ALEXANDRE LADEIRA DE SOUSA

PREDIÇÃO DA ANÁLISE DE CAL LIVRE NA PRODUÇÃO DE CIMENTO POR MEIO DE APRENDIZADO DE MÁQUINA E USO DE DADOS SINTÉTICOS

LAVRAS – MG

2024

ALEXANDRE LADEIRA DE SOUSA

PREDIÇÃO DA ANÁLISE DE CAL LIVRE NA PRODUÇÃO DE CIMENTO POR MEIO DE APRENDIZADO DE MÁQUINA E USO DE DADOS SINTÉTICOS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação para obtenção do título de Mestre.

Prof. DSc. Demóstenes Zegarra Rodríguez
Orientador

LAVRAS – MG
2024

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Sousa, Alexandre Ladeira de

Predição da análise de cal livre na produção de cimento por meio de aprendizado de máquina e uso de dados sintéticos / Alexandre Ladeira de Sousa. – 2024.

81 p. : il.

Orientador(a): Prof. DSc. Demóstenes Zegarra Rodríguez.

Dissertação (mestrado acadêmico)– Universidade Federal de Lavras, 2024.

Bibliografia.

1. Aprendizado de máquina. 2. Cal livre. 3. Cimento. 4. Clínquer. 5. Qualidade. 6. Dados sintéticos. I. Rodríguez, Demóstenes Zegarra. II. -. III. Título.

ALEXANDRE LADEIRA DE SOUSA

PREDIÇÃO DA ANÁLISE DE CAL LIVRE NA PRODUÇÃO DE CIMENTO POR MEIO DE APRENDIZADO DE MÁQUINA E USO DE DADOS SINTÉTICOS

PREDICTION OF FREE LIME ANALYSIS IN CEMENT PRODUCTION THROUGH MACHINE LEARNING AND USE OF SYNTHETIC DATA

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação para obtenção do título de Mestre.

APROVADA em 17 de Dezembro de 2024.

Prof. DSc. Dick Carrillo Melgarejo	LUT/Nokia
Prof. DSc. Vinicius Vitor dos Santos Dias	DCC/UFLA

Prof. DSc. Demóstenes Zegarra Rodríguez
Orientador

LAVRAS – MG
2024

AGRADECIMENTOS

Primeiramente deixo o agradecimento à minha família, pelo apoio irrestrito para os meus estudos desde tão novo. Tudo que até aqui aprendi devo a eles.

Agradeço também ao meu orientador Prof. Demóstenes Zegarra Rodríguez pelas horas dedicadas à minha orientação no trabalho de pesquisa. Somente pelos seus direcionamentos este trabalho tornou-se possível.

Preciso agradecer também à UFLA, que vem fomentando meus estudos desde a graduação, pela excelente estrutura que tornou possível a conclusão de mais essa etapa acadêmica.

Por fim, agradeço aos amigos do trabalho, Rangel Correa e Edson Barboza, que muito contribuiram nas análises das bases de dados a fim de entender o melhor caminho a seguir no presente trabalho.

Sem objetivos bem definidos, somente por um acaso chegaremos a algum lugar.
Pai

RESUMO

A produção de cimento é um processo complexo que envolve etapas de mineração, moagem de matérias-primas e aquecimento de materiais em fornos de clinquerização. Ao longo dessa cadeia produtiva, ocorrem etapas de homogeneização e reações químicas que alteram a estrutura dos compostos para se obter um produto final conforme os padrões de qualidade exigidos por normas. Na etapa de clinquerização, um dos parâmetros cruciais a ser monitorado é o teor de cal livre, que impacta diretamente na qualidade do cimento e na eficiência do processo. No entanto, essa análise é pontual, realizada geralmente a cada duas horas, e inclui etapas de coleta, preparação da amostra e aferição em aparelhos de raio X. Diante disso, o presente trabalho tem como objetivo desenvolver um modelo preditivo para a análise de cal livre nos fornos de clínquer, utilizando técnicas de aprendizado de máquina e geração de dados sintéticos. Foi realizada uma consulta a especialistas da área de processo e a pesquisas relacionadas ao tema para a definição das variáveis com maior impacto no valor de cal livre, seus respectivos limites que caracterizam a estabilidade na operação, bem como casos em que a amostragem das variáveis pudesse ser prejudicada. Com base nisso, foi elaborada uma base de dados histórica das grandezas selecionadas, seguida de tratamento de dados e aumento da base com a geração de dados sintéticos por meio de técnicas baseadas nos dados reais, sendo interpolação entre amostras e perturbação por meio de ruído Gaussiano. Em seguida, foram aplicados algoritmos de aprendizado de máquina para prever o teor de cal livre, avaliando o desempenho de cada um e visando a orientar ajustes proativos no processo. Na análise de desempenho, o modelo preditivo proposto obteve os índices de $R^2 = 0{,}966$, MSE = 0,02 e RMSE = 0,141, em comparação com $R^2 = 0{,}73$ e MSE = 0,1162 da referência bibliográfica mais simular à presente aplicação. Os resultados experimentais obtidos indicam que é possível a predição da cal livre, evidenciando a relevância do trabalho para a melhoria da eficiência energética da planta, redução de desperdícios, maior estabilidade na qualidade do clínquer e, consequentemente, do cimento produzido.

Palavras-chave: aprendizado de máquina; cal livre; cimento; clínquer; qualidade; dados sintéticos.

ABSTRACT

Cement production is a complex process that involves mining, grinding of raw materials and heating of materials in clinker kilns. Throughout this production chain, homogenization and chemical reactions occur that alter the structure of the compounds to obtain a final product that meets the quality standards required by regulations. In the clinkering stage, one of the crucial parameters to be monitored is the free lime content, which directly impacts the quality of the cement and the efficiency of the process. However, this analysis is punctual, usually performed every two hours, and includes collection, sample preparation and calibration steps in X-ray equipment. Therefore, this work aims to develop a predictive model for the analysis of free lime in clinker kilns, using machine learning techniques and synthetic data generation. A consultation with process experts and related research was carried out to define the variables with the greatest impact on the free lime value, their respective limits that characterize stability in the operation, as well as cases in which the sampling of the variables could be impaired. Based on this, a historical database of the selected quantities was created, followed by data processing and increase of the database with the generation of synthetic data through techniques based on real data, being interpolation between samples and disturbance through Gaussian noise. Then, machine learning algorithms were applied to predict the free lime content, evaluating the performance of each one and aiming to guide proactive adjustments in the process. In the performance analysis, the proposed predictive model obtained the indices of $R^2 = 0.966$, MSE = 0.02 and RMSE = 0.141, compared to $R^2 = 0.73$ and MSE = 0.1162 of the bibliographic reference most similar to the present application. The experimental results obtained indicate that it is possible to predict free lime, highlighting the relevance of the work for improving the energy efficiency of the plant, reducing waste, and providing greater stability in the quality of the clinker and, consequently, of the cement produced.

Keywords: machine learning; free lime; cement; clinker; quality; data augmentation.

INDICADORES DE IMPACTO

Este trabalho de mestrado teve como objetivo principal desenvolver um modelo preditivo para a análise de cal livre em fornos de clínquer, utilizando técnicas de aprendizado de máquina e geração de dados sintéticos. A produção de cimento é um processo complexo, e o monitoramento do teor de cal livre é crucial para a qualidade do cimento e a eficiência do processo. A análise tradicional de cal livre é pontual, realizada a cada duas horas, e envolve etapas de coleta, preparação da amostra e aferição em equipamentos de raio-X, um processo que leva cerca de 40 minutos na planta estudada. Este estudo propõe uma abordagem preditiva para otimizar a tomada de decisão dos operadores, garantindo maior padronização e regularidade nos parâmetros de qualidade e operação da planta. Foram consultados especialistas da área de processo, e foi construída uma base de dados histórica com as variáveis mais relevantes para o cálculo da cal livre. Essa base de dados foi expandida com a geração de dados sintéticos através de interpolação e perturbação por ruído gaussiano, e algoritmos de aprendizado de máquina foram aplicados para prever o teor de cal livre. Os resultados experimentais obtidos demonstraram a viabilidade da predição da cal livre, com índices de R^2 = 0,966, MSE = 0,02 e RMSE = 0,141. Este modelo preditivo tem um impacto tecnológico significativo, pois permite a otimização do processo de produção de cimento, com potencial para redução de desperdícios e maior estabilidade na qualidade do clínquer e, consequentemente, do cimento produzido. Do ponto de vista econômico, a melhoria da eficiência energética da planta e a redução de desperdícios se traduzem em menores custos de produção. O impacto social reside na diminuição da exposição a riscos de segurança dos analistas de laboratório durante a coleta de amostras. Em termos de extensão, este trabalho envolveu a participação de especialistas da área de processo, promovendo uma troca de conhecimento entre a academia e a indústria, e tem como público-alvo os operadores de plantas de cimento. O estudo também se enquadra na área temática de tecnologia e produção e está alinhado com o Objetivo de Desenvolvimento Sustentável (ODS) 9 - Indústria, Inovação e Infraestrutura, que busca construir infraestruturas resilientes, promover a industrialização inclusiva e sustentável, e fomentar a inovação. A implementação deste modelo em plantas de cimento tem potencial para otimizar processos, reduzir custos, melhorar a qualidade do produto e reduzir riscos no trabalho.

IMPACT INDICATORS

This master's thesis aimed to develop a predictive model for the analysis of free lime in clinker kilns, using machine learning techniques and synthetic data generation. Cement production is a complex process, and monitoring free lime content is crucial for cement quality and process efficiency. Traditional free lime analysis is done sporadically, typically every two hours, involving collection, sample preparation, and measurement on X-ray equipment, a process that takes about 40 minutes at the plant studied. This study proposes a predictive approach to optimize decision-making by operators, ensuring greater standardization and consistency in quality and plant operation parameters. Process area experts were consulted, and a historical database was built with the most relevant variables for calculating free lime. This database was expanded with the generation of synthetic data through interpolation and Gaussian noise perturbation, and machine learning algorithms were applied to predict the free lime content. The experimental results obtained demonstrated the feasibility of predicting free lime, with R^2 = 0.966, MSE = 0.02, and RMSE = 0.141. This predictive model has a significant technological impact, as it enables the optimization of the cement production process, with the potential to reduce waste and increase stability in the quality of clinker and, consequently, of the cement produced. From an economic point of view, improving the plant's energy efficiency and reducing waste translates into lower production costs. The social impact lies in the reduction of safety risks for laboratory analysts during sample collection. In terms of outreach, this work involved the participation of process area experts, promoting a knowledge exchange between academia and industry, and targets cement plant operators. The study also falls under the thematic area of technology and production and is aligned with Sustainable Development Goal (SDG) 9 - Industry, Innovation, and Infrastructure, which seeks to build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation. The implementation of this model in cement plants has the potential to optimize processes, reduce costs, improve product quality, and reduce workplace risks.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 Introdução

A palavra eficiência, conforme Priberam (2024), denota a "Qualidade de algo ou alguém que produz com o mínimo de erros ou de meios". Esse termo pode ser encontrado hoje nas mais diversas áreas ligadas a produtos e serviços. Desde as lâmpadas que saíram do modelo incandescente para LED a fim de produzir mais luminosidade com menor consumo de energia, carros que vêm fazendo cada vez mais quilômetros com um litro de gasolina, ou até plantações das mais diversas culturas que geram hoje muito mais toneladas por hectare do que há alguns anos atrás. Esse cenário também ocorre com as indústrias, que são forçadas a levar o maquinário ao limite de produção, em concomitância com recursos financeiros restritos, estoques reduzidos, dentre outros fatores limitantes que não podem afetar o produto final, em qualidade e volume. Com isso, todo o ecossistema relacionado ao ambiente industrial passou por modificações a fim de proporcionar a tão buscada eficiência. Nesse sentido, métricas foram e continuam sendo criadas e aprimoradas para dar números à eficiência de todos os processos das organizações.

Em uma indústria cimenteira, por exemplo, a eficiência para o departamento financeiro é encontrada calculando cada centavo gasto para produção de cimento, enquanto para a manutenção se traduz na disponibilidade dos equipamentos para a equipe de produção. Este time, por sua vez, tem seu rendimento medido pela quantidade de toneladas produzidas por hora, enquanto a logística é considerada mais eficiente quanto mais consegue colocar cimento no mercado. Ligado a toda essa dinâmica, a equipe de controle de qualidade precisa assegurar que todas as etapas de produção sigam critérios que assegurem que os produtos e subprodutos tenham suas características ótimas conforme normas internas e externas à empresa.

> O processo produtivo do cimento é, resumidamente, uma combinação de exploração e beneficiamento de substâncias minerais não metálicas, sua transformação química em clínquer (produto intermediário do cimento) em um forno a cerca de 1.450 ºC e posterior moagem e mistura a outros materiais, conforme o tipo de cimento. (CNI, 2017)

O clínquer figura como o principal componente do cimento e sua composição química, conforme Taylor (1997), é de 67% de CaO, 22% de SiO_2, 5% de Al_2O_3, 3% de Fe_2O_3 e 3% de produtos em concentrações menores. Em suma, combina-se variantes de cálcio, sílica, alumínio e ferro para a sua obtenção. Ao passar pelo forno de clínquer, essa mistura chega a cerca de 1450 ºC, sendo fundida e, então, resfriada rapidamente. O autor explica que nesse momento o clínquer pode ser observado quimicamente nas fases: Alita (50 - 70 %), belita (15 – 30 %), aluminato (5 – 10 %), ferrita (5 – 15 %) e alguns componentes em menor concentração, dentre

eles o CaO, limitado entre 3 e 5 %. Esse óxido de cálcio (CaO) é a parte que não reagiu com os outros componentes químicos do processo, e popularmente é denominada Cal livre.

Esse componente é altamente indesejado no processo, uma vez que pode reagir lentamente com a água e causar grande expansibilidade. Como o clínquer é o maior componente do cimento, o excesso de cal livre causaria a expansibilidade do cimento já solidificado, afetando diretamente seu propósito.

Para controlar os índices de cal livre, usualmente é responsabilidade da equipe de qualidade a coleta periódica e normalmente manual (em média de 2 em 2 horas) de uma amostra de clínquer que é preparada e analisada em aparelhos de raio X, processo que leva ao todo cerca de 40 minutos na planta que foi estudada no presente trabalho. Mediante o resultado, os operadores atuam no processo de forma a controlar os níveis de cal livre, modificando a alimentação do forno, injeção de combustível, dentre outras variáveis. Por se tratar de amostras pontuais, as atuações são realizadas sempre de maneira reativa, uma vez que o indicador reflete a qualidade de um clínquer que, no momento do resultado, já está estocado no silo. Além disso, existe o grande impacto na rotina dos analistas de laboratório, bem como a exposição a riscos de segurança nos momentos de coleta.

Por se tratar de uma variável química do processo, entende-se ser possível determinar a cal livre por meio da análise de dados já existentes, que são medidos a todo tempo como parâmetros de operação da planta. Porém, a química envolvida na produção de clínquer e cimento é extremamente complexa, o que torna basicamente impossível a análise manual. Por isso, em casos como este em que não é possível inferir informações por análises humanas, algoritmos de *machine learning* (ML) são utilizados. Estes algoritmos são alimentados com todos os dados referentes ao processo que se deseja analisar, configurados adequadamente por meio de seus parâmetros, gerando como saída uma estimativa da variável que se deseja controlar.

Ainda, quando não estão disponíveis informações em volume ou representatividade adequadas, é possível utilizar técnicas de geração de dados sintéticos, aumentando a base de dados sem prejudicar os resultados dos modelos. Com isso, a etapa de criação da base de dados se torna mais facilitada, alimentando os algoritmos de ML com um maior volume de informações para processamento. A geração de dados sintéticos figura, então, como técnica que melhora a robustez dos resultados sem dificultar a etapa de construção da base de dados.

Nesse sentido, muito pouco é encontrado de produções científicas relacionadas à predição em fábricas de cimento. Do levantamento bibliográfico realizado, dois estudos foram loca-

lizados realizando a predição da cal livre por meio de algoritmos de aprendizado de máquina. O primeiro, de Li et al. (2015), obteve um algoritmo robusto ao utilizar dados de variáveis de processo associados à imagens de uma câmera que visualizava a chama do forno, mas acabou tendo problemas na precisão dos resultados à medida que a lente da câmera sujava ou não era possível obter uma boa imagem devido a variações inerentes de processo. Já Magalhães (2019) buscou realizar a predição somente com dados de processo, contribuindo na escolha de variáveis representativas para a base de dados (BD), bem como cenários de coleta de valores defasados no tempo em razão de se tratar de um processo contínuo. Como cada processo de produção de cimento é diferente dos outros, ainda que o trabalho da autora obteve resultados expressivos, é preciso realizar uma análise detalhada de cada linha de produção.

Dos demais trabalhos avaliados, com menor semelhança com a presente proposta, também foi possível aprender sobre um contexto geral dos dados aplicados aos algoritmos de *Machine Learning*, entendendo o impacto de dados lançados incorretamente ou disponíveis em intervalos longos entre amostras, que prejudicam fatalmente a qualidade dos resultados de aplicação de modelos de ML.

Portanto, este estudo dispôs-se a realizar a aplicação de algoritmos de aprendizado de máquina capazes de predizer o valor da cal livre baseado em coletas de variáveis de processo que são medidas para operação de um forno de clínquer.

1.1 Objetivos

O objetivo principal do presente trabalho foi propor um modelo para prever o comportamento da cal livre, baseados em algoritmos de aprendizado de máquina, para auxiliar a tomada proativa de decisão dos operadores do processo, bem como garantindo maior padronização e regularidade nos parâmetros de qualidade e operação da planta. Para isso, o estudo abrange as etapas de consulta de especialistas da área de processo, obtenção de bases de dados reais referentes ao processo produtivo de clínquer em uma fábrica de cimento, expansão dos dados coletados por meio da geração de dados sintéticos, desenvolvimento de modelos com aprendizado de máquina, avaliação de eficiência e posterior validação dos modelos obtidos.

Com a dinâmica acima descrita, objetivou-se contribuir com a construção de uma ferramenta que atue de maneira preditiva, de forma a nortear o trabalho dos operadores para atuação precisa no equipamento responsável pela alteração dos índices da variável monitorada.

1.1.1 Objetivos específicos

Os objetivos específicos deste trabalho de mestrado são:

a) entender quais variáveis são diretamente ligadas ao comportamento da cal livre por meio da consulta a especialistas da área de processo;
b) construir uma base de dados do processo de um forno real de clínquer;
c) incrementar a base de dados por meio de dados sintéticos;
d) desenvolver algoritmo para predição da cal livre por meio da análise da base de dados anteriormente construída;
e) aferir o grau de adesão dos resultados preditos com os valores reais para comparação da eficiência dos modelos;
f) avaliar os resultados obtidos.

1.2 Organização do trabalho

Este documento foi particionado para melhor entendimento do trabalho: o referencial teórico, apresentado no Capítulo 2, avalia referências nacionais e internacionais dos temas correlatos a produção de cimento e clínquer, algoritmos de aprendizado de máquina e métricas de avaliação de resultados de modelos de predição. O Capítulo 3 pondera o que já foi realizado no sentido de avaliação de predições e aplicações de aprendizado de máquina dentro e fora da indústria cimenteira, enquanto o Capítulo 4 apresenta como foi operacionalizada a execução do presente trabalho. No Capítulo 5, avaliam-se os resultados obtidos com a aplicação dos algoritmos para predição da cal livre. Por fim, no Capítulo 6, são elucidados os resultados e possibilidades de trabalhos futuros baseados no que foi aqui desenvolvido.

2 Referencial teórico

Este capítulo tem por objetivo discorrer sobre todos os temas que permeiam a execução do trabalho de forma a pavimentar o caminho necessário ao estudo da aplicação de algoritmos de aprendizado de máquina em uma planta de produção de clínquer e cimento.

