UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

VINICIUS MATHEUS PIMENTEL ARIZA

ESTIMAÇÃO DA PRODUÇÃO DE BIOGÁS NA CODIGESTÃO ANAERÓBIA DE RESÍDUOS ORGÂNICOS DA INDÚSTRIA SUCROALCOOLEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

LONDRINA

2022

VINICIUS MATHEUS PIMENTEL ARIZA

ESTIMAÇÃO DA PRODUÇÃO DE BIOGÁS NA CODIGESTÃO ANAERÓBIA DE RESÍDUOS ORGÂNICOS DA INDÚSTRIA SUCROALCOOLEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Estimation of biogas production in anaerobic codigestion of organic residues from the sugar-alcohol industry using Machine Learning algorithms

Trabalho de conclusão de curso de graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Produção da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Bruno Samways dos Santos

LONDRINA

2022

VINICIUS MATHEUS PIMENTEL ARIZA

ESTIMAÇÃO DA PRODUÇÃO DE BIOGÁS NA CODIGESTÃO ANAERÓBIA DE RESÍDUOS ORGÂNICOS DA INDÚSTRIA SUCROALCOOLEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Trabalho de conclusão de curso de graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Produção da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 02 de junho de 2022

_____

Bruno Samways dos Santos
Doutor
Universidade Tecnológica Federal do Paraná

_____

Rafael Henrique Palma Lima
Doutor
Universidade Tecnológica Federal do Paraná

_____

Marco Antonio Ferreira
Doutor
Universidade Tecnológica Federal do Paraná

LONDRINA

2022

AGRADECIMENTOS

Agradeço à minha família por todo o apoio e por compreenderam a minha ausência enquanto eu me dedicava à realização deste trabalho.

Ao meu orientador por ter desempenhado tal função, pela dedicação e pelo imprescindível direcionamento.

A todos aqueles que contribuíram, de alguma forma, para a realização deste trabalho.

RESUMO

O desenvolvimento de modelos preditivos confiáveis que estimam a produção de biogás em instalações de escala industrial tem-se mostrado um desafio. A utilização de um modelo de Aprendizado de Máquina combinado com ferramentas analíticas pode apoiar a tomada de decisão e aprimorar o desempenho das usinas de biogás. Neste contexto, o presente trabalho teve como objetivo desenvolver modelos de Aprendizado de Máquina para estimar a produção de biogás, utilizando dados diários de análises e de sensores de uma usina produtora de biogás a partir de resíduos da indústria sucroalcooleira. A metodologia envolveu a utilização dos algoritmos *Random Forest*, Redes Neurais Artificiais, *Support Vector Machine e Least Absolute Shrinkage And Selection Operator*, sendo ajustados e comparados através das métricas R-quadrado, *Mean absolute error* (MAE) e *Mean absolute percentage error* (MAPE). Os resultados mostraram que o modelo *Random Forest* obteve os melhores resultados, com R-quadrado final de 0,83 para o conjunto de teste. Também verificou-se que os principais atributos para a estimação da produção de biogás foram a quantidade de resíduos dosado (sólidos e líquidos) e o pH. Por fim, também foi desenvolvida uma aplicação *web* para realizar as estimações utilizando o melhor modelo gerado.

Palavras-chave: Aprendizado de Máquina; Produção de Biogás; Estimação; Regressão.

ABSTRACT

The development of reliable predictive models that estimate biogas production in industrial-scale facilities has been a challenge. The use of a Machine Learning model combined with analytical tools can support decision-making and improve the performance of biogas plants. In this context, the present work aimed to develop Machine Learning models to estimate the production of biogas, using daily data from analysis and sensors of a plant producing biogas from residues from the sugar and ethanol industry. The methodology involved the use of Random Forest, Artificial Neural Networks, Support Vector Machine, and Least Absolute Shrinkage And Selection Operator Algorithms, being adjusted and compared using the R-squared, Mean absolute error (MAE) and Mean absolute percentage error (MAPE) metrics. The results showed that the Random Forest model obtained the best results, with a final R-squared of 0,83 for the test set. It was also found that the main attributes for the estimation of biogas production were the amount of waste dosed (solid and liquid) and pH. Finally, a web application was also developed to perform the estimations using the best model generated.

