UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA METALÚRGICA E DE MATERIAIS

CURSO DE GRADUAÇÃO EM ENGENHARIA METALÚRGICA

DANIEL DOS SANTOS SOUZA

DESENVOLVIMENTO DE MODELO DE REGRESSÃO PARA PREVISÃO DO TEOR DE FÓSFORO NA ETAPA DE REFINO PRIMÁRIO DE ACIARIA A OXIGÊNIO

FORTALEZA

2022

DANIEL DOS SANTOS SOUZA

DESENVOLVIMENTO DE MODELO DE REGRESSÃO PARA PREVISÃO DO TEOR DE FÓSFORO NA ETAPA DE REFINO PRIMÁRIO DE ACIARIA A OXIGÊNIO

Trabalho de conclusão de curso apresentado ao curso de Engenharia Metalúrgica do Departamento de Engenharia Metalúrgica e de Materiais da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Engenharia Metalúrgica.

Orientador: Prof. Dr. Ing. Jeferson Leandro Klug.

FORTALEZA

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S237d Souza, Daniel dos Santos.
Desenvolvimento de modelo de regressão para previsão do teor de fósforo na etapa de refino primário de aciaria a oxigênio / Daniel dos Santos Souza. – 2022.
33 f.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia Metalúrgica, Fortaleza, 2022.
Orientação: Prof. Dr. Jeferson Leandro Klug.

1. Aprendizado de máquina. 2. Modelo preditivo. 3. Regressão linear múltipla. I. Título.

CDD 669

DANIEL DOS SANTOS SOUZA

DESENVOLVIMENTO DE MODELO DE REGRESSÃO PARA PREVISÃO DO TEOR DE FÓSFORO NA ETAPA DE REFINO PRIMÁRIO DE ACIARIA A OXIGÊNIO

Trabalho de conclusão de curso apresentado ao curso de Engenharia Metalúrgica do Departamento de Engenharia Metalúrgica e de Materiais da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Engenharia Metalúrgica.

Aprovada em: __/__/____.

BANCA EXAMINADORA

_____
Prof. Dr. Ing. Jeferson Leandro Klug (Orientador)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Elineudo Pinho de Moura
Universidade Federal do Ceará (UFC)

_____
Eng. Carlos André Ruy Carneiro
Companhia Siderúrgica do Pecém

À Deus.

Aos meus pais, Edilson e Luiza e a minha esposa, Ester.

AGRADECIMENTOS

Agradeço primeiramente a Deus que sempre me deu forças para seguir nessa caminhada. Soli Deo Gloria

Aos meus pais por todo o apoio durante a minha vida, sempre possibilitando os meios para que eu realizasse os meus sonhos.

À minha esposa, Ester, por todo o apoio e pela pessoa que ela é.

Ao meu irmão, Lucas, por todo o apoio. E ao meu irmão Rafael, pelas incontáveis caronas e ajudas durante minha vida acadêmica.

Aos meus amigos de curso: Lucas, Cesanildo, Henry, Matheus e Bruno Ribeiro.

Ao Prof. Dr. Ing. Jeferson Leandro Klug, pela excelente orientação e pelo profissional que és.

À Companhia Siderúrgica do Pecém, pelo fornecimento de dados para o trabalho.

“Três coisas são necessárias para a salvação do homem: Saber o que deve crer, O que deve querer, O que deve fazer! Crer em Deus Pai..., Querer a Vida Eterna (Jesus Cristo) e, Fazer o bem.”

São Tomás de Aquino

RESUMO

O aprendizado de máquina é uma área da inteligência artificial e que tem uma aplicação bastante ampla. Há diversos modelos existentes como por exemplo, a regressão linear. É comum encontrarmos aplicações de aprendizado de máquina nas áreas de medicina, segurança pública, empresas do segmento financeiro, entre tantas outras áreas. O presente trabalho tem por objetivo evidenciar a aplicação de um algoritmo de regressão dentro do contexto da engenharia metalúrgica prevendo o teor de fósforo no processo de refino primário, em uma siderúrgica, utilizando regressão. Os dados foram fornecidos por uma siderúrgica com todas as variáveis que compõem o processo de refino primário. A base de dados foi normalizada e então dividida em dados de treino e dados de teste para implementação do modelo de Regressão Linear. O modelo apresentou uma parcial assertividade, entretanto, os parâmetros de avaliação do resíduo estatístico tiveram baixos valores reforçando a possibilidade de modelagem do fenômeno através de um modelo preditivo.

