Relatório Final MSI I
Pesquisa Científica

Thiago Campos[1], **Adriano Veloso (Orientador)**[2]

[1]Departamento de Ciencia da Computaçao
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

thiago.a.campos@hotmail.com, adrianov@dcc.ufmg.br

Abstract. *Machine learning models have been used to contribute to problem solving in different contexts of our society. In the food industry, specifically in the deodorization stage during the oil refining process, an increase in deodorizer pressure may occur, due to several factors, which leads to the phenomenon called vacuum break. This phenomenon causes problems in this process and damages the production chain of these industries, so it is important to identify and anticipate such a situation. This work presents supervised machine learning models that help in the prediction process of this phenomenon. The preliminary results obtained show a satisfactory error rate that contributes to the solution to the presented problem.*

Resumo. *Os modelos de aprendizado de máquina têm sido utilizados para contribuir na resolução de problemas em diferentes contextos da nossa sociedade. No ramo industrial alimentício, especificamente na etapa de desodorização durante o processo de refino de óleos, pode ocorrer um aumento da pressão do desodorizador, por fatores diversos, o que leva ao fenômeno denominado quebra de vácuo. Esse fenômeno causa problemas neste processo e prejudica a cadeia produtiva destas indústrias, por isso é importante identificar e antecipar tal situação. Este trabalho apresenta modelos de aprendizado de máquina supervisionados que auxiliam no processo de predição desse fenômeno. Os resultados preliminares obtidos mostram uma taxa de erro satisfatória que contribui na solução para o problema apresentado.*

1. Introdução

1.1. Óleos e Gorduras Vegetais

Os óleos e gorduras vegetais são componentes importantes e presentes em nosso dia a dia, sendo uma fonte de energia e de ácidos essenciais no nosso organismo [Rajko Vidrih 2010]. Vários alimentos que consumimos possuem essas substâncias mais evidentes em sua composição, como o óleo de soja e óleo de canola, assim como o seu uso em elementos base para produção de alimentos mais refinados como é o caso da margarina e da manteiga de cacau, esta última presente em alimentos como chocolates, doces e coberturas [M.R. Norazlina 2021]. O óleo bruto pode ser extraído de frutas ou sementes e possui substâncias como ácidos graxos livres, triglicerídeos, fosfatídeos entre outros. Alguns desses elementos são benéficos para os seres humanos, como vitaminas, diglicerídeos e polifenóis, outros como pesticidas, vestígios de metais e alguns óleos minerais podem fazer mal a saúde. Ainda, alguns itens não fazem mal a saúde, mas têm um

efeito negativo na qualidade e estabilidade de óleos, como ceras, pigmentos e produtos de oxidação [Gharby 2022].

Indústrias que utilizam esses óleos, como a indústria alimentícia, possuem um processo elaborado para a eliminação de substâncias não desejáveis e tóxicas do óleo cru que é o refinamento do óleo [John Wiley & Sons 2005]. Esse processo possui métricas que devem ser seguidas para que a qualidade dos óleos atendam a especificações definidas para a segurança dos consumidores de produtos finais que possuam esse item em sua composição. As duas formas mais comuns para realização desse procedimento são o refinamento químico ou físico dos óleos, que possuem uma série de etapas [Syed Sherazi 2016]. Este refinamento busca dar uma aparência mais agragável para o item, promover um gosto neutro e prover uma resistência maior a oxidação, além da eliminação dos componentes não desejados.

Após a extração do óleo cru é iniciado o fluxo de degomagem, que visa eliminar sustâncias como fosfolipídios e gomas mucilaginosas, elementos que podem dificultar o refinamento devido a insolubilidade no processo de hidratação feito posteriormente. A etapa de neutralização é feita no refinamento químico e consiste na eliminação de ácidos graxos, metais e clorofilas, sendo importante na regulação da acidez de determinados óleos [Kamel Essid 2009]. Em seguida, ocorre a lavagem e secagem, operação que elimina resíduos presentes no óleo resultantes na turbina da etapa de neutralização. Na fase de clareamento são eliminados materias solúveis do óleo, removidos pigmentos de cor, sabões e produtos de oxidação.

O processo de desodorização de óleos é uma das últimas etapas ao longo do processo refino de óleos. Como o próprio nome sugere, trata-se de um processo que elimina os mau odores presentes no óleo, que são resultado das etapas anteriores do refinamento. Nessa etapa ocorre uma destilação a vapor a vácuo, por meio da injeção de vapor de água no óleo. Esse procedimento envolve a passagem de vapor por diferentes camadas do óleo que são mantidas a temperaturas altíssimas, assim como a utilização de um vácuo muito alto [Gupta 2017].

As etapas de refinamento descritas envolvem vários equipamentos e máquinas com diferentes funcionalidades, componentes e características que modificam ou contribuem de alguma forma para o processo de refino apresentado [Gabriele Landucci 2013]. Então, é possível que ocorram configurações ou situações específicas nestas máquinas que levem a problemas ou ao insucesso dessas etapas. Assim, buscar mecanismos para identificar os fatores associados a essas configurações é uma tarefa desejável porém, ao mesmo tempo, desafiadora, que pode levar ao aumento da eficiência operacional nestas indústrias.

1.2. Aprendizado de Máquina

Aprendizado de Máquina é um ramo da Ciência da Computação que está inserido no campo da Inteligência Artificial [Shinde and Shah 2018]. Esta última têm como objetivo fazer com que máquinas realizem funções associadas ao cérebro humano, como o aprendizado e a resolução de problemas. Um dos principais objetivos da sub área de Aprendizado de Máquina é fazer com que algoritmos aprendam e não somente recebam instruções e produzam um resultado esperado [Iqbal Muhammad 2015]. Nesse sentido, existem diferentes tipos de aprendizados que ajudam a resolver problemas de naturezas

diversas, muitas vezes com o objetivo de obter uma descrição ou informação não conhecida a partir do dado observado ou, ainda, realizar uma predição de um fator conhecido a partir dos dados disponíveis, sendo alguns deles pontuados brevemente a seguir.

Os chamados modelos de aprendizado de máquina não-supervisionados são interessantes na resolução de problemas que buscam obter uma descrição dos dados sem um conhecimento prévio de uma variável de saída esperada. Assim, o seu objetivo é identificar dimensões, componentes, agrupamentos ou trajetórias contidas na estrutura dos dados em questão [Tammy Jiang 2020]. O seu uso pode ser visto em aplicações que visam a identificação de grupos ou dimensões não observadas na saúde mental de pacientes ou ainda na representação das relações de dados urbanos, descobrindo novos padrões e representações de cidades que podem auxiliar na tomada de decisão [Wang and Biljecki 2022].