Keywords: *Machine Learning; Biogas Production; Estimation; Regression.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADM1	*Anaerobic Digestion Model nº 1*
AM	Aprendizado de Máquina
ANFIS	*Adaptive Neuro Fuzzy Inference System*
API	*Application Programming Interface*
AutoML	*Automated Machine Learning*
COP26	Conferência do Clima da Organização das Nações Unidas de 2021
ETL	*Extract, transform and load*
GEE	Gases de efeito estufa
GerDA	*Generalized Discriminant Analysis*
IA	Inteligência Artificial
JSON	*JavaScript Object Notation*
KDD	*Knowledge Discovery in Databases*
KNN	*K-Nearest Neighbors*
LASSO	*Last absolute shrinkage and selection operator*
LDA	*Linear Discriminant Analysis*
LSSVM	*Least-Squares Support-Vector Machine*
MAD	*Median Absolute Deviation*
MAE	*Mean absolute error*
MAPE	*Mean absolute percentage error*
MSE	*Mean Squared Error*
PDI	*Pentaho Data Integration*

R^2	Coeficiente de Determinação
RAE	*Relative Absolute Error*
RF	*Random Forest*
RMSE	*Root Mean Square Error*
RNA	Rede Neural Artificial
RNN	*Recurrent Neural Network*
RSE	*Relative Squared Error*
RSM	*Response Surface Methodological*
RSS	*Residual Sum of Squares*
SVM	*Support Vector Machine*
TPOT	*Tree-based Pipeline Optimization Tool*
VPN	*Virtual Private Network*
XGBoost	*Extreme Gradient Boosting*

LISTA DE SÍMBOLOS

°C	Grau Celsius
A	Ampere
$C_2H_4O_2$	Ácido acético
CH_4	Metano
CO_2	Dióxido de carbono
H	Hidrogênio
mbar	Milibar
mg/L	Miligrama por litro
NH_3	Amoníaco
Nm^3	Normal metro cúbico
pH	Potencial Hidrogeniônico
ppm	Parte por milhão
t	Tonelada

SUMÁRIO

1. INTRODUÇÃO

Conforme Clercq *et al.* (2019), têm-se feito diversos estudos experimentais em escala laboratorial com o objetivo obter a maior eficiência produtiva de biogás através da mistura de resíduos orgânicos. Porém, por se tratar de um processo biológico, as características do local da instalação industrial, como por exemplo impurezas e o clima, podem influenciar no processo de digestão dos resíduos e são difíceis de serem simuladas em laboratório. Além disso, devido à complexidade do processo e as limitações da compreensão humana, torna-se difícil construir modelos baseados na experiência para estimações precisas, podendo o aprendizado de máquina ajudar neste sentido (WANG *et al*., 2021).

Diariamente, em uma usina de biogás são gerados diversos dados de indicadores de processo e de qualidade por meio de sensores e análises físico-químicas. Esses dados podem ser resumidos em três diferentes momentos: antes, durante e após a biodigestão. Os parâmetros controláveis antes da biodigestão, como quantidade, tipo e característica de cada resíduo dosado, são muito importantes para o processo, pois esses indicadores poderão causar mudanças ambientais dentro do biodigestor, no qual estão envolvidas diversas espécies de microrganismos que precisam de condições específicas que favoreçam a produção de biogás (KARLSSON, 2014).

Com a finalidade de aumentar o desempenho dos digestores anaeróbios, é recomendada a digestão simultânea de dois ou mais resíduos sob condição anaeróbia, ou seja, a codigestão anaeróbia de resíduos. Dessa forma, a escolha do melhor substrato e a proporção ideal da mistura pode proporcionar um efeito sinergético positivo no meio, conduzindo à digestão estável e a otimização do rendimento de biogás (SILVEIRA, 2017; ALVES, 2016).

O aprendizado de máquina (AM) é um método baseado em dados para desenvolver a construção de modelos que permitem fazer estimações de sistemas complexos, onde há, por exemplo, diversos parâmetros que podem influenciar no resultado esperado (JEONG, *et al.*, 2021). Assim, com a utilização deste método, espera-se construir um modelo confiável que auxilie na tomada de decisão, potencializando a produção dessa matriz renovável e a redução da poluição ambiental.

1.1 Objetivo geral

Desenvolver modelos de AM para estimar a produção de biogás, utilizando dados de análises e de sensores de uma usina produtora de biogás por meio de resíduos da indústria sucroalcooleira e integrar o melhor modelo a uma interface gráfica de usuário.