Palavras-chave: Aprendizado de Máquina; Modelo Preditivo; Regressão Linear Múltipla

ABSTRACT

Machine learning is an artificial intelligence area widely applicable. There are many kinds of it, for example linear regression model. Commonly we find machine learning aplications in healthcare systems, public safety, financial services and several others. This papel will evidence an aplication of an algorithm of machine learning that will predict the phosphorus content in the primary refining process, in a steel mill, using linear regression. Data were provided by an steel mills industry with all variables used in this case. Inicially the data base was normalized, then divided in two data packets and finally was implemented a linear regression model. This model presented a partial assertivity, although the adopted metrics have had low values which indicates that is possible an adoption of a predictive models in similar works.

Keywords: Machine Learning; Predictive Model; Multiple Linear Regression.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

KDD	Knowledge Discovery in Databases
DW	Data Warehouse
MSE	Erro Quadrático Médio
RMSE	Raiz do Erro Quadrático Médio

LISTA DE SÍMBOLOS

%	Porcentagem

SUMÁRIO

1 INTRODUÇÃO

Com o avanço tecnológico, uma grande quantidade de informação é gerada, coletada e armazenada todos os dias para diversos fins, dentre objetivos científicos, sociais e econômicos. E devido ao grande volume de dados, concluísse que é humanamente impossível o ser humano analisar e extrair informações desse grande volume de dados que são gerados todos os dias. Devido a limitação humana para obter informações a cerca desse grande volume de dados, foram desenvolvidos algoritmos computacionais que geram informações com uma maior expertise e capacidade. Nesta perspectiva, surgiram diversas linhas de pesquisa, dentre essas, uma denominada de KDD – Knowledge Discovery in Databases (Descoberta de conhecimento em Bases de Dados) definida por (FAYYAD et al. 1996) como um processo não trivial, interativo e iterativo, que visa melhorar os processos de análise de grandes massas de dados compostos das seguintes etapas: Pré-Processamento, Mineração de Dados e Pós-Processamento.

A etapa de Pré-Processamento consiste em tratar e preparar os dados para serem usados na etapa posterior. É nesta etapa que será determinado a qualidade dos dados, tendo em vista que é um fator determinante para a eficiência de qualquer algoritmo. Após esta etapa, temos a etapa de Mineração dos Dados definida por (NEVEZ, 2005) como uma etapa da descoberta do conhecimento dos dados buscando padrões e aplicando algoritmos e técnicas computacionais específicas. Atualmente existem diversos algoritmos que possuem aplicações praticamente em inúmeras áreas. A etapa posterior, será o Pós-Processamento que terá por objetivo a interpretação do conhecimento descoberto. O conhecimento extraído pode ser simplificado; avaliado por meio de critérios como precisão e compreensibilidade entre outros; visualizado, ou simplesmente documentado para o usuário final (DOMINGUES; REZENDE, 2005). É nessa perspectiva que algoritmos têm sido utilizados para prever o risco de crédito em empresas do segmento financeiro. Na metalurgia, modelos têm sido utilizados para prever falhas estruturais em equipamentos metálicos (DANTAS, 2015). E o aço por ser um material presente em quase tudo nas nossas vidas, a otimização do seu processo de fabricação dentro da indústria siderúrgica, é uma das possibilidades que os algoritmos fornecem.

A siderúrgica é a indústria responsável pela produção do aço. O processo siderúrgico inicia-se com o minério de ferro beneficiado no alto forno, onde irá ocorrer a redução dos óxidos de ferro através da utilização de um redutor que deverá ser um material a base carbono. Desse modo, obtém-se uma solução líquida chamada de ferro-gusa, composta de Fe e outros elementos como C, S e P. Após esta etapa, temos o refino primário em que ocorre a elaboração do aço através da diminuição dos teores de carbono, fósforo e enxofre. Posteriormente, ocorre o refino secundário que tem por objetivo a homogeneização da temperatura, bem como o aumento do grau de pureza do aço necessário para atender os requisitos de qualidade. É nesta perspectiva que será analisada uma base de dados do processo de refino secundário de uma siderúrgica e com o objetivo de atender os requisitos de qualidade, o teor de fosforo no término do processo será predito por um algoritmo de aprendizado de máquina. Por fim, temos a etapa de lingotamento onde ocorrerá a solidificação do aço.

2 OBJETIVOS

Este trabalho tem por objetivo prever o teor de fósforo do aço líquido após a realização do processo de conversão, através da aplicação de um modelo de regressão utilizando variáveis do processo de refino primário. Além disso, é esperado mostrar a aplicação de modelos de regressão dentro da engenharia metalúrgica.

