UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
UNIDADE ACADÊMICA DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

DISSERTAÇÃO

OTIMIZAÇÃO DO PROCESSO DE PRODUÇÃO DE CLORO E SODA CÁUSTICA COM APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING

PEDRO AUGUSTO SILVA DE FREITAS

Campina Grande – PB
2023

PEDRO AUGUSTO SILVA DE FREITAS

OTIMIZAÇÃO DO PROCESSO DE PRODUÇÃO DE CLORO E SODA CÁUSTICA COM APLICAÇÃO DE TÉCNICAS DE *MACHINE LEARNING*

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Química da Universidade Federal de Campina Grande, pertencente à linha de pesquisa em Modelagem e Simulação, como requisito para a obtenção do título de Mestre em Engenharia Química.

Orientador: Prof. Dr. Heleno Bispo da Silva Junior

Coorientador: Prof. Dr. Eudésio Oliveira Vilar

Campina Grande – PB
2023

F866o Freitas, Pedro Augusto Silva de.
Otimização do processo de produção de cloro e soda cáustica com aplicação de técnicas de *Machine Learning* / Pedro Augusto Silva de Freitas. - Campina Grande, 2023.
52 f. : il. color.

Dissertação (Mestrado em Engenharia Química) - Universidade Federal de Campina Grande, Centro de Ciências e Tecnologia, 2023.
"Orientação: Prof. Dr.Heleno Bispo da Silva Junior, Prof. Dr. Eudésio Oliveira Vilar."
Referências.

1. Otimização Operacional. 2. Aprendizado de Máquina. 3. Indústria Eletrointensiva. 4. Cloro e Soda Cáustica. I. Silva Junior, Heleno Bispo da. II. Vilar, Eudésio Oliveira. III. Título.

CDU 66.01(043)

FICHA CATALOGRAFICA ELABORADA PELA BIBLIOTECARIA ITAPUANA SOARES DIAS GONÇALVES CRB-15/093

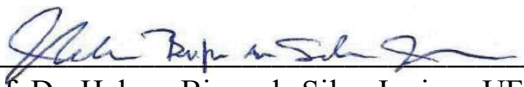
PEDRO AUGUSTO SILVA DE FREITAS

OTIMIZAÇÃO DO PROCESSO DE PRODUÇÃO DE CLORO E SODA CÁUSTICA COM APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Química da Universidade Federal de Campina Grande, pertencente à linha de pesquisa em Modelagem e Simulação, como requisito para a obtenção do título de Mestre em Engenharia Química.

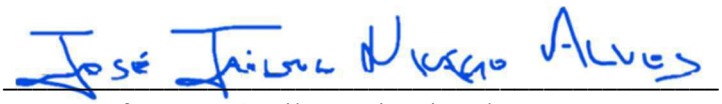
Aprovado em: 26/09/2023

BANCA EXAMINADORA:



Prof. Dr. Heleno Bispo da Silva Junior - UFCG

Prof. Dr. Eudésio Oliveira Vilar - UFCG



Prof. Dr. José Jailson Nicacio Alves - UFCG

Prof. Dr. Sidinei Kleber da Silva - UFCG

AGRADECIMENTOS

A Deus, por me dar forças e confiança em todos os momentos da minha vida, me ajudando a conquistar mais essa etapa tão importante e desejada.

Aos meus pais Francinete e Enaldo, por todo incentivo, apoio, amor e dedicação desde as primeiras palavras até a pessoa que sou hoje. Estendo esse agradecimento também a meus queridos irmãos, Paulo, Valesca e Conceição que sempre me estimularam e torceram pelo êxito desse trabalho.

A esta universidade, seu corpo docente, direção e administração que viabilizaram mais essa etapa tão importante na minha formação pessoal e profissional.

A meus orientadores, prof. Eudésio e prof. Heleno, pelo suporte, orientação, incentivos, paciência, que me ajudaram durante toda a condução e execução desse trabalho.

A empresa Braskem, pelo apoio e incentivo prestados durante todas as etapas de desenvolvimento desse estudo.

A Nayana, minha esposa, que esteve sempre ao meu lado, com conselhos e apoio, e foi a principal incentivadora para o sucesso na realização desse trabalho.

FREITAS, PEDRO AUGUSTO SILVA. **Otimização do Processo de Produção de Cloro e Soda Cáustica com Aplicação de Técnicas de *Machine Learning***. 2023. 52 p. Dissertação (Mestrado em Engenharia Química) - Universidade Federal de Campina Grande, Paraíba, 2023.

RESUMO

Na indústria de produção de cloro, soda cáustica e hidrogênio no Brasil, três tecnologias de células eletrolíticas são utilizadas comercialmente para esse fim: célula de mercúrio, diafragma e membrana. Nas células com tecnologia de diafragma, este desempenha um papel decisivo na eficiência da célula. Sua importância vai desde a eficiência energética da célula até aspectos relacionados à segurança operacional. Neste trabalho foram construídos modelos de aprendizado de máquina (ML) para predição do desempenho das células eletrolíticas, a partir dados industriais relativos as células à diafragma produzido pela UCS (Unidade de Cloro-Soda) da fábrica Braskem S/A, implantada no estado de Alagoas. Os modelos treinados apresentaram desempenho satisfatório na previsão do desempenho da célula a partir dos dados de fabricação do diafragma e de operação das células. O modelo *Random Forest* obteve desempenho superior com relação aos demais modelos, com acurácia superior a 90%, resultado importante que serve de base para a busca pela melhoria da performance do diafragma. Esse resultado confirma a viabilidade de aplicação de técnicas de aprendizado de máquinas na indústria de produção de cloro e soda cáustica, possibilitando a melhorias nos processos de produção. Outro resultado importante da modelagem desenvolvida foi a obtenção das variáveis de maior relevância para os modelos, viabilizando o controle dessas variáveis nos processos de fabricação e operação das células eletrolíticas, contribuindo para o incremento da sua performance.

Palavras-chave: Aprendizado de Máquina, Industria Eletrointensiva*, Cloro e Soda Cáustica, Otimização operacional.

FREITAS, PEDRO AUGUSTO SILVA DE. **Otimização do Processo de Produção de Cloro e Soda Cáustica com Aplicação de Técnicas de Machine Learning**. 2023. 52 p. Dissertação (Mestrado em Engenharia Química) - Universidade Federal de Campina Grande, Paraíba, 2023.

ABSTRACT

In the chlorine, caustic soda and hydrogen production industry in Brazil, three electrolytic cell technologies are used commercially for this purpose: mercury cell, diaphragm and membrane. In cells with diaphragm technology, this plays a decisive role in cell efficiency. Its importance ranges from the energy efficiency of the cell to aspects related to operational safety. In this work, machine learning (ML) models were built to predict the performance of electrolytic cells, based on industrial data relating to diaphragm cells produced by the UCS (Chlorine Soda Unit) of the Braskem S/A factory, located in the state of Alagoas. The trained models showed satisfactory performance in predicting cell performance based on diaphragm manufacturing and cell operation data. The Random Forest model achieved superior performance in relation to the other models, with an accuracy greater than 90%, an important result that serves as a basis for the search for improving the performance of the diaphragm. This result confirms the feasibility of applying machine learning techniques in the chlorine and caustic soda production industry, enabling improvements in production processes. Another important result of the modeling developed was the obtaining of the most relevant variables for the models, making it possible to control these variables in the manufacturing and operation processes of the electrolytic cells, contributing to an increase in their performance.

Keywords: Machine Learning, Electrical Intensive Industry*, Chlorine and Caustic Soda, Operational optimization.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Símbolo	Descrição	Unidade
x_i	Atributo	-
i_G	índice de Gini	-
p	Proporção	%
C_j	Classe	-
P	Probabilidade	%
Abreviações		
ML	Machine Learning	
RF	Modelo Random Forest	
KNN	Modelo K-Nearest Neighbours	
SVM	Modelo Máquina de vetores de suporte	
NE	Quantidade de Árvores	
TP	Verdadeiros Positivos	
TN	Verdadeiros Negativos	
FP	Falso Positivos	
FN	Falso Negativos	
TD	Profundidade Máxima das Árvores	
GPU	Unidade de Processamento Gráfico	

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

A grande maioria das industriais químicas utilizam como insumos a soda cáustica e o cloro (e derivados) durante seus processos produtivos. Segundo Lopez (2018) a demanda mundial de soda cáustica é de aproximadamente 75 milhões de toneladas por ano, destinada principalmente para indústrias de produção de alumina, papel e celulose e o setor têxtil. Já para o cloro, a demanda anual é estimada em 70 milhões de toneladas, utilizadas na sua maioria nas indústrias de produção de plástico, tratamento de água e produção de óxido de propileno.

Na indústria de produção de cloro, soda cáustica e hidrogênio, três principais tecnologias são comercialmente utilizadas para esse fim: Célula eletrolítica à Mercúrio, diafragma e membrana. Independente da tecnologia utilizada, uma solução saturada em cloreto de sódio é alimentada no eletrolisador, que juntamente com energia elétrica, proporcionam as reações de oxidação e redução no anodo e catodo, respectivamente.

O presente trabalho foi realizado a partir de dados relativos à tecnologia de células eletrolíticas à diafragma, sendo esta tecnologia então o foco no desenvolvimento e discussões. A tecnologia de produção à diafragma caracteriza-se pela utilização de diafragma (separador) poroso no interior da célula, responsável por garantir a segurança operacional e eficiência de produção.

Estudos realizados por Hine et al. (1976) avaliaram o transporte de massa em células com diafragma composto de amianto e polímero, bem como a resistência desse tipo de material para aplicação nas células eletrolíticas. Modelos matemáticos de estado estacionário simplificado propostos por Van Zee (1986) foram utilizados para estimar a zona neutra de pH no interior do diafragma a partir de variáveis associadas ao separador, com o objetivo de prever a difusão dos íons hidroxilas e hidrogênio pelo diafragma, bem como a concentração de soda cáustica produzida pela célula.

Modelos de redes neurais elaborados por Júnior (2006) foram treinados para simular o processo de uma célula à diafragma com objetivo de elevação no ganho financeiro da produção. Os modelos foram construídos a partir de variáveis de operação e construção do diafragma. Os resultados mostraram que a modelagem da célula eletrolítica não apresentou consistência nos resultados, indicando falta de ajuste da rede neural, conforme mencionado nas conclusões do trabalho.

Filho et al (2011) propôs um modelo estatístico para estimar a eficiência de corrente e concentração de soda cáustica produzida em uma célula eletrolítica com diafragma constituído de amianto e polímero ramificado SM-2®, a partir de dados referentes a deposição do diafragma,

como o tipo de fibra utilizada e as concentrações de NaCl e NaOH no banho de deposição. Os modelos apresentaram boa aplicabilidade nas condições operacionais da planta.

Todos esses trabalhos citados não levaram em consideração o impacto das variáveis operacionais aliadas as propriedades dos diafragmas no desempenho das células eletrolíticas, ou não obtiveram sucesso ao associar os dois conjuntos de variáveis.

1.2 Objetivos e Organização da Dissertação

Construção e aplicação de modelos *de machine learning* de classificação para estimar o desempenho de células eletrolíticas de produção de cloro e soda cáustica, na busca da otimização do processo de produção, pelo fato de não se ter conhecimento desse tipo de modelos de classificação aplicados ao processo industrial em questão. Com isso, este trabalho teve como objetivo principal a **otimização do processo de produção de cloro e soda cáustica via tecnologia à diafragma com o uso de técnicas de *machine learning*.** Para obtenção do objetivo principal, alguns itens foram considerados e compuseram os **objetivos específicos** do trabalho:

1) Construção de modelos de classificação binária e determinação do modelo com melhor desempenho (acurácia/precisão).
2) Otimização do modelo de classificação que obteve o melhor desempenho.
3) Determinar quais variáveis são mais relevantes para o desempenho dos modelos.

No Capítulo 2 são apresentados os principais trabalhos desenvolvidos para otimização das células com tecnologia á diafragma e a metodologia utilizada. É abordado também uma revisão das principais técnicas de *machine learning*, com os tipos de modelos de classificação que foram construídos e utilizado no trabalho.

O Capítulo 3 apresenta a metodologia aplicada na modelagem utilizada para construção dos modelos de classificação, incluindo as etapas para o desenvolvimento dos modelos. É exposto também as variáveis utilizadas para o desenvolvimento dos modelos, além da base de dados utilizada.

Já no Capítulo 4 é apresentado os principais resultados do trabalho com base nos objetivos específicos mencionados anteriormente. Os resultados são discutidos e avaliados no intuito de atender aos objetivos citados nesse trabalho.

Por fim, o Capítulo 5 traz as principais conclusões obtidas.

