UNIVERSIDADE FEDERAL DE VIÇOSA

MODELAGEM DE ANALISADORES VIRTUAIS (SOFT SENSORS) PARA O PROCESSO DE POLIMERIZAÇÃO DE POLIETILENO EM ALTA PRESSÃO

Gabriel Alves da Silva Abreu
Magister Scientiae

VIÇOSA - MINAS GERAIS
2025

GABRIEL ALVES DA SILVA ABREU

MODELAGEM DE ANALISADORES VIRTUAIS (SOFT SENSORS) PARA O PROCESSO DE POLIMERIZAÇÃO DE POLIETILENO EM ALTA PRESSÃO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Química, para obtenção do título de *Magister Scientiae*.

Orientador: Fabio de Avila Rodrigues

VIÇOSA - MINAS GERAIS
2025

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
A162m
2025

Abreu, Gabriel Alves da Silva, 1999-
Modelagem e simulação de analisadores virtuais (*soft sensors*) para o processo de polimerização de polietileno em alta pressão / Gabriel Alves da Silva Abreu. – Viçosa, MG, 2025.
1 dissertação eletrônica (61 f.): il. (algumas color.).

Orientador: Fábio de Ávila Rodrigues.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Química, 2025.
Referências bibliográficas: f. 59-61.
DOI: https://doi.org/10.47328/ufvbbt.2025.244
Modo de acesso: World Wide Web.

1. Polimerização. 2. Ciência de dados. 3. Controle de processo. 4. Aprendizado do computador. 5. Análise de regressão. I. Rodrigues, Fábio de Ávila, 1981-. II. Universidade Federal de Viçosa. Departamento de Química. Programa de Pós-Graduação em Engenharia Química. III. Título.

CDD 22. ed. 547.28

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

GABRIEL ALVES DA SILVA ABREU

MODELAGEM DE ANALISADORES VIRTUAIS (SOFT SENSORS) PARA O PROCESSO DE POLIMERIZAÇÃO DE POLIETILENO EM ALTA PRESSÃO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Química, para obtenção do título de *Magister Scientiae*.

APROVADA: 30 de janeiro de 2025.

Assentimento:

Gabriel Alves da Silva Abreu
Autor

Fabio de Avila Rodrigues
Orientador

AGRADECIMENTOS

Agradeço primeiro a Deus, que sempre esteve comigo, me amparando e sendo o meu melhor amigo. Por ter me acolhido mesmo nos momentos de maior dificuldade. Por Ele ter me abençoado com saúde, capacidade, família, amigos e educação, para que eu pudesse atuar em todas as metas pessoais e profissionais. A minha mãe, Auleni Alves, ao meu pai, Adilson Emídio e sua esposa Sueli Aparecida pela companhia, parceria e apoio em todos os momentos.

Agradeço à Universidade Federal de Viçosa, por me acolher desde a graduação, oferecendo ensino de qualidade e permitindo o meu crescimento profissional.

Agradeço ao meu professor orientador Fabio de Ávila, por tantos anos me ensinando, orientando e aconselhando.

A todos que, diretamente ou indiretamente, contribuíram para a realização desse trabalho.

Este trabalho foi realizado com o apoio das seguintes agências de pesquisa brasileiras: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

RESUMO

ABREU, Gabriel Alves da Silva, M.Sc., Universidade Federal de Viçosa, janeiro de 2025. **MODELAGEM DE ANALISADORES VIRTUAIS (SOFT SENSORS) PARA O PROCESSO DE POLIMERIZAÇÃO DE POLIETILENO EM ALTA PRESSÃO**. Orientador: Fabio de Avila Rodrigues.

A evolução contínua da indústria ao longo dos anos tem sido impulsionada por avanços tecnológicos, resultando na chamada Indústria 4.0, caracterizada pela integração de coleta intensiva de dados, inteligência artificial e aprendizado de máquina. Nesse contexto, o uso de analisadores virtuais (*soft sensors*) surge como uma alternativa eficiente para o controle de processos, sendo desenvolvidos a partir de dados operacionais de processo e/ou equações fenomenológicas. A literatura oferece uma variedade de algoritmos preditivos, ajustados conforme o processo estudado, incluindo métodos supervisionados e não supervisionados, como árvores de decisão, regressões lineares e redes neurais. Este trabalho tem como objetivo avaliar o desempenho de modelos preditivos aplicados ao processo de polimerização em alta pressão para a predição do Índice de Fluidez (IF) do Polietileno de Baixa Densidade (PEBD). Os dados do processo foram fornecidos por uma empresa brasileira do setor petroquímico e tratados de forma sigilosa por meio da normalização. A análise focou em dados de produção de PEBD, divididos em 3 grupos representativos de diferentes patamares operacionais. Após a remoção de *outliers* (4 desvios padrão), foram gerados modelos *baseline* e realizadas otimizações por meio de *feature selection* e ajuste de hiperparâmetros, utilizando os algoritmos *Gradient Boosting*, *LightGBM* e LSTM. Os modelos foram avaliados com base em métricas de acurácia e RMSE. O modelo do Grupo 1 apresentou alta aderência aos dados reais, com erro relativo médio inferior a 12%, devido à maior disponibilidade de dados. Em contraste, os Grupos 2 e 3 apresentaram menor aderência, com erro relativo médio entre 20% e 30%. Os resultados destacam a importância da disponibilidade e qualidade dos dados no desempenho de modelos preditivos para aplicações industriais, além de evidenciar a possibilidade de aplicar modelos de *machine learning* na indústria de polimerização.

Palavras-chave: Ciência de dados; Controle de processos; Aprendizado de máquina; Polimerização; Modelos de regressão

ABSTRACT

ABREU, Gabriel Alves da Silva, M.Sc., Universidade Federal de Viçosa, January, 2025. **MODELING AND SIMULATION OF VIRTUAL ANALYZERS (SOFT SENSORS) FOR THE HIGH-PRESSURE POLYETHYLENE POLYMERIZATION PROCESS**. Adviser: Fabio de Avila Rodrigues.

The continuous evolution of industry over the years has been driven by technological advancements, resulting in the Industry 4.0, characterized by the integration of intensive data collection, artificial intelligence, and machine learning. In this context, the use of virtual analyzers (soft sensors) emerges as an efficient alternative for process control, developed from operational process data and/or phenomenological equations. The literature offers a variety of predictive algorithms, adjusted by the process studied, including supervised and unsupervised methods such as decision trees, linear regressions, and neural networks. This study aims to evaluate the performance of predictive models applied to the high-pressure polymerization process for predicting the Melt Index (MI) of Low-Density Polyethylene (LDPE). Process data were provided by a Brazilian company in the petrochemical sector and handled confidentially through normalization. The analysis focused on LDPE production data, divided into three groups representing different operational ranges. After outlier removal (4 standard deviations), baseline models were generated, and optimizations were performed through feature selection and hyperparameter tuning using Gradient Boosting, LightGBM, and LSTM algorithms. The models were evaluated based on accuracy and RMSE metrics. The model for Group 1 showed high adherence to real data, with a mean relative error below 12%, due to the greater data availability. In contrast, Groups 2 and 3 exhibited lower adherence, with mean relative errors between 20% and 30%. The results underscore the importance of data availability and quality in the performance of predictive models for industrial applications while also highlighting the feasibility of applying machine learning models in the polymerization industry.

Keywords: Data science; Process control; Machine learning; Polymerization; Regression models

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

BB - *Black Box*

EQM - Erro Quadrático Médio

GB - *Grey Box*

GBR - *Gradient Boosting Regressor*

IF - Índice de Fluidez

IW - *Input weight*

LSTM - *Long Short-Term Memory*

ML - *Machine Learning*

NaN - *Not a Number*

PEAD - Polietileno de Alta Densidade

PEBD - Polietileno de Baixa Densidade

PEBDL - Polietileno de Baixa Densidade Linear

PP - Polipropileno

PVC - Policloreto de vinila

RMSE - *Root Mean Square Error*

RNA - Redes Neurais Artificias

RW - *Recurrent weight*

WB - *White Box*

SUMÁRIO

1. INTRODUÇÃO

O desenvolvimento industrial destaca o melhor da tecnologia disponível, refletindo políticas, pesquisa, apoio às instituições, pesquisa, capacidade técnica dos profissionais e cultura (CHANG; BHADURI, 2003). Considerando o crescimento acelerado da indústria e as extensas mudanças nos últimos anos, conceitua-se que a indústria evoluiu de sua versão inicial 1.0 para a nova e crescente versão 4.0, repleta de tecnologia (THANGARAJ; NARAYANAN, 2018).

A primeira revolução industrial corresponde à transição do uso de mão de obra mecânica para métodos de produção utilizando água e vapor, da segunda metade do século XVIII ao início do século XIX (VINITHA *et al.,* 2020). Durante este período, o carvão foi intensamente utilizado como fonte de energia, principalmente devido a sua abundância em diversas regiões da Europa. Em contrapartida, a segunda revolução industrial foi marcada pela utilização da eletricidade como fonte primária. Além do uso de energia elétrica, observou-se a ampla utilização de políticas de gestão da produção em busca de melhor eficiência, impulsionada por políticas como *just-in-time* e manufatura enxuta (SHARMA; SINGH, 2020).

A Indústria 3.0 foi marcada pela utilização de computadores para automação industrial, com lógica de controle que exigia controle humano (SHARMA; SINGH, 2020). A Tecnologia da Informação (TI) foi introduzida e foram utilizados transistores e circuitos cada vez mais eficientes (THANGARAJ; NARAYANAN, 2018).

A Indústria 4.0 representa a quarta revolução industrial, caracterizada pela integração de tecnologias avançadas que transformam os sistemas de produção. Essa revolução é impulsionada por sistemas integrados de produção, IoT (Internet das Coisas), inteligência artificial e big data, permitindo uma produção autônoma e interconectada (BARRETO; AMARAL; PEREIRA, 2017). A capacidade de armazenamento e análise massiva de dados, aliada ao uso de máquinas inteligentes treinadas para operar com mínima ou nenhuma assistência humana, viabiliza processos altamente eficientes e flexíveis. Essa abordagem promove um ambiente de manufatura inteligente, onde os sistemas são capazes de tomar decisões em tempo real, otimizando recursos e reduzindo custos operacionais (THANGARAJ; NARAYANAN, 2018).

À medida que a indústria cresceu, a ciência de dados tornou-se complementar a diversos segmentos, facilitando cada vez mais a vida humana (MAHESH, 2019). O conceito de ciência de dados não é recente, datado por volta da década de 1960 para

descrever uma nova profissão capaz de suportar e compreender um grande conjunto de dados recolhidos e armazenados (HAYASHI *et al.,*1998). Segundo HAYASHI *et al.* (1998), é possível definir o estudo da ciência de dados em três tópicos: planejamento amostral (desenho de dados), coleta e armazenamento de dados e análise de dados.

Para o desenho dos dados é importante questionar, entender o problema a ser resolvido e definir quais dados devem ser coletados. Para a coleta de dados é necessário entender como coletar dados, evitar dados com vieses, além de armazená-los evitando a perda de qualidade (HAYASHI *et al.,*1998). Existe, portanto, a expectativa que a análise dos dados seja realizada com métodos claros e sem construções matemáticas muito sofisticadas, havendo um foco na compreensão dos dados, representação e significado. A ciência de dados começou com matemática e estatística, mas incluiu práticas e conceitos de inteligência artificial (MAHESH, 2019). Dentro deste contexto, o *machine learning* tem sido utilizado para ensinar máquinas a lidar com dados de forma eficiente (JAEHYUN *et al.*, 2021). O objetivo do *machine learning* é aprender com os dados de entrada, seja por um conceito previamente definido pelo usuário ou na busca por padrão de regra de negócio (JAEHYUN *et al.*, 2021). O crescente volume de dados, o rápido processamento computacional e o armazenamento adequado de dados têm impulsionado cada vez mais o uso do *machine learning* (FANG *et. al.,* 2022).

De maneira similar, a indústria de polímeros vem destacando-se por atender as demandas de materiais da sociedade, com ampla produção de polímeros como o PP, PVC e PEBD (CANEVAROLO, 2013). Desse modo, têm-se a oportunidade de utilizar práticas de ciência de dados e *machine learning* para otimizar o processo de polimerização e reduzir a produção *off-spec*. Algoritmos de *machine learning* podem ajudar no controle de qualidade em tempo real e na automação das indústrias, tornando a produção mais eficiente e reduzindo o desperdício. O Índice de Fluidez (IF) é o principal parâmetro utilizado para definir a qualidade e o grupo do polímeros produzido, sendo que os polímeros produzidos fora do range de IF são considerados *off-spec* (GUERREIRO, 2012).

Desse modo, o objetivo deste estudo é apresentar como a ciência de dados e a indústria 4.0 se relacionam, baseado no desenvolvimento de um analisador virtual para a indústria de polimerização, prevendo o IF. Neste contexto, serão aplicados algoritmos de *machine learning* para a predição do IF a partir de dados industriais coletados ao longo do tempo.

2. OBJETIVOS

2.1 OBJETIVO GERAL

Este trabalho tem como objetivo desenvolver a modelagem do processo de polimerização de Polietileno de Baixa Densidade (PEBD) em reatores de alta pressão, utilizando dados de sensores coletados ao longo do tempo. A partir dessa modelagem, busca-se criar um *soft sensor* capaz de predizer o IF do polímero.

2.2 OBJETIVOS ESPECÍFICOS

- Efetuar o tratamento e verificação da qualidade dos dados de processo, utilizando de métodos estatísticos como a remoção de *outliers* a partir da análise de desvio padrão;

- Realizar o estudo dos principais modelos preditivos disponíveis nas bibliotecas (modelos de árvores, regressões, redes neurais etc.);

- Estudar o processo produtivo de polimerização a partir de reatores de alta pressão, compreendendo quais os principais equipamentos de processo que interferem no Índice de Fluidez (IF);

- Selecionar métricas capazes de avaliar corretamente o modelo preditivo.

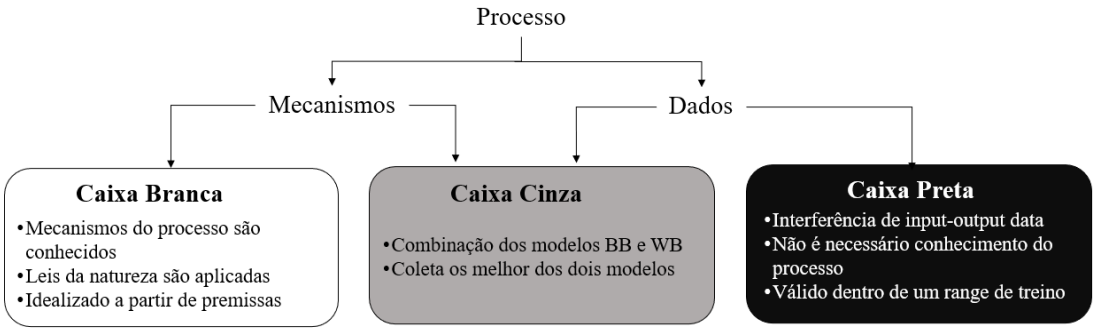
3. REVISÃO DE LITERATURA

3.1 *SOFT SENSORS*

Segundo KADLEC; GABRYS; STRANDT (2009), os dados de sensores de *hardware* são amplamente coletados, mas foi apenas há três décadas que a ideia de desenvolver modelos de previsão se expandiu. Modelos preditivos treinados a partir de dados são conhecidos como *soft sensors*, uma combinação de "*software*" (usando programas de computador) e "sensores", pois tem finalidade semelhante aos sensores de *hardware* implementados (AHMAD *et al.,* 2020). Esses sensores podem ser utilizados para medir parâmetros necessários ao controle do processo que são difíceis de medir em sensores convencionais, ou são caros ou mesmo, funcionam com atraso (AHMAD *et al.,* 2020). O *soft sensor* pode substituir o uso de sensores de hardware, mas em alguns casos pode ser utilizado como complemento (SUN *et al.,* 2014).