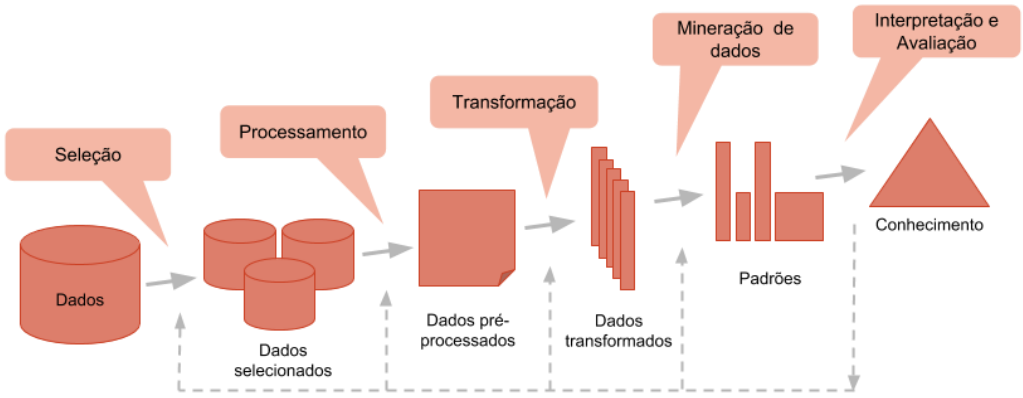
3 REFERENCIAL TEÓRICO

O objetivo desta seção é abordar como o processo de descoberta de conhecimento será aplicado e principalmente, como o modelo de aprendizado de máquina trabalha os dados para fornecer um resultado, bem como os seus métodos de avaliação.

3.1 Processo de descoberta de conhecimento em bancos de dados (DCBD)

Podemos definir o processo de descoberta de conhecimento (Knowledge Discovery in Databases), o KDD, como sendo um processo de busca e extração de conhecimento em bases de dados com o intuito de extrair insights novos, úteis e pertinentes ao negócio. Esse processo pode ocorrer de forma interativa e iterativa. (ELMASRI, NAVATHE 2011) divide o processo em seis etapas: seleção de dados, limpeza de dados, enriquecimento de dados, transformação de dados, mineração de dados e exibição da informação descoberta, conforme pode ser observado na figura 1.

Figura 1: Etapas do processo KDD.



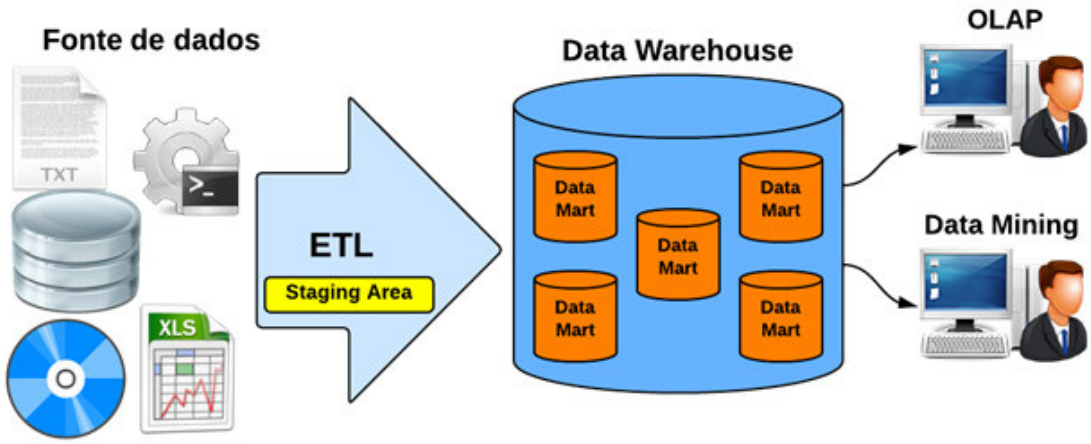
FAYYAD et al. (1996, com adaptações)

3.1.1 Seleção de dados

De acordo com (FAYYAD et al., 1996), é nesta fase que iremos definir o conjunto de dados que será utilizado no processo de descoberta do conhecimento. Nesta etapa, os dados serão organizados em uma base de dados, entretanto, ainda sem uma estruturação adequada para o processo de mineração de dados. Os conjuntos de dados podem ser obtidos a partir de várias fontes de dados, comumente utilizadas no mercado, como por exemplo: Planilhas, Data Warehouse e Base de Dados Externa.

Segundo (INOMN, 1997), um Data Warehouse é uma coleção de dados agrupados por assunto, integrados, variáveis com o tempo e não voláteis que auxiliam o processo de tomada de decisão nas organizações. O Data Warehouse terá dados de diferentes fontes agrupando as informações dos sistemas da empresa em um único lugar. Uma Base de Dados externa são os bancos de dados, fora do domínio. Muito comumente, esses bancos de dados estão hospedados na web. Na figura 2, temos uma exemplificação de como funciona um Data Warehouse.

Figura 2: Estrutura de um Data Warehouse.