2.1 Produção de cimento

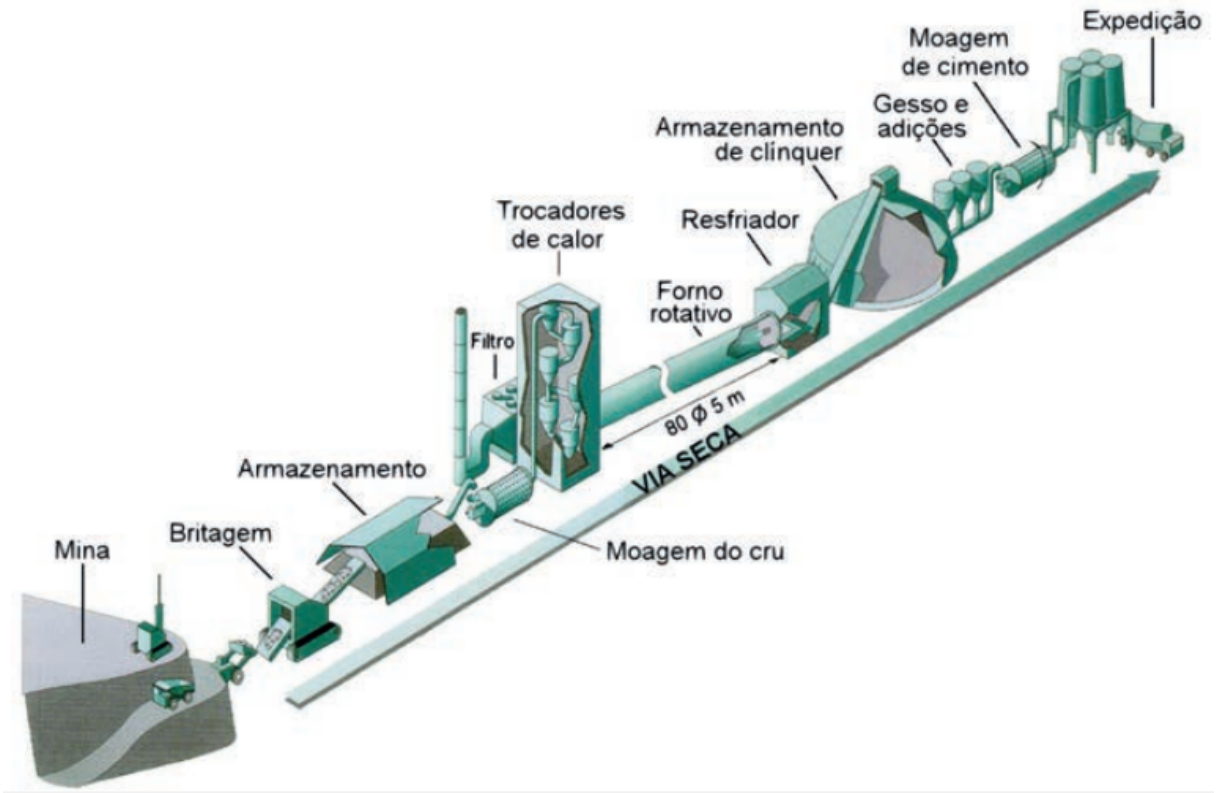
Para iniciar os estudos de todas as áreas de conhecimento a serem aplicadas no presente trabalho, cujo cenário é uma linha de produção de cimento, cabe conceituar inicialmente o processo de fabricação do cimento e um pouco da química que o envolve. Ilustrado na figura 2.1, sua etapa inicial baseia-se na extração de calcário e argila, que correspondem, respectivamente, a 75-80% e 20 a 25% da proporção para fabricação do clínquer (CNI, 2017). Estes materiais são obtidos em minas próximas ao local da fábrica. O calcário britado e a argila são armazenados em pátios ou galpões de forma que o empilhamento e retomada dessas matérias-primas atuem como um primeiro processo de homogeneização.

Na moagem de cru (ou farinha), o calcário e argila são misturados com outros materiais como o minério de ferro a fim de atender a composição química ideal do clínquer, demonstrada por Taylor (1997), de 67% de CaO, 22% de SiO_2, 5% de Al_2O_3, 3% de Fe_2O_3 e 3% de produtos em concentrações menores. Essa mistura é enviada a um moinho e ganha o nome de farinha ou cru. Este material, estocado em silo, é também homogeneizado no processo para seguir para a próxima etapa da fabricação.

Inicia-se então a etapa da clinquerização. A farinha é enviada para a torre de pré calcinação, onde começa a troca de calor e consequente descarbonatação e reações químicas da mistura. A torre culmina em um forno rotativo onde o material chega à fase líquida, devido à temperatura que chega a cerca de 1450 ºC.

Após passar pelo forno, o material é rapidamente resfriado, gerando o clínquer, a base do cimento. Esse material já é um aglomerante hidráulico como o cimento. Porém, em contato com água, endurece muito rapidamente, o que impossibilitaria o trabalho com o material. Trata-se, portanto, de um produto intermediário na fabricação do cimento, sendo então armazenado. A Figura 2.1 ilustra o fluxo de produção do cimento.

Figura 2.1 – Fluxograma de produção de cimento.



Fonte: Caillon Rouge/Roger Rivet apud CNI (2017).

Para dar ao clínquer a trabalhabilidade necessária para não solidificar durante o manuseio, ele é misturado com gesso e outras adições, que dependem do tipo de cimento, normatizado pela NBR 16697 ((ABCP), 2018). Dosados em quantidades conforme as receitas de norma, clínquer, gesso e adições são enviadas para um moinho, e o produto que sai dele é, enfim, o cimento, que é armazenado em silos para posterior venda como ensacado, granel ou *big bag*.

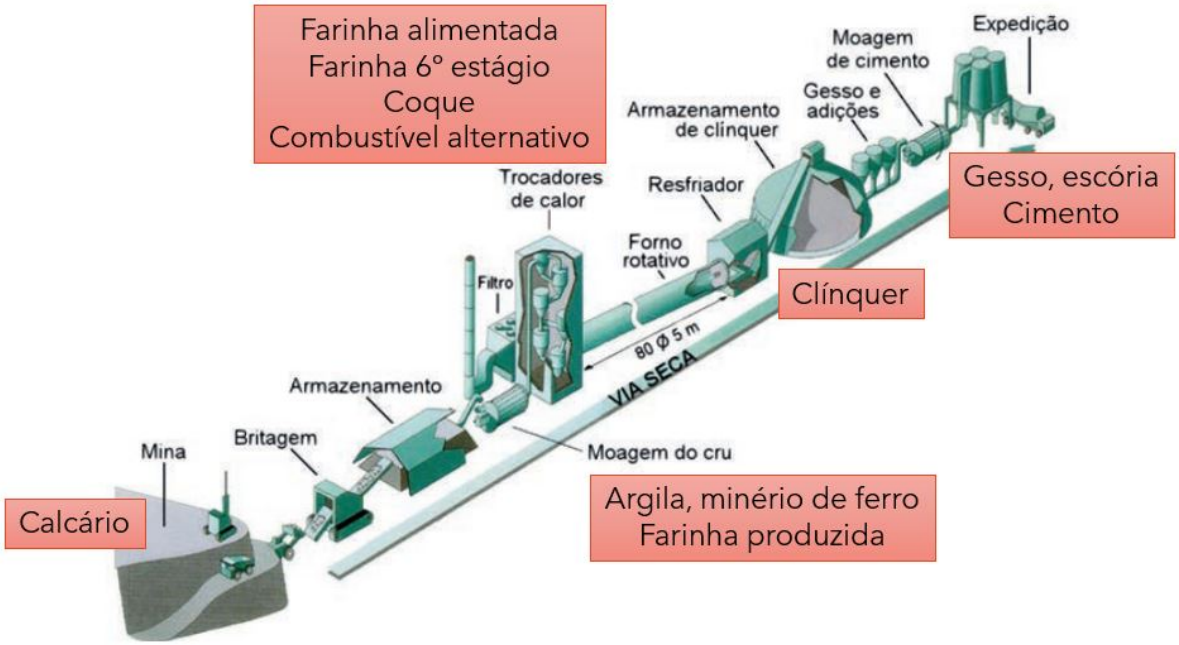
2.1.1 Controle de qualidade de clínquer e cimento

Conforme a (ABCP) (2018), a NBR 16697 rege quanto ao percentual permitido de cada material na fabricação de cimento, bem como parâmetros físico-químicos que os cimentos devem cumprir. Conforme as receitas definidas pela norma, o clínquer anteriormente produzido é misturado ao gesso e outras matérias-primas como calcário e escória, a depender do tipo de cimento, e essa mistura passa por moinhos para, enfim, ser chamada de cimento, que é estocado para posterior venda.

Para manter a composição química ideal, a equipe de qualidade realiza análises físico-químicas em várias fases do processo. Alguns exemplos de elementos analisados são os teores

de cálcio e enxofre do calcário britado na mina, saturação, módulo de sílica e alumínio da farinha que sai do moinho e também da alimentada ao forno, análise na farinha em pontos intermediários da torre de pré-calcinação, do pó oriundo do sistema de filtragem dos gases da torre, e também do clínquer que sai do resfriador em direção ao silo. Por fim, o cimento produzido também é analisado para manter as características exigidas por norma. A Figura 2.2 indica os principais materiais analisados durante o fluxo de produção de cimento.

Figura 2.2 – Fluxograma de análises de qualidade na fabricação de cimento.



Fonte: Adaptado de Caillon Rouge/Roger Rivet apud CNI (2017).

Tamanha é a importância dessas análises que é possível encontrar no mercado soluções para automação de alguns desses processos de análise. Um exemplo é o sistema QCX (FLS-MIDTH, 2024), que fornece processos automatizados de controle e garantia da qualidade em plantas de cimento. Este sistema, além de gerenciar cada análise do laboratório, pode possuir estações automáticas ao longo da planta para coletar e enviar amostras periodicamente ao laboratório, podendo inclusive analisar de maneira autônoma e corrigir as dosagens de matéria-prima ao longo da cadeia produtiva. Tudo isso para garantir maior linearidade e padronização da produção. Embora sejam soluções amplamente conhecidas entre as cimenteiras, estes produtos são extremamente caros, o que limita o acesso às tecnologias de controle automático de qualidade. Por isso o presente trabalho busca avaliar a aplicação de técnicas de análise dos dados sem necessidade de aquisição de novas ferramentas de *hardware* ou *software*, que via de regra são caras, figurando como uma possibilidade de geração de uma ferramenta de baixo custo quando comparado às ferramentas de automação do mercado.

2.2 A cal livre do clínquer

O presente trabalho dispõe-se a estudar a análise de Cal livre do clínquer produzido no forno, uma das várias atividades realizadas para assegurar a qualidade do cimento ao fim do processo. Taylor (1997), Lea (1970) e Hewlett (2019) mostram que o calcário aquecido na torre de pré-calcinação e no forno transforma-se em carbonato de cálcio, óxido de cálcio e dióxido de carbono. Estes materiais devem reagir com os outros presentes no processo, com as variantes de alumínio, sílica e ferro, para formar as fases principais do clínquer: Alita, belita, aluminato e ferrita. Nem todo óxido de cálcio (CaO) reage, e a essa quantidade remanescente é dado o nome de CaO (ou cal) livre. Lea (1970) ainda indica que os níveis elevados de Cal livre podem indicar temperaturas inadequadas no processo, tempo de residência curto do clínquer no forno ou até composição inadequada da farinha. Para melhor ilustrar essas reações químicas, a Figura 2.3 mostra os componentes gerados a partir do aumento da temperatura no processo.

Figura 2.3 – Reações químicas da produção de clínquer.

Acima de 1073 K (800 °C)	Início do processo de fabricação de CaO
Entre de 1073 K (800 °C) e 1473 K (1200 °C)	Formação do C_2S ($2CaOSiO_2$)
Entre 1368 K (1095 °C) e 1478K (1205 °C)	Formação do C_3A ($3CaOAl_2O_3$) e C_4AF ($4CaOAl_2O_3Fe_2O_3$)
Entre 1533 K (1260 °C) e 1728 K (1455 °C)	Formação do C_3S a partir do C_2S com quase extinção da cal livre (CaO)
Entre 1728 K (1455 °C) e 1573 K (1300 °C)	Cristalização da fase líquida do C_3A e do C_4AF.

Fonte: Souza et al. (2015).

Independente da origem da perturbação dos seus índices, o impacto da cal livre elevada está diretamente relacionado às propriedades mecânicas do cimento, uma vez que pode ocasionar expansões indesejadas e consequentes fissurações no cimento já solidificado. Dessa forma, a mensuração regular da cal livre norteia a atuação dos operadores do processo de forma a otimizar injeções de combustível, receitas da farinha produzida, rotação e alimentação do forno em prol da estabilidade dessa variável.

2.3 Coprocessamento de resíduos em fornos de clínquer

O coprocessamento (ou coincineração) de resíduos figura como uma rota tecnológica que visa tratar refugos e efluentes dos mais variados processos por meio da destruição térmica em fornos de clínquer.

> "A alta temperatura e o tempo de residência dos gases no forno, além dos robustos sistemas de tratamento de emissões atmosféricas necessários para o licenciamento desse tipo de unidade industrial, possibilitam a destruição de compostos perigosos presentes nos resíduos e níveis adequados de emissões atmosféricas."(BEER, 2017; BOURTSALAS, 2019) apud (TORRES; LANGE, 2022)

Regulamentado no Brasil em 1999 (TORRES; LANGE, 2022), o coprocessamento é sustentado por dois viéses na indústria cimenteira. Do ponto de vista ambiental, o processo de incineração de resíduos, em substituição parcial de combustíveis fósseis não renováveis, reduz a pegada de carbono das organizações, podendo inclusive ter ganhos convertidos em créditos de carbono para serem vendidos no mercado. Já do ponto de vista econômico, embora usados em maior quantidade proporcional do que o combustível principal, os resíduos custam muito menos às empresas, o que torna vantajoso todo este processo para as companhias que geram refugos que eram passivos ambientais, como para as cimenteiras que convertem o poder calorífico desses materiais para economizar com combustível.

Porém, vários são os desafios que envolvem a incineração desses materiais. A utilização de resíduos como matéria prima ou diretamente como combustível na produção de cimento exige o ajuste das receitas de produção de forma a compatibilizar os parâmetros de qualidade aos novos componentes adicionados. Dessa forma, estes novos materiais precisam ter critérios mínimos de qualidade, bem como disponibilidade constante.

Torres e Lange (2022) avaliaram metodologias de processamento de resíduo sólido urbano justamente com objetivo de melhorar suas características de qualidade de queima. Já Mazur, Demito e Watanabe (2024) examinaram o poder calorífico de um óleo rejeitado pelo processo de fabricação de latas, principalmente aferindo o impacto da exposição desse resíduo à umidade e deixando claro o impacto negativo desse processo na queima do material no forno de clínquer. Por fim, Rocha, Lins e Santo (2011) realizaram uma pesquisa bibliográfica abrangente sobre tipos de resíduos utilizados, impactos econômicos, sociais e ambientais do coprocessamento.

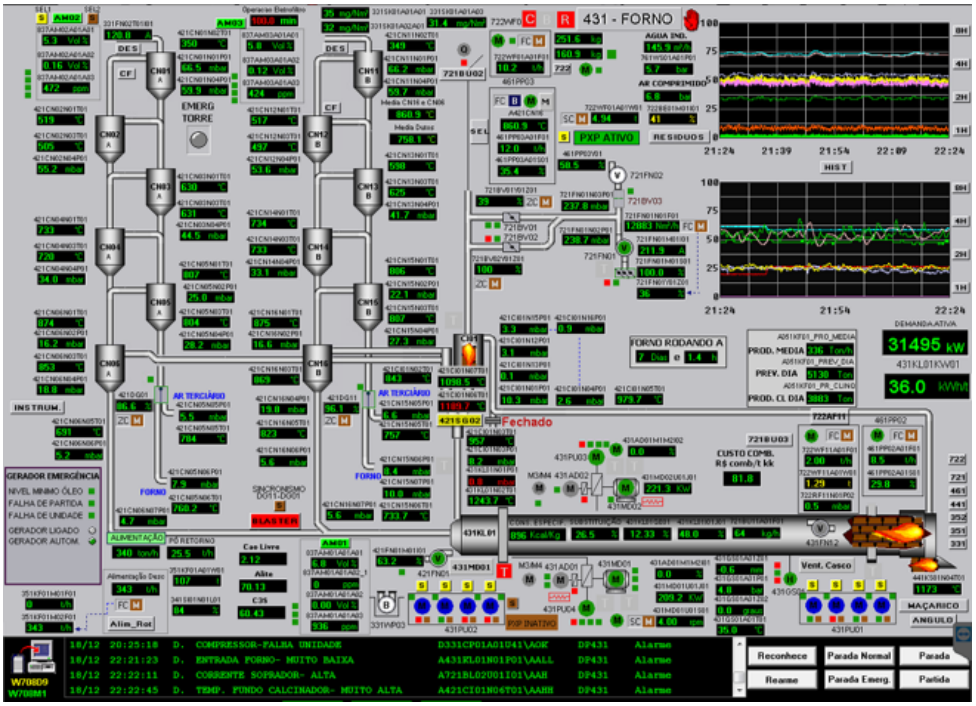
Do trabalho de Rocha, Lins e Santo (2011) verificou-se a necessidade da construção deste tópico individual sobre o coprocessamento, pois em um dos referencias avaliados ponderou-se sobre o impacto do uso de resíduos no teor de cal livre, uma vez que todas as reações químicas são afetadas pela queima de combustíveis alternativos. Portanto, espera-se que seja possível verificar a existência desse impacto durante a execução do presente estudo.

2.4 Sistemas de monitoramento industrial

Os sistemas de monitoramento industrial são fundamentais para garantir uma operação eficiente e padronizada de cadeias produtivas, desde processos manuais aos 100% automatizados. Esse processo envolve a coleta de dados em campo por meio de sensores, transdutores e outro dispositivos, eventuais interpretações feitas por lógicas de programação e também atuações manuais de operadores, para culminar com a produção dos mais variados tipos de subprodutos e produtos.

Santos (2014) aborda, em seu livro, a supervisão de processos industriais. Nesse sentido, ele conceitua o termo SCADA (*supervisory control and data aquisition*) como uma "representação virtual do processo ou planta física a controlar que possui uma representação em tempo real das principais variáveis monitoradas ou controladas". Ainda, este sistema pode conter funções de supervisão, operação e controle.

Figura 2.4 – Sistema SCADA de forno de clínquer.



Fonte: Do autor (2024).

Santos (2014) ainda comenta que estes sistemas, em sua forma primitiva, eram basicamente uma telemetria de condições ocorridas em campo, e hoje se tornaram uma ferramenta para "automatizar a monitoração e controle de processos industriais".

> Os sistemas SCADA podem também verificar condições de alarme, identificadas quando o valor do *tagname* ultrapassa uma faixa ou condição preestabelecida, sendo possível programar a gravação de registros em bancos de dados,

> ativação de som, mensagem, mudança de cores, envio de mensagens, *e-mail*, celular etc. (SANTOS, 2014)

Ainda, é no SCADA que são monitoradas em tempo real todas as variáveis analógicas do processo produtivo, sendo possível também a montagem de gráficos, etc. Para as plantas de cimento, atualmente todas são operadas via sistemas SCADA, de maneira a garantir operação contínua, com qualidade e otimização de todo o processo.

2.5 Armazenamento de dados históricos de processo

Além do sistema supervisório (SCADA), muitas empresas buscam também maneiras de guardar o comportamento das variáveis de processo ao longo do tempo de forma a possibilitar análises posteriores com base em histórico. Uma das ferramentas mais comuns nesse ambiente industrial é o PI System (AVEVA (2024)). Trata-se de um historiador de variáveis capaz de armazenar o histórico de variáveis analógicas, digitais, *status* de equipamentos, dentre outras informações de processo.

Figura 2.5 – *PI System.*

Fonte: AVEVA (2024).

Aliado ao armazenamento das variáveis, o *PI System* possui ferramentas que possibilitam a consulta dos dados armazenados por meio de gráficos e planilhas, sendo possível adicionar fórmulas de filtro, médias, dentre outros cálculos. Para a planta em que o presente trabalho foi realizado, o sistema possui mais de 10 anos de dados históricos armazenados, e a base de dados será montada em excel por meio da ferramenta acessória *PI Data Link*.

2.6 Big data

Como exposto acima, ao considerar que estão sendo armazenados dados históricos da operação da planta por mais de 10 anos, é preciso conceituar também o termo *big data*. Para isso, Gandomi e Haider (2015) esclareceram os conceitos que cercam este tema, bem como explicaram de maneira a facilitar o entendimento das vantagens que podem ser obtidas com o *big data*. Para melhor entendimento, Gandomi e Haider (2015) mostraram os dados classificam-se pelo seu volume, ou quantidade, pela variedade de informações, pela velocidade em que são geradas e, por fim, pela veracidade dos dados.

No cenário das fábricas de cimento, atualmente encontram-se plantas com alto nível de instrumentação, para que seja possível o controle cada vez mais automático do processo. Por isso, são medidas temperaturas, pressões, velocidades e outras grandezas em todas as etapas da cadeia produtiva. Com o alto nível de instrumentação associado às ferramentas de armazenamento de dados históricos, entende-se que já estão disponíveis dados suficientes para uma análise detalhada em busca de indicadores proativos do processo. Porém, por se tratar de uma grande quantidade de dados, armazenados por muito tempo, compreende-se a existência do contexto de *big data* aplicado ao cenário de produção de cimento, o que exige que sejam utilizadas técnicas de processamento automático de dados.

A esse montante de registros são aplicadas técnicas de processamento e análise de forma a transformar dados em informação útil para tomadas de decisão, etc. Nesse sentido, a inteligência artificial é uma das ferramentas mais utilizadas perante a necessidade de processamento de um grande volume de informações.

2.7 Inteligência artificial

A inteligência artificial (IA) aparece como solução capaz de processar um grande volume de dados e proporcionar conclusões e previsões baseadas no que foi coletado. Para isso, são empregados vários tipos de algoritmos conforme a base de dados disponível. Russel e Norvig (2020) conceitua a IA como um "campo universal", um universo de mentes brilhantes virtuais com objetivo de entender e construir entidades inteligentes na sua aplicação.

A IA vem sendo entendida como uma das grandes responsáveis pela inovação tecnológica atualmente. Isso por conseguir processar grandes volumes de dados independente de qual seja o assunto e gerando novas conclusões para temas que outrora poderiam estar saturados.

