UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

PAOLLA MARLENE CAETANO DA CUNHA

CONSTRUÇÃO DE UM SENSOR VIRTUAL PARA A CLASSIFICAÇÃO DE EMISSÕES DE DIÓXIDO DE ENXOFRE EM UMA CALDEIRA KRAFT VIA ALGORITMO *k*-NN (*k-Nearest Neighbours*)

BELO HORIZONTE - MG
2021

PAOLLA MARLENE CAETANO DA CUNHA

CONSTRUÇÃO DE UM SENSOR VIRTUAL PARA A CLASSIFICAÇÃO DE EMISSÕES DE DIÓXIDO DE ENXOFRE EM UMA CALDEIRA KRAFT VIA ALGORITMO *k*-NN (*k-Nearest Neighbours*)

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Química da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do Grau de Mestre em Engenharia Química.
Linha de Pesquisa: Engenharia de Sistemas em Processos.
Orientador: Prof. Dr. Gustavo Matheus de Almeida.

BELO HORIZONTE – MG
2021

Cunha, Paolla Marlene Caetano da.
C972c Construção de um sensor virtual para a classificação de emissões de dióxido de carbono em uma caldeira kraft via algoritmo *k*-NN (*k-Nearest Neighbours*) [recurso eletrônico] / Paolla Marlene Caetano da Cunha. - 2021.
1 recurso online (xiv, 75 f. : il., color.) : pdf.

Orientador: Gustavo Matheus de Almeida.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Apêndice: f. 74-75.
Bibliografia: f. 68-73.

Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia química - Teses. 2. Polpação alcalina por sulfato – Teses. 3. Sensor virtual – Teses. I. Almeida, Gustavo Matheus de. II. Universidade Federal de Minas Gerais. Escola de Engenharia. III. Título.

CDU: 66.0(043)

Ficha catalográfica elaborada pela Bibliotecária Letícia Alves Vieira - CRB-6/2337
Biblioteca Prof. Mário Werneck - Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

FOLHA DE APROVAÇÃO

"CONSTRUÇÃO DE UM SENSOR VIRTUAL PARA A CLASSIFICAÇÃO DE EMISSÕES DE DIÓXIDO DE ENXOFRE EM UMA CALDEIRA KRAFT VIA ALGORITMO K-NN (K-NEAREST NEIGHBOURS)"

Paolla Marlene Caetano da Cunha

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Química da Escola de Engenharia da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de **MESTRE EM ENGENHARIA QUÍMICA**.

285ª DISSERTAÇÃO APROVADA EM 30 DE AGOSTO DE 2021 POR:

Documento assinado eletronicamente por **Gustavo Matheus de Almeida**, **Professor do Magistério Superior**, em 30/08/2021, às 11:44, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Roberto da Costa Quinino**, **Coordenador(a) de curso**, em 30/08/2021, às 11:45, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Giovani Guimarães Rodrigues**, **Usuário Externo**, em 30/08/2021, às 12:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0911178** e o código CRC **1129EC05**.

Referência: Processo nº 23072.243467/2021-37 SEI nº 0911178

AGRADECIMENTOS

Agradeço primeiramente à Deus, o Criador, pela vida e o caminho que me levou até aqui. Agradeço a minha família e ao meu namorado pelo apoio e confiança no meu potencial. Realizar esse programa de mestrado não foi fácil, principalmente em meio a uma pandemia, como nunca vista. Obrigada pela ajuda de vocês.

Agradeço também ao meu orientador, pela paciência demonstrada e todo o ensinamento sobre ciência de dados e o processo *kraft*. Ao grupo de pesquisa por me influenciar na escolha da linguagem de programação. Estou muito feliz por tudo que desenvolvi pessoal e profissionalmente. Tenho orgulho do meu trabalho e espero que ele possa contribuir no desenvolvimento da indústria e no aprendizado de mais pessoas.

“A new, a vast, and a powerful language is developed for the future use of analysis, in which to wield its truths so that these may become of more speedy and accurate practical application for the purposes of mankind than the means hitherto in our possession have rendered possible.”
Ada Lovelace

RESUMO

O desenvolvimento industrial é um dos principais fatores para regulamentações ambientais em geral. Certificações que padronizam a gestão e parâmetros do processo foram consolidados no mundo todo, e hoje são vistos como obrigatórios para a operação industrial sustentável. O desenvolvimento de modelos baseados em dados se fortaleceu a partir da geração massiva de dados pelos processos industriais em geral em todo o mundo, dado o avanço das áreas de instrumentação, informática e banco de dados. Esse avanço contribui então para a disseminação de aplicações de ciência de dados no setor industrial. Uma sub-área da ciência de dados diz respeito aos métodos de aprendizado de máquina, que geralmente são de simples implementação e obtidos diretamente a partir de conjuntos de dados. Esse trabalho explorou um problema de classificação pelo método supervisionado denominado de k-vizinhos mais próximos (k-NN; k-*Nearest Neighbours*), com o foco em emissões de gases poluentes à saúde humana e ao meio ambiente. De modo específico, propõe a construção de um sensor virtual baseado em dados para o monitoramento e, por conseguinte, para o controle de emissões de dióxido de enxofre (SO_2) em caldeiras de recuperação química do setor de celulose *kraft*. Em relação à metodologia, foi apresentado o comportamento do método com uma variável de entrada, com subconjuntos de preditores e com um comitê de modelos (*ensemble learning*), para seis classes de SO_2. Os resultados apresentados são satisfatórios, o que é importante para gerar confiança em implementações industriais. O comitê de modelos apresentou a melhor performance, com acurácia de 92% e média geométrica de 94,75% sobre o conjunto (independente) de teste.

Palavras-chave: Sensor virtual; Caldeira kraft; Dióxido de Enxofre; k-Nearest Neighbours (KNN); Ensemble Learning; Processo Kraft; Estudo de caso real.

ABSTRACT

Industrial development is one of the main factors for environmental regulations in general. Certifications that standardize management and process parameters have been consolidated worldwide and are now seen as mandatory for sustainable industrial operation. The development of data-driven models has been strengthened by the massive generation of data by industrial processes in general all over the world, given the advances in the areas of instrumentation, informatics, and databases. This advance then contributes to the spread of data science applications in the industrial sector. One sub-area of data science concerns machine learning methods, which are usually simple to implement and are directly obtained from data sets. This work explored a supervised k-nearest neighbours (KNN) classification problem, focusing on gas emissions that pollute human health and the environment. Specifically, it proposes the construction of a soft sensor based on data for monitoring and, consequently, controlling sulphur dioxide (SO_2) emissions in chemical recovery boilers in the kraft pulp industry. Regarding the methodology, the behaviour of the method with one input variable, with subsets of predictors, and with an ensemble learning, for six classes of SO_2 was presented. The results presented are satisfactory, which is important to generate confidence in industrial implementations. The ensemble learning showed the best performance, with accuracy of 92% and geometric mean of 94.75% over the (independent) test set.

Keywords: Soft sensor; Kraft boiler; Sulphur dioxide; k-Nearest Neighbours (KNN); Ensemble Learning; Pulping process; Real case study.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

EBNN – Rede Neural de Camada Central Aprimorada (Enhanced Bottleneck Neural Network)

ETL – Extração, Transformação e Carregamento (Extract, Transform and Load)

ICA – Análise por Componentes Independentes (Independent Component Analysis)

k-NN – k-vizinhos mais próximos (k-Nearest Neighbours)

KPCA – kernel por componentes principais (Kernel Principal Component - Aanalysis)

LHT – Tratamento térmico (Liquor Heat Treatment)

LSTM – Memória de curto prazo longo (Long-Short Term Memory)

MSE – Erro Quadrático Médio (Mean Squared Error)

NNG – Garrote não negativo (NonNegative Garrote)

NOx – Óxidos de nitrogênio

PCA – Análise por Componentes Principais (Principal Component Analysis)

RBF – Função de base radial (Radial Basis Function)

SCR – Reator de redução catalítica seletiva (Selective Catalytic Reduction Reactor)

SVM – Máquina de Vetores de Suporte (Support Vector Machine)

SUMÁRIO

1. INTRODUÇÃO

Em 1972, foi criado pela ONU (Organização das Nações Unidas), o Programa das Nações Unidas para o Meio Ambiente (PNUMA), o qual marcou o início das discussões a nível global sobre proteção ambiental. Com o decorrer dos anos, essa preocupação foi se consolidando até o desenvolvimento de certificações de qualidade, como forma de estabelecer padrões, diretrizes claras e melhor controle dos efeitos que as indústrias causam no meio em que se situam. Entre essas certificações, há a série ISO 14000 (em português, Organização Internacional de Padronização série 14000), que zela pela gestão ambiental. Ela engloba desde o sistema de gestão a se utilizar até como tratar desvios do processo (DAMINELLI, 2019).

Além de certificar uma gestão padronizada, a ISO 14000 oferece um diferencial competitivo ao mercado. Junto a isso, o aumento de eventos globais para discussões climáticas e de sustentabilidade, direcionam o setor industrial a cada vez mais se atentar aos impactos ambientais, como emissões de gases nocivos ao meio ambiente (MENEZES et al., 2012).

Em se tratando do processo industrial, o controle dos resíduos se faz importante tanto para garantir a maior e mais eficiente produção final, quanto para atender as expectativas de sustentabilidade ambiental. Uma das formas de atender essa necessidade é o uso de sensores para a medição de variáveis-chave em campo, os quais proporcionam um diagnóstico contínuo e instantâneo sobre o processo, em comparação com amostras extraídas para análise em laboratório.

O aumento desses instrumentos nas indústrias vem gerando uma quantidade massiva de dados. E o armazenamento destes já não é mais uma preocupação, e sim, como usá-los ou como transformá-los também em produto. A partir dessas questões, investigações estatísticas cada vez mais complexas são usadas a fim de extrair informações dos processos reais. E a partir destas informações, é possível gerar conhecimento para tomadas de decisões multiformes, ou seja, com objetivos variados (ALMEIDA; PARK, 2017).

Em paralelo a essa necessidade, há os desafios dos modelos fenomenológicos, os mais presentes nas simulações e controle de processos

atuais. Essas dificuldades estão relacionadas, por exemplo, com a não linearidade e alta quantidade de variáveis, a interação entre componentes químicos, a dependência do processo com o tempo, o desgaste de equipamentos e entre outros. E esses desafios tornam a modelagem das operações complexa e cara, em relação ao tempo e ao conhecimento específico do processo necessários (LIMA et al., 2016).

Essas limitações e a alta geração de dados foram, como principais fatores, impulsionadores do uso e desenvolvimento de pesquisas em análise ou ciência de dados no campo da engenharia de processos industriais. Essa ciência já vem sendo amplamente usada na área de negócios e *marketing*, por exemplo. O setor econômico prevê uma transformação significativa no mercado em geral ao longo do tempo, após expressiva utilização de ferramentas digitais desenvolvidas especificamente para análise de dados. Gibert et al. (2018) afirmam que esse avanço também irá ocorrer no campo da engenharia e meio ambiente.

Um exemplo de que já está acontecendo é apresentado por Yun et al., (2021), que compararam a atuação de um modelo fenomenológico e um modelo por rede neural artificial na predição da concentração de micro poluentes, oriundos da aplicação de pesticidas em plantações agrícolas. Eles apresentaram as limitações e simplificações do modelo fenomenológico e a alta precisão da rede neural, mesmo com diferentes tipos de pesticidas usados nas plantações.

Já no campo de controle ambiental, a literatura reporta estudos como o de Kumar, Pandey e Sarkar, (2019), que utilizaram diferentes bancos de dados de qualidade do ar ambiente para comparar duas abordagens: aprendizado profundo (rede neural artificial) e por valores limites, para a detecção de poluição aceitável ou não. Mele e Magazzino, (2020) também utilizaram rede neural artificial, do tipo LSTM (do inglês, *Long Short Term Memory*), para estudar a relação entre o desenvolvimento das indústrias de ferro e aço com a poluição do ar e o crescimento econômico do país. Eles reportaram que o decréscimo na poluição por emissão das indústrias está diretamente ligado com o fortalecimento de responsabilidade sustentável do país em questão.

Na representação de processos contínuos, lidar com uma base de dados crua se mostra um desafio devido a fatores, como por exemplo, informações redundantes ou irrelevantes, amostras em tempos não sincronizados, diferentes tipos de variáveis e diferentes fontes de dados, como análises de laboratório e

sensores de campo. Ainda, uma característica inerente dos dados de processo é a dispersão, que é função dos dispositivos de medição, dos equipamentos e do próprio processo. Devido a todos esses fatores, a análise de dados requer investigação minuciosa, antes de ser capaz de gerar qualquer informação relevante (LIMA et al., 2016).

Com isso, Pani e Mohanta (2011) revisaram diferentes técnicas para diferentes problemas enfrentados no pré-processamento dos dados, ou seja, na etapa anterior à obtenção do modelo. Esses autores apresentaram métodos heurísticos, bayesianos, gráficos, entre outras ferramentas, para tratar valores discrepantes, nulos, colinearidades entre variáveis e a normalização de variáveis. Os exemplos relatados no trabalho têm em comum o uso dos dados para a construção de sensores virtuais, conforme o objetivo do presente do trabalho.

Os sensores virtuais são modelos matemáticos preditivos, em tempo real, que utilizam medições de sensores físicos, do processo, como variáveis de entrada. Eles podem complementar análises de laboratório antecipando medições para a tomada mais rápida de decisões, ou, em conjunto com medições por sensores físicos, podem apresentar informações sobre as operações em diferentes pontos do processo (LOTUFO; GARCIA, 2008). Uma vantagem do sensor virtual em relação ao físico é a redução de manutenção local e de calibração do aparelho (sem necessidade de parar o processo) e, por conseguinte, de material de reposição (KANO; FUJIWARA, 2013).

Nesse sentido, tem-se o trabalho de Sainlez e Heyen (2013) comparando diferentes técnicas no desenvolvimento de um sensor virtual para emissões de óxidos de nitrogênio em uma caldeira de recuperação química do processo de celulose *kraft*. Eles apresentaram rede neural artificial, árvore de decisão, regressão linear múltipla e composições entre essas técnicas, para o treinamento e teste dos modelos de regressão.

Como brevemente descrito, algumas pesquisas foram e estão sendo desenvolvidas em processos industriais com técnicas de Ciência de Dados. E a grande maioria delas, principalmente em relação a emissões de gases, empregam redes neurais artificiais. Os exemplos com outras técnicas, como o algoritmo *k*-vizinhos mais próximos (k-NN, do inglês *k-nearest neighbours*), são apresentados para modelos de detecção de falha, ou seja, sob condição

operacional normal ou de falha (CHENG et al., 2019; SONG et al., 2020). Assim, este trabalho visa apresentar um sensor virtual para a classificação de emissões de dióxido de enxofre a partir de uma caldeira industrial. O número usual de classes nas aplicações industriais é igual a dois. Neste trabalho, utilizaram-se seis classes, de modo a se ter um melhor reconhecimento a respeito do ponto operacional do processo.

2. OBJETIVO

2.1 OBJETIVO GERAL

O objetivo deste trabalho é construir um sensor virtual para a classificação de níveis de emissões em uma caldeira de celulose *kraft*.

Essa concepção empregará técnicas de Ciência de Dados, a fim de potencializar o processo de tomada de decisão em relação à melhoria de processo. Esse sensor virtual será baseado no algoritmo denominado k-vizinhos mais próximos (k-NN; *k-Nearest Neighbours*). O estudo de caso é referente às emissões de SO_2 da caldeira de uma fábrica de celulose *kraft* no Brasil.

2.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos são:

- Análise do comportamento dos dados coletados, a fim de identificar correlações e particularidades entre as variáveis de processo e a variável de interesse, a “concentração de dióxido de enxofre”;
- Obtenção de modelos k-NN para cada variável como única preditora, a fim de avaliar o potencial individual de classificação para a variável de interesse;
- Obtenção de modelos k-NN a partir de prévia seleção de sub-conjuntos de variáveis de entrada;
- Obtenção de comitês de modelos k-NN através da abordagem de *ensemble learning*, utilizando diferentes sub-conjuntos de variáveis de entrada;
- Comparativo entre as três diferentes abordagens descritas anteriormente;
- Análise e seleção do modelo k-NN com melhor desempenho para ser usado como sensor virtual em relação à classificação da variável de interesse.

3. REVISÃO BIBLIOGRÁFICA

3.1. CIÊNCIA DE DADOS APLICADA À INDÚSTRIA QUÍMICA

A Revolução Industrial no século XVIII trouxe a realidade de máquinas como expansores da produção, e deu ao mundo a expectativa da tecnologia como meio de alcançar maiores taxas de produtividade. Não só no meio industrial, mas também em transportes, comunicação, entretenimento, educação e outras áreas que, direta e indiretamente, contribuíram para a expansão do setor industrial. Nos séculos seguintes, houve o uso do aço e eletricidade marcando a segunda revolução e a automação, a terceira. Todas essas revoluções foram acompanhadas de pesquisas científicas. O século XXI traz a quarta revolução industrial baseada na digitalização dos processos, ou seja, na disponibilização do controle da produção de forma virtual e online (SANTOS; SANTOS; SILVA JUNIOR, 2019).

A automação permitiu uma geração contínua e precisa de dados e, junto aos computadores e a Internet, o seu armazenamento em grande escala. Assim, veio a necessidade de transformar esses dados em informações, e essas informações em conhecimento, a serem utilizadas como suporte à tomadas de decisões mais inteligentes ou racionais (ALMEIDA; PARK, 2017). Isto é, agregando conhecimentos computacionais (programação) e ferramentas estatísticas, os dados gerados têm o poder de prover uma visão diferente, até então, do processo. E o aumento massivo de dados ao longo das últimas décadas, e a necessidade de correlacioná-los, levou pesquisadores e profissionais em geral a desenvolverem diversas técnicas que podem ser reunidas em três áreas principais, estatística, inteligência computacional, e processamento de sinais. São exemplos de técnicas: análise por componentes principais, redes neurais artificiais, árvores de decisão, análise de regressão, modelo oculto de Markov, máquina de vetores de suporte, entre outras.

A Figura 1 apresenta um exemplo do ciclo da Análise de Dados, que se inicia na definição do problema ou o que se deseja transformar com a Ciência de Dados. Na sequência, tem-se as etapas de Extração, Transformação e Carregamento (ETL, do inglês, *Extract, Transform and Load*), a fim de identificar

quais tipos de dados são necessários e então de extraí-los de sua fonte original para compor a base de dados que será utilizada no projeto; se necessário, faz-se transformações (como o que acontece em dados de imagens). A terceira etapa denomina-se Análise Exploratória de Dados (AED), na qual busca-se entender o comportamento das variáveis, possíveis correlações entre elas, e o seu nível de representatividade no processo a ser descrito. Também é onde se faz um tratamento sobre valores nulos e discrepantes. A segunda e terceira fases são as que demandam maior tempo, pois referem-se à montagem e preparação da base de dados, e quanto melhor executadas, maior a chance de maior eficiência do modelo.

Figura 1 - Ciclo da Análise de Dados. Fonte: (*SLIDE TEAM*, 2020) modificado.

Um exemplo de pesquisa concentrada nessas etapas é a de Ajami e Daneshvar (2012), que utilizaram PCA (Análise por Componentes Principais) e ICA (Análise por Componentes Independentes) para selecionar um subconjunto de variáveis com maior influência, a fim de detectar e diagnosticar falhas em turbinas em uma planta térmica. Além de identificar as variáveis com maior influência sobre a operação, esses autores detectaram os pontos de falhas e diminuíram os ruídos da resposta.

A última etapa diz respeito à geração de modelos candidatos, em que métodos de aprendizado de máquina serão testados e comparados, a fim de se selecionar aquele de melhor desempenho. Com a implementação, enxergam-se

novos pontos e lacunas a serem estudados e melhorados; por isso, se tratar de um ciclo (BONTHU; HIMA BINDU, 2018).

Inicialmente, foram investigados diversos trabalhos que aplicam a Ciência de Dados em processos químicos em geral. Por exemplo, Osorio et al. (2008) desenvolveram um sensor virtual para determinar a concentração de álcool destilado a partir de quatro medidas de temperatura usando uma rede neural, que se mostrou um método mais simples do que relacionar as equações da coluna de destilação com o processo físico de separação de líquidos.

Seguindo a linha de estatística multivariada, Bouzenad e Ramdani (2017) combinaram PCA não-linear e uma rede neural de gargalo (EBNN; *Enhanced Bottleneck Neural Network*) para o manuseio de dados com comportamento não-gaussiano. Esse modelo gerou os parâmetros de entrada para um sensor virtual voltado à detecção de falhas, com a redução da taxa de alarmes falsos.