Soft sensor usam dados brutos de sensores de *hardware* ou são calculados por equações fenomenológicas. Eles podem ser classificados em três grupos, separados pela estrutura dos dados: modelos caixa branca (WB), caixa preta (BB) e caixa cinza (GB). Os modelos WB são aqueles baseados apenas no conhecimento do processo e em equações fenomenológicas. Dessa forma, utilizam conceitos como modelos mecanicistas, analíticos, fenomenológicos, físicos e paramétricos, equacionados de acordo com as leis da ciência e da engenharia (ZENDEHBOUDI; REZAEI; LOHI, 2018). Os modelos BB são construídos a partir da coleta de dados de processos, desenvolvidos sem explicação física clara e possuem correlação principalmente estatística. Porém, os modelos BB podem não capturar a dinâmica dos processos industriais e não se adaptar às diferentes condições do processo (AHMAD *et al.,* 2020), além de serem modelos que necessitam de muitos dados de treinamento (ZENDEHBOUDI; REZAEI; LOHI, 2018). Os modelos GB integram os modelos WB e BB e são conhecidos como modelos híbridos (HM). A classificação dos modelos é apresentada na Figura 1.

Figura 1 - Tipos de *soft sensors*



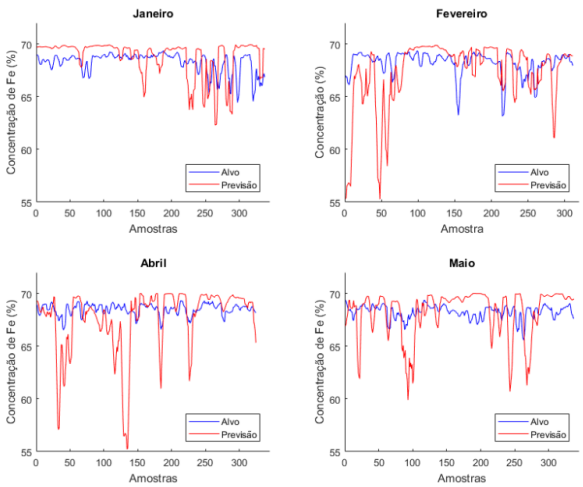
Fonte: ZENDEHBOUDI; REZAEI; LOHI, 2018.

3.2 APLICAÇÃO DE *SOFT SENSORS* EM PROCESSOS INDUSTRIAIS

O estudo de GUEDES (2020) investigou a aplicação de um modelo baseado em redes neurais artificiais e *random forest* para a predição do teor de ferro no concentrado de flotação de minério de ferro, com fomento do Instituto Tecnológico Vale (ITV). Para o estudo, foram coletas 73 variáveis de processo durante 4 meses.

Para o estudo da correlação das variáveis de processo com o teor de ferro, o autor utilizou o algoritmo RReliefF. Esse método indica a relevância de cada uma das variáveis, mas considera a sua relevância a partir do contexto dos vizinhos mais próximos (amostras similares). Após a seleção de variáveis, o autor treinou os modelos utilizando algoritmos de Rede Neural Artificial (RNA) e *Random Forest*. A Figura 2 apresenta os resultados de teor de ferro obtidos para cada um dos meses.

Figura 2 - Teor do concentrado de ferro em % nos meses de janeiro, fevereiro, abril e maio obtida em laboratório e obtida pelo *soft sensor* usando RNA

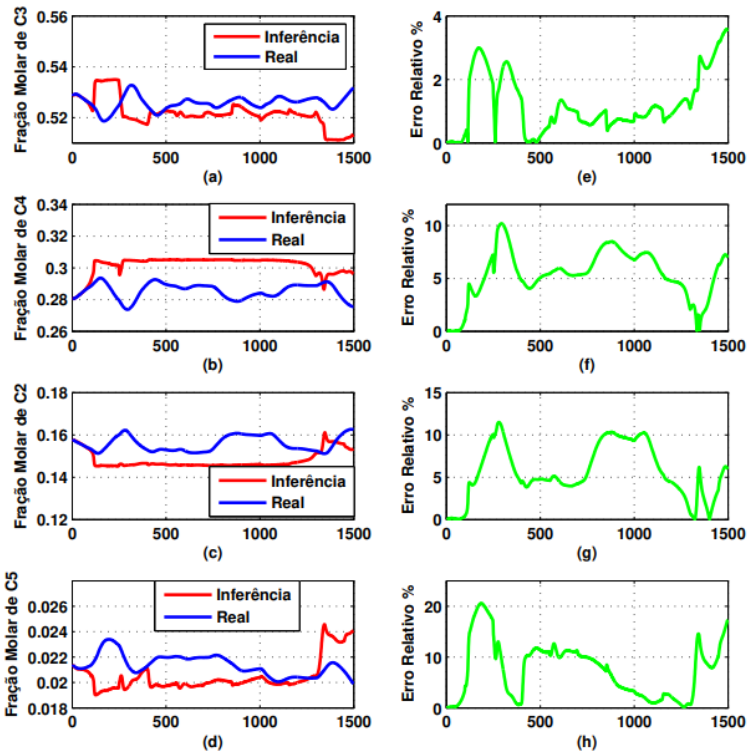


Fonte: GUEDES, 2020.

O autor destaca que apesar dos picos observados na previsão, o erro médio do modelo de predição continuou baixo e foi possível observar uma indicação da tendência de subida ou descida da concentração de ferro nos gráficos. O estudo contribui significativamente para o tema abordado nesse trabalho, exemplificando gráficos comparativos dos dados preditos pelo modelo e dados reais, além de discutir que modelos preditivos incorretamente treinados podem predizer valores médios na maior parte do tempo, limitando o potencial do modelo de predizer tendências de subida ou descida da variável predita. O autor destaca que a indicação de tendência para esse estudo foi crucial para a engenharia considerar a utilização do modelo.

O estudo de LIMA (2018) teve como objetivo a criação de um *soft sensor* aplicado ao processamento de gás natural. O autor destaca que a qualidade do gás natural é dependente da composição química resultante do processo, que é obtida a partir de uma análise cromatográfica em longos intervalos de medição. O trabalho foi desenvolvido utilizando dados desenvolvidos em uma simulação de processos no simulador Aspen Hysys®, com foco na predição da Fração Molar do Propano (C3), Butano (C4), Etano (C2) e Pentano (C5). O modelo de RNA foi escolhido para o treino de cada um dos modelos, e os resultados obtidos para um conjunto de testes pode ser lido na Figura 3.

Figura 3 - Valores reais e de inferência dos componentes do GLP estimados



Fonte: LIMA, 2018.

É possível notar que o erro relativo para alguns grupos foi consideravelmente alto, e, portanto, o autor desenvolve um módulo de correção. Esse módulo ajusta o patamar de predição a partir do último dado real, retroalimentando o módulo neural e corrigindo-o. O estudo contribui com esse trabalho por demonstrar graficamente a visualização do erro relativo entre os dados preditos pelo modelo e os dados reais simulados, além de demonstrar que, durante a modelagem é necessário considerar que produtos diferentes uma vez que podem existir diferentes condições de processo. Desse modo, um conjunto de modelos preditivos pode ser desenvolvido para atender as predições de uma determinada planta industrial.

O estudo de DOMÍNGUEZ-NIÑO *et. al.* (2017) investigou a aplicação de um modelo neural caixa cinza (GB) em um sistema preditivo de operação de um secador rotativo, a fim de criar um esquema de Controle Preditivo de Modelo Não Linear (NMPC). Como modelo caixa cinza, DOMÍNGUEZ-NIÑO *et. al.* (2017) aplicou conceitos fenomenológicos além da RNA, diferente dos estudos de GUEDES (2020) e LIMA (2018). Para a descrição fenomenológica foi desenvolvida uma formulação de balanço de massa e energia das fases gasosa e sólida, além de associar tempo de retenção, cinética de secagem e transferência de calor. As equações fenomenológicas consideradas para este sistema foram:

$$X_{out}(k+1) = 0{,}181(X_{in}(k) - 300Nd) + 0{,}819\, X_{out}\,(k) \tag{1}$$

$$T_{g,out}(k+1) = \left(\frac{93{,}31\, U_a + 0{,}114 T_{g,in}(k) + 12776{,}7Nd}{1.33\, U_a + 0.114}\right)\left[1 - e^{(-60A)}\right] + T_{g,out}(k)e^{(-60A)} \tag{2}$$

E,

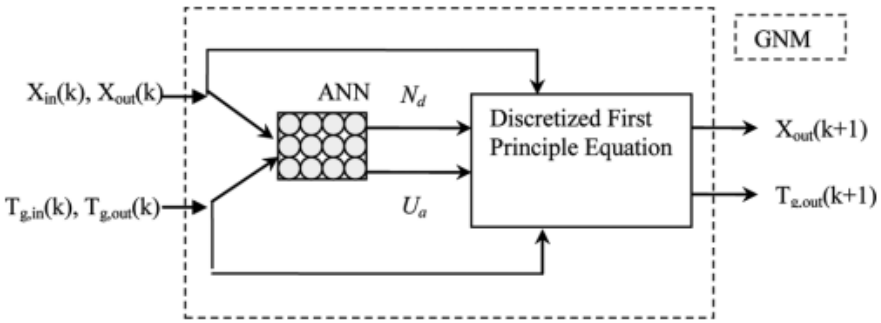
$$A = \ 1{,}33\, U_a + 0{,}114 \tag{3}$$

$$B = 3{,}31\, U_a + 0{,}114 T_{g,in}(k) + 12776{,}7Nd \tag{4}$$

Na qual o ar de secagem tem o subíndice g e o sólido o subíndice s, $T_{i,j}$ a temperatura, Nd é a taxa de secagem, U_a são o coeficiente volumétrico de transferência de calor e X_{out} é a umidade sólida. O subíndice j é utilizado para especificar a entrada ou saída

do secador rotativo. Considerando a possibilidade de utilização de modelos baseados em dados, o estudo conceitua a possibilidade de aplicação de modelos de RNA para prever o coeficiente volumétrico de transferência de calor (U_a) e a taxa de secagem (Nd). A metodologia de treinamento da RNA para obtenção dos parâmetros esperados, utilizando 120 pontos experimentais de entrada-saída, é apresentado na Figura 4. Os resultados foram avaliados sob a perspectiva do controle frente a um distúrbio causado. O autor conclui o estudo definindo que o modelo preditivo gerado tem a capacidade de manipular a temperatura de entrada do gás em diferentes perturbações. Para este trabalho, o estudo de DOMÍNGUEZ-NIÑO *et. al.* (2017) contribui com a exemplificação de modelagem para algoritmo de RNA, e como equações fenomenológicas podem ser incorporadas nos modelos preditivos.

Figura 4 - Esquema da estimativa de temperatura e concentração



Fonte: DOMÍNGUEZ-NIÑO *et. al.,* 2017.

3.3 PROCESSO DE POLIMERIZAÇÃO EM ALTA PRESSÃO

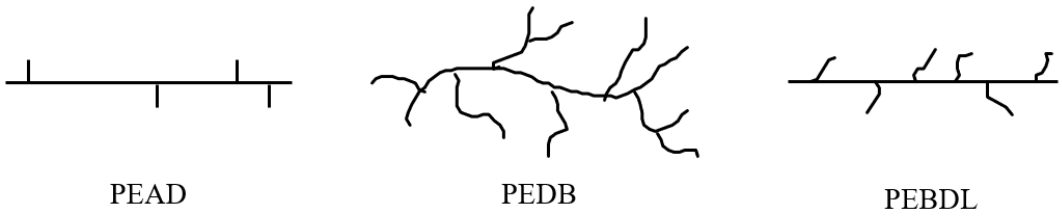
O processo de polimerização escolhido varia conforme o tipo de polímero que se deseja produzir. O polietileno (PE) é uma resina termoplástica aplicada em bens de consumo não-duráveis, que não há requisito de suporte de grandes carga (PEACOCK, 2000). O polietileno pode ser classificado de acordo com os seguintes grupos:

1. Polietileno de Alta Densidade (PEAD): possui densidade entre 0,935 e 0,960 g/cm^3 e não há ramificações na sua cadeia.
2. Polietileno de Baixa Densidade (PEBD): possui densidade entre 0,910 e 0,925 g/cm^3 e muitas ramificações na cadeia principal.

3. Polietileno de Baixa Densidade Linear (PEBDL): possui densidade na faixa de 0,918 a 0,940 g/cm^3, com uma cadeia menos ramificada e mais curta que o PEBD (SILVA, 2012).

A Figura 5 ilustra a diferença entre os três tipos de Polietileno produzidos.

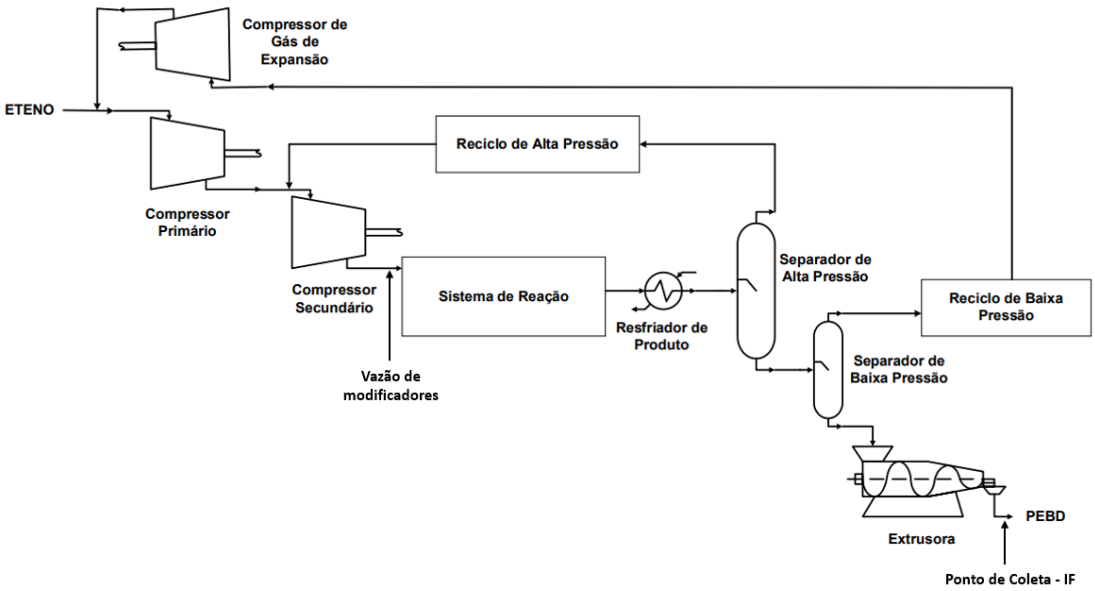
Figura 5 - Estrutura molecular do PEAD, PEBD, PEBDL



Fonte: SILVA, 2012.

Dentro os processos comerciais para a produção de PEBD, destaca-se o processo de alta pressão. O reatores operam a pressões na faixa de 1000 a 3.500 bar, e temperaturas entre 150 e 300°C, de modo que o processo seja não catalítico, mas a partir de radicais livres. O reator autoclave pode ser descrito como um vaso agitado por múltiplas pás para promover uma mistura completa, com múltiplos pontos de alimentação. Para iniciar a reação de polimerização, peróxidos orgânicos podem ser utilizados, em processo reacional próximo ao regime adiabático (SILVA, 2012). A temperatura é a principal variável de acompanhamento da reação, com pontos de medição espalhados pelo reator. A alimentação do processo ocorre a partir de eteno gasoso no compressor, que será comprimido até a pressão de operação do reator, juntamente com os iniciadores de reação (ex. oxigênio). Espera-se que ao entrar no reator a conversão seja de 20 a 30%, e que o gás não convertido seja separado por um Sistema de Alta Pressão (SAP), pelo Sistema de Baixa Pressão (SBP) e retorne para o compressor (SILVA, 2012). As correntes de fundo dos vasos de alta e baixa são enviadas para extrusora, onde o polímero formado é resfriado e transformado em *pellets* para a comercialização. A Figura 6 resume o processo de alta pressão para a produção de PE.

Figura 6 - Polimerização via alta pressão para a produção de PE



Fonte: SILVA, 2012.