Antes de abordar assuntos mais complexos relacionados ao tema, há de se considerar uma obra de mais de 70 anos atrás. Em 1950, Turing (1950) iniciou seu artigo perguntando se as máquinas podiam pensar, o que alguns estudiosos consideram como o pontapé inicial da inteligência artificial. Várias questões chamam atenção em seu estudo, e a primeira é que ele foi publicado na revista Mind, que se intitulava como uma revisão trimestral em psicologia e filosofia. Ora, um tema de tal complexidade ainda no meio do século XX era de fato considerado como filosófico, uma vez que as máquinas digitais ainda estavam em incipiente desenvolvimento e não chegam nem perto do nível de processamento de um simples relógio dos dias atuais. Ainda nesse sentido filosófico da questão, Turing (1950) questiona logo no início o conceito consolidado para as palavras "pensar" e "máquina". Dessa forma, ele prefere modificar sua pergunta fundamental para uma questão se as máquinas poderiam imitar o comportamento humano de forma indistinguível.

O autor propõe então um teste que ficou conhecido como Teste de Turing, e, ainda em 1950, Turing (1950) conclui sua argumentação dizendo que, com o avanço tecnológico, recursos de memória, processamento e dados suficientes, as máquinas poderiam sim imitar o comportamento humano de tal maneira a não ser possível distinguir o pensamento humano do "pensamento" das máquinas.

Portanto, a inteligência artificial envolve o desenvolvimento de algoritmos capazes de aprender com dados de qualquer natureza (medicina, engenharia, química, etc.) e executar análises que antes requeriam a inteligência humana. Sua eficiência vem crescendo junto com o avanço da computação e hoje pode ser encontrada em diversas áreas como o aprendizado de máquina (*Machine Learning - ML*) e redes neurais.

No presente trabalho, a inteligência artificial encontra aplicação para interpretação de dados históricos do processo de produção de cimento sem a necessidade de modelar o sistema, que por si só é complexo e dinâmico. Há de se entender que o volume de dados já armazenados torna basicamente impossível a análise manual em busca de padrões de comportamento que norteiem a atuação dos operadores, mesmo que a avaliação fosse realizada por especialistas de processo que tenham pleno conhecimento da química envolvida na produção de cimento. Sob este olhar, a IA permite uma análise empírica, ou seja, não é preciso conhecimento avançado de química, mas sim de análise de dados, entendimento que pode ser aplicado a várias áreas de conhecimento e diversos segmentos da indústria.

2.8 *Machine learning*

Um dos grandes campos da inteligência artificial é o aprendizado de máquina (*machine learning*). Basicamente são algoritmos que permitem com que as máquinas aprendam através dos dados que estão coletando de forma a desempenhar uma tarefa cada vez melhor ao longo do tempo.

O aprendizado de máquina parte do pressuposto de que um sistema computacional pode tomar decisões e fazer previsões de acordo com os dados que foram utilizados no seu treinamento, ou seja, a máquina toma decisões com base no que experimentou anteriormente, e ainda melhora seu desempenho quando é apresentada a mais dados (JORDAN; MITCHELL, 2015). Taurion (2023) trata com muita cautela a aplicação de algoritmos de IA em qualquer cenário, pontuando, inclusive, que se "Implemente governança de dados. A velha máxima, entra lixo, sai lixo, continua valendo". Um bom exemplo dado inclusive em palestras pelo autor é que se um algoritmo for treinado para identificar vacas utilizando milhares de fotos delas em pastos, pode ser que quando se deparar com uma foto de uma vaca em um outro ambiente, reportará a negativa da presença do animal na foto, pois na verdade o algoritmo pode ter notado a grande presença da cor verde em seu treinamento, e não as vacas.

Já quanto à natureza dos dados, modelos que utilizam dados discretos são conhecidos como modelos de classificação, enquanto os que usam dados dinâmicos são nomeados como regressão.

2.8.1 Tipos de aprendizado

A forma de aprendizado dos algoritmos de ML é muito importante para ser estudada. Basicamente este aprendizado pode ser supervisionado, não supervisionado ou realizado por reforço, podendo ainda serem encontrados modelos que compartilham características entre os 3 grandes grupos de aprendizado..

Nesse contexto, por exemplo, os modelos de aprendizado supervisionado recebem um conjunto de dados rotulados (ou nomeados) e que contenham a resposta desejada para realizar o treinamento. (LUDERMIR, 2021).

Já no aprendizado semi-supervisionado, existe uma pequena base de dados completa e rotulada e outra sem rótulos que, ao serem interpretadas juntas, obtêm bons resultados de predição (SANCHES, 2003).

Por outro lado, no estudo de Santos e Rossi (2020) o autor utiliza metodologias de aprendizado não supervisionado, agrupando textos em grupos de semelhança. Como os dados disponíveis para aprendizado não supervisionado não têm rótulos, a alternativa é utilizar modelos que façam a análise e separação das informações por grupos de semelhança, usando para isso bibliotecas e algoritmos, como por exemplo o "k-means".

Por fim, existe também o aprendizado por reforço. Basicamente o modelo vai recebendo recompensas positivas para ganhos no processo e negativas para perdas, enquanto sua busca é sempre por maximizar as recompensas. Dessa forma, com base nos dados de recompensa que o próprio algoritmo vai experimentando, acaba ocorrendo o ajuste direcionado às escolhas que trazem melhor resultado. Para isso, pode ser aplicado o algoritmo *Q-learning*, tendo também citado o SARSA como opção de método.(ZANETTI; HARMENDANI, 2020)

Sob a perspectiva do aprendizado, depreende-se que ao utilizar as bases de dados históricos de processo de produção de cimento, tem-se pleno conhecimento das variáveis aplicadas. Por isso, utiliza-se o aprendizado supervisionado, fornecendo aos modelos uma base de dados com entrada e saída rotulados com objetivo de se realizar a predição da variável cal livre.

2.8.2 Técnicas de aplicação de ML

Os modelos de ML buscam obter uma tradução a partir dos dados de entrada do modelo para se obter uma estimativa do valor de saída, sendo que essa deve ser mais próxima possível do valor real. Para isso, cada tipo de técnica tem uma metodologia em específico. De acordo com a natureza dos dados, são aplicadas técnicas de regressão ou classificação, que são escolhidas, respectivamente, se a variável alvo é contínua ou categórica.

Na regressão linear, podem ser utilizados modelos com uma (regressão simples) ou mais (regressão múltipla) variáveis de entrada para prever a variável de saída. Obtem-se uma equação que atribui coeficientes a cada uma das variáveis de entrada X1, X2, ..., Xn de forma a se calcular a variável Y (CARVALHO, 2023). A equação é dada por:

$$Y_i = a + b_1X_1 + b_2X_2 + ... + b_nX_n + e \tag{2.1}$$

Na fórmula acima, os coeficientes X e Y são os dados de entrada e saída da base, respectivamente, enquanto a e b são fatores calculados pelo modelo de regressão e "e"é o erro para a observação.

Já para a regressão polinomial, como explicam Hwang, Lee e Lee (2021), a relação entre as variáveis de entrada e a variável de saída é calculada por meio de expoentes de potência N. Nesse caso, a relação entre entrada e saída não é linear, por isso a necessidade da exponenciação. Para isso, as relações podem ser encontradas conforme a fórmula abaixo:

$$Yi = a + b_1 X_1 + b_2 X_2^2 + ... + b_n X_n^n \tag{2.2}$$

A ordem das potências varia de acordo com a complexidade e relação dos modelos. O significado de cada variável é o mesmo da equação de regressão linear.

O presente trabalho trata de dados que são contínuos, como será melhor detalhado no capítulo 4, mas é preciso também citar que o aprendizado de máquina encontra aplicações também quando os dados disponíveis são categóricos. Nesse caso de quando os dados não são contínuos, é preciso usar metodologias de classificação. Alguns exemplos de aplicação são encontrados na separação de *e-mails* entre *spam* ou não, por exemplo. Além disso, podem ser utilizadas para reconhecimento de padrões, a exemplo do que fizeram Santos e Rossi (2020). Para estas metodologias ainda se aplicam os conceitos de bases de dados rotulados ou não, conforme já foi descrito acima.

A escolha do algoritmo de classificação ideal depende do problema a ser tratado. Por exemplo, casos de classificação binária costumam ter tratamento eficaz por algoritmos de regressão logística (CESSIE; HOUWELINGEN, 1992) enquanto máquinas de vetores de suporte reconhecem padrões por meio da busca dos melhores limiares para separação de classes (CORTES, 1995).

2.8.3 Algoritmos de ML

Para aplicar o ML, vários algoritmos podem ser utilizados a depender de características da base de dados, do objetivo do treinamento, etc.

Uma das opções é realizar a parametrização de uma árvore de decisão para processamento dos dados. Em adição à essa metodologia, Breiman (2001) publicou em seu trabalho uma abordagem que agrupava várias arvores de decisão no que chamou de *random forest*. Com esse conjunto de árvores, o modelo passava a ser mais robusto, melhorando sua generalização ao mesmo tempo que prevenia o *overfitting*.

Já o algoritmo de Máquina de vetores de suporte (SVM - *Support Vector Machine*) busca por um hiperplano que melhor separa as classes de um problema de classificação, ou por uma reta que melhor se aproxima dos pontos de dados em problemas de regressão. Essa metodologia foi melhor entendida pelo estudo do trabalho de Negreiros et al. (2020), que utilizaram a ferramenta na predição de resistências de compressão de cimento. Um modelo de SVM é capaz de capturar relações complexas entre os dados, mas ficou também claro o impacto de incoerências das bases de dados utilizadas.

Ainda, figuram os algoritmos de *boosting*, cuja intenção é reunir múltiplos modelos com suas qualidades e fraquezas de forma a resultar em uma predição mais robusta (FRIEDMAN, 2001). Dentre estes modelos é possível citar o *Gradient Boosting* e *LightGBM*.

No *Gradient Boosting*, as iterações, ao invés de realizarem o ajuste dos dados diretamente, buscam minimizar os erros de estimação. Já o modelo *LightGBM* é uma evolução do anterior que busca ser mais eficiente, demandar menos recursos computacionais sem penalizar os resultados do modelo. Ke et al. (2017) propõem no seu trabalho que não é necessário avaliar todas as instâncias de dados como o *Gradient Boosting*, bastando identificar quais são as instâncias com maior variação (e normalmente com maior impacto na variável alvo) e avaliar somente elas. Os autores obtiveram experimentalmente que esta abordagem aumenta em até 20 vezes a velocidade de treinamento sem penalizar índices de acurácia dos resultados.

Por fim, além dos algoritmos responsáveis pelas avaliações dos dados, são amplamente aplicados os algoritmos de validação. Eles são responsáveis por avaliar com mais critério e de maneira menos enviesada o erro de generalização do modelo (ARLOT; CELISSE, 2010). Ao passo de que a elaboração de qualquer modelo de ML passa por etapas de treino e teste, depreende-se que o modelo será mais aderente aos dados que foram fornecidos no treino do que no teste. Ao utilizar um algoritmo de validação, a intenção é variar essas duas divisões de forma que o modelo se torne mais abrangente ao universo de dados disponível globalmente.

Dentre os métodos de validação, a abordagem mais simples, conhecida como *Holdout*, consiste em separar a base de dados em duas frações, uma para treino e outra para teste, em proporção que costuma ficar entre 60/4% até 70/30%, respectivamente, para treino e teste, Ainda estão disponíveis abordagens similares, já otimizadas, como o que foi realizado por Azulay et al. (2020). Em suma, o modelo obtido é treinado com a base de treino, enquanto a base de teste serve para mensurar as métricas de qualidade dos resultados, como o R^2, MSE e RMSE. Já entre metodologias mais robustas, uma das mais populares é a *k-folds*, que foi citada por Arlot

e Celisse (2010), e que subdivide a BD em várias folhas e realiza várias iterações, cada uma com uma folha usada para teste enquanto as demais são usadas para treino. Embora o esforço computacional seja muito maior (uma vez que são realizados múltiplos treinamentos), é possível otimizar os resultados globais do modelo uma vez que são aplicados múltiplos cenários de treinamento e teste, quando são medidas as métricas de qualidade dos resultados.

2.9 Redes neurais

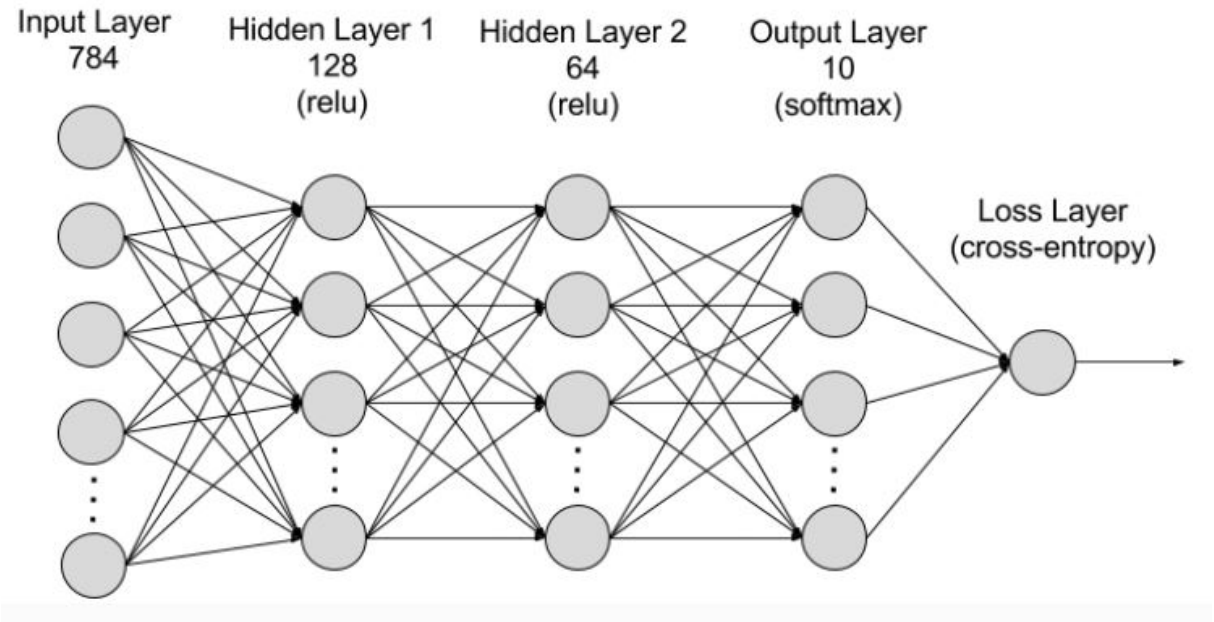
As redes neurais artificiais (*Neural Networks* - NN) são modelos criados tendo como inspiração as conexões existentes dentro do cérebro humano, de forma a processar dados e tirar conclusões. Sua estrutura concatena camadas de neurônios de forma a processar as bases de dados fornecidas.

As aplicações baseadas em redes neurais são "algoritmos que otimizam o tempo, solucionam problemas e realizam tarefas com maior precisão do que as realizadas por seres humanos, além de criar uma interação comunicativa entre computadores contribuindo para o avanço da indústria 4.0" (TURKIEWICZ; FRACAROLLI, 2019). Nesse levantamento bibliográfico os autores mostraram um pouco do universo de possibilidades de aplicações das NN bem como a baixa quantidade de artigos da sua aplicação em ambientes industriais. Por fim, a exemplo do que foi visto nos tópicos mais abrangentes de inteligência artificial, mostrou-se ser criticamente importante a qualidade da base de dados dos processos industriais com os quais se deseja aplicar este tipo de algoritmo.

2.9.1 Estrutura

A estrutura das redes neurais consiste basicamente em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. A camada de entrada é responsável pelo primeiro processamento dos dados fornecidos à rede neural. Após esse processamento, que pode contemplar inclusive a categorização, ela envia para a próxima camada. Já o bloco de camadas ocultas é responsável por receber os dados da camada anterior, processá-los mais um pouco, e encaminhá-los à próxima camada oculta ou à camada de saída. Por fim, a camada de saída é responsável por fornecer o resultado do processamento realizado pela rede neural. (Amazon Web Services, 2023)

Figura 2.6 – Estrutura redes neurais.



Fonte: Amazon Web Services (2023).

2.9.2 Funções de ativação

Segundo Fleck et al. (2016), que cita Haykin (2001), diversos tipos de funções de ativação são aplicadas nas camadas de uma rede neural. Essas funções tem por finalidade permitir que as redes neurais consigam capturar as relações entre as variáveis das bases de dados, mesmo que sejam não lineares, bem como para definir um formato adequado da saída do modelo, a depender das necessidades de cada caso. Algumas das funções estão dispostas abaixo:

a) limiar: Restringe a saída para níveis binários (0 ou 1), sendo 0 quando o resultado da camada é negativo e 1 quando é positivo;
b) linear por partes: Quando a saída é linear ao menos em um intervalo de dados;
c) sigmoidal: função crescente com balanceamento adequado entre um comportamento linear ou não linear, sendo dado por:

$$f(\mu) = \frac{1}{1+\exp(-a\mu)} \tag{2.3}$$

d) tangente hiperbólica: Diferencia-se da função sigmoidal pois esta tem resultados entre 0 e 1 enquanto a tangente varia, inclusive, entre valores positivos e negativos. Portanto, sua notação matemática é:

$$f(\mu) = tanh(\mu) \tag{2.4}$$

2.9.3 Parâmetros

Rauber (2005) mostra que os neurônios são responsáveis por atribuir pesos às entradas em relação às saídas de cada camada. Esse parâmetro define a importância de cada informação recebida pela rede neural frente ao resultado que se espera obter. Durante o treinamento da rede, os pesos vão sendo ajustados de maneira a minimizar as perdas observadas ao comparar o valor predito com o valor real, possibilitando que, iterativamente, vá se obtendo valores com menor erro à medida que o treinamento vai sendo realizado.

2.9.4 Redes neurais profundas - DNN

As redes neurais profundas (DNNs) são uma forma mais complexa das redes neurais tradicionais. Trata-se de uma ferramenta avançada de aprendizado supervisionado que se difere de uma rede neural comum pois possui mais camadas ocultas entre as camadas de entrada e saída. A estrutura mais elaborada possibilita análises igualmente mais complexas para as bases de dados fornecidas. (GOODFELLOW, 2016). As camadas ocultas são responsáveis pela modelagem de relações não lineares entre as variáveis, bem com tornam possível a extração de detalhes implícitos dos dados (LOCA; RAUBER, 2019).

A aplicação de redes neurais de aprendizado profundo aliada a recursos computacionais suficientes é, portanto, capaz de atingir níveis de sucesso antes nunca vistos com redes neurais simples.Krizhevsky, Sutskever e Hinton (2012)

A escolha de funções de ativação deve ser criteriosa com objetivo de possibilitar a detecção de não linearidades entre os dados da base de dados (BD), otimizando o resultado da rede neural. Uma das opções estudadas é a função de ativação ReLU (*Rectified Linear Unit*), uma vez que ela preserva os gradientes da entrada de dados, facilitando a interpretação da rede neural (NAIR; HINTON, 2010).

Ainda, é imperativo na aplicação de modelos de aprendizado de máquina a aplicação de metodologias que evitem o *overfitting*, ou seja, a aderência exagerada do modelo aos dados de treino. Para isso, podem ser aplicadas técnicas de regularização, como *dropout* (SRIVASTAVA et al., 2014) e *batch normalization* (IOFFE; SZEGEDY, 2015). A associação dessas técnicas permite que o modelo obtido tenha maior similaridade com os dados reais do processo em questão.

Ainda, as DNNs abriram caminhos para outras abordagens de aprendizado profundo, como as redes neurais convolucionais (CNNs) e redes neurais recorrentes (RNNs) (LOCA; RAUBER, 2019).

2.10 Dados sintéticos

No mundo real, a montagem de uma base de dados robusta exige uma análise criteriosa sobre a fonte dos dados, sensibilidade das informações (atualmente mais regulamentada pela lei geral de proteção de dados – LGPD), custos de obtenção, bem como sobre a qualidade das amostras registradas. Em razão desses fatores, a montagem de uma base de dados representativa pode se tornar uma tarefa extremamente penosa, bem como seu insucesso pode se traduzir também no êxito ou fracasso de todo trabalho de interpretação de uma base de dados com uso de ferramentas de inteligência artificial e redes neurais. Nesse sentido, algoritmos de geração de dados sintéticos vêm sendo utilizados a fim de aumentar a robustez dos modelos obtidos. Em suma, dados sintéticos são amostras integrantes de uma BD provenientes não da coleta de dados, mas de simulações que buscam representar os valores ou estados reais sem as dificuldades a eles relacionados.

2.10.1 Contexto de utilização de dados sintéticos

Os dados sintéticos, como disse Andrews (2021) "são criados em mundos digitais em vez de coletados ou medidos no mundo real". Porém, eles são estatisticamente semelhantes aos dados reais e tão eficientes quanto eles para treinamento de modelos de IA, trazendo ainda a vantagem de serem mais baratos que os processos de obtenção dos dados reais, bem com têm o benefício de serem totalmente aderentes às políticas de proteção de dados cada vez mais incisivas no mundo todo, o que é uma dificuldade com os dados reais. Portela (2022) pontua que as tecnologias para análises avançadas de dados já não são mais um gargalo nas linhas de pesquisa em inteligência artificial, mas sim a obtenção de dados de qualidade para essas análises, em quantidade razoável, bem como a dificuldade de tratamento dos dados obtidos para posterior treinamento dos modelos. Ele ainda comenta que um estudo da consultoria Gartner projeta que ainda em 2024 os dados sintéticos já serão 60% das bases de dados utilizadas em modelos de IA, ante 1% em 2021.