Com o intuito de controlar uma fermentação em batelada, Li, Meng e Song, (2016) aprimoraram um algoritmo duo-heurístico, baseado na diferença de mínimos quadrados e gradiente descendente, para a correção de modelos da alimentação de um reator, de forma mais rápida em relação a outros algoritmos. Já Bo et al. (2010) combinaram PCA, um modelo de árvore binária, e máquina de vetores de suporte (SMV, do inglês, *Support Vector Machines*), para diagnosticar falhas em um processo de destilação na indústria de butadieno. Os autores reportaram que o modelo se mostrou adequado para processos de destilação em geral.

Bachnas et al. (2014) utilizaram uma coluna de destilação de alta pureza para testar e analisar diferentes modelos com variação de parâmetros lineares (LPV, do inglês *Linear Parameter Variation*), a partir de duas abordagens. Primeiramente, por interpolação linear com tempo invariável, em dados locais, demonstrando ser uma boa opção para pequenas perturbações no processo. A segunda abordagem foi por parametrização da variação dos parâmetros por uma base de dados global. Já Neves et al. (2018) focaram na destilação extrativa de etanol, a partir de um estudo simulado, para desenvolver um sistema de inferência de uma variável de interesse, seguido de seu controle com duas redes neurais recorrentes, uma para cada faixa do processo. Canete et al. (2012) partiram de um mesmo processo simulado por Neves (autor citado anteriormente); porém utilizaram PCA para reduzir a dimensionalidade dos

dados, rede neural artificial, para predizer as composições de saída da coluna e modelar o processo, e uma rede neurogenética para o seu controle.

Outro exemplo é o estudo de Adams et al. (2020), que analisaram a variação na emissão dos gases óxidos de enxofre e nitrogênio a partir da variação de conversão do combustível, em uma caldeira de leito fluidizado. Segundo os autores, os resultados apresentaram eficiências de 89% e 99% nos modelos preditivos para óxidos de enxofre e nitrogênio, respectivamente.

3.2. SENSOR VIRTUAL

Um sensor virtual é um modelo preditivo, desenvolvido a partir de uma quantidade massiva de dados medidos em um processo industrial. A principal função desse sistema é a predição *online* de uma variável, que é significativa para o controle de qualidade do produto ou de segurança do processo. Esse tipo de sensor vem sendo desenvolvido a décadas pela razão de não haver um modelo físico adequado ou que esse seja de alto custo à indústria (THAM et al., 1991). Como exemplo, no mercado brasileiro atual, encontra-se disponível um modelo de sensor físico para a medição do teor de dióxido de enxofre, que é indicado para indústrias petroquímicas, metalúrgicas, de mineração e agricultura. Comparando com um medidor de pressão – um dos mais comuns na indústria - com as mesmas especificações de temperatura e pressão suportadas, o sensor de dióxido de enxofre pode chegar a ser nove vezes mais caro (ALIBABA, 2021).

Devido a isso, o mercado de sensores e medidores industriais têm a necessidade de alternativas, como são os sensores virtuais, onde técnicas de estatística e de aprendizado de máquina são usadas para a sua construção. Yang et. al (2020) utilizaram SMV para predizer a concentração de NO_x na entrada de um reator de redução catalítica seletiva (SCR, do inglês *selective catalytic reduction reactor*). Foram consideradas como variáveis de entrada para o modelo, o tempo de residência e o estado dinâmico do processo, desenvolvendo-se um sensor de predição em tempo real com altas acurácia e capacidade de generalização, segundo os autores.

Estudos como esses, além de desenvolverem métodos, também contribuíram para disseminar a aplicação de sensores virtuais. Por exemplo,

Kadlec, Grbic e Gabrys (2009) apresentaram a detecção e o diagnóstico de falhas das faixas de operação de um processo como outra função para os sensores. Os autores discutiram também sobre os desafios enfrentados na construção desses modelos, como por exemplo, o grande volume de dados, mas com pouca representatividade sobre as informações do processo industrial.

Um sensor virtual pode ser modelado a partir de dados históricos, ou seja, no modo *offline*. Mas, para isso, a base de dados deve conter os principais cenários do processo, tanto de condição operacional normal quanto de desvios em relação à normalidade. Esses desvios ocorrem, por exemplo, devido a mudanças no ambiente externo e em correntes de alimentação, desgastes em componentes, e variações de processo. Outra dificuldade se faz na seleção dos parâmetros do modelo, a fim de que esse compreenda todas as diferentes condições do processo. Há ainda a variação de tempo que todo processo apresenta em relação à frequência de coleta de dados, necessitando assim, de uma estratégia para o seu uso *online* (KADLEC; GABRYS; STRANDT, 2009).

Entre publicações dessa área, tem-se, por exemplo, Liu e Xie (2020), que compilaram vários modelos baseados na função *kernel*, com aplicações nas etapas de pré-processamento, seleção de amostras e variáveis, construção do modelo, e análise de confiabilidade dos sensores virtuais. A discussão vai além de uma revisão sobre predições de variáveis de difícil medição, mas também para a detecção de falhas e o controle avançado em processos industriais.

Investigando trabalhos de sensores virtuais para a detecção de gases nocivos, tem-se por exemplo, Iliyas et al. (2013), que usaram dados simulados de uma caldeira a gás natural para treinar uma rede neural de função de base radial (RBF, do inglês *radial basis function*), a fim de inferir a emissão de óxidos de nitrogênio e oxigênio. Eles testaram o modelo obtido com dados industriais reais, atingindo igual performance de um analisador físico; porém, com menores custo e manutenção do equipamento.

Sun et. al (2019) apresentaram um modelo para o processo de dessulfurização em uma termoelétrica. A proposta foi de uma rede perceptron de múltiplas camadas, composta com o método Garrote não negativo (NNG, do inglês, *nonnegative garrote*), para maior redução dos pesos na rede e otimização extrema para seleção de variáveis locais. Os autores compararam os resultados

com aqueles de outros modelos, mostrando melhor acurácia para o modelo proposto.

Como apresentado nessa seção, sensores virtuais são utilizados, principalmente, como medidores de concentração, detecção e diagnóstico de falhas. Os modelos desenvolvidos a partir de dados enfrentam dificuldades geradas pelas características das bases de dados. Com isso, uma etapa essencial em qualquer tarefa de análise de dados é o pré-processamento, de modo a se ter um conjunto de dados de trabalho, conforme o objetivo da aplicação. Curreri, Graziani e Xibilia (2020) fizeram um trabalho apontando a seleção de variáveis como a etapa mais importante durante o pré-processamento. Eles apresentaram diferentes técnicas de medição de grau de correlação, de extração e de seleção de características, tanto para modelos orientados a dados, quanto fenomenológicos. Além disso, os autores propuseram uma abordagem híbrida e realizaram testes com dados industriais reais.

3.3. k-VIZINHOS MAIS PRÓXIMOS

No aprendizado de máquina, há três principais grupos de métodos de aprendizagem, aprendizado supervisionado, aprendizado não-supervisionado, e aprendizado por reforço. No primeiro, o computador usa, além das variáveis de entrada, também aquelas de saída (supervisão). No segundo, não há variáveis de saída ou rótulos para os dados (sem supervisão). No terceiro, a máquina interage com o ambiente, e a cada passo do programa, há uma recompensa ou punição, mostrando acertos e erros (RUSSELL et al., 2016).

O algoritmo k-vizinhos mais próximos utiliza aprendizado supervisionado de forma simples e prática, como citam Feng e Li (2020). A Figura 2 apresenta um exemplo. Ela correlaciona duas variáveis representadas pelos eixos da abscissa e coordenada, em que os dados foram divididos em dois grupos, A e B (pontos pretos e cinzas, respectivamente). O método está avaliando o ponto tracejado, a fim de classificá-lo.

Portanto, para k igual a um (menor círculo), tem-se apenas um ponto da classe A, e assim, o modelo classifica o ponto tracejado como sendo de classe A. Com k igual a três (círculo mediano), consideram-se duas amostras da classe

B e uma da classe A, e com isso o novo ponto pertencerá à classe B. Já para k igual a cinco, há três pontos da classe A e dois da classe B dentro do círculo maior; por conseguinte, a nova amostra é classificada como sendo da classe A.

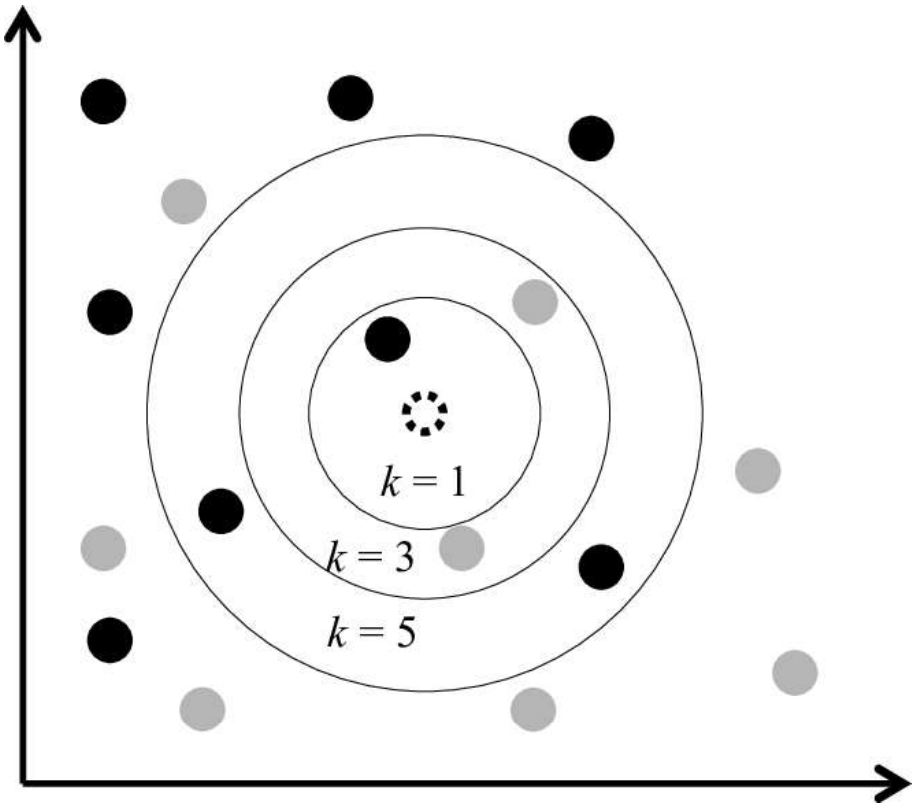


Figura 2 – Exemplo de funcionamento do modelo k-vizinhos mais próximos (ALMEIDA; BALANCO; DANTAS, 2016).

Além de implementação simples, esse método não necessita de informações prévias dos dados, ou seja, é um algoritmo não paramétrico. Ele não desenvolve modelos explícitos como de regressões lineares; e pode ser usado para problemas de regressão e classificação, entre outras vantagens. Porém, há desvantagens, como por exemplo, alta sensibilidade para valores discrepantes e para diferentes grandezas das variáveis, e aumento de esforço computacional a partir do aumento da quantidade de dados e de dimensionalidade (BISHOP, 2006).

A seguir, tem-se alguns exemplos de aplicações de k-NN. Harrou et al. (2020) apresentaram o método em conjunto com o gráfico de controle de Shewhart, a fim de monitorar tráfegos de trânsito; e Sarmadi e Karamodin (2020), para monitoramento na área de saúde. Madeti e Singh (2018) aplicaram k-NN para classificar falhas em uma planta fotovoltaica para produção de energia

elétrica. Notam-se distintas áreas de aplicabilidade do método, cujas adaptações são de acordo com as características do conjunto de dados e do comportamento dos sistemas de interesse.

Um exemplo de aplicação na área de engenharia é aquele de Pandya et al. (2013), que desenvolveram um detector de falhas para rolamentos de motores a partir da caracterização de efeitos sonoros. Os autores usaram sete métodos diferentes para classificar as falhas, sendo o k-NN o mais eficiente, que, com adaptações, alcançou uma eficiência superior a 96%. Outro exemplo em que o método k-NN apresentou melhores resultados foi aquele publicado por Song et al. (2019) utilizando dados com multimodos de operação. Primeiro, os autores normalizaram os dados para média igual a zero e desvio padrão igual a um. Em seguida, padronizaram os dados pelos k-vizinhos mais próximos, de acordo com os modos de operação, para após a detecção da falha, sinalizar a variável que causou o distúrbio. O modelo foi denominado k-vizinhos mais próximos padronizados (SkNN, do inglês, *standardized k-nearest neighbours*).

Outras publicações relevantes em Engenharia Química foram a de Yang et al. (2017) e de Balram, Lian e Sebastian (2020). No primeiro trabalho, utilizou-se um coeficiente de correlação para separar as variáveis entre correlacionadas e independentes, e aplicaram o algoritmo k-vizinhos mais próximos para as últimas variáveis e análise por componentes principais via *kernel* (KPCA; *kernel principal component analysis*) para as primeiras variáveis, compondo um sensor virtual multifuncional para detecção de falhas. Já no segundo trabalho, desenvolveu-se um sistema de alerta sobre a qualidade do ar utilizando k-vizinhos mais próximos, com uma das variáveis sendo concentração de ozônio ao nível do solo, calculada a partir de um sensor virtual, além de concentrações de dióxido de enxofre, dióxido de nitrogênio, monóxido de carbono e material particulado.

Jain e Lella (2020) também propuseram um modelo de k-vizinhos mais próximos para a predição da concentração de óxidos de nitrogênio; porém, com pesos atribuídos a partir do coeficiente de correlação de Pearson, que apresentou menores erros que o mesmo método sem a atribuição de pesos. Rezzadeh et al. (2021) apresentaram um sensor virtual para a predição de emissões de NO_x em turbinas a gás natural de uma planta de geração de energia elétrica, usando k-NN como técnica de regressão. Os autores obtiveram um

coeficiente de determinação (R^2) de 0,8634 para os dados de teste de todos os anos (2011 a 2015) e para os modelos anuais, o melhor obteve R^2 igual a 0,9244.

Nota-se que a maioria das pesquisas com k-NN foram, ou com abordagem de regressão ou com apenas duas classes (Sim e Não). Sharma et al. (2021) propuseram o desafio de usar a técnica para seis classes. Eles compararam diversas técnicas de classificação na predição do índice de qualidade do ar na Índia, e o k-NN se destacou com uma acurácia de 97,92%.

3.4. PROCESSO *KRAFT* DE PAPEL E CELULOSE

Para a produção de papel e celulose tem-se quatro tipos de processos de polpação: o mecânico (ou termomecânico), o termoquímico mecânico, o semiquímico e o químico (BAJPAI, 2010). A diferença entre eles define a aplicação do papel resultante. O estudo de caso neste trabalho diz respeito ao processo químico *kraft*, que a Figura 3 apresenta de forma geral.

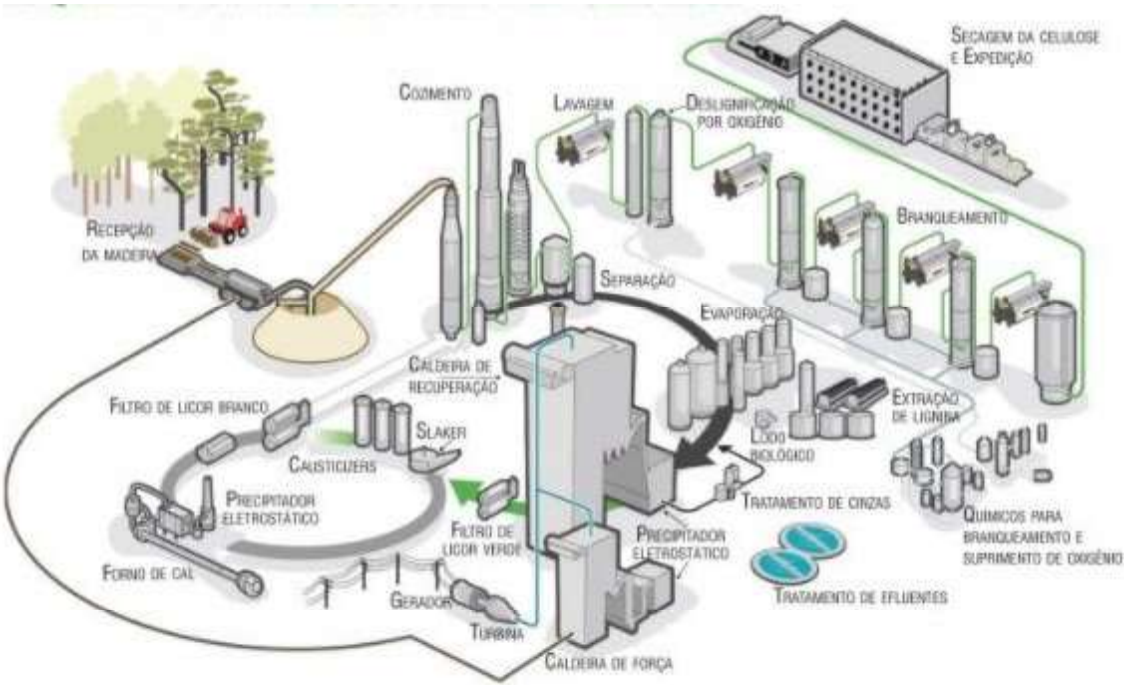


Figura 3 – Ilustração do Processo *Kraft*. Fonte: (TIMMER, 2020).

O processo inicia-se na recepção da madeira e seu corte em partes menores, chamadas de cavacos, para melhor rendimento da reação química de deslignificação. Em seguida, tem-se o cozimento desses cavacos junto com o

licor branco, que é uma solução aquosa rica em hidróxido de sódio (NaOH) e sulfeto de sódio (Na_2S), responsável pela solubilização da lignina e consequente liberação das fibras de celulose e hemicelulose. O produto resultante é composto por fibras suspensas em um líquido denominado licor preto. Após uma etapa de lavagem, a polpa celulósica seguem para a etapa de secagem e enfardamento, para venda, ou segue para a etapa de branqueamento e fabricação de papel, em caso de fábrica integrada (TIMMER, 2020).

Já o licor preto residual da etapa de polpação dos cavacos de madeira é enviado para o ciclo de recuperação química, composto pelas etapas de evaporação, incineração e caustificação. Ao final, tem-se a recuperação do licor branco de cozimento, responsável pela etapa de polpação dos cavacos de madeira (TIMMER, 2020). Em suma, a recuperação química é um ciclo fechado, com baixas perdas e geração de energia térmica e elétrica. Nas seções a seguir, tem-se a descrição de cada etapa, de forma mais completa.

3.4.1. Preparo da Madeira

O preparo da madeira é composto por quatro subprocessos: descascador, lavagem, picador e seletor. O primeiro retira as cascas das toras de madeira, pois estas possuem baixo teor de fibras e só pioram a qualidade do produto, sendo usadas na produção de energia por queima. Em seguida, as toras são lavadas, a fim de eliminar qualquer impureza para o processo, como terra, folhas, entre outros (CASTRO, 2009).

No picador, a madeira lavada é triturada em medidas calculadas, que proporciona melhor reação em menor tempo de cozimento. Esses pedaços são chamados de cavacos, e medem entre 15 e 20 mm de comprimento e entre 3 e 6 mm de espessura. Na última etapa, os cavacos são selecionados de acordo com o tamanho, através de peneiras vibratórias. Aqueles com tamanho inferior ao padrão são enviados junto com as cascas para queima em uma caldeira auxiliar (PAULA, 2017).

3.4.2. Etapa de Cozimento

A etapa de cozimento pode ser em regime batelada ou contínuo, sendo a última a mais usada. No digestor, há zonas de temperatura crescentes até o cozimento propriamente dito, onde se mantém entre 140 e 180 ºC. O licor branco, ou seja, a solução aquosa rica em hidróxido de sódio e sulfeto de sódio, age rompendo as ligações da lignina. A soda cáustica é responsável pelo cozimento mais uniforme e menos drástico dos carboidratos. Já o sulfeto de sódio evita uma alta concentração da soda cáustica na fase de impregnação inicial do cozimento, evitando uma maior degradação das fibras de celulose (PAULA, 2017).

Segundo Carvalho (1999), a duração do cozimento depende do grau de deslignificação que se pretende atingir, geralmente traduzido pelo número *kappa*. O número *kappa* aponta o teor de lignina presente na pasta celulósica, ou seja, a facilidade de branqueamento da polpa marrom. No caso do Eucalipto, o tipo de madeira mais consumido no Brasil pelo processo *kraft*, o número *kappa* ideal varia entre 14 e 20, com álcali residual de 5 g/L a 10 g/L para evitar reprecipitação de lignina na superfície das fibras, sendo a relação entre o volume de licor e a massa de madeira seca, de 3,5:1.