O IF é uma propriedade amplamente utilizada na indústria de polímeros para caracterizar e identificar os diferentes grades do polímero produzido. Para a realização das análises de laboratório para a obtenção do IF, pontos de coleta são instalados após a extrusora, conforme a Figura 6. O IF é inversamente proporcional à viscosidade do material fundido a um certa temperatura e taxa de cisalhamento, que será expressa junto a unidade de medida da análise (ROCHA; COUTINHO; BALKE, 1994). O índice de fluidez é dependente de propriedades como a quantidade de ramificações e o peso molecular, e representa a facilidade de fluxo do polímero fundido.

As análises de IF são desenvolvidas em laboratório, sendo necessário a coleta da amostra na linha de produção e a sua correta identificação. A análise é realizada a partir da extrusão do polímero em um reômetro capilar aquecido, operado a uma pressão imposta (ROCHA, COUTINHO, BALKE, 1994). Desse modo, o índice de fluidez é obtido a partir da taxa de fluxo do fluido sob a carga imposta em um tempo total de 10 min, e, portanto, a unidade de medida é g/10 min. Essa técnica é especificada na ASTM D1238 ("*Standard Test Method for Melt Flow Rates of Thermoplastics by Extrusion Plastometer*") e ISO1133 ("*Plastics — Determination of the Melt Mass-Flow Rate* (MFR) *and Melt Volume-Flow Rate* (MVR) *of Thermoplastics*"). Para a operação da planta é importante que os dados sejam obtidos, validados e disponibilizados o mais rápido possível, de modo que ajustes no processo sejam realizados para manter o produto dentro

da especificação esperada. Considerando que a análise é custosa e requer tempo, muitas indústrias executam a análise a cada duas ou quatro horas, dificultando o controle adequado da qualidade do polímero (SONG, *et al.,* 2022*)*.

3.4 *MACHINE LEARNING*

O estudo de PAINULI *et. al.* (2021) apresenta *machine learning* como o campo de estudo que fornece aos computadores a capacidade de aprender sem serem explicitamente programados, possivelmente sendo auto programáveis. A aprendizagem pode ocorrer por meio de observações, exemplos, dados históricos, semelhança paterna e outros (FANG *et. al.,* 2022). O treinamento de modelos de *machine learning* seguem uma sequência de etapas sistemáticas: seleção e limpeza dos dados, estudo da correlação entre as variáveis e seleção do tipo de algoritmo de treino com base no serviço.

A seleção dos dados corretos é um dos aspectos mais cruciais no processo de treinamento de modelos de *machine learning*, além da própria seleção do tipo de algoritmo. Dados de qualidade formam a base para a criação de modelos eficazes, precisos e úteis. Sem dados adequados, as etapas seguintes do estudo podem falhar ou fornecer modelos insatisfatórios. A inclusão de variáveis irrelevantes ou desbalanceadas pode comprometer a capacidade do modelo de capturar os padrões associados ao fenômeno em estudo. Além disso, dados coletados em contextos industriais frequentemente apresentam ruídos e inconsistências provenientes de medições imprecisas, variações entre *batchs* ou discrepâncias nas condições de processamento. É importante destacar que alguns sensores no campo foram instalados no passado e não possuem manutenção frequente, e desse modo, os resultados podem não apresentar alta confiabilidade. Portanto, apesar da riqueza de dados disponível na indústria de polímeros, o sucesso no desenvolvimento de modelos preditivos confiáveis e precisos exige não apenas a quantidade, mas também a qualidade desses dados. Além da seleção criteriosa, o pré-processamento rigoroso é uma etapa indispensável para explorar o potencial dos dados no avanço de *soft sensors* no setor, tratando dados de baixa qualidade e removendo *outliers*.

A escolha adequada das variáveis que serão utilizadas para treinar o modelo pode ser efetuada a partir do estudo da correlação entre as variáveis. A utilização de muitas variáveis, nem todas necessariamente relevantes, pode aumentar a dimensionalidade do modelo que por sua vez aumenta o custo computacional (GUEDES, 2020). Além disso,

a seleção de variáveis evita a multicolinearidade, que ocorre quando duas ou mais variáveis independentes estão altamente correlacionadas (GUEDES, 2020). No âmbito da ciência de dados, a técnica de matriz de correlação é amplamente conhecida para a seleção de variáveis, e utiliza bibliotecas como Pandas, NumPy e Seaborn (LIDIANE *et al.*, 2018) A matriz de correlação é uma tabela que mostra os coeficientes de correlação entre pares de variáveis em um conjunto de dados. Esses coeficientes variam de -1 a 1, onde:

- 1: Correlação positiva perfeita.
- 0: Nenhuma correlação.
- -1: Correlação negativa perfeita.

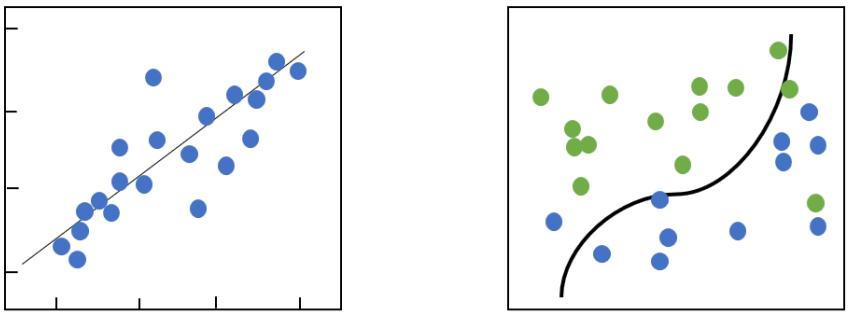
Ao longo dos anos, diversos algoritmos foram desenvolvidos para realizar o aprendizado de acordo com as necessidades do serviço. Porém, não existe um algoritmo único que funcione para todos os problemas e, da mesma forma, não é possível definir que um tipo de algoritmo será sempre melhor que outro (MAHESH, 2019), uma vez que os modelos são resultados de fatores como quantidade de dados, dispersão, desvio padrão e necessidades específicas de representação (FLASK; DOCKER; KUBERNETES, 2021).

Dentre os algoritmos de processamento, dois grupos podem ser destacados:

- Aprendizado supervisionado;
- Aprendizado não supervisionado.

Existem também dois tipos de modelos preditivos: classificação e regressão. Se a natureza da previsão for numérica, podem ser utilizados modelos de regressão. Caso a necessidade seja categorizar/agrupar os resultados, selecionam-se modelos de classificação (MAHESH, 2019). A Figura 7 foi desenvolvida para evidenciar a diferença entre os modelos.

Figura 7 - Regressão *versus* Classificação



Fonte: Elaborado pelo autor.

3.4.1 APRENDIZADO NÃO-SUPERVISIONADO

O aprendizado não supervisionado constrói seus modelos com base na semelhança entre os dados (MAHESH, 2019). O objetivo é descobrir padrões ocultos pela similaridade do conjunto de dados e mais de uma resposta pode ser considerada adequada (LIKAS; VLASSIS; VERBEEK, 2003). Assim, se os parâmetros de treinamento forem diferentes entre os usuários, resultados totalmente diferentes podem ser obtidos e ainda assim considerados aceitáveis (LIKAS; VLASSIS; VERBEEK, 2003).

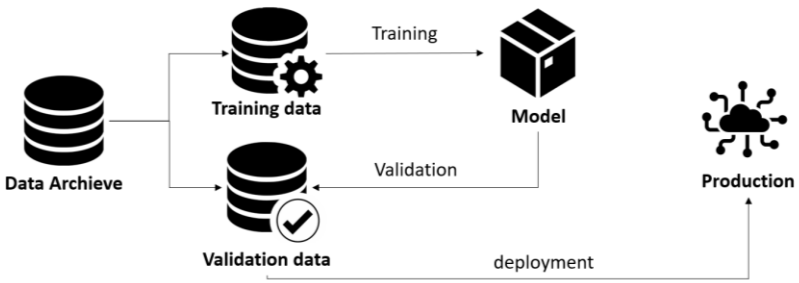
O algoritmo *k-means* é um modelo não supervisionado disponível na literatura. O algoritmo trabalha para agrupar os dados, fornecendo uma estrutura por similaridade de pontos para cada grupo ou subgrupo. O algoritmo separa os dados em um número de grupos especificado pelo usuário. O desempenho do algoritmo é baseado na atribuição de pontos a um cluster de forma que a soma da distância quadrada entre os dados e o centroide (média aritmética de todos os pontos de dados no *cluster*) seja a mínima possível e atuando para fornecer a maior distância entre diferentes clusters (MAHESH, 2019; ALAM, 2022).

3.4.2 APRENDIZADO SUPERVISIONADO

O aprendizado supervisionado é definido por algoritmos que buscam classificar ou prever resultados com certa precisão a partir de um conjunto de dados de entrada e

uma variável conhecida (FANG *et. al.,* 2022). Normalmente, os dados são separados em dados de treino e de teste. O modelo é treinado para testar sua capacidade preditiva em um novo conjunto de dados de resposta conhecidos. O treino ocorre comparando o valor previsto e o valor esperado atual, a fim de minimizar o erro obtido (MAHESH, 2019). A Figura 8 exemplifica o fluxo de trabalho desse tipo de algoritmo.

Figura 8 - *Workflow* de algoritmos de aprendizado supervisionado

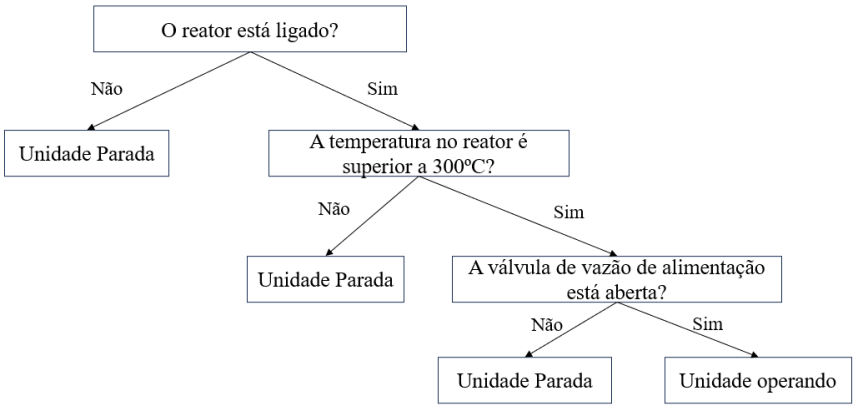


Fonte: Elaborado pelo autor.

O fluxo de trabalho apresenta a implantação (*deploy*) do modelo de *Machine Learning*. *Deploy* pode ser definido como o processo de finalização do modelo, exportando o modelo para o sistema de arquitetura industrial designado (FLASK; DOCKER; KUBERNETES, 2021).

Dentro dos principais algoritmos de aprendizado supervisionado, podem ser destacados modelos de árvores de decisão e de regressão. Árvores de decisão são algoritmos cuja estrutura é semelhante a um fluxograma, onde cada um dos nós internos representa um teste (KAJAL *et al.,* 2016). Um exemplo seria tomar uma decisão se o valor é maior ou menor que um determinado número. Assim, cada nó do fluxograma é responsável por apresentar os atributos do grupo, e as ramificações, o valor que pode ser assumido (KAJAL *et al.,* 2016). Por exemplo, cada nó pode agrupar um sim ou um não. A construção de árvores de decisão geralmente é feita com uma abordagem *top-down*, tomando uma decisão em cada etapa, a fim de reduzir a impureza das divisões que estão sendo realizadas. A Figura 9 exemplifica o conceito de árvore de decisão.

Figura 9 – Exemplo de árvore de decisão



Fonte: Elaborado pelo autor.

Existem dois tipos principais de árvores de decisão em aprendizado de máquina, que são árvores de classificação e árvores de regressão (YETURU, 2020). Em geral, as árvores de classificação são aquelas utilizadas para categorização, e as árvores de regressão, para obtenção de um resultado numérico. Dentre os principais benefícios da utilização de uma árvore de decisão, está a facilidade de entendimento mesmo para *stakeholders* sem conhecimento técnico, permitindo a utilização de dados categóricos ou numéricos (MITTAL; KHANDUJA; TEWARI, 2017). Segundo Maulud e Abdulazeez (2020), a árvore de regressão é um algoritmo baseado em duas técnicas: para prever e para determinar a relação entre variáveis independentes e dependentes. Os modelos de regressão em *machine learning* são responsáveis por estimar uma variável dependente "y" a partir de uma variável independente "x" dentro de um período de treino. A expressão genérica para regressão linear é dada por:

$$y = a_0 + a_1 x + \mathcal{E} \qquad (5)$$

Onde y é a variável dependente, a_0 uma constante que intercepta o eixo vertical, a_1 é o coeficiente de regressão e $\mathcal{E}$ é o erro aleatório (MAULUD; ABDULAZEEZ, 2020). Além da regressão linear, podem ser citados algoritmos de regressão polinomial e de regressão vetorial de suporte (SVR).

Para esse trabalho, foram considerados os algoritmos de *Gradient Boosting*, o *LightGBM* e o *Long Short-Term Memory* (LSTM).

3.4.2.1 *GRADIENT BOOSTING REGRESSOR*

O *Gradient Boosting Regressor* (GBR) é um algoritmo de *machine learning* supervisionado disponível na biblioteca do *scikit-learn*. A abordagem do modelo baseia-se na combinação sequencial de vários modelos simples, geralmente árvores de decisão rasas, para formar um modelo robusto e preciso. O processo do GBR começa com a construção de um modelo inicial simples, que geralmente prevê uma constante, como a média dos valores da variável objetivo no conjunto de treinamento. Em seguida, o erro residual pode ser definido pelo hiperparâmetro escolhido pelo usuário, sendo que usualmente aplica-se o erro quadrático (NATEKIN; KNOLL, 2013). O cálculo do erro residual pode ser observado na equação 6, sendo que y_i é a variável objetivo e $\widehat{y_i}$, a predição do modelo.

$$r_i = (y_i - \widehat{y_i})^2 \tag{6}$$

Este erro é utilizado como base para o treinamento de um novo modelo, que se ajusta aos resíduos deixados pelo modelo anterior. Desse modo, o modelo seguinte é treinado para prever o resíduo do modelo anterior, ou seja, para relacionar as variáveis de entrada do modelo com os erros constantes.

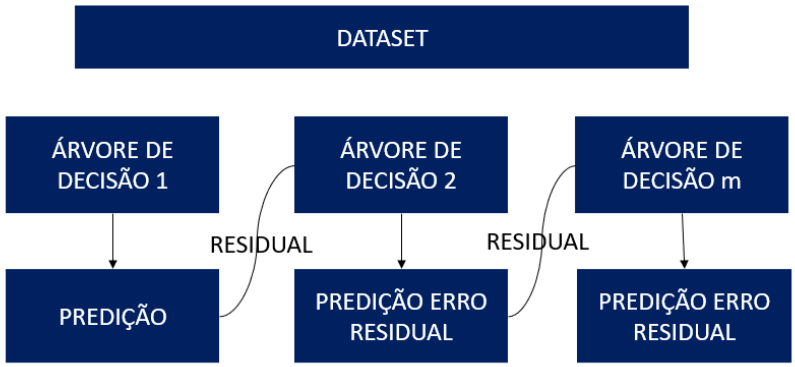
O novo modelo é então adicionado ao modelo global. A combinação dos modelos é feita atribuindo um peso (ν, chamado de taxa de aprendizado) ao novo modelo, que controla o impacto de cada etapa no resultado (FRIEDMAN; HASTIE; TIBSHIRANI, 2000) . A fórmula geral para a previsão final após n iterações é:

$$\widehat{y_i}^{(m)} = \widehat{y_i}^{(m-1)} + \nu * f_m(x_i) \tag{7}$$

Na qual, $\widehat{y_i}^{(m)}$ é a previsão atualizada após m iterações, $f_m(x_i)$ é a saída do m-ésimo modelo ajustado aos resíduos, ν é a taxa de aprendizado - valores menores tornam o ajuste mais gradual e podem melhorar a generalização (FRIEDMAN; HASTIE; TIBSHIRANI, 2000).