Machado (2004, com adaptações)

Importante salientar que a seleção dos dados é uma etapa muito importante no processo de descoberta do conhecimento, tendo em vista que os dados escolhidos são determinantes para a escolha da estratégia do negócio.

3.1.2 Pré-processamento dos dados

A etapa de pré-processamento consiste em fornecer qualidade aos dados selecionados na etapa anterior. Os dados obtidos na etapa anterior possuem diferentes formatos (estruturados, semiestruturados e não-estruturados). Desse modo, faz-se necessário processos de transformação dos dados. Corroborando (NAVEGA, 2002), as bases de dados são dinâmicas, incompletas, redundantes, ruidosas e esparsas, necessitando de um pré-processamento para “limpá-las”. Operações muito comuns nesta etapa são: remoção de ruídos e desvios, definição de informação necessária para o modelo, adequação do ruído, estratégia para tratar os campos ausentes, normalização e indexação.

3.1.3 Transformação de dados

Nesta etapa, a base de dados tratada, se necessário, será combinada com outras diferentes bases de dados alterando o seu formato tornando a informação mais rica e útil. A ferramenta que será utilizada no processo de mineração dita como os dados devem ser representados, por esse motivo, pode-se dizer que o objetivo da etapa de transformação de dados é eliminar quaisquer limitações nos algoritmos que serão utilizados posteriormente.

3.1.3.1 Normalização

O processo de normalização é um processo de transformação e nada mais é do que transformar os valores dos atributos de seu intervalo original para um intervalo de nosso interesse. Utiliza-se esta técnica de transformação quando o objetivo é calcular distâncias entre atributos, isso porque os algoritmos, tendem a dar mais importância para os atributos com um intervalo maior de valores.

3.1.4 Mineração de dados

De acordo com (THOMÉ, 2002), a etapa de data mining é a implementação de modelos computacionais que possuem a capacidade de identificar e revelar padrões desconhecidos, existentes em dados que pertencem a uma ou mais bases de dados. É nesta etapa que será descoberto novas informações de correlações, padrões e tendências do conjunto de dados analisado.

Segundo (JESUS; MOSER; OGLIARI, 2011), as técnicas de mineração de dados adaptam-se melhor a alguns problemas que a outros, ou seja, os métodos são escolhidos e aplicados com base no problema e no objetivo de tomada de decisão que se busca. Dentre as técnicas de mineração de dados, podemos destacar: agrupamento, regressão, associação e classificação.

3.1.5 Avaliação

A última etapa do processo de descobertas do conhecimento consiste em interpretar e aplicar os conhecimentos obtidos na tomada de decisão. É nesta fase que são apresentadas as medidas de desempenho (REZENDE, 2003). Os resultados podem ser apresentados e interpretados de diferentes formas como por exemplo, gráficos e tabelas.

3.2 Aprendizado de Máquina

Temos que o aprendizado de máquina é um sub-campo da inteligência artificial e que aborda questões de como tornar as máquinas aptas a aprender. Portanto, o aprendizado de máquina se refere a inferência indutiva. (RÄTSCH, G. A, 2004).

O aprendizado de máquina generaliza além dos exemplos existentes no conjunto de treinamento. Importante salientar que as máquinas não aprendem tão bem quanto os seres humanos, porém há diversos algoritmos que são eficientes para várias tarefas de aprendizado.

Um dos motivos de tornarem os seres humanos limitados quanto a análise de problemas que o aprendizado de máquina se propõe a resolver, é devido a intuição humana ser treinada em um universo tridimensional, desse modo, surge uma dificuldade natural para resolver problemas de dimensões maiores sem a utilização de ferramentas de auxílio. (DOMINGOS, 2012). A dificuldade humana não se limita a problemas de dimensões maiores. As máquinas possuem uma grande capacidade de armazenamento e dada a grande quantidade de dados, o ser humano não consegue processar esse volume de dados sem o auxílio de computadores.

Além disso, o aprendizado de máquina pode ser dividido em dois grandes grupos: aprendizagem supervisionada e aprendizagem não-supervisionada.

3.2.1 Aprendizagem supervisionada

Na aprendizagem supervisionada existe um supervisor, portando, o algoritmo aprende a partir de um conjunto de dados rotulado. Esses rótulos ou variáveis-objetivo são as variáveis que se deseja predizer a partir dos dados existentes. Segundo (SATHYA; ABRAHAM, 2013), as variáveis-objetivo devem ser escolhidas de modo a representar a resposta para o problema.

3.2.2 Aprendizagem não-supervisionada

Diferentemente do aprendizado supervisionado, o aprendizado não-supervisionado não utiliza as variáveis de saída. Os algoritmos leem os dados e de acordo com sua proximidade, são agrupados.