2.10.2 Desvio de conceito e desbalanceamento de classes

O trabalho de Pereira e Murai (2021) analisou o comportamento das redes neurais baseadas em grafos para detecção de fraudes no mercado financeiro. Embora seja aplicado a uma área bem diferente do ambiente industrial que se pretende tratar neste trabalho, os autores abordaram várias técnicas que se aplicam para este desenvolvimento, como a geração de dados sintéticos. Além disso, foram conceituados termos muito importantes que foram avaliados no contexto das transações financeiras e que servem para os dados aqui avaliados. O primeiro é o desvio de conceito, que caracteriza o fato de dados de treinamento se tornarem obsoletos com o tempo em razão de mudanças do padrão de comportamento do sistema avaliado. Nesse sentido, os dados sintéticos permitem aumentar a robustez de uma BD pequena que represente de maneira eficaz o padrão de operação atual de qualquer processo, uma vez que uma base maior pode acabar englobando padrões diferentes e prejudicar o modelo. O segundo termo é o desequilíbrio de classes, que no trabalho dos autores significava que dentre o universo de dados existem muito menos casos de fraude do que de transações legítimas. Com isso, no decorrer do trabalho Pereira e Murai (2021) avaliaram vários algoritmos para determinar qual teria o melhor desempenho, o que pode servir de inspiração na escolha dos algoritmos que serão aqui avaliados.

No contexto do desbalanceamento de classes e dados sintéticos, foi possível observar uma aplicação prática no trabalho de Fonseca et al. (2023), no qual os autores realizaram uma classificação de imagens de radiografias de torax para criar um modelo de detecção de tuberculose, atuando como ferramenta de saúde pública. No contexto da construção da base de dados de radiografias, era muito maior a quantidade de imagens de pulmões sadios do que doentes, o que era um fator de risco à qualidade do treinamento do modelo. Por isso, os autores aplicaram a técnica SMOTE (sobreamostragem de minoria sintética), que reamostra dados de uma base já criada tomando os valores vizinhos como base para criação de uma nova instância. As amostras sintéticas foram criadas até que o número de imagens de casos de tuberculose fosse igual ao número de casos sem tuberculose. Depois da geração dos dados, os autores aplicaram um modelo de redes neurais do tipo *feed forward*, em razão da existência de vários casos de sucesso em aplicações médicas, e validação cruzada com 10 dobras, avaliando ao final os índices de acurácia, taxa de verdadeiros positivos e taxa de verdadeiros negativos. Com isso, foi possível

confirmar que a geração de dados sintéticos ajudou nos resultados, especificamente na taxa de verdadeiros positivos, em comparação à execução do modelo com a BD original.

2.10.3 Algoritmos de geração de dados sintéticos

A depender da natureza das bases de dados disponíveis e que se deseja obter, várias metodologias de geração de dados sintéticos podem ser aplicadas. Passando por métodos estatísticos, utilização de redes neurais e *machine learning*, modificação de dados reais ou simulações computacionais, todas as aplicações têm um objetivo em comum: aumentar a base de dados com intenção de melhorar os resultados de métodos de predição sem a necessidade de se buscar mais dados reais, processo muitas vezes trabalhoso.

Uma das opções é a utilização de métodos baseados em estatística, como o método de Monte Carlo, utilizado por Antunes (2020). O método consiste em definir um modelo computacional que represente a natureza dos dados e então efetuar uma distribuição de probabilidade que caracterize a variável que se deseja obter. Dessa forma, são recebidos vários possíveis valores mediante os diversos cenários analisados.

Também são encontrados algoritmos de modelagem generativa, utilizando redes neurais e aprendizados de máquina, por exemplo, para geração de amostras adicionais. Silva e Rodrigues (2020), por exemplo, avaliaram brevemente a utilização dessas metodologias na ampliação de bases de dados médicas, especificamente em imagens.

Também é possível gerar dados sintéticos por meio de abordagens que se baseiam em dados reais e geram novas amostras por meio da perturbação dos dados (com ruído, por exemplo) ou até pela interpolação e extrapolação, combinando linear ou não linearmente os dados disponíveis. Com esse viés, Fonseca et al. (2023) aplicaram a metodologia SMOTE, que identifica a classe minoritária com objetivo de gerar mais amostras e diminuir o desequilíbrio de classes. Para isso, o algoritmo seleciona uma instância da classe minoritária e calcula a distância entre alguns vizinhos mais próximos (quantidade que é definida no algoritmo). Depois, são geradas novas amostras por meio da interpolação entre estes pontos até que seja atingido o equilíbrio entre as classes da base de dados. Nesse contexto, Junior et al. (2021) aplicaram o algoritmo interpolando entre 2 instâncias vizinhas para balanceamento entre as classes do trabalho. Ainda, o algoritmo só foi aplicado após uma limpeza da base de dados por meio da remoção de *outliers*.

Outra opção é a geração de ruído Gaussiano, quando são criados dados sintéticos com condições controladas de variação em comparação com os dados reais. Birnie e Ravasi (2021) aplicaram a geração de ruído gaussiano em uma base real de dados sísmicos. Nesse contexto em questão foi ponderado que os dados gerados precisam ser realistas de forma a não prejudicar os modelos criados, o que não era uma realidade quando utilizado ruído branco. Por isso, os autores ainda enfatizam que, basicamente, não pode ser possível distinguir entre os dados reais e os gerados sinteticamente, o que foi alcançado com o ruído gaussiano, controlando índices como média e desvio padrão. No contexto de dados sintéticos para variáveis de instrumentação industrial, cujo ruído já é inerente à medição, o ruído Gaussiano possibilita gerar um comportamento muito semelhante aos dados reais, trazendo mais coerência para a BD sintética.

Ainda, alguns autores utilizam técnicas de simulação de processos para obter dados fictícios que possam ser utilizados como BD. Neste caso, as simulações precisam ser realistas de forma que os dados obtidos sejam, de fato, representativos. Um exemplo é o que foi mostrado por Júnior e Setti (2010), que realizaram a criação de dados sintéticos por meio de um simulador microscópico de tráfego. A intenção do trabalho era gerar uma base de dados similar à obtida por meio da observação presencial de uma estrada ou pelos dados coletados por sensores diretamente instalados nas vias. Embora não fosse um processo tão caro, a intenção era gerar dados tão bons quanto os dos sensores sem necessitar dessa etapa de amostragem. No trabalho, a geração de dados sintéticos se deu, portanto, por meio da utilização de um simulador de tráfego cujos parâmetros foram calibrados por meio de algoritmos genéticos de forma a minimizar as diferenças entre os dados reais e os sintéticos. Estes algoritmos genéticos são baseados nas teorias de Darwin e vão eliminando os indivíduos que não conseguem se adaptar bem ao meio, convergindo o modelo para soluções ótimas. Esta técnica aplicada por Júnior e Setti (2010) obteve êxito na geração de dados sintéticos de tráfego, mostrando que foi uma abordagem eficaz de geração de dados sintéticos.

2.11 Métricas de avaliação dos modelos

Após a aplicação dos algoritmos de aprendizado de máquina, é preciso avaliá-los para entender a eficácia de cada modelo, bem como nortear os passos seguintes do desenvolvimento dos códigos. Para isso, diversas métricas são calculadas e interpretadas de forma a transformar

em números o grau de aderência dos resultados preditos pelos algoritmos frente ao valor real da base de dados.

2.11.1 Coeficientes de correlação

Wang et al. (2023) utilizaram metodologias de aprendizado de máquina para prever indicadores de qualidade de água. Para verificar a eficácia dos modelos, foram aplicadas medições do coeficiente de correlação de Pearson com um método de peso de entropia, resultando em indicativos de quais são as variáveis mais relevantes nessa medição de qualidade. O coeficiente de correlação de Pearson utilizado avalia a força e a direção da correlação das variáveis e varia entre -1 (correlação negativa perfeita), passando pelo 0 (ausência de correlação) e chegando até 1 (correlação positiva perfeita). Nesse sentido, o coeficiente se mostrou uma boa ferramenta de medição da eficácia na aplicação de modelos de aprendizado de máquina. Sua equação também foi demonstrada pelos autores e segue abaixo:

$$R = \frac{\sum_{(X_i \breve{} X_{mean})(Y_i \breve{} Y_{mean})}}{N * X_{std} * Y_{std}} \tag{2.5}$$

Na equação, as variáveis tem o seguinte significado:

X_i = Valor individual da variável X (entrada);

X_{mean} = Valor médio de X;

Y_i = Valor individual da variável Y (saída ou alvo);

Y_{mean} = Valor médio de Y;

X_{std} = Desvio padrão de X;

Y_{std} = Desvio padrão de Y;

N = número de pares de dados;

Com o cálculo e interpretação dos resultados do coeficiente de correlação de Pearson ou Spearman, por exemplo, torna-se possível, inclusive, otimizar a base de dados antes da aplicação de qualquer algoritmo de aprendizado de máquina, uma vez que já é possível determinar variáveis que são mais ou menos relevantes para a grandeza de saída do modelo (LOCA; RAUBER, 2019).

2.11.2 Índices R², MSE e RMSE

Ainda, equações mais populares de aferição de desempenho de modelos de inteligência artificial foram tabulados conforme a tabela 2.1

Tabela 2.1 – Tabela de métodos de ajuste de modelos

Métodos	Equação	Intervalo	Valor Ótimo
R^2	$R^2 = \frac{\sum(P_{iy}-\overline{P_i})^2}{\sum(P_{iy}-\overline{P_{iy}})^2}$	-∞ a 1	1
MSE	$MSE = \frac{1}{N}\sum(P_i - P_{iy})^2$	0 a ∞	0
RMSE	$RMSE = \left[\frac{1}{N}\sum(P_i - P_{iy})^2\right]^{1/2}$	0 a ∞	0

Fonte: Filho et al. (2020).

O termo R² refere-se ao coeficiente de determinação e mede o quanto uma variável de saída está sendo explicada pelas variáveis de entrada do modelo utilizado, ainda conforme Filho et al. (2020). Quando essa métrica tem valores negativos, depreende-se que as variáveis de entrada não conseguem explicar o comportamento da variável de saída, sendo a aplicação do modelo pior do que o uso de uma simples regressão linear. À partir de 0 até 1, quanto maior o valor, maior foi a captura das relações entre as variáveis da BD.

O coeficiente MSE (*mean squared error*) refere-se ao valor médio do quadrado das diferenças entre os valores reais e os preditos pelo modelo aplicado. Pela exponenciação, acaba amplificando grandes erros, o que ajuda a demonstrar a presença de *outliers*. Porém, é preciso um maior esforço de interpretação dos resultados justamente por essa sensibilidade. Já o RMSE (*root mean squared error*) é basicamente a raiz do MSE. Com isso, o valor de erro passa a ser dado na mesma unidade das variáveis analisadas, o que torna a interpretação mais amigável. Com ele entende-se o erro no mesmo contexto da base de dados (CHAI; DRAXLER, 2014).

Já os demais termos das equações são:

P_i = Valor real;

P_i = Valor predito;

P_iy = Média dos valores observados;

2.12 Considerações Finais

Como já foi citado anteriormente, o aprendizado de máquina, por mais avançado que já esteja, ainda tem pela frente uma grande curva de crescimento impulsionada pela indústria 4.0 e o avanço da computação em geral. Esse cenário dividido pelas ferramentas traz maiores

ganhos sem abrir mão da segurança associada aos processos em que elas estão aplicadas. É evidente a grande quantidade de pesquisas realizadas nesse campo disponível para consulta na internet. Ao mesmo tempo, ainda não é tão comum encontrar esse tipo de pesquisa aplicada diretamente ao ambiente de indústria pesada. Dessa forma, o presente referencial teórico foi feito de maneira a entender o que vem sendo usado desse grande campo, com os pontos fortes e fracos, de forma a traçar um paralelo das melhores ferramentas a serem aplicadas ao universo prático da indústria de fabricação de cimento. Por fim, foi realizada uma análise geral do fluxo de produção de clínquer e cimento para contextualizar a área de atuação do trabalho.

3 Trabalhos relacionados

Diversos trabalhos na literatura demonstram o uso de aprendizado de máquina para aumento de índices de disponibilidade e eficiência nas mais variadas cadeias de produção. Em maior ou menor escala, essas metodologias têm sido utilizadas de maneira a garantir previsibilidade e antecipação nos processos de produção. Neste capítulo, são apresentados alguns trabalhos que promoveram o uso do aprendizado de máquina em ambientes industriais ou fora deles, mas que forneceram o embasamento necessário para entender o campo de possibilidades dado pela inteligência artificial.

Antonelli e Neitzel (2015) realizaram a aplicação de redes neurais no processo de produção de fios de algodão. Para isso, uma base de dados foi construída com informações das fibras utilizadas e parâmetros de produção para prever a qualidade do fio. Os autores encontraram dificuldades na construção da BD pois as informações de qualidade das fibras não eram contínuas, além de que as mudanças de qualidade do fio só eram percebidas 48 horas depois da alteração da nova mistura de fibras. Da execução do trabalho concluiu-se que o modelo gerado não era muito satisfatório na previsão de dados absolutos de qualidade do fio, mas se mostrou eficiente para definir variações na qualidade do produto final frente a variações da matéria-prima, gerando mais informações para montagem ideal da mistura utilizada na produção do fio.

Girelli e Corso (2020) utilizaram redes neurais convolucionais para detecção de falhas em contentores industriais. Popularmente essas redes vêm sendo utilizadas para detecção e classificação em bases de dados compostas por imagens. Nesse sentido, uma base com fotos de contentores normais e defeituosos foi utilizada para treinamentos e testes do modelo, obtendo satisfatoriamente a detecção inclusive do local de defeito. Dadas as devidas proporções, o estudo conseguiu bons indicadores de precisão para o tipo de contentor avaliado, uma vez que a empresa na qual foi realizado o estudo possuía vários modelos em produção.

Já Freitas et al. (2023) conseguiram reunir uma ampla gama de algoritmos no trabalho. Baseado na ideia de separação em grupos do aprendizado não supervisionado, os autores reuniram dados financeiros de instituições de ensino superior e aplicaram as metodologias k-means e hierárquico aglomerativa para separação das instituições em *clusters*. Dos resultados da partição já foi possível inferir quais instituições tinham números financeiros atípicos, sendo elas as que estavam em grupos mais vazios. Depois, para corroborar ainda mais a análise, aplicou-se as metodologias de detecção de outliers "*Angle-based Outlier Detection, Feature Bagging, Isolation*

Forest, K-Nearest Neighbors, Local Outlier Factor e One Class SVM"(FREITAS et al., 2023). Com isso, os autores conseguiram explicar com o trabalho todo o cerne do aprendizado não supervisionado. Como não havia necessidade de um determinismo absoluto de valores (como de uma coluna alvo de uma base de dados), dividir computacionalmente os dados em grupos foi uma excelente estratégia para detectar discrepâncias por meio das técnicas complementares de remoção de *outliers*, deixando clara uma ótima abordagem do aprendizado sem supervisão.

O trabalho de Gois et al. (2019) tomou como base 19 variáveis de entrada para prever uma variável de consumo de combustível em um alto forno de produção de ferro fundido. Para isso, as informações colhidas de um processo real foram aplicados em uma rede neural artificial com 19 neurônios de entrada e de camada intermediária e 1 neurônio de saída. Este número foi escolhido por ser exatamente o número de variáveis de entrada e de saída do processo. Ao final do estudo, os autores demonstraram a eficácia do modelo obtido para previsão do consumo de combustível neste processo siderúrgico.

Li et al. (2015) elaboraram um modelo composto por várias metodologias de aprendizado de máquina baseados em imagens da chama dentro do forno e dos dados de processo, conforme o quadro 3.1 que segue:

Quadro 3.1 – Dados de entrada do modelo de predição.

Flame images features (input)	Color of ROI Global configuration of ROI Local configuration of ROI
Kiln operating variables (input)	Coal feeding (Wc) Opening degree of induced draft fan (Od) Kiln main motor current(Ik) Raw material pump current(Im) Kiln tail temperature(Tt) Kiln head temperature(Th) Kiln head pressure (Ph)
Raw material quality (input)	Lime saturation factor (K H) Silicic acid rate (S M) Alumina mudulua (A M) Granularity (Gr)
Clinker quality (output)	f-CaO content

Fonte: Li et al. (2015).

Foram aplicados algoritmos de KPLS (*Kernel Partial Least Squares*), RVFL (*Random Vector Functional Link*), SVR (*Support Vector Regression*) e PLS (*Partial Least Squares*) individualmente e em pares para obtenção dos melhores resultados. Para avaliar a eficácia dos modelos, os autores utilizaram os indicadores RMSE (erro quadrático médio) e R^2 (coeficiente

de determinação) e obtiveram como melhores métricas os valores de RMSE = 0.092 ± 0.01 e R2 = 0.958 ± 0.01 para os dados de teste.

Magalhães (2019) desenvolveu em seu trabalho o que intitula como "sensor virtual para predição do teor de cal livre no clínquer". Para isso, utilizou técnicas de regressão linear múltipla, redes neurais artificiais do tipo Perceptron de Múltiplas Camadas (MLP) e Função de Base Radial (RBF) para elaboração do modelo de predição baseado em 15 variáveis de entrada e 1 de saída, conforme a Tabela 3.1.

Tabela 3.1 – Base de dados utilizada no trabalho de Magalhães (2019).

Nº	Variável	Unidade	Classificação	Tipo de coleta
1	Temperatura na entrada do forno	ºC	Entrada	Online
2	Pressão na entrada do forno	mmca	Entrada	Online
3	NOx na entrada do forno	ppm	Entrada	Online
4	CO na entrada do forno	ppm	Entrada	Online
5	O2 na entrada do forno	%	Entrada	Online
6	Temperatura de saída da torre de ciclones	ºC	Entrada	Online
7	Temperatura do 6º estágio da torre de ciclones	ºC	Entrada	Online
8	Consumo térmico do forno	kcal/kg de clínquer	Entrada	Online
9	Taxa de combustível no forno	t/h	Entrada	Online
10	Taxa de combustível no pré-calcinador	t/h	Entrada	Online
11	Alimentação de farinha no forno	t/h	Entrada	Online
12	Temperatura na zona de queima	ºC	Entrada	Online
13	Temperatura do ar secundário	ºC	Entrada	Online
14	Corrente do motor do forno	%	Entrada	Online
15	Rotação do forno	RPM	Entrada	Online
16	Cal livre no clínquer	%	Saída	Análise de laboratório

Fonte: Magalhães (2019).

A autora também considerou o tempo de residência do clínquer no forno e no resfriador, norteando a coleta de dados defasados com relação à análise de cal livre entre 30 e 150 minutos, o que foi definido empiricamente. Para calcular a eficiência dos modelos foram utilizadas as métricas de R^2 e MSE, cujos melhores valores encontrados foram de 0,73 e 0,1162, respectivamente.

Para melhorar os resultados do modelo de redes neurais, Loca e Rauber (2019) aplicaram o coeficiente de correlação de Spearman para determinar quais variáveis do seu *dataset* eram mais interessantes no treinamento da rede. Com isso, depois da parametrização e testes de uma

rede neural profunda com a base de dados completa, executaram iterações em 3 cenários de filtro da BD, sendo eles: coeficientes menores que 30%, menores que 50% e, por fim, maiores que 70%. Os resultados tiveram maior precisão à medida que os níveis de correlação ficaram mais restritos, mostrando que essa foi uma boa estratégia de seleção de dados. A seleção de dados com maior coeficiente de correlação auxilia o modelo em dois viéses de interpretação. Primeiro, quanto maior o coeficiente de correlação, mais fácil ficou do modelo captar as variações dos resultados. Além disso, as bases filtradas ficaram menores do que a original, o que também economiza processamento computacional durante a execução dos modelos.

Souza et al. (2015) utilizaram uma técnica de minimização da energia livre de Gibbs para prever a composição do clínquer produzido em uma planta real. Os resultados se mostraram muito satisfatórios principalmente para as componentes que ocorrem em maior quantidade, enquanto foi observado um erro maior na mensuração das componentes que são minoria no clínquer.

Negreiros et al. (2020) adotaram o modelo baseado em *Support Vector Regression* (SVR) para prever a resistência à compressão a 28 dias do cimento CP-II de uma planta de produção de cimento por meio de uma base de dados com mais de 20000 ensaios realizados. Após um trabalho de limpeza e organização da base de dados, com posterior aplicação dos códigos, foi obtido um valor de RMSE (raiz do erro quadrático médio) na casa de 5%, indicando ótima precisão do modelo. Porém, os autores ressaltam a necessidade de ter muito critério na interpretação dos resultados, uma vez que consideraram que a base de dados tem falhas importantes na sua construção em razão de dados mal coletados no decorrer do tempo.

Com base nos estudos avaliados nas seções 2,de referencial teórico, e 3, de trabalhos correlatos, depreende-se que a natureza dos dados disponíveis na base do presente trabalho será útil para aplicação de modelos de regressão de dados. Dessa forma, a intenção é ter uma ferramenta de análise preditiva do teor de cal livre no clínquer, capaz também de nortear a atuação das equipes de operação, qualidade e processo.

Os quadros 3.2, 3.3 e 3.4 abaixo resumem os trabalhos cujas abordagens foram mais semelhantes à proposta deste estudo, com objetivo principal de mostrar os resultados obtidos para posterior comparação com o que foi alcançado na presente dissertação.

Quadro 3.2 – Comparação dos principais trabalhos - Parte 1.