Ao final do cozimento, no fundo do digestor, ocorre a diluição do licor preto fraco, oriundo da pré-lavagem e da descarga da polpa no tanque de descarga, onde a despressurização da polpa permite que as fibras fiquem suspensas na solução.

3.4.3. Etapa de lavagem

O produto resultante do cozimento é transferido para o tanque de descarga. A polpa marrom é então enviada para o sistema de depuração, para a separação dos materiais estranhos às fibras por processo mecânico (como nós de madeira, pequenos palitos, cavacos não cozidos). O material de aceite é transferido para os filtros lavadores, que têm, por finalidade, remover o licor residual que poderia contaminar a pasta em processos subsequentes, recuperar o máximo de reagentes químicos com uma diluição mínima, e recuperar os constituintes da madeira dissolvidos no licor para utilizá-los como combustível.

A polpa resultante segue para o processo de fabricação de papel ou pode ser preparada para armazenagem, através da secagem e enfardamento. Em paralelo, a solução residual do cozimento, denominada de licor preto fraco, contendo entre 12 e 20% de sólidos, é encaminhada para a etapa de evaporação, a fim de se elevar a sua concentração de sólidos acima 65% em massa (BAJPAI, 2010).

3.4.4. Etapa de Evaporação

Essa operação é separada em três fases. Primeiro, há a evaporação de parte da água contida no licor preto residual, formando condensado e licor preto mais concentrado. Evaporadores de múltiplo-efeito são normalmente usados nesta etapa, por reaproveitarem o evaporado produzido em um efeito (tanque) como meio de aquecimento para o efeito seguinte (BAJPAI, 2010).

Para reduzir a viscosidade do licor preto, devido ao aumento de sua concentração, um processo de tratamento térmico (LHT, do inglês, *liquor heat treatment*) pode ser adicionado. Esse procedimento trata o licor à elevadas temperaturas por um longo período para que as moléculas pesadas de polissacarídeos e lignina sejam quebradas e a viscosidade seja reduzida (RAUSCHER; KAILA; JAAKKOLA, 2006).

3.4.5. Caldeira de recuperação química

O licor preto chega à caldeira de recuperação química com pelo menos 65% de teor de sólidos, em que 1/3 são compostos inorgânicos e 2/3 são orgânicos dissolvidos, aproximadamente. O processo na caldeira é complexo, envolvendo diversas reações físico-químicas em diferentes regiões. As três principais etapas na caldeira são: combustão do material orgânico para a geração de energia elétrica; redução dos compostos inorgânicos para a redução do licor branco de cozimento, como compostos de enxofre para sulfeto de sódio; e recuperação de compostos inorgânicos na sessão superior do equipamento, em função de seu valor econômico.

A Figura 4 esquematiza uma caldeira de recuperação química.

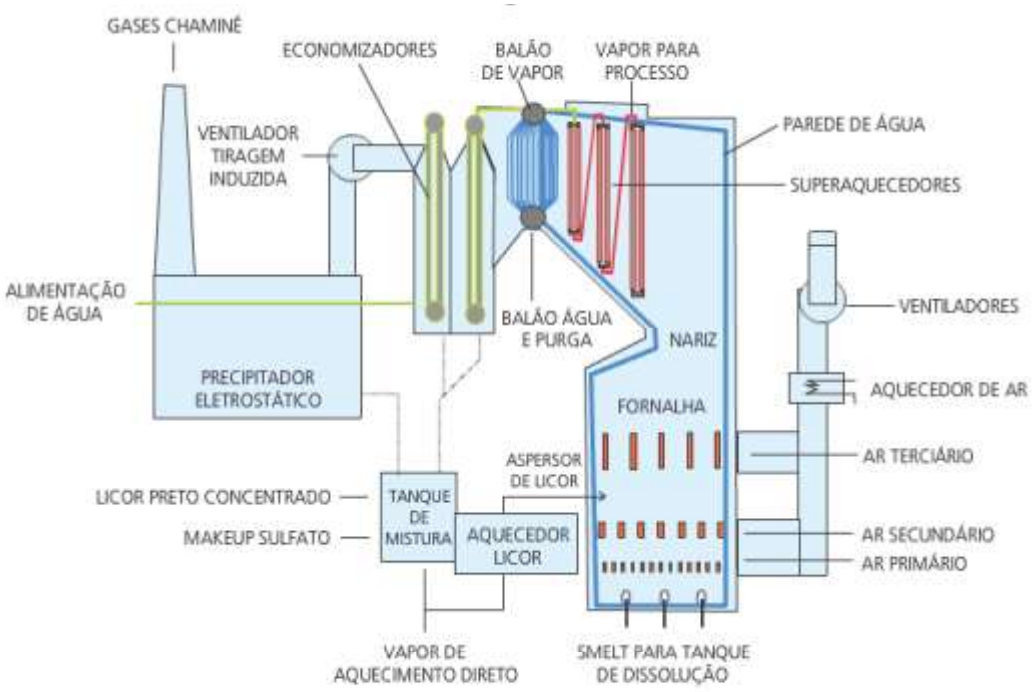


Figura 4 – Modelo esquemático de uma caldeira de recuperação química. Fonte: (SILVA, 2016).

A fornalha contém a zona de oxidação na parte superior e de redução na parte inferior. O licor preto é introduzido nessa última e o ar de combustão é introduzido em três (ou quatro) níveis a diferentes temperaturas e vazões. O ar primário é responsável por uniformizar o leito, e principalmente conduzir a redução dos compostos inorgânicos; já o ar secundário, controla a altura desse leito reacional e realiza o processo de combustão do licor. Por último, o ar terciário finaliza o processo de combustão do licor e evita o arraste de compostos para a sessão superior da caldeira (VAKKILAINEN, 2000).

Na zona da oxidação, enxofre é oxidado a dióxido de enxofre, que reagindo com sódio, formam sulfato de sódio e, de forma indesejada, o sulfeto. Quando isso ocorre, o sulfeto de hidrogênio também é formado, e uma parte é carregada com os gases (especialmente se a alimentação de ar for insuficiente ou incompleta). Uma maior quantidade de sólidos secos leva a maiores temperaturas na caldeira, e assim, menor emissão de sulfeto de hidrogênio e maior de sódio. Por conseguinte, tem-se mais enxofre se ligando a sulfato de sódio e menor emissão de SO_2 (BAJPAI, 2010). Ari e Tarja Tamminen (2015) apresentaram uma explicação sobre as reações envolvendo a emissão de dióxido de enxofre na caldeira e mostraram que a temperatura do leito é

fundamental para o nível de emissão. Porém, a temperatura também depende dos fluxos de entrada, tanto do licor preto quanto do ar de combustão (nesse caso, ar secundário).

É importante ressaltar que durante os processos físico-químicos, há a liberação de gases, como SO_x, CO_2, CO, CH_4, H_2O e TRS (enxofre total reduzido). Esse último sofre oxidação com o ar terciário, também produzindo SO_2. O controle da produção desses gases é importante não só devido a emissão para a atmosfera, mas também a fim de se evitar incrustações nos superaquecedores, que são trocadores de calor na sessão superior do equipamento (COSTA; BISCAIA; LIMA, 2008).

Os sais fundidos são resultantes do processo da queima e constituem o *smelt* (termo em inglês que representa a extração de metais por processo de aquecimento e fusão), que é rico em sulfeto de sódio e carbonato de sódio. O *smelt* flui para o tanque de dissolução por meio de bicas resfriadas na parte inferior do equipamento. Neste tanque, provido de agitação, resulta o licor verde que possui esta cor devido aos sais ferrosos formados (PAULA, 2017).

3.4.6. Etapa de Caustificação

Nesta etapa, o carbonato de sódio (Na_2CO_3) no licor verde é convertido em hidróxido de sódio (NaOH) e carbonato de cálcio ($CaCO_3$; ou lama de cal), a partir de reação com hidróxido de cálcio ($Ca(OH)_2$), que é oriundo da reação de apagamento do óxido de cálcio (CaO) com água (DARÉ ALVES et al., 2015).

O licor resultante desta etapa, rico em NaOH e Na_2S, é clarificado, ou seja, lavado a fim de ser recuperado e então retornado ao digestor, fechando-se o ciclo de recuperação dos compostos inorgânicos da etapa de cozimento. Quanto maior a atividade do licor branco, menor a quantidade de inertes no digestor e para a caldeira de recuperação, aumentando a produção de celulose (MORAES, 2011). A lama de cal extraída é filtrada e encaminhada para o forno de cal.

3.4.7. Forno de Cal

A lama de cal ou carbonato de cálcio se junta com calcário no forno de cal para a regeneração do óxido de cálcio. A alimentação no forno passa por regiões de aquecimento a fim de favorecer a evaporação da água da mistura até que chegue na região de calcinação. Nessa região, o carbonato de cálcio é convertido em óxido de cálcio (cal recuperada). Essa cal reagirá com a água para a formação de hidróxido de cálcio, que será reutilizado na caustificação do licor verde (PAULA, 2017). Com isso, fecha-se o ciclo de recuperação da cal.

Segundo Bajpai (2010), a maior fonte de emissões de gases na indústria de Papel e Celulose, do tipo *Kraft*, é a caldeira de recuperação química, e grande parte das emissões são de dióxido de enxofre. Devido a isso, se faz necessário um constante aperfeiçoamento do processo a fim de se manter a sua produção; porém, com maior sustentabilidade.

4. METODOLOGIA

O presente trabalho analisa dados industriais de pressão, temperatura, vazão e concentração, a fim de descrever o processo. Assim, o banco de dados inicialmente coletado é denominado de conjunto de dados crus. Esse conjunto foi coletado em um certo período, independentemente do estado do processo, o que acarreta muitos valores nulos ou não representativos. Devido a isso, se faz necessário preparar essa base de modo a se obter informações relevantes, ou seja, um pré-processamento com técnicas selecionadas de acordo com o perfil dos dados e o objetivo do trabalho.

Após a etapa de pré-processamento, o conjunto resultante de dados deve estar adequado para a construção do modelo do processo; nesse caso, um sensor virtual para as emissões de SO_2. Essa etapa é denominada de processamento. Ao seu final, ainda pode ser necessário uma etapa de pós-processamento, para melhor validação dos resultados.

Destaca-se que toda a metodologia foi construída utilizando-se a linguagem de programação Python (VAN ROSSUM, 1995), e o Jupyter Notebook como ambiente de desenvolvimento integrado (IDE, do inglês *Integrated Development Environment*) (KLUYVER et al., 2016). Nas subseções a seguir, as técnicas utilizadas nessas três etapas de análise de dados são descritas.

4.1 ETAPA DE PRÉ PROCESSAMENTO

Para o pré-processamento dos dados, foram utilizadas técnicas de visualização, de identificação e exclusão de dados anômalos e nulos, de rotulação das categorias da variável resposta, e de seleção dessas para o modelo final, conforme descrição a seguir.

4.1.1. Visualização dos Dados

Antes de qualquer tentativa de desenvolvimento de modelos com os dados, é necessário determinar o seu comportamento e interações entre variáveis. Esse trabalho utilizou técnicas de visualização, como gráficos de tendência, dispersão e matriz de correlações.O gráfico de tendência relaciona os

dados ao longo do tempo, apresentando o comportamento temporal da variável. O gráfico de dispersão é útil para verificar o relacionamento entre pares de variáveis. Já a matriz de correlações exibe uma matriz de gráficos correlacionando as variáveis, a fim de apresentar alguma tendência entre os dados, e a diagonal principal traz gráficos de frequência para cada variável, o que pode revelar um comportamento específico, como o gaussiano por exemplo.

Com a análise gráfica, é possível reconhecer padrões nos dados que ajudem a modelá-los, de forma a obter melhores resultados com a aplicação do moledo. Porém, não se trata de modificações que alterem o processo que os dados representam, e sim que destaquem as características importantes de acordo com o objetivo do modelo. Neste trabalho, foram realizadas algumas tratativas que serão descritas no capítulo Resultados e Discussões.

4.1.2. Filtro de Hampel

Valores anômalos (do inglês, *outliers*) são observações que se desviam, significativamente, da grande maioria dos dados. Eles podem se originar devido a ruído nos sensores, perturbações no processo, degradação dos instrumentos e/ou erros humanos. Em geral, não é recomendável realizar uma análise quando os dados contêm anomalias, pois elas podem gerar um modelo com um desempenho inferior, estimações tendenciosas de parâmetros, e uma análise incorreta dos resultados (LIU; SHAH; JIANG, 2004).

Para medir a robustez de um estimador em relação a uma anomalia, Hampel (1971) introduziu o conceito de ponto de repartição (*breakdown point*), que corresponde à menor porcentagem de dados anômalos que podem causar valores tendenciosos em parâmetros. Ao final, para a identificação de dados anômalos, o autor apresenta a Equação (1) e (2), para a mediana, e o desvio absoluto da mediana (MAD) representada na Equação (3), em que n é o tamanho do conjunto de dados.

$$\text{mediana } \tilde{X} = X_{\frac{n+1}{2}},\ para\ n\ ímpar \tag{1}$$

$$mediana\ \tilde{X} = \frac{Xn_{/2} + Xn_{/2+1}}{2},\ para\ n\ par \tag{2}$$

$$MAD\ (\tilde{X}) = mediana(\left|x_1 - mediana\tilde{X}\right|, \ldots, \left|x_n - mediana\tilde{X}\right|) \quad (3)$$

Hampel sugere a mediana como estimador de posição central e o MAD, como estimador de dispersão. Assim, a observação (x) é identificada como anômala se a sua diferença em relação à mediana é maior do que MAD vezes uma constante, como apresentado na Equação (4), em que L representa uma constante de valor igual a 1,4826 (para distribuição aproximadamente normal) e t é o fator de correção. Um alto fator faz o filtro mais tolerável a dados discrepantes. Pearson (2002) considerou o Filtro de Hampel significativamente efetivo na prática. Esse filtro é aplicado para cada variável em separado. Se pelo menos uma observação, de uma variável em particular, é considerada um dado anômalo, o vetor de observações é classificado como um *outlier*.

$$\left|x_1 - medianaX_N\right| \geq L\ t\ MAD(X_N) \quad (4)$$

Após identificar os valores discrepantes, o tratamento deles pode se dar de diversas maneiras, como por exemplo, eliminação da observação, substituição pelo valor médio da variável ou média móvel e entre outros. Essa etapa dependerá do objetivo do trabalho e como os dados representam o processo (PEARSON, 2002).

4.1.3. Divisão entre conjuntos de Treinamento e Teste

Um modelo a ser construindo a partir de uma base de dados não deve apresentar problemas de subajuste ou de sobreajuste. Deve ser tal que explique a variável resposta de modo adequado, a partir de novas observações até então desconhecidas. Um modelo sobreajustado é aquele que explica parte do ruído contido no conjunto de dados de treinamento, o que reduz a sua capacidade de generalização, em relação ao conjunto de teste. Já um modelo subajustado não consegue explicar, minimamente, o comportamento da variável de interesse, sendo inadequado para predição (HASTIE; TIBSHIRANI; TIBSHIRANI, 2017).

De modo a mitigar esses problemas, divide-se o conjunto de dados, de forma aleatório, em subconjuntos de identificação (aproximadamente 75% dos registros), para treinamento do modelo, e de teste (do inglês, *Train/Test split*),

para avaliação dele. O primeiro é usado para estimar os parâmetros do modelo, e o segundo, para verificar a sua capacidade de generalização, de modo a evitar o sub- ou sobreajuste dos dados. Ainda, deve-se garantir que o maior e o menor valor de cada variável estejam no subconjunto de treinamento, a fim de se evitar extrapolações (JAMES et al., 2013).

4.2. ETAPA DE PROCESSAMENTO

Para o efetivo processamento dos dados, as técnicas utilizadas foram relacionadas ao problema em questão ou aos objetivos parciais desejados. Assim, na construção do sensor virtual, se tornou adequado o algoritmo k-vizinhos mais próximos.

4.2.1. k-Vizinhos mais Próximos

Um dos mais antigos e simples algoritmos para classificação é o k-vizinhos mais próximos (k-NN; *k-nearest neighbours*). Ele classifica uma observação para a classe que contém a maioria das k observações vizinhas, a partir do cálculo de uma métrica de distância (GUL et al., 2018). Ou seja, o método computa a distância de todas as amostras de treinamento para cada nova amostra, a fim de selecionar os k vizinhos mais próximos. Apesar de ser simples, k-NN gera resultados competitivos, e não raramente, até com melhor performance comparado a algoritmos de aprendizagem complexas (GOLDBERGER et al., 2004).

Porém, ele é influenciado negativamente por variáveis não informativas, em casos de alta dimensionalidade. Ou seja, o custo da complexidade do tempo linear sobre o tamanho da amostra limita a ação desse método para uma base de dados com muitas variáveis e/ou muitas amostras. Ainda, se uma das classes contém uma maior quantidade de valores do que as demais, o modelo pode se tornar tendencioso para essa classe majoritária. Outro ponto de atenção é que o valor de k depende das características dos dados e a sua determinação é geralmente a partir de testes. Em geral, um valor alto de k reduz o efeito do ruído; porém, torna as fronteiras de classificação mais complexas (DENG et al., 2016).

Neste trabalho, esse método foi programado a partir da biblioteca *Scikit-learn* (PEDREGOSA et al., 2011), com a variação dos hiperparâmetros a seguir:

- valor do k, que é a quantidade de pontos vizinhos a serem utilizados, e pode ser entre 1 e n (número total de observações);
- métrica para o cálculo da distância, variando entre as mais conhecidas: Manhattan, também chamada de *Cityblock* (Equação 5), e Euclidiana (Equação 6), para o caso deste trabalho.

$$\sum_{i=1}^{n}(|x_i - y_i|) \tag{5}$$

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{6}$$

Em que x_i e y_i representam valores da i-ésima variável nos vetores de observações x e y, e n refere-se ao número máximo de variáveis (GOLDBERGER et al., 2004).

Para a seleção das melhores opções para cada parâmetro, existe uma grande variedade de métricas de performance. Neste trabalho, serão utilizadas aquelas usuais que têm como base a matriz confusão: a acurácia, a precisão, a sensibilidade (mais conhecido como *recall*), a F1-score, e a média geométrica (do inglês, *G-mean*). Também será usada a função perda, Entropia Cruzada (do inglês, *Cross Entropy*), para medir o custo computacional em relação aos erros do modelo.

4.2.2. Métricas de avaliação de performance

A matriz confusão é uma tabela que correlaciona as taxas de acertos e erros em relação à cada classe do modelo. A Figura 5 mostra a matriz confusão para um problema de classificação binária.

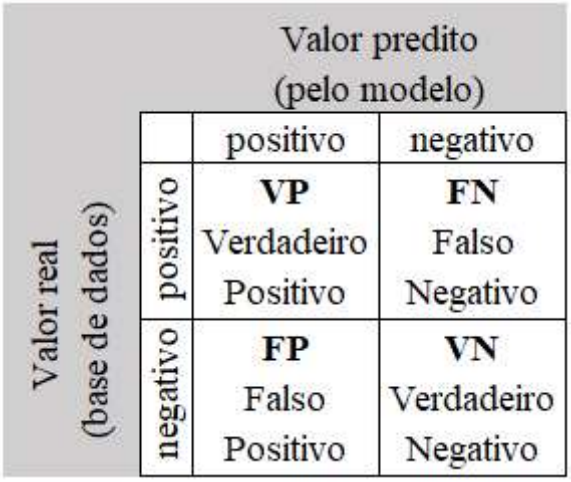


Figura 5 – Modelo de matriz confusão. Fonte: próprio autor.

O verdadeiro positivo (VP) refere-se à classificação correta da classe-alvo, enquanto o falso positivo (FN), à sua classificação incorreta. O verdadeiro negativo (VN) é a classificação correta da classe não-alvo, enquanto que o falso negativo (FP), à sua classificação incorreta (JAMES et al., 2013).