3.3 Regressão Linear

A regressão, seja linear simples ou múltipla, é do que o estudo da dependência de uma variável dependente (Y) em relação a uma ou mais variáveis explicativas (X_1, X_2, …, X_3). Deste modo, é possível conhecer o comportamento da variável (Y).

A regressão, portanto, irá modelar e tentar descrever como as variáveis estão relacionadas entre si, através de modelos estatísticos que expliquem essa dependência.

3.3.1 Regressão Linear Simples

O modelo de regressão linear simples pode ser definido de acordo com a expressão abaixo.

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad i = 1, \dots, n \qquad Eq.3.1$$

onde:

y_i: É o valor da variável dependente para o i-ésimo elemento da amostra;

x_i: É o valor conhecido da variável independente ou preditora para o i-ésimo elemento da amostra;

$\beta_1 \; e \; \beta_2$: São os parâmetros desconhecidos;

e_i: É o erro associado ao modelo.

Além disso, temos que o interceptor $\beta_0$representa o ponto inicial de y, quando o valor de x é zero. E x_i representa cada observação da variável explicativa e β_1 representa o ângulo em que a reta faz em relação ao eixo x. Por fim, o e_i é o erro associado a cada observação.

3.3.2 Regressão Linear Múltipla

De acordo com (Montgomery; Runger, 2012), um modelo de regressão que contenha mais de uma variável explicativa é chamado de regressão múltipla. O termo linear é utilizado porque a equação é função linear dos parâmetros desconhecidos $\beta_0, \; \beta_1 \; e \; \beta_2$.

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_n x_n + e_i \qquad i = 1, \dots, n \qquad Eq.3.2$$

onde:

y_i: É o valor da variável dependente para o i-ésimo elemento da amostra;

x_i: É o valor conhecido da variável independente ou preditora para o i-ésimo elemento da amostra;

β_n: São os parâmetros desconhecidos;

e_i: É o erro associado ao modelo.

Segundo (Montgomery, 2001) o erro aleatório é caracterizado como uma variável aleatória contínua, independentemente distribuída, com média nula e variância constante ao longo dos valores das varáveis do modelo. Desse modo, temos que Y também é definido como uma variável aleatória e X uma variável normal.

Podemos representar o modelo de regressão de forma matricial, em que temos K variáveis e n observações.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} \qquad Eq.\,3.3$$

onde:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} e\ \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Comumente, $\boldsymbol{Y}$ será um vetor (n x 1), o que representa n observações da variável dependente. $\boldsymbol{X}$ é uma matriz (n x p) e $\boldsymbol{\beta}$ é um vetor (p x 1) dos coeficientes da equação de regressão. Assim como a variável dependente, o erro $\boldsymbol{e}$ é um vetor (n x 1).

3.3.2.1 Estimativa de mínimos quadrados de parâmetros

A função escolhida para o fenômeno será aquela que melhor se ajustar aos dados. A escolha dessa função será de acordo com a diferença entre os valores reais e os estimados, de tal forma que seja a menor possível. Ou seja, a melhor função é a que propicia o menor resíduo estatístico. Segundo (CORRAR, 2009), o método de predição mais comumente utilizado em regressão linear é o Metodo dos Mínimos Quadrados (MMQ), cujo objetivo é obter a menor soma de quadrados dos resíduos (SQR).

Portanto, deseja-se encontrar o vetor dos estimadores quadrados que minimize:

$$L = \sum_{i=1}^{n} e_i^2 = e'e = (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) \qquad Eq.\,3.4$$

Derivando e igualando a zero, temos:

$$\boldsymbol{\beta}' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \qquad Eq.3.5$$

Os elementos da diagonal de $\boldsymbol{X}'\boldsymbol{X}$ são a soma dos quadrados dos elementos nas colunas de $\boldsymbol{X}$ e os elementos fora da diagonal são a soma dos quadrados dos elementos nas colunas de $\boldsymbol{X}$. Os elementos fora da diagonal são a soma dos produtos cruzados dos elementos nas colunas de $\boldsymbol{X}$ e dos valores observados na variável controle.

Podemos escrever em notação matricial, temos:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \qquad Eq.3.6$$

A diferença entre o que se estima e os valores real é denominado de resíduo de regressão.

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} \qquad Eq.3.7$$

3.3.2.2 Correlação

De acordo com (Montgomery, 2001), o coeficiente de correlação linear (R) é uma medida de correlação que indica o nível de intensidade que ocorre na correlação entre as variáveis. Uma forma de medir o coeficiente de correlação linear foi desenvolvida por Pearson e mede o grau de ajustamento dos valores em torno de uma reta e pode ser determinado de acordo com a equação abaixo.

$$R = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum xi^2 - (\sum xi)^2] * [n\sum yi^2 - (\sum yi)^2]}} \qquad Eq.3.7$$

Desse modo, (Montgomery, 2001) diz que o coeficiente de correlação irá variar entre -1 e +1, sendo que positivo indica correlação positiva entre às variáveis, valor negativo indica correlação negativa entre as variáveis e o valor zero indica falta de correlação entre as variáveis.