O aprendizado supervisionado é caracterizado principalmente pela predição ou classificação de valores de saída conhecidos. Nesse sentido, são utilizados dados como entrada que servem para o aprendizado desses modelos com o objetivo de inferir uma função alvo que melhor mapeia dados de entrada para valores de saída reais esperados [Amanpreet Singh 2016]. Esse tipo de aprendizado têm sido aplicado em variados contextos como na área da saúde, social, financeira e alimentícia, assim como na predição de eventos ecológicos catastróficos [Crisci et al. 2012].

Existe ainda uma abordagem denominada aprendizado por reforço. Nesse cenário, agentes podem realizar ações e interagir com um ambiente ao qual eles estão inseridos com o objetivo de maximizar um esquema de recompensas, de forma que cada ação realizada mapeia uma recompensa para o agente [Alharin et al. 2020]. A aplicação dessa vertente de aprendizado está presente na robótica, desenvolvimento de carros autônomos, automação de agentes em jogos, entre várias outras.

1.3. Objetivos

Uma vez apresentados os principais conceitos pertinentes ao trabalho proposto, este trabalho tem como objetivo a aplicação de modelos de aprendizado de máquina supervisionados no âmbito do processo de refino de óleos comestíveis, especificamente, na etapa de desodorização do oléo em seu refinamento. Para isso, serão utilizados dados de sensores presentes em máquinas que atuam durante esta etapa, contendo informações como temperatura, pressão e nível de substâncias. Assim, espera-se que as técnicas de aprendizado de máquina empregadas sejam capazes de capturar relações presentes entre esses diferentes sensores de forma a predizer o valor de um determinado sensor alvo.

1.4. Estrutura do trabalho

O trabalho apresenta uma primeira seção de introdução, contendo uma motivação e explicações relacionadas aos óleos e gorduras vegetais, assim como o processo de refinamento desses óleos e uma breve descrição de suas etapas. Em seguida, são apresentados os conceitos e principais tipos de modelos de aprendizado de máquina, com suas características e relevância. Ainda na introdução, são discutidos os objetivos e a estrutura na qual o trabalho está organizado. No segundo tópico deste trabalho é apresentado o referencial teórico, contendo uma descrição das características e aplicações dos óleos já desenvolvidas na literatura, assim como uma apresentação mais específica de trabalhos

em que modelos de aprendizados de máquina foram utilizados para solucionar problemas em variados contextos.

Na seção de Metodologia são descritos os principais conceitos e o que foi desenvolvido no trabalho proposto. Descrevendo as biliotecas utilizadas, como o dado foi preparado, quais premissas foram assumidas, quais modelos foram utilizados, quais variações foram realizadas com o intuito de obter melhores resultados entre outros pontos. Na etapa de Resultados, são mostrados os resultados obtidos no trabalho, com tabelas mostrando o desempenho de modelos, discorrendo e analisando as métricas de erro obtidas. Por fim, é realizada a conclusão do trabalho, discorrendo sobre o que foi realizado, uma reflexão sobre os resultados obtidos e o que pode ser realizado em trabalhos futuros, assim como a exposição das referências bibliográficas utilizadas.

2. Referencial Teórico

2.1. Características e Aplicações dos Óleos

Um estudo realizado em [Rajko Vidrih 2010] mostra um comparativo entre as composições de ácidos graxos e outros parâmetros de qualidade de óleos refinados e não refinados. Para isso, foram coletadas amostras de óleos vegetais refinados e não refinados de diferentes localidades como Áustria, Eslovênia e Holanda. Então, tais óleos foram submetidos a uma série de processos a fim de determinar a quantidade dos componentes citados em sua composição, além da sua estabilidade oxidativa. Os resultados obtidos pelos autores mostram que os óleos vegetais inclusos na pesquisa possuem níveis aceitáveis de parâmetros de qualidade aceitos por entidades reguladoras e que a estabilidade oxidativa de óleos não-refinados foi melhor do que nos óleos refinados, sendo óleos com essa estabilidade mais adequados para frituras. Tal estudo exemplifica a importância e diversidade de substâncias contidas em óleos vegetais e a necessidade da sua regulação para diferentes fins.

Em [M.R. Norazlina 2021], por sua vez, óleos e gorduras vegetais são utilizadas na produção das chamadas alternativas a manteiga de cacau. A manteiga de cacau é um recurso natural de gordura que é utilizado na produção de chocolates, doces e coberturas. O estudo mostra que com o aumento da demanda por essa substância, além da sua disponibilidade que é limitada, ela têm se tornado uma gordura cara. Dessa forma, a exploração de alternativas a ela vêm sendo cada vez mais frequente. Uma dessas alternativas é feita a partir da mistura de frações de gorduras e óleos com o intuito de melhorar propriedades da manteiga de cacau, com o objetivo de substituir ou diminuir a necessidade dessa substância, de forma a reduzir custos para indústrias que a utilizam. Os autores concluem que vegetais puros, possuem limitações no fornecimento de propriedades desejáveis a manteiga de cacau, porém a mistura desses componentes melhorou essas propriedades significativamente, além da possibilidade do seu uso na fabricação de produtos de confeitaria e chocolates.

Sob uma outra perspectiva, diferentes tipos de óleos podem apresentar composições variadas e uma concentração maior ou menor de determinadas propriedades. Possuir um detalhamento e um comparativo dessas variações é importante para sua aplicação em diferentes processos. A partir de um processo de extração, limpeza, filtro e armazenamento realizado em [Brahmi et al. 2020], foi produzida uma análise comparativa entre a composição química e biológica de óleos das sementes *Pistacia lentiscus*

L.(PL), Opuntia ficus-indica (L) e Argania spinosa L. Skeels (AS). Nesse estudo, os autores verificam a diferença entre componentes como ácidos graxos, fitoesteróis e perfis fenólicos e destacam uma baixa atuação antioxidante e atividades antimicrobianas nos óleos. Porém, apesar disso, concluem que a presença dos componentes citados mostram um potencial para a saúde humana, seja de forma nutricional ou em contextos médicos.

2.2. Aplicações de Aprendizado de Máquina

Os modelos de aprendizado de máquina tem sido cada vez mais adotados na resolução de diversos problemas em diferentes contextos da nossa sociedade. Na área da saúde, por exemplo, em [Lv Cai-Xia and Wei 2021] é realizada uma predição do número de casos de febre hemorrágica com síndrome renal na China. Para isso, foram coletados dados que representam a incidência da doença ao longo do período de 2004 a 2018 e foi realizado um comparativo entre a estimativa ou predição utilizando um modelo de aprendizado supervisionado e um modelo baseado em estatística denominado ARIMA (Autoregressive Integrated Moving Average). Os resultados mostraram uma melhor acurácia e estabilidade nas métricas avaliadas no modelo de aprendizado supervisionado.