CAPÍTULO 2

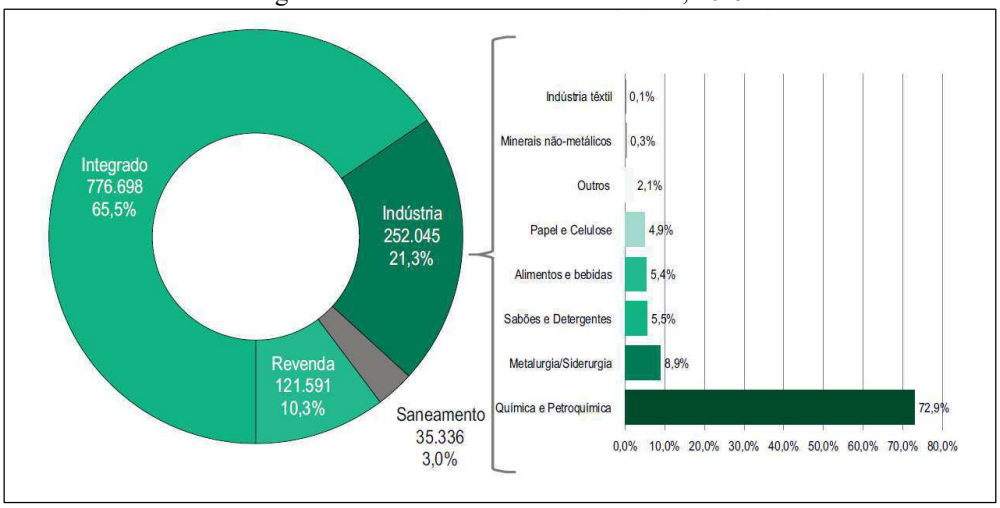
REVISÃO DE LITERATURA

2.1 INDÚSTRIA DE CLORO E SODA CÁUSTICA

A indústria de produção de cloro e soda cáustica está presente em diversos setores da economia, sendo que os produtos derivados desse setor são essencialmente indispensáveis para utilização da sociedade, como em tratamento de águas de abastecimento e esgoto doméstico. Esses insumos são da mesma forma necessários para a obtenção de produtos e a prestação de serviços que proporcionam a melhor qualidade de vida, incluindo a saúde, habitação, transportes, alimentação, vestuário e lazer (Abiclor, 2020).

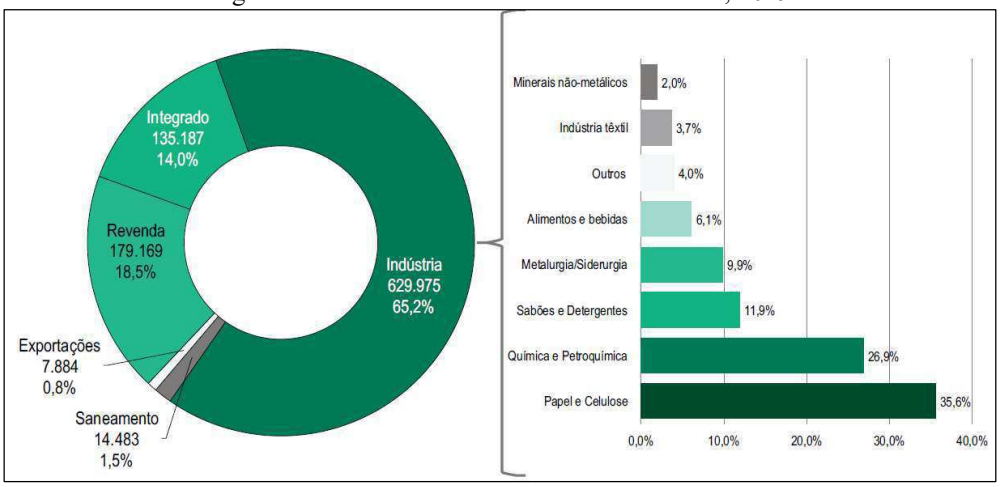
Dentre os diversos setores da economia que utilizam o cloro, soda cáustica e derivados, podemos destacar as indústrias química e petroquímica, metalurgia e siderurgia, sabões e detergentes, alimentos e bebidas e papel e celulose, conforme ilustrado nas figuras 2.1 e 2.2.

Figura 2.1: Demanda de Cloro no Brasil, 2019



Fonte: Abiclor, 2020

Figura 2.2: Demanda de Soda cáustica no Brasil, 2019



Fonte: Abiclor, 2020

Conforme Lopez (2018, a américa latina possui a segunda maior capacidade de produção de cloro do mundo (aproximadamente 19%), o que corresponde a 18,5 milhões de toneladas por ano. Dessa capacidade, 44,1% da produção é realizada via diafragma e 39,8% via membrana (Abiclor, 2020).

No Brasil, a tecnologia à diafragma é responsável por aproximadamente 63% da capacidade instalada de produção de cloro (Abiclor, 2020).

Tabela 2.1: Capacidade instalada de produção de cloro por diferentes tecnologias no Brasil, 2020

Mercúrio	Diafragma	Membrana
14%	63%	23%

Fonte: Abiclor, 2020

A planta da Braskem S/A, no estado de Alagoas, possui a segunda maior capacidade instalada de produção de cloro no Brasil, atrás apenas da empresa Dow (Abiclor, 2020).

Tabela 2.2: Capacidade instalada de produção, em mil toneladas de cloro, no ano de 2019

Empresa	Cidade	Capacidade Instalada
Dow Brasil	Aratu/BA	415,0
Braskem	Maceió/AL	409,4
Unipar Carbocloro	Cubatão/SP	355,0

Fonte: Abiclor, 2020

2.2 TECNOLOGIA À DIAFRAGMA

As primeiras células eletrolíticas com tecnologia à diafragma foram utilizadas comercialmente a partir de 1888 pela Griesheim Company, para a produção de cloro e hidróxido de potássio. Após isso, diversas outras tecnologias de células á diafragma foram desenvolvidas, como a célula *Hargreaves-Bird* em 1890 e a célula *Le Sueur*, que utilizava um diafragma de cimento poroso (Braga, 2009).

Historicamente, o amianto foi o principal material utilizado na construção de diafragmas em células eletroquímicas (Junior et al, 2021). Inicialmente, o amianto era empregado como diafragma em forma de papel ou filtros. Em 1928, Stuart desenvolveu o diafragma de amianto depositado, gerando ampla flexibilidade no design das células, possibilitando a deposição dos diafragmas em catodos de diversos formatos, procedimento que se tronou padrão na indústria (McIntyre, 2002). Esse tipo de diafragma convencional foi empregado até a década de 1970, quando o diafragma modificado com polímero de fluorocarbono foi desenvolvido (Schulz et al, 1989). O diafragma modificado com polímero apresentou expressivo aumento na resistência mecânica e integridade estrutural, além da redução no consumo de energia, se comparado com os diafragmas preparados utilizando as

técnicas convencionais (Schulz et al, 1989). O polímero atua estabilizando as fibras naturais do amianto, reduzindo sua dissolução antecipada (Cunha, 2015).

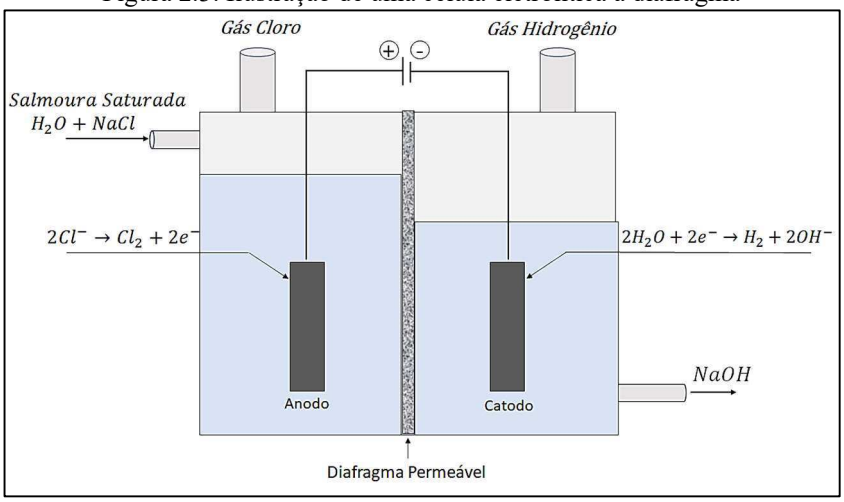
Diversas características foram importantes para a escolha do amianto como constituinte principal para os diafragmas, tais como a estabilidade química, sua propriedade de troca iônica, baixo custo e abundância natural (Andrade, 2006).

Conforme Cunha (2015), as células com tecnologia à diafragma são constituídas por um catodo perfurado de aço ou ferro e um anodo de titânio recoberto de platina ou óxido de platina. Fazendo a separação entre as regiões do catodo e anodo, existe um diafragma poroso de fibras de asbesto (amianto), misturado com outras fibras, como as de politetrafluoretileno (PTFE).

Na produção de cloro e soda cáustica a partir da tecnologia à diafragma, este tem papel decisivo na eficiência de produção. Sua importância vai desde a eficiência energética da célula até aspectos relativos à segurança operacional.

O diafragma atua mantendo a separação entre os dois compartimentos principais que compõem a célula eletrolítica, os compartimentos anódico e catódico. Esta separação assegura que a mistura entre os gases cloro (produzido no compartimento anódico) e o hidrogênio (produzido no compartimento catódico) seja a mínima possível dentro de critérios operacionais evitando que determinadas concentrações gerem uma mistura explosiva. O diafragma é responsável também por garantir a redução da formação dos subprodutos como o hipoclorito de sódio e clorato de sódio os quais se presentes na soda produzida, diminui sua qualidade e seu custo comercial.

Figura 2.3: Ilustração de uma célula eletrolítica à diafragma



Fonte: elaborado pelo autor (2023)

Atualmente os diafragmas têm base polimérica, ou seja, sem a presença de amianto na sua composição. Pode-se citar por exemplo, o Polyramix®, da DeNora, e o Tephram®, da Westlake, comercializados para utilização nas células de produção de cloro e soda cáustica.

A formação dos diafragmas citados, sejam eles a base de amianto ou não, ocorre a partir da mistura com uma composição pré-definida das fibras que o formam. Essa mistura ou dispersão é depositada sobre o catodo da célula eletrolítica formando o diafragma que irá separar os compartimentos anólito e católito (Figura 2.3). Nesse processo, diversas variáveis são acompanhadas para garantir a qualidade adequada do diafragma.

O processo de obtenção do diafragma a base de amianto é uma atividade essencialmente semiempírica, sendo considerado um trabalho quase que inerentemente artesanal, onde a suspensão de fibras de amianto é misturada em uma solução contendo NaCl e NaOH, sendo depositadas sob vácuo em um catodo metálico, Filho et al (2011).

2.3 MACHINE LEARNING

Com a crescente quantidade de dados coletados e armazenados em indústrias e diversos outros setores da engenharia, se faz necessário cada vez mais a utilização de técnicas adequadas para o processamento e obtenção de modelagens matemáticas com satisfatória acurácia. Esses modelos são importantes para o maior entendimento e implementação de otimizações nos processos industriais. Conforme Çinar et al (2020), o surgimento da indústria 4.0 gerou uma demanda cada vez maior por sistema inteligentes e autônomos, como o aprendizado de máquinas, com intuito de avaliar as condições de operação de equipamentos e sistemas, contribuindo para o diagnóstico e detecção de falhas. As aplicações de aprendizado de máquina na indústria proporcionam uma série de vantagens, como redução nos custos de manutenção, aumento do tempo de vida útil de sobressalentes, redução de estoque e aumento de produção. Para Sugahara (2020), o uso de técnicas de *Machine Learning* para organizar e obter informações de grandes volumes de dados proporcionam integração dos sistemas de produção, com acesso a dados de produção em tempo real, planejamento de manutenções preventivas e controle logístico. Segundo Ge (2017) a aprendizagem autônoma de máquinas possui função crucial no treinamento de modelos baseados em dados de processos industriais, sendo possível obter informações essenciais, com a identificação de padrões e facilidade na previsão de resultados para novos dados, promovendo maior rapidez nas tomadas de decisões.

O significado de *machine learning* (ML) pode ser traduzida de forma geral como sendo “aprender”, ou seja, o aprendizado de máquina. De forma mais específica, a denominação de

ML é indicar o aprendizado autônomo das máquinas com base em informações consolidadas, resultado em conclusões o mais precisas possíveis (Silva, 2021). *ML* é um campo que envolve as ciências cibernética e de computação, que atualmente atrai a atenção de profissionais da área e o público em geral, impulsionado principalmente pelo desenvolvimento e capacidade computacional cada vez mais acessível (Fradkov, 2020).

As técnicas de abordagem de ML são classificadas em *supervisionada, não-supervisionada e reforço*, sendo as duas primeiras as mais utilizadas. Na abordagem *supervisionada*, os modelos de regressão ou classificação são construídos a partir de dados de entrada e saída, com base no treinamento interativo, sendo capazes de predizer atributos não vistos no processo de treinamento do modelo (Silva, 2021).