De acordo com o valor do coeficiente de correlação entre as variáveis X e Y, podemos ter os seguintes indicativos, conforme tabela 1.

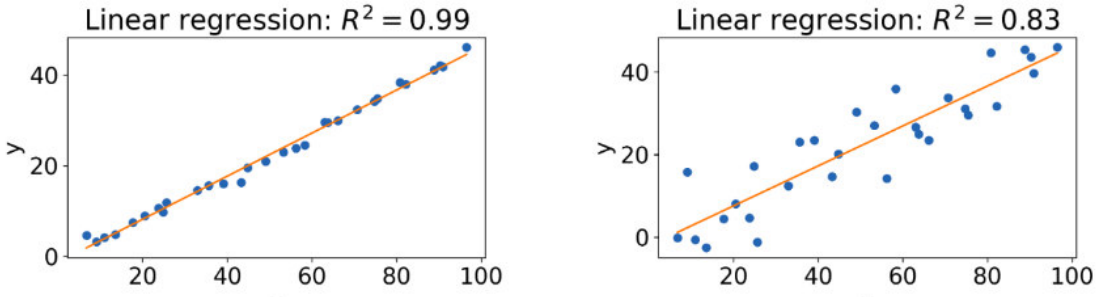
Tabela 1: Grau de relação Linear entre X e Y

Valor de R	Relação Linear
0	nula
0 < R ≤ 0,30	fraca
0,30 < R ≤ 0,60	média
0,60 < R ≤ 0,90	forte
0,90 < R ≤ 0,99	fortíssima
1	perfeita

Hochhein (2011, com adaptações)

Segundo (Montgomery, 2001), R^2 mede quanto da variabilidade dos dados é explicada pelo modelo de regressão. Quanto maior R^2, mais a variação total de Y é explicada. Essa observação pode ser visualizada na figura 3.

Figura 3: Exemplos de cálculo de R^2.



Fonte: Disponível no blog Turing Talks.

3.3.2.3 Parâmetros de validação e avaliação da performance do modelo

Conforme (CORRAR, 2009), com o intuito de minimizar o risco de decisões erradas com base nas informações fornecidas pelo algoritmo, podemos validar o modelo através de alguns métodos estatísticos e irão descrever a habilidade do modelo.

A avaliação do modelo será de acordo com os seguintes métodos estatísticos:

- R^2: Também chamado de coeficiente de determinação, este parâmetro expressa a quantidade da variância dos dados que é explicado pelo modelo construído.

- Erro Quadrático Médio (MSE): Consiste na média do erro das previsões ao quadrado. Quanto maior o valor de MSE, pior é o modelo. É calculado de acordo com a equação: $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad Eq.3.8$
- Raiz do Erro Quadrático Médio (RMSE): É a raiz do erro quadrático médio e visa melhorar a interpretação do parâmetro, ajustando a unidade. Entretanto, o RMSE, assim como o MSE, penaliza predições mais distantes do valor real. O RMSE pode ser calculado segundo a equação a seguir: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad Eq.3.9$

4 METODOLOGIA

A análise foi feita de acordo com as etapas previstas no processo de descoberta de conhecimento. Além disso, todas as etapas de pré-processamento, mineração de dados e avaliação foram feitas com a IDE Jupyter Notebook.

A amostra foi fornecida por uma siderúrgica e possui o formato descrito na tabela 2.

Tabela 2: Formato da amostra.

Observações	Variáveis
4.549	34

Fonte: Autor (2022)

Na amostra de dados, observou-se que havia um grande volume de dados, cujas informações estavam faltantes, conforme pode ser observado na tabela 3.

Tabela 3: Volume de dados faltantes presentes na amostra.

Variável	Porcentagem de dados Faltantes
Mn_Final_Sopro	87,34%
C_Final_Sopro	77,56%
P_Final_Sopro	77,56%
Mn_CDy	36,34%
Volume_Gas_Recuperado_(Nm³)	13,67%
OFS	9,17%
P_Gusa	7,41%
Si_Gusa	7,32%
Temp_Gusa	6,07%
TFS	5,12%
P_CDy	4,92%
MgO	3,80%
FeT	3,80%
Basicidade	3,80%
Temp_CDy	2,22%
%_Instante_Medição_CDy	1,41%
Mn_Gusa	1,38%
CAL_Total_(t)	0,59%
BSSF_(t)	0,59%
CAL_1º_Sopro_(t)	0,59%
Sinter_Total_(t)	0,59%
Adicao_Sinter	0,59%
Ventaneiras	0,59%

Fonte: Autor (2022)

Desse modo, para não interferir na implementação do modelo, optou-se por excluir todas as observações com algum atributo ausente o que reduziu a amostra de dados para 402 entradas para as 34 variáveis. Além disso, após discussão com engenheiros da própria siderúrgica, determinou-se que apenas as variáveis da tabela 4 iriam compor o modelo, dada a sua influência no processo, de acordo com a literatura.