No campo da agricultura, um estudo realizado na China [Cao et al. 2023], realiza uma previsão de temperatura em estufa com base em recursos de séries temporais e modelos de aprendizado de máquina supervisionados. A proposta apresentada discute a problemática da dificuldade de previsão acurada da temperatura em estufas, considerando as mudanças em diferentes componentes internos e externos a esse espaço, como a umidade e pressão do ar, além da utilização de *features* computadas como a diferença de temperatura ou umidade no ambiente interno e externo em um intervalo de tempo. Assim, a partir de um comparativo entre diferentes algoritmos de aprendizado, como *LightGBM* baseado no algoritmo *GBDT(Gradient Boosting Decision Tree), SVM(Support Vector Machine) e Linear Regression* foi verificado que o modelo com LightGBM performou melhor, assim como apresentou um tempo de treino menor em comparação com os outros para o problema apresentado. Os autores concluem que os resultados são importantes e prestam um papel importante na melhora da qualidade e desenvolvimento de uma agricultura inteligente.

Nesse mesmo contexto, modelos de aprendizado de máquina podem ser utilizados na previsão de melhores ou colheitas mais favoráveis considerando determinadas características do ambiente [Kalimuthu et al. 2020]. O sistema proposto pelo trabalho citado possibilita que agricultores forneçam parâmetros como a temperatura e sua localização em um aplicativo e este possa indicar qual semente é mais apropriada para plantio considerando uma safra mais satisfatória e rentável. Para isso, o aplicativo utiliza como suporte um modelo baseado no algoritmo de *Naive Bayes* que recebe como parâmetros atributos como a temperatura, umidade do ar e pluviosidade e possui como variável alvo a categoria de semente a ser plantada. A contribuição destacada pelos autores é referente a ajuda proporcionada a agricultores que possuem menos conhecimento na predição de safras e também propõem uma extensão do trabalho para sugerir fertilizantes e orientações adequadas para terras cultiváveis.

Em [Gan et al. 2021], por sua vez, o modelo *LightGBM* é aplicado na predição dos níveis de água do Baixo Rio Columbia. Nesse estudo, é discutida a relação não linear entre a descarga dos rios e as marés nos estuários, sendo estes não estacionários e considerando que seus mecanismos ainda não foram totalmente compreendidos. A proposta

de aprendizado de máquina nesse contexto se baseia nas características desses modelos de capturar relações desconhecidas entre essas variáveis. Como entradas para o modelo foram utilizadas descargas dos rios Columbia e Willamette, assim como as características e constituintes das marés. Os resultados obtidos se mostraram satisfatórios em comparação com outras abordagens com o mesmo objetivo e foi destacada a revelação da importância dos contituintes das marés em sua interação com os rios nos estuários.

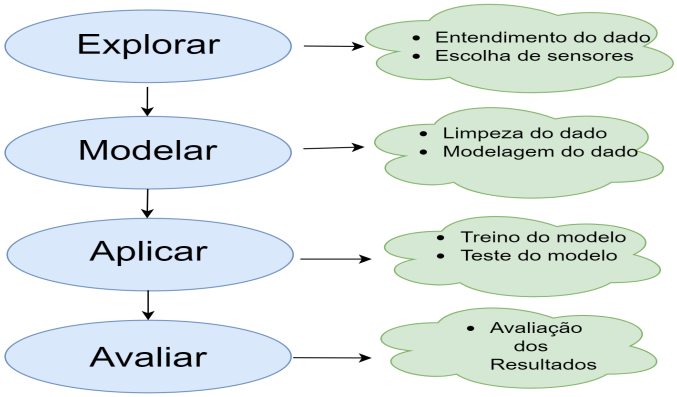
A partir dos trabalhos apresentados, fica evidente a aplicação, importância e contribuição dos modelos de aprendizado de máquina em variados contextos como na saúde, agricultura e ambiental. No melhor que conhecemos, não foram encontrados trabalhos que relacionam o aprendizado de máquina com o processo de refino de óleos, especificamente na etapa de desodorização de óleos. Dessa forma, esse trabalho busca contribuir nesse processo a partir do uso de aprendizado supervisionado e com o intuito de melhorar a eficiência operacional das indústrias que atuam nesse ramo.

3. Metodologia

A metodologia a ser utilizada no trabalho pode ser dividida em quatro etapas, sendo elas: Explorar, Modelar, Aplicar e Avaliar. A etapa de exploração tem como processo uma avaliação dos dados disponíveis e seu entendimento, buscando entender as características e como o dado está organizado. Já a etapa de modelagem, visa realizar uma limpeza e sanitização do dado, buscando adaptá-lo e deixá-lo pronto para a aplicação de modelos de aprendizado de máquina, sendo realizados processamentos em sua composição e criando novos formatos de dados a partir do original.

Em seguida, temos a aplicação de modelos de aprendizado de máquina supervisionado no dado pré modelado disponibilizado na etapa anterior, realizando o treino dos modelos e o seu teste a partir dos subconjuntos de treino e teste extraídos e disponibilizados na etapa de modelagem. Por fim, ocorre uma avaliação dos resultados obtidos com o uso de métricas, analisando a acurácia dos modelos com relação aos valores reais verificados no dado e realizando um comparativo entre os resultados obtidos. A Figura 1 mostra um fluxo simplificado que ilustra cada uma dessas etapas.

Figura 1. Etapas da Metodologia do Trabalho



3.1. Explorar

Os dados utilizados no trabalho são referentes a valores capturados por sensores presentes no processo de desodorização de óleos de uma indústria. Foram coletados aproximadamente 1 milhão, duzentos e trinta e quatro mil e novecentos registros. Conforme citado, cada registro, representa o valor de um conjunto de sensores em um instante de tempo, sendo valores que representam a temperatura, vazão, pressão, entre outros. Na tabela 1 mostrada abaixo é possível verificar um resumo de informações dos sensores utilizados.