Essa abordagem iterativa continua até que o número máximo de iterações seja alcançado ou até que não haja melhorias significativas na redução do erro. A Figura 10 ilustra o modelo.

Figura 10 – Diagrama representativo do algoritmo *Gradient Boosting Regressor*



Fonte: Elaborado pelo autor.

Sendo que o *Boosted Model* pode ser finalmente descrito pela Equação 8.

$$F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma(x_j) \tag{8}$$

Sendo F_m a predição do modelo, γ a predição do erro residual da árvore de predição x_j e ν a taxa de aprendizado. A taxa de aprendizado, por exemplo, é um dos hiperparâmetros que pode ser ajustado para aumentar o desempenho dos modelos durante o treino. São considerados hiperparâmetros os parâmetros específicos de cada algoritmo como a taxa de aprendizado, o número de árvores em uma floresta aleatória, a profundidade máxima das árvores ou o número de neurônios em uma rede neural. A Tabela 1 resume os hiperparâmetros do modelo de *Gradiente Boosting Regressor*. É importante destacar que existe interdependência entre os hiperparâmetros, por exemplo o *learning_rate* e o *n_estimators* devem ser ajustados em conjunto uma vez que valores menores de *learning_rate* geralmente requerem mais árvores.

Tabela 1 – Hiperparâmetros do algoritmo *Gradient Boosting Regressor*

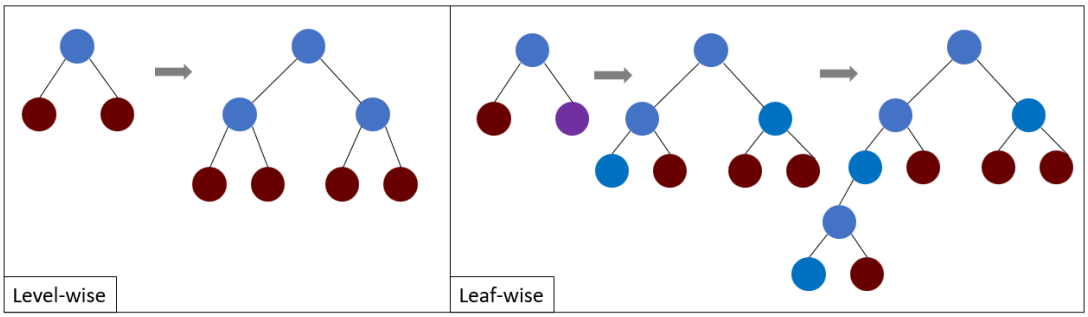
Hiperparâmetro	Descrição	Valores Comuns
n_estimators	Número de árvores ou modelos fracos a serem treinados.	100 a 500
learning_rate	Taxa de aprendizado, controla o impacto de cada árvore no modelo final.	0.01 a 0.2
max_depth	Profundidade máxima das árvores, controla a complexidade das árvores individuais.	3 a 10
min_samples_split	Número mínimo de amostras necessárias para dividir um nó.	2 a 20
min_samples_leaf	Número mínimo de amostras necessárias em uma folha.	1 a 10
loss	Função de perda a ser minimizada.	"squared_error", "huber"
random_state	Semente para gerar resultados reproduzíveis.	Inteiro (ex.: 42)
max_leaf_nodes	Número máximo de folhas em cada árvore.	Inteiro ou None

Fonte: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html, acessado em 09/12/2024.

3.4.2.2 *LIGHTGBM REGRESSOR*

O modelo de *Light*GBM foi desenvolvido para lidar com um conjunto grande de dados e alta dimensionalidade, desenvolvida pela Microsoft Research. O modelo é baseado na técnica de *Gradient Boosting*, mantendo o conceito de previsão da variável alvo a partir de uma combinação de múltiplas arvores de decisão, ajustadas iterativamente para minimizar erros (ALABDULLAH *et al.*, 2022). Entretanto, o método utilizada do crescimento na forma de *leaf-wise*, ou seja, a árvore é expandida apenas nas folhas que trazem o maior ganho de informação. Árvores de decisão normalmente crescem uniformemente em cada nível, mas o método de *leaf-wise* permite que o sistema computacional mantenha em foco apenas nos nós relevantes (ALABDULLAH *et al.*, 2022).

Figura 11 – Diferença nas construções da árvore de decisão



Fonte: ALABDULLAH *et al.*, 2022

A Tabela 2 resume os hiperparâmetros do modelo de *Light*GBM.

Tabela 2 - Hiperparâmetros do algoritmo *Light*GBM

Hiperparâmetro	Descrição	Valores Comuns
n_estimators	Número de árvores ou modelos fracos a serem treinados.	100 a 500
learning_rate	Taxa de aprendizado, controla o impacto de cada árvore no modelo final.	0.01 a 0.2
max_depth	Profundidade máxima das árvores, controla a complexidade das árvores individuais.	3 a 10
num_leaves	Número máximo de folhas por árvore	20 a 50
feature_fraction	Proporção de características (features) usadas para treinar cada árvore.	0.8 a 1.0
bagging_fraction	Proporção dos dados amostrados para cada iteração	0.8 a a1.0

Fonte: https://lightgbm.readthedocs.io/en/latest/Parameters.html, acessado em 09/12/2024.

3.4.2.3 *LSTM REGRESSOR*

O *LSTM Regressor* é um modelo de *machine learning* baseado em *Recurrent Neural Network* (RNN) que utiliza *Long Short-Term Memory* para realizar predições. Os modelos de RNN foram desenvolvidos a partir de RNA, entretanto, são recomendados para lidar com dados sequenciais, utilizando-se de memória para aprendizado. O modelo de LSTM é uma variante das RNNs pois permite que a rede aprenda dependências de

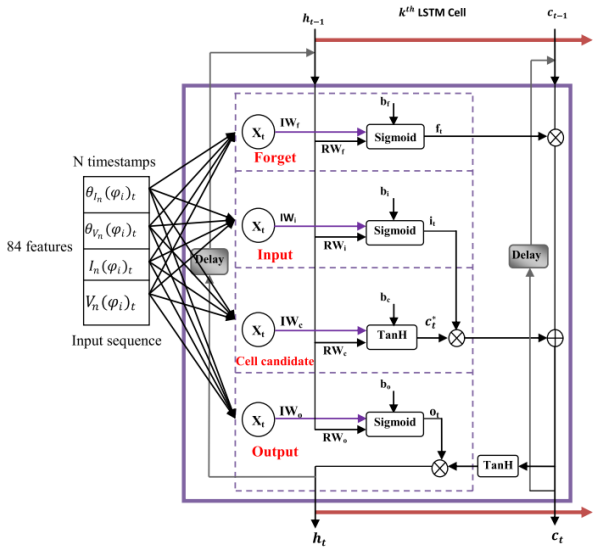
dados de longo prazo. Um modelo convencional de RNN decorre do processamento unidirecional, ou seja, da entrada para a saída em cada neurônio, sem considerar a memória do que ocorreu no passo anterior (AL-SELWI *et al.*, 2023). Por outro lado, LSTM considera estados de memória, para carregar informações entre os passos temporais anteriores denominadas portas (BELAGOUNE *et al.*, 2021). Para esse estudo, LSTM foi considerada pois os conjuntos de dados acompanham as tendências de processo e a dependência temporal das variáveis deve ser levada em consideração.

As portas determinam como cada informação será manipulada no neurônio e ajudam a controlar o fluxo de informações, sendo:

- Porta do esquecimento: decide quais informações serão descartadas;
- Porta de entrada: quais novas informações serão adicionas a memória de cálculo;
- Porta de atualização: combina informações antigas e novas para atualizar a memória;
- Porta de saída: decisão de quais informações de memória devem ser usadas para o cálculo da saída atual (BELAGOUNE *et al.*, 2021).

A Figura 12 exemplifica a arquitetura do modelo de LSTM, demonstrando cada uma das portas para uma célula k da rede, na qual cada sigmoide é uma função de saída específica do tipo de porta, e utiliza parâmetros como: RW (*Recurrent weight*), IW (*Input weight*) e b (*bias*). O resultado $C(t)$ representa a memória da célula e $h(t)$ é a predição passada de uma célula LSTM a outra (BELAGOUNE *et al.*, 2021).

Figura 12 – Sequência de conexões entre $\boldsymbol{k}$ LSTM células em um algoritmo LSTM



Fonte: BELAGOUNE *et al.*, 2021.

A Tabela 3 resume os hiperparâmetros do modelo de LSTM.

Tabela 3 - Hiperparâmetros do algoritmo LSTM

Hiperparâmetro	Descrição	Valores Comuns
Units	Número de neurônios/células LSTM na camada.	50 a 200
Learning Rate	Controle da magnitude dos ajustes nos pesos.	0.001 a 0.1
Batch Size	Quantidade de amostras em cada atualização de pesos.	16 a 128
Optimizer	Método para atualizar pesos através de gradientes.	Adam, SGD, RMSprop.
Number of Layers	Quantas camadas LSTM são empilhadas no modelo.	1 a 3

Fonte: https://keras.io/api/layers/recurrent_layers/lstm/, acessado em 09/12/2024.

3.5 MÉTRICAS DE AVALIAÇÃO DOS MODELOS

Durante o teste de adequação do melhor modelo *de machine learning*, diferentes métricas podem ser utilizadas. O uso de métricas como o RMSE (*Root Mean Squared Error*) e a acurácia é amplamente conhecido na ciência de dados, sendo nativos de diferentes bibliotecas (BORGES, 2020).

O RMSE é amplamente utilizado em problemas de regressão para medir a diferença média entre os valores previstos pelo modelo e os valores reais do conjunto de dados. O RMSE é calculado pela Equação 9:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (9)$$

Sendo que, RMSE mede a média da magnitude dos erros entre as previsões do modelo ($\hat{y}_i$) e os valores reais (y_i), sendo particularmente sensível a grandes desvios

devido à elevação dos erros ao quadrado (BORGES, 2020). Como o RMSE eleva os erros ao quadrado antes de calcular a média, erros maiores têm um impacto desproporcional, tornando-o mais sensível a *outliers* em comparação a outras métricas como o *Mean Absolute Error* (MAE). Um RMSE mais baixo indica que o modelo faz previsões mais próximas dos valores reais, e o modelo em questão se adequa melhor ao conjunto de dados utilizado.

A acurácia é uma métrica frequentemente empregada em problemas de classificação e regressão, pois representa a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões, conforme pode ser visto na Equação 10 (ARAGAO, 2022) .

$$Acurácia = \frac{Número\ de\ previsões\ corretas}{Total\ de\ previsões} \tag{10}$$

Considerando que a variável predita é proveniente de análises de laboratório, torna-se razoável considerar que mesmo os valores resultados das análises podem possuir um erro de medição. A metodologia para análise do índice de fluidez é descrita na norma ASTM D1238 ("*Standard Test Method for Melt Flow Rates of Thermoplastics by Extrusion Plastometer*") e ISO1133 ("*Plastics — Determination of the Melt Mass-Flow Rate* (MFR) *and Melt Volume-Flow Rate* (MVR) *of Thermoplastics*"), que fornecem orientações detalhadas para a realização do teste de índice de fluidez, incluindo tolerâncias aceitáveis para erros e variações. O erro de medição do Índice de Fluidez em polímeros pode variar dependendo de fatores como o equipamento utilizado, a calibração, a metodologia do operador e as condições experimentais. Considerando a reprodutibilidade dos testes em laboratórios diferentes, a variação aceitável é de até 5%, e para a repetibilidade em um mesmo laboratório os valores não devem exceder 3%. Para o cálculo de acurácia neste estudo, foi considerado um erro aceitável de até 5% do valor real.

4. MATERIAIS E MÉTODOS

4.1 CONFIGURAÇÃO DO AMBIENTE E RECEBIMENTO DOS DADOS

O conjunto de dados deve ser processado unicamente em ferramentas de leitura capazes de comportar o entendimento de milhares de linhas simultaneamente. Para isso, foi utilizado o Jupyter Notebook, instalado em conjunto ao Anaconda. O Anaconda Navigator é uma interface gráfica, que permite o acesso e gerenciamento dos ambiente sem utilizar o terminal. O Jupyter Notebook é um aplicativo baseado na web e possui suporte em linguagens como Python e R.

4.2 SELEÇÃO DOS DADOS

A indústria de polímeros dispõe de uma vasta base de dados acumulada ao longo de décadas, abrangendo propriedades físico-químicas, parâmetros de processamento, condições ambientais e desempenho dos materiais em aplicações específicas. Contudo, foi realizada uma seleção criteriosa dos dados para a construção de modelos preditivos.

Após o estudo do processo por meio de documentações de engenharia e apresentações operacionais, foram selecionados dados da seguintes áreas de processamento: reação, trocadores de calor atrelados ao reator, extrusão e injeção de modificadores de processo. Os dados de sensor foram extraídos de minuto a minuto entre o período de 24 de novembro de 2020 a 09 de janeiro de 2022, na qual foram recebidos em formato ".CSV". A Tabela 4 apresenta quantidade de dados selecionadas por áreas conforme apresentado na Figura 13.

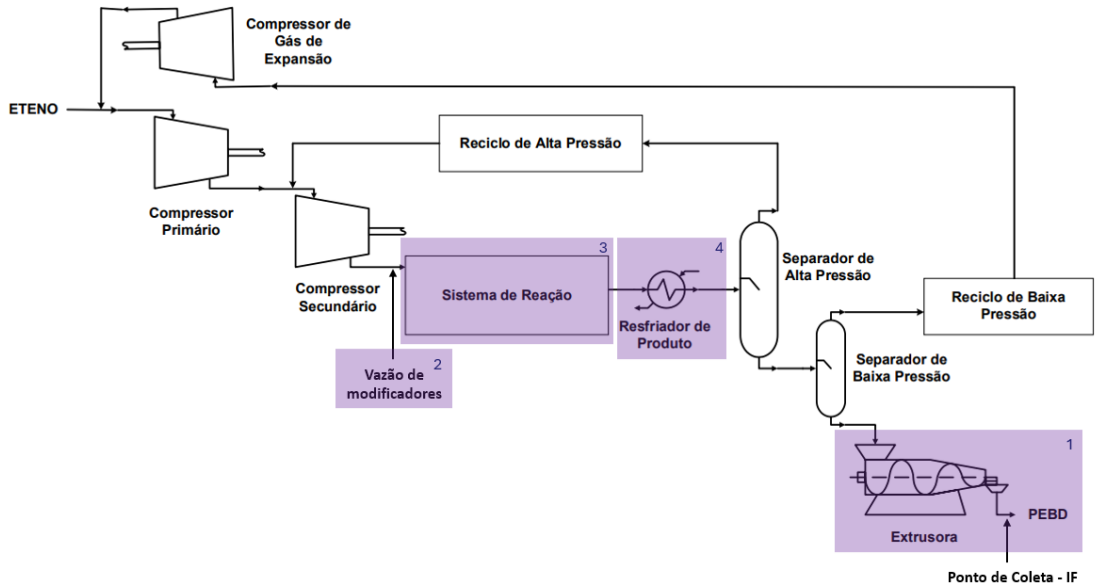
Tabela 4 - Dados de sensor selecionados por área

ID	Área	Quantidade de variáveis	Descrição
1	Extrusão	9	Variáveis relacionadas a corrente, pressão do filtro, temperatura dos passes e vazões de aditivos

2	Modificadores	4	Vazão e concentração de modificadores (agente de transferência de cadeia e inibidores)
3	Reator	25	Pressão e Temperatura
4	Trocador de Calor do Reator	2	Temperaturas
	Total	40	

Fonte: Elaborado pelo autor.

Figura 13 – Seleção de variáveis de processo com base na área funcional (ID)



Fonte: Elaborado pelo autor.