Já na modelagem *não-supervisionada*, não há o conhecimento da relação entre os dados de entrada e a saída correspondente. Os modelos são construídos com base na relação entre os dados de entrada utilizados (Silva, 2021).

Diante das diversas técnicas de *machine learning supervisionada*, as principais, e que foram abordadas nesse trabalho, são:

- Árvore de decisão;
- Random Forest (RF);
- Naive Bayes;
- K-Nearest Neighbours (KNN);
- Máquina de vetores de suporte (SVM).

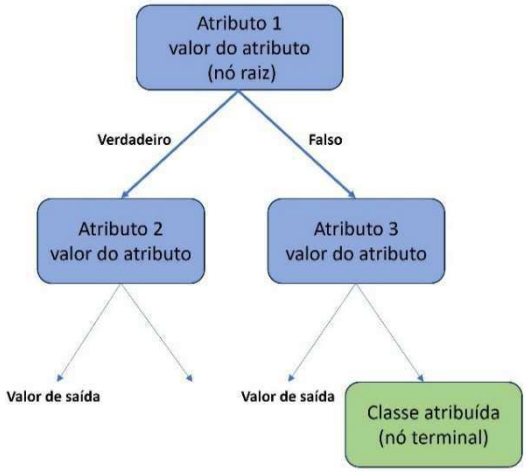
Nesse trabalho foi priorizado a obtenção, otimização e utilização dos modelos de **árvore de decisão** e **Random Forest**, por se tratar de modelos que são mais facilmente interpretáveis e aplicáveis à realidade do processo industrial em questão. Os modelos de Naive Bayes, KNN e SVM também foram treinados a fim de comparação de desempenho com os demais modelos.

2.3.1 Árvore de Decisão

Os modelos de árvore de decisão utilizam uma estrutura semelhante a uma árvore, onde cada ramo do modelo é dividido de forma a minimizar o erro de previsão (Shimizu, 2021). As árvores de decisão são compostas dos nós chamados de *raiz*, *terminal* e *interno*, sendo que este último é particionado em dois segmentos (árvore binária) com o uso de regra do tipo *Se-Então*, a partir de um atributo x_i e um valor de corte. A direção à esquerda do nó interno representa a condição verdadeira com base no valor de corte definido para cada nó (Júnior, 2018). Os nós que não são divididos posteriormente na árvore são chamados de nós *finais* ou *terminais* (Shimizu, 2021).

Nos nós internos de uma árvore de decisão, o espaço é dividido em dois (partição binária) ou mais subespaços, a partir dos valores de entrada do modelo. Em cada nó terminal, é atribuída uma determinada classe, com a respectiva probabilidade de ocorrer. As instâncias são classificadas percorrendo o caminho desde o nó raiz até os nós terminais, avaliando a decisão entre cada nó interno (Nasteski, 2018).

Figura 2.4: Ilustração de uma árvore de decisão



Fonte: elaborado pelo autor (2023)

2.3.1.1 Métrica de Impureza

Na divisão realizada em cada nó interno, durante o treinamento de uma árvore de decisão, é possível determinar o grau de impureza gerado em cada divisão, sendo que a pureza do nó está diretamente relacionada com a melhor capacidade de predição (Júnior, 2018). A impureza baseada no *índice de Gini*, i_G, é uma medida muito utilizada nos problemas de treinamento de arvores de decisão.

O i_G estima a probabilidade de um atributo, escolhido aleatoriamente, ser classificado incorretamente para determinada classe (Louppe, 2014).

Considerando o treinamento de árvores de decisão de classificação binária, o índice de Gini pode ser representado pela seguinte equação (Júnior, 2018):

$$i_G = 2p(1-p) \tag{2.1}$$

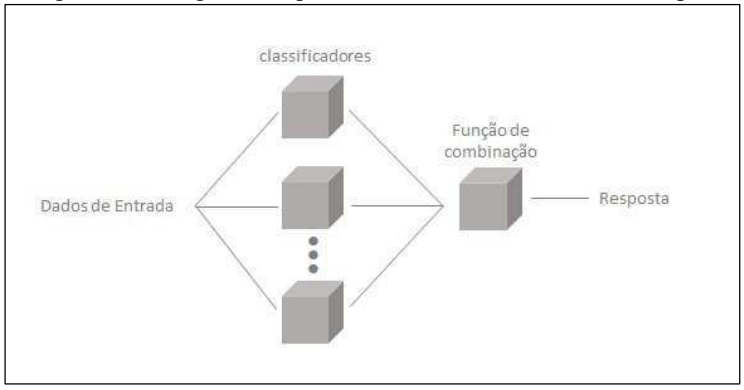
Onde p representa a proporção de uma das classes. O *índice de Gini* apresenta valores entre 0 (nó com maior pureza) e 1.

2.3.2 Random Forest

Diante das diversas técnicas de *machine learning* supervisionada, a *Random Forest* (RF) caracteriza-se pela aplicabilidade e acessível utilização, não sendo necessário a normalização de dados. A técnica possibilita ainda a determinação das variáveis de maior relevância na predição, a partir de um processo de votação durante o processo de treinamento, o que proporciona a obtenção de modelos mais precisos e com interpretação simplificada (Silva, 2019).

Na forma tradicional de obtenção de modelos de classificação, apenas um modelo classificador é utilizado para realizar a predição dos dados. Contudo, a acurácia do modelo pode ser aumentada com o uso de um conjunto de classificadores, cada um destes auxiliando na melhor tomada de decisão dos outros classificadores, em um processo de votação. As estratégias mais utilizadas são *bagging* e *boosting* (Briem, 2002).

Figura 2.5: Diagrama esquemático de um classificador múltiplo



Fonte: Adaptado de Briem (2002)

Na estratégia *Bagging*, vários preditores são construídos com base em conjuntos de dados gerados com o uso da técnica *bootstrap*, produzindo assim um preditor agregado (Breiman, 1996). A média para essa agregação é calculada, proporcionado redução do erro do modelo na fase de treinamento. No método *bootstrap*, são construídos randomicamente conjunto de dados a partir de dados de treinamento, no qual cada conjunto tem a mesma dimensão do conjunto de dados originais (Hastie, 2001).

Avaliações realizadas com árvores de classificação e regressão em conjuntos de dados reais e simulados demostraram a eficácia da técnica *Bagging* para aumentar a precisão dos modelos, com base na modificação do conjunto de dados (Breiman, 1996). De forma geral (Panov et al., 2007), a associação de um conjunto de classificadores proporciona melhor performance se comparada ao desempenho individual de cada classificadores, principalmente em modelos de aprendizagem instáveis, como árvores de decisão.

O método *Boosting* também é utilizado para aumentar a precisão de modelos de regressão ou classificação. Nessa técnica, todas as amostras iniciais possuem o mesmo peso, formando assim uma distribuição uniforme. Após a primeira rodada de treinamento do modelo, os pesos de cada amostras são redistribuídos, de forma que as classificadas incorretamente são priorizadas, com aumento do seu respectivo peso. O método então segue focando na melhoria da precisão do modelo priorizando as amostras classificadas erroneamente (Briem, 2002).

2.3.2.1 Hiperparâmetros

Nos problemas de *machine learning*, os modelos são construídos a partir do treinamento em uma base de dados onde parâmetros do modelo são definidos. Nesse tipo de modelagem, outros parâmetros podem ser pré-definidos antes do processo de treinamento, são os chamados *Hiperparâmetros* (Bishop, 2006). Com isso, há a possibilidade de determinar os valores ótimos dos *hiperparâmetros* dos modelos para obtenção do melhor desempenho de predição (Pellicer, 2020).

Considerando, por exemplo, o *hiperparâmetro* para os modelos de *Random Forest* chamado de número de preditores ou árvores (*NE*). Para o *NE*, em geral, quanto maior é seu valor, melhor a acurácia de predição do modelo. No entanto, esse incremento na quantidade de árvores é vantajoso, mas possui limite, pelo fato de a maior quantidade de árvores exige maior recurso computacional, além do fato de que o aumento na performance do modelo estabiliza a partir de determinada quantidade do *NE* (Júnior, 2018).

2.3.3 Naive Bayes

Esse modelo de classificação utiliza uma suposição que nem sempre é realidade em problemas de aprendizado de máquina, a de que há inexistência de correlação entre as variáveis de entrada do modelo. Embora essa suposição não seja sempre válida na prática, em problemas envolvendo classificação de dados categóricos, o método demonstrou rendimento satisfatório (Domingos e Pazzani, 1997).

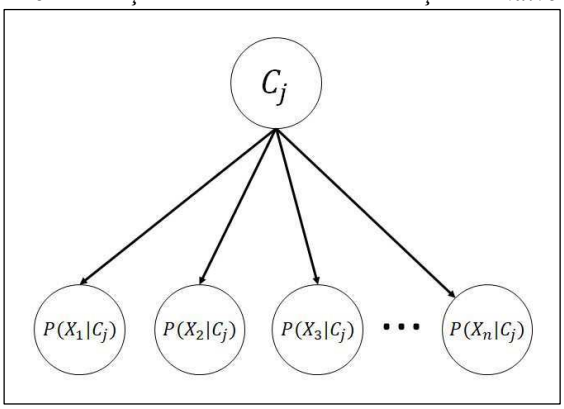
Na construção de modelos de Naive Bayes de classificação, as probabilidades para cada classe são calculadas e utilizadas na predição de novas observações (Taheri, 2013). Conforme Islam (2007), uma das grandes vantagens do modelo de Naive Bayes é que eles são autocorretivos, gerando resultados diferentes a medida em que os dados utilizados para treinar o modelo variam.

Supondo que $\boldsymbol{C_j}$ seja uma determinada classe e $\boldsymbol{X}$ uma observação aleatória, considerando que todas as variáveis de entrada são independentes, a determinação da classe

mais provável dada uma observação de teste $\boldsymbol{X_n}$ é calculada conforme equação 2.2 (Taheri, 2013):

$$P(C_j|\boldsymbol{X}) = \frac{P(C_j)\prod_{i=1}^{n} P(X_i|C)}{P(\boldsymbol{X})} \tag{2.2}$$

Figura 2.6: Ilustração de modelo de classificação de *Naive Bayes*



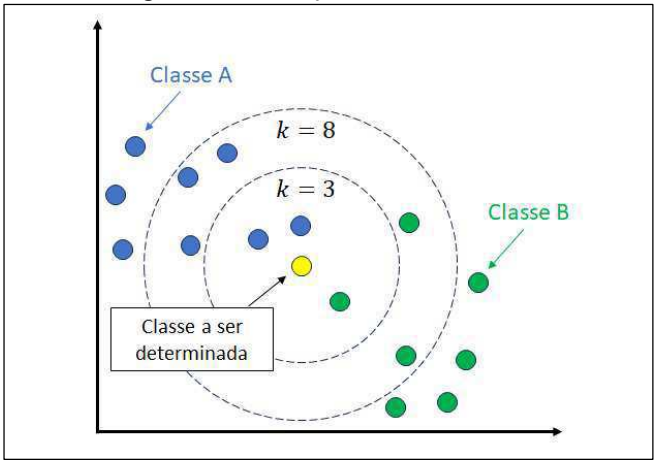
Fonte: Taheri, 2013

2.3.4 K-Nearest Neighbours (KNN)

Em problemas de classificação utilizando esse algoritmo, os atributos são classificados com base na classe dos atributos mais próximos, sendo necessário as vezes considerar mais de um atributo próximo, devido isso, essa técnica é chamada de classificação *k-vizinho mais próximo* (Cunningham e Delany, 2021).

A proximidade é a medida de quão similar as amostras são, em que, quanto menor a distância calculada, mais similar serão as duas amostras, sempre comparadas aos pares. As distâncias entre as classes que se tem conhecimento e os atributos de teste são medidas e o ponto de menor distância é chamado de k-vizinho mais próximo (Dhanabal, 2011).