Tabela 1: Atributos do dado

Atributos do Dado		
Tag	Descrição	Código
time	Momento de coleta dos valores dos sensores	T
gme-ref-03-desodorizador-03-AT-832B-1.EU	pH da caixa barométrica	PH
gme-ref-03-desodorizador-03-COMANDOS-GENERALES-AUTO.25	NA	CGA25
gme-ref-03-desodorizador-03-COMANDOS-GENERALES-AUTO.26	NA	CGA26
gme-ref-03-desodorizador-03-ESTADOS-GENERALES-AUTO.25	NA	EGA25
gme-ref-03-desodorizador-03-ESTADOS-GENERALES-AUTO.26	NA	EGA26
gme-ref-03-desodorizador-03-FT-846E-1.EU	Medidor de Vazão de Vapor Direto	FT-846
gme-ref-03-desodorizador-03-FT-878GL-1.EU	Medidor	FT-878
gme-ref-03-desodorizador-03-FT-881W-1.EU	Vazao da Caixa Barométrica	FT-881
gme-ref-03-desodorizador-03-FT-P801-1.EU	Medidor de Vazão de Entrada	FT-P801
gme-ref-03-desodorizador-03-FT-P814AG-1.EU	Medidor de vazão de ácido graxo	FT-P814AG
gme-ref-03-desodorizador-03-FT-SAC-R-1.EU	Medidor de vazão de vapor geral	FT-SAC-R-1
gme-ref-03-desodorizador-03-Job-List-802M.Code	Categórico	802M
gme-ref-03-desodorizador-03-Job-List-821HP.Code	Categórico	821HP
gme-ref-03-desodorizador-03-Job-List-821TS1.Code	Categórico	821TS1
gme-ref-03-desodorizador-03-Job-List-821TS2.Code	Categórico	821TS2
gme-ref-03-desodorizador-03-Job-List-822A1.Code	Categórico	822A1
gme-ref-03-desodorizador-03-Job-List-822A2.Code	Categórico	822A2
gme-ref-03-desodorizador-03-Job-List-880B.Code	Categórico	880B
gme-ref-03-desodorizador-03-Job-List-880TS1.Code	Categórico	880TS1
gme-ref-03-desodorizador-03-Job-List-880TS2.Code	Categórico	880TS2
gme-ref-03-desodorizador-03-Job-List-880WC.Code	Categórico	880WC
gme-ref-03-desodorizador-03-PT-814AG-2.EU	Pressão de circulação de ácido graxo	PT-814AG
gme-ref-03-desodorizador-03-PT-814-23P-1.EU	Vácuo após o scrubber	PT-814-23P
gme-ref-03-desodorizador-03-PT-846E-1.EU	Pressão de vapor direto	PT-846E-1
gme-ref-03-desodorizador-03-PT-846F-1.EU	Pressão de vácuo do desodorizador	PT-846F-1
gme-ref-03-desodorizador-03-PT-846F-2.EU	Pressão de vácuo do desodorizador	PT-846F-2
gme-ref-03-desodorizador-03-PT-881W-1.EU	Pressão da bomba da caixa barométrica	PT-881W-1
gme-ref-03-desodorizador-03-PT-SAC-R-3.EU	Pressão de vapor geral do deso 03	PT-SAC-R-3
gme-ref-03-desodorizador-03-PT-WCD-R-2.EU	Pressão do coletor da água barométrica	PT-WCD-R-2
gme-ref-03-desodorizador-03-Program:DESO-III.PT-88.EU	Pressão ejetor Z	DESO-III-PT-88
gme-ref-03-desodorizador-03-Program:DESO-III.PT-89.EU	Pressão ejetor X	DESO-III-PT-89
gme-ref-03-desodorizador-03-Program:DESO-III.PT-90.EU	Pressão ejetor Y	DESO-III-PT-90
gme-ref-03-desodorizador-03-TT-814-23P-1.EU	Temperatura do scrubber	TT-814-23P-1
gme-ref-03-desodorizador-03-TT-814-23P-2.EU	Temperatura dos gases do scrubber	TT-814-23P-2

Atributos do Dado		
Tag	Descrição	Código
gme-ref-03-desodorizador-03-TT-821HP-1.EU	Temperatura da bandeja 4.2	TT-821HP-1
gme-ref-03-desodorizador-03-TT-821HP-2.EU	Temperatura da bandeja 4.1	TT-821HP-2
gme-ref-03-desodorizador-03-TT-822A1-1.EU	Temperatura da bandeja 5	TT-822A1-1
gme-ref-03-desodorizador-03-TT-822A2-1.EU	Temperatura da bandeja 6	TT-822A2-1
gme-ref-03-desodorizador-03-TT-841A-C1-1.EU	Temperatura condensador 1	TT-841A-C1-1
gme-ref-03-desodorizador-03-TT-841A-C2-1.EU	Temperatura condensador 2	TT-841A-C2-1
gme-ref-03-desodorizador-03-TT-841A-C3-1.EU	Temperatura condensador 3	TT-841A-C3-1
gme-ref-03-desodorizador-03-TT-878GL-1.EU	Temperatura do tanque de água gelada	TT-878GL-1
gme-ref-03-desodorizador-03-TT-880B-1.EU	NA	TT-880B-1
gme-ref-03-desodorizador-03-TT-881AG-2.EU	Temperatura de circulação do ácido graxo	TT-881AG-2
gme-ref-03-desodorizador-03-TT-881W-1.EU	Temperatura da caixa barométrica	TT-881W-1
gme-ref-03-desodorizador-03-TT-881W-3.EU	Temperatura da água gelada do chiller	TT-881W-3
gme-ref-03-desodorizador-03-TT-WCD-R-2.EU	Temperatura de água de condensação	TT-WCD-R-2

Conforme pode ser verificado, existem sensores que coletam informações variadas. Além disso, não foi possível coletar uma descrição para alguns dos sensores, sendo preenchido com a sigla NA (Não se Aplica). O sensor alvo do trabalho, ou seja, que pretendemos predizer o seu valor é o sensor com código PT-846F-2, representando a pressão de vácuo do desodorizador. Um ponto importante que vale ser destacado é que alguns desses sensores não foram utilizados como atributos para os modelos desenvolvidos. Isso se deve ao fato desses sensores, por conhecimento empírico prévio, não impactarem ou não possuírem relação com o sensor alvo, ou, ainda, que coletam informações muito similares a ele, como é o caso do sensor PT-846F-1.

Além disso, para fins práticos do trabalho e levando em consideração o processo industrial da empresa ao qual os dados foram coletados, assumimos que um valor para o sensor alvo PT-846F-2 maior que 10, representa uma quebra de vácuo. Essa expressão se refere ao problema na qual a pressão do desodorizador está acima de um limite aceitável para um pleno funcionamento do processo de desodorização. Dessa forma, espera-se que os modelos desenvolvidos sejam capazes de capturar ou predizer esse fenômeno.