De modo a garantir a confidencialidade das informações, os dados de sensor foram normalizando utilizando o método *Min-Max Scalling*. Essa técnica é amplamente utilizada em ciência de dados e aprendizado de máquina para transformar os valores de uma variável para um intervalo específico entre 0 e 1. O principal objetivo do *Min-Max Scaling* é reescalar os valores de forma que todos fiquem dentro de um intervalo padronizado, preservando as relações proporcionais entre os dados. A fórmula matemática para a transformação é apresentada na Equação 11.

$$x_{scaled} = \frac{(x - x_{min})}{x_{max} - x_{min}} \tag{11}$$

Nessa equação, x_{scaled} representa o valor original da variável, x_{min} é o menor valor observado na variável, e x_{max} é o maior valor observado. O resultado, x_{scaled} é o valor normalizado, que estará no intervalo [0,1].

Foram também coletadas análises de laboratório para todos os produtos da unidade de polimerização entre o período de 24 de novembro de 2020 a 09 de janeiro de 2022. Diferente dos dados de sensor, os dados de laboratório possuem um intervalo de aproximadamente 2-3h. Isso ocorre porque, diferente dos dados dos sensores que são coletados automaticamente pelos instrumentos em campo, as análises de laboratório são obtidas a partir de análises laboratoriais realizadas a partir de um conjunto amostral do produto. Atualmente, a unidade de polimerização produz 10 tipos de material, sendo que 4 configuram produção de PEBD (Material 1, 4, 5 e 10). A quantidade de dados recebidos por tipo de material pode ser observada na Tabela 5.

Tabela 5 – Quantidade de dados de laboratório coletados

Material	**Quantidade de Amostras**
1	1296
2	850
3	897
4	681
5	584
6	351
7	288
8	273
9	235
10	203

Fonte: Elaborado pelo autor.

De modo a garantir a confidencialidade das informações, os dados de laboratório também foram normalizando utilizando o método *Min-Max Scalling.*

4.3 PROCESSAMENTO E TRATAMENTO DOS DADOS

O processamento dos dados consistiu em fazer a leitura e conversão dos arquivos contendo as informações dos sensores ou de laboratório para um *dataset*. O *dataset* (ou conjunto de dados) é uma coleção de dados organizados de forma a facilitar a análise, a visualização ou a aplicação de algoritmos, especialmente em contextos como estatística, ciência de dados e *machine learning*. A leitura dos dados resultou em um *dataset* estruturado, ou seja, um conjunto de dados organizados em um formato tabular (como uma planilha), onde há uma clara estrutura com colunas e linhas. A primeira coluna denominada "*index"* permanece com o *timestamp* da coleta dos dados, sendo que cada linha representa um observação no tempo e cada coluna, representa uma variável.

Durante a leitura dos dados, os valores faltantes foram transformados em indicadores de texto ou código de erro (ex. *bad value*) compatível com as técnicas utilizadas. A linguagem de programação *python* é capaz de transformar valores faltantes no símbolo NaN (*Not a Number*). Desse modo, durante a montagem do *dataset* têm-se um conjunto de dados no tempo, contendo os dados de cada variável ou o código NaN, conforme pode ser observado na Figura 14.

Figura 14 – Exemplificação *dataset* obtido

Timestamp	Variável 1	Variável 2		Variável X
01/01/2023 12:00	23,02	NaN		120,00
01/01/2023 12:01	27,00	2,09	...	NaN
01/01/2023 12:02	27,05	2,01		122,07
01/01/2023 12:03	NaN	2,00		122,02

Fonte: Elaborado pelo autor.

De posse dos dados processados, a etapa seguinte resumiu-se ao tratamento e limpeza dos dados. O tratamento adequado dos dados é um dos aspectos cruciais para o sucesso no treinamento de modelos ML, uma vez que dados brutos frequentemente contêm inconsistências, informações irrelevantes ou até mesmo códigos de erro. A detecção e remoção de *outliers* é considerada uma etapa crítica no tratamento dos dados, pois evita o treinamento do modelo a partir de valores atípicos. *Outliers* são dados que se desviam substancialmente do comportamento geral das observações no conjunto de dados.

Uma das principais razões para remover *outliers* é que eles podem afetar a construção de modelos, principalmente aqueles baseados em métodos sensíveis a variações extremas, como regressão linear, árvores de decisão e até redes neurais. Por exemplo, em modelos de regressão, *outliers* podem influenciar a estimativa dos coeficientes, tornando o modelo mais propenso ao *overfitting* e perdendo sua capacidade de generalização para novos dados. Isso pode resultar em uma performance ruim, com previsões imprecisas.

A remoção de *outliers* foi realizada a partir do desvio padrão, sendo uma técnica estatística simples e eficaz para identificar e tratar valores que estão fora de um intervalo considerado "normal" em um conjunto de dados. Essa abordagem baseou-se na suposição de que os dados seguiam uma distribuição aproximadamente normal (ou gaussiana), onde a maioria dos valores está concentrada em torno da média e a probabilidade de ocorrência diminui à medida que se afastam dela (ARANTES, 2017). O processo iniciou-se com o cálculo da média (μ) e do desvio padrão (σ) do conjunto de dados. A média representa o ponto central dos valores, enquanto o desvio padrão mede a dispersão dos dados em torno dessa média.

4.4 ANÁLISE DE CORRELAÇÃO ENTRE AS VARIÁVEIS

Para analisar a correlação entre as variáveis, a função nativa *pandas.DataFrame.corr()* da biblioteca pandas foi executada no Jupyter Notebook, aplicando as metodologias de correlação de Pearson e Spearman. A relação de Pearson mede a força e a direção da relação linear entre duas variáveis numéricas, baseando-se na covariância e normalizada pelo produto dos desvios padrão. A equação abaixo apresenta a fórmula de cálculo utilizada.

$$r = \frac{\sum(X_i - \hat{X})(Y_i - \hat{Y})}{\sqrt{\sum(X_i - \hat{X})^2 \sum(Y_i - \hat{Y})^2}} \tag{12}$$

A correlação de Spearman por outro lado mede a força e a direção da relação monotônica (quando uma variável aumenta ou diminui consistentemente com outra), e não necessariamente linear. O valor de correlação é a calculada com base na ordem

crescente dos dados (*ranking*) ao invés dos valores originais, permitindo que posições relativas importem mais do que valores extremos, tornando o modelo menos sensível a *outliers*. A equação abaixo apresenta a correlação de Spearman utilizada, na qual d_i é a diferença entre a ordem das duas variáveis.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{13}$$

Para esse estudo, a matriz de correlação foi desenvolvida na forma de mapa de calor pela função *sns.heatmap*() para cada um dos grupos. O mapa de calor permite observar não apenas a correlação das variáveis de processo em relação a variável objetivo, mas também entre as variáveis consideradas. Cada célula do *heatmap* representa o coeficiente de correlação entre duas variáveis, podendo variar de -1 a 1. No caso de +1, entende-se que existe uma correlação positiva perfeita com sua intensidade representada por cores quentes (ex. vermelho), enquanto -1 representa uma correlação negativa perfeita e normalmente apresentada por cores frias (ex. azul). A coluna diagonal apresenta a correlação entre uma variável e ela mesma, enquanto o triângulo superior é uma reflexão dos dados do triângulo inferior.

4.5 TREINAMENTO E AVALIAÇÃO DOS MODELOS DE *MACHINE LEARNING*

Após a seleção prévia de um conjunto de variáveis que se correlaciona com a variável objetivo, deve-se executar o treinamento do modelo. Considerando que os dados de entrada devem predizer o IF, modelos de aprendizado supervisionado foram utilizados. A primeira etapa no treinamento de modelos é a escolha do algoritmo apropriado, uma vez que a literatura dispões de diversos grupos para esse tipo de aprendizado: Regressão Linear, Árvores de Decisão, *Boosting*, ou Redes Neurais. Os algoritmos de *machine learning* podem ser acessados de diversas maneiras, geralmente por meio de bibliotecas e *frameworks* que fornecem implementações eficientes e de fácil utilização. Dentre as bibliotecas utilizadas estão a *scikit-learn* (*sklearn*) que aplica a linguagem de

programação *python*, *TensorFlow* e *Keras* queforam utilizadas para treinamento usando redes neurais.

A partir da seleção do modelo, os dados foram divididos em dois subconjuntos: treino e teste. O conjunto de treino foi utilizado para ajustar os parâmetros internos do modelo (como os coeficientes na regressão ou os pesos em uma rede neural), hiperparâmetros do modelo (como a taxa de aprendizado ou a profundidade das árvores em modelos de árvore de decisão) e seleção de variáveis de processo. Já o conjunto de teste é utilizado para avaliar a performance do modelo em dados não vistos, verificando sua capacidade de generalização. A divisão dos dados foi 80% para treinamento, e 20% para teste.

Um modelo baseline foi obtido considerando os algoritmos selecionados para esse trabalho: *Gradient Boosting*, o *LightGBM* e o LSTM. O modelo baseline foi treinado para cada um dos grupos, de modo a entender quais dos algoritmos teriam potencial de aderência aos dados coletados, permitindo avaliar o desempenho do modelo, sem otimização de hiperparâmetros.

Em sequência, foi realizada a seleção de variáveis. Para essa etapa, o modelo foi submetido a um processo iterativo no qual diferentes combinações de variáveis foram testadas, priorizando as variáveis com maior correlação para cada grupo, conforme obtido no item 4.4. Durante essa etapa, variáveis foram adicionadas ou removidas com o objetivo de identificar o conjunto que proporcionasse o melhor desempenho, conforme as métricas de avaliação especificadas no item 4.6. O refinamento contínuo priorizou a redução do número total de variáveis, preservando o melhor valor da métrica de avaliação.

Para a seleção dos hiperparâmetros, foi utilizado o *GridSearchCV*. A função é uma ferramenta da biblioteca *Scikit-learn* que permite otimizar os hiperparâmetros de um modelo de *machine learning* de forma sistemática, a partir de um *range* estipulado pelo usuário. Cada algoritmo possui hiperparâmetros específicos de ajuste, otimizando o modelo obtido.

Em todas as etapas de treinamento do modelo, foram utilizadas as métricas de avaliação selecionadas para esse trabalho: Acurácia e RMSE. De modo a facilitar a comparação das métricas de avaliação e a tendência do modelo, gráficos foram gerados evidenciando na parte superior as métricas de avaliação, além da tendência dos dados reais de laboratório acompanhadas da incerteza de medição juntamente com o resultado predito pelo modelo.

5. RESULTADOS E DISCUSSÃO

5.1 TRATAMENTO DOS DADOS DE LABORATÓRIO

O tratamento adequado dos dados de laboratório é fundamental, uma vez que contêm a variável objetivo a ser predita. Os dados de laboratório são resultado do *input* manual do operador após a análise de laboratório, e, portanto, não foram detectados valores negativos desproporcionais. Para a remoção dos *outliers*, foi considerada a remoção de *outliers* a 4 desvios padrão, resultando em uma limpeza de 0,88% dos dados.

Além disso, compreende-se que os dados de laboratório contemplam todos os materiais produzidos pela unidade de polimerização. Considerando os patamares de IF para cada um dos 4 materiais de PEBD, foram atribuídos 3 grupos conforme apresentado na Tabela 6.

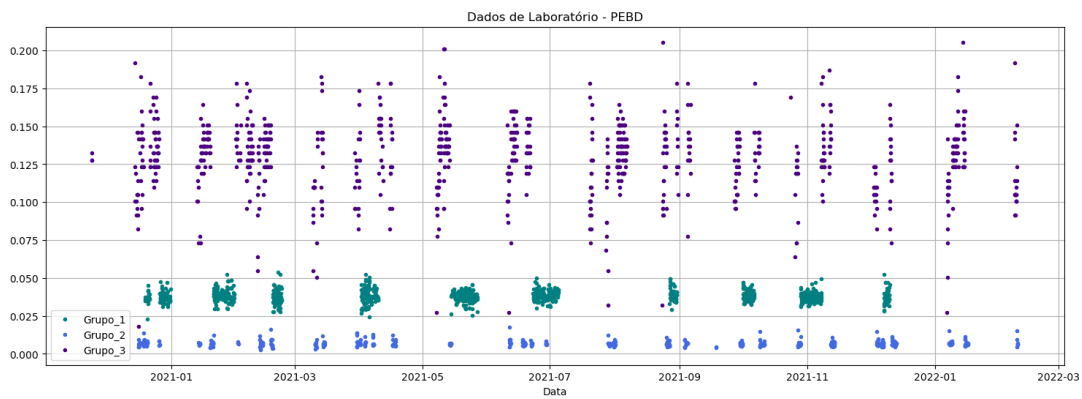
Tabela 6 – Agrupamento dos materiais por grupo de IF

Material	Grupo
1	Grupo 1
5	Grupo 2
4, 10	Grupo 3

Fonte: Elaborado pelo autor.

A normalização dos resultados de laboratório foi desenvolvida agrupando os dados por tipo de Material. A Figura 15 apresenta a dispersão dos dados de laboratório normalizados.

Figura 15 – Dispersão dos dados de laboratório



Fonte: Elaborado pelo autor.

5.2 TRATAMENTO DOS DADOS DE SENSOR

O tratamento adequado dos dados de sensor antes do treinamento de modelos de *machine learning* é uma etapa fundamental para garantir a eficácia e a precisão dos resultados. Os dados de sensor extraídos contêm códigos de erro (valores negativos altos), ruídos e valores ausentes. Os dados de sensor processados geraram um *dataset* com uma coluna *index* representando um *timestamp* de minuto a minuto, e 40 colunas com as variáveis de processo recebidas. Todas as variáveis de processo foram renomeadas conforme a sua área, ou seja, EXT_sensor_X (extrusão), MD_sensor_X (modificadores), RCT_sensor_X (reator) e RCT_TC_sensor_X (trocadores).

Para a primeira limpeza dos dados foi considerada que durante o período de extração de dados ocorreram cenários de *shutdown*. Esse cenário refere-se ao processo de interrupção planejada ou não planejada das operações de produção de uma fábrica ou instalação industrial, seja por motivos de manutenção, atualizações tecnológicas, falhas imprevistas, questões de segurança ou outros fatores que exigem que a planta pare de operar. Para essa situação foi utilizado o corte especificado na Tabela 7.

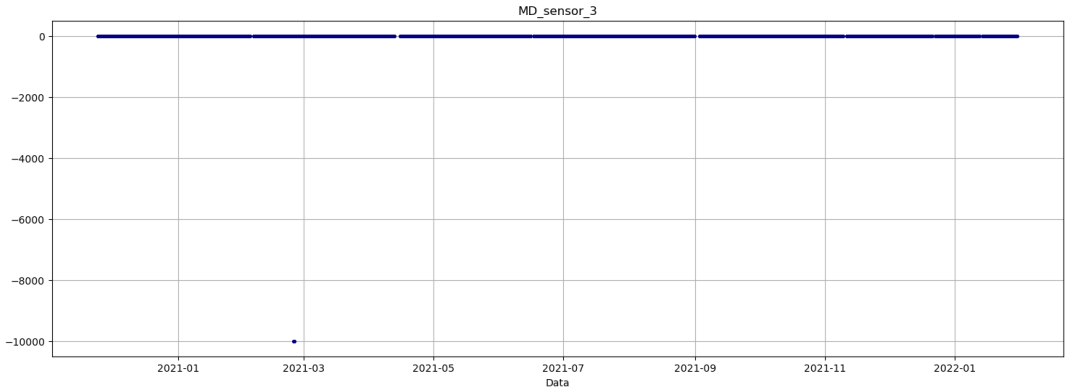
Tabela 7 – Remoção de *plant shutdown*

Variável	Mínimo	% Removido
RCT_sensor_1	1375	8,87

Fonte: Elaborado pelo autor.

A etapa seguinte consistiu na remoção de *outliers*. A Figura 16 demonstra a dificuldade de visualização de dados de sensores devido a *outliers*, e a magnitude da diferença entre eles e o padrão observado para os demais dados.

Figura 16 – Dados de sensor antes da remoção de *outliers*



Fonte: Elaborado pelo autor.

Entretanto, o cálculo é dependente da obtenção de um valor médio representativo, que sofre impacto de valores negativos elevados como os observados na Figura 16. Desse modo, todos os valores negativos foram transformados em NaN. Para o cálculo da média, foi levado em consideração que cada produto possui uma condição de processo específica, e a média deve ser realizada por tipo de material produzido. Para obtenção do material produzido foram utilizados dados de laboratório brutos interpolados de minuto a minuto, para adequar-se à série temporal dos dados de sensores. A transformação dos dados de laboratório foi exemplificada na Figura 17.