O *recall* é a taxa de positivos verdadeiros e indica o quanto de tal classe o modelo acerta. Já a precisão indica o quão certo o modelo está, ou seja, a taxa de acertos do que o modelo prediz de tal classe (PROVOST; FAWCETT, 2013). Essas métricas são apresentadas nas Equações (7) e (8) para uma classe (positiva ou alvo) em um caso binário.

$$\text{Recall} = \frac{\text{VP}}{\text{VP+FN}} \tag{7}$$

$$\text{Precisão} = \frac{\text{VP}}{\text{VP+FP}} \tag{8}$$

A métrica F1-score é a média harmônica, de peso um, entre precisão e *recall* de uma mesma classe, como mostra a Equação (9).

$$\text{F1} = \frac{\text{precisão x recall}}{\text{precisão+recall}} \tag{9}$$

Recall, precisão e *F1-score* são as métricas de performance mais comumente utilizadas para análise das classes individualmente. Já acurácia e *G-mean* são as mais comuns como métricas médias (padrão ou ponderadas)

para o modelo. A medida de acurácia é uma porcentagem resultante da matriz confusão, e refere-se ao percentual de acertos, ou classificações corretas, do modelo. A expressão para o seu cálculo está apresentada na Equação (10).

$$\text{Acurácia} = \frac{\text{VP+VN}}{\text{VP+FP+FN+VN}} = \frac{\text{predições corretas}}{\text{todas as predições}} \quad (10)$$

Há também a acurácia balanceada, que pondera a acurácia com a quantidade total de observações em cada classe. Quando o modelo se torna tendencioso para a classe majoritária, ela reduz essa vantagem fazendo a média das sensitividades das classes, como a Equação (11) apresenta.

$$\text{Acurácia}_{\text{balanceada}} = \frac{1}{M}\left(\frac{\text{VP}}{\text{VP+FN}} + \frac{\text{VN}}{\text{VN+FP}}\right) \quad (11)$$

Em que M significa a quantidade de classes.

A métrica *g-mean*, ou média geométrica, propõe maximizar a acurácia de cada classe enquanto mantem o balanceamento entre o tamanho das classes. Ela utiliza especificidade e sensitividade. Essa primeira é o mesmo que o *recall* da classe alvo, já a última é o *recall* da outra classe (negativa) ou, para classificação em problemas multiclasses, é o *recall* do agrupamento das outras classes (BARANDELA et al., 2003). A Equação (12) apresenta um exemplo para classificação binária.

$$\text{G}_{\text{mean}} = \sqrt{\text{Sensibilidade x Especificidade}} = \sqrt{\frac{\text{VP}}{\text{VP+FN}} \text{X} \frac{\text{VN}}{\text{VN+FP}}} \quad (12)$$

Para o cálculo da média entre as classes, em uma classificação multiclasses, para qualquer uma das métricas apresentadas, pode ser feito de três modos: global, média simples ou ponderada. O modelo global representa a aplicação da métrica sobre as amostras, por exemplo, para a precisão, seria a divisão da quantidade total de verdadeiros positivos pela soma destes com o total de falsos negativos. Já o modelo em média simples, representa uma média aritmética entre as classes, ou seja, calcula-se a métrica para cada classe e, em seguida, calcula-se a média. Por último, a média ponderada é como a anterior,

só que com pesos, que são função da quantidade de amostras em cada classe. A preferência no uso de um desses modos será discutida no capítulo 6 (Resultados e Discussões).

A função de custo, Entropia Cruzada (*Cross Entropy*), é usada não apenas como uma medida de erro, mas também para medir o esforço dispendido para acertar e errar cada amostra. Em outras palavras, essa métrica é uma medida do número médio de *bits* necessários para a identificação de um evento, ou seja, para acertar a predição. A Equação (13) apresenta o cálculo da função.

$$L_{log}(Y, P) = -\log Pr(Y|p) = -\frac{1}{n}\sum_{I=0}^{n-1}\sum_{C=0}^{C-1} y_{i,C} \log p_{i,C} \quad (13)$$

Sendo n, a quantidade de amostras, C, a quantidade de classes, i, uma amostra em particular, Y, a matriz resposta esperada, e P, a matriz probabilidade estimada. Quanto maior o valor da função custo, maior foi o esforço dispendido e, consequentemente, os erros cometidos (BISHOP, 2006).

4.2.3. Seleção de melhor subconjunto de variáveis

A seleção do melhor subconjunto de variáveis preditoras (do inglês, *best subset selection*) é um método clássico de seleção das variáveis que melhor predizem a resposta, ou seja, com menor erro. Neste trabalho, utilizou-se a abordagem de regressão linear, com o intuito de analisar o potencial das variáveis de entrada em predizer a variável alvo.

O método desenvolve diferentes regressões com todas as possibilidades de combinações entre as variáveis disponíveis, a fim de compor o melhor subconjunto de fatores com maior capacidade de predição da variável resposta de interesse; nesse caso, o nível de emissões de SO_2.Para essa verificação, utilizou-se o erro quadrático médio (EQM). Por exemplo, tendo-se 10 variáveis disponíveis; porém, com um objetivo de se utilizar apenas 5 delas, verifica-se todas as possíveis combinações entre as 10 variáveis com grupos de 5, e obtém-se os modelos de regressão. Aquele que apresentar o menor EQM indica o melhor subconjunto de preditores, segundo esse critério (HASTIE; TIBSHIRANI; TIBSHIRANI, 2017).

4.2.4. Aprendizado por conjunto (Comitê de modelos)

O Aprendizado por conjunto (do inglês, *Ensemble Learning*) caracteriza-se como uma abordagem do aprendizado de máquina, com o intuito de melhorar a performance do modelo, a partir da combinação de múltiplos modelos. Esse procedimento pode usar técnicas diferentes para cada modelo ou uma única técnica; porém, com variações em sua parametrização. As principais classes de aprendizado por agrupamento são: métodos de média - geralmente feito com preditores homogêneos, cada um de forma independente e paralela em relação ao outro. O resultado é a média do que foi obtido a partir dos modelos individuais; métodos de reforço - também feito com preditores homogêneos; porém, aplicados de forma sequencial e depois combinados no modelo final; e métodos de empilhamento - usa preditores heterogêneos, treinando-os em paralelo, e depois aplica um modelo na saída, que decide o resultado a partir de um aprendizado dos modelos anteriores (ZHOU, 2012).

Neste trabalho, foi usado a abordagem de média com seleção de subespaço, no qual só a técnica k-NN foi utilizada para a geração de comitês de modelos. Os hiperparâmetros estudados foram, a quantidade de modelos e de variáveis em cada modelo, o número de vizinhos (*k*) e a métrica da distância (*cityblock* e euclidiana). Deve-se ressaltar que a seleção das variáveis para cada treinamento foi aleatória e sem reposição, e que todas as observações no grupo de treinamento foram usadas em cada modelo. A decisão da resposta da classificação foi feita por voto majoritário, ou seja, a classe com mais resultados entre os modelos foi a escolhida (BREIMAN, 1996; MOSAVI et al., 2021).

5. ESTUDO DE CASO

A metodologia discutida nesse trabalho foi aplicada a uma base de dados de uma caldeira de recuperação química em funcionamento, pertencente a uma indústria brasileira de celulose que utiliza o processo do tipo *kraft*.

5.1 DESCRIÇÃO DA BASE DE DADOS

O processo ocorreu em uma caldeira de recuperação química, e teve os dados coletados entre os meses de Junho e Setembro de 2001 em funcionamento normal (sem paradas do equipamento), constituindo 2.860 observações com 16 variáveis ao todo. O processo é contínuo, impossibilitando qualquer folga no tempo para ajustes ou reparos, ou seja, um monitoramento regular é essencial para controlar a qualidade do produto. Qualquer necessidade de parada não programada afeta toda a planta industrial. Os dados contemplam características do licor preto injetado na fornalha, dos três níveis de ar alimentados e da variável de interesse, a emissão do dióxido de enxofre.

A Figura 6 esquematiza a caldeira e as variáveis coletadas. Notam-se duas ou mais medições para uma mesma variável. É o caso das pressões do licor preto medidas em diferentes paredes da fornalha, e os teores de sólidos, que são medidos em dois pontos distintos. Assim, essas variáveis foram resumidas por média para serem usadas no modelo, por suas estatísticas serem semelhantes.

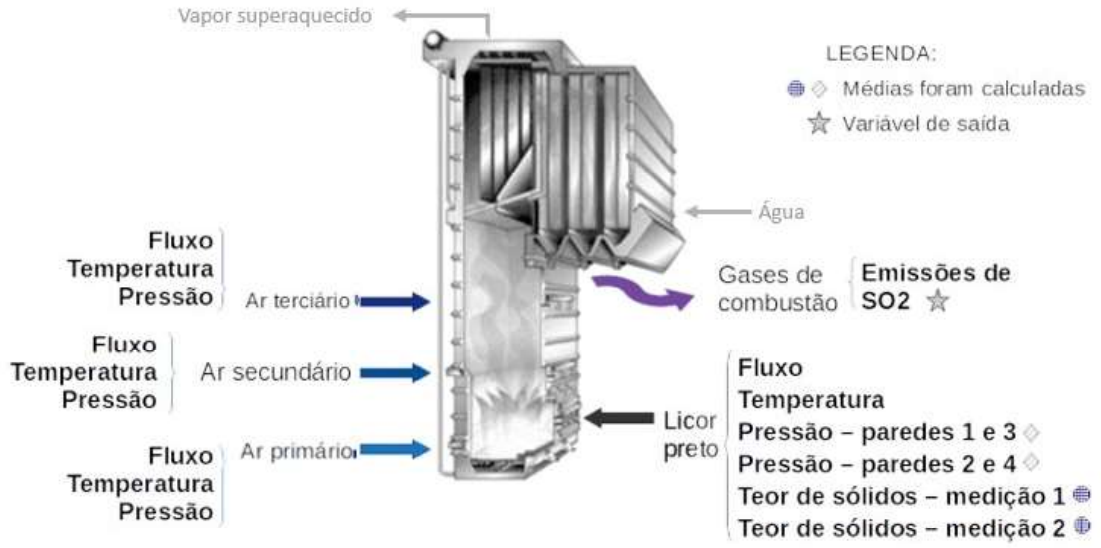


Figura 6 – Representação das variáveis e seus pontos de coleta na caldeira química. Fonte: adaptado de Belizário (2020).

A Tabela 1 lista as variáveis com seus valores de média, desvio-padrão, mínimo e máximo, além da unidade de medida. Observa-se que a maior parte das variáveis possuem desvios-padrão relativamente baixos, o que pode caracterizar um processo com poucas perturbações, ou seja, em condições de operação normal.

Tabela 1 - Listagem das variáveis coletadas no processo com dados estatísticos.

	Variável (código)	Média	Desvio-padrão	Mínimo	Máximo	Unidade
Entrada (Combustível)	Fluxo de entrada de licor (F_lp)	109,43	9,13	51,7	125,8	ton/h
	Temperatura do licor (T_lp)	126,45	0,98	115,4	131,7	°C
	Pressão do licor (paredes 2 e 4) (P_lp)	0,86	0,08	0,6	1,4	mmH2O
	Pressão do licor (paredes 1 e 3)	0,9	0,07	0,7	1,4	mmH2O
	Teor de sólidos do licor (medição 1) (S_lp)	68,16	1,66	62,4	73,6	%
	Teor de sólidos do licor (medição 2)	68,24	1,62	62,4	74,1	%
Entrada (Ar)	Fluxo de ar primário (F_a1)	153,91	6,97	134,2	176,2	ton/h
	Temperatura do ar primário (T_a1)	150,03	1,95	136,5	158,6	°C
	Pressão do ar primário (P_a1)	37,37	7,04	16,4	105,4	mmH2O
	Fluxo de ar secundário (F_a2)	187,51	20,96	114,4	260,3	ton/h
	Temperatura do ar secundário (T_a2)	166,99	3,93	129,2	173,3	°C
	Pressão do ar secundário (P_a2)	202,77	18,54	118	265,9	mmH2O
	Fluxo de ar terciário (F_a3)	48,44	3	33,6	54,3	ton/h
	Temperatura do ar terciário (T_a3)	29,9	4,58	17,8	44,7	°C
	Pressão do ar terciário (P_a3)	200,23	17,11	100,8	263,5	mmH2O
Saída	Emissões de SO_2	119,42	6,46	0	630	ppm

A variável de saída possui a maior parte dos valores na faixa de 80 à 140 ppm, como pode ser visto na Figura 7, apresentando uma distribuição em grupos de 10 em 10ppm, como por exemplo, [80-90[ppm e [90-100[ppm. Nota-se que a quantidade de valores só é expressiva a partir de 80 ppm. Por apresentar faixas distintas, esse modelo de categorização foi escolhido para a variável resposta. Ou seja, a faixa [80-90[ppm contempla a classe 1, [90-100[ppm, a classe 2, [100-110[ppm, a classe 3, [110-120[ppm, a classe 4, [120-130[ppm, a classe 5, e [130-140[ppm, a classe 6. No próximo capítulo, será apresentado de forma mais clara o porquê dessa forma de categorização dos dados.

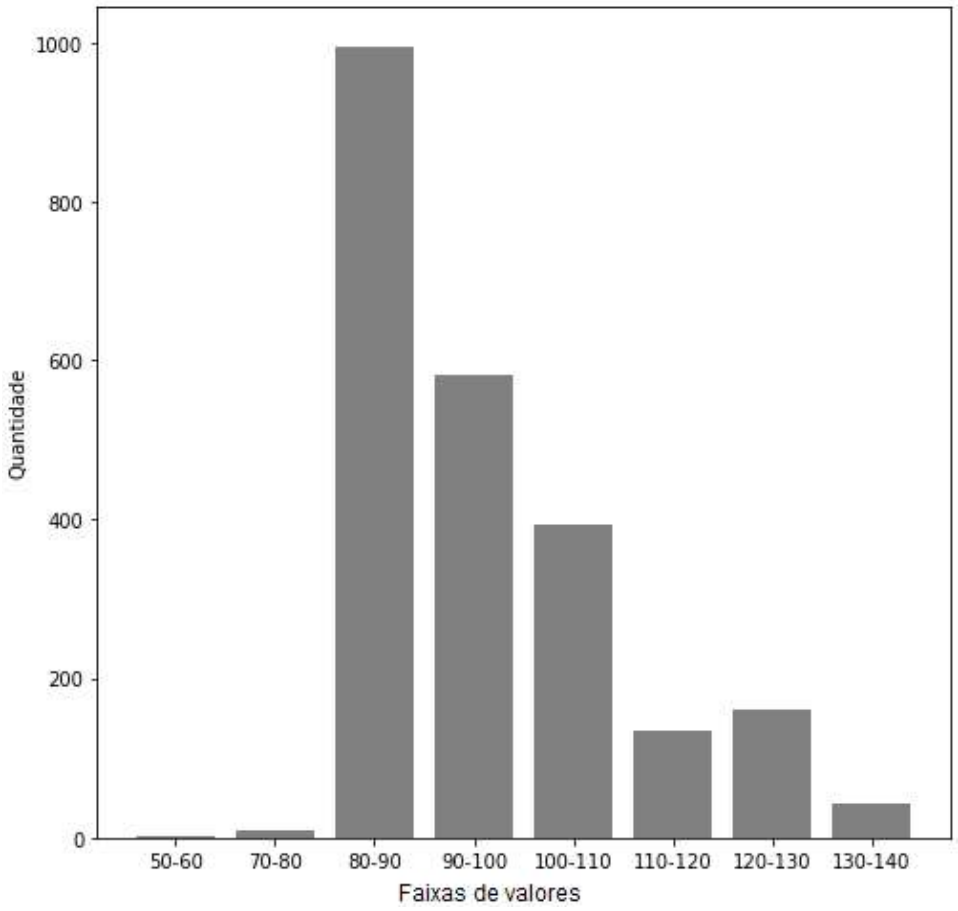


Figura 7 – Contagem de dados da Emissão de SO_2, por faixa de valores. Fonte: autoria própria.

6. RESULTADOS E DISCUSSÕES

Segundo Gibert et al. (2018), para escolher o melhor método a ser aplicado aos dados, é necessário ter duas informações: qual o problema a ser solucionado, e qual a estruturação do conjunto de dados. Como apresentado no Capítulo 2 (Objetivo), a questão a ser resolvida é como antecipar a tomada de decisão a fim de se aprimorar o controle ambiental do equipamento.

Nessa direção, o primeiro passo tomado foi a construção de um sensor virtual para a emissão de dióxido de enxofre. O tratamento dos dados foi determinado a partir de um estudo gráfico da variável resposta, determinando sua utilização de forma contínua ou discreta, com quais variáveis preditoras há correlações, e se houve necessidade de limpeza dos dados. Assim, com o objetivo e o comportamento dos dados determinados, foi aplicada a técnica de classificação para cada variável, a fim de observar o potencial de cada uma delas em predizer a variável alvo, além de discutido o modelo que melhor ajustou aos dados a partir de métricas de performance. Na sequência, o método de seleção do melhor subconjunto foi aplicado com regressão linear. Como terceiro objeto de comparação, foi desenvolvido um comitê de modelos com seleção aleatória de subconjuntos de variáveis. Ao final, apresenta-se uma comparação entre as opções descritas acima.

6.1. SENSOR VIRTUAL PARA EMISSÃO DE DIÓXIDO DE ENXOFRE

6.1.1. Pré-processamento de dados

Definido o objetivo como sendo a construção de um sensor virtual para a predição das emissões de dióxido de enxofre (SO_2) na caldeira de recuperação química, o primeiro passo no tratamento dos dados foi a junção das variáveis pressão do licor preto, com medições em duas paredes da fornalha, e a porcentagem de sólidos no licor preto, também com dois pontos de medição. Assim, tem-se, ao final, treze variáveis de entrada e uma variável de saída, como listado na Tabela 1.

Visualizando a variável de saída ao longo das observações coletadas, apresenta-se, na Figura 8, um grupo de observações com valores

sugnificativamente altos, acima de 500 ppm, destacados pelo círculo em vermelho. Como o intuito desse trabalho não é a investigação de possíveis operações anormais, esse grupo discrepante foi descartado.

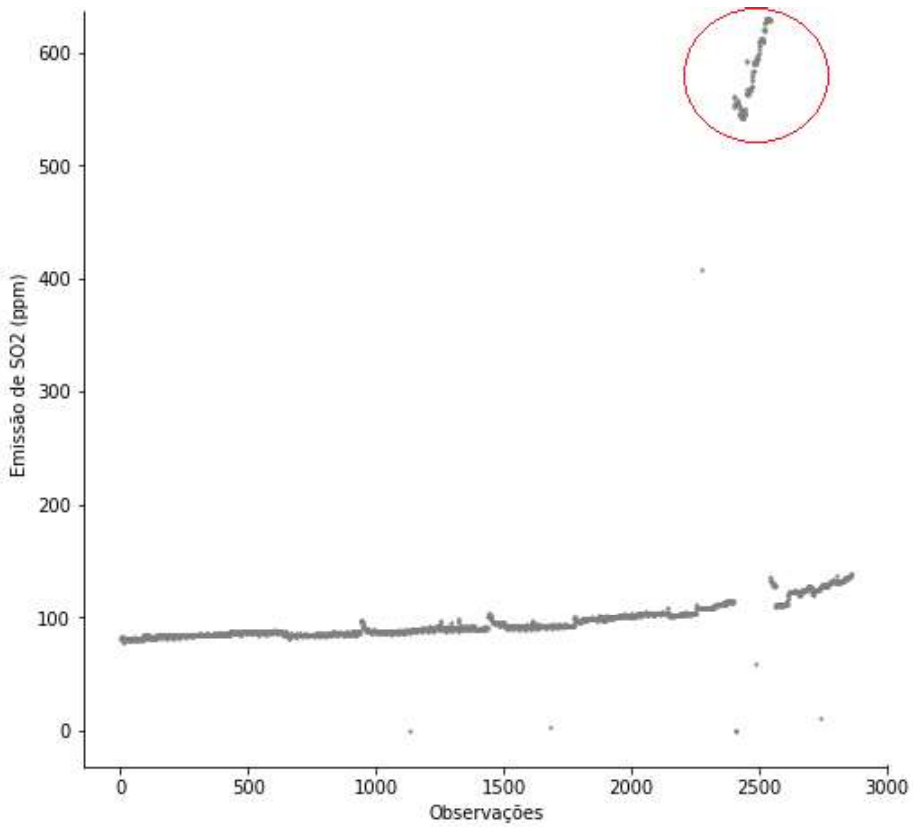


Figura 8 – Gráfico com valores da variável de saída, Emissão de SO_2, em função do tempo.

A partir dessa primeira exclusão, foi analisada a dispersão e possível correlação entre as variáveis, conforme as Figuras de 9 à 12. A Figura 9 mostra uma matriz de gráficos de disperssão entre as variáveis relacionadas ao licor preto e a emissão de SO_2, com a diagonal principal sendo o histograma de cada variável. Já a Figura 10 mostra as características coletadas do ar primário e as suas relações com a varíavel de saída. A Figura 11 retrata o ar secundário, e a Figura 12, o ar terciário.

Em geral, visualizam-se pontos isolados em relação ao maior agrupamento de dados. Nesse sentido, faz necessário a aplicação de um filtro para selecioná-los e removê-los. Também é possível observar algumas correlações entre a vazão e a temperatura, e entre a vazão e a pressão, dos ares de combustão. Porém, nessa visualizações, não foi possível identificar correlações diretas com a variável de saída.