3.2. Modelar

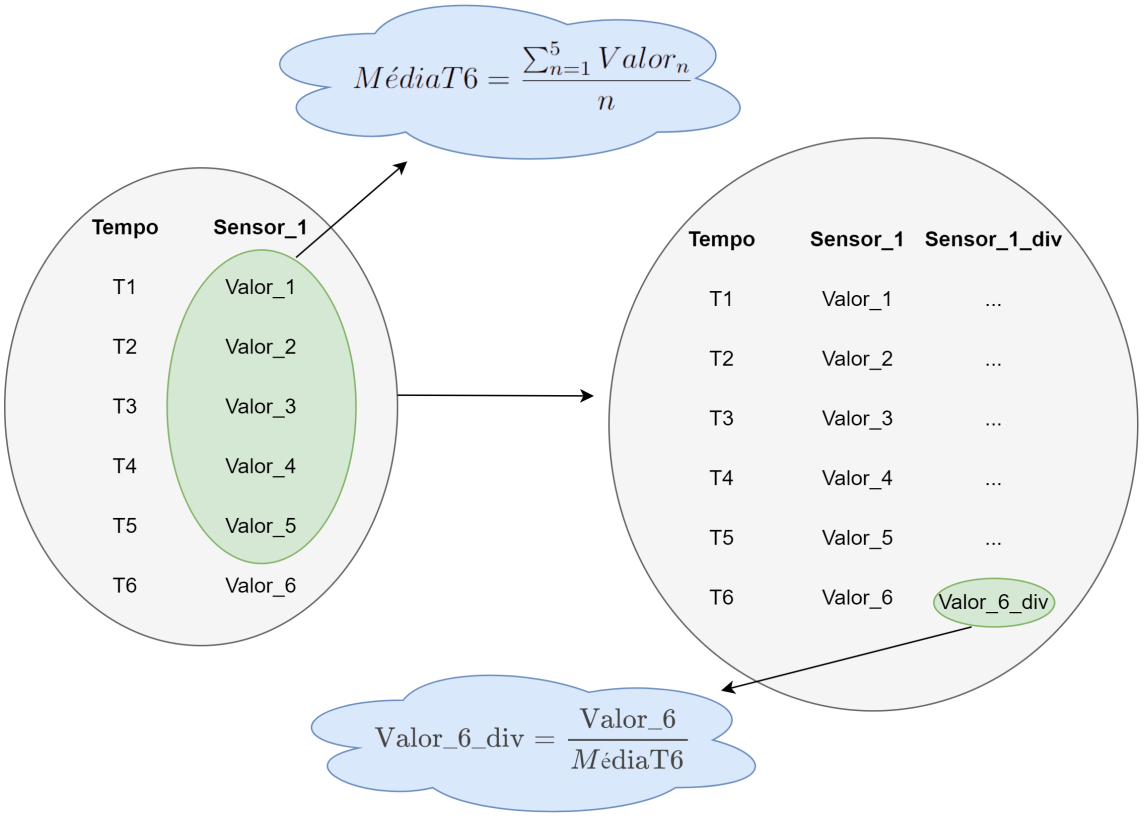
Na etapa de modelagem foram realizadas operações no dado observado, com o intuito de deixá-lo preparado para a aplicação de modelos de aprendizado de máquina. Uma dessas operações foi a remoção de registros não desejáveis para utilização no modelo, sendo esses registros caracterizados por possuírem o sensor alvo PT-846F-2 com um valor maior do que 20. A justificativa para essa remoção se deve ao fato destes registros refletirem momentos onde as máquinas não estavam em seu pleno funcionamento, portanto esses valores se tornariam ruídos para os modelos construídos. Além disso, os nomes das colunas que representam os nomes dos sensores foram renomeados de forma a remover caracteres especiais, permitindo somente letras e números sem acento.

Uma outra atividade realizada foi a separação e remoção das colunas que não

foram utilizadas nos modelos. As colunas com código T, PT-846F-1, PT-814-23P, FT-881, FT-878 e PT-846F-2. Conforme comentado na seção anterior, esses sensores tem um comportamento similar ou diretamente afetado pelo sensor alvo ou não influenciam no processo apresentado. Ademais, os dados foram separados em um conjunto de dados de treino e um conjunto de dados de teste, a serem utilizados na aplicação dos modelos, sendo que 80% dos dados foram destinados ao treino dos modelos e 20% para teste.

Na subseção de Aplicar, ainda nesta seção de Metodologia, será apresentado um modelo que possui como variação a utilização de atributos computados em sua composição. Assim, a Figura 2, mostra um esquema detalhando de como o dado é preparado para a aplicação desse modelo e como é feito o cálculo desses novos atributos, que será discutido a seguir.

Figura 2. Esquema de Computação de Novos Atributos



A partir da análise da Figura 2, é possível verificar o cálculo de uma média no tempo 6 para o sensor 1 exemplificado na imagem. Essa média leva em consideração os cinco valores do mesmo sensor nos cinco registros passados, considerando que o dado está organizado no formato de uma série temporal ordenada de forma crescente. Assim, foi adicionada uma nova coluna com o mesmo nome da coluna do sensor, porém com a terminação _div, de forma que os valores para essa coluna são preenchidos a partir da divisão do valor do sensor naquele momento e a média daquele sensor, calculada levando em consideração os cinco registros passados. Essa operação é feita para todos os sensores utilizados no modelo, ou seja, ao final teremos o dobro de colunas iniciais, pois para cada sensor será adicionada uma nova coluna.

Os novos atributos computados, representam uma espécie de histórico para os valores dos sensores, de forma que esses atributos indicam se o valor atual do sensor está aumentando ou diminuindo em relação a sua média histórica, que representa um intervalo de 15 segundos, aproximadamente. Dessa forma, o modelo criado com essa variação recebe mais atributos para sua predição que podem contribuir em seu desempenho.

3.3. Aplicar

Uma vez que o dado passou por um processo de limpeza e modelagem nas etapas anteriores, é possível utilizá-lo em modelos de aprendizado de máquina. Neste trabalho foram desenvolvidos três modelos de aprendizado supervisionado, apresentando como variações a utilização de diferentes implementações e uma versão contendo atributos computados. Para isso, foram utilizados *frameworks* que implementam os modelos LightGBM e XGBoost que serão discutidos brevemente a seguir.

O modelo XGBoost [Chen and Guestrin 2016] é um algoritmo baseado em *Boosting* que utiliza o algoritmo *Gradient Boosting Decision Tree (GBDT)* e um modelo linear para alcançar uma predição mais acurada. Um dos fatores que o tornam mais rápido que outros algoritmos que utilizam GBDT é devido a computação paralela em uma única máquina que ele propõe. A proposta de algoritmos baseados em *Boosting* se baseia na ideia da utilização de modelos mais fracos ou simples, de forma a utilizar os seus erros para construção de um modelo mais forte, no caso do GBDT, são utilizados modelos baseados em árvore de decisão. Para o trabalho desenvolvido, o XGBoost foi utilizado para a tarefa de regressão, ou seja, predizer o valor da variável alvo discutida anteriormente, utilizando como entradas os registros da base apresentada.

O modelo LightGBM foi originalmente proposto pela Microsoft [Ke et al. 2017], sendo um modelo que também utiliza o algoritmo GBDT e possui uma proposta similar ao XGBoost, também apresentando modelos de árvore de decisão como modelos fracos. Como diferencial, o LightGBM proporciona outros três algoritmos que ajudam a acelerar o seu processo de treino, assim como reduzir o consumo de memória, sendo eles o *Histogram-Based, Gradient-Based One-side Sampling (GOSS) e Exclusive Feature Bundling (EFB).* Assim como no XGBoost, o LightGBM foi utilizado para a tarefa de regressão na predição da variável alvo apresentada.