Figura 2.7: Ilustração do Método KNN

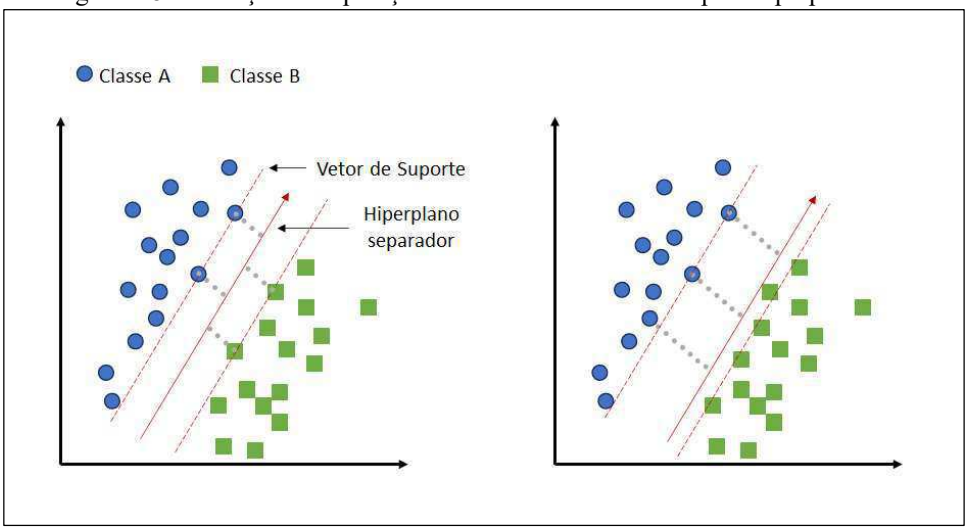


Fonte: elaborado pelo autor (2023)

2.3.5 Máquina de vetores de suporte (SVM)

O aprendizado supervisionado usando o algoritmo SVM é tradicionalmente utilizado para problemas de classificação, podendo também ser utilizado para problemas de regressão. Nesse método, é buscado um hiperplano separador (geralmente um traçado linear) que maximize a distância (margem) entre os vetores de suporte das classes que estão sendo utilizadas (Pisner e Schnyer, 2020). A distância ótima é aquela em que a separação das duas classes, realizada pelo hiperplano separador, é equidistante. Na figura 2.8, a ilustração da esquerda representa uma separação ótima.

Figura 2.8: Ilustração da separação de duas classes de dados pelo hiperplano



Fonte: elaborado pelo autor (2023)

2.3.6 Métricas de performance

A função das métricas de avaliação é determinar a assertividade do modelo de classificação a partir da resposta gerada na modelagem e o rótulo verdadeiro das classes utilizadas (Júnior, 2018). A determinação da performance de modelos de *machine learning* é realizada pelo comparativo entre as previsões realizadas pelo modelo treinado e a resposta de interesse (Santos, 2018). Conforme Kuhn e Jonhson (2013), problemas envolvendo modelos de classificação decorrem de dois tipos de predição: A contínua, que envolve probabilidade de classes, e a categórica, que se refere ao rótulo de cada observação.

A métrica principal utilizada nesse trabalho para avaliar o desempenho dos modelos foi a acurácia, que indica o acerto médio geral do preditor para as classes de um modelo de classificação ou regressão (Júnior, 2018). Ao utilizar a acurácia como métrica de avaliação, um cuidado adicional deve ser considerado, pois em problemas com desbalanceamento de

classes (que é discutido no item 3.1.2) a acurácia não fornece informação adequada sobre a capacidade de predição dos classificadores (De Castro, 2011).

Além da acurácia, também foi utilizado a métrica de *f1-score* para mensurar a performance de cada modelo construído. Como mencionado a seguir, o *f1-score* é obtido a partir de outras duas métricas. Tanto a acurácia quanto o *f1-score* utilizam os termos destacados a seguir (Silva, 2021):

True Positive (TP) – quantidade de predições da classe positiva corretamente previstos.
True Negative (TN) – quantidade de predições da classe negativa corretamente previstos.
False Positive (FP) – quantidade de predições da classe positiva erradamente previstos.
False Negative (FN) – quantidade de predições da classe negativa erradamente previstos.

O cálculo do valor da acurácia é obtido conforme descrito na equação (2.3):

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

A métrica *f1-score*, que possui valores entre 0 e 1, é obtida a partir da média harmônica de duas outras métricas, a sensibilidade e a especificidade.

$$\text{Sensibilidade} = \frac{TP}{TP + FP} \tag{2.4}$$

$$\text{Especificidade} = \frac{TP}{TP + FN} \tag{2.5}$$

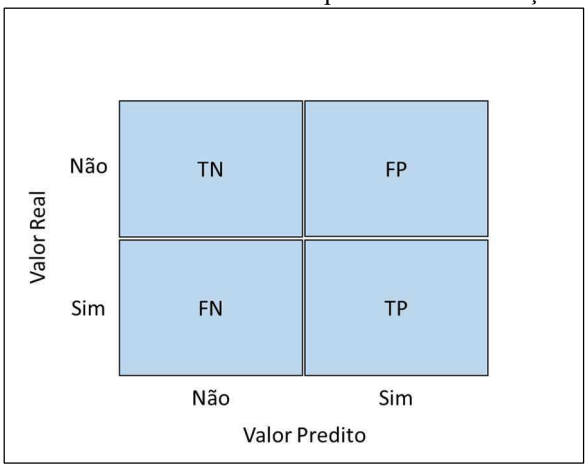
Considerando a *sensibilidade* como sendo a proporção das classificações corretas que o modelo fez da classe positiva e considerando a *especificidade* como a medida da proporção de falso negativos previstos pelo modelo, o valor do *f1-score* pode ser obtido conforme equação (2.6) (Silva, 2021):

$$f1 - score = 2 * \frac{sensibilidade * especificidade}{sensibilidade + especificidade} \tag{2.6}$$

Os termos *TP*, *TN*, *FP* e *FN* também são utilizados para construção das chamadas *matrizes de confusão*, que corresponde a uma tabela com os termos mencionados, onde são geradas e utilizadas para avaliar o desempenho dos modelos, a partir de um conjunto binário de dados (Navin e Pankaja, 2016). A avaliação da performance de modelos de classificação em problemas de *machine learning* é realizada, em sua grande maioria, com uso de métricas

que tem a matriz de confusão como base (Junior et al., 2022). A matriz de confusão ou matriz de erros, possui versatilidade e pode ser utilizada para uma série de técnicas estatísticas descritivas e analíticas, sendo a precisão geral a estatística mais simples, determinada para os classificadores, obtida pela divisão entre o total de previsões corretas e o somatório de todas as observações (Congalton,1991).

Figura 2.9: Ilustração de uma matriz de confusão para uma classificação binária "Sim"/"Não"



Fonte: elaborado pelo autor (2023)

A matriz de confusão é eficaz para avaliar a distinção dos erros (ou acertos) cometidos pelo modelo de classificação para cada classe. A diagonal principal da matriz (representado pelos valores de *TN* e *TP*) representa todas as decisões corretas do modelo, enquanto os termos não pertencentes a esse conjunto indicam os erros cometidos, expresso pelos valores de *FP* e *FN* (De Castro, 2011). Embora matrizes de confusão normalizadas, com números percentuais, sejam úteis para diversas avaliações, o uso das matrizes com valores absolutos das quantidades de observações para cada classes é importante no caso de conjunto de dados desequilibrados (Heydarian et al., 2022).

CAPÍTULO 3

METODOLOGIA

3.1 Modelagem

Devido ao processo industrial específico e tipo de tecnologia envolvida na qual esse trabalhou foi desenvolvido, a construção de modelos de classificação supervisionado foi a metodologia enxergada como a mais adequada para a busca e implementação de otimizações nesse seguimento.

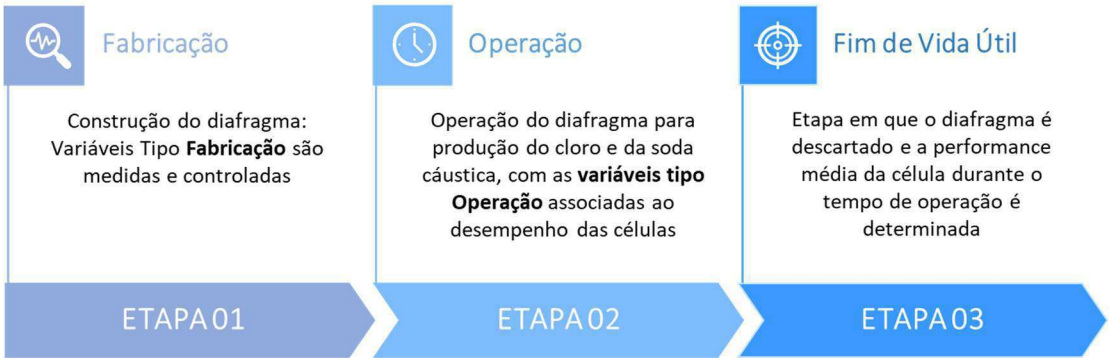
Para a construção e treinamento dos modelos de classificação, algumas etapas foram necessárias para garantir a obtenção de preditores com desempenho satisfatório. A etapa inicial foi o levantamento, tratamento e consolidação dos dados industriais relacionados a fabricação e operação de célula eletroquímicas com a tecnologia de diafragmas de amianto. Mesmo sendo uma tecnologia que entrou em desuso na planta industrial em que os dados foram coletados, a motivação principal foi a aplicação da metodologia de *machine learning* no processo industrial em estudo.

Após a consolidação da base de dados, etapa primordial para garantir uma modelagem adequada, os modelos foram treinados e suas respectivas performances foram avaliadas a partir do conjunto de dados de teste, com uso de métricas adequadas para o tipo de modelagem desenvolvida.

3.1.1 Descrição das Variáveis

Os dados utilizados nesse trabalho para treinamento dos modelos de classificação são referentes as variáveis associadas a duas etapas que envolvem os diafragmas utilizados em células eletrolíticas de produção de cloro e soda cáustica: variáveis de **Fabricação** e variáveis de **Operação**. A Figura 3.1 ilustra as etapas envolvidas durante a fabricação e o fim de vida útil dos diafragmas de Amianto.

Figura 3.1: Etapas e variáveis associadas à operação do diafragma de Amianto



Fonte: elaborado pelo autor (2023)

As variáveis chamadas de **fabricação** estão relacionadas ao processo de deposição do diafragma de amianto modificado que foi utilizado nas células eletrolíticas de produção de cloro e soda cáustica. Esse processo se caracteriza pela inserção das fibras de amianto na tela metálica do catodo, sob a ação de vácuo. Diversas variáveis estão associadas a essa etapa, sendo medidas e controladas, para garantia de uma deposição adequada do diafragma de amianto. Com a finalização do processo de fabricação do diafragma, este é montado em uma célula eletrolítica que é posta em operação. Nessa etapa as variáveis descritas como de **operação** são coletadas e o desempenho das células é mensurado. Após o tempo de vida útil das células (regida na maioria das vezes pelo desgaste e necessidade de substituição dos diafragmas), esta é substituída e o desempenho médio de operação da célula é determinado. Nessa etapa, foi unificado os dados de fabricação, operação e desempenho finais para cada célula, sendo possível obter o conjunto consolidado dos dados de treinamento e teste.

A Tabela 3.1 mostra a descrição das variáveis envolvidas nos processos de fabricação e operação do diafragma e empregadas na modelagem, ou seja, as *features* utilizadas na construção dos modelos de classificação.

Tabela 3.1: Variáveis (*features*) utilizadas para construção dos modelos

Tipo da variável	Descrição da variável	Unidade
Fabricação	Amianto	g/L
Fabricação	Massa SM-2	Kg
Fabricação	NaCl	g/L
Fabricação	NaOH	g/L
Fabricação	Tempo de Deposição	minutos
Fabricação	Vácuo máximo Deposição	mmHg
Fabricação	Vácuo máximo Secagem	mmHg
Fabricação	Viscosidade	cP
Operação	Carga	kA
Operação	pH	-
Operação	Temperatura	°C
Operação	NaCl Licor	g/L
Operação	NaCl Salmoura	g/L
Operação	Dureza	ppm
Operação	Ferro	ppm

Fonte: elaborado pelo autor (2023)

Para cada conjunto de variáveis de fabricação e operação, o desempenho da célula foi determinado com base nos parâmetros operacionais descritos na tabela 3.2. Vale ressaltar que as variáveis descritas na tabela 3.2 não foram utilizadas como *features* para a construção dos modelos, mas sim para determinar a classe das células.

A performance das células foi categoriza em classes: **classe 0** e **classe 1**, sendo a primeira, a classe de células com pior desempenho, e a última, a classe com maior rendimento. Ou seja, o interesse maior está associado as células tipo classe 1, que obtiveram maior desempenho.

A classificação de desempenho das células foi realizada com base em 4 variáveis associadas à sua operação, descritas na tabela 3.2. Se os quatro parâmetros fossem alcançados durante o tempo de operação da célula, esta foi classificada como sendo ótima (classe 1), caso contrário, a célula foi classificada como não ótima (classe 0). Os três últimos parâmetros da tabela 3.2 foram calculados a partir da média dos resultados durante o tempo de operação da célula.