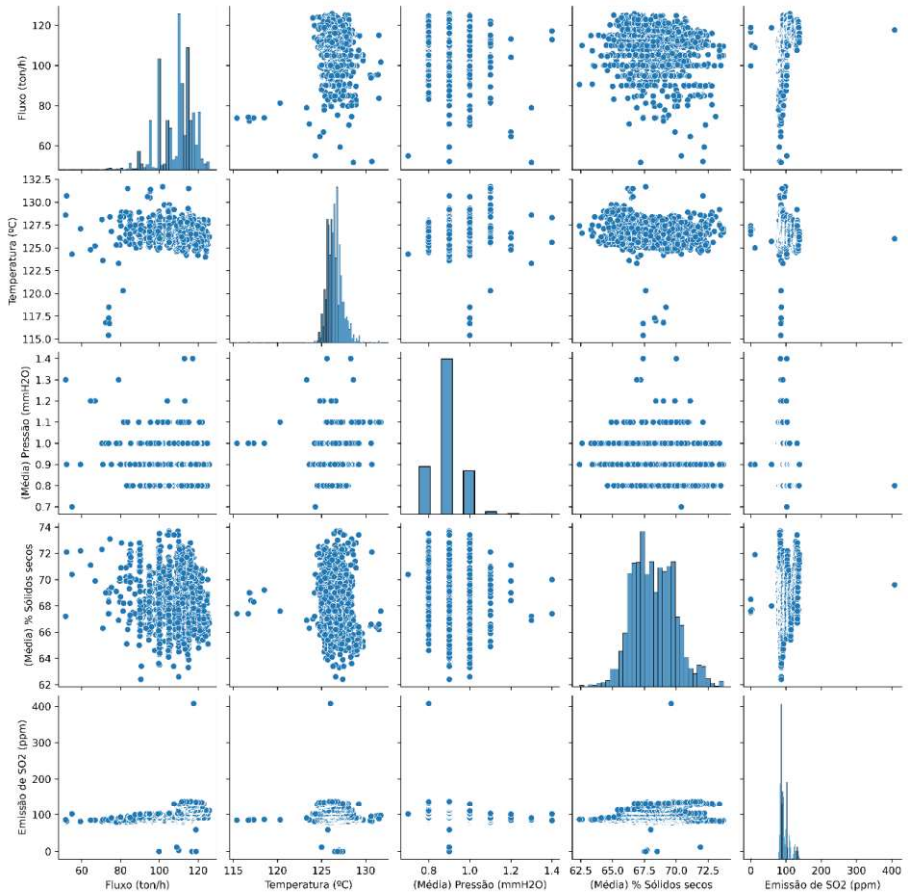


Figura 9 - Gráficos de correlação-comportamento das variáveis do licor preto e a resposta.

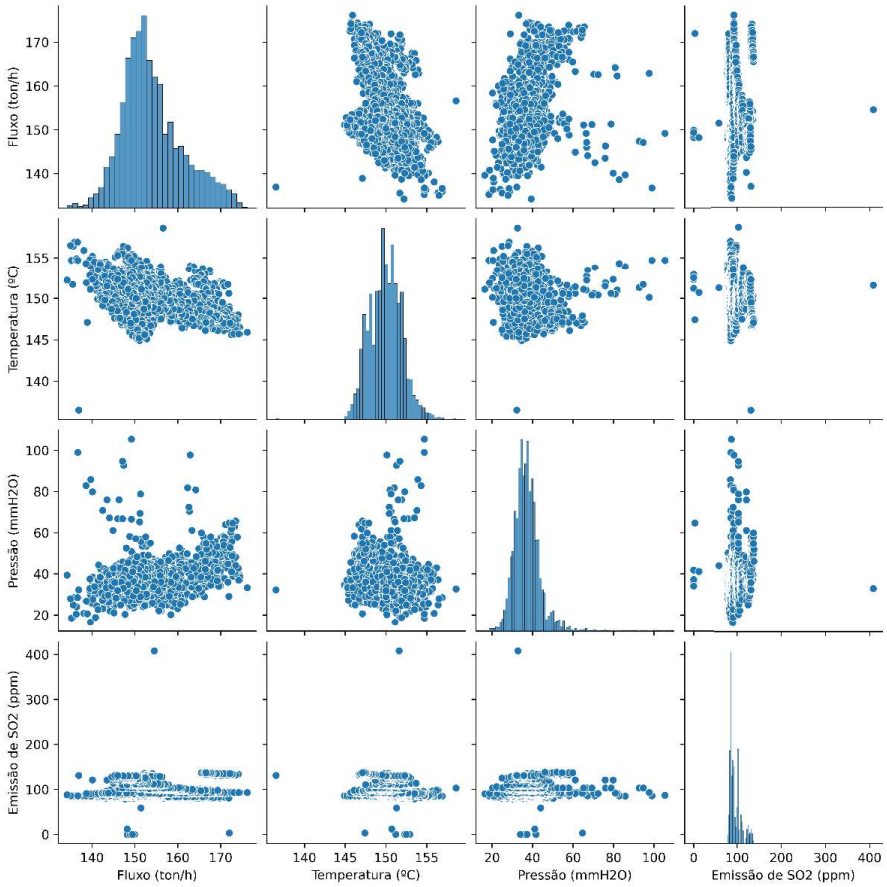


Figura 10 - Gráficos de correlação-comportamento das variáveis do ar primário e a resposta.

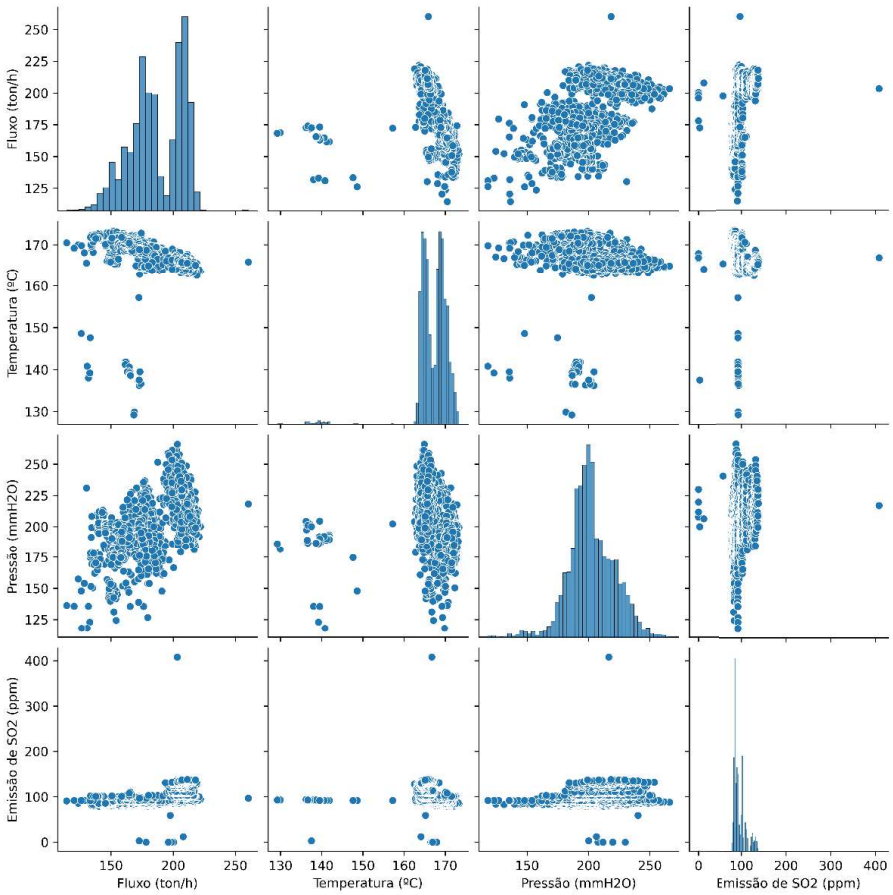


Figura 11 - Gráficos de correlação-comportamento das variáveis do ar secundário e a resposta.

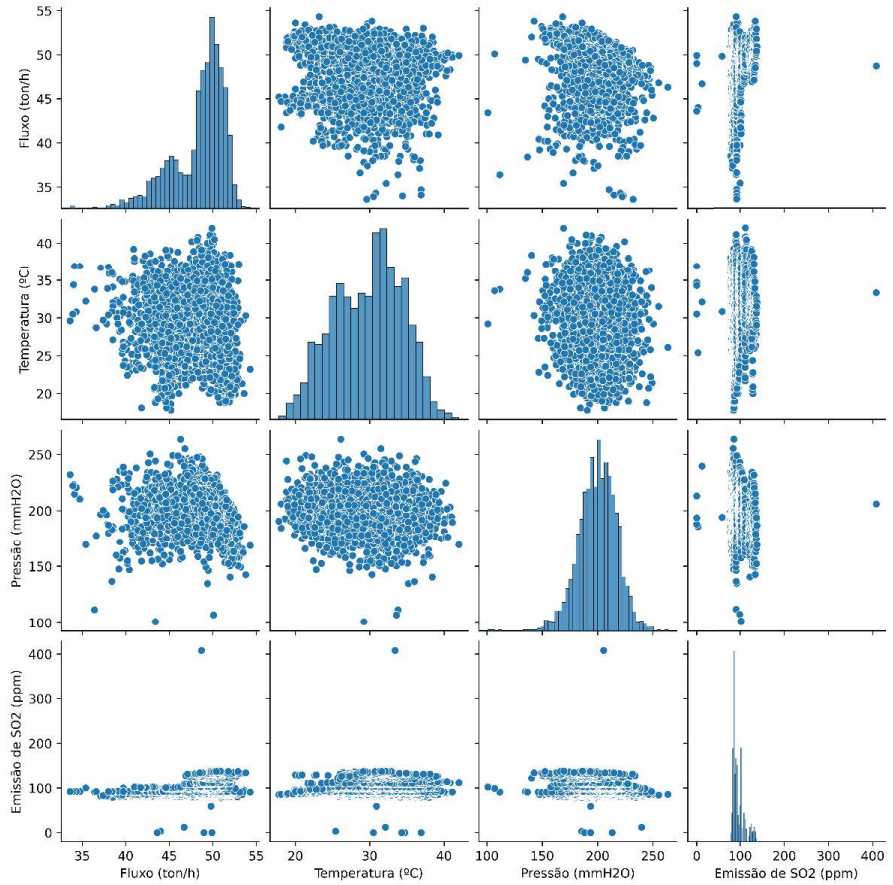


Figura 12 - Gráficos de correlação-comportamento das variáveis do ar terciário e a resposta.

O Filtro de Hampel foi aplicado em todas as variáveis a fim de identificar os valores discrepantes, baseado na mediana e com fator de correção igual a 3 (Equação 4). A Tabela 2 mostra a quantidade de dados identificados pelo Filtro de Hampel como discrepantes, para cada variável. A variável que mais contém dados discrepantes, 370 observações, é a emissão de SO_2, o que acontece devido ao grupo de valores acima de 500 ppm, já mostrados na Figura 8. Em seguida, tem-se as vazões do ar terciário e do ar primário e a pressão do ar primário, com 185, 100 e 93 valores discrepantes, respectivamente.

Tabela 2 – Contagem de valores discrepantes pelo Filtro de Hampel.

Variáveis	Quantidade de Observações retidas
Fluxo de entrada de licor	43
Temperatura do licor	53
Pressão do licor (média)	9
Teor de sólidos do licor (média)	6
Fluxo de ar primário	58
Temperatura do ar primário	12
Pressão do ar primário	84
Fluxo de ar secundário	0
Temperatura do ar secundário	34
Pressão do ar secundário	57
Fluxo de ar terciário	154
Temperatura do ar terciário	0
Pressão do ar terciário	15
Emissões de SO_2	240
Total de observações discrepantes	765
Total de linhas removidas	543
Tamanho final da base de dados	2317 linhas e 14 colunas

A Figura 13 apresenta, em destaque, os pontos identificados e retirados, usando como exemplo, essas quatro variáveis. Observam-se alguns pontos isolados, em azul, em cada gráfico, e também outros pontos em azul entre os grupos de dados aglomerados, isto é, o filtro separa os dados discrepantes para cada variável; porém, todo o vetor de observações (com as 14 variáveis) é excluído do conjunto de dados.

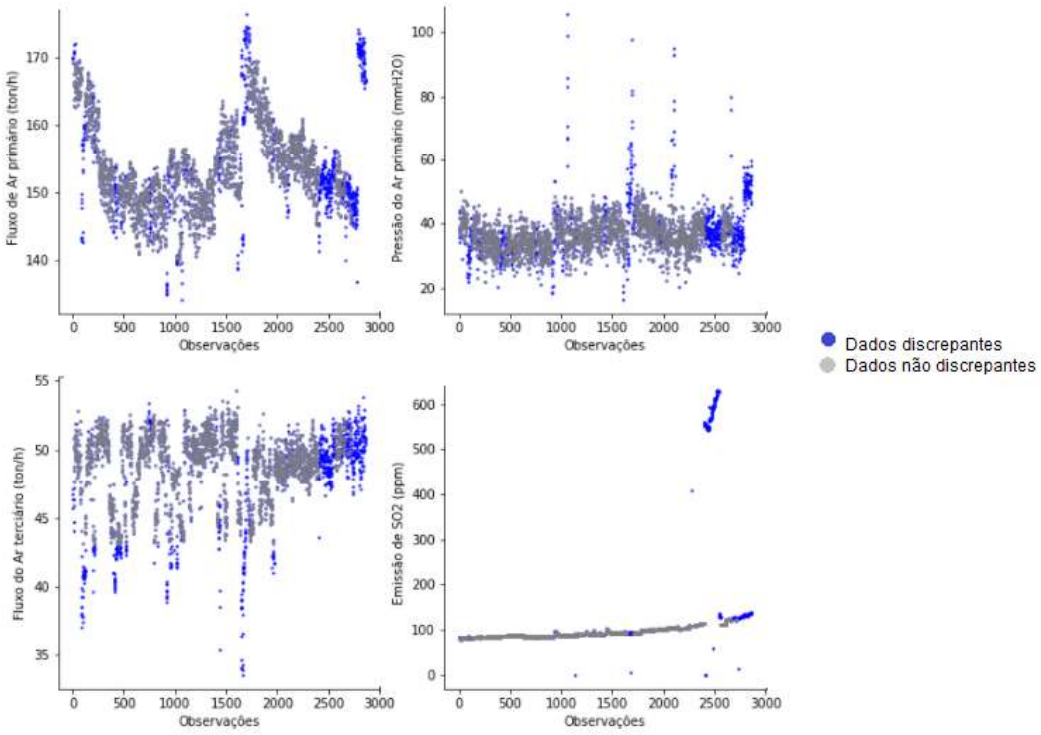


Figura 13 – Destaque para os dados retidos pela aplicação do Filtro de Hampel em toda a base de dados, apresentando as quatro variáveis com mais dados discrepantes.

Para melhor visualização das correlações entre as variáveis, a Figura 14 apresenta dois mapas de calor com matrizes de correlações, usando coeficiente de correlação de Pearson antes e após a aplicação do filtro de Hampel. É possível observar maiores valores de correlações após a aplicação do filtro, como por exemplo entre a vazão de licor preto e a temperatura do ar secundário, e a vazão do ar secundário e a emissão de SO_2. Porém, é necessário ressaltar que não houve mudança significativa nas correlações, apenas um pequeno aumento, mostrando que o filtro de Hampel não alterou a natureza e as relações entre as variáveis.

No geral, a Figura 14 mostra relações entre as variáveis de entrada e delas com a variável de saída. Juntamente com o conhecimento do fenômeno do processo, pode-se sugerir que um modelo preditivo com acurácia satisfatória seria possível com essas variáveis.

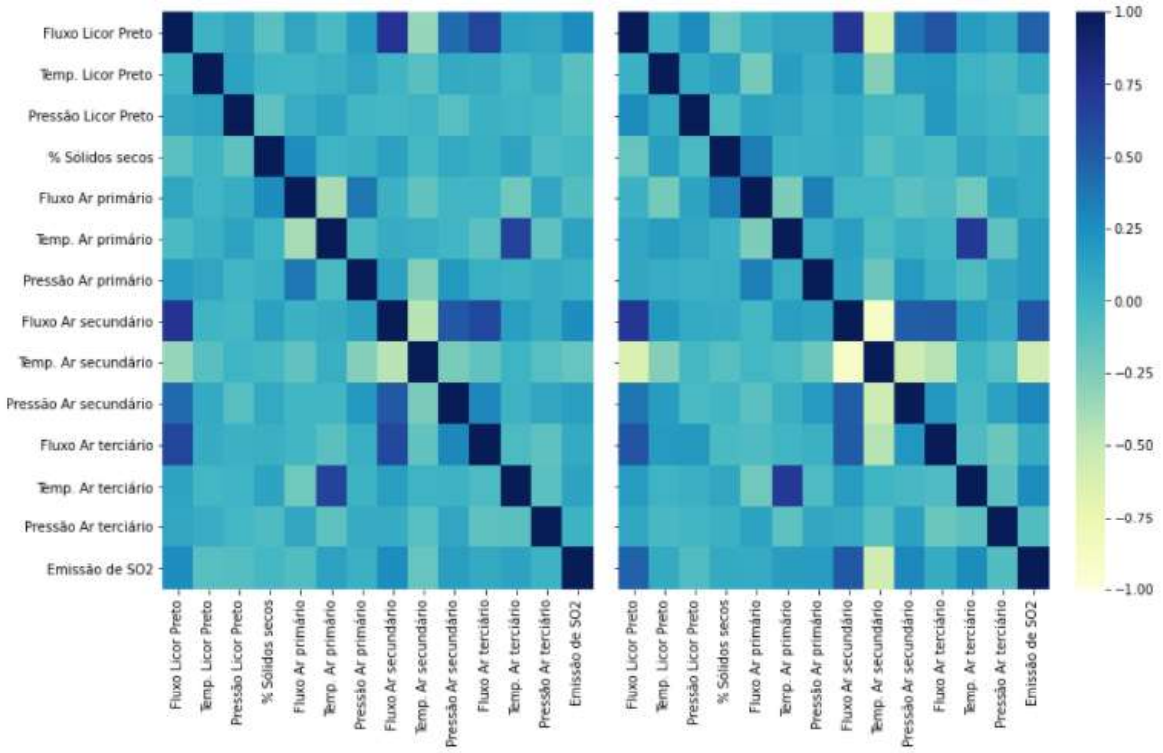


Figura 14 – Mapa de calor apresentando a correlação entre as variáveis. À esquerda, base de dados original e à direita, base de dados após filtro de Hampel.

Na Tabela 1, foram apresentados os valores mínimos e máximos das variáveis originais. Para a emissão de SO_2, tem-se zero e 680 ppm, respectivamente. Após o tratamento com o filtro de Hampel, esses limites foram reduzidos para 50 e 140 ppm, e com isso a distribuição em classes de 10 em 10 ppm, conforme mostrado na Figura 7, para o uso de uma abordagem classificatória, como é proposto neste trabalho.

As duas primeiras faixas, 50-60 ppm e 70-80 ppm, foram desconsideradas, por conterem uma quantidade consideravelmente menor de observações.

Como foi apresentado no capítulo 3, Revisão Bibliográfica, no estudo do processo de emissão de dióxido de enxofre na caldeira química, um dos destaques para o controle e menor emissão é a temperatura da fornalha. Quanto mais quente a caldeira, menor o nível de emissões de SO_2 (TAMMINEN; TAMMINEN, 2015). As Figuras 15 e 16 mostram, respectivamente, os comportamentos da vazão de licor e da temperatura do ar secundário, em relação às classes da variável de saída. Essa avaliação depende de uma análise simultânea de um conjunto de variáveis; porém, é possível visualizar, em ambos

os casos, associações entre as faixas dessas variáveis e as classes de SO_2, o que é importante para a sua discriminação.