Figura 17 – Interpolação dos dados de laboratório para corte de *outlier* das variáveis de sensor

Timestamp	Material
01/01/2023 12:00	1
01/01/2023 15:00	2
01/01/2023 17:00	3

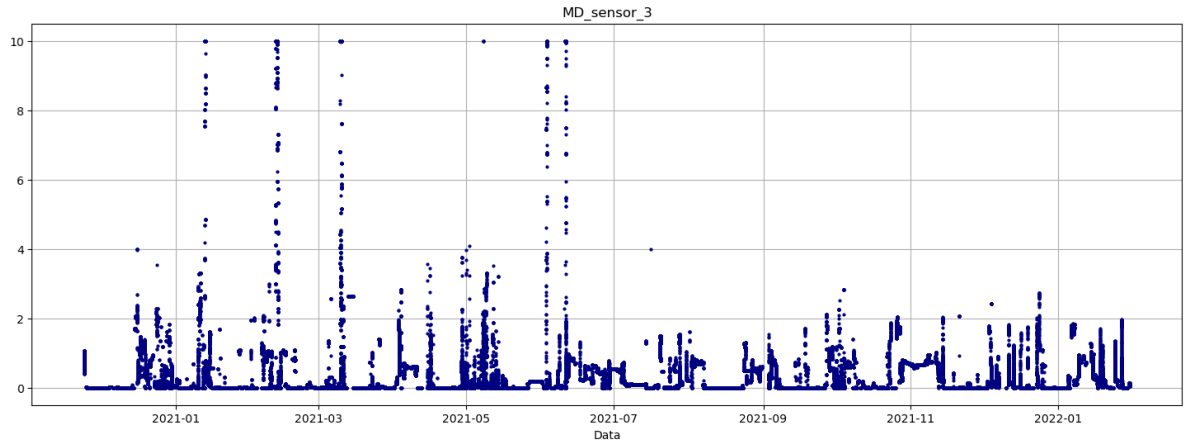
→

Timestamp	Material
01/01/2023 12:00	1
01/01/2023 12:01	1
01/01/2023 12:02	1
...	...
01/01/2023 14:59	1
01/01/2023 15:00	2
01/01/2023 15:00	2

Fonte: Elaborado pelo autor.

Desse modo, a média foi calculada por produto e foi considerada uma remoção conservadora de 4 σ, uma vez que remoções mais restritivas propiciariam um corte superior a 3% dos dados disponíveis para algumas variáveis. A Figura 18 mostra a exemplificação da variável MD_sensor_3 após a remoção de *outliers*.

Figura 18 – Dados sem *outliers*



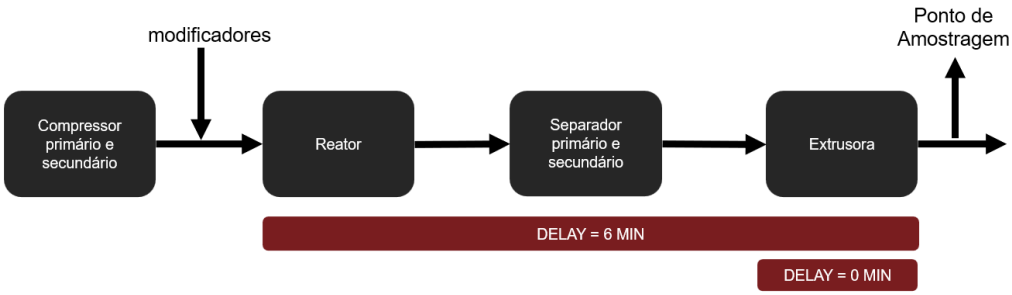
Fonte: Elaborado pelo autor.

5.3 INTEGRAÇÃO DOS DADOS DE SENSOR E DE LABORATÓRIO

Os dados de sensor são coletados de minuto a minuto, enquanto os dados de laboratório são coletados em intervalos de aproximadamente 3h. Tanto as amostras dos sensores quanto os resultados laboratoriais estão indexados pela data e hora da extração e, portanto, a junção dos dois conjuntos pode ser feita a partir interseção das datas e horas presentes em cada conjunto.

Para a junção dos *datasets* é importante considerar o atraso entre o efeito das variáveis de sensor sobre o valor da variável de laboratório. Como exemplo, pode-se tomar um caso de alteração de alguma variável de processo no reator por decisão operacional. O resultado da alteração do processo não será visto na qualidade do produto no mesmo tempo, pois existe um *delay* relacionado ao tempo de residência dos equipamentos e a saída do produto. Considerando a velocidade do processo, foram admitidas como premissas desse estudo os valores de *delay* apresentados na Figura 19.

Figura 19 – Delay considerado para as etapas de processo



Fonte: Elaborado pelo autor.

Após a aplicação do *delay* dos dados de sensor baseado nos valores de tempo de residência, os *datasets* foram unificados considerando as datas indexadas no *dataset* de laboratório. O perfil do *dataset* obtido se assemelha ao apresentado na Figura 20.

Figura 20 – *Dataset* obtido da integração entre os dados de sensor e laboratório

Timestamp	Resultado Laboratório	Variável de sensor 1	...	Variável de sensor X
01/01/2023 12:00	0.05	0.43		0.55
01/01/2023 15:00	0.27	0.44		0.51
01/01/2023 17:00	0.58	0.37		NaN
01/01/2023 20:00	0.32	0.40		0.49

Fonte: Elaborado pelo autor.

No treinamento de modelos de *machine learning*, a presença de valores ausentes, comumente representados como NaN, é um problema crítico que precisa ser resolvido antes de prosseguir com a modelagem. A maioria dos algoritmos de *machine learning* não consegue lidar diretamente com valores NaN, pois dependem de entradas completas e numéricas para realizar cálculos como distâncias, gradientes ou probabilidades. A presença de NaN interrompe o treinamento e causa erros computacionais. Desse modo, a função *pandas.DataFrame.dropna* foi utilizada. A quantidade de dados disponíveis para o treinamento, por grupo de material, está descrita na Tabela 8.

Tabela 8 – Quantidade de dados utilizados para treino dos modelos por grupo

Grupo	Quantidade de dados
Grupo 1	589
Grupo 2	273
Grupo 3	339

Fonte: Elaborado pelo autor.

5.4 TREINAMENTO DOS MODELOS

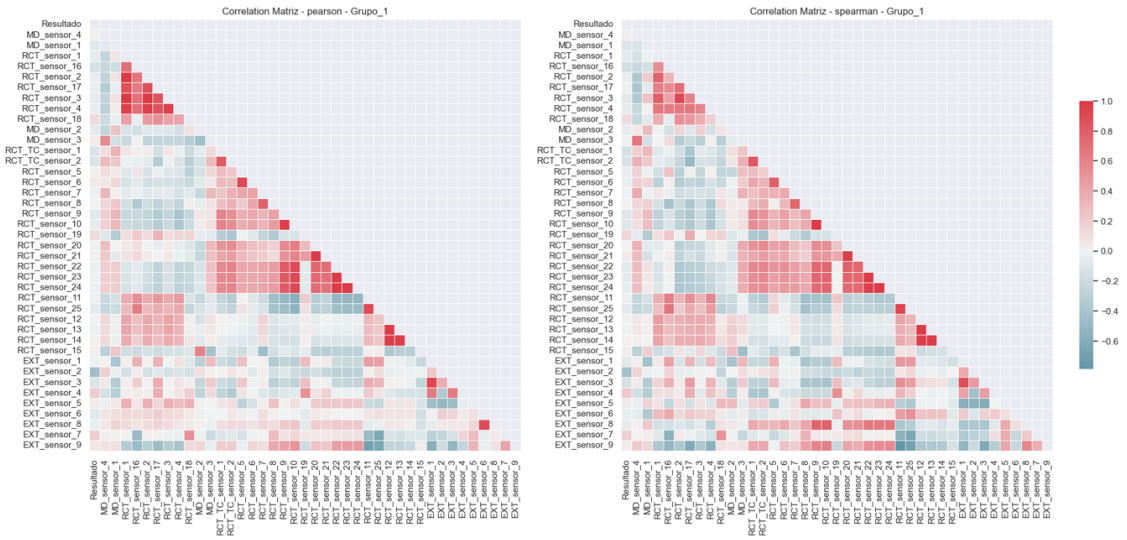
Para explorar as características do conjunto de dados obtido foram obtidos dados de correlação das variáveis de processo com o valor de IF real de laboratório. Em seguida, os modelos de ML foram treinados para a definição do melhor algoritmo.

5.4.1 ANÁLISE DA CORRELAÇÃO DAS VARIÁVEIS

Os patamares existentes para cada um dos grupos dos dados de laboratório, representam uma condição de processo ajustada para a obtenção da especificação do polímero processado, e, portanto, variáveis altamente correlacionadas para um grupo não necessariamente serão as variáveis altamente correlacionadas para outro grupo. Isso ocorre porque diferentes concentrações de modificadores podem ser aplicadas, e as variáveis de processo como temperatura e pressão podem ser ajustadas. Para a análise de correlação, o *dataset* foi dividido em Grupo 1 (G1), Grupo 2 (G2) e Grupo 3 (G3).

A Figura 21 apresenta o mapa de calor desenvolvido para o G1, omitindo o triângulo superior e a diagonal central para facilitar a visualização, e a variável "Resultado" representa a análise de laboratório do IF real.

Figura 21 – *Heatmap* obtido para o Grupo 1 considerando os métodos de Pearson e Spearman



Fonte: Elaborado pelo autor.

A tabela 9 foi desenvolvida com base na correlação específica de cada uma das variáveis de sensor em relação a variável de laboratório. A tabela não considera variáveis que tiver correlação inferior a ±0.15. É possível notar que as existe certa similaridade na correlação das variáveis, sendo as mais significativas RCT_TC_sensor_1, RCT_sensor_15, EXT_sensor_5, EXT_sensor_2, EXT_sensor_3, RCT_sensor_2 e RCT_sensor_3.

Tabela 9 – Correlação de Pearson e Spearman para cada um dos grupos

Variável	Correlação G1 Pearson	Correlação G1 Spearman	Correlação G2 Pearson	Correlação G2 Spearman	Correlação G3 Pearson	Correlação G3 Spearman
Resultado	1	1	1	1	1	1
RCT_TC_sensor_1	0.4468	0.2952	0.4468	0.2952	0.4468	0.2952
RCT_sensor_15	0.3533	0.3879	0.3533	0.3879	0.3533	0.3879
EXT_sensor_5	0.2892	0.0459	0.2892	0.0459	0.2892	0.0459
RCT_sensor_7	0.2800	0.0276	0.2800	0.0276	0.2800	0.0276
RCT_sensor_6	0.2680	0.0948	0.2680	0.0948	0.2680	0.0948
RCT_sensor_5	0.2599	0.0725	0.2599	0.0725	0.2599	0.0725

Variável	Correlação G1 Pearson	Correlação G1 Spearman	Correlação G2 Pearson	Correlação G2 Spearman	Correlação G3 Pearson	Correlação G3 Spearman
RCT_sensor_8	0.2582	0.0845	0.2582	0.0845	0.2582	0.0845
MD_sensor_3	0.2127	0.1807	0.2127	0.1807	0.2127	0.1807
EXT_sensor_4	0.1565	0.0060	0.1565	0.0060	0.1565	0.0060
RCT_sensor_12	-0.1539	0.1486	-0.1539	0.1486	-0.1539	0.1486
RCT_sensor_14	-0.2193	0.0871	-0.2193	0.0871	-0.2193	0.0871
MD_sensor_1	-0.3338	-0.3438	-0.3338	-0.3438	-0.3338	-0.3438
RCT_sensor_17	-0.4092	-0.4986	-0.4092	-0.4986	-0.4092	-0.4986
RCT_sensor_4	-0.4311	-0.5632	-0.4311	-0.5632	-0.4311	-0.5632
RCT_sensor_18	-0.4441	-0.2977	-0.4441	-0.2977	-0.4441	-0.2977
RCT_sensor_16	-0.4753	-0.5567	-0.4753	-0.5567	-0.4753	-0.5567
RCT_sensor_3	-0.4763	-0.6144	-0.4763	-0.6144	-0.4763	-0.6144
RCT_sensor_2	-0.4912	-0.6300	-0.4912	-0.6300	-0.4912	-0.6300
EXT_sensor_6	-0.4974	-0.5063	-0.4974	-0.5063	-0.4974	-0.5063
EXT_sensor_8	-0.5040	-0.6117	-0.5040	-0.6117	-0.5040	-0.6117
RCT_sensor_1	-0.5078	-0.6437	-0.5078	-0.6437	-0.5078	-0.6437
EXT_sensor_3	-0.5295	-0.6015	-0.5295	-0.6015	-0.5295	-0.6015
EXT_sensor_2	-0.7072	-0.7398	-0.7072	-0.7398	-0.7072	-0.7398

Fonte: Elaborado pelo autor.

4.6.2 TREINAMENTO DOS MODELOS

O treinamento de modelos de *machine learning* para predição do IF foi realizado utilizando os algoritmos selecionados.

Para o modelo *baseline* do Grupo 1, as métricas de acurácia e RMSE podem ser vistas na Tabela 10.

Tabela 10 – Métricas de Acurácia e RMSE para o *baseline* do Grupo 1

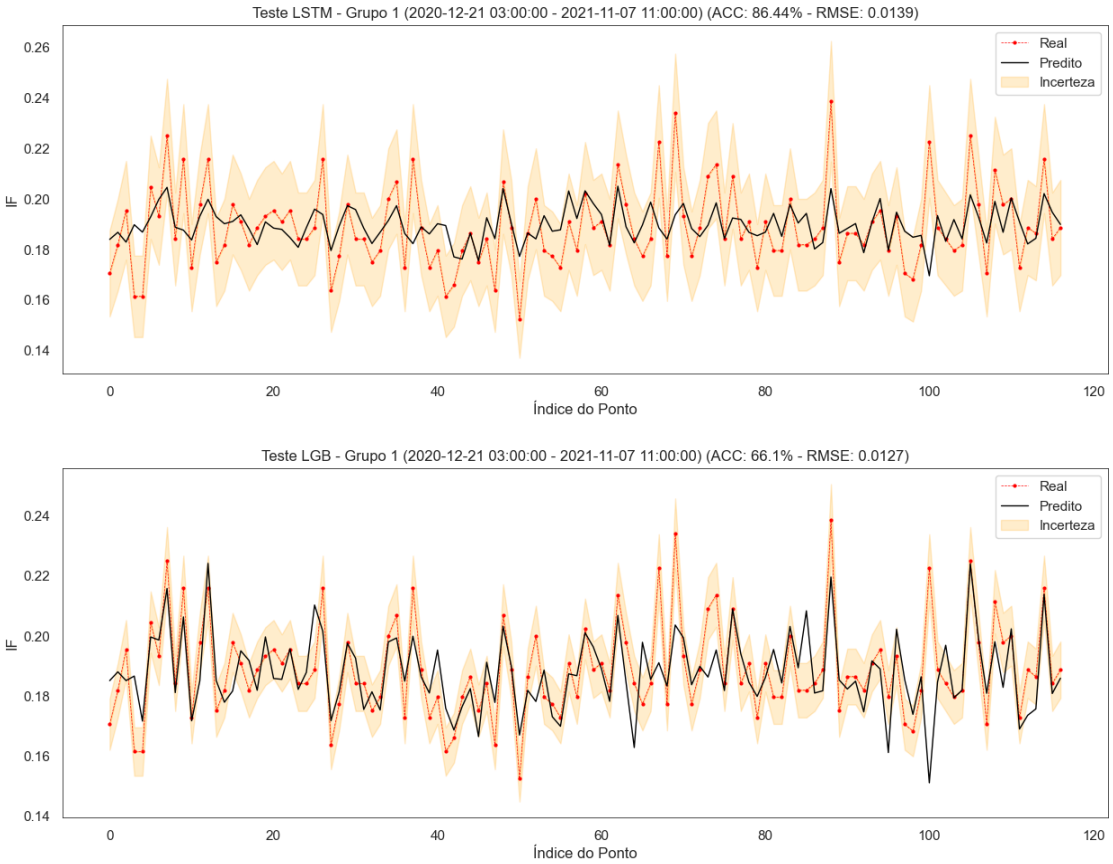
	ACC - Baseline	RMSE - Baseline
ACC GB	61,02	0,013
ACC LGBM	66,1	0,0127
ACC LSTM	86,44	0,0139

Fonte: Elaborado pelo autor.