Tabela 3.2: Descrição das variáveis envolvidas na determinação da classe da célula

Parâmetro	Faixa Ótima
DOL – Tempo de operação da Célula (dias)	>=250
Concentração de Hidrogênio no Gás Cloro	<=0,3*
Concentração de Soda (g/L)	>=135*
Fator R (Relação entre Clorato de Sódio e NaOH no produto)	<=4,0*

Fonte: elaborado pelo autor (2023)

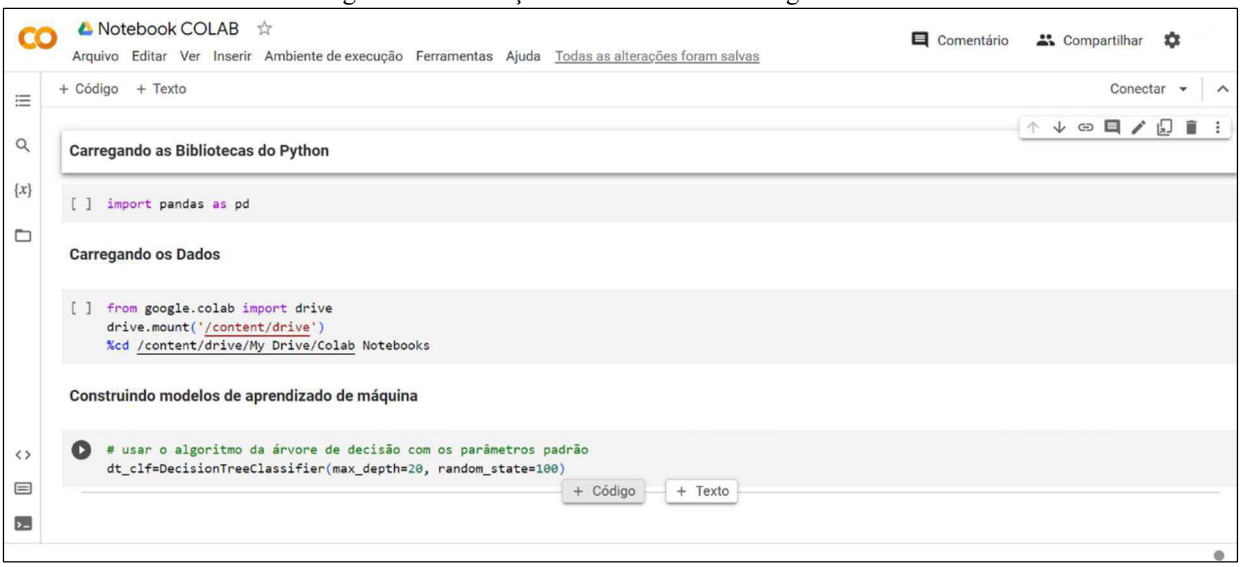
3.1.2 Construção dos Modelos

Os modelos de *machine learning* de classificação utilizados nesse trabalho foram obtidos com o uso da plataforma *Google Colaboratory (Colab)*, um serviço em nuvem que utiliza a linguagem *Python* como código fonte. Essa plataforma é um produto do *Google Research* que permite a construção de códigos em Python pelo navegador, mais tecnicamente, é um serviço de notebooks hospedados do *Jupyter* que não requer nenhuma configuração para uso e oferece acesso, sem custo financeiro, à recursos de computação como *GPU's*, sendo muito utilizado em problemas de *machine learning* (Google, 2023). O *Jupyter* Notebook é uma aplicação da web de código aberto que permite a criação e compartilhamento de documentos que contêm código ativo, equações, visualizações e texto narrativo. Dentre os diversos usos, podemos destacar a transformação de dados, simulação numérica, modelagem estatística, visualização de dados e aprendizado de máquina (Naik, 2022).

A etapa inicial para utilização do *Google Colab* é a criação de um ambiente interativo, chamado *notebook Colab,* que possibilita a escrita dos códigos em Python. No espaço do notebook é possível importar conjuntos de dados de diversas formas. Através das *células de códigos*, toda a modelagem é construída, utilizando as diversas bibliotecas Python disponíveis

(Colab, 2023). A medição da performance dos modelos também é realizada nos ambientes criados nos notebooks, sendo possível a exportação de gráficos, tabelas e figuras.

Figura 3.2: Ilustração da tela inicial do Google Colab



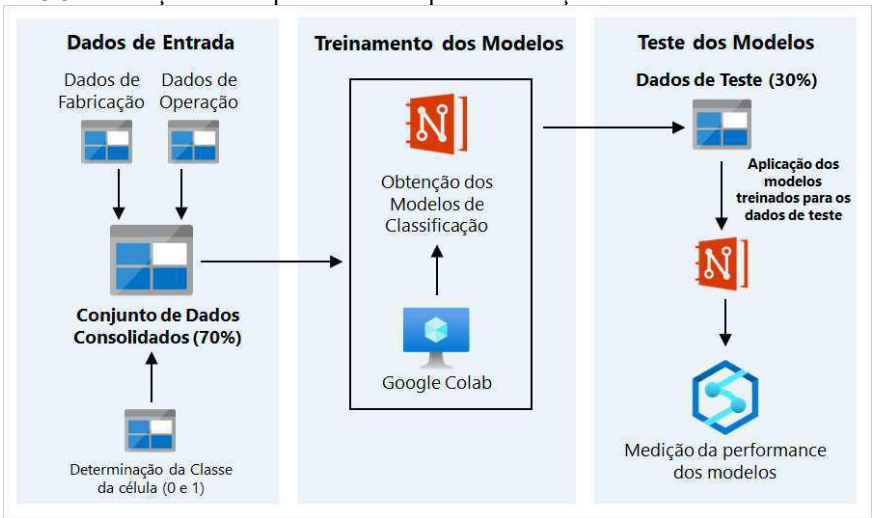
Fonte: Colab (2023)

Nesse trabalho, foram construídos os 5 modelos de classificação supervisionados mais utilizados em problemas de *machine learning*:

- Árvore de decisão;
- Random Forest (RF);
- Naive Bayes;
- K-Nearest Neighbours (KNN);
- Máquina de vetores de suporte (SVM);

As etapas para a construção dos modelos de classificação estão ilustradas na figura 3.3. De início, foram realizados o pré-processamento e a consolidação dos dados de entrada para os modelos utilizando o Microsoft Excel, com o carregamento dos dados direto no notebook do *Colab*, com uso de código *Python* específico. Ainda no próprio ambiente do *Colab*, os dados de entrada foram divididos em atributos e rótulos, ou seja, as variáveis de entrada e a classe das células. A construção dos modelos de classificação foi realizada utilizando o *default* de parâmetros para cada um dos modelos, sem nenhum processo de otimização inicial. Em seguida, a performance de cada preditor foi avaliada para o conjunto de dados de treinamento, com base nos resultados das métricas de avaliação e a construção das matrizes de confusão, que indicaram uma visão geral do desempenho de cada modelo. Por fim, foram obtidas as arvores de decisão para os modelos de árvore de decisão e *Random forest*, bem como as variáveis de maior importância e seus respectivos intervalos de variação.

Figura 3.3: Ilustração das etapas realizadas para construção dos modelos de classificação

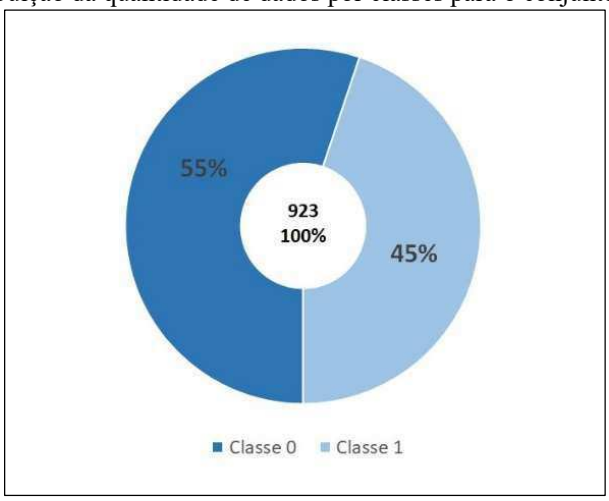


Fonte: elaborado pelo autor (2023)

Para a construção dos modelos foi utilizado um conjunto com 1333 dados, destes, 923 foram utilizados para treinamento e 410 para teste, relação de partição que seque o padrão de divisão treino/teste de 70%/30% utilizado em trabalhos envolvendo *machine learning* (Hastie et al., 2008).

A figura 3.4 mostram a distribuição de classes para o conjunto de dados de treinamento. É possível observar que o conjunto de dados está balanceado (equilibrados) para as duas classes, ou seja, uma quantidade equilibrada de dados representando cada classe da célula, fato que é importante para o bom desempenho dos modelos (Mohammed, 2020). Modelos treinados com conjuntos de dados desbalanceados são tendenciados para a classe majoritária, gerando modelos enviesados (Krawczyk, 2016). O conjunto de dados utilizados para teste dos modelos também estava balanceado nas duas classes.

Figura 3.4: Distribuição da quantidade de dados por classes para o conjunto de treinamento



Fonte: elaborado pelo autor (2023)

CAPÍTULO 4

RESULTADOS E DISCUSSÃO

4.1 Resultado da Modelagem

Após as etapas de treinamento e teste (conforme ilustrado na figura 3.3), a performance de cada modelo de classificação foi medida com base nas métricas de acurácia e *f1-score*, para o conjunto de dados de teste. A tabela 4.1 resume o desempenho de cada preditor para a classificação binária das células eletrolíticas como sendo ótima e não-ótima.

Tabela 4.1: Resumo do resultado de desempenho dos modelos

Modelo	Acurácia	*f1-score*
Árvore de decisão	0,788	0,765
Random Forest	**0,880**	**0,866**
Naive Bayes	0,744	0,743
KNN	0,707	0,689
SVM	0,846	0,828

Fonte: elaborado pelo autor (2023)

Como visto na tabela 4.1, o modelo de *Random Forest* obteve o maior desempenho, em comparação aos outros modelos, conforme resultados da acurácia e *f1-score*. Considerando uma avaliação mais geral, o valor da acurácia é relevante para avaliação dos modelos por ser uma medida intuitiva e direta, além do fato de que os dados utilizados para construção e validação dos modelos estarem balanceados, o que traz mais robustez a essa métrica.

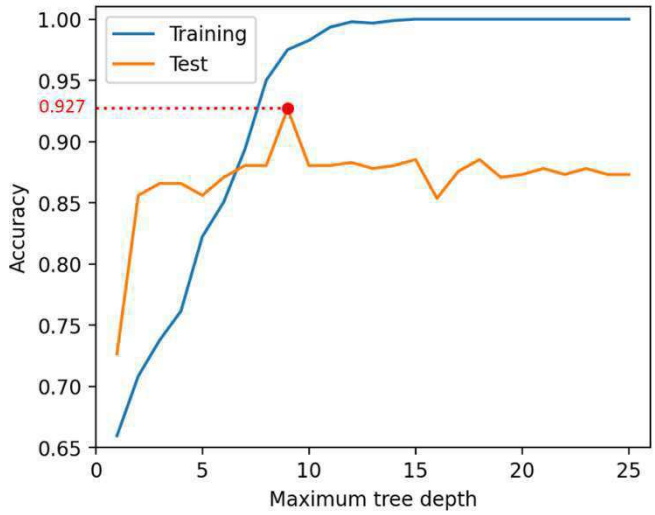
O maior desempenho do modelo de *Random Forest* pode ser atribuído a características especificas desse modelo, como robustez, bom desempenho com variáveis numéricas e categóricas, além da capacidade de modelar interações complexas (Louppe, 2014).

Para o modelo de Random Forest, que possui maior facilidade de interpretação e implementação, além de ter sido um dos modelos em foco para o desenvolvimento desse trabalho, foi aplicado a metodologia de otimização de dois de seus *hiperparâmetros*, que são a *profundidade máxima das árvores (TD)* e o *número de preditores ou árvores (NE)*, com intuito de se obter os valores de *TD* e *NE* que proporcionassem a maior acurácia para o modelo. Conforme podemos observar no gráfico 4.1 e gráfico 4.2, a profundidade máxima da árvore de 9 (ponto em vermelho destacado no gráfico), bem como o número de preditores de 85, proporcionaram aumento no desempenho do modelo de *Random Forest*, com resultado de **92,7%** de acurácia após o processo de otimização.

Outra constatação observada após a otimização dos *hiperparâmetros* foi que a medida em que a quantidade de preditores (*NE*) ou a profundidade máxima das árvores (*TD)* é acrescida, a acurácia do modelo para o conjunto de treinamento aumentou até próximo de 100%, juntamente com o aumento e consistência da acurácia para o conjunto de dados de teste, o que mostra que não houve *sobreajuste* (*overfitting*) no modelo, que segundo Júnior (2018),

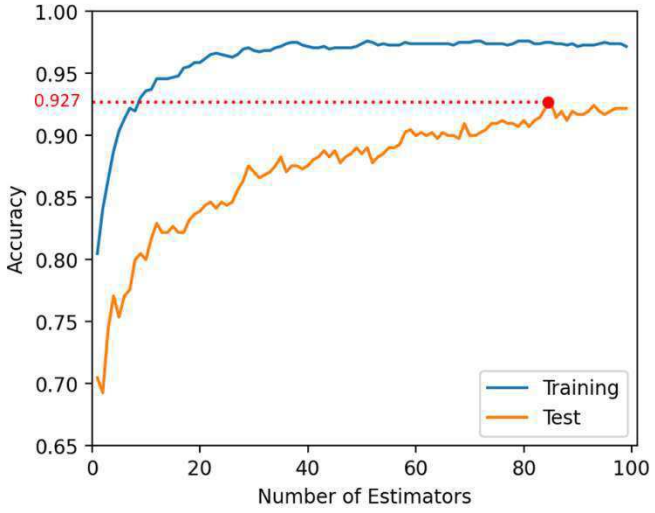
ocorre quando o modelo obtém baixo erro durante o processo de treinamento associado a elevado erro durante o teste com dados nunca vistos. Os modelos de *RF* conseguem reduzir o *overfitting* devido a geração de árvores de decisões aleatórias, que compõem o modelo (Silva, 2019). No Gráfico 4.1, por exemplo, é possível observar uma elevação e estabilização da acurácia do modelo à medida que o *hiperparâmetro TD* é acrescido, considerando o conjunto de dados de teste. O mesmo resultado é observado durante otimização da quantidade de preditores do modelo, ilustrado no Gráfico .