Tabela 4: Variáveis que compõem o modelo preditivo.

Origem	Variável
Ferro Gusa	Mn_Gusa
	P_Gusa
	Si_Gusa
	Temp_Gusa
Análise de medição intermediária do processo	Mn_CDy
	P_CDy
	Temp_CDy
	VO2_CDy
Demais variáveis que compõem o processo	Sucata_Total
	Adicao_Sinter
	Vida_BOF
	Vida_Lanca
	Volume_O_Soprado
	Ventaneiras

Fonte: Autor (2022)

Entretanto, aplicando-se o teste ANOVA de significância ($\alpha = 0{,}05$) para as variáveis sugeridas pelos engenheiros de processo da siderúrgica, concluiu-se que apenas as variáveis da tabela 5 possuem um relacionamento estatisticamente significativo. O teste foi aplicado através da utilização do método SelectKBest do pacote Feature_Selection disponível no Scikit-Learn.

Tabela 5: Variáveis que compõem o modelo preditivo após a aplicação do teste ANOVA.

Origem	Variável
Ferro Gusa	Mn_Gusa
	P_Gusa
	Temp_Gusa
Análise de medição intermediária do processo	Mn_CDy
	P_CDy
	Temp_CDy
	VO2_CDy
Demais variáveis que compõem o processo	Sucata_Total
	Volume_O_Soprado

Fonte: Autor (2022)

Na tabela 6, temos o dicionário, bem como a descrição de cada variável da amostra utilizada na implementação do modelo.

Tabela 6: Dicionário de variáveis

Variável	Descrição
Mn_Gusa	Teor de manganês do ferro gusa
P_Gusa	Teor de fósforo do ferro gusa
Temp_Gusa	Temperatura do ferro gusa
Mn_CDy	Teor de manganês obtido na análise do controle dinâmico
P_CDy	Teor de fósforo obtido na análise do controle dinâmico
Temp_CDy	Temperatura obtida na análise do controle dinâmico
VO2_CDy	Volume de oxigênio obtido na análise do controle dinâmico
Sucata_Total	Quantidade de sucata total utilizada no processo
Volume_O_Soprado	Volume de oxigênio utilizado em todo o processo

Fonte: Autor (2022)

Desse modo, o formato da amostra utilizada para implementação do modelo era a seguinte, conforme tabela 7.

Tabela 7: Formato da amostra utilizada no modelo.

Observações	Variáveis
402	9

Fonte: Autor (2022)

Determinado os atributos que iriam compor o modelo, os dados da amostra passaram por uma etapa denominada de normalização mínimo-máximo, em que foi atribuído o intervalo de nosso interesse (-1, 1) de forma que as variáveis não fossem penalizadas de forma diferente pelo modelo. Posteriormente, dividimos a amostra em dados de treino e dados de teste na proporção 70/30, em que 70% dos dados foram agrupados em dados de treino e 30% dos dados foram agrupados em dados de teste. Utilizou-se o algoritmo de regressão linear múltipla disponível em um pacote da linguagem de programação Phyton, chamado Scikit-learn.

5 RESULTADOS E DISCUSSÕES

Executamos o algoritmo 200 vezes e obtivemos o seguinte resultado para a equação proposta pelo modelo, como pode ser visualizado na tabela 8.

Tabela 8: Coeficientes determinado pelo modelo.

Coeficiente	Valor
β_1	0,05003
β_2	0,2199
β_3	0,0458
β_4	0,2689
β_5	0,4295
β_6	-1,0712
β_7	-0,0892
β_8	-0,0292
β_9	0,0506

Fonte: Autor (2022)

O valor encontrado para o intercepto segue na tabela 9.

Tabela 9: Coeficiente determinado pelo modelo para o intercepto

Coeficiente	Valor
β_0	0,8008

Fonte: Autor (2022)

O coeficiente de determinação encontrado foi de 63,14%, o que significa que o valor unitário é explicado pela equação de regressão. Todavia, ainda há 36,86% que não foi explicado e que podem estar relacionados a variáveis que não foram consideradas no modelo.

Na tabela abaixo, temos a avaliação com base nos parâmetros propostos na seção 3.3.2.3, conforme tabela 10.