Ao analisar as métricas geradas, têm-se a expectativa de maior aderência do modelo de LSTM devido à alta acurácia. Entretanto, a Figura 22 mostra a diferença entre o modelo LSTM e *LightGBM*. É possível observar que o modelo LSTM não acompanha

as tendências de processo, e desse modo, o valor do modelo resultante é próximo a média da variável objetivo. Desse modo, o modelo *LightGBM* foi selecionado para as etapas de otimização.

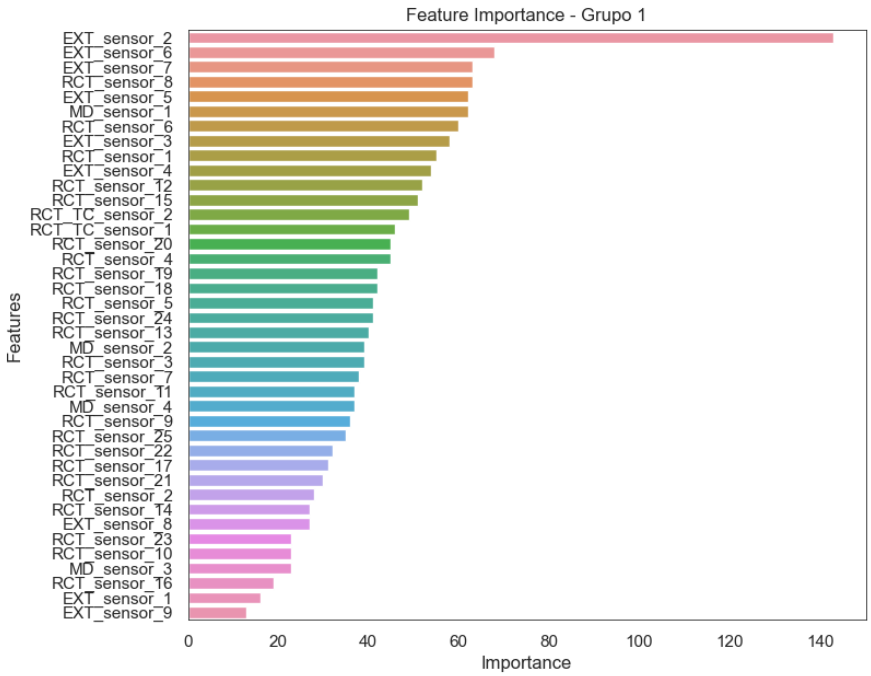
Figura 22 – Modelos *baseline* de LSTM e *LightGBM* obtidos para o Grupo 1



Fonte: Elaborado pelo autor.

Para a otimização do modelo, foi considerada a redução das variáveis do modelo. Essa etapa é importante pois variáveis irrelevantes ou redundantes podem introduzir ruído nos dados, reduzindo a capacidade do modelo de capturar padrões importantes. A Figura 23 apresenta a *feature importance* do modelo *baseline* selecionado, na qual é possível notar que variáveis de extrusão 2, 6 e 7 apresentaram grande importância para o modelo.

Figura 23 – *Feature Importance* obtida no modelo *baseline LightGBM* para o Grupo 1



Fonte: Elaborado pelo autor.

A Tabela 11 apresenta os hiperparâmetros considerados e o range de teste, sendo que a métrica usada para o desempenho foi o Erro Quadrático Médio (EQM).´

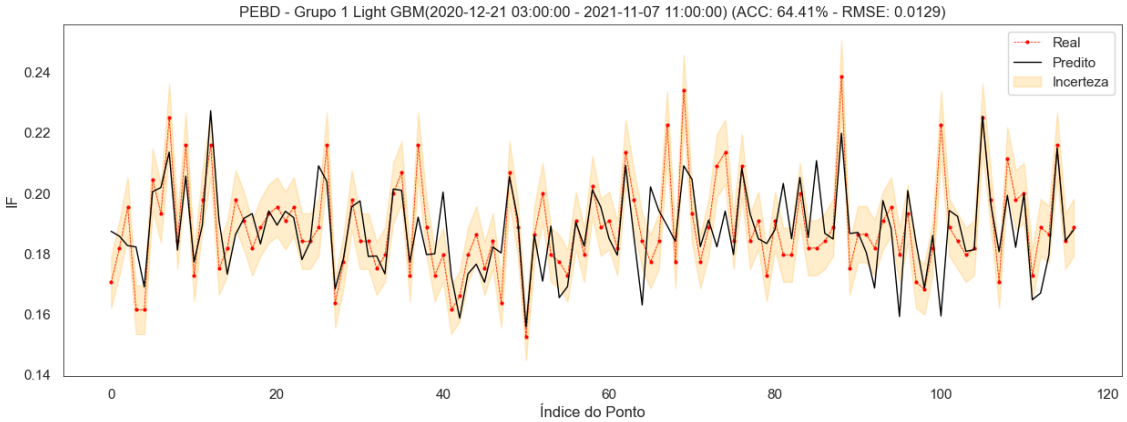
Tabela 11 – Faixa de Hiperparâmetros considerados para o modelo do Grupo 1

Hiperparâmetro	**Valores considerados**
learning_rate	[0.01, 0.05, 0.1],
n_estimators	[50, 100, 200],
max_depth	[3, 5, 7],
num_leaves	[20, 31, 50],
feature_fraction	[0.8, 0.9, 1.0],
bagging_fraction	[0.8, 0.9, 1.0]

Fonte: Elaborado pelo autor.

O modelo final obtido para o Grupo 1 pode ser visto na Figura 24 , juntamente com as variáveis selecionadas para o modelo e os hiperparâmetros.

Figura 24 – Modelo final Grupo 1, Variáveis selecionadas e Hiperparâmetros otimizados



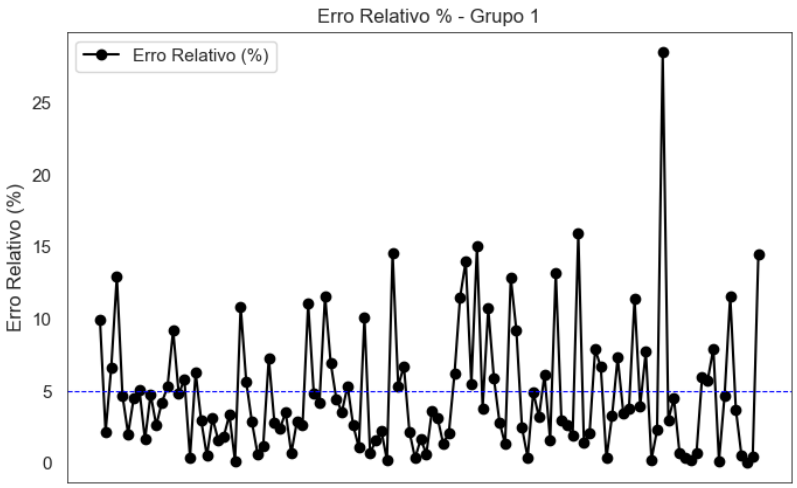
Variáveis selecionadas	
MD_sensor_1	RCT_sensor_20
RCT_sensor_1	RCT_sensor_21
RCT_sensor_3	RCT_sensor_24
RCT_sensor_4	RCT_sensor_12
RCT_sensor_18	RCT_sensor_13
MD_sensor_2	RCT_sensor_14
RCT_TC_sensor_1	RCT_sensor_15
RCT_TC_sensor_2	EXT_sensor_2
RCT_sensor_5	EXT_sensor_3
RCT_sensor_6	EXT_sensor_4
RCT_sensor_8	EXT_sensor_5
RCT_sensor_9	EXT_sensor_6
RCT_sensor_19	EXT_sensor_7

Hiperparâmetros	
bagging_fraction	0.8
feature_fraction	0.9
learning_rate	0.15
max_depth	7
n_estimators	200
num_leaves	20

Fonte: Elaborado pelo autor.

É possível notar que o modelo possui alta aderência as tendências de processo, conseguindo acompanhar a inclinação de valores altos e baixos da variável objetivo, como os pontos aproximadamente 10, 25, 60 e 105. Entretanto, existem pontos que o modelo parece atrasar a predição em relação aos valores reais, como os pontos 18, 40 e 95. É possível que o *delay* considerado para as variáveis de processo não esteja corretamente ajustado, e nos pontos mencionados, a relevância dessas variáveis seja alta, o que promove o descompasso entre a variável objetivo e o resultado de laboratório. Para entender a inclinação do modelo ao erro, a Figura 25 foi desenvolvida. Essa figura demonstra a variação do erro relativo do modelo, sendo que a maioria do erro relativo é menor que 15%, demonstrando a alta aderência do modelo aos dados utilizados.

Figura 25 - Erro relativo obtido para o Grupo 1



Fonte: Elaborado pelo autor.

Com relação ao Grupo 2, a Tabela 12 apresenta os resultados obtidos dos modelos *baseline* para os três algoritmos considerados. É possível notar que o valor obtido para as métricas consideradas é baixo, o que inicialmente o modelo demonstra baixa aderência dos dados de processo em relação a variável objetivo.

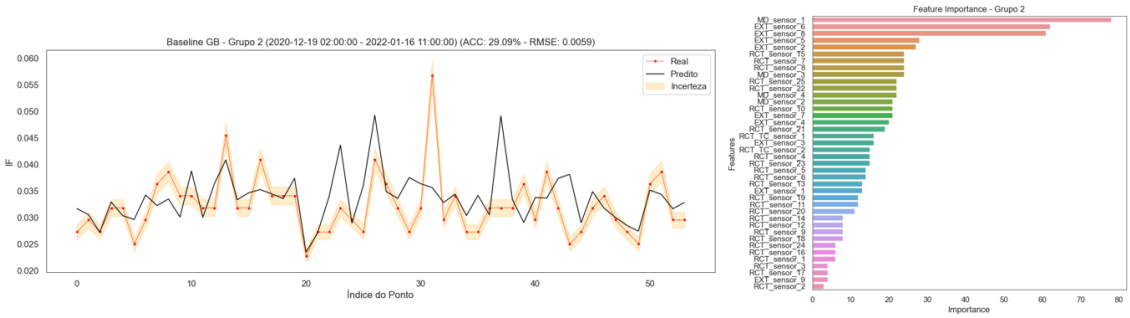
Tabela 12 - Métricas de Acurácia e RMSE para o *baseline* do Grupo 2

	ACC - Baseline	**RMSE - Baseline**
ACC GB	29,09	0,0059
ACC LGBM	25,45	0,0056
ACC LSTM	18,18	0,0071

Fonte: Elaborado pelo autor.

Dentre os modelos estudados, o modelo de *GradientBoosting* apresentou o melhor desempenho e sua tendência pode ser observada na Figura 26, juntamente com a *feature importance* desse modelo.

Figura 26 - Modelo *Baseline GradientBoosting* obtido para o Grupo 2



Fonte: Elaborado pelo autor.

Para a otimização do modelo, foi considerada a redução das variáveis do modelo, removendo variáveis irrelevantes ou redundantes. A Tabela 13 apresenta os hiperparâmetros considerados e o range de teste.

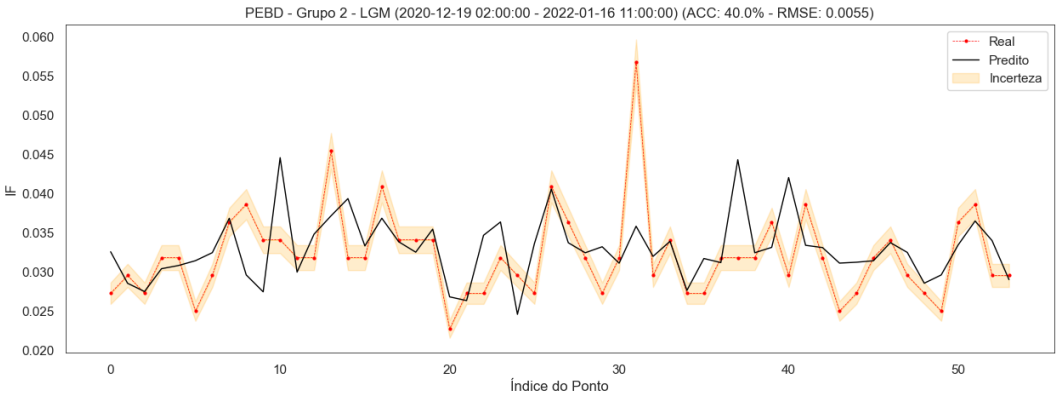
Tabela 13 - Faixa de Hiperparâmetros considerados para o modelo do Grupo 2

Hiperparâmetro	Valores considerados
n_estimators	[100, 200]
learning_rate	[0.01, 0.1, 0.2]
max_depth	[3, 5, 7]
min_samples_split	[2, 5, 10]
subsample	[0.8, 1.0]

Fonte: Elaborado pelo autor.

O modelo final obtido para o Grupo 2 pode ser visto na Figura 27 , juntamente com as variáveis selecionadas para o modelo e os hiperparâmetros.

Figura 27 - Modelo final do Grupo 2, variáveis selecionadas e hiperparâmetros otimizados



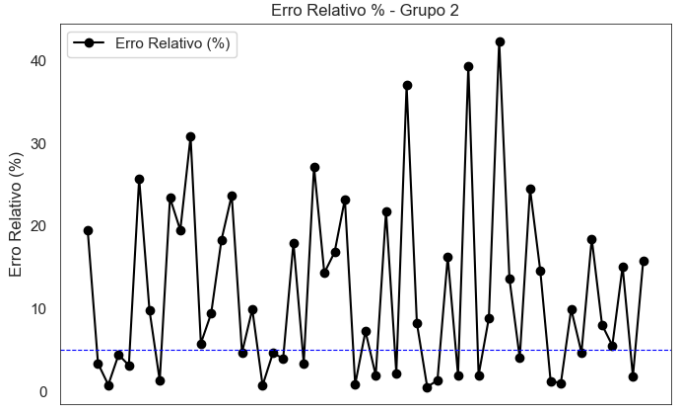
Variáveis selecionadas
MD_sensor_1
EXT_sensor_6
EXT_sensor_8
EXT_sensor_5
RCT_sensor_10
MD_sensor_3
EXT_sensor_2

Hiperparâmetros	
learning_rate	0.1
n_estimators	100
min_samples_split	2
subsample	1.0
max_depth	3

Fonte: Elaborado pelo autor.

Apesar da seleção das variáveis e otimização dos hiperparâmetros aumentar a acurácia e reduzir o RMSE, o modelo não conseguiu acompanhar corretamente as tendências da variável predita. É possível notar que próximo ao ponto 10 o modelo adianta o pico real, sendo que o mesmo acontece próximo ao ponto 40. A baixa aderência pode ser explicada pela limitação de quantidade de dados, demonstrando que o modelo não conseguiu aprender suficiente para executar boas predições. Em comparação ao Grupo 1, o modelo tem apenas 50% dos dados disponíveis. A Figura 28 evidencia o erro relativo do modelo, atingindo até 40%.

Figura 28 - Erro relativo obtido para o Grupo 2



Fonte: Elaborado pelo autor.

O modelo *baseline* do Grupo 3, foi selecionado a partir da Tabela 14, na qual é possível notar que o modelo de *LightGBM* apresentou maior potencial em relação aos demais.

Tabela 14 - Métricas de Acurácia e RMSE para o *baseline* do Grupo 3

	ACC - Baseline	**RMSE - Baseline**
ACC GB	35,29	0,1054
ACC LGBM	35,29	0,0967
ACC LSTM	29,41	0,1029

Fonte: Elaborado pelo autor.