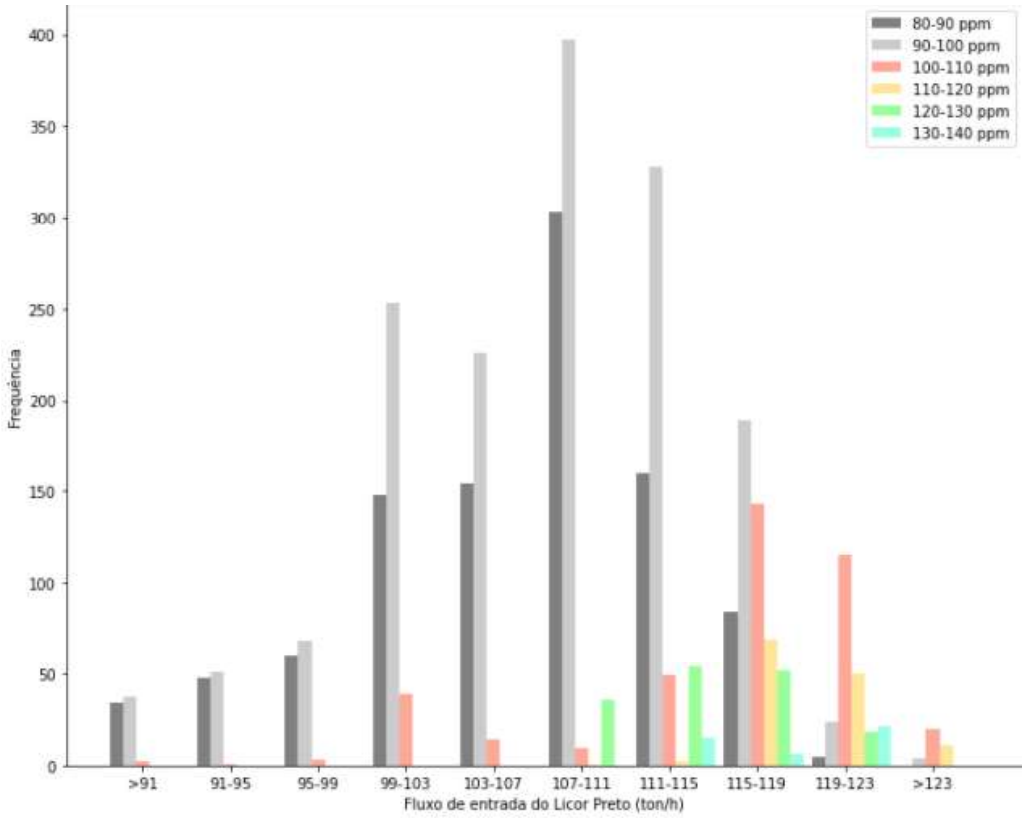


Figura 15 – Distribuição dos valores de vazão do licor preto, pelas classes da emissão de SO_2.

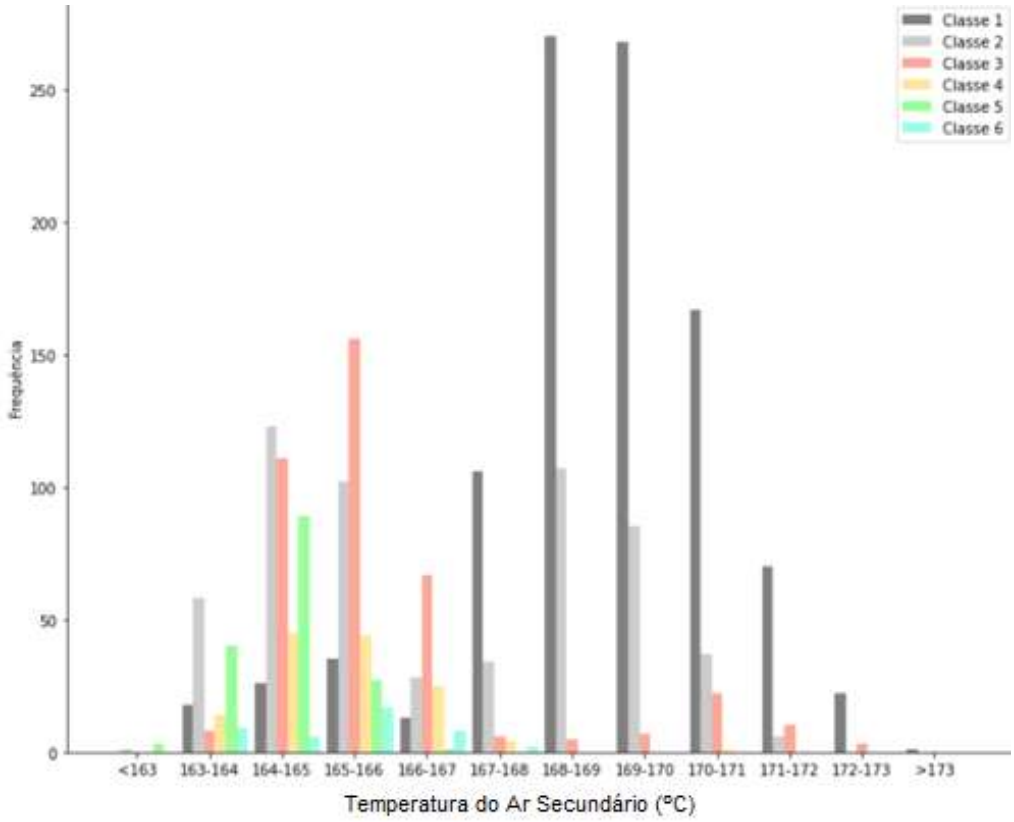


Figura 16 – Distribuição da temperatura do ar secundário, pelas classes da variável resposta.

Após essa fase de análise do comportamento de cada variável em relação às demais e o tratamento dos dados discrepantes, foi realizada a divisão dos dados para a geração dos modelos. 75% dos dados compuseram o grupo de treinamento, e os outros 25%, o grupo de teste. Essa divisão foi aleatória, salvo os valores de máximo e mínimo de cada variável, que foram incluídos no grupo de identificação propositalmente, a fim de se evitar extrapolações pelo modelo.

Após essa divisão, os dados de identificação foram padronizados com média zero e desvio-padrão um, para impedir que o modelo desenvolva um viés devido à diferença entre as grandezas das variáveis. Esse procedimento também foi aplicado sobre os dados de teste.

6.1.2. Processamento de dados (construção de modelo)

Para a construção do modelo classificatório da emissão de dióxido de enxofre, com as variáveis destacadas na subseção anterior, foi utilizada a técnica k-vizinhos mais próximos (k-NN), já descrita no capítulo de Metodologia. Faz-se necessário ressaltar que os modelos classificatórios de aprendizado de máquina, e a própria técnica usada neste trabalho, são mais frequentemente utilizados com duas classes (YANG et al., 2017); (FENG; LI, 2020); (HARROU; ZEROUAL; SUN, 2020). Assim, o fato de a variável alvo apresentar seis classes, como apresenta a Tabela 3, acarreta um desafio pouco encontrado na literatura de processos químicos.

Tabela 3 – Faixas de valores para cada classe da variável resposta.

Classe	Faixa de Valores
1	$80 \leq y < 90$
2	$90 \leq y < 100$
3	$100 \leq y < 110$
4	$110 \leq y < 120$
5	$120 \leq y < 130$
6	$130 \leq y < 140$

Com o intuito de analisar a capacidade das variáveis de entrada em predizer a emissão de SO_2, primeiramente foi treinado um modelo k-NN com cada uma das treze variáveis em separado, como apresenta o item a seguir.

6.1.2.1. Modelo k-NN univariável

No capítulo de “Revisão Bibliográfica”, foi discutida a desvantagem da técnica k-NN em relação à quantidade de dimensões, ou seja, variáveis de entrada. Devido a isso, o primeiro passo do processamento dos dados foi treinar o modelo com apenas uma variável, obtendo assim, 13 modelos para cada combinação entre os parâmetros (k vizinhos, e métrica de distância).

Nesta primeira etapa, a quantidade de vizinhos foi variada entre 1 e 5, e a métrica de distância, entre Manhattan (*cityblock*) e Euclidiana. Para cada combinação entre os parâmetros, o algoritmo foi executado 5 vezes, a fim de se testar a estabilidade do modelo, gerando resultados médios. Para cada vez, dividiu-se o conjunto de dados entre conjunto de treinamento (75%) e conjunto de teste (25%), de forma aleatória. A Figura 17 apresenta os gráficos com acurácia média para cada modelo, para as métricas de distância, (a) Manhattan e (b) Euclidiana.

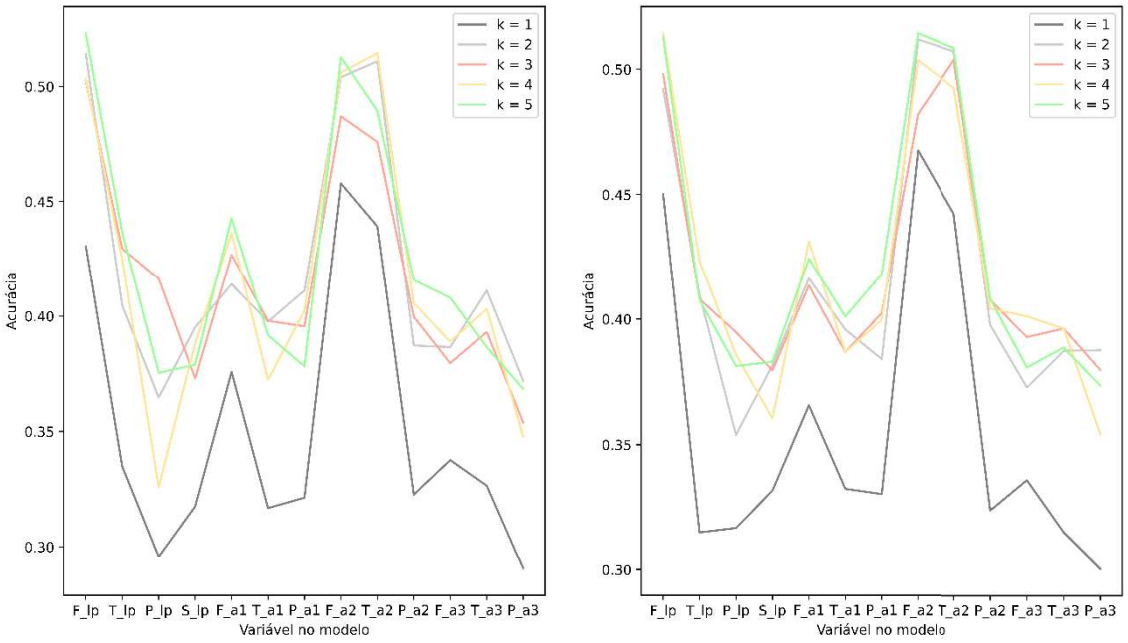


Figura 17 – Curvas da acurácia média em função das variáveis para cada modelo e valor de k, para as métricas de distância, (a) Manhattan e (b) Euclidiana.

Analisando a influência de cada variável na classificação da resposta (emissão de SO_2), tem-se uma média mínima de 35% de acurácia, o que é considerando satisfatório, dada a complexidade do processo para que apenas

uma única variável posso descrevê-lo. As variáveis com maiores acurácias foram, as vazões de licor preto e de ar secundário, e a temperatura do ar secundário. Essas variáveis estão entre aquelas de maior efeito no controle da temperatura da fornalha. Além de que, os modelos para k = 2, 3, 4 e 5, praticamente não se diferenciaram. Por fim, analisando o desempenho das métricas de distância, também não houve diferenças significativas.

Esses resultados indicam a possibilidade de não usar todas as variáveis da base de dados para desenvolver um classificador satisfatório para as emissões de SO_2. Por isso, a próxima etapa consiste em uma seleção de subconjuntos de variáveis preditoras para explicar a emissão de dióxido de enxofre.

6.1.2.2. Seleção de subconjunto de variáveis preditoras

Como foi apresentado na Figura 17, algumas variáveis possuem significativo potencial em classificar a variável resposta. Por isso, se fez necessário testar possíveis subconjuntos de variáveis de entrada. Neste trabalho, empregou-se o modelo de regressão linear múltipla (MONTGOMERY; RUNGER, 2009).

A seleção do melhor subconjunto ocorre a partir do menor erro quadrático médio (EQM), dadas as combinações entre as variáveis preditoras. A Tabela 4 destaca as variáveis selecionadas para os melhores subconjuntos com 1, 2, 3 até com 13 preditores. Nela tem-se a temperatura do ar secundário como a variável com maior percentual de explicação sobre o SO_2, entre os subconjuntos com apenas um preditor. Essa variável é uma daquelas que apresentaram maior valor de acurácia média no modelo de k-NN univariável (Figura 17).

Tabela 4 – Variáveis selecionadas para o melhor subconjunto do modelo de n preditores (continua).

n	F_lp	T_lp	P_lp	S_lp	F_a1	T_a1	P_a1	F_a2	T_a2	P_a2	F_a3	T_a3	P_a3
1									X				
2					X				X				
3					X			X		X			
4			X		X			X		X			
5					X	X		X		X		X	
6					X	X		X		X		X	X
7			X		X	X		X		X		X	X

Tabela 4 – Variáveis selecionadas para o melhor subconjunto do modelo de n preditores (conclusão).

n	F_lp	T_lp	P_lp	S_lp	F_a1	T_a1	P_a1	F_a2	T_a2	P_a2	F_a3	T_a3	P_a3
8			X	X	X	X		X		X		X	X
9			X	X	X	X		X		X	X	X	X
10		X	X	X	X	X		X		X	X	X	X
11		X	X	X	X	X		X	X	X	X	X	X
12		X	X	X	X	X	X	X	X	X	X	X	X
13	X	X	X	X	X	X	X	X	X	X	X	X	X

A Figura 18 mostra o coeficiente de determinação obtido pelo modelo de regressão para cada subconjunto selecionado. O subconjunto com todas as variáveis é o de maior coeficiente de determinação na regressão linear; por isso foi desenvolvido novamente um modelo de k-NN, para cada subconjunto selecionado.

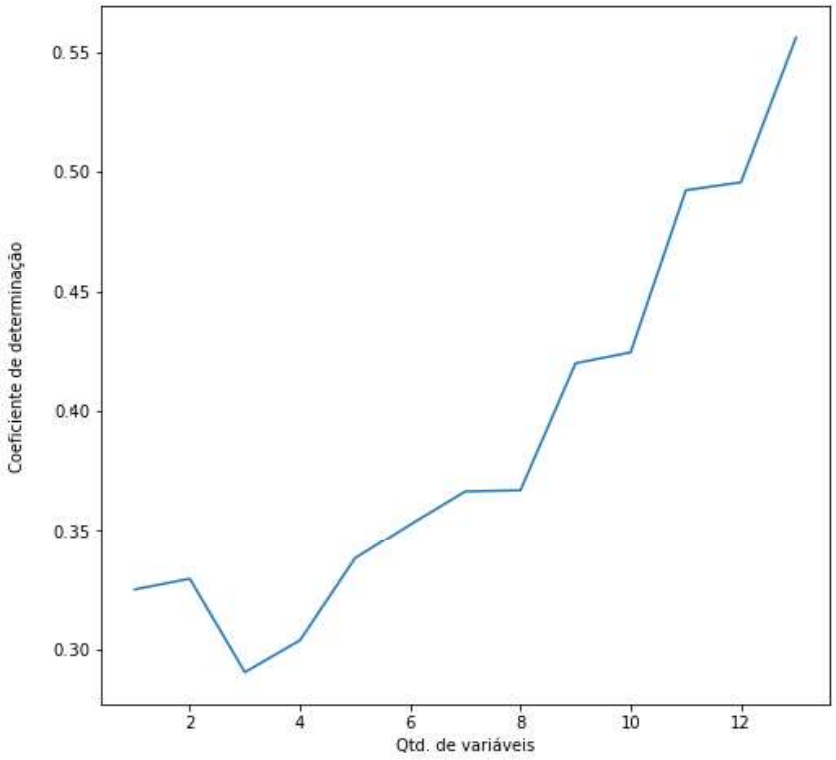


Figura 18 – Coeficiente de determinação em função da quantidade de preditores.

A Figura 19 apresenta os valores médios de acurácia em função do número de preditores, no qual as curvas distinguem-se pela quantidade de vizinhos mais próximos (k).

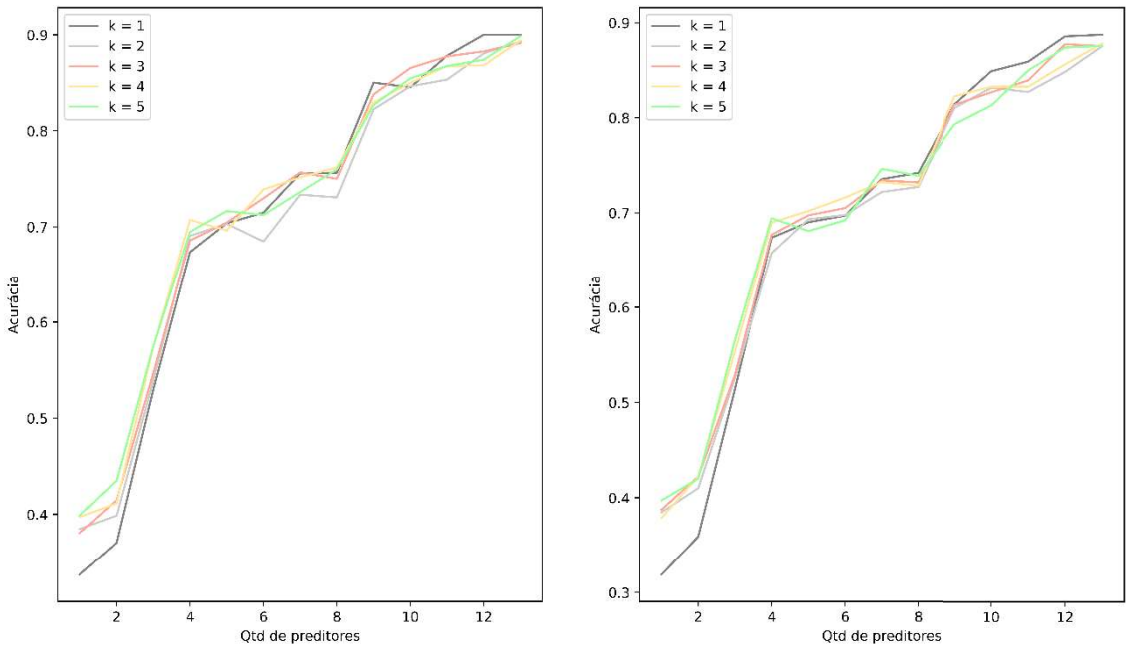


Figura 19 – Acurácia média em função da quantidade de preditores no modelo k-NN, em que (a) distância Manhattan e (b) distância Euclidiana.

A Figura 19(a) difere-se da 19(b) pela métrica de distância, Manhattan e Euclidiana, respectivamente. Todas as curvas demonstram comportamentos similares; porém, a distância de Manhattan apresenta resultados um pouco superiores. Nota-se ainda que, assim como mostrado na Figura 18, o conjunto com as 13 variáveis preditoras apresentou o melhor resultado.

A primeira abordagem, de modelos com preditores em separado, resultou em um indício do potencial de poucas variáveis para compor o modelo. Já pela segunda abordagem, a melhora da capacidade de classificação é crescente até o uso de todas as 13 variáveis. Esse fato é parcialmente devido à ordenação dessas variáveis no conjunto de dados. Com isso, uma terceira abordagem foi investigada, denominada de comitê de modelos.

6.1.2.3. Comitê de modelos (*Ensemble Learning*)

Como foi apresentado no capítulo de Metodologia, o comitê de modelos baseia-se em, a partir de uma seleção aleatória de subconjuntos de amostras, os dados de treinamento são divididos para o treinamento de n modelos, um para cada um desses subconjuntos. Pode-se empregar técnicas diferentes ou a mesma técnica de aprendizado de máquina. Neste trabalho empregou-se a

mesma técnica k-NN. Nesse procedimento o resultado final é função de uma votação, em que a classe mais frequente na saída da maioria dos modelos define o resultado final (BREIMAN, 1999).

Assim, foram desenvolvidos comitês de modelos variando-se a quantidade de vizinhos mais próximos (k), a métrica de distância, a quantidade de sub-modelos e a quantidade de preditores. A Tabela 5 resume todos os valores utilizados.

Tabela 5 – Parâmetros e suas variações para a abordagem do comitê de modelos.

Hiperparâmetro	Valores
K – vizinhos mais próximos	1, 2, 3, 4, 5
Métrica de distância	Manhattan Euclidiana
Quantidade de sub-modelos	1, 5, 15, 25, 100
Quantidade de preditores	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

A Figura 20 apresenta a acurácia média para os modelos com 1 vizinho mais próximo e distância de Manhattan, em função da quantidade de sub-modelos e de preditores.

