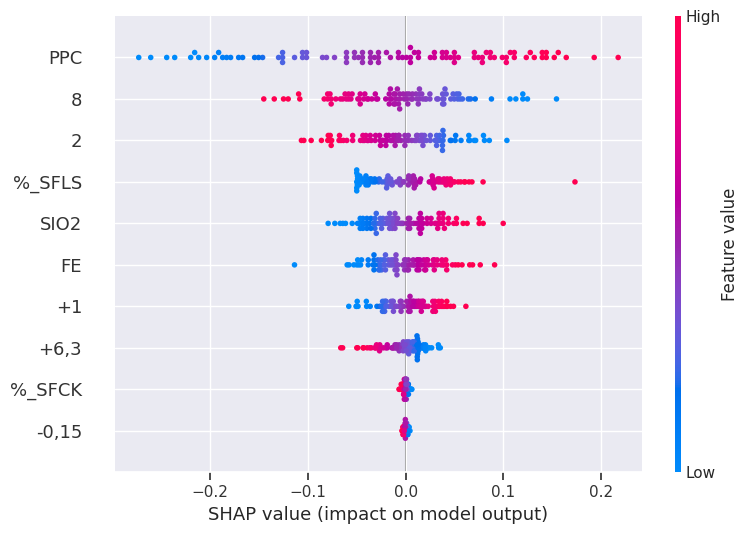
A seguir, partiu-se para o ajuste de hiper parâmetros. A figura 30 mostra o código utilizado e o resultado obtido.

**Figura 30: Ajuste de hiper parâmetros (LinearRegression)**



Por fim, é apresentado o gráfico resumo gerado pela função SHAP mostrando como cada característica (variável) contribui para as previsões nesse modelo.

**Figura 31: Influência das variáveis pela função SHAP (LinearRegression)**



# 6. Apresentação dos Resultados

A interpretação de resultados em algoritmos de machine learning requer uma abordagem multidisciplinar, considerando desde a definição do problema e, em alguns casos, até a ética envolvida na aplicação dos resultados. A compreensão clara das métricas, a qualidade dos dados, a análise de erros e a consulta a especialistas do domínio são fundamentais para uma correta interpretação, contribuindo para tomadas de decisão mais eficazes e responsáveis. Á luz de todos esses fatores, a seguir as conclusões obtidas.

O desenvolvimento do trabalho mostrou a importância da incorporação de novas variáveis na melhoria da performance dos modelos. Iniciou-se apenas considerando as variáveis granulométricas, depois foram adicionadas as variáveis químicas e , finalmente, composição por mina de origem.

Por outro lado, pensando em um processo industrial, a redução de dimensionalidade do modelo a partir das variáveis mais importantes pode ajudar a direcionar medidas no mundo real, isto é, para o caso concreto, colocar em prática ações que evitem a redução do TML. Para auxiliar essa tarefa lançou-se mão da matriz de correlação (figura 13). Aplicado o critério de escolha das variáveis de maior correlação, chegou-se a um conjunto de 10 variáveis para compor o modelo, destaque para a variável PPC (figura 14).

A seguir foi utilizada a biblioteca Lazy Predict para auxiliar a comparação rápida de diversos algoritmos de ML. Dos modelos performados (figura 15) escolhemos três modelos de regressão com diferentes características para avaliar o desempenho de cada um: i) SVR (Regressão de Suporte Vetorial) ii) Regressão Lasso e iii) Regressão Linear.

Nos passos seguintes, todos os modelos foram submetidos a etapas de cross validation e ajuste de hiper parâmetros em busca de validar a capacidade de generalização do modelo e melhoria da performance, respectivamente.