Como ambiente operacional do trabalho, foi utilizada a plataforma Google Colab [1] que provisiona um ambiente em nuvem para implementação e execução de códigos. Neste trabalho foi utilizada a linguagem de programação Python e suas bibliotecas auxiliares para manipulação de conjuntos de dados como *pandas* e *numpy*. Para implementação dos modelos discutidos, foram utilizadas as bibliotecas *xgboost* e *lightgbm* disponíveis para a linguagem Python. Além disso, foi utilizada a biblioteca de aprendizado de máquina *scikit-learn* para cálculo de métricas e separação dos dados de treino e teste, assim como *matplotlib* para exibição de gráficos do trabalho.

Os dois modelos iniciais desenvolvidos utilizando LightGBM e XGBoost, foram implementados de forma que os hiperparâmetros para os modelos não fossem alterados, ou seja, permaneceram como padrão das bibliotecas. Assim, foi realizado o treino dos modelos com o subcojunto de treino dos dados e suas avaliações foram feitas utilizando o

[1] https://colab.research.google.com/

subconjunto de teste dos dados, que representam 20% de sua composição. Já no terceiro modelo implementado, foi utilizado o XGBoost juntamente com os atributos computados descritos na seção 3.2 (Modelar), também realizando o treino e validação do modelo.

3.4. Avaliar

Para avaliação dos modelos desenvolvidos, foram utilizadas algumas métricas de erros comuns em modelos de aprendizado supervisionados de regressão que estão dispostas a seguir.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - x_i}{y_i}|.100\%$$

Em todas as equações o valor de *y_i* representa o valor observado ou real para cada amostra, já o valor de *x_i* representa o valor previsto pelo modelo para cada amostra. A primeira delas é a *Mean Absolute Error (MAE)* que realiza a média de erros absolutos levando em consideração o total de amostras. A métrica *Root Mean Squared Error (RMSE)* é bem similar a MSE, porém devido a raiz e elevação quadrática dos erros observados ela é interessante na captação de erros maiores. Já a *Mean Squared Error (MSE)* é mais sensitiva a erros maiores, porém o seu resultado muitas vezes não é apresentado em termos da variável alvo que está sendo prevista, podendo dificultar sua análise. A *Mean Absolute Percentage Error (MAPE)*, por sua vez, é geralmente utilizada em relatórios devido ao erro ser medido em percentual, portanto ela fornece uma ideia de acurácia do modelo em termos percentuais.

Após o treino e teste dos modelos, foram aplicadas cada uma das métricas descritas acima em cada um dos modelos, com o objetivo de verificar as diferenças entre cada um deles. A Tabela 2 mostrada abaixo exibe essa relação dos erros para cada um dos modelos.

Tabela 2. Métricas por Modelos

Métricas	LightGBM	XGBoost	XGBoost Final
MAE	1.14	0.99	0.92
RMSE	1.52	1.36	1.27
MSE	2.31	1.86	1.63
MAPE	0.19	0.17	0.16

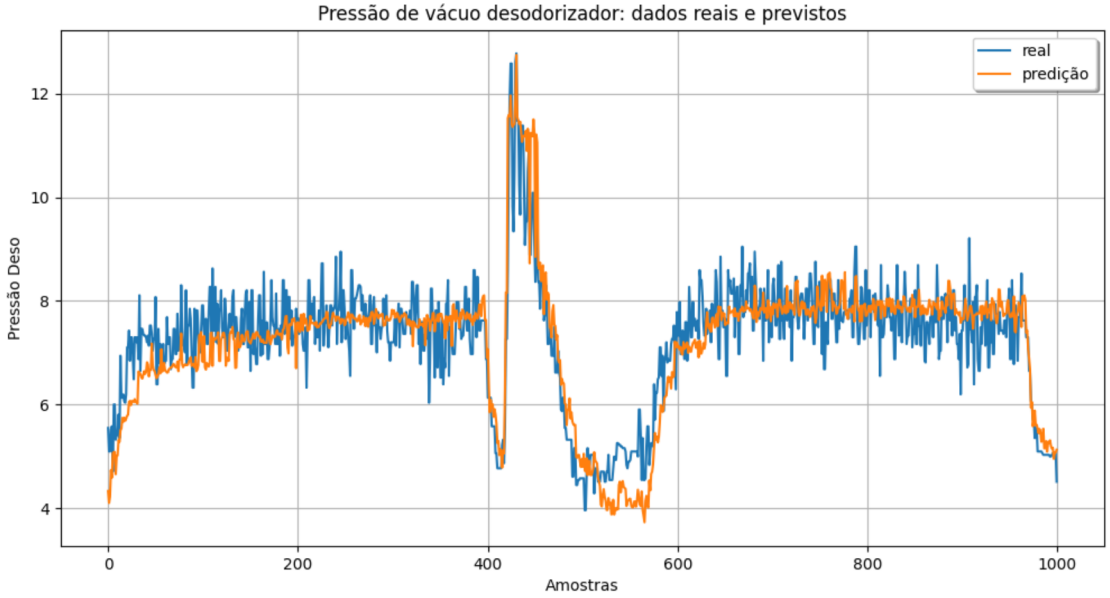
Conforme podemos verificar na Tabela 2, o modelo implementado com o LightGBM foi o que apresentou o pior desempenho quando comparado com os outros dois

modelos. Já o modelo implementado com o XGBoost obteve uma melhora significativa em todas as métricas de avaliação, com destaque para o MSE com uma redução de 2.31 para 1.86. Por fim, o modelo denominado XGBoost Final foi a implementação realizada com o XGBoost juntamente com os atributos computados descritos na seção 3.2 (Modelar). O resultado dessa implementação foi superior as versões sem os atributos computados, indicando que os atributos tiveram um impacto positivo no desempenho do modelo.

4. Resultados Preliminares

A partir da avaliação das métricas obtidas com cada um dos modelos descrita na seção 3.4, foi utilizado como modelo final o XGBoost com os atributos computados, por ter apresentado o melhor desempenho. Então, nesta seção serão avaliados diferentes intervalos de tempo no qual o modelo final apresentou comportamentos variados discutidos a seguir.

Figura 3. Intervalo com quebra de vácuo capturada pelo modelo final



A Figura 3 mostra um intervalo onde estão representados os valores reais para a variável alvo destacada pela linha azul no gráfico, assim como qual foi a predição do modelo final para essa variável ao longo da linha laranja dentro desse montante de mil amostras em um intervalo de tempo definido. É possível observar que próximo a amostra de número quatrocentos, ocorre uma quebra de vácuo, já discutida anteriormente, na qual a variável alvo passa a possuir um valor maior que dez. O modelo nesse caso foi capaz de predizer e acompanhar o valor da pressão do desodorizador, além de apresentar um valor de predição bem condizente do valor real ao longo do intervalo, embora não oscile tanto como ele.

Por outro lado, a Figura 4 mostra um intervalo de tempo no qual não houve uma quebra de vácuo e, mesmo assim, o modelo foi capaz de acompanhar o valor real da pressão do desodorizador. Porém, aqui vale destacar que nos intervalos aproximados de 0 a 200 e 700 a 800 o modelo obteve um descolamento maior com os valores reais, indicando pontos de melhoria que ainda podem ser explorados.