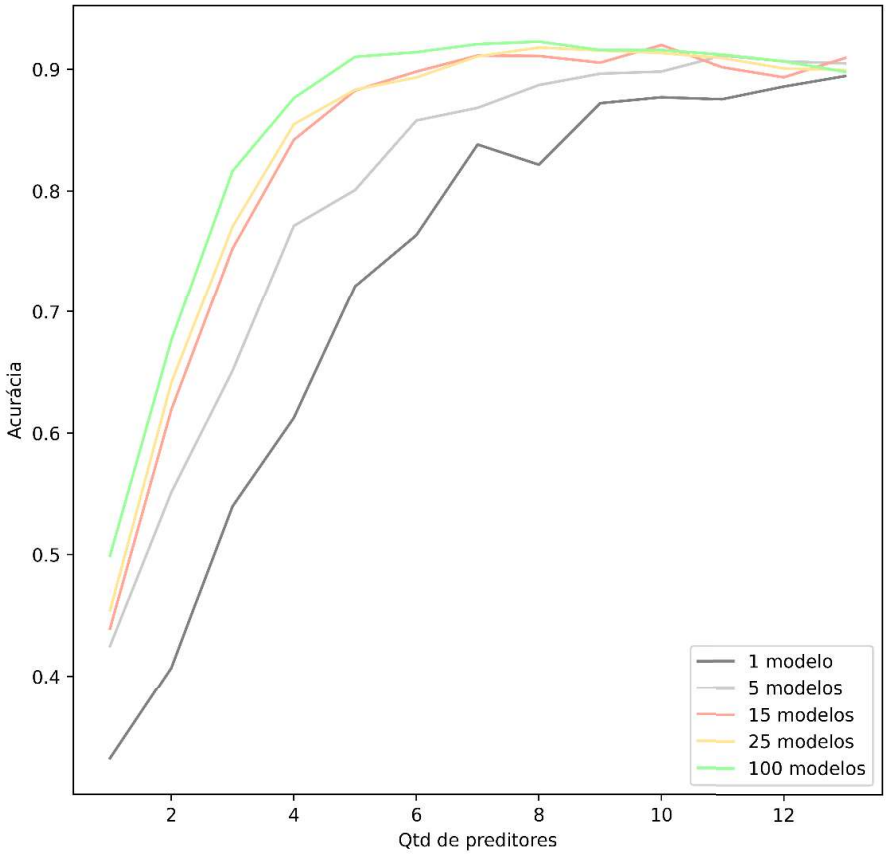


Figura 20 – Acurácia média em função da quantidade de preditores nos modelos. As curvas distinguem-se pela quantidade de modelos no comitê.

Analisando a Figura 20, percebe-se que o aumento do número de sub-modelos, além de aumentar o valor da acurácia, reduz a quantidade de preditores para os melhores resultados. Tem-se que a curva de 100 sub-modelos apresenta o melhor resultado geral. Devido a isso, a Figura 21 apresenta a variação de k e da distância para essa abordagem.

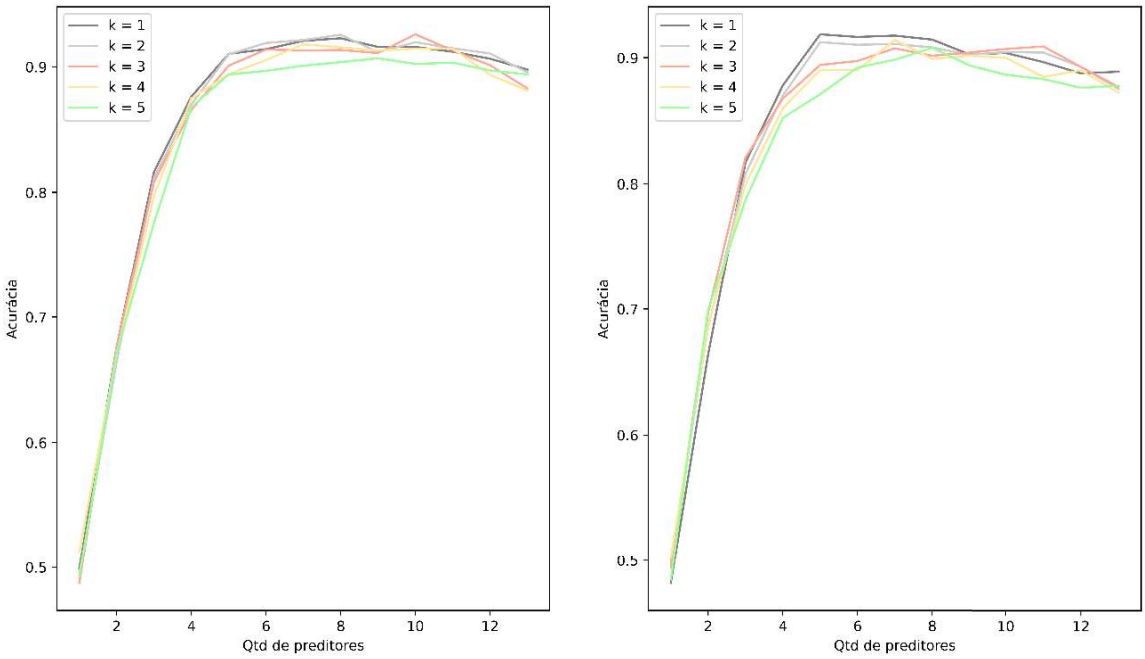


Figura 21 – Médias da acurácia em função da quantidade de preditores e as curvas se diferem pelo valor de k. (a) distância Manhattan e (b) distância Euclidiana.

Na Figura 21, nota-se o aumento expressivo da acurácia média de 1 até 5 preditores; após isso, até 9 preditores, o ganho é relativamente pequeno; em seguida, o valor da métrica começa a decair. Assim, pelo princípio da parcimônia, em que o mais simples deve ser o selecionado, tem-se k igual a 1 e sete preditores aleatórios com acurácia de 92%.

Diante dessas três abordagens, foi realizado uma comparação usando outras métricas de performance, além da acurácia.

6.1.2.4. Comparação das três abordagens

Nos itens anteriores, foram apresentadas três abordagens para a construção do sensor virtual. Primeiro, usando somente uma única variável, depois, com o melhor sub-conjunto de preditores, selecionados a partir da regressão linear, e por último, um comitê de modelos com seleção de variáveis

de forma aleatória. Assim, nesta sessão, será discutida a comparação entre essas três abordagens de obtenção de modelos k-NN.

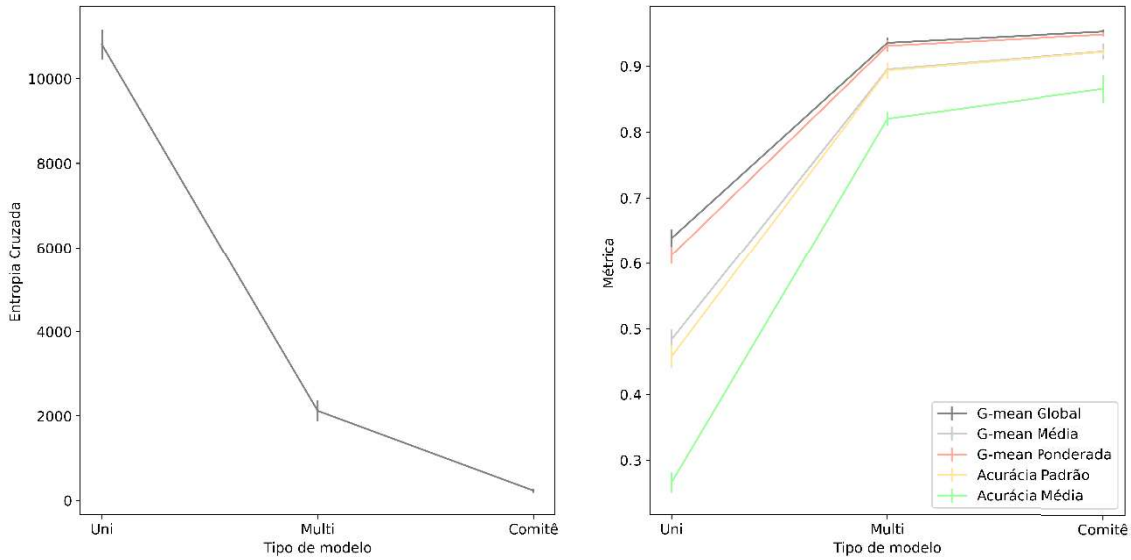
No geral, o aumento da quantidade de vizinhos mais próximos não demonstra diferenças significativas no resultado da acurácia, por isso, serão apresentados somente os resultados para os modelos com k igual à 1. A Tabela 6 relembra as melhores características selecionadas por cada abordagem, e se faz importante ressaltar que os dados foram divididos de forma aleatória em 75% para treinamento e 25% para teste, e que as métricas de performance resultam de médias após 5 execuções dos algoritmos para cada abordagem.

Tabela 6 – Parâmetros de cada abordagem do modelo com o k-NN.

Univariável	**Melhor sub-conjunto**	**Comitê de modelos**
Vazão de ar secundário	13 variáveis	7 variáveis aleatórias
k = 1	k = 1	100 modelos
Distância de Manhattan	Distância de Manhattan	k = 1
		Distância de Manhattan

No Apêndice, encontram-se os valores médios e os respectivos desvios-padrão de todas as métricas, juntamente com a matriz confusão média de cada abordagem.

A Figura 22 exibe os gráficos da função custo, acurácia e média geométrica (*g-mean*), e no eixo x estão representados os tipos de abordagem, em que “Uni” refere-se ao modelo k-NN com uma única variável de entrada, “Multi”, ao modelo com todas as 13 variáveis, e “Comitê”, ao conjunto de modelos com 7 variáveis aleatórias.

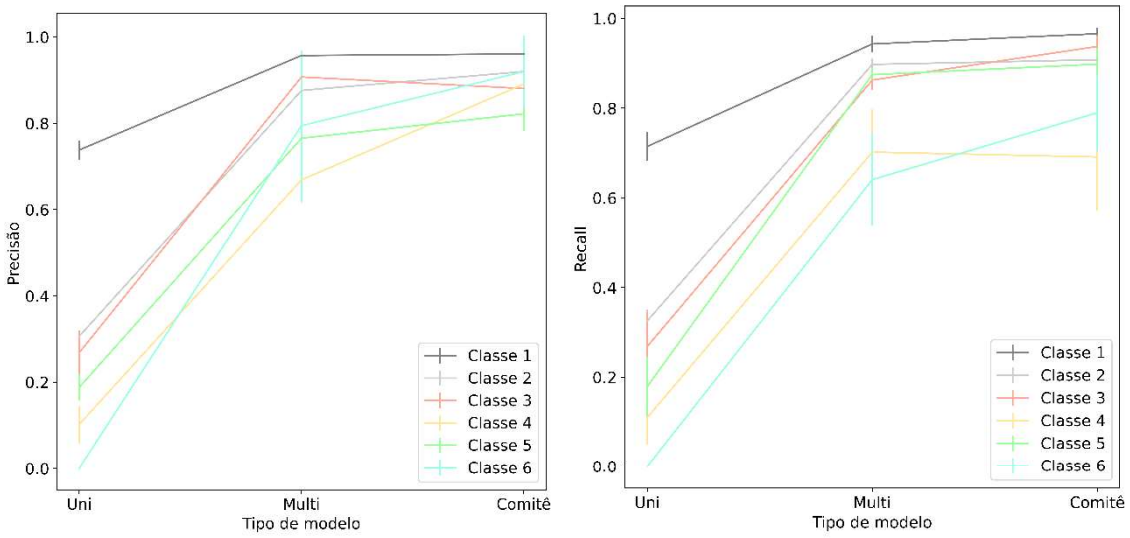


(a) (b)

Figura 22 – Médias com desvio-padrão das métricas em função do tipo de abordagem com k-NN. (a) Função Custo Entropia Cruzada (b) G-mean e Acurácia.

O comitê de modelos é mais complexo, por treinar 100 modelos k-NN, apresentando uma assertividade maior, e por isso, um valor menor para a função custo (Figura 22(a)), em comparação com as outras abordagens. Percebe-se também que para essa abordagem o desvio-padrão das métricas é, em geral, relativamente menor. Na Figura 22(b), a curva de *g-mean* global está próxima à curva de *g-mean* ponderada, pois a classe 1, com maior quantidade de observações, foi aquela com mais acertos, como mostra a Figura 23, com a precisão e *recall* de cada classe.

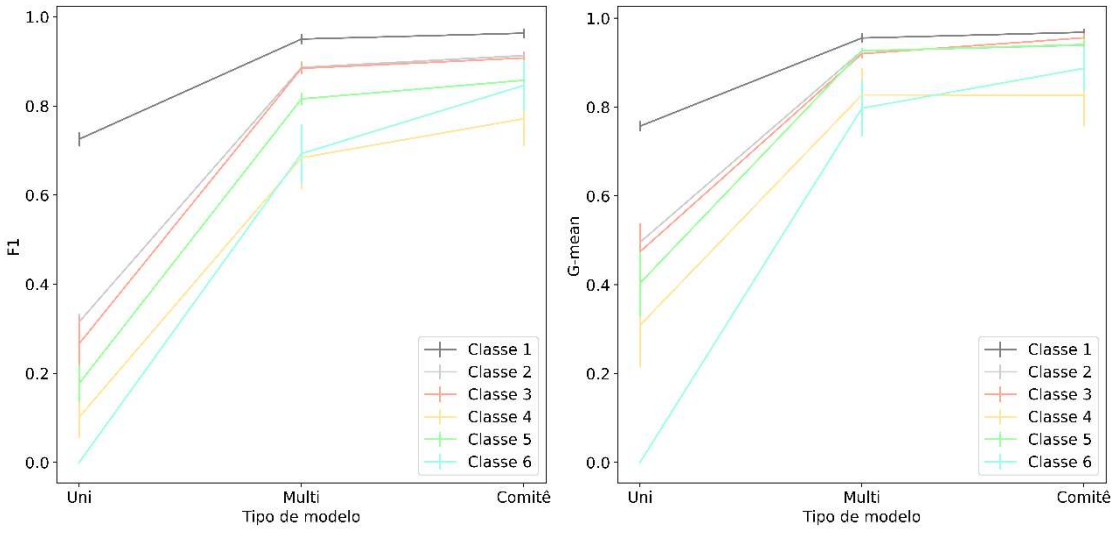
A acurácia média detém os menores valores por contabilizar a média de somente os acertos de cada classe, ou seja, uma classe com poucas observações torna seus erros mais expressivos, se comparado a uma de maior quantidade de observações.



(a) (b)

Figura 23 – Médias com desvio-padrão das métricas de cada classe em função do tipo de abordagem com k-NN. (a) Precisão (b) Recall.

Para ambas as métricas apresentadas na Figura 23, a classe 1 demonstra ser a de melhores resultados, 73% para o modelo univariável e acima de 90% para os outros dois. Avaliando somente o gráfico da precisão, nota-se que os valores para o modelo de comitê são satisfatórios em todas as classes, e que em comparação com o modelo "multi", a classe 4 obteve o melhor crescimento. Isso também ocorreu com a classe 6, para a métrica *recall*, como mostra a Figura 23(b). A métrica *F1-score* é a média harmônica entre precisão e *recall*, resumindo o comportamento das classes, como mostra a Figura 24(a).



(a) (b)

Figura 24 – Médias com desvio-padrão das métricas de cada classe em função do tipo de abordagem com k-NN. (a) *F1-score* (b) *G-mean.*

Nos valores da métrica *F1-score*, nota-se o maior crescimento entre as abordagens “multi” e “comitê” para as classes 4 e 6 em comparação com as outras classes, e que também elas são as de menores valores, ou seja, menor assertividade, como também mostra o gráfico de *g-mean* (Figura 24(b)). Deve-se ressaltar que as classes 4 e 6 são as com menores quantidades de observações na base de dados do processo.

Como a métrica *g-mean* relaciona a sensibilidade da classe com a sua sensitividade, têm-se as classes 2, 3 e 5 com as curvas quase sobrepostas, mostrando que os modelos aprenderam sobre elas de forma semelhante, isto é, o grau de assertividade para elas é similar.

Em suma, o comitê de modelos apresentou-se como a abordagem mais satisfatória, com uma média geométrica ponderada de 94,75% e uma acurácia de 92%. Para o modelo univariável, que consistiu no uso da vazão de ar secundário, para a classe 1 (com maior quantidade de observações), obtiveram-se valores satisfatórios de precisão e *recall*, iguais a 73,78% e 71,44% respectivamente. Esses resultados refletiram na média geométrica global, que foi de 63,89%, em comparação com o *g-mean* médio de 48,37%, que calcula a média simples de *g-mean* para cada classe. Para o modelo de melhor sub-conjunto de preditores, que consistiu no k-NN aplicado a todas as variáveis, a classe 4 se destacou por apresentar melhores resultados que no modelo de comitê, apesar da diferença ter sido inferior a 2%.

O propósito de apresentar neste trabalho mais do que as métricas de performance usuais, foi de comparar o entendimento e o desempenho de cada uma para um modelo com seis classes, o que, como descrito anteriormente, não é usual nos trabalhos de processos químicos. No geral, independentemente da métrica, os resultados foram melhores na abordagem “comitê”, e em específico, para a classe 1. Porém, deve-se ressaltar que cada métrica explicou e demonstrou uma característica específica resultante do modelo. Não há como selecionar as melhores no uso avaliativo de técnicas classificatórias, mas pode-se afirmar que a avaliação de todas enriquece a pesquisa.

Como próximos passos neste estudo, nota-se o desempenho inferior das classes 4 e 6, se comparadas com as demais classes. Talvez por possuírem as menores quantidades de observações. Um caminho a seguir seria a geração de dados sintéticos, através de técnicas comumente apresentadas na literatura,

como SMOTE (do inglês, *synthetic minority oversampling technique*) ou por abordagens pouco exploradas, como pela rede neural artificial Redes Adversárias Generativas (GAN, do inglês *Generative Adversarial Network*).

7. CONCLUSÃO

O uso de máquinas, equipamentos e instrumentos, possibilitou a padronização do processo, maiores eficiências com menores desperdícios, gerando maiores lucros. E a introdução da “produção sustentável” e de regulamentações ambientais proporcionaram aperfeiçoamentos nos processos, como por exemplo, o reaproveitamento dos resíduos, os menores danos ao meio ambiente, e de forma indireta, mais lucros, proporcionando valor agregado a todos, ou quase todos os subprodutos. Assim, cada vez mais, as indústrias buscam formas de aumentar a produção com menores custos, insumos e resíduos, ao mesmo tempo que mantêm a qualidade.

O objetivo deste trabalho é, em suma, analisar dados industriais do processo de recuperação química nas indústrias de celulose e desenvolver uma nova abordagem classificatória, com o foco em controle de emissão de gases nocivos a partir de técnicas de aprendizado de máquina. Foram investigadas as características do licor preto e dos ares de combustão, de uma caldeira do processo *kraft* de celulose.

O mercado oferece sensores físicos para a emissão de gases; porém, com um custo muito acima do que a maioria das indústrias pode arcar, além de exigir manutenção do dispositivo e de sua lógica. Daí a motivação para a construção de um sensor virtual eficiente, a fim de medir a emissão de dióxido de enxofre.

Foi utilizada a técnica k-vizinhos mais próximos - um algoritmo de aprendizado de máquina supervisionado de simples implementação – em três abordagem distintas para comparação. O comitê de modelos, nesse caso, uma abordagem de *ensemble learning* por média, se mostrou o mais eficiente, pois, como descrito na seção anterior, para um conjunto de dados considerável, que não fez parte da construção do modelo, obteve-se uma média geométrica, ponderada pelo tamanho das classes, de 94,75% e acurácia de 92%.

REFERÊNCIAS BIBLIOGRÁFICAS

ADAMS, D. et al. Prediction of SOx–NOx emission from a coal-fired CFB power plant with machine learning: Plant data learned by deep neural network and least square support vector machine. **Journal of Cleaner Production**, v. 270, p. 122310, out. 2020.

AJAMI, A.; DANESHVAR, M. Data driven approach for fault detection and diagnosis of turbine in thermal power plant using Independent Component Analysis (ICA). **International Journal of Electrical Power & Energy Systems**, v. 43, n. 1, p. 728–735, dez. 2012.

ALIBABA. **Mbar De Pressão Calibre,Medidor De Pressão De Medidor De Pressão Baixa Cápsula Cápsula 100 Milímetros - Buy Mbar Pressure Gauge,Low Pressure Gauge,Capsule Pressure Gauge Product on Alibaba.com**. E-commerce. Disponível em: <https://portuguese.alibaba.com/product-detail/mbar-pressure-gauge-capsule-pressure-gauge-low-capsule-pressure-gauge-100mm-60834124332.html?spm=a2700.galleryofferlist.normal_offer.d_image.7abb221a3TZRFw&s=p.%20Acessado%20em%20Maio%20de%202021.>. Acesso em: 26 jul. 2021.

ALMEIDA, G. M.; PARK, S. W. Big Data Analytics em Engenharia Química. **Uma Engenharia Química 4.0 - Revista Brasileira de Engenharia Química**, v. 33, n. 1, p. 6, 2017.