Remetendo à meta definida no início do trabalho, alcançar um MAE < 0.15, obteve-se os seguintes melhores resultados para cada modelo:

SVR: MAE = 0.1119 (figura 19)

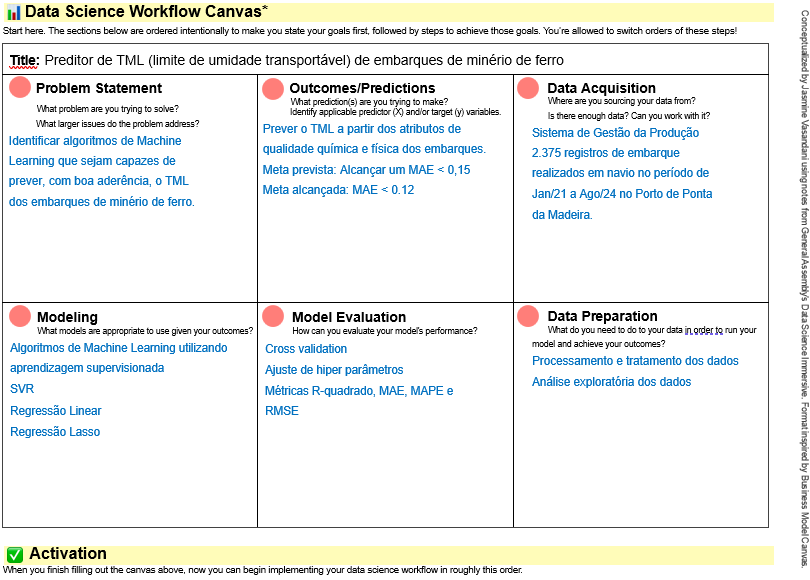
Regressão Lasso: MAE = 0,1113 (figura 21)

Regressão Linear: MAE = 0,1127 (figura 26)

Alcançado o objetivo traçado, cabem algumas considerações.

* Com base nos gráficos de TML previsto x TML realizado (figuras 17, 22 e 27), é possível perceber que os modelos tiveram uma boa capacidade de acompanhar a tendência do TML realizado, todavia se mostraram conservadores em ambas as direções.
* Mesmo executando o ajuste de hiper parâmetros com a busca em grade (*grid\_search*) não se conseguiu obter um melhor ajuste para curva, exceto para o modelo SVR (figura 19), que travou na execução da busca em grade e teve que ser limitado a um parâmetro de “C” e “épsilon” por vez, mas o ganho obtido foi muito pequeno.
* Para o propósito final do uso da predição do TML, superestimar o valor do TML não é um problema, mas o contrário sim. Nesse sentido, verificamos que o modelo, apesar de acompanhar a tendência, não foi capaz de acompanhar a amplitude dos valores de maior dispersão em relação à média.
* Considerando melhorias futuras, os efeitos acima podem ser mitigados com a implementação de algumas medidas tais como: i) aumentar a base de dados de treinamento, o que só será possível à medida que novos embarques forem realizados e ii) estudar a incorporação de novas variáveis que possam ajudar a explicar melhor as variações no TML.

Os resultados do projeto encontram-se resumidos na figura abaixo utilizando o modelo Canvas de Vasandani.



# 7. Links

Aqui você deve disponibilizar os links para o vídeo com sua apresentação de 5 minutos e para o repositório contendo os dados utilizados no projeto, scripts criados, etc.

Link para o vídeo: <https://youtu.be/LW1baSazg4U?si=SwsfTwGK1uIhXUxD>

Link para o repositório: <https://github.com/luciofbarboza/TCCPUCMG2024>

# REFERÊNCIAS

CLAUS O. Wilke. Fundamentals of Data Visualization. O’Reilly Media. https://clauswilke.com/dataviz/, 2019.

PYLE, Dorian. Data Preparation for Data Mining. The Morgan Kaufmann Series in Data Management Systems. 1999. ISBN-10: 1558605290.

Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems, 2011. ISBN-10: 9780123814791.

MULLER, Andreas C., and Sarah Guido. Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.", 2016.

Kour, Herleen, and Naveen Gondhi. "Machine Learning Techniques: A Survey." International Conference on Innovative Data Communication Technologies and Application. Springer, Cham, 2019.

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro, RJ: LTC, 2011. xvi, 378 p. ISBN 9788521618805.