	(GOIS et al., 2019)	**(LI et al., 2015)**
Resumo	RN para predição de consumo de combustível de alto forno	Sensor virtual de predição de cal livre usando dados e imagens
Resultados	R^2 = 0,837; RMSE = 11,8 a 12,7	RMSE = 0.068 ± 0.01
Contribuições	- Obteve resultados usando dados de processo; - Métricas de avaliação MSE, RMSE e R^2.	- Provou factibilidade do projeto; - Utilização de RMSE para avaliação.
Lacunas	- Utilizou média diária dos dados; - Processo de alto forno é diferente do forno de clínquer.	- Uso de imagens da chama do forno, sensibilidade à qualidade da imagem; - Uso de variáveis de processo não disponíveis na amostragem adequada.

Fonte: Do Autor (2024).

Quadro 3.3 – Comparação dos principais trabalhos - Parte 2.

	(ANTONELLI; NEITZEL, 2015)	**(NEGREIROS et al., 2020)**
Resumo	Predição da qualidade de fios de algodão	Predição da resistência à compressão 28 dias de cimento CPII
Resultados	R^2 = 0,9518.	R^2 = 0,95; RMSE = 0,05
Contribuições	- Montagem ideal da RNA; - Nº de camadas igual ao número de dados de entrada.	- Utilização de ML (SVR)
Lacunas	- Não conseguiu prever a qualidade propriamente dita; - Teve grande impacto negativo em razão da qualidade da amostragem.	- Necessidade de boas bases de dados; - Cálculo da resistência à compressão é dependente de ensaios e lançamentos manuais.

Fonte: Do Autor (2024).

Quadro 3.4 – Comparação dos principais trabalhos - Parte 3.

	(LOCA; RAUBER, 2019)	**(MAGALHãES, 2019)**
Resumo	CNN para detecção de falhas em processos industriais	Sensor virtual de predição de cal livre
Resultados	Acurácia entre 50 e 100%	R^2 = 0,73; MSE = 0,1162
Contribuições	- Influência da seleção de variáveis com menor ou maior correlação com a variável alvo.	- Variáveis de interesse para BD; - Factibilidade da abordagem de predição da Cal livre; - Defasagem na coleta de dados;
Lacunas	- Foi aplicado somente em ambiente de simulação e focado em falhas de segurança nos processos.	- Os fornos de clínquer não são iguais em tamanho, nº de estágios, alimentação de farinha, tipos de combustíveis, etc..

Fonte: Do Autor (2024).

As lacunas detectadas nos trabalhos acima listados têm causas em comum em alguns casos. É evidente que todos os dados cujo processo de coleta ou lançamento é manual têm erros associados, e ficou claro o impacto da qualidade dos dados nos resultados dos modelos implementados. Ainda, por se tratar de um processo químico muito específico, foi preciso focar em estudos da área de produção de cimento, que não foram encontrados em grandes quantidades. Para resolver estas lacunas, ficou claro ser necessário priorizar a construção da base de dados com informações coletadas *online*, por sensores ou outros dispositivos encontrados no processo, bem como escolher, dentre essas informações, as que não tem tanta perturbação por sujeira ou outros fatores, como aconteceu no caso das imagens no trabalho de Li et al. (2015).

O presente trabalho difere-se dos demais em razão do parque fabril brasileiro ser muito heterogêneo. As diferenças entre as fábricas de cimento fazem com que não exista uma solução universal para todas as plantas. Essa afirmativa se justifica pois os distintos arranjos físicos das fábricas gera condições de processo com várias diferenças, o que por si só já corrobora uma análise personalizada para cada fábrica. Em adição a este fator, observa-se uma busca crescente por combustíveis alternativos para queima, os chamados resíduos, abordados no tópico 2.3, e consequentemente as condições de processo são afetadas diretamente durante a operação da planta. A maneira como cada resíduo afeta a produção é definida por aspectos físico-químicos das matérias-primas utilizadas e também dos resíduos, o que novamente gera uma perspectiva personalizada para cada forno de clínquer, justificando assim a análise individual de cada processo. Por fim, todas as matérias primas utilizadas na produção de cimento também variam química e fisicamente de acordo com a localização geográfica da planta, corroborando ainda mais a análise individual de cada processo.

A análise individual de cada linha de produção de clínquer deve ser feita no sentido de capturar as particularidades de comportamento de cada variável especificamente no cenário estudado. Por isso, depreende-se que o modelo elaborado para uma fábrica não tem interoperabilidade com outra. No entanto, as técnicas de processamento e análise de dados, bem como a escolha de variáveis, pode ser melhor aproveitada entre aplicações de diferentes fornos.

Por fim, outro diferencial está em utilizar somente os dados de processo já coletados normalmente e gerar resultados melhores do que dos trabalhos estudados, o que será aferido por meio das métricas comuns em aplicações de IA, e que foram abordadas no item 2.11.2.

4 Metodologia

O trabalho foi realizado aplicando algoritmos de aprendizado de máquina utilizando uma base de dados de uma fábrica de cimento real, com produção nominal de 5000 toneladas de clínquer por dia.

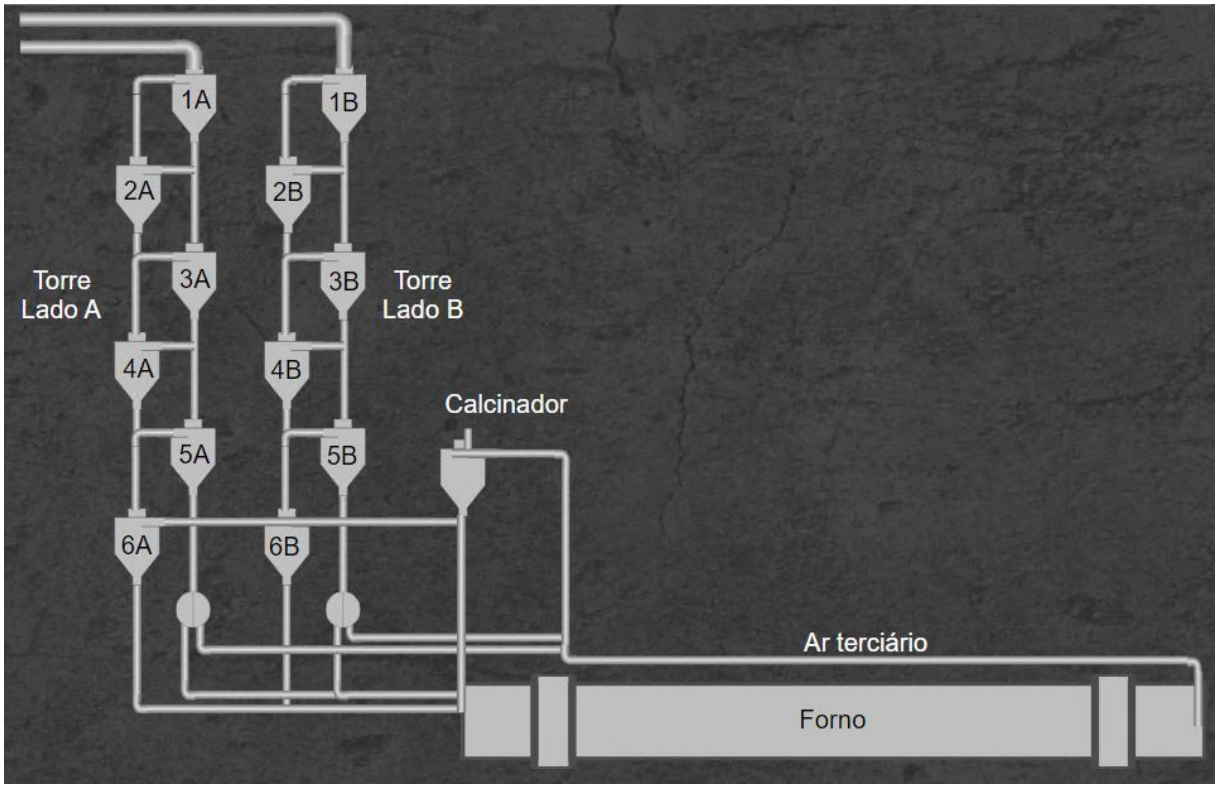
4.1 Pesquisa bibliográfica

Na etapa de pesquisa bibliográfica foram consultados livros e artigos relacionados à área de atuação do presente trabalho. A busca foi realizada com objetivo de se entender parte da química relacionada ao processo de produção de clínquer e cimento, bem como de conhecer o que vem sendo pesquisado quanto ao tema da cal livre no clínquer e predições em geral aplicadas ao processo produtivo de cimento.

4.2 Escolha de variáveis

Embora a química envolvida na produção de clínquer seja uma só, cada fábrica de cimento tem suas características individuais. Essas diferenças entre fábricas podem estar na composição de matérias-primas, nos tipos de combustíveis utilizados (principalmente os combustíveis alternativos), número de estágios da torre de pré-calcinação, comprimento do forno, periodicidade das coletas de amostras, etc. Para a fábrica onde foi desenvolvido este trabalho, a torre de pré-calcinação é dupla com 6 estágios e um calcinador, culminando em um forno de 62,5 metros de comprimento e 5 de diâmetro. A Figura 4.1 contém um desenho esquemático da planta em questão.

Figura 4.1 – Esquemático do forno estudado.



Fonte: Do autor (2024)..

Nessa planta, a alimentação de farinha do forno é realizada por meio dos dois ramais da torre, conhecidos como Lado A e Lado B. É possível modificar a proporção dessa divisão a critério do operador. O material então desce a torre realizando a troca de calor com o gás quente que está subindo no processo, iniciando a descarbonatação da farinha.

O calcinador é responsável por fornecer mais energia térmica ao processo, de forma a acelerar as reações de descarbonatação do material que foi inserido na torre. Já no forno são atingidas as maiores temperaturas de todo o processo. A farinha, agora bem aquecida, assume um estado similar à lava de um vulcão, fluindo vagarosamente pelo forno rotativo até sua saída, que culmina no resfriador.

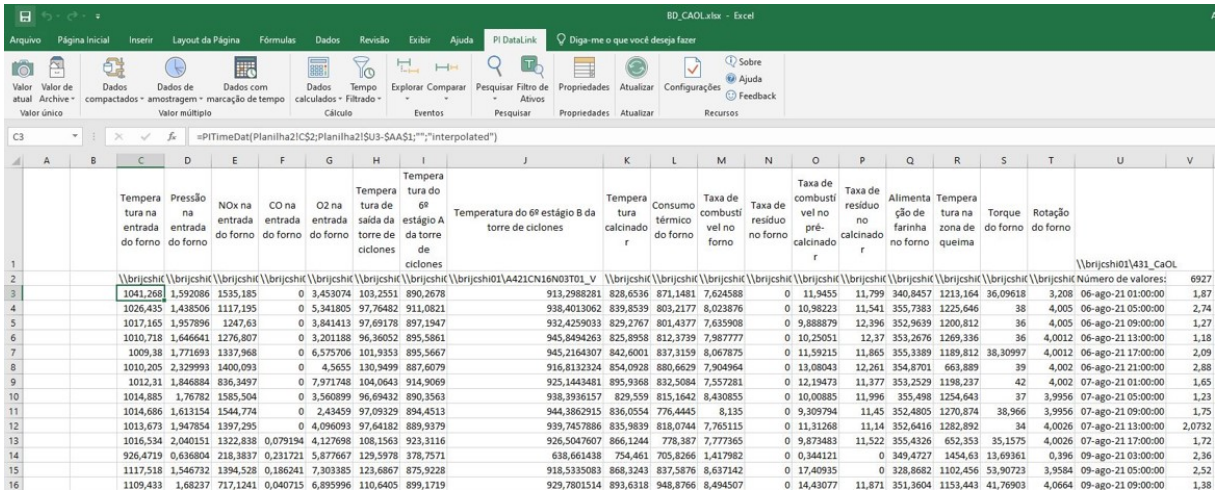
Vale ressaltar que todo o processo é contínuo. Portanto, a farinha obrigatoriamente faz o caminho acima descrito. Ao chegar no resfriador, o material passa por resfriamento rápido, gerando o clínquer. Na saída do resfriador, amostras de clínquer são coletadas periodicamente para monitoramento de parâmetros de qualidade.

Antes de iniciar a coleta de dados para construção dos modelos, foi conduzida uma conversa com profissionais que detém grande conhecimento técnico do processo da fábrica de cimento na qual os trabalhos foram realizados. Com base nas variáveis por eles selecionadas, foi realizada a montagem da base de dados para implementação dos algoritmos.

4.3 Levantamento da base de dados

A etapa de levantamento da base de dados envolve a escolha de variáveis anteriormente realizada e a extração dos dados históricos. Para construção da base, os dados foram extraídos do historiador *PI System*, instalado na fábrica, por meio da ferramenta *Data Link*, que permite a exibição dos dados diretamente em planilhas, como mostra a Figura 4.2. foram extraídas as medições de cal livre e o valor das variáveis selecionadas pelos especialistas de processo exatamente com a mesma marcação de tempo da cal livre ou em instantes anteriores, o que ficará mais claro no tópico 5.

Figura 4.2 – Interface de obtenção dos dados.



Fonte: Do autor (2024)..

4.4 Implementação de algoritmos de geração de dados sintéticos

Além dos dados reais de processo extraídos por meio da ferramenta *Data Link*, foram aplicados algoritmos de geração de dados sintéticos com objetivo de promover maior aderência dos modelos de ML aos resultados esperados de cal livre. Foram avaliadas diversas opções de geração de dados e comparados os resultados entre os modelos.

4.5 Testes com modelos

Após a montagem das bases de dados, foram executados testes com modelos de aprendizado de máquina de forma a entender qual a melhor abordagem a ser realizada. Como observado no referencial teórico (Capítulo 2) existem amplas possibilidades de utilização de algoritmos de

regressão, redes neurais, dentre outras metodologias, e estas tiveram seus resultados comparados por meio dos índices MSE, RMSE, R^2, coeficiente de Pearson e Spearman. Com os resultados dessas métricas, será feita uma comparação entre as metodologias e iterações realizadas a fim de se encontrar o melhor ajuste conforme os dados coletados.

4.6 Métricas de avaliação dos algoritmos

As primeiras execuções de códigos baseados em aprendizado de máquina foram realizadas somente com intuito de validar a representatividade dos dados selecionados para a predição da cal livre, no sentido de figurar como uma confirmação que o trabalho tinha chance de êxito. Dessa forma, foram adotadas somente métricas mais básicas observadas durante o levantamento bibliográfico. Com o cálculo de MSE e R^2, já foi possível observar a evolução à medida que a base de dados começa a ser organizada e limpa, bem como as diferenças de aplicação dos algoritmos de aprendizado de máquina. Já com os modelos consolidados, serão medidos os coeficientes MSE, RMSE e R^2 para possibilitar a comparação com os trabalhos consultados no Capítulo 2, de referencial teórico.

4.7 Validação do projeto

Uma vez validada a escolha preliminar dos dados e a possibilidade de melhoria nos resultados dos modelos, foram aplicados modelos mais complexos em busca do arranjo que melhor captou o comportamento da cal livre frente às variáveis escolhidas da base de dados. Ao final foram verificados resultados melhores que os dos trabalhos avaliados no referencial teórico. Em uma análise de ganhos que podem ser auferidos com a implementação de uma ferramenta para predição em tempo real, o maior destaque fica para o cenário de tornar possível, baseado na confiabilidade dos resultados preditos, aumentar o intervalo entre as amostras de clínquer, diminuindo a exposição dos profissionais aos riscos da coleta na área industrial, bem como otimizando as suas rotinas do laboratório.

5 Resultados

Para execução do presente trabalho, várias etapas foram realizadas em busca da melhor combinação de base de dados, algoritmos de *machine learning* (ML) e dados sintéticos. As subseções a seguir descrevem o que foi executado em cada etapa.

5.1 Escolha de variáveis para montagem da base de dados

O bom resultado de um algoritmo de aprendizado de máquina (ML) depende diretamente da qualidade dos dados disponibilizados aos modelos para treinamento. Por outro lado, a análise de dados feita pelos modelos independe da sua origem, ou seja, não existe um modelo de ML específico para processos químicos e outro para físicos. Uma vez fornecidas informações que sejam representativas para a variável que se pretende predizer, os algoritmos captam as relações independente de sua natureza. Nessa ótica, entende-se que não seria necessário um conhecimento detalhado da química que descreve a produção de clínquer e cimento. Porém, a montagem de uma base de dados de processo sem nenhum conhecimento das suas relações implica certamente na seleção de dados irrelevantes e a falta de dados importantes. Por isso, foram consultados especialistas da área de processo da fábrica, que contribuíram com seu conhecimento aprofundado das reações químicas que ocorrem na produção de clínquer.

Primeiramente, foram perguntados sobre quais variáveis de processo, na opinião de cada um, tinham influência direta na cal livre do forno. A Tabela 5.1 ilustra as variáveis listadas pelos especialistas e, no final, caracteriza também a variável alvo, cal livre.

Embora todas as variáveis da Tabela 5.1 foram consideradas pelos especialistas como importantes e representativas na mensuração da cal livre, é preciso levar em consideração o que foi aprendido durante a etapa de levantamento bibliográfico, que mostrou o impacto negativo causado pelo uso de variáveis com alto erro associado à sua aferição, registro ou qualquer outra etapa do processo de amostragem. Nesse sentido, os especialistas novamente foram consultados para indicarem variáveis cujo lançamento é mais suscetível a erros, cujo período entre amostragens é muito grande ou caso existam quaisquer condições que possam impactar diretamente o valor medido.

Assim, os resultados das análises de cinzas e PCI (itens 23, 24, 25 e 26 da Tabela 5.1 foram desconsiderados por serem obtidos por meio de um processo extremamente manual e demorado, além de imprecisões na informação exata de qual material está sendo consumido a

Tabela 5.1 – Variáveis levantadas para montagem da base de dados.

Nº	Variável	Unidade	Classe	Coleta
1	Temperatura entrada forno	ºC	Entrada	Online
2	Pressão entrada forno	mbar	Entrada	Online
3	NOx entrada forno	ppm	Entrada	Online
4	CO entrada forno	%	Entrada	Online
5	O2 entrada forno	%	Entrada	Online
6	Temperatura saída torre	ºC	Entrada	Online
7	Temperatura 6º estágio A torre	ºC	Entrada	Online
8	Temperatura 6º estágio B torre	ºC	Entrada	Online
9	Temperatura calcinador	ºC	Entrada	Online
10	Consumo térmico do forno	kcal	Entrada	Online
11	Taxa de combustível no forno	ton/h	Entrada	Online
12	Taxa de resíduo no forno	ton/h	Entrada	Online
13	Taxa de combustível no pré-calcinador	ton/h	Entrada	Online
14	Taxa de resíduo no calcinador	ton/h	Entrada	Online
15	Alimentação de farinha no forno	ton/h	Entrada	Online
16	Temperatura na zona de queima	ºC	Entrada	Online
17	Torque do forno	%	Entrada	Online
18	Rotação do forno	rpm	Entrada	Online
19	Fator de saturação farinha alimentada	%	Saída	Laboratório
20	Módulo de sílica farinha alimentada	N/D	Saída	Laboratório
21	Módulo de alumínio farinha alimentada	N/D	Saída	Laboratório
22	Fator de saturação de clínquer	%	Saída	Laboratório
23	Cinzas do resíduo do forno	%	Saída	Laboratório
24	PCI do resíduo do forno	%	Saída	Laboratório
25	Cinzas do resíduo do calcinador	%	Saída	Laboratório
26	PCI do resíduo do calcinador	%	Saída	Laboratório
27	Cal livre	%	Saída	Laboratório

Fonte: Do Autor (2024).

cada momento no forno. Outras variáveis que foram excluídas são as relacionadas à farinha alimentada ao forno (itens 19, 20 e 21) pois, embora amostrada automaticamente, é feita somente a cada 12 horas. Por fim, o fator de saturação de clínquer (item 22) também foi excluído por ser uma medida obtida no mesmo ensaio da cal livre. Essa decisão baseia-se no sentido de que, se está sendo elaborada uma ferramenta de predição, entende-se que, caso os valores preditos sejam suficientemente aderentes aos valores reais, é possível aumentar o intervalo entre amostras sem prejudicar o controle de qualidade, e caso o modelo fosse dependente de alguma variável medida nesse processo, a qualidade da predição seria diretamente afetada.

Portanto, a natureza da base de dados ficou baseada somente em variáveis coletadas online, ou seja, diretamente de algum instrumento ou equipamento de campo cujas medidas são contínuas e não dependem de intervenção humana no processo de obtenção.

A Tabela 5.2 mostra todos os dados que foram selecionados para composição da base de dados após a análise dos especialistas de processo da fábrica.

Tabela 5.2 – Variáveis selecionadas para montagem da base de dados.

Nº	Variável	Unidade	Classe	Coleta	Origem
1	Temperatura na entrada do forno	ºC	Entrada	Online	Termopar
2	Pressão na entrada do forno	mbar	Entrada	Online	Transdutor de pressão
3	NOx na entrada do forno	ppm	Entrada	Online	Analisador de gases
4	CO na entrada do forno	%	Entrada	Online	Analisador de gases
5	O2 na entrada do forno	%	Entrada	Online	Analisador de gases
6	Temperatura de saída da torre	ºC	Entrada	Online	Termopar
7	Temperatura do 6º estágio A da torre	ºC	Entrada	Online	Termopar
8	Temperatura do 6º estágio B da torre	ºC	Entrada	Online	Termopar
9	Temperatura calcinador	ºC	Entrada	Online	Termopar
10	Consumo térmico do forno	kcal	Entrada	Online	Cálculo no CLP
11	Taxa de combustível no forno	ton/h	Entrada	Online	Balança dosadora
12	Taxa de resíduo no forno	ton/h	Entrada	Online	Balança dosadora
13	Taxa de combustível no pré-calcinador	ton/h	Entrada	Online	Balança dosadora
14	Taxa de resíduo no calcinador	ton/h	Entrada	Online	Balança dosadora
15	Alimentação de farinha no forno	ton/h	Entrada	Online	Balança dosadora
16	Temperatura na zona de queima	ºC	Entrada	Online	Pirômetro
17	Torque do forno	%	Entrada	Online	Inversor de frequência
18	Rotação do forno	rpm	Entrada	Online	Inversor de frequência
19	Cal livre	%	Saída	Laboratório	Raio X

Fonte: Do Autor (2024).