1.2 Objetivos específicos

- Implementar e avaliar modelos de AM para estimar a produção de biogás
- Compreender os principais fatores que influenciam na produção de biogás.
- Desenvolver uma interface gráfica de usuário integrada ao modelo de AM para possibilitar o uso do modelo gerado nas operações diárias da empresa.

1.3 Justificativa

O desenvolvimento de modelos preditivos confiáveis que estimam o rendimento de biogás em função do tipo de matéria-prima tem se mostrado um desafio. Modelos baseados em processo para a estimação da produção, como o *Anaerobic Digestion Model 1* (ADM1), requerem conhecimento de muitas variáveis relacionadas as concentrações para componentes detalhados de substratos, o que exige uma análise contínua e extensa, limitando a aplicabilidade em instalações industriais onde esses dados não são coletados regularmente. Além disso, embora seja teoricamente bem compreendida, a microbiologia por trás do processo de biodigestão é altamente complexa e, devido a essa complexidade, torna-se difícil construir modelos baseados na experiência para estimações precisas (WANG *et al.*, 2021).

Abordagens baseadas em AM podem ser superiores às condutas teóricas quando os sistemas de destino possuem uma maior complexidade do ambiente por causa de vários parâmetros de entrada (CLERCQ *et al.*, 2019). Portanto, a busca por modelos apropriados para análise preditiva são, atualmente, uma prioridade para auxiliar no controle de processos de biodigestão anaeróbia em instalações de biogás

em escala industrial, evidenciando a importância da realização de pesquisas que contribuam com este objetivo (CLERCQ *et al.*, 2019).

1.4 Estrutura do trabalho

Após a introdução, em que foi apresentada a contextualização, os objetivos gerais e específicos e a justificativa, o restante do trabalho está dividido em mais cinco capítulos.

O referencial teórico (Capítulo 2) aborda os temas de descoberta de conhecimento em banco de dados, aprendizado de máquina, técnicas de regressão, métricas de avaliação e uma breve tratativa sobre energias alternativas, dando foco ao biogás.

Em seguida, no Capítulo 3, são apresentados alguns trabalhos correlatos que tratam da aplicação de modelos de AM para a análise de fatores que influenciam a produção de biogás ou de seus compostos.

No Capítulo 4 sobre a metodologia, são descritos o conjunto de dados e a sequência da pesquisa, bem como as ferramentas utilizadas em cada uma das etapas.

Os resultados obtidos foram apresentados no Capítulo 5. Nesta sessão foi feita a avaliação dos modelos treinados para estimar a produção o de biogás, a análise dos resultados obtidos e a explicação sobre as etapas para a implantação do modelo em uma plataforma *web*.

Por fim, no Capítulo 6, são expostas as conclusões e considerações sobre o trabalho realizado.

2. REFERENCIAL TEÓRICO

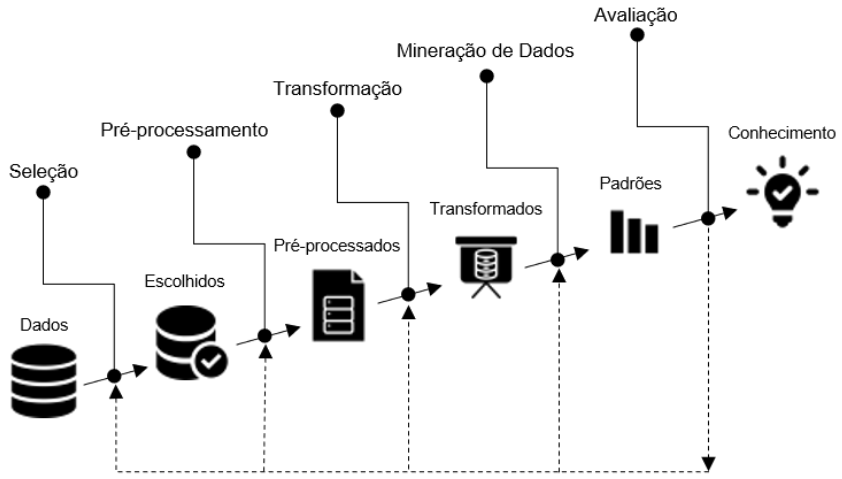
Esta seção descreve brevemente os principais conceitos de AM e o processo de produção de biogás.