Figura 4. Intervalo sem quebra de vácuo com o modelo final

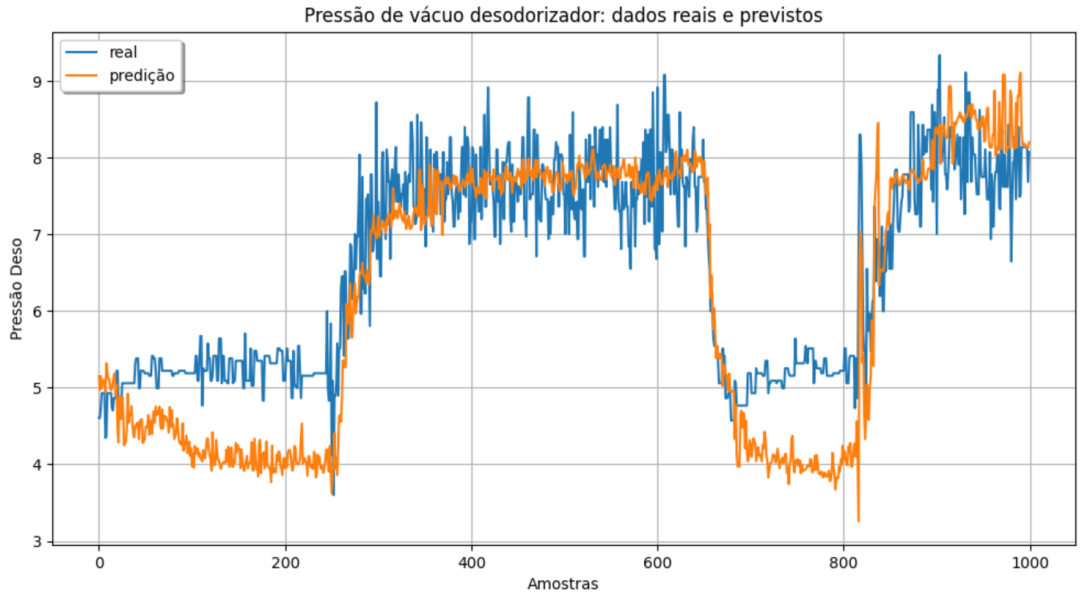
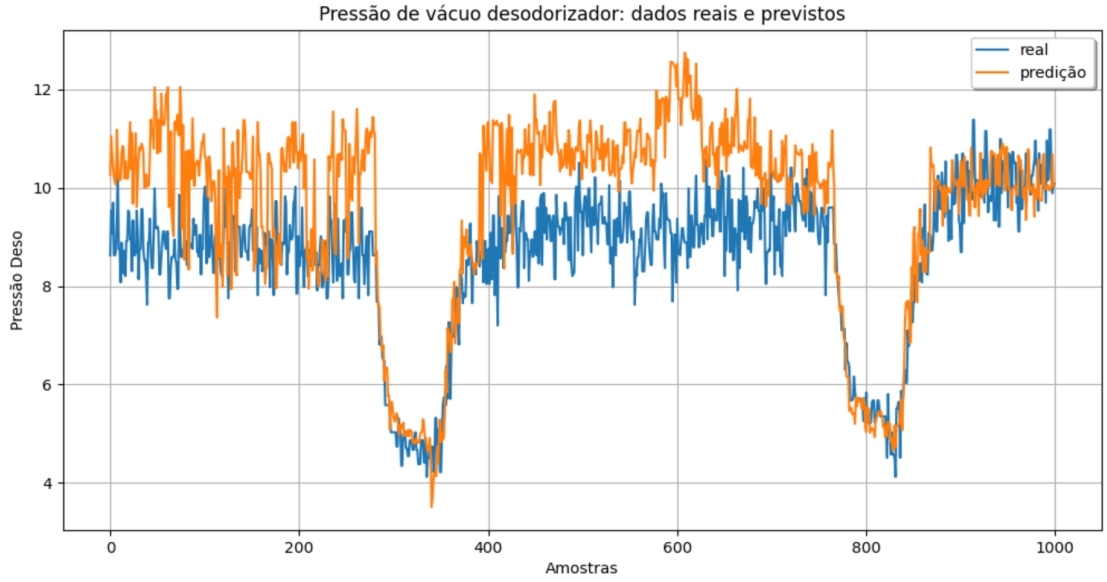


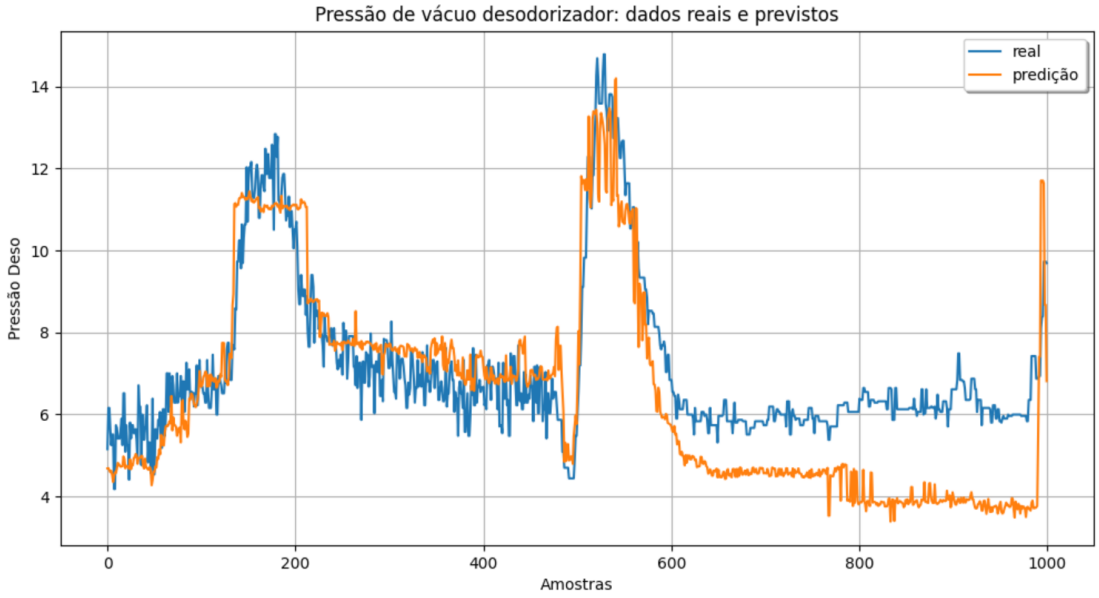
Figura 5. Intervalo mostrando índice elevado de erros com o modelo final



Na Figura 5 foi recortado um intervalo mostrando um índice de erro maior ou um maior descolamento do valor previsto pelo modelo com o valor real da variável alvo. Pela figura podemos observar que em vários momentos o modelo indicou uma quebra de vácuo, porém na verdade comparando com o valor real não houve essa quebra. Esse caso mostra que quando próximo do limite do valor da quebra de vácuo é importante o modelo performar bem, caso contrário são reportados muitos falso positivos como foi verificado.