Para a otimização do modelo, foi considerada a redução das variáveis do modelo, removendo variáveis irrelevantes ou redundantes. A Tabela 15 apresenta os hiperparâmetros considerados e o range de teste.

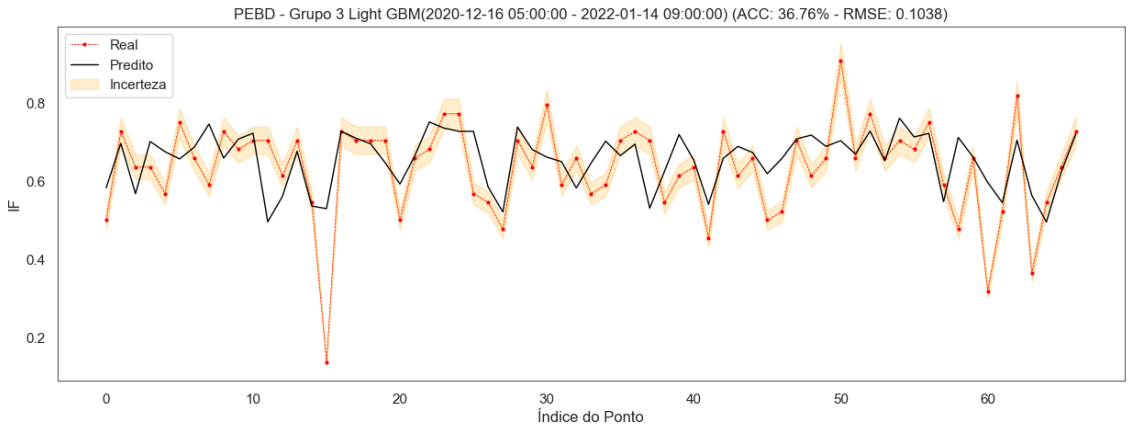
Tabela 15 - Faixa de Hiperparâmetros considerados para o modelo do Grupo 3

Hiperparâmetro	**Valores considerados**
learning_rate	[0.01, 0.05, 0.1],
n_estimators	[50, 100, 200],
max_depth	[3, 5, 7],
num_leaves	[20, 31, 50],
feature_fraction	[0.8, 0.9, 1.0],
bagging_fraction	[0.8, 0.9, 1.0]

Fonte: Elaborado pelo autor.

O modelo final obtido para o Grupo 3 pode ser visto na Figura 29, juntamente com as variáveis selecionadas para o modelo e os hiperparâmetros.

Figura 29 - Modelo final do Grupo 3, variáveis selecionadas e hiperparâmetros otimizados



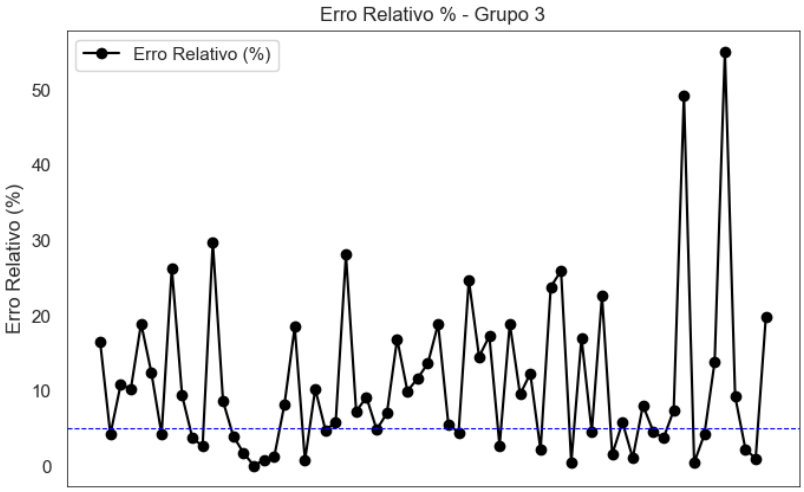
Variáveis selecionadas
RCT_sensor_8
RCT_sensor_15
EXT_sensor_5
EXT_sensor_3
RCT_sensor_5
RCT_sensor_18
EXT_sensor_2

Hiperparâmetros	
bagging_fraction	1
feature_fraction	1
learning_rate	0.1
max_depth	0
n_estimators	100
num_leaves	20

Fonte: Elaborado pelo autor.

Dentre os modelos estudados, o Grupo 3 é o que apresenta maior oscilação nos valores de IF. A oscilação é de ±0.2 do índice de fluidez (IF) normalizado, fator que decorre do agrupamento de dois materiais (Materiais 4 e 10). Apesar da oscilação dos dados, o modelo acompanha algumas tendências de processo, mas apresenta um patamar *flat* entre os pontos 40 e 50. Para casos em que os dados não refletem a variável objetivo, é necessário entender junto a engenharia de processos quais cenários incomuns de operação ocorreram durante o período de dados extraídos. O processo de polimerização é complexo e possui um conjunto de equipamentos em série, sendo suscetível a oscilações e cenários fora do padrão de operação. Considerando a baixa disponibilidade de dados, o modelo visivelmente possui dificuldade ao lidar com oscilações abruptas, observado em transição de valores de baixo IF para alto IF. A Figura 30 evidencia o erro relativo do modelo, destacando que um número considerável de predições o erro obtido foi inferior a 20%.

Figura 30 - Erro relativo obtido para o Grupo 3



Fonte: Elaborado pelo autor.

6. CONCLUSÃO

Este estudo apresentou uma nova metodologia para a predição de Índice de Fluidez (IF) no processo de polimerização em Alta Pressão, utilizando diferentes algoritmos. Os dados reais de laboratório foram utilizados para treinar e validar o modelo, treinado a partir de variáveis de sensores: extrusão, reação, trocador de calor da reação e injeção de modificadores.

O tratamento dos dados contribuiu para um modelo mais confiável devido a substituição de medições não representativas do processo ou códigos de erro. Os dados de sensor e de laboratório foram limpos considerando 4 desvios padrão, removendo os *outliers* presentes. A integração entre os dados foi realizada a partir de um *delay* de 6 minutos no reator e 0 minutos na extrusora, mantendo o *timestamp* dos dados de laboratório. Para a modelagem preditiva, foram utilizados algoritmos de aprendizado supervisionado, sendo que os dados foram divididos em 3 grupos, conforme a faixa do Índice de Fluidez. Um modelo baseline foi obtido para cada um dos grupos, sendo uma etapa importante para definir o modelo com maior potencial dentre os algoritmos estudados: *GradientBoosting*, *LightGBM* e LSTM. A seleção da variáveis foi considerada para otimizar o modelo, visando a redução de ruídos e do risco associado ao número alto de variáveis de processo. O risco decorre da dependência do modelo em relação as variáveis de sensor, e caso ocorra algum erro com uma das variáveis, o modelo não consegue efetuar predições. A seleção de variáveis foi efetuada por tentativa e erro, buscando aumentar as métricas selecionadas (Acurácia e RMSE), bem como melhorar a tendência das predições. Além disso, a otimização foi realizada a partir dos hiperparâmetros de cada um dos algoritmos, automatizado pela função do *scikit-learn GridSearchCV*.

Dentre os *soft-sensors* propostos, o Grupo 1 apresentou o melhor desempenho, atingindo uma acurácia de 64,41% e um forte acompanhamento as tendências de processo. Esse resultado condiz com a maior disponibilidade de dados para treino e teste. Por outro lado, os Grupos 2 e 3 não apresentaram predições aderentes. O Grupo 3 apesar de acompanhar as tendências de processo, falha em atingir as predições no mesmo tempo da análise de laboratório. Um estudo mais aprofundado do tempo de residência para a melhor seleção do *delay* das variáveis de processo deve ser conduzido, uma vez que foram utilizadas variáveis como temperatura, pressão e concentração.

Dentre os algoritmos estudados, 2 modelos foram selecionados utilizando *LightGMB* e 1 utilizando *GradientBoosting*. O modelo de *LightGMB* apresentou boa aderência em grupos com alta variabilidade (Grupo 1 e 3), enquanto o *GradientBoosting* foi mais adequado quando a variável predita oscilou discretamente. O algoritmo LSTM não se adequou ao tipo de processo, que constitui de um número grande de variáveis de sensor e comportamentos bastante oscilatórios.

Os resultados desse trabalho indicam haver uma significativa correlação entre os dados disponíveis em tempo real no processo da polimerização em alta pressão e os resultados da análise laboratorial. A melhora dos modelos pode ocorrer a partir de estudos futuros no ajuste de *delay* dos conjuntos de variáveis, além da coleta e disponibilidade maior de dados – em especial Grupo 2 e 3.

7. REFERÊNCIAS BIBLIOGRÁFICAS

AHMAD I, *et al*. Gray-box Soft Sensors in Process Industry: Current Practice, and Future Prospects in Era of Big Data. Processes, Vol. 8, 2020.

AKBAS, A; BUYRUKOGLUS, S. Machine Learning based Early Prediction of Type 2 Diabetes: A New Hybrid Feature Selection Approach using Correlation Matrix with Heatmap and SFS. Balkan journal of electrical & computer engineering, Vol. 10, No. 2, 2022.

ALABDULLAH, A.A *et al.* Prediction of rapid chloride penetration resistance of metakaolin based high strength concrete using light GBM and XGBoost models by incorporating SHAP analysis. Construction and Building Materials, Vol. 345, 2022.

ALAM, B. Implementation of K-means clustering algorithm using Python. Hands-on.cloud. Available in: https://hands-on.cloud/implementation-of-k-means-clustering-algorithm-using-python/ . Acesso em 10/12/2022.

Al-Selwi, S. M. *et al.* LSTM Inefficiency in Long-Term Dependencies Regression Problems. Journal of Advanced Research in Applied Sciences and Engineering Technology, Vol. 30, Issue 3, 16-31 p, 2023.

ARAGAO, M. V. C., MAFRA, S. B; FIGUEIREDO, F. A. P. Análise de Tráfego de Rede com Machine Learning para Identificação de Ameaças a Dispositivos IoT. XL Simpósio brasileiro de telecomunicações e processamento de sinais - SBrT. 25-28 p. 2022.

ARANTES, C.A. Análise da depreciação de áreas com uso da distribuição gaussiana. congresso brasileiro de engenharia de avaliações e perícias. Foz do Iguaçu - PR, 2017.

BARRETO, L., AMARAL, A., PEREIRA, T. Industry 4.0 implications in logistics: an overview. Procedia Manufacturing. Vol. 13. 1245-1252 p. 2017.

BELAGOUNE, S. *et al*. Deep learning through LSTM classification and regression for transmission line fault detection, diagnosis and location in large-scale multi-machine power systems. Measurement, Vol. 177. 2021.

BORGES, M. M. Machine learning como ferramenta gerencial para predição de indicadores e detecção de anomalias. Dissertação de mestrado. Porto Alegre. 2020.

CHANG, H.J., BHADURI, A. Rethinking Development Economics. Anthem Frontiers of Global Political Economy and Development Series. Anthem Press. New York. 544 p. 2003.

DOMÍNGUEZ-NIÑO, A., *et. al*. Energy Requirements and Production Cost of the Spray Drying Process of Cheese Whey. Drying Technology. Vol. 36. Edition 10. Guadalupe. 2017.

FANG, J. *et. al.* Machine learning accelerates the materials Discovery. Materials Today Communications. Vol. 33. 2022.

FLASK, W., DOCKER, S., KUBERNETES. Deploy Machine Learning Models to Production. Pramod Singh. India. 150 p. 2021.

FRIEDMAN, J. H., HASTIE, T., TIBSHIRANI, R. Additive Logistic Regression: A Statistical View of Boosting. The Annals of Statistics. Vol. 28. 337-407 p. 2000.

GUEDES, E. V. O. Aplicação de soft sensor baseado em redes neurais artificiais e random forest para predição em tempo real do teor de ferro no concentrado da flotação de minério de ferro. Dissertação de Mestrado. Ouro Preto MG. 2020.

HAYASHI, C. *et al.* What is Data Science? Fundamental concepts and a heuristic example. In Data Science, Classification, and Related Methods. Springer: Heidelberg. Germany. 40–51 p. 1998.

JAEHYUN, K. *et al.* Catalyze Materials Science with Machine Learning. American Chemical Society. Vol. 8. 1151- 1171 p. 2021.

KADLEC, P.; GABRYS, B.; STRANDT, S. Data-driven soft sensors in the process industry. Comput. Chem. Eng. Vol. 33. 795–814 p. 2009.

KAJAL, R. *et al.* Decision Tree Based Algorithm for Intrusion Detection. International Journal of Advanced Networking and Applications. Vol. 7. 2828-2834 p. 2016.

KERAS. Disponível em https://keras.io/api/layers/recurrent_layers/lstm/. Acessado em 09/12/2024.

LIDIANE, L. M. K. *et al*. Análise Fatorial por Meio da Matriz de Correlação de Pearson e Policórica no Campo das Cisternas. Engineering and Science, Vol. 7, n. 1. 58–70 p, 2018.

LIGTHGBM. Disponível em: https://lightgbm.readthedocs.io/en/latest/Parameters.html. Acessado em 09/12/2024.

LIKAS, A., VLASSIS, N., VERBEEK. J. J. The global k-means clustering algorithm. Pattern Recognition. Vol. 36. Issue 2. 451-461 p. 2003.

LIMA, J. M. M. Soft Sensor Aplicado a Plantas de Processamento de Gás Natural Baseado em Redes Neurais Artificiais. Dissertação de Mestrado. Natal – RN. 2018.

MAHESH B. Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR), 381-386 p. 2019.

MAULUD, D. H., ABDULAZEEZ M. A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends. Vol. 1. 140-147 p. 2020.

MITTAL, K., KHANDUJA, D., TEWARI, P. C. An Insight into “Decision Tree Analysis. WWJRMD. 111-115 p. 2017

PAINULI, D., MISHRA, D., BHARDWAJ, S., MAYANK, A. Forecast and prediction of COVID-19 using machine learning. Data Science for COVID-19. Academic Press. 381-397 p. 2021.

PEACOCK, A. J. Handbook of Polyethylene: Structes, Properties, and Applications. New York: Marcel Dekker, Inc. 534 p. 2000.

ROCHA, M. C. G., COUTINHO, F. M. B., BALKE, S. Indice de Fluidez: Uma Variável de Controle de Processos de Degradação Controlada de Polipropileno por Extrusão Reativa. Polímeros: Ciência e Tecnologia. 33-37 p. 1994.

SCIKIT-LEARN. GradientBoostingRegressor. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html. Acesso em: 09/12/2024.

SHARMA, A., SINGH, D. Evolution of Industrial Revolutions: A Review. International Journal of Innovative Technology and Exploring Engineering. Vol. 9. 2020.

SILVA, J. L. Modelagem e Simulação de reatores autoclave para produção de PEBD. Dissertação de mestrado. Porto Alegre - RS, 2012.

SONG, M. J. *et al.* Soft Sensor for Melt Index Prediction Based on Long Short-Term Memory Network. IFAC-PapersOnLine, Vol. 55, Issue 7. 857-862 p. 2022.

SUN, K. *et al.* Soft Sensor Development with Nonlinear Variable Selection Using Nonnegative Garrote and Artificial Neural Network. In Computer Aided Chemical Engineering; Elsevier: Amsterdam, The Netherlands. Vol. 33. Pages 883–888 p. 2014.

THANGARAJ, J., NARAYANAN, R. L. Industry 1.0 to 4.0: the evolution of smart factories. 2018. Vol. 22. Issue 9. 1624-1636 p. 2012.

VINITHA, K. *et. Al*. Review on industrial mathematics and materials at Industry 1.0 to Industry 4.0. Materials Today: Proceedings. Vol. 33, Part 7. 3956-3960 p. 2020.

YETURU, K. Chapter 3 - Machine learning algorithms, applications, and practices in data Science. Handbook of Statistics. Elsevier. Vol. 43. 81-206 p. 2020

ZENDEHBOUDI, S., REZAEI, N., LOHI, A. Applications of hybrid models in chemical, petroleum, and energy systems: A systematic review Appl. Energy. 2539–2566 p. 2018