2.1 Descoberta de conhecimento em banco de dados

A descoberta de conhecimento em banco de dados, também designado como processo *Knowledge Discovery in Databases* (KDD), é um conjunto de etapas que tem como objetivo extrair informações potencialmente úteis de bases de dados. De acordo com Fayyad *et al*. (1996, p. 2), KDD é definido como: “o processo não-trivial de identificação válida, em dados novos, potencialmente úteis e finalmente com padrões compreensíveis”.

Esse processo é composto por cinco etapas, a saber: a seleção dos dados, o pré-processamento, a transformação, a mineração de dados e a avaliação dos resultados, conforme ilustrado na Figura 1.

Figura 1 - Etapas do processo KDD



Fonte: Fayyad, Piatetsky-Shapiro e Smyth, 1996 (tradução pelo autor).

A compreensão das etapas do processo KDD pode auxiliar na criação de modelos confiáveis e consistentes. A seguir, serão descritas cada uma das etapas do método.

2.1.1 Seleção

A seleção de dados é a etapa onde são identificadas quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo de KDD (GOLDSCHMIDT; PASSOS, 2005). Esses dados são originados de diversas fontes, como transações, sensores e dados de mídias sociais, podendo ser estruturados, semiestruturados ou não estruturados (AYSWARRYA, 2019).

A escolha de atributos e dos registros que serão avaliados no processo do KDD deve ser cuidadosa, selecionando-se aqueles considerados fundamentais para se chegar ao objetivo da construção do modelo.

2.1.2 Pré-Processamento

Devido ao grande volume de dados obtidos e a integração de múltiplas fontes heterogêneas, a seleção de dados é suscetível à obtenção de dados incompletos, ruidosos e inconsistentes que podem levar a resultados de baixa qualidade.

Dessa forma, o pré-processamento envolve investigar os detalhes da base de dados, como o tamanho, a qualidade dos dados e os tipos de variáveis, a fim de detectar e corrigir anomalias de forma a proporcionar dados que otimizem a eficiência das etapas posteriores, aumentando a possibilidade de adequação à tarefa que será utilizada (SCHMITT, 2005). Além disso, nesta etapa também é verificada a possibilidade de diminuir o número de variáveis envolvidas no processo, compreendendo, dentre os dados selecionados, quais estão realmente aptos a serem analisados durante o KDD, visando melhorar o desempenho dos algoritmos de análise (GOLDSCHMIDT; PASSOS, 2005).

2.1.3 Transformação

Após o pré-processamento dos dados, passa-se para a transformação. A transformação de dados é a etapa na qual o conjunto de dados bruto é convertido em uma forma padrão de uso (GOLDSCHMIDT; PASSOS, 2005). Nesta etapa, os dados são processados aplicando fórmulas matemáticas aos valores dos atributos, buscando obter informações de forma apropriada para a posterior modelagem, satisfazendo premissas de modelos ou prevenindo erros (FERREIRA, 2005).

Entre as transformações mais utilizadas, está a homogeneização da variabilidade das variáveis por meio da normalização ou padronização e a codificação de variáveis categóricas em variáveis numéricas, de acordo com a característica dos algoritmos aplicados nas etapas posteriores (FERREIRA, 2005).

Nessa etapa, é comum a aplicação de técnicas de discretização e binarização. A discretização é a técnica que converte um atributo contínuo em categórico, estabelecendo um número de categorias a serem usadas para os valores definidos. Por outro lado, a binarização é o mecanismo que converte atributos contínuos ou discretos em binários, por exemplo, substituindo as instâncias “desligado” e “ligado” pelos binários 0 e 1 (MENDES, 2011).

Portanto, diversos métodos podem ser utilizados para a transformação dos dados, sendo aplicados conforme os objetivos pretendidos, não existindo um único critério.

2.1.4 Mineração de Dados

A Mineração de Dados pode ser definida como um campo multidisciplinar, que envolve um conjunto de técnicas de exploração de grandes massas de dados de forma a descobrir padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano (AMORIM, 2006). Essa fase tem por objetivo a descoberta de padrões que possam representar informações úteis, descrevendo características do passado e predizendo tendências para o futuro (SFERRA; CORRÊA, 2003 apud AMORIM, 2006).

Em decorrência da grande diversidade de métodos de pré-processamento, são muitas as alternativas possíveis de combinações (GOLDSCHMIDT; PASSOS, 2005). Portanto, assim como nas etapas anteriores, existem diferentes técnicas e algoritmos que podem ser aplicados, e que devem ser escolhidos levando em consideração o tipo da tarefa em Mineração de Dados que é aplicada. Por fim, apresenta-se a avaliação de resultados.