Já a variação mostrada na Figura 6 exibe a correta previsão da quebra de vácuo pelo modelo em dois momentos distintos dentro de um mesmo intervalo. Isso mostra que a previsão é feita corretamente dentro de intervalos menores e não somente em momentos

Figura 6. Intervalo com mais de uma quebra de vácuo capturada pelo modelo final



espaçados. Apesar disso, a partir da amostra seiscentos desse intervalo, podemos verificar novamente um descasamento entre o valor real e previsto pelo modelo, seguido por uma aproximação ao valor real no final da amostra apresentada.

Considerando todas as figuras com as amostras apresentadas é possível observar que, de forma geral, o modelo prediz ou estima bem a variável alvo. Porém, conforme foi visto, em determinados momentos ainda existem pontos que podem ser melhorados.

A partir da análise dos resultados obtivemos uma taxa de erro satisfatória para o modelo, considerando as métricas avaliadas. Contudo, mesmo com os resultados obtidos, o tempo de antecipação para a quebra de vácuo é pequeno, portanto não é trivial gerar um valor para as indústrias na prevenção da quebra de vácuo. Nesse sentido, o principal desfecho da nossa avaliação é o seguinte questionamento: Até que ponto conseguimos um modelo que seja eficaz e antecipe a quebra de vácuo o máximo possível ou com a maior antecedência possível levando em conta o erro?

5. Conclusão

Neste trabalho, estudamos o conceito, aplicações e as etapas do processo de refino de óleos. Foram discutidas diversas aplicações nas quais os modelos de aprendizado de máquina são úteis na resolução de problemas. No contexto do trabalho, foi discutido o problema denominado quebra de vácuo, que ocorre quando a pressão do desodorizador assume valores maiores do que o desejado e isso acaba por prejudicar o processo de desodorização de óleos nas indústrias.

Os resultados que obtivemos embora sejam preliminares, são promissores e contribuem na solução desse problema. No futuro, planejamos estudar a dependência entre o tempo de antecipação da quebra de vácuo e o erro obtido pelos modelos, de forma a encontrar um limiar aceitável entre esses dois fatores.

6. Etapas e Cronograma

Cronograma																
Atividade	Abril				Maio				Junho				Julho			
Explorar	X	X	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Modelar	1	2	X	X	X	2	3	4	1	2	3	4	1	2	3	4
Aplicar	1	2	3	4	1	X	X	X	1	2	3	4	1	2	3	4
Avaliar	1	2	3	4	1	2	3	4	X	X	3	4	1	2	3	4
Relatório Final	1	2	3	4	1	2	3	4	1	2	X	X	X	2	3	4

Referências Bibliográficas

Alharin, A., Doan, T.-N., and Sartipi, M. (2020). Reinforcement learning interpretation methods: A survey. *IEEE Access*, 8:171058–171077.

Amanpreet Singh, Narina Thakur, A. S. (2016). A review of supervised machine learning algorithms. In *IEEE, Institute of Electrical and Electronics Engineers*.

Brahmi, F., Haddad, S., Bouamara, K., Yalaoui-Guellal, D., Prost-Camus, E., de Barros, J.-P. P., Prost, M., Atanasov, A. G., Madani, K., Boulekbache-Makhlouf, L., and Lizard, G. (2020). Comparison of chemical composition and biological activities of algerian seed oils of pistacia lentiscus l., opuntia ficus indica (l.) mill. and argania spinosa l. skeels. *Industrial Crops and Products*, 151:112456.

Cao, Q., Wu, Y., Yang, J., and Yin, J. (2023). Greenhouse temperature prediction based on time-series features and lightgbm. *Applied Sciences*, 13(3).

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Crisci, C., Ghattas, B., and Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling*, 240:113–122.

Gabriele Landucci, Gabriele Pannocchia, L. P. C. N. (2013). Analysis and simulation of an industrial vegetable oil refining process. In *Elsevier, Journal of Food Engineering*.

Gan, M., Pan, S., Chen, Y., Cheng, C., Pan, H., and Zhu, X. (2021). Application of the machine learning lightgbm model to the prediction of the water levels of the lower columbia river. *Journal of Marine Science and Engineering*, 9(5).

Gharby, S. (2022). Refining vegetable oils: Chemical and physical refining. In *Hindawi, The Scientific World Journal*.

Gupta, M. K. (2017). Chapter 8 - deodorization. In Gupta, M. K., editor, *Practical Guide to Vegetable Oil Processing (Second Edition)*, pages 217–247. AOCS Press, second edition edition.

Iqbal Muhammad, Z. Y. (2015). Supervised machine learning approaches: A survey. In *ICTACT JOURNAL ON SOFT COMPUTING*, volume 05.

John Wiley & Sons, I. (2005). In *Bailey's Industrial Oil and Fat Products*, volume 2, pages 371–378.

Kalimuthu, M., Vaishnavi, P., and Kishore, M. (2020). Crop prediction using machine learning. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 926–932.

Kamel Essid, Manef Chtourou, M. T. M. H. F. (2009). Influence of the neutralization step on the oxidative and thermal stability of acid olive oil. In *Journal of Oleo Science*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lv Cai-Xia, An Shu-Yi, Q. B.-J. and Wei, W. (2021). Time series analysis of hemorrhagic fever with renal syndrome in mainland china by using an xgboost forecasting model. *BMC Infectious Diseases*, 21.

M.R. Norazlina, M.H.A. Jahurul, M. H. A. M.-J. N. M. P. M. R. R. G. A. N. J. L. H. F. (2021). Trends in blending vegetable fats and oils for cocoa butter alternative application: A review. In *Elsevier, Trends in Food Science Technology*.

Rajko Vidrih, Sergeja Vidakovič, H. A. (2010). Biochemical parameters and oxidative resistance to thermal treatment of refined and unrefined vegetable edible oils. In *Department of Food Science and Technology, Biotechnical Faculty, University of Ljubljana*.

Shinde, P. P. and Shah, S. (2018). A review of machine learning and deep learning applications. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6.

Syed Sherazi, Sarfaraz Mahesar, S. (2016). Vegetable oil deodorizer distillate: A rich source of the natural bioactive components. In *Journal of Oleo Science*.

Tammy Jiang, Jaimie L. Gradus, A. J. R. (2020). Supervised machine learning: A brief primer. In *Elsevier Ltd, Behavior Therapy*.

Wang, J. and Biljecki, F. (2022). Unsupervised machine learning in urban studies: A systematic review of applications. *Cities*, 129:103925.