Vale ressaltar que, para cada variável selecionada para compor a base de dados, serão coletados os valores no momento de lançamento da cal livre e também nos instantes de de-

fasagem que foram definidos, ou seja, cada variável aparece mais de uma vez na BD, sendo diferenciada pelo prefixo XXmin/, que representa o tempo de defasagem da coleta.

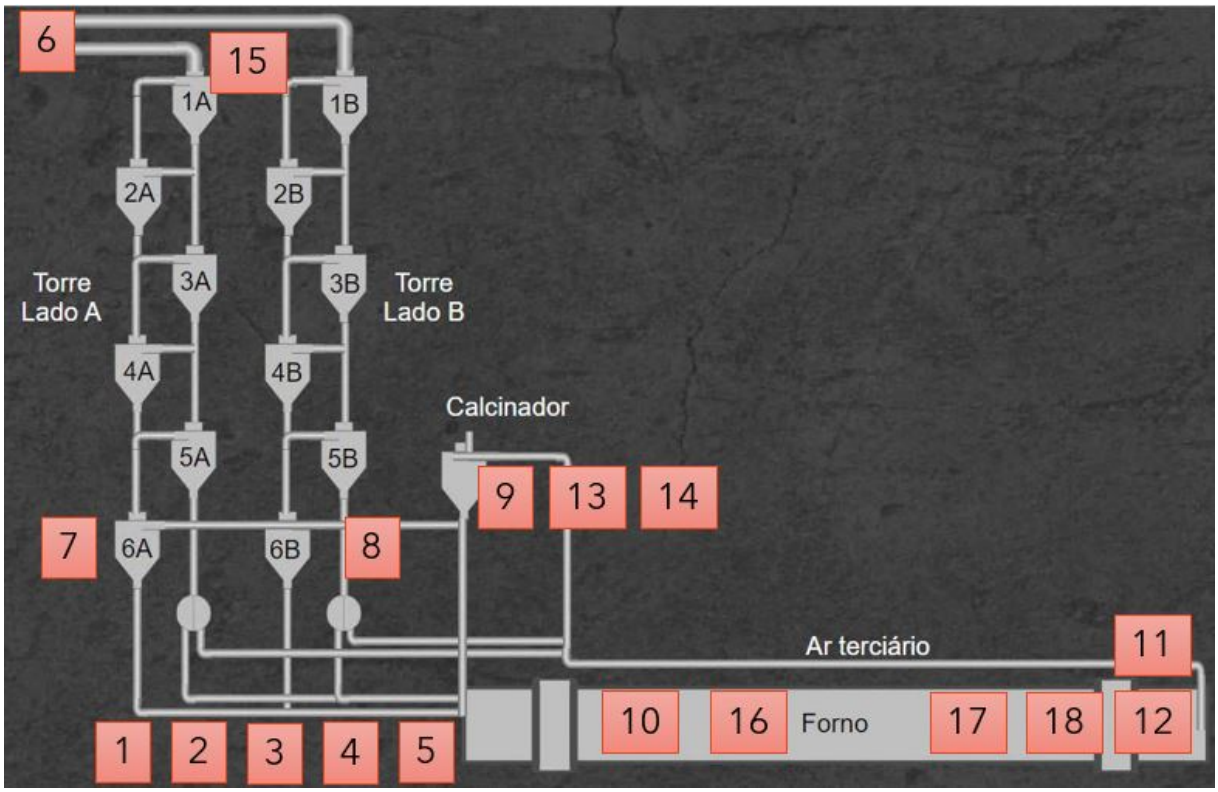
5.2 Construção da base de dados de processo

Para a construção dos algoritmos de ML, o primeiro passo é a obtenção de uma base de dados com informações suficientes em número e relevância para o modelo desenvolvido. Por isso, foram extraídos, com uso da ferramenta *PI Data Link*, todos os lançamentos manuais de análise de Cal livre em um horizonte de 3 anos corridos, dispostos em uma coluna. Para cada um deles, foram criadas mais 4 colunas referentes às 4 amostras consecutivamente anteriores. A intenção de coletar amostras anteriores da medição da cal livre visava garantir a visualização de um gradiente entre as coletas e nortear a variação possível entre uma amostra e outra. Em outras palavras, o valor da cal livre não dá um salto de 2 para 8% entre uma amostra e outra, mas vai subindo ou descendo gradativamente, o que fica mais fácil de entender ao analisar amostras consecutivamente anteriores à atual. Para as demais variáveis foram coletados os valores instantâneos do momento do lançamento da análise da cal livre e também defasados no tempo. Essa defasagem se justifica pois há de se considerar que o clínquer no ponto de coleta, na saída do resfriador, representa os dados de processo do forno e da torre de ciclones em momentos anteriores. O tempo gasto entre a farinha estar fluindo na torre e forno, até sair transformada em clínquer no resfriador, é estimado em 40 minutos pelos profissionais especialistas da área de processo na planta, conhecimento que foi obtido por meio da observação constante do processo. Este valor é conhecido como tempo de residência do clínquer. Baseado nessa informação, e com objetivo de promover melhor exatidão nessa estimativa do tempo de transformação de farinha em clínquer, foram coletados os valores das variáveis entre 0 e 195 minutos antes da amostra de cal livre, com passos de 15 minutos entre elas, seguindo o exemplo visto no trabalho de Magalhães (2019). Ou seja, cada variável foi coletada no instante do lançamento da cal livre, 15 minutos antes, 30 minutos antes, e assim sucessivamente até 195 minutos de antecedência. Com a obtenção da base de dados de 3 anos de operação do forno, os dados foram organizados em uma tabela que serviu de base para execução dos algoritmos de aprendizado de máquina, com 6920 amostras de cal livre e das demais variáveis de processo.

5.3 Confirmação do tempo de residência do clínquer

A primeira iteração realizada com os dados teve por objetivo buscar entender mais sobre o tempo de residência do clínquer na torre, forno e resfriador específicos da fábrica em que o trabalho foi realizado. Isso ocorre pois, como já foi explicado na seção 2, a farinha é alimentada no topo da torre (com cerca de 160 metros de altura) e vai descendo pelos ciclones enquanto as reações químicas vão ocorrendo. Ao chegar no pé da torre, a farinha já aquecida entra no forno, onde ocorre sua fusão. O material então vai escoando no forno rotativo de 62,5 metros até chegar ao resfriador, responsável pela brusca queda de temperatura do material. O clínquer percorre todo o resfriador (cerca de 40 metros) até cair em um arrastador, responsável pela transferência do material até o silo de armazenamento. Considerando que o clínquer é coletado já no arrastador, enquanto as variáveis da base de dados são medidas na torre e no forno, depreende-se que os valores obtidos de coleta online na base de dados correspondem a um estado futuro de cal livre. Por isso, a base de dados foi separada em conjuntos contendo todas as variáveis de um mesmo instante (isto é, todas com 15 minutos de defasagem, depois todas com 30 minutos, e assim sucessivamente) com a última coluna sendo sempre a cal livre do instante sem defasagem. Para facilitar o entendimento do tempo de residência, a Figura 5.1 abaixo ilustra os locais de obtenção de cada variável no forno estudado:

Figura 5.1 – Pontos de aferição das variáveis da BD.



Fonte: Do autor (2024)..

Os dados foram inseridos em uma rede neural simples, utilizando a biblioteca Keras, construída com 64 neurônios de entrada e função de ativação RELU, uma camada intermediária com 32 neurônios também com ativação RELU e uma camada de saída com 1 neurônio e ativação linear. Nenhum mecanismo de otimização foi utilizado nessa etapa. Vale ressaltar que essa iteração não tinha por objetivo obter altos índices de aderência dos resultados, mas sim verificar graficamente em qual intervalo de defasagem estava a maior representatividade dos dados da BD em relação à cal livre, para confirmar se já estavam disponíveis dados adequados para continuidade dos trabalhos. Das execuções realizadas se obtiveram os resultados apresentados na Tabela 5.3.

Tabela 5.3 – Resultados de representatividade das BDs defasadas.

Defasagem	**MSE**	**RMSE**	**R²**
15 min	0,83	0,91	-0,12
30 min	0,78	0,88	-0,05
45 min	0,73	0,85	0,019
60 min	0,73	0,85	-0,06
75 min	0,73	0,85	-0,07
90 min	0,71	0,84	-0,03
105 min	0,78	0,88	-0,05
120 min	0,85	0,92	-0,14
135 min	0,88	0,93	-0,18
150 min	0,85	0,92	-0,14
165 min	0,85	0,92	-0,23
180 min	0,82	0,91	-0,2
195 min	0,91	0,95	-0,32

Fonte: Do Autor (2024).

Com os resultados da rede neural aplicada aos vários cenários de defasagem da BD, interpretando o coeficiente de correlação (R²) percebe-se um indicativo de que o tempo de residência da farinha da fase inicial até culminar em clínquer da fase final gira em torno dos 45 minutos, corroborado pelo melhor valor observado para o R² sendo exatamente nessa defasagem, bem como pelo valor informado pelos especialistas de processo. De qualquer maneira, todos os cenários de deslocamento de tempo serão utilizados, devido a ser uma estratégia comprovadamente válida, o que foi elucidado no Capítulo 3 no detalhamento do trabalho de Magalhães (2019).

Alterando-se empiricamente as camadas de neurônios e as funções de ativação, foi possível chegar a um coeficiente de correlação de 0,15 utilizando a BD defasada de 45 minutos. As configurações da rede neural para obtenção dos resultados acima descritos estão na Tabela 5.4 que segue:

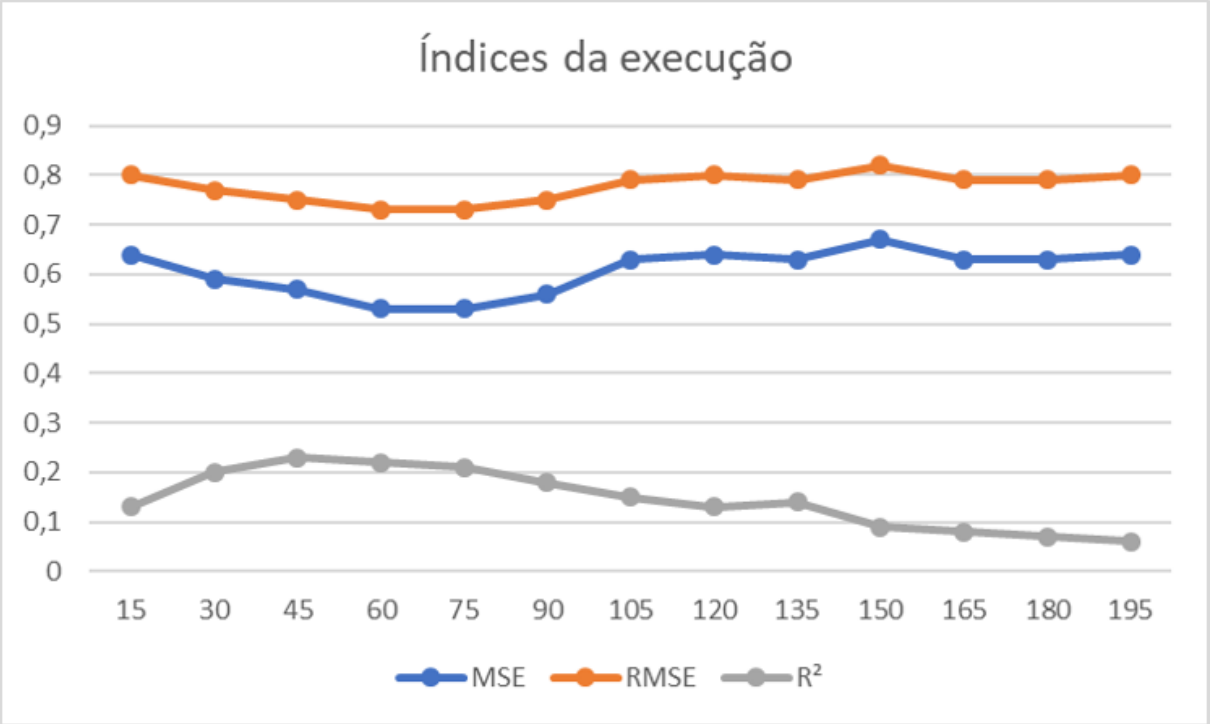
Tabela 5.4 – Configuração da rede neural.

Parâmetros	Valor
Nº épocas	150
Tamanho de lotes	64
Neurônios camada 1	64
Ativação camada 1	Relu
Neurônios camada 2	32
Ativação camada 2	Relu
Neurônios Camada 3 (saída)	1
Otimizador	Adam
Função de perda	MSE
Téc. Adicionais	Normalização, dropout
BD Teste	20%

Fonte: Do Autor (2024).

Nesse momento, os resultados foram dispostos no gráfico da Figura 5.2, cujo eixo Y é o valor dos índices e o eixo X é a defasagem da BD em minutos:

Figura 5.2 – Resultados iniciais obtidos nos cenários de defasagem.



Fonte: Do autor (2024)..

O melhor resultado, encontrado na defasagem de 45 minutos, foi de:

MSE: 0,62

RMSE: 0,79

R²: 0,15

Com os resultados acima, ficou claro que a base de dados já continha instantes de defasagem que captavam a maior representatividade dos dados de entrada perante à variável de objetivo, chancelando a passagem para etapas de otimização e testes de algoritmos para obter valores altos de coeficiente de determinação e minimizar os erros.

5.4 Aplicação de algoritmos de ML com a base de dados consolidada

Depois de analisar cada cenário de defasagem de tempo de maneira isolada, todos os dados foram reunidos em uma só tabela, sendo executado um programa de regressão linear múltipla baseado nas bibliotecas Pandas em Python, no ambiente Anaconda. Como esse algoritmo foi o mais adequado nos trabalhos de Magalhães (2019), julgou-se procedente iniciar com esse processo a fim de verificar a qualidade dos dados obtidos. O primeiro objetivo foi avaliar o comportamento variando o tamanho à proporção de dados de teste. Iniciando com 10% da base de dados até 90%, sendo que a BD era dividida cronologicamente. Ou seja, de o teste era de 10%, os primeiros 90% da base eram dados de treino e os 10% seguintes de teste.

Tabela 5.5 – Resultados de regressão linear múltipla.

Proporção Dados de teste na BD	**MSE**	**R²**
10%	0,564	0,240
20%	0,665	0,041
30%	2,038	-1,746
40%	0,601	0,183
50%	7,438	-9,184
60%	6,489	-7,592
70%	1,734	-1,273
80%	5,204	-5,850
90%	100,914	-130,952

Fonte: Do Autor (2024).

Com os resultados da Tabela 5.5 foi possível observar melhores resultados iniciais em comparação com o modelo de redes neurais, bem como o aumento da qualidade dos resultados a medida que a base de dados de treino foi aumentando. Porém, em geral os índices de MSE ainda se mantinham altos e o R² baixo (próximo de 0) ou negativo, indicando que o modelo não seria adequado para os dados fornecidos. Por isso, os esforços de otimização dos algoritmos foram interrompidos para realizar uma melhoria nos dados da base antes de continuar.

5.5 Remoção inicial de *outliers* baseada em dados de processo

Para melhorar a base de dados utilizada, era preciso definir uma abordagem a ser realizada para definir o que seria excluído ou mantido na base. Então foi proposto realizar a retirada de períodos em que a cal livre estava muito prejudicada. Em nova consulta aos especialistas da área de processo e produção da fábrica, foi unânime a opinião de que anomalias no processo de operação do forno geram consequente instabilidade nos valores de cal livre. Essas instabilidades podem ocorrer por variações na matéria prima, distúrbios no processo, falha em equipamentos ou até quedas da energia elétrica. Pensando nisso, a BD foi completada adicionando um bit de indicação se o forno estava parado ou rodando em momentos concomitantes com os lançamentos de cal livre anteriormente adicionados à base. Para iniciar os trabalhos de remoção de *outliers*, foram excluídas as linhas em que algum dos valores de cal Livre foi enviado em momentos que o forno estava parado, uma vez que as retomadas de funcionamento do forno são caracterizados como períodos de instabilidade operacional. Com isso, foram excluídas 134 linhas da base de dados, totalizando agora 6786 amostras.

Com a nova execução do programa de regressão linear com os dados filtrados somente para momentos em que o forno estava rodando, foram obtidos os resultados abaixo:

Tabela 5.6 – Resultados de regressão linear múltipla com remoção de *outliers*.

Proporção Dados de teste na BD	**MSE**	**R^2**
10%	0,5	0,342
20%	0,519	0,286
30%	0,531	0,267
40%	0,548	0,268
50%	0,554	0,254
60%	0,566	0,229
70%	0,59	0,199
80%	3,748	-4,052
90%	13,480	-17,136

Fonte: Do Autor (2024).

Com essa nova execução foi possível observar a melhor aderência dos modelos conforme os dados de treino foram aumentando, bem como o melhor desempenho continuou sendo observado com os dados de teste em 10% da base. Por isso, o trabalho prosseguiu com os dados de teste fixados em 10% e foram realizadas novas iterações com modelos de ML para entender

a melhoria apropriada aos resultados mediante aos processos de remoção de *outliers* antes de filtrar ainda mais a BD.

5.6 Execução de novas iterações com BD filtrada pelo bit de forno rodando

O trabalho prosseguiu com os dados de teste fixados em 10% e foram aplicados códigos de regressão linear múltipla, regressão Ridge e Lasso, a fim de verificar a possibilidade de melhoria dos resultados com estas abordagens de regularização de dados por meio da análise de importância das variáveis, obtendo-se os resultados abaixo:

Tabela 5.7 – Teste de metodologias de regressão Linear, Ridge e Lasso.

Modelo de regressão	**MSE**	**R^2**
Linear	0,5	0,342
Ridge	0,493	0,35
Lasso	0,613	0,193

Fonte: Do Autor (2024).

Foi adotada uma busca iterativa pelo melhor hiperparâmetro *alpha* da regressão Lasso em busca de melhorias dos seus resultados. Com o valor ajustado em 0,01 foram obtidos os seguintes resultados:

MSE: 0,49

R^2: 0,354

A busca iterativa pelo melhor ajuste da regressão de Lasso demandou muito tempo de execução e o resultado tido como melhor foi basicamente o mesmo da regressão Ridge sem nenhum esforço de otimização. Por isso, ela foi excluída e substituída pela metodologia de Gradient Boosting, obtendo-se o melhor resultado até então:

Tabela 5.8 – Comparação dos melhores resultados obtidos entre modelos de regressão Linear, Ridge e *Gradient Boosting*.

Modelo de regressão	**MSE**	**R^2**
Linear	0,5	0,342
Ridge	0,493	0,35
Gradient boosting	0,416	0,452

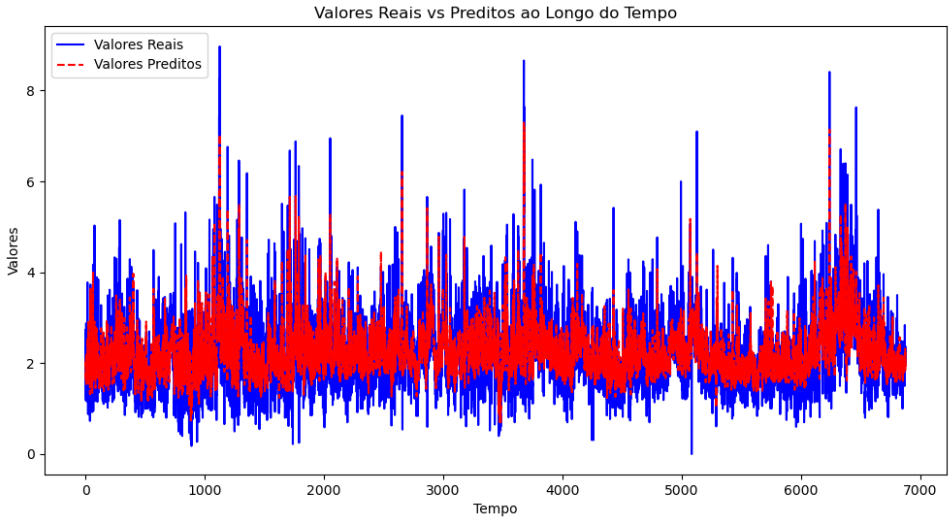
Fonte: Do Autor (2024).

Com as iterações e testes realizados acima já foi possível observar a melhoria dos resultados dos coeficientes, iniciada pela remoção de alguns *outliers* da base. Dessa forma, podem

ser aplicadas metodologias de ajustes dos hiperparâmetros dos algoritmos em busca de otimizar ainda mais as respostas dos modelos.