ALMEIDA, L. M. L.; BALANCO, P.; DANTAS, E. “Gestão” da dívida pública e bloco no poder: uma análise comparativa entre os meses de governo FHC, Lula e Dilma — Outubro Revista. **Revista Outubro**, v. 25, 2016.

BACHNAS, A. A. et al. A review on data-driven linear parameter-varying modeling approaches: A high-purity distillation column case study. **Journal of Process Control**, v. 24, n. 4, p. 272–285, abr. 2014.

BAJPAI, P. **Environmentally Friendly Production of Pulp and Paper: Bajpai/Pulp and Paper Production**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2010.

BALRAM, D.; LIAN, K.-Y.; SEBASTIAN, N. A novel soft sensor based warning system for hazardous ground-level ozone using advanced damped least squares neural network. **Ecotoxicology and Environmental Safety**, v. 205, p. 111168, dez. 2020.

BARANDELA, R. et al. Strategies for learning in class imbalance problems. **Pattern Recognition**, v. 36, n. 3, p. 849–851, mar. 2003.

BELISÁRIO, A. B. **ANÁLISE DE EMISSÕES EM CALDEIRAS DE RECUPERAÇÃO QUÍMICA DE FÁBRICAS DE CELULOSE KRAFT: PREDIÇÃO E ANÁLISE DE SENSIBILIDADE COM REDES NEURAIS ARTIFICIAIS**. BELO HORIZONTE - MG: UNIVERSIDADE FEDERAL DE MINAS GERAIS, 2020.

BISHOP, C. M. **Pattern recognition and machine learning**. New York: Springer, 2006.

BONTHU, S.; HIMA BINDU, K. Review of Leading Data Analytics Tools. **International Journal of Engineering & Technology**, v. 7, n. 3.31, p. 10, 24 ago. 2018.

BOUZENAD, K.; RAMDANI, M. Multivariate Statistical Process Control Using Enhanced Bottleneck Neural Network. **Algorithms**, v. 10, n. 2, p. 49, 29 abr. 2017.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, ago. 1996.

BREIMAN, L. Pasting Small Votes for Classification in Large Databases and On-Line. **Machine Learning**, v. 36, n. 1/2, p. 85–103, 1999.

CANETE, J. F. et al. Dual composition control and soft estimation for a pilot distillation column using a neurogenetic design. **Computers & Chemical Engineering**, v. 40, p. 157–170, maio 2012.

CARVALHO, M. DA G. V. DE S. **Efeito das variáveis de cozimento nas características químicas de pastas kraft de eucalyptus globulus.** Coimbra: Universidade de Coimbra, 1999.

CASTRO, H. **Processos Químicos Industriais II Apostila 4 PAPEL E CELULOSE**UNIVERSIDADE DE SÃO PAULO Esco la de Engenharia de Lorena, 2009.

CHENG, T. et al. Monitoring Influent Measurements at Water Resource Recovery Facility Using Data-Driven Soft Sensor Approach. **IEEE Sensors Journal**, v. 19, n. 1, p. 342–352, 1 jan. 2019.

COSTA, A. O. S.; BISCAIA, E. C.; LIMA, E. L. Chemical Composition Determination at the Bottom Region of a Recovery Boiler Furnace by Direct Minimization of Gibbs Free Energy. **The Canadian Journal of Chemical Engineering**, v. 83, n. 3, p. 477–484, 19 maio 2008.

CURRERI, F.; FIUMARA, G.; XILIBIA, M. Input selection methods for data-driven Soft sensors design: Application to an industrial process. **Information Sciences**, v. 537, p. 1–17, 1 out. 2020.

DAMINELLI, M. E. P. **Sustentabilidade e a indústria química: estratégia e competividade no mercado interno e externo**. Trabalho de Conclusão de Curso—Criciúma - SC: Universidade do Extremo Sul Catarinense, 11 mar. 2019.

DARÉ ALVES, É. et al. Estudo do processo de obtenção celulose Kraft com ênfase no forno de cal. **Revista Liberato**, v. 16, n. 26, p. 205–218, 2015.

DENG, Z. et al. Efficient k NN classification algorithm for big data. **Neurocomputing**, v. 195, p. 143–148, jun. 2016.

FENG, J.; LI, K. MRS-kNN fault detection method for multirate sampling process based variable grouping threshold. **Journal of Process Control**, v. 85, p. 149–158, jan. 2020.

GIBERT, K. et al. Which method to use? An assessment of data mining methods in Environmental Data Science. **Environmental Modelling & Software**, v. 110, p. 3–27, dez. 2018.

GOLDBERGER, J. et al. Neighbourhood Components Analysis. **Advances in Neural Information Processing Systems**, n. 17, p. 8, 2004.

GUL, A. et al. Ensemble of a subset of kNN classifiers. **Advances in Data Analysis and Classification**, v. 12, n. 4, p. 827–840, dez. 2018.

HAMPEL, F. R. **A general qualitative definition of robustness**. Annals of Mathematics Statistics. **Anais**...: 6.1971. Disponível em: <http://scholar.google.com.br/scholar_url?url=https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-42/issue-6/A-General-Qualitative-Definition-of-Robustness/10.1214/aoms/1177693054.pdf&hl=pt-BR&sa=X&ei=D-XhYP_qG4jCmwG3s6r4Cg&scisig=AAGBfm2XsVt5SrqiWBJC89ri_sTJy21OQA&nossl=1&oi=scholarr>

HARROU, F.; ZEROUAL, A.; SUN, Y. Traffic congestion monitoring using an improved kNN strategy. **Measurement**, v. 156, p. 107534, maio 2020.

HASTIE, T.; TIBSHIRANI, R.; TIBSHIRANI, R. J. Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso. 29 jul. 2017.

ILIYAS, S. A. et al. RBF neural network inferential sensor for process emission monitoring. **Control Engineering Practice**, v. 21, n. 7, p. 962–970, jul. 2013.

JAIN, A.; LELLA, R. L. **Pearson Correlation Coefficient Based Attribute Weighted k-NN for Air Pollution Prediction**. 2020 IEEE 17th India Council International Conference (INDICON). **Anais**... In: 2020 IEEE 17TH INDIA COUNCIL INTERNATIONAL CONFERENCE (INDICON). New Delhi, India: IEEE, 10 dez. 2020. Disponível em: <https://ieeexplore.ieee.org/document/9342275/>. Acesso em: 24 jul. 2021

JAMES, G. et al. **An Introduction to Statistical Learning**. New York, NY: Springer New York, 2013. v. 103

KADLEC, P.; GABRYS, B.; STRANDT, S. Data-driven Soft Sensors in the process industry. **Computers & Chemical Engineering**, v. 33, n. 4, p. 795–814, abr. 2009.

KANO, M.; FUJIWARA, K. Virtual Sensing Technology in Process Industries: Trends and Challenges Revealed by Recent Industrial Applications. **JOURNAL OF CHEMICAL ENGINEERING OF JAPAN**, v. 46, n. 1, p. 1–17, 2013.

KLUYVER, T. et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. **University of Southampton Institutional Repository**, 2016.

LI, D.; MENG, N.; SONG, T. Learning control of fermentation process with an improved DHP algorithm. **Chinese Journal of Chemical Engineering**, v. 24, n. 10, p. 1399–1405, out. 2016.

LIMA, R. N. et al. Trend modelling with artificial neural networks. Case study: Operating zones identification for higher SO 3 incorporation in cement clinker. **Engineering Applications of Artificial Intelligence**, v. 54, p. 17–25, set. 2016.

LIU, H.; SHAH, S.; JIANG, W. On-line outlier detection and data cleaning. **Computers & Chemical Engineering**, v. 28, n. 9, p. 1635–1647, ago. 2004.

LOTUFO, F.; GARCIA, C. **Sensores Virtuais ou Soft Sensors: Uma introdução**. . In: 7 BRAZILIAN CONFERENCE ON DYNAMICS, CONTROL AND APPLICATION. Presidente Prudente, São Paulo: set. 2008.

MADETI, S. R.; SINGH, S. N. Modeling of PV system based on experimental data for fault detection using kNN method. **Solar Energy**, v. 173, p. 139–151, out. 2018.

MELE, M.; MAGAZZINO, C. A Machine Learning analysis of the relationship among iron and steel industries, air pollution, and economic growth in China. **Journal of Cleaner Production**, v. 277, p. 123293, dez. 2020.

MENEZES, U. G. et al. MANAGEMENT INNOVATION FOR SUSTAINABLE DEVELOPMENT: BEHAVIOR AND REFLECTIONS ON THE CHEMICAL INDUSTRY. **Review of Administration and Innovation - RAI**, v. 8, n. 4, p. 88–116, 27 jan. 2012.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. Rio de Janeiro (RJ): LTC, 2009.

MORAES, F. **Modelo para avaliação do consumo específico de madeira e insumos energéticos no processo de produção de celulose e papel**. Mestrado Profissional em Engenharia de Produção—Araraquarara - SP: Centro Universitário de Araraquara, 2011.

MOSAVI, A. et al. Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction. **Water Resources Management**, v. 35, n. 1, p. 23–37, jan. 2021.

NEVES, T. G. et al. Intelligent control system for extractive distillation columns. **Korean Journal of Chemical Engineering**, v. 35, n. 4, p. 826–834, abr. 2018.

OSORIO, D. et al. Soft-sensor for on-line estimation of ethanol concentrations in wine stills. **Journal of Food Engineering**, v. 87, n. 4, p. 571–577, ago. 2008.

PANDYA, D. H.; UPADHYAY, S. H.; HARSHA, S. P. Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN. **Expert Systems with Applications**, v. 40, n. 10, p. 4137–4145, ago. 2013.

PANI, A. K.; MOHANTA, H. K. A Survey of Data Treatment Techniques for Soft Sensor Design. **Chemical Product and Process Modeling**, v. 6, n. 1, 3 jan. 2011.

PAULA, K. R. DE. **Análise da sulfidez no processo kraft em uma indústria de celulose**. Trabalho de Conclusão de Curso (Engenharia Química)—Ponta Grossa - PR: Universidade Tecnológica Federal do Paraná, 10 nov. 2017.

PEARSON, R. K. Outliers in process modeling and identification. **IEEE Transactions on Control Systems Technology**, v. 10, n. 1, p. 55–63, jan. 2002.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, n. 85, p. 2825–2830, 2011.

PROVOST, F.; FAWCETT, T. **Data science for business: what you need to know about data mining and data-analytic thinking**. 1. ed., 2. release ed. Beijing Köln: O'Reilly, 2013.

RAJA KUMAR, J. R.; PANDEY, R. K.; SARKAR, B. K. Pollutant Gases Detection using the Machine learning on Benchmark Research Datasets. **Procedia Computer Science**, v. 152, p. 360–366, 2019.

RAUSCHER, J.; KAILA, J.; JAAKKOLA, H. **State of the Art Evaporation System**Atlanta GA, 2006.

REZAZADEH, A. Environmental Pollution Prediction of NOx by Predictive Modelling and Process Analysis in Natural Gas Turbine Power Plants. **Pollution**, v. 7, n. 2, abr. 2021.

RUSSELL, S. J. et al. **Artificial intelligence: a modern approach**. Third edition, Global edition ed. Boston Columbus Indianapolis New York San Francisco Upper Saddle River Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo: Pearson, 2016.

SAINLEZ, M.; HEYEN, G. Comparison of supervised learning techniques for atmospheric pollutant monitoring in a Kraft pulp mill. **Journal of Computational and Applied Mathematics**, v. 246, p. 329–334, jul. 2013.

SANTOS, I. L. DOS; SANTOS, R. C. DOS; SILVA JUNIOR, D. DE S. Análise da Indústria 4.0 como Elemento Rompedor na Administração de Produção. **Future Studies Research Journal: Trends and Strategies**, v. 11, n. 1, p. 48–64, 6 jan. 2019.

SARMADI, H.; KARAMODIN, A. A novel anomaly detection method based on adaptive Mahalanobis-squared distance and one-class kNN rule for structural health monitoring under environmental effects. **Mechanical Systems and Signal Processing**, v. 140, p. 106495, jun. 2020.

SHARMA, M. et al. Forecasting And Prediction Of Air Pollutants Concentrates Using Machine Learning Techniques: The Case Of India. **IOP Conference Series: Materials Science and Engineering**, v. 1022, p. 012123, 19 jan. 2021.

SILVA, R. M. DA. Estudo de aumento de eficiência e produção de uma caldeira de recuperação química. **Aleph**, p. 59 f., 9 dez. 2016.

SLIDE TEAM. **Data Science Lifecycle Problem Learning Model Deployment**, set. 2020. Disponível em: <https://www.slideteam.net/data-science-lifecycle-problem-learning-model-deployment.html>. Acesso em: 1 set. 2020

SONG, B. et al. Fault detection and diagnosis via standardized k nearest neighbor for multimode process. **Journal of the Taiwan Institute of Chemical Engineers**, v. 106, p. 1–8, jan. 2020.

SUN, K. et al. Development of a new multi-layer perceptron based soft sensor for SO2 emissions in power plant. **Journal of Process Control**, v. 84, p. 182–191, dez. 2019.

TAMMINEN, A.; TAMMINEN, T. Sulfur emissions from kraft recovery boilers – a short review of measurement techniques and boiler characteristics on SO2 and TRS emissions. 2015.

THAM, M. T. et al. Soft-sensors for process estimation and inferential control. **Journal of Process Control**, v. 1, n. 1, p. 3–14, jan. 1991.

TIMMER, T. **Produção de ácido sulfúrico a partir de gases odoríferos não condensáveis – Uma nova abordagem para a indústria de papel e celulose**. Curitiba - PR: Valmet, 2020. Disponível em: <www.quimica.com.br/producao-de-acido-sulfurico-a-partir-de-gases-odoriferos>.

VAKKILAINEN, E. K. Recovery Boiler. In: **Papermaking Science - Technology Book 6B**. Helsinki - Finland: Fapet Oy, 2000. p. 95.

VAN ROSSUM, G. **Python tutorial**. Amsterdam - Netherlands: Centrum voor Wiskunde en Informatica (CWI), 1 jan. 1995. Disponível em: <https://ir.cwi.nl/pub/5007>. Acesso em: 26 jul. 2021.

YANG, J. et al. Data validation of multifunctional sensors using independent and related variables. **Sensors and Actuators A: Physical**, v. 263, p. 76–90, ago. 2017.

YANG, T. et al. Real-time dynamic prediction model of NOx emission of coal-fired boilers under variable load conditions. **Fuel**, v. 274, p. 117811, ago. 2020.

YUN, D. et al. Developing a deep learning model for the simulation of micro-pollutants in a watershed. **Journal of Cleaner Production**, v. 300, p. 126858, jun. 2021.

ZHOU, Z.-H. **Ensemble methods: foundations and algorithms**. Boca Raton, FL: Taylor & Francis, 2012.

APÊNDICE

Apresenta-se nas Tabelas 1, 2 e 3, as matrizes confusão das abordagens univariável, multivariável e comitê de modelos, respectivamente.

Tabela 1 – Matriz Confusão média para o modelo, com vazão de ar secundário, k-NN com k = 1 e métrica de distância de Manhattan, treinado e testado 5 vezes.

		Predito					
		1	2	3	4	5	6
Real	1	180,4	52,8	14	2,4	2,6	0,4
	2	46,8	46,8	28	10,6	9	2,4
	3	11,4	32	26,2	11	13,2	3,8
	4	1,6	9,4	10,4	3,6	5,2	2,2
	5	3	9,6	13,4	5,8	7,2	1,2
	6	1,4	2	4,8	1	1,4	0

Tabela 2 – Matriz Confusão média para o modelo k-NN com as 13 variáveis, k = 1 e métrica se distância de Manhattan, treinado e testado 5 vezes.

		Predito					
		1	2	3	4	5	6
Real	1	234,2	12,4	1,4	0	0,2	0,2
	2	8,6	131,4	4,2	1,8	0,4	0
	3	1,2	4,6	87,8	7	1,2	0
	4	0	1	3	23,2	5,8	0,2
	5	0,2	0,8	0,4	2,4	33,4	1
	6	0,6	0	0	0,2	2,6	5,6

Tabela 3 – Matriz Confusão média para o modelo, com 100 modelos e 7 variáveis selecionadas aleatoriamente, k-NN com k = 1 e métrica de distância de Manhattan, treinado e testado 5 vezes.

		Predito					
		1	2	3	4	5	6
Real	1	238,4	7,6	0,2	0	0,6	0
	2	7,6	136	5,4	0	1	0
	3	2,2	3,2	89,4	0,6	0	0
	4	0	0,4	6,4	23,4	3,8	0,2
	5	0	0,8	0,2	2,4	35,4	0,6
	6	0	0	0	0	2,4	8,8

A Tabela 4 apresenta os valores médios das métricas analisadas nos três modelos.

Tabela 4 – Valores médios e desvio-padrão das métricas avaliadas.

Métrica	Comitê de modelos		Melhor Sub-conjunto		Univariável	
	Média	Desvio-padrão	Média	Desvio-padrão	Média	Desvio-padrão
Precisão classe 1	0,9604	0,0157	0,9569	0,0119	0,7378	0,0226
Precisão classe 2	0,9197	0,0320	0,8759	0,0317	0,3061	0,0123
Precisão classe 3	0,8803	0,0318	0,9071	0,0165	0,2680	0,0524
Precisão classe 4	0,8906	0,0460	0,6687	0,0530	0,1018	0,0428
Precisão classe 5	0,8218	0,0391	0,7651	0,0318	0,1878	0,0299
Precisão classe 6	0,9196	0,0841	0,7942	0,1744	0,0000	0,0000
Precisão média	0,8987	0,0134	0,8280	0,0374	0,2669	0,0114
Precisão global	0,9210	0,0051	0,8936	0,0127	0,4579	0,0177
Precisão ponderada	0,9230	0,0048	0,8966	0,0131	0,4643	0,0182
Recall classe 1	0,9659	0,0132	0,9431	0,0181	0,7144	0,0323
Recall classe 2	0,9078	0,0303	0,8975	0,0129	0,3254	0,0252
Recall classe 3	0,9375	0,0254	0,8626	0,0218	0,2682	0,0772
Recall classe 4	0,6911	0,1202	0,7019	0,0961	0,1097	0,0600
Recall classe 5	0,8981	0,0360	0,8745	0,0143	0,1788	0,0672
Recall classe 6	0,7897	0,0847	0,6405	0,1014	0,0000	0,0000
Recall média	0,8650	0,0214	0,8200	0,0107	0,2661	0,0158
Recall global	0,9210	0,0051	0,8936	0,0127	0,4579	0,0177
Recall ponderada	0,9210	0,0051	0,8936	0,0127	0,4579	0,0177
F1 classe 1	0,9631	0,0108	0,9499	0,0112	0,7253	0,0164
F1 classe 2	0,9129	0,0053	0,8862	0,0146	0,3153	0,0179
F1 classe 3	0,9074	0,0149	0,8841	0,0130	0,2664	0,0602
F1 classe 4	0,7711	0,0614	0,6836	0,0707	0,1015	0,0470
F1 classe 5	0,8573	0,0194	0,8157	0,0146	0,1772	0,0423
F1 classe 6	0,8458	0,0572	0,6928	0,0664	0,0000	0,0000
F1 média	0,8763	0,0160	0,8187	0,0171	0,2643	0,0136
F1 global	0,9210	0,0051	0,8936	0,0127	0,4579	0,0177
F1 ponderada	0,9198	0,0056	0,8940	0,0126	0,4597	0,0164
G-mean classe 1	0,9682	0,0088	0,9553	0,0108	0,7568	0,0116
G-mean classe 2	0,9392	0,0105	0,9264	0,0067	0,4954	0,0151
G-mean classe 3	0,9558	0,0112	0,9199	0,0113	0,4741	0,0648
G-mean classe 4	0,8265	0,0704	0,8273	0,0595	0,3088	0,0943
G-mean classe 5	0,9406	0,0177	0,9262	0,0067	0,4036	0,0739
G-mean classe 6	0,8870	0,0480	0,7972	0,0638	0,0000	0,0000
G-mean média	0,9219	0,0116	0,8951	0,0065	0,4837	0,0152
G-mean global	0,9521	0,0032	0,9352	0,0078	0,6389	0,0136
G-mean ponderada	0,9475	0,0032	0,9306	0,0089	0,6128	0,0135
Entropia Cruzada	230,39	49,62	2120,68	252,16	10803,73	352,57
Acurácia padrão	0,9210	0,0051	0,8936	0,0127	0,4579	0,0177
Acurácia ponderada	0,8650	0,0214	0,8200	0,0107	0,2661	0,0158