Gráfico 4.1: Profundidade máxima (TD) ótima que proporciona a maior acurácia para o modelo *RF*



Fonte: elaborado pelo autor (2023)

Gráfico 4.2: Número máximo de árvores (NE) ótima que proporciona a maior acurácia para o modelo *RF*

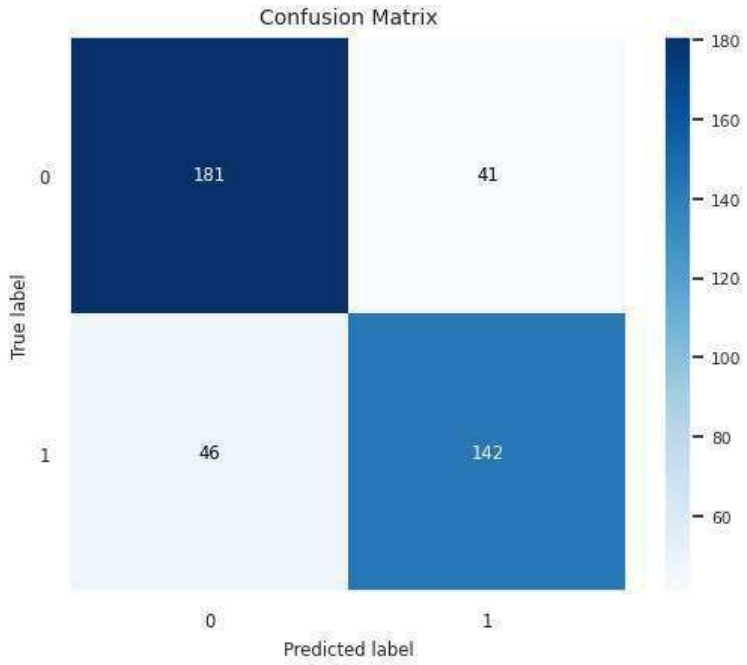


Fonte: elaborado pelo autor (2023)

As matrizes de confusão para cada modelo, vistas na figura 4.1 à figura 4.5, foram obtidas a partir do conjunto de dados de teste. Como podemos observar na figura 4.2, a menor quantidade nos valores de $FN = 19$ e $FP = 11$ para o modelo de Random Forest (RF)

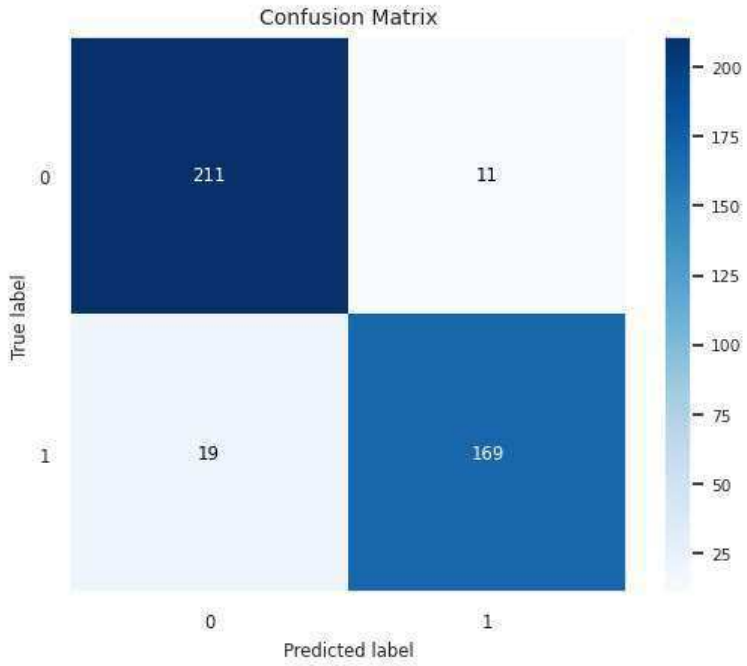
corroboram o melhor desempenho desse preditor para a classificação das células. Além disso, é possível observar a taxa de acerto similar do modelo para as duas classes, com desempenho levemente superior para a classe 0.

Figura 4.1: Matriz de confusão para o modelo de *Árvore de Decisão*



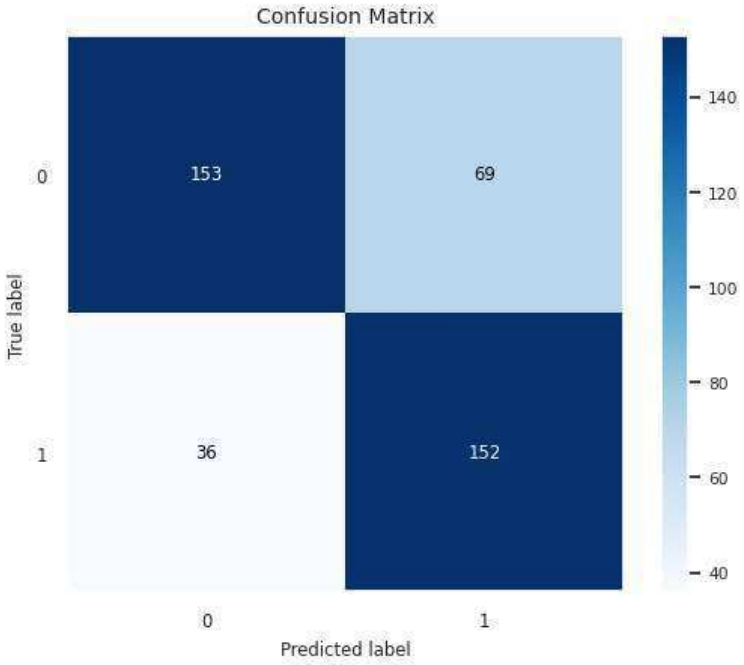
Fonte: elaborado pelo autor (2023)

Figura 4.2: Matriz de confusão para o modelo de *Random Forest*



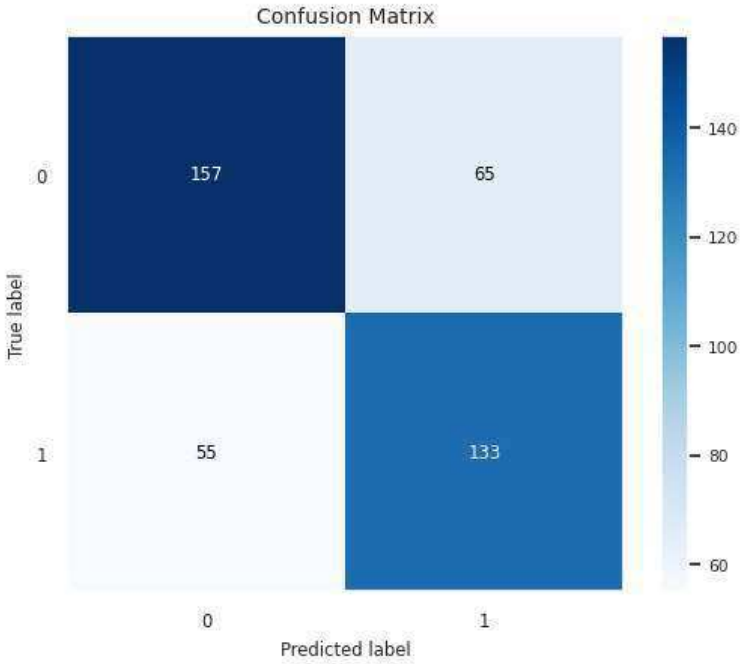
Fonte: elaborado pelo autor (2023)

Figura 4.3: Matriz de confusão para o modelo de *Naive Bayes*



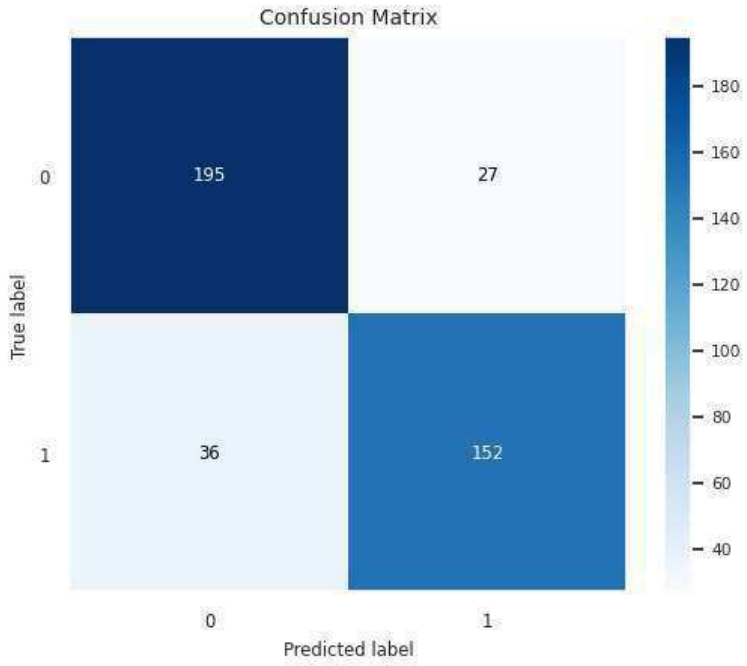
Fonte: elaborado pelo autor (2023)

Figura 4.4: Matriz de confusão para o modelo de *KNN*



Fonte: elaborado pelo autor (2023)

Figura 4.5: Matriz de confusão para o modelo de *SVM*



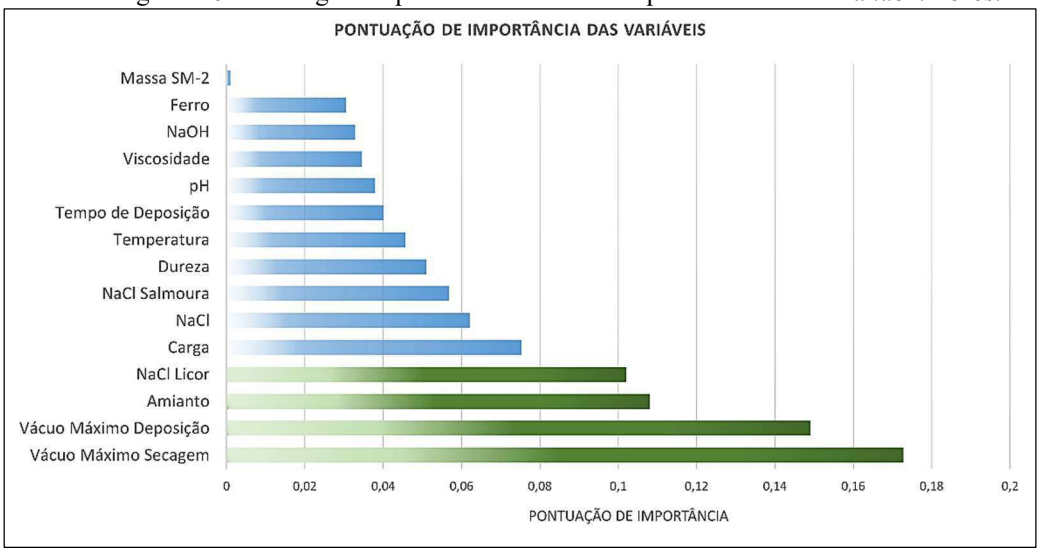
Fonte: elaborado pelo autor (2023*)*

Além da obtenção de modelos com acurácia adequada, outro resultado importante em problemas de *machine learning* é determinar a importância de cada variável de entrada (*features importance*) no desempenho dos modelos, proporcionando maior conhecimento do processo em estudo.

A obtenção das *features importance* é fundamental para problemas que surgem devido ao fato de que cada vez mais cresce a quantidade de variáveis de entrada que são introduzidas para construção dos modelos, o que pode trazer variáveis preditoras ruidosas, mistura de variáveis numéricas e categóricas, bem como a criação de interações complexas. Os benefícios da determinação das variáveis mais importantes para o modelo podem ser o menor esforço computacional, aumento na precisão e mais fácil interpretação dos modelos (Louppe, 2014).