Com a implementação preliminar de códigos de aprendizado de máquina, foi possível corroborar a possibilidade de se realizar a predição de cal livre com base nas informações obtidas *online* da cadeia de forno de clínquer. Os melhores resultados até então foram obtidos nos algoritmos de regressão linear multivariável e também utilizando a metodologia *Gradient Boosting*, obtendo-se o gráfico abaixo, que ilustra a comparação entre os valores reais da base de dados e os preditos pelo modelo:

Figura 5.3 – Resultados preliminares obtidos com metodologia *Gradient Boosting*.



Fonte: Do autor (2024)..

Antes de voltar à adoção de técnicas para melhoria da base de dados, foram realizados testes também com a metodologia LightGBM, que é uma variante otimizada do Gradient Boosting, uma vez que ela vinha apresentando os melhores resultados entre todos os que foram obtidos. Após várias iterações, os melhores resultados obtidos estão dispostos a seguir:

Tabela 5.9 – Comparação dos melhores resultados obtidos entre as metodologias *Gradient Boosting* e *LightGBM*.

Modelo de regressão	**MSE**	**R^2**
Gradient boosting	0,416	0,452
Light GBM	0,392	0,483

Fonte: Do Autor (2024).

Dessa forma, preliminarmente ficou validada a estratégia de predição e a seleção que foi realizada de variáveis, ficando como próximos passos o aprimoramento das metodologias de remoção de *outliers* com base nos parâmetros operacionais de processo ou também pela

avaliação dos resultados realizando uma triagem das variáveis por meio dos coeficientes de correlação de Pearson e Spearman.

5.7 Remoção inicial de *outliers* baseada nos coeficientes de Pearson e Spearman

Dando sequência aos trabalhos, foram calculados os coeficientes de Pearson e Spearman das variáveis em comparação com a cal livre medida. Na definição das variáveis, os itens Cal livre -1, -2, -3 e -4 referem-se aos lançamentos consecutivamente anteriores ao lançamento atual de cada linha. Ainda, as variáveis iniciadas por xxmin\ definem a coleta de cada variável defasada em xx minutos. As Tabelas 5.10 e 5.11 abaixo exibem os 20 maiores coeficientes:

Tabela 5.10 – Maiores coeficientes de Pearson entre as variáveis da BD.

Variável	**Módulo de coef. Pearson**
Cal livre -1	0,315339
Cal livre -2	0,26144
75min\Taxa de combustível no pré-calcinador	0,252142
90min\Taxa de combustível no pré-calcinador	0,252142
60min\Taxa de combustível no pré-calcinador	0,23251
Cal livre -3	0,224885
Cal livre -4	0,215532
60min\NOx na entrada do forno	0,18572
105min\Taxa de combustível no pré-calcinador	0,17497
60min\Rotação do forno	0,171568
60min\Temperatura na zona de queima	0,168757
45min\Rotação do forno	0,164875
150min\Taxa de combustível no pré-calcinador	0,164471
190min\Taxa de combustível no pré-calcinador	0,164112
75min\Rotação do forno	0,163831
90min\Rotação do forno	0,163831
120min\Taxa de combustível no pré-calcinador	0,162987
135min\Taxa de combustível no pré-calcinador	0,162263
15min\Rotação do forno	0,158514
30min\Rotação do forno	0,157606

Fonte: Do Autor (2024).

Tabela 5.11 – Maiores coeficientes de Spearman entre as variáveis da BD.

Variável	Módulo de coef. Spearman
Cal livre -1	0,325419
60min\Temperatura na zona de queima	0,29465
75min\Temperatura na zona de queima	0,276218
90min\Temperatura na zona de queima	0,276218
Cal livre -2	0,268724
Cal livre -3	0,238491
Cal livre -4	0,231327
45min\Temperatura na zona de queima	0,176956
60min\NOx na entrada do forno	0,174527
30min\Temperatura na zona de queima	0,167304
120min\Temperatura na zona de queima	0,161214
150min\Temperatura na zona de queima	0,160905
105min\Temperatura na zona de queima	0,160294
135min\Temperatura na zona de queima	0,158533
15min\Temperatura na zona de queima	0,149992
75min\Taxa de combustível no pré-calcinador	0,148838
90min\Taxa de combustível no pré-calcinador	0,148838
75min\NOx na entrada do forno	0,148685
90min\NOx na entrada do forno	0,148685
60min\Taxa de combustível no pré-calcinador	0,147661

Fonte: Do Autor (2024).

Depois do cálculo dos coeficientes, foram realizadas iterações com as bases de dados excluindo os coeficientes menores que 0,1. No trabalho de Loca e Rauber (2019), a medida que os coeficientes menores foram sendo excluídos, os resultados do modelo foram melhorando, e a intenção foi verificar a existência ou não desse comportamento no cenário do presente trabalho. A tabela que segue mostra os resultados obtidos:

Tabela 5.12 – Iterações realizadas com filtro de coeficientes de Pearson e Spearman.

Metodologia	MSE	R^2
Gradient Boosting BD Original	0,416	0,452
Gradient Boosting BD Pearson > 0,1	0,434	0,429
Gradient Boosting BD Spearman > 0,1	0,433	0,429
Light GBM BD Original	0,392	0,483
Light GBM BD Pearson > 0,1	0,447	0,412
Light GBM BD Spearman > 0,1	0,441	0,419

Fonte: Do Autor (2024).

Verificando os resultados dispostos na Tabela 5.12, ficou claro que a estratégia de diminuir a base de dados excluindo variáveis com menores coeficientes de correlação não se mostrou uma boa abordagem, uma vez que os índices de erro oscilaram para cima enquanto o R^2 oscilou para baixo, indicando menor qualidade dos resultados. Dessa maneira, não foi possível

perceber a mesma melhora obtida no trabalho de Loca e Rauber (2019). Vale ressaltar que os autores aplicaram modelos baseados em simulação de processos químicos, enquanto o presente trabalho trata de um processo real.

Da análise das Tabelas 5.10 e 5.11, foi possível perceber que os coeficientes não indicam grande correlação entre as variáveis de entrada e a variável cal livre, o que já é um indicativo das dificuldades encontradas para melhorar o resultado dos modelos experimentados. Além disso, o coeficiente de Pearson demonstrou maior ocorrência de maiores coeficientes da taxa de combustível no calcinador, enquanto o de Spearman demonstrou maior incidência da temperatura da zona de queima. Para melhor entender estes resultados, foi preciso recorrer novamente aos especialistas que detém conhecimento detalhado acerca da química do processo.

A influência do combustível alternativo na cal livre já havia sido mencionada no Tópico 2.3. Dessa forma, os resultados apontados pelo coeficiente de Pearson corroboram com essa influência e unem os princípios de correlação e causalidade, uma vez que a correlação foi elucidada pelo coeficiente ao mesmo tempo que as variações de combustível alterativo realmente causam as perturbações da cal livre, existindo a causalidade neste processo.

Já interpretando os resultados do coeficiente de correlação de Spearman, ficou destacada a presença da temperatura da zona de queima. De fato, quando a temperatura sobe, a cal livre sobe, e quando cai, o mesmo ocorre com a cal livre. Assim, existe de fato uma correlação. Porém, a temperatura da zona de queima é uma consequência dentro do processo, uma vez que ela pode ser influenciada pelo aumento ou diminuição da taxa de combustíveis ou resíduos no processo, ou também pela quantidade de farinha alimentada. Nesse caso, portanto, existe uma correlação sem causalidade, considerando o fato da temperatura ser consequência no processo.

A partir dessa análise, entende-se que o coeficiente de correlação de Pearson conseguiu capturar melhor as correlações químicas do processo pelo indicativo de causalidade, de forma a dar mais ênfase aos reais causadores de variações da qualidade do clínquer, medidas pela cal livre. Porém, como os modelos não apresentaram melhora com os filtros de coeficientes, essa abordagem não foi mais utilizada.

5.8 Avaliação detalhada da BD e desenvolvimento de novas estratégias de remoção de *outliers*

Depois das tentativas sem sucesso de aplicação dos coeficientes de correlação, foi realizada uma análise da composição dos valores de cada variável selecionada para utilização, a fim de entender os maiores desvios e oportunidades de tratamento nos valores. Com isso, foi construída a Tabela 5.13 abaixo:

Tabela 5.13 – Composição das variáveis da BD.

Variável	Mínimo	Médio	Máximo	Unidade
Temperatura na entrada do forno	0,00	1054,69	1243,66	ºC
Pressão na entrada do forno	0,00	1,77	13,48	mbar
NOx na entrada do forno	0,00	1144,42	2704,81	ppm
CO na entrada do forno	0,00	0,01	4,02	%
O2 na entrada do forno	0,00	5,57	26,06	%
Temperatura de saída da torre de ciclones	42,68	113,32	245,28	ºC
Temperatura do 6º estágio A da torre	384,20	879,37	1102,22	ºC
Temperatura do 6º estágio B da torre	137,04	881,71	1035,73	ºC
Temperatura calcinador	430,13	911,79	1351,21	ºC
Consumo térmico	0,10	830,53	104430,03	kcal
Taxa de combustível no forno	0,01	7,86	13,01	ton/h
Taxa de resíduo no forno	0,00	1,12	5,11	ton/h
Taxa de combustível no pré-calcinador	0,00	10,08	19,97	ton/h
Taxa de resíduo no calcinador	0,00	10,50	21,05	ton/h
Alimentação de farinha no forno	0,00	330,32	375,67	ton/h
Temperatura na zona de queima	650,65	1172,98	1528,03	ºC
Torque do forno	0,58	48,62	76,35	%
Rotação do forno	0,00	3,76	4,11	rpm
Cal livre	0,00	2,21	8,97	%

Fonte: Do Autor (2024).

Como a abordagem dos coeficientes de determinação não funcionou bem para essa aplicação, foi definido outro critério de limpeza da BD baseando-se na taxa de alimentação do forno. Em momentos de instabilidade do processo, como em falhas de combustível, problemas em algum equipamento da cadeia ou retomadas de operação, a alimentação do forno sempre é reduzida para manter mínima estabilidade operacional do processo. Por isso, da BD extraída foram excluídas as linhas com alimentação abaixo de 290 ton/h, depois abaixo de 300, 310 e 320 ton/h. Estes valores foram escolhidos empiricamente e considerando que a alimentação nominal do forno é de 330 ton/h. Para as bases de dados extraídas foram obtidos os resultados a seguir:

Tabela 5.14 – Resultados com filtros de alimentação do forno.

Critério	Nº Amostras	MSE	R^2
Alimentação > 290 ton/h	6407	0,378	0,431
Alimentação > 300 ton/h	6135	0,348	0,46
Alimentação > 310 ton/h	5814	0,335	0,465
Alimentação > 320 ton/h	4934	0,33	0,413

Fonte: Do Autor (2024).

Com os dados da tabela 5.13 ficou clara a grande variação dos valores de todas as grandezas selecionadas no trabalho, bem como as tentativas de melhorar os resultados mostradas na tabela 5.14 não surtiram grandes efeitos. Por isso foi tomada a decisão de construir uma nova base de dados com uma melhor metodologia de extração do *PI Data Link*, para depois retornar aos processos de limpeza da base de dados.

5.9 Obtenção de nova base de dados de processo

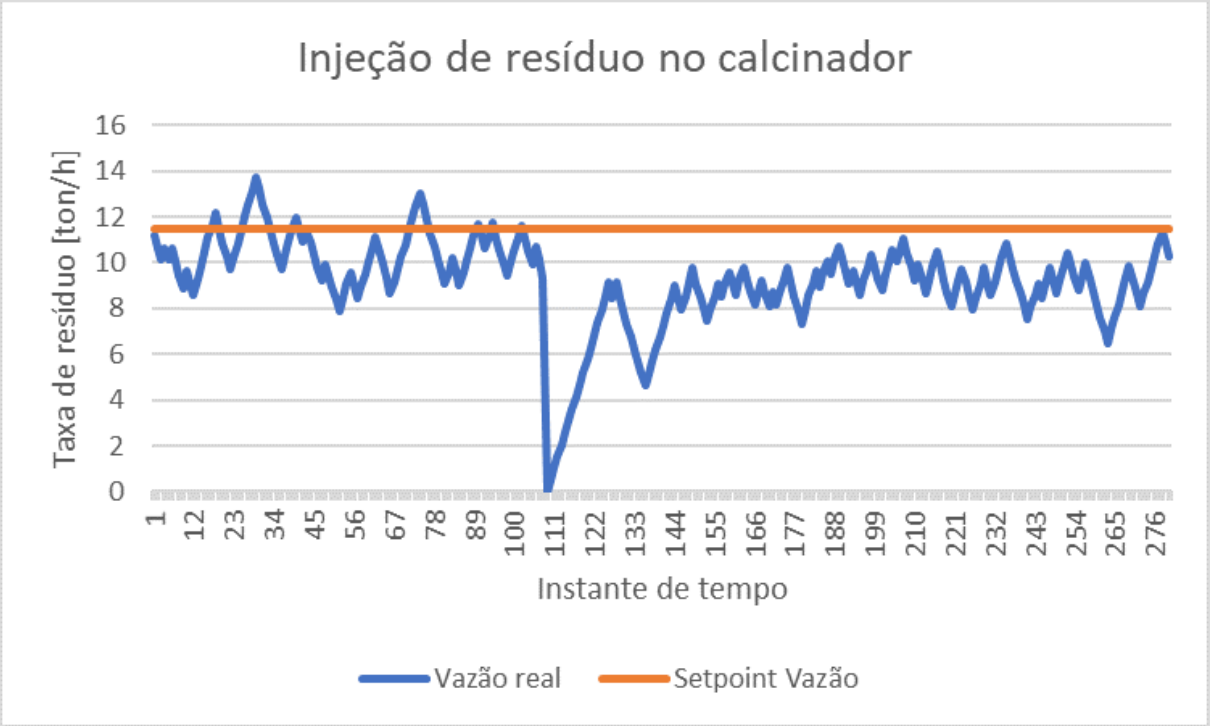
Na primeira BD construída, foram coletados os valores instantâneos das variáveis nos momentos de defasagem definidos. Porém, o valor instantâneo pode não representar a dinâmica de operação daquele período, pois todas as variáveis tem uma característica oscilatória comum. Além disso, foram extraídos todos os valores disponíveis de cal livre para o período escolhido, sem nenhuma validação dos especialistas quanto à qualidade das medições.

A nova base de dados construída teve a primeira modificação na escolha das medições de cal livre. A aferição dessa grandeza pode ser feita em dois aparelhos de raio-X, um que utiliza o princípio de fluorescência e outro o princípio de difração. Entre os dois, o raio-X de difração é bem mais preciso, fornecendo valores mais coerentes. Por isso, foi definido extrair somente medições realizadas por este equipamento. Além disso, é unânime entre os especialistas a opinião de que o processo é muito dinâmico no sentido de que frequentemente são modificadas metodologias de operação que, consequentemente, têm seus impactos nos indicadores de qualidade. Por isso, a montagem de uma BD de 3 anos acaba englobando várias mudanças de estratégia, o que prejudica as análises. Portanto, a nova BD englobou somente indicadores de somente 8 meses de operação, período que foi escolhido devido à disponibilidade de uma base de dados detalhada das análises de cal livre pelo raio X de difração, bem como extraiu exclusivamente as medições de cal livre feitas por este equipamento de análise, totalizando 1731 amostras.

Escolhidas as amostras de cal livre, o próximo passo foi definir os intervalos de defasagem a serem utilizados para coleta das variáveis de processo. Baseado na figura 5.2, ficou

definido ir até a defasagem máxima de 75 minutos, entendendo que a maior representatividade da BD está nesse intervalo. Ainda, quanto à natureza oscilatória das informações de processo, a Figura 5.4 mostra duas questões relevantes. A primeira é a diferença entre o setpoint(objetivo) de injeção de combustível alternativo e a taxa de injeção efetivamente alcançada ao longo do tempo, o que explica o uso das duas informações na base de dados do trabalho. Ainda, é possível visualizar o quando a injeção oscila ao longo do tempo por questões inerentes ao dosador utilizado.

Figura 5.4 – Variação dos registros de vazão de combustível alternativo no calcinador.



Fonte: Do autor (2024)..

Por isso, ao invés de adotar novamente a extração do valor instantâneo das variáveis ao longo dos instantes de defasagem definidos, optou-se por extrair um valor médio referente a 5 minutos de operação de cada variável dentro do intervalo de defasagem definido de 75 minutos. Ou seja, foi extraída a média de cada variável entre o horário de lançamento da cal livre e 5 minutos antes, entre 5 e 10 minutos, 10 e 15 minutos, e assim sucessivamente até 70 a 75 minutos de defasagem. Com essa estratégia objetivou-se ter menor variação das grandezas selecionadas.

5.10 Execução de algoritmos de *Light GBM* com a nova BD

Com a construção da nova base de dados, foram realizadas iterações excluindo novamente as taxas de alimentação do forno para acompanhamento dos resultados do modelo *Light GBM*, que foram dispostos abaixo:

Tabela 5.15 – Resultados com filtros de alimentação do forno e BD de médias.

Critério	Nº Amostras	MSE	R^2
BD Original	1731	0,439	0,397
Alimentação > 290 ton/h	1609	0,391	0,443
Alimentação > 300 ton/h	1497	0,359	0,479
Alimentação > 310 ton/h	1397	0,229	0,495
Alimentação > 320 ton/h	1181	0,336	0,479

Fonte: Do Autor (2024).

A tabela 5.15 demonstrou o melhor desempenho do modelo na base de dados nova e com restrição de alimentação do forno maior que 310 ton/h. Porém, os resultados ainda precisavam melhorar para chegar ao menos nos que foram observados na seção de trabalhos relacionados, o que não estava sendo alcançado com exaustivos ajustes dos hiperparâmetros dos modelos. Por isso, iniciou-se a etapa de geração de dados sintéticos para a base de dados. Desse ponto em diante, foi utilizada sempre a base de dados filtrada pela alimentação do forno maior que 310 ton/h.

5.11 Testes de algoritmos de geração de dados sintéticos

Para geração de dados sintéticos para o trabalho, foram aplicadas algumas das metodologias aplicadas a dados reais de processo, como foi visto no tópico 2.10. A fim de garantir igualdade de execução dos códigos, foram definidos os seguintes parâmetros para o modelo *Light GBM* executado. Os parâmetros utilizados foram:

Tabela 5.16 – Parâmetros do modelo LightGBM.

Parâmetro	Valor
Número de árvores	1000
Taxa de aprendizado	0,01
Número de folhas das árvores	40
Profundidade máxima	5
Peso mínimo das instâncias	10
Regularização Lasso	0,1
Regularização Ridge	0,1

Fonte: Do Autor (2024).

Inicialmente foi implementado um algoritmo de geração de dados sintéticos por meio de interpolação, aumentando a base de dados em 20x. A interpolação realizada foi linear, gerando 5 novas amostras entre 2 pontos da base de dados real. Para este caso, os resultados obtidos foram dispostos na tabela 5.17.

Tabela 5.17 – Comparativo de resultados com dados sintéticos via interpolação.

Construção da BD	Validação	MSE	RMSE	R^2
BD conjunta (real + sintético)	*Holdout*	0,097	0,312	0,669
Treino com BD Sintética e teste BD real	*Holdout*	0,186	0,432	0,685
BD conjunta (real + sintético)	5 k-folds	0,079	0,281	0,744
BD real	5 k-folds	0,043	0,207	0,927

Fonte: Do Autor (2024).

A utilização de dados sintéticos conseguiu impor uma melhora consistente nos resultados, corroborada também pela iteração que usou a base sintética para treino e a real para teste. Aliado aos dados sintéticos, a metodologia de validação K-folds também possibilitou a melhora dos resultados com a sua aplicação. No entanto, ao retirar a geração de dados sintéticos por interpolação e manter só a validação K-folds, que ainda não havia sido experimentada anteriormente, os resultados melhoraram expressivamente, o que demonstrou que a estratégia de geração de dados sintéticos por interpolação não estava adequada para a base de dados.

Então, foi implementado um código de geração de dados sintéticos por meio da inserção de ruído gaussiano com desvio padrão de 0,01, tendo sido obtidos os resultados da tabela 5.18 referentes à execução de algoritmo em uma base de dados única composta pelos dados reais e sintéticos:

Tabela 5.18 – Resultados com dados sintéticos de ruído gaussiano e validação k-folds 5 dobras.

Multiplicação da BD	Desvio padrão	MSE	RMSE	R^2
5x	0,01	0,022	0,147	0,963
10x	0,01	0,021	0,144	0,965
20x	0,01	0,021	0,146	0,964
50x	0,01	0,021	0,145	0,965
50x	0,001	0,021	0,144	0,965
50x	0,1	0,02	0,141	0,966
150x	0,1	0,02	0,14	0,967

Fonte: Do Autor (2024).

Com os resultados acima descritos ficou claro o melhor ajuste que a técnica de geração de dados sintéticos por meio de ruído gaussiano deu ao modelo, obtendo-se um erro na casa de 0,14 para a cal livre, que em regime normal de operação gira em torno de 2. Este erro médio representa 7% do valor mais comum para a cal livre. Ainda, o valor de R^2 em torno de 0,96

confirma que o modelo, aliado ao erro baixo, conseguiu captar bem as variações da cal livre por meio das variáveis que foram fornecidas como dados de entrada.

Com isso, para o cenário de geração de dados sintéticos de 20x o tamanho da BD real, aliado à metodologia de validação dos 5 K-folds, foi obtida a Tabela 5.19 que segue:

Tabela 5.19 – Comparação de resultados com validação 5 K-folds e dados sintéticos de 20x a BD real.