2.1.5 Avaliação dos resultados

Os resultados do processo de descoberta do conhecimento podem ser representados de diversas formas, como por meio de informações gráficas ou modelos disponibilizados em diferentes aplicações.

Devido a isso, é importante se certificar de que o modelo criado conseguirá alcançar os objetivos de negócio. Isso pode ser feito a partir da aplicação de testes, os quais deverão analisar os resultados obtidos de forma criteriosa. Desse modo, será possível identificar se há a necessidade de retornar a qualquer um dos estágios anteriores do processo KDD, visando obter a confiabilidade nos modelos (GOLDSCHMIDT; PASSOS, 2005).

Portanto, a participação de conhecedores do negócio e tomadores de decisão é importante, uma vez que, no fim desta etapa, espera-se que se tenha extraído informações úteis da base de dados, e que uma decisão sobre o uso dos resultados da mineração possa ser tomada. O entendimento de todas estas etapas do KDD, auxiliará na criação de modelos confiáveis e consistentes de AM, o qual será definido a seguir.

2.2 Aprendizado de máquina

O AM é um subcampo da ciência da computação onde são criados modelos capazes de aprender com os dados (GÉRON, 2019). Pode-se definir, de forma simplificada, como a tarefa de predizer o futuro com base em fatos que aconteceram no passado (DAUMÉ III, 2012). Conforme Cerri e Carvalho (2017, p. 2), o AM é considerado “uma área de pesquisa da Inteligência Artificial que visa o desenvolvimento de programas de computador com a capacidade de aprender a executar uma dada tarefa com sua própria experiência”.

Esses protótipos estão cada vez mais sofisticados e hábeis em resolver problemas complexos graças ao apoio da ciência básica, especialmente, a Matemática e a Estatística (FREITAS; SANTANA, 2019).

Embora não seja um tema recente, a sua aplicação no âmbito organizacional tem vindo a suscitar um maior interesse nos últimos anos, e só agora passaram a perceber as verdadeiras potencialidades e impactos desta junção. Observa-se o poder

de extração de informações de forma eficiente, o que faz com que o AM seja considerado como um dos principais pilares dessa nova era da indústria (FREITAS; SANTANA, 2019).

Diante de um problema que envolve AM, deve-se decidir por qual caminho iniciar, identificando o grupo que deve ser seguido. Os principais grupos de AM são o supervisionado, semi-supervisionado, não supervisionado e por reforço. A seguir, serão descritos cada um deles.

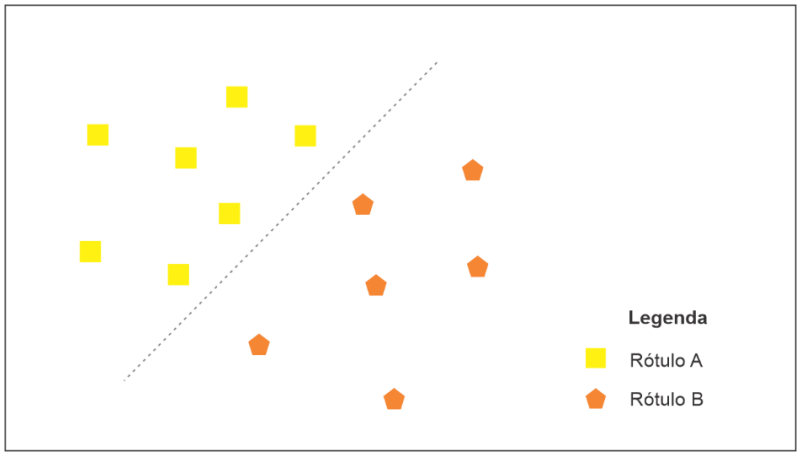
2.2.1 Aprendizagem supervisionada

Os algoritmos de aprendizagem supervisionada relacionam uma saída a uma entrada com base em dados rotulados, ou seja, ela compreende a abstração de um modelo de conhecimento a partir das informações apresentadas na forma de pares ordenados (GOLDSCHMIDT; PASSOS, 2005).

Dessa forma, para cada saída é atribuído um rótulo, que pode ser um valor numérico ou uma classe. O algoritmo determina uma forma de prever qual o rótulo de saída com base em uma entrada informada (FONTANA, 2020).

A Figura 2 ilustra o funcionamento de um classificador utilizando a abordagem supervisionada.

Figura 2 - Classificação por meio de