A partir da figura 4.6, podemos concluir que as variáveis: *vácuo máximo na secagem* e *vácuo máximo na deposição,* apresentaram maior pontuação de importância relativa para o modelo de *Random Forest*. Esse resultado vai em concordância com o conhecimento empírico da importância dessas variáveis na performance da célula. As variáveis NaCl no licor e dureza também se destacaram na composição do modelo de *RF*.

Figura 4.6: Ranking de importância das variáveis para o modelo de *Random Forest*



Fonte: elaborado pelo autor (2023)

Após a construção, treinamento e otimização do modelo *Random Forest*, a obtenção da árvore que compõe o modelo é fundamental para avaliação e interpretação do preditor, além de determinar as faixas das variáveis em que há a maior probabilidade de uma célula ser classificada como sendo ótima (classe 1), ou seja, ter um desempenho satisfatório durante operação. Conforme observado na figura 4.7, o nó raiz (nó #0) foi definido com a variável *vácuo máximo na deposição*, justamente uma das variáveis de maior importância para o modelo, segundo o ranking de pontuação ilustrado na figura 4.6.

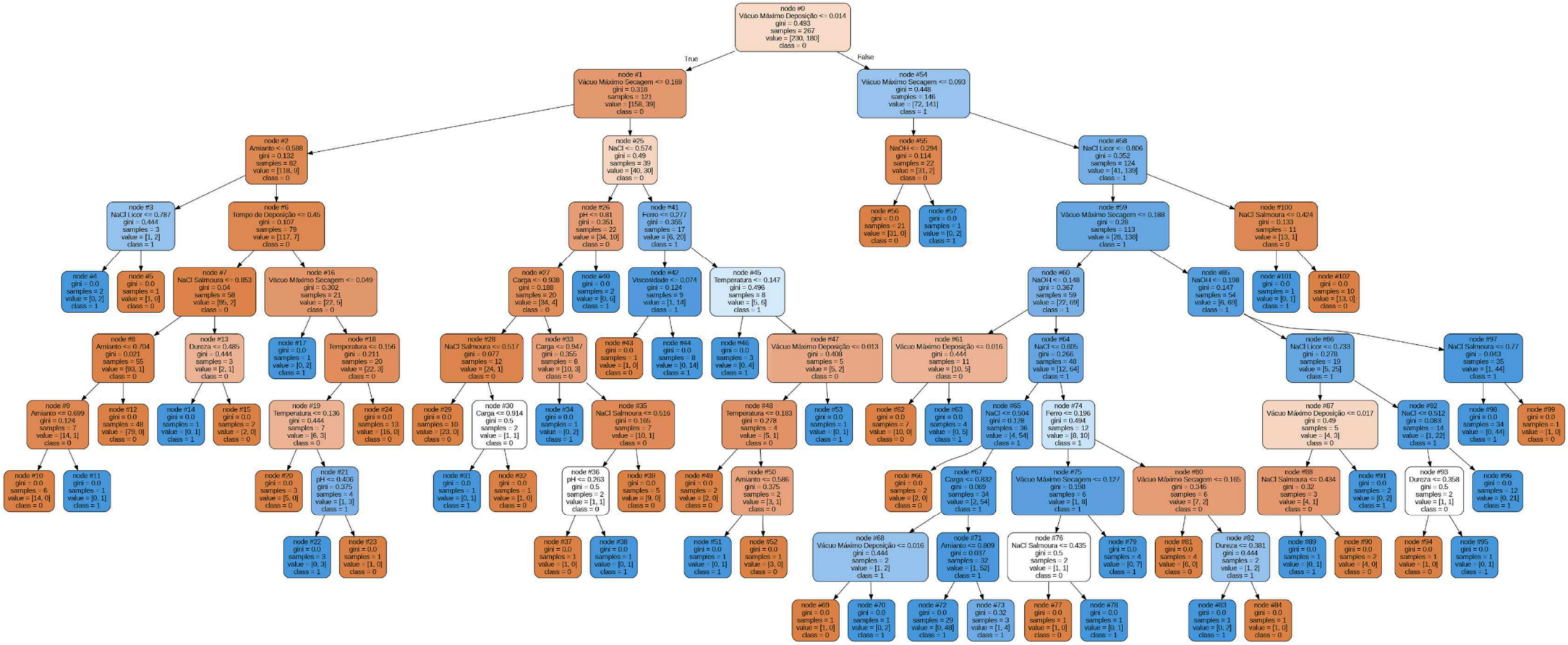
A partir da árvore de decisão ilustrada na figura 4.7 e considerando que o nó *#72* como o nó terminal alvo, por possuir a maior quantidade de amostras aliado a maior pureza possível (índice $Gini = 0$), percorrendo o caminho desde o nó raiz até os nó #72 é possível determinar o range de trabalho das variáveis para uma célula aleatória possuir a maior probabilidade de ser classificada como ótima. Com isso, mantendo o intervalo de trabalho das variáveis descrito na tabela 4.2, obtemos a maior probabilidade de uma célula aleatória ser classificada como ótima, ou seja, ter o maior desempenho durante operação. Devido sigilo de informações industriais, os ranges das variáveis descrito na tabela 4.2 e figura 4.7 foram normalizados.

Tabela 4.2: Resumo do intervalo de trabalho das variáveis para obtenção de uma célula ótima

Variável	Intervalo
Vácuo máximo na Deposição	$\geq 0{,}014$
Vácuo Máximo na Secagem	$\geq 0{,}09$ e $\leq 0{,}188$
NaCl no Licor	$\leq 0{,}806$
NaCl	$\leq 0{,}504$
NaOH	$\leq 0{,}148$
Carga	$\leq 0{,}832$
Amianto	$\leq 0{,}806$

Fonte: elaborado pelo autor (2023)

Figura 4.7: Árvore obtida a partir do modelo de *Random Forest*



Fonte: elaborado pelo autor (2023)

CAPÍTULO 5

CONCLUSÕES

5.1 Conclusões

Nesse trabalho foram construídos modelos de classificação supervisionado para a predição da performance e otimização das células eletrolíticas com diafragma de amianto. Os modelos treinados apresentaram desempenho satisfatório na previsão do desempenho das células a partir de dados de fabricação do diafragma e operação da célula. O modelo *Random Forest* destacou-se com relação ao demais modelos, com acurácia de 92,7%, resultado importante que serve de base para a busca da melhoria de performance do diafragma. Esse resultado confirma a viabilidade da aplicação de técnicas de aprendizado de máquinas na indústria de produção de cloro e soda cáustica, possibilitando a implementação de melhorias nos processos de produção, tais como aumento do tempo de vida útil dos diafragmas, maior tempo de campanha dos materiais sobressalentes, menor consumo de energia e redução de custos de manutenção e operação.

A otimização de *hiperparâmetros* também se mostrou aplicável no tipo de modelagem desenvolvida, sendo possível o incremento ($\approx 5\%$) no desempenho do modelo de *Random Forest* treinado. Foi possível constatar também que não houve *sobreajuste* (*overfitting*) do modelo de *RF* durante o processo de otimização dos *hiperparâmetros*, o que assegura a manutenção da performance do modelo para novos dados.

Outro resultado importante da modelagem desenvolvida foi a obtenção das variáveis de maior relevância para os modelos de classificação, viabilizando o controle dessas variáveis nos processos de fabricação e operação dos diafragmas, contribuindo para a otimização da performance das células eletrolíticas. O presente estudo mostrou que o vácuo resultante aplicado no diafragma de amianto (*vácuo máximo na secagem* e *vácuo máximo na deposição*) possui maior relevância no desempenho final dos diafragmas, resultado que é tido empiricamente como verdadeiro, não sendo mensurado sua faixa ótima até então.

Com o resultado do trabalho foi possível concluir também que a técnica utilizada para construção e treinamento dos modelos de classificação pode ser aplicada a outros processos ou problemas que envolvam classificação com dimensão elevada de variáveis de entrada.

CAPÍTULO 6

REFERÊNCIAS BIBLIOGRÁFICAS

ABICLOR. *Balanço socioeconômico da indústria de cloro-álcalis no brasil.* 2020. Disponível em: https://www.abiclor.com.br/cloro/. Acesso em 1 ago. 2023.

ANDRADE, M. H. S. *Estudo e Otimização da Fluidodinâmica do Anólito de Celas de Cloro-Soda com Tecnologia de Diafragma*. 2006. Tese (Doutorado em Engenharia de Processos, Pós-Graduação em Engenharia de Processos). Universidade Federal de Campina Grande, Campina Grande, 2006.

BARELLA, V. H. *Técnicas para o problema de dados desbalanceados em classificação hierárquica*. 2015. 85 p. Dissertação (Mestrado em Ciências). Universidade de São Paulo, São Paulo, 2015.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. 1st ed. Cambridge: Springer, 2006. 738 p.

BRAGA, J.M.F. Análise da Viabilidade Econômica da Integração de Sistemas de Célula a Combustível, nas Plantas de Cloro-Soda, para Utilização do Hidrogênio Gerado no Processo. 2009. Tese (Doutorado em Engenharia Química), Universidade Federal do Rio de Janeiro - UFRJ, Rio de Janeiro – Rio de Janeiro.

BRIEM, G. J., BENEDIKTSSON, J. A., SVEINSSON, J. R. *Multiple Classifiers Applied to Multisource Remote Sensing Data*. IEEE. v. 40, Oct. 2002.

BREIMAN, L. *Bagging Predictors.* Machine Learning. Boston. v. 24, p. 123-140, 1996.

BREIMAN, L. *Random Forests.* Machine Learning. Boston. 1999. v. 45, p. 5-32, Apr. 2001.

CABRAL, B. R. *Processos Gaussianos para Aprendizado Supervisionado*. 2021. 106 p. TCC (Bacharelado em Matemática e Computação Científica). Universidade Federal de Santa Catarina, Florianópolis, 2021.

CAVALIN, P., OLIVEIRA, L. *Confusion Matrix-Based Building of Hierarchical Classification.* Springer Nature Switzerland. Rio de Janeiro. p. 271–278, 2019.

ÇINAR, Z. M., ABDUSSALAM NUHU, A., ZEESHAN, Q., KORHAN, O., ASMAEL, M., SAFAEI, B. *Machine Learning in Predictive Maintenance towards Sustainable Smart Manufacturing in Industry 4.0*. Sustainability 2020, 12, 8211. https://doi.org/10.3390/su12198211.

COLAB, 2023. Conheça o Colab. Página inicial. Disponível em < https://colab.research.google.com/>. Acesso em: 18 de set. 2023.

CONGALTON, R. G. *A review of assessing the accuracy of classifications of remotely sensed data*. 1990. Remote Sensing of Environment, v. 37, p. 35-46. Apr. 1991.

CORREIA, P. H. O. *Desenvolvimento de novas tecnologias aplicáveis a produção de cloro-álcalis no Brasil frente as necessidades ambientais atuais: Uma revisão bibliográfica.* 2022. 36 p. TCC (Bacharelado em Engenharia Química). Universidade Federal da Paraíba, João Pessoa, 2022.

CUNHA, C. T. C. *Desenvolvimento de diafragmas poliméricos aplicáveis na produção eletrolítica de cloro-soda*. 2015. 113 p. Tese (Doutorado em Ciência e Engenharia de Materiais). Universidade federal de campina grande, Campina Grande, 2015.

CUNHA, R. O. O. *Modelo de regressão por processos gaussianos Aplicado a problemas de otimização estrutural via Metaheurísticas*. 2018. 59 p. Trabalho de conclusão de curso (Bacharelado em Estatística). Universidade Federal de Juiz de Fora, Juiz de Fora, 2018.

CUNNINGHAM, P., DELANY, S. J. *k-Nearest Neighbour Classifiers - A Tutorial*. 2021. ACM Computing Surveys, Dublin, v. 54, n. 6, p. 128 – 153. Jul. 2021.

DA SILVA, L. A., PERES, S. M., BOSCARIOLI, C. *Introdução à mineração de dados: com aplicações em R*. 1. ed. Rio de Janeiro: Elsevier, 2016. 468 p.

DE CASTRO, C. L. *Novos critérios para seleção de modelos neurais em problemas de classificação com dados desbalanceados*. 2011. 154 p. Tese (Doutorado em Engenharia Elétrica). Universidade Federal de Minas Gerais. Minas Gerais, 2011.

DHANABAL, S. CHANDRAMATHI, S. *A Review of various k-Nearest Neighbor Query Processing Techniques*. 2020. International Journal of Computer Applications, Coimbatore, v. 31, n. 7, p. 975 – 8887. Oct. 2011.

DOMINGOS, P., PAZZANI, M. *On the Optimality of the Simple Bayesian Classifier under Zero-One Loss*. 1997. Machine Learning, Irvine. v. 29, p. 103 – 130. 1997.

DONADIA, D. D. E. A. *Comparação entre as técnicas de regressão logística, árvore de decisão, bagging e random forest aplicadas a um estudo de concessão de crédito.* 2013. 67 p. TCC (Bacharel em Estatística). Universidade federal do paraná, Curitiba, 2013.