Tabela 10: Resultado dos parâmetros de avaliação do modelo

Parâmetro	Valor
R^2	63,14%
Erro Quadrático Médio (MSE)	0,0452
Raiz do erro quadrático médio (RMSE)	0,0047

Fonte: Autor (2022)

Como já informado anteriormente, o modelo explica apenas 63,14% do fenômeno observado. Entretanto, através da comparação dos dados reais e dados previstos pelo modelo, observou-se que para o erro quadrático médio, tivemos um valor de 4,52%, o que significa que embora 36,86% dos dados da amostra estejam sem explicação, o erro permaneceu dentro dos limites esperados. O valor apresentado pela raiz do erro quadrático médio em torno de 0,47% comprovou a hipótese acima.

Para os dados de teste, observou-se que o modelo apresentou um erro de até 4,52% para cerca de 72% das observações dos dados de teste. O restante, cerca de 28% das observações dos dados de teste, tiveram erros de até 9,32%. Podemos visualizar essas informações na tabela 11, abaixo.

Tabela 11: Diferença percentual dos dados de teste

Erro	Quantidade (%)
Até 4,52%	72%
Acima de 4,52%	28%

6 Conclusão

O presente trabalho tem por objetivo apresentar a aplicação de um modelo de regressão dentro do contexto da engenharia metalúrgica implementado uma solução simples para um problema de previsão de fósforo no refino primário, em uma aciaria.

O uso do algoritmo de regressão linear múltipla mostrou-se parcialmente aderente ao fenômeno, tendo em vista o valor de 63,14% apresentando pelo coeficiente de determinação. Tal fato pode ser explicado por variáveis que não estavam presentes na amostra ou o não mapeamento de demais variáveis dentro da siderúrgica. Entretanto, a implementação do modelo se mostrou assertiva, tendo em vista o baixo valor encontrado para o erro quadrático médio, além disso, verificou-se que cerca de 72% dos dados possuíam o erro de até 4,52% em relação ao valor original dos dados de teste. Dessa forma, o objetivo do presente trabalho foi atingido evidenciando a aplicação de modelos de regressão dentro da área da engenharia metalúrgica.

7 Trabalhos Futuros

- Mapeamento de outras variáveis dentro da siderúrgica que possam aumentar o grau de confiabilidade da solução;
- Implementação de uma solução mais robusta com uso de Deep Learning;
- Realizar o deploy do modelo em um software embarcado.

REFERÊNCIAS

BISHOP, C. M. **Pattern recognition and machine learning**. [S.l.]: springer, 2006

DANTAS, Guilherme Vieira. **Utilização de classificador random forest na detecção e previsão de falhas em máquinas rotativas**. Monografia (Graduação em Engenharia Eletrônica e de Computação) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

CORRAR, L.; Paulo, E,; Filho,J. **Análise Multivariada para os Cursos de Administração, Ciências Contábeis e Economia**. Atlas, São Paulo, 2009.

DOMINGOS, P. **A few useful things to know about machine learning**. Communications of the ACM, ACM, v. 55, n. 10, p. 78–87, 2012.

DOMINGUES, M. A.; REZENDE, S. O. **Pós-Processamento de Regras de Associação usando Taxonomias**. Artigo. Journal of Computer Science. v. 4, p. 26-37. 2005.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de banco de dados.** 6.ed. São Paulo: Pearson, 2011.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in Knowledge Discovery and Data Mining**. AAAI/MIT Press, 1996.

INMON, W. H.; HACKATHORN, R. D. **Como usar o Data Warehouse.** Rio de Janeiro: Infobook, 1997.

MACHADO, F. N R. **Tecnologia e Projeto de Data Warehouse Uma Visão**

MONTGOMERY, D.; RUNGER G. **Estatística aplicada e probabilidade para engenheiros.** LTC, São Paulo, 2012.

MONTGOMERY, D.; VINING, G., PECK, A. **Introduction to Linear Regression Analysis**. John Wiley & Sons New York, New York, 2001.

NAVEGA, Sérgio. **Princípios essenciais do Data Mining**. São Paulo, SP: 2002. Publicada nos Anais do Infoimagem.

NEVEZ, R. C. D. **Pré-Processamento no Processo de Descoberta de Conhecimento em Banco de Dados**. Universidade Federal do Rio Grande do Sul, Instituto de Informática, Programa de Pós-Graduação em Computação, 2005. Dissertação de Mestrado. Multidimensional.São Paulo: Editora Érica, 2004.

RÄTSCH, G. **A brief introduction into machine learning**. In: 21st Chaos Communication Congress. [S.l.: s.n.], 2004.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. [S.l.]: Editora Manole Ltda, 2003.

SATHYA, R.; ABRAHAM, A. **Comparison of supervised and unsupervised learning algorithms for pattern classification**. Int. J. Adv. Res. Artificial Intell, Citeseer, v. 2, n. 2,p. 34–38, 2013.