Construção da BD	**MSE**	**RMSE**	**R^2**
BD real	0,043	0,207	0,927
BD conjunta (real + 20x sintética via interpolação)	0,079	0,281	0,744
BD conjunta (real + 20x sintética via ruído Gaussiano)	0,01	0,021	0,964

Fonte: Do Autor (2024).

A fim de confirmar a qualidade dos dados gerados, uma última iteração foi realizada utilizando somente a base de dados sintética para treinamento do modelo e realizando o teste com a base de dados real. Os resultados foram:

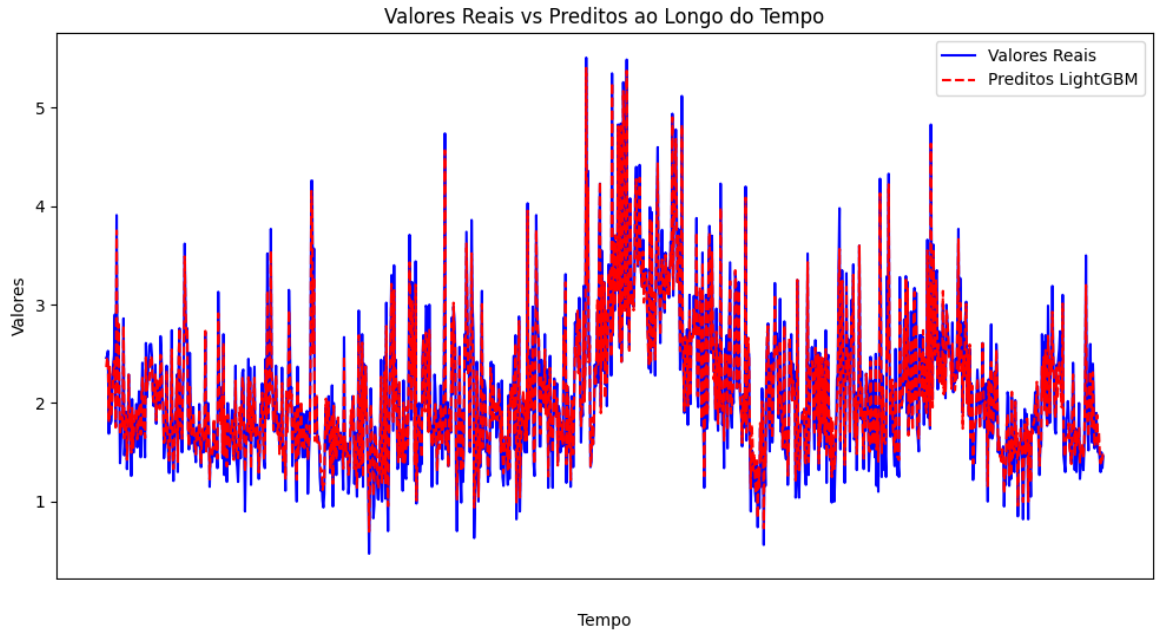
Tabela 5.20 – Validação das metodologias aplicadas.

Cenário	**Desvio padrão**	**MSE**	**RMSE**	**R^2**
Treino e teste com BD única (real + sintético)	0,1	0,02	0,141	0,966
Treino BD sintética e teste BD real	0,1	0,02	0,143	0,965

Fonte: Do Autor (2024).

Desse último teste foi obtido o gráfico disposto na Figura 5.5, que demonstra a aderência dos dados preditos aos dados reais de processo e confirmam a qualidade do modelo que foi construído.

Figura 5.5 – Resultado final obtido.



Fonte: Do autor (2024)..

Com a comparação dos dois cenários dispostos na Tabela 5.20, foi possível perceber que os dados sintéticos gerados conseguiram captar as variações da base real sem prejudicar os resultados de MSE, RMSE e R^2 do modelo, corroborando que foi possível de fato captar as variações da cal livre por meio de dados históricos do processo do forno.

5.12 Considerações Finais

A aplicação de ferramentas de aprendizado de máquina para as bases de dados criadas no presente trabalho mostrou-se capaz de predizer com bons índices de exatidão o valor da cal livre de um processo de forno de clínquer. Vale ressaltar que a escolha de montar a base de dados somente com variáveis que são medidas online torna o modelo mais robusto pela mínima interferência humana que é possível de ocorrer nas variáveis, o que também contribui na robustez da predição. Ainda, a complementação da BD com dados sintéticos foi a ferramenta que permitiu melhorar expressivamente os resultados dos algoritmos, bem como permite simplificar a etapa de obtenção de dados de processo do ponto de vista de representatividade e volume de dados. Com isso, abre-se um cenário de possibilidade de aprendizado mais rápido, com menor esforço para levantamento das bases de dados reais sem prejudicar os indicadores de qualidade dos modelos. Quanto aos resultados obtidos, confirma-se a possibilidade de predição da cal livre por meio de dados de processo, tornando factível a criação de uma ferramenta proativa de

indicação da cal livre, de maneira a possibilitar um controle mais rígido das suas variações, que são inerentes ao processo produtivo.

6 Conclusões

Da análise dos resultados do trabalho, foi possível corroborar com a estratégia de predição de cal livre baseada em dados online de processo de produção de clínquer. O estudo demonstrou a aplicabilidade de algoritmos de aprendizado de máquina no ambiente industrial como forma de desenvolvimento de ferramentas proativas a serem aplicadas no meio fabril.

O conhecimento prévio do processo de produção de clínquer e cimento possibilitou melhor entendimento dos resultados de aplicação de modelos de aprendizado de máquina. A familiaridade com o processo e com as variáveis associadas permitiu que a interpretação dos padrões identificados pelos modelos fosse mais fácil, além de gerar melhor compreensão da relevância real de cada variável para o desempenho dos modelos. Esse conhecimento prévio foi, portanto, fundamental para validar os resultados e interpretá-los de forma mais crítica e eficaz, tornando o processo de análise menos complicado.

No que tange às metodologias de regressão aplicadas, foi possível observar a grande influência nos resultados causada pela modificação de hiperparâmetros dos códigos e também pela alteração da base de dados (como no caso da remoção de *outliers*). Por isso, é importante realizar um mapeamento sistemático das iterações realizadas para melhor rastreabilidade das ações com retorno positivo ou negativo nos resultados.

As metodologias de geração de dados sintéticos foram decisivas para o êxito do trabalho dada a grande contribuição que tiveram para melhoria dos índices MSE, RMSE e R^2 dos modelos treinados. Além da contribuição específica no presente estudo, seus benefícios são ainda mais claros quando o processo de construção das bases de dados é caro, trabalhoso ou trata de dados sensíveis.

Além dos dados sintéticos, as metodologias de validação também são parte muito importante para o resultado final do modelo. A escolha da melhor metodologia pode se dar de forma empírica, observando a mudança de resultados mediante a aplicação das diferentes técnicas de validação.

No que tange à interoperabilidade do modelo criado para ser usado em outras plantas de cimento, alguns fatores precisam ser considerados. Em todos os cenários de fábrica de cimento, as variáveis que foram escolhidas para compor a base de dados são de natureza online, o que facilita a implementação do modelo em qualquer unidade fabril. Porém, o cenário de

coprocessamento de resíduos é muito desafiador ao aproveitamento dos modelos por outras plantas devido ao alto impacto na química do processo de produção de clínquer.

Ainda, é inerente ao processo de produção de clínquer e cimento a mudança periódica de modelos de operação, de matérias primas, limites de qualidade, dentre outras variáveis que acabam modificando a dinâmica das reações químicas e de todo o processo que ali ocorre. Dessa forma, os modelos de predição criados precisam ser retroalimentados constantemente de maneira a captar a realidade mais recente do modo de operação adotado na planta. Essa retroalimentação precisa vir desde as primeiras execuções dos modelos novamente, passando pela modificação de hiperparâmetros conforme necessário, para garantir sempre a minimização dos erros de estimação e a qualidade da predição como um todo. Caso não ocorra este processo periódico de retroalimentação, certamente o modelo passará a apresentar maiores erros de estimação no decorrer do tempo. Associando essa questão aos dados sintéticos, mesmo que no presente trabalho não existiam questões de custo ou grandes dificuldades na montagem da base de dados, há de se considerar o aspecto das mudanças de modo de operação da fábrica. Por exemplo, se a cada mudança de metodologia de operação fosse preciso esperar algumas semanas para criação de uma BD que represente o novo modelo, uma ferramenta de predição acabaria caindo em descrédito. Por isso, com uma base de poucos dias de operação, somada aos dados sintéticos, rapidamente seria possível gerar um novo modelo aderente nos resultados de predição.

Como trabalhos futuros, algumas perspectivas se destacam após a avaliação do presente estudo. A primeira perspectiva é a dos combustíveis alternativos. Por serem uma realidade cada vez mais presente nas fábricas, aliado à necessidade de coprocessar quantidades cada vez maiores de resíduos, pode se obter bons resultados analisando aspectos individuais da qualidade dos resíduos de maneira a prever o impacto da sua queima nos fornos de clínquer. Outra abordagem importante é a operacionalização da predição da cal livre dentro do sistema de automação das plantas. O desenvolvimento de uma ferramenta capaz de acessar diretamente as bases de dados, predizer o valor da cal livre e escrevê-lo dentro do ambiente de automação (como dentro de um CLP ou de um servidor OPC) completa o trabalho que foi aqui realizado, uma vez que, dessa maneira, os resultados da predição estarão visíveis aos operadores para nortear efetivamente sua atuação no processo.

REFERÊNCIAS

(ABCP), A. B. de C. P. **A nova norma de especificação de cimento ABNT NBR 16697: saiba o que mudou e o que não mudou**. 2018. <https://abcp.org.br/a-nova-norma-de-especificacao-de-cimento-abnt-nbr-16697-saiba-o-que-mudou-e-o-que-nao-mudou/>.

Amazon Web Services. **What is a Neural Network?** 2023. Accessed: 2024-08-25. Disponível em: <https://aws.amazon.com/pt/what-is/neural-network/>.

ANDREWS, G. **O que são Dados Sintéticos?** 2021. Acessado em Abril de 2024. Disponível em: <https://blog.nvidia.com.br/blog/o-que-sao-dados-sinteticos/>.

ANTONELLI, G. C.; NEITZEL, I. Aplicação de redes neurais artificiais na indústria de fios de algodão: Deterimnação do índice de fibras imaturas. **Revista Gestão Industrial**, v. 11, n. 2, 2015. Disponível em: <https://periodicos.utfpr.edu.br/revistagi/article/view/2662/2169>.

ANTUNES, V. d. S. F. **Proposta de Metodologia para Otimização da Variabilidade em um Parque Híbrido Eólico-Fotovoltaico Utilizando o Método de Monte Carlo e Programação Não-Linear**. 2020.

ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. **Statistics Surveys**, Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada, v. 4, n. none, p. 40 – 79, 2010. Disponível em: <https://doi.org/10.1214/09-SS054>.

AVEVA. **PI System: AVEVA World 2024**. 2024. <https://events.aveva.com/aw-2024/pi-system>.

AZULAY, S. et al. **Holdout SGD: Byzantine Tolerant Federated Learning**. 2020. Disponível em: <https://arxiv.org/abs/2008.04612>.

BIRNIE, C.; RAVASI, M. Geometry-independent realistic noise models for synthetic data generation. In: EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS. **82nd EAGE Annual Conference & Exhibition**. [S.l.], 2021. v. 2021, n. 1, p. 1–5.

BREIMAN, L. Random forests. **Machine Learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

CARVALHO, A. Análise de regressões lineares e redes neurais artificiais: Um estudo comparativo de ténicas de inteligência artificial aplicadas na previsão de preço das ações da nvidia. **Apoena Revista Eletrônica**, 2023.

CESSIE, S. l.; HOUWELINGEN, J. V. Ridge estimators in logistic regression. **Journal of the Royal Statistical Society Series C: Applied Statistics**, Oxford University Press, v. 41, n. 1, p. 191–201, 1992.

CHAI, T.; DRAXLER, R. R. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. **Geoscientific model development**, Copernicus Publications Göttingen, Germany, v. 7, n. 3, p. 1247–1250, 2014.

CNI. **Indústria brasileira de cimento: base para a construção do desenvolvimento**. 2017. Acessado em março de 2024. Disponível em: <https://static.portaldaindustria.com.br/media/filer_public/d2/4b/d24b904d-c672-43f5-a0b7-9644ad97091a/abcp.pdf>.

CORTES, C. Support-vector networks. **Machine Learning**, 1995.

FILHO, D. F. F. et al. Análise pluviométrica no estado do pará: comparação entre dados obtidos de estações pluviométricas e do satélite gpcc. **Revista Brasileira de Climatologia**, v. 26, 2020.

FLECK, L. et al. Redes neurais artificiais: Princípios básicos. **Revista Eletrônica Científica Inovação e Tecnologia**, v. 1, n. 13, p. 47–57, 2016.

FLSMIDTH. **QCX/RoboLab® Systems**. 2024. <https://www.flsmidth-cement.com/products/qcx-robolab-systems>.

FONSECA, A. U. et al. Diagnosticando tuberculose com redes neurais artificiais e recursos bppc. **Journal of Health Informatics**, v. 15, n. Especial, 2023.

FREITAS, N. C. et al. Detecçao de outliers em finanças de instituições de ensino superior brasileiras utilizando aprendizado de máquina nao supervisionado. In: SBC. **Anais do XXXIV Simpósio Brasileiro de Informática na Educação**. [S.l.], 2023. p. 1489–1500.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189 – 1232, 2001. Disponível em: <https://doi.org/10.1214/aos/1013203451>.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. **International journal of information management**, Elsevier, v. 35, n. 2, p. 137–144, 2015.

GIRELLI, C.; CORSO, L. Detecção de falhas em contentores plásticos utilizando redes neurais convolucionais. **Scientia cum Industria**, v. 8, p. 156–163, 10 2020.

GOIS, G. A. et al. Redes neurais artificiais para predição do consumo total de combustível de um alto-forno. **Tecnologia em metalurgia, materiais e mineração**, v. 16, n. Especial, 2019. ISSN 2176-1515.

GOODFELLOW, I. **Deep learning**. [S.l.]: MIT press, 2016. v. 196.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2001.

HEWLETT, P. C. (Ed.). **Lea's Chemistry of Cement and Concrete**. 5th. ed. [S.l.]: Butterworth-Heinemann, 2019.

HWANG, J.; LEE, J.; LEE, K.-S. A deep learning-based method for grip strength prediction: Comparison of multilayer perceptron and polynomial regression approaches. **PLOS ONE**, Public Library of Science, v. 16, n. 2, p. 1–12, 02 2021. Disponível em: <https://doi.org/10.1371/journal.pone.0246870>.

IOFFE, S.; SZEGEDY, C. **Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift**. 2015. Disponível em: <https://arxiv.org/abs/1502.03167>.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015.

JUNIOR, A. C. d. S. et al. Geração de dados sintéticos para classificação de disléxicos por meio de aprendizado de máquina. **Journal of Health Informatics**, v. 13, n. 1, mar. 2021. Disponível em: <https://www.jhi.sbis.org.br/index.php/jhi-sbis/article/view/764>.

JúNIOR, J. E. B.; SETTI, J. R. Produção de dados de tráfego sintéticos através de algoritmo genético e simulação microscópica. **TRANSPORTES**, v. 18, n. 3, jul. 2010. Disponível em: <https://www.revistatransportes.org.br/anpet/article/view/447>.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2012. v. 25. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.

LEA, F. M. **The Chemistry of Cement and Concrete**. 4th. ed. [S.l.]: Arnold, 1970.

LI, W. et al. An improved multi-source based soft sensor for measuring cement free lime content. **Information Sciences**, v. 323, 06 2015.

LOCA, A.; RAUBER, T. Uso de uma rede neural convolucional unidimensional para detecção de falhas em processos industriais. In: SBC. **Anais da XIX Escola Regional de Computação Bahia, Alagoas e Sergipe**. [S.l.], 2019. p. 42–47.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. **Estudos Avançados**, SciELO Brasil, v. 35, p. 85–94, 2021.

MAGALHãES, R. d. S. L. Desenvolvimento de sensor virtual para predição do teor de cal livre no clínquer em uma fábrica de cimento. **Dissertação de Mestrado, Universidade Federal de Minas Gerais**, 2019. Disponível em: <https://repositorio.ufmg.br/bitstream/1843/RAOA-BCTLSR/1/disserta__o_de_mestrado___roberta_de_souza_lima_magalhaes.pdf>.

MAZUR, D. L.; DEMITO, M. L.; WATANABE, E. R. L. d. R. Avaliação de óleo residual oriundo da fabricação de latas destinado a coprocessamento em fornos de cimento e os efeitos de sua mistura. **Miscellaneous**, v. 22, n. 8, p. e6320, 2024.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: **Proceedings of the 27th international conference on machine learning (ICML-10)**. [S.l.: s.n.], 2010. p. 807–814.

NEGREIROS, K. et al. Aplicação de machine learning na predição da resistência do cimento. **REGRAD - Revista Eletrônica de Graduação do UNIVEM - ISSN 1984-7866**, v. 13, n. 01, p. 1–15, 2020. ISSN 1984-7866. Disponível em: <https://revista.univem.edu.br/REGRAD/article/view/3013>.

PEREIRA, R.; MURAI, F. Quão efetivas são redes neurais baseadas em grafos na detecção de fraude para dados em rede? In: **Anais do X Brazilian Workshop on Social Network Analysis and Mining**. Porto Alegre, RS, Brasil: SBC, 2021. p. 205–210. ISSN 2595-6094. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/16141>.

PORTELA, M. Dados sintéticos: a chave para a inovação sustentável. **MIT Technology Review**, MIT Technology Review, 2022.

PRIBERAM. **Eficiência**. 2024. Acessado em 12 de Abril de 2024. Disponível em: <https://dicionario.priberam.org/efici%C3%AAncia>.

RAUBER, T. W. Redes neurais artificiais. **Universidade Federal do Espírito Santo**, v. 29, 2005.

ROCHA, S. D. F.; LINS, V. d. F. C.; SANTO, B. C. d. E. Aspectos do coprocessamento de resíduos em fornos de clínquer. **Miscellaneous**, v. 16, n. 1, p. 1, 2011.

RUSSEL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach.** 4 ed.. ed. Pearson, 2020. Disponível em: <https://books.google.com.br/books?id=lEKMkgEACAAJ>.

SANCHES, M. K. **Aprendizado de máquina semi-supervisionado: proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados**. Tese (Doutorado) — Universidade de São Paulo, 2003.

SANTOS, J.; ROSSI, R. Aprendizado de máquina não supervisionado baseado em redes heterogêneas para agrupamento de textos. In: SBC. **Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)**. [S.l.], 2020. p. 35–46.

SANTOS, M. M. D. **Supervisão de Sistemas - Funcionalidades e Aplicações**. São Paulo, Brasil: Editora Érica, 2014. ISBN 978-85-365-2037-7.

SILVA, V. d. A.; RODRIGUES, P. S. S. **Comparação de metodologias para amplicação de bases de dados médicas**. 2020.

SOUZA, H. et al. Predição da composição do clínquer industrial utilizando minimização da energia livre de gibbs. v. 61, n. 357, p. 23–30, 2015.

SRIVASTAVA, N. et al. Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, JMLR. org, v. 15, n. 1, p. 1929–1958, 2014.

TAURION, C. **Do hype para o ROI: da experimentação às aplicações reais nos negócios**. 2023. <https://c-taurion.medium.com/do-hype-para-o-roi-da-experimenta%C3%A7%C3%A3o-%C3%A0s-aplica%C3%A7%C3%B5es-reais-nos-neg%C3%B3cios-fb5095693487#:~:text=Do%20hype%20para%20o%20ROI%3A%20da%20experimenta%C3%A7%C3%A3o%20%C3%A0s%20aplica%C3%A7%C3%B5es%20reais%20nos%20neg%C3%B3cios,-Cezar%20Taurion&text=Muitos%20fatores%20afetam%20a%20capacidade,empresa%20s%C3%A3o%20dos%20mais%20cr%C3%ADticos.>

TAYLOR, H. **Cement Chemistry**. Emerald Publishing Limited, 1997. ISBN 9780727725929. Disponível em: <https://books.google.com.br/books?id=1BOETtwi7mMC>.

TORRES, V. A.; LANGE, L. C. Rotas tecnológicas, desafios e potencial para valoração energética de resíduo sólido urbano por coprocessamento no brasil. **Miscellaneous**, v. 27, n. 1, p. 25, 2022.

TURING, A. M. I.— Computing machinery and intelligence. **Mind**, LIX, n. 236, p. 433–460, 10 1950. ISSN 0026-4423. Disponível em: <https://doi.org/10.1093/mind/LIX.236.433>.

TURKIEWICZ, M. S.; FRACAROLLI, R. L. Redes neurais aritificiais: Importância da aplicação na indústria brasileira. **Anais do Encontro Nacional de Engenharia de Produção - Enegep**, 2019.

WANG, X. et al. Water quality prediction based on machine learning and comprehensive weighting methods. v. 25, n. 8, p. 1186, 2023.

ZANETTI, M. C. V.; HARMENDANI, P. H. N. Aprendizado de máquina por reforço aplicado no jogo de cartas uno. **Revista de Sistemas e Computação-RSC**, v. 9, n. 2, 2020.