FONTES, D. O. L. *Desenvolvimento de soft sensor para predição do estado térmico do ferro gusa em alto-forno usando fuzzy c-médias e modelo exógeno auto-regressivo não linear.* Dissertação (Mestrado em Engenharia Química). Universidade federal de campina grande, Campina Grande, 2020.

FRADKOV, A. L. *Early History of Machine Learning*. 2020. IFAC Papers On Line. v. 53, p. 1385 – 1390. 2020.

ESTOPIER, L. M. O., GOURVÉNEC, S. A, CAHORS, R., BEHARA, N., SCELLIER, J. *Prediction of flooding in distillation columns using machine learning*. 2023. Digital Chemical Engineering. 2023. França. 2023.

FILHO, E. M. A. *Caracterização Físico Química e Modelagem Estatística de Diafragmas de Células Eletrolíticas Utilizadas para Produção de Clorossoda*. 2009. 60 p. Dissertação (Mestrado em Engenharia Química). Universidade Federal de Campina Grande, Campina Grande, 2009.

FILHO, E. M. A., VILAR, E. O., FEITOZA, A. C. O. *Physical–chemical characterization and statistical modeling applied in a chlor-alkali diaphragm-cell process*. 2011. Chemical Engineering Research and Design, Maceió. v. 89, p. 491 – 498. 2011.

FRANK, E., TRIGG, L., HOLMES, G., WITTEN, I. H. *Technical Note: Naive Bayes for Regression*. 2000. Machine Learning, Hamilton. v. 41, p. 5 – 25. 2000.

GOOGLE, 2023. Colaboratory. Página inicial. Disponível em < https://research.google.com/colaboratory/intl/pt-BR/faq.html>. Acesso em: 18 de set. 2023.

GOMES, J. C. B. *Estimação Não-Paramétrica para Função de Covariância de Processos Gaussianos Espaciais*. 2009. 138 p. Dissertação (Mestrado em estatística). Universidade Estadual de Campinas, Campinas, 2009.

GE, Z., SONG, Z., DING, S. X., HUANG, B. *DATA MINING AND ANALYTICS IN THE PROCESS INDUSTRY: THE ROLE OF MACHINE LEARNING*. IEEE Access, vol. 5, p. 20590-20616, 2017, DOI: 10.1109/ACCESS.2017.2756872.

GUYON, I., ELISSEEFF, A. *An Introduction to Variable and Feature Selection.* 2002. Journal of Machine Learning Research. v. 3, p. 1157 – 1182. 2003.

HASTIE, T., TIBSHIRANI, R., E FRIEDMAN, J. H. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag. Second Edition.

HEYDARIAN, M., DOYLE, T. E., SAMAVI, R. *MLCM: Multi-Label Confusion Matrix*. 2021. IEEE Access. v. 10, p. 19083 – 19095. Feb. 2022.

HINE, F., YASUDA, M., TANAKA, T., 1977. *Mass transfer through the deposited asbestos diaphragm in chlor-alkali cells*. Electrochim. Acta 22, 429–437.

ISLAM, M. J., WU, Q. M. J., AHMADI, M., SID-AHMED, M. A. *Investigating the performance of Naive-Bayes classifiers and K-nearest neighbor classifiers*. 2007. International Conference on Convergence Information Technology, p. 1541-1546, Nov. 2007.

JUNIOR, A. N. *Aplicação de redes neurais utilizando o software Matlab*. Monografia (Graduação em ciências da computação). Centro universitário Eurípides de Marília. Marília, 2005.

JÚNIOR, R. N. J. *Modelagem matemática de um Processo Industrial de Produção de Cloro e Soda por Eletrólise de Salmoura visando sua Otimização*. 2006. Dissertação de Mestrado. 139 p. Dissertação (Mestrado em Engenharia). Universidade de São Paulo. São Paulo, 2006.

JÚNIOR, W. J. A. *Métodos de otimização hiperparamétrica: um estudo Comparativo utilizando árvores de decisão e florestas Aleatórias na classificação binária*. 2018. Dissertação (Mestrado em Engenharia Elétrica.). Universidade Federal de Minas Gerais. Belo Horizonte, 2018.

JÚNIOR, W. J. A. *Metamodelagem kriging dinâmica aplicada em trocadores de calor*. 2019. Dissertação de Mestrado. 2019. 94 p. Dissertação (Mestrado em Engenharia química). Universidade Federal de Campina Grande. Campina Grande, 2019.

JUNIOR, C. F. M.; SILVA, R. T. C.; VILAR, E. O. *Development and evaluation of uhmwpe microfibers polymeric Diaphragms for electrolytic production of sodium hydroxide*. 2021. International Journal of Development Research, vol. 11. p. 50428-50434, Sep. 2021.

JUNIOR, G. B. V. et al. *Determinação das Métricas Usuais a Partir da Matriz de Confusão de Classificadores Multiclasses dm Algoritmos Inteligentes das Ciências do Movimento Humano*. 2022. CPAQV Journal. v. 14, p. 1-9. 2022.

KAVEHA, N. S., MOHAMMADIB, F., ASHRAFIZADEHA, S.N. *Prediction of cell voltage and current efficiency in a lab scale chlor-alkali membrane cell based on support vector machines*. 2007. Chemical Engineering Journal. vol. 147. p. 161-172, Jun. 2008.

KRAWCZYK, B. *Learning from imbalanced data: open challenges and future directions*. 2016. Springer. Wrocław, v. 5. p. 221-232, Apr. 2016.

KUHN, M., JOHNSON, K. *Applied Predictive Modeling*. 2013. 600 p. Springer New York, NY. DOI: https://doi.org/10.1007/978-1-4614-6849-3.

LANGSETH, H., NIELSEN, T. D. *Classification using Hierarchical Naive Bayes models*. 2005. Mach Learn, vol. 63. p. 135-139, Mar. 2006.

LOUPPE, G. *Understanding random forests: From theory to practice*. Tese de Doutorado, Universidade de Lieja. 2014.

LOPEZ, A. *Global Chlor-alkali Market Outlook*. 2018. Clorosur Technical Conference. Disponível em: https://www.clorosur.org/seminar2018/presentations/15-10.pdf. Acesso em 8 ago. 2023.

MCINTYRE, J. *100 Years of Industrial Electrochemistry*. 2002. Journal of The Electrochemical Society, vol. 149. p. 79-83, 2002.

MOHAMMED, H., HAMEED, I. A., SEIDU, R. *Random Forest Tree for Predicting Fecal Indicator Organisms in Drinking Water Supply*. Ålesund, 2017. ResearchGate. 7 p. DOI: 10.1109/BESC.2017.8256398. Disponível em: https://www.researchgate.net/publication/32251704. Acesso em 10 ago. 2023.

MOHAMMED, R., RAWASHDEH, J., ABDULLAH, M. *Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results*. 2020. International Conference on Information and Communication Systems. Wrocław, p. 243-248, 2020.

NAIK, P., NAIK, G., PATIL, M. *Conceptualizing Python in Google COLAB*. India: Shashwat Publication, 2022.

NASTESKI, V. *An overview of the supervised machine learning methods*. 2017. ResearchGate. 11 p. DOI: 10.20544/HORIZONS.B.04.1.17.P05. Disponível em: https://www.researchgate.net/publication/328146111. Acesso em 10 Jul. 2023.

NAVIN, M. J. R., PANKAJA, R. *Performance Analysis of Text Classification Algorithms using Confusion Matrix*. 2016. International Journal of Engineering and Technical Rese, vol. 6. p. 75-78, Dec. 2016.

PANOV, P., DŽEROSKI, S. Combining Bagging and Random Subspaces to Create Better Ensembles. 2007. Advances in Intelligent Data Analysis VII. IDA 2007. vol 4723, p 118-129. 2007.

PARMAR, A., RAKESH, K., VATSAL, P. *A Review on Random Forest: An Ensemble Classifier*. 2018. Springer Nature Switzerland. vol. 26. p. 758-763, 2019.

PASSOS, A. G. *Otimização Multiobjetivo Com Base Em Processo Gaussiano De Regressão (Kriging)*. 2020. 162 p. Tese (Doutorado em Engenharia). Universidade Tecnológica Federal Do Paraná, Curitiba, 2020.

PELLICER, L. F. A. O. *Otimização de Hiperparâmetros de Modelos de Machine Learning com Baryseach*. 2020. Dissertação de Mestrado. 2019. 94 p. Dissertação (Mestrado em Ciências). Universidade de São Paulo. São Paulo, 2020.

PISNER, D. A., SCHNYER, D. M. *Support vector machine*. Elsevier. University of Texas, Austin, 2020. https://doi.org/10.1016/B978-0-12-815739-8.00006-7. p. 101– 121. 2020.

RAY, S. A *Quick Review of Machine Learning Algorithms*. India. 2019. J. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, p. 35-39, Feb. 2019.

RODRIGUES, W., CANNAVALE, V., TREVISAN, D. M. Q., PRATA, D. *Uso De Machine Learning Para A Análise De Projetos Legislativos De Desenvolvimento Regional: O Caso Da Zona Franca De Manaus.* 2022. J. Informe GEPEC, vol. 26. p. 127-140, Jun. 2022.

SANTOS, H. G. *Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina*. 2018. Tese (Doutorado em Epidemiologia). Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2018. DOI:10.11606/T.6.2018.tde-09102018-132826. Acesso em: 2023-09-17.

SCALCO, F. F. *Visualização De Dados Em Processos De Machine Learning*. 2021. 118 p. TCC (Bacharelado em Ciência da Computação). Universidade de Caxias do Sul, Caxias do Sul, 2021.SHIMIZU, N., KANEKO. H. *Constructing Regression Models with High Prediction Accuracy and Interpretability Based on Decision Tree and Random Forests*. 2020. J. Comput. Chem, vol. 20. p. 71-87, Aug. 2021.

SCHULZ, A. C., BOMMARAJU, T. V, KISZEWKI, R., KELLER, U. Diaphragm for an electrolytic cell. 1987. (ELTECH) Patente US 4810345, Nov. 1989.

SILVA, E. R. *Técnicas de Metamodelagem Aplicadas à Otimização de Turbomáquinas*. 2011. 142 p. Tese (Doutorado em Engenharia Mecânica). Universidade Federal de Itajubá, Itajubá, 2011.

SILVA, I. S. *Inteligência Artificial Para Avaliação Da Qualidade Da Água*. 2019. 105 p. Dissertação (Mestrado em Recursos Hídricos). Universidade Federal De Sergipe, São Cristóvão, 2019.

SILVA, D. F. B. F. *Pré-processamento de Dados e Comparação entre Algoritmos de Machine Learning para a Análise Preditiva de Falhas em Linhas de Produção para o Controlo de Qualidade*, 2021.

SUGAHARA, J A. S. *Machine Learning e Data Science na indústria: aplicações e desafios.* 2020. 55 p. TCC (Bacharelado em Engenharia de Produção Mecânica). Universidade Estadual Paulista, São Paulo, 2020.

TAHERI, S., MAMMADOV, M. *Learning The Naive Bayes Classifier With Optimization Models*. 2012. Int. J. Appl. Math. Comput. Sci. New Delhi, v. 23. No. 4, p. 787 – 795. 2013.

VAN ZEE, J.W., WHITE, R.E., WATSON, A.T. 1986. Simple models for diaphragm-type chlorine/caustic cells. J. Electrochem. Soc. 133 (3), p501.

VAIDA, K., GHOSEB, U. *Predictive Analysis of Manpower Requirements in Scrum Projects Using Regression Techniques*. 2020. Procedia Computer Science, New Delhi, v. 173, p. 335 – 344. 2020.

VEIGA, R. K. S. *Metamodelo para estimar o desempenho térmico de edificações residenciais multifamiliares naturalmente ventiladas*. Dissertação de Mestrado. 2021. 130 p. Dissertação (Mestrado em Engenharia Civil). Universidade Federal De Santa Catarina. Florianópolis, 2021.

VIANA. K. M. S. *Diafragmas de PEUAPM para aplicação no processo de produção eletrolítica de cloro-soda*. 2009. 113 p. Tese (Doutorado em Engenharia de Processos). Universidade federal de campina grande, Campina Grande, 2009.

VILLAR, S. B. B, *et al*. *Otimização utilizando metamodelo kriging: uma aplicação à separação de propeno por destilação*. 2016. Revista Interdisciplinar de Pesquisa em Engenharia, Brasília. 2016.

VILLAR, S. B. B, *et al*. *Metamodelagem Kriging E Sua Aplicação Na Otimização De Uma Unidade De Separação De Propeno Por Destilação*. 2016. 81 p. Dissertação (Mestrado em Engenharia Química). Universidade Federal de Campina Grande. Campina Grande, 2016.

Yumashev, A. V., Fateminasab, S. M., Marjani, A., Lirgeshas, A. B. *Development of computational methods for estimation of current efficiency and cell voltage in a Chlor-alkali membrane cell*. 2021. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2021. DOI: https://doi.org/10.1080/15567036.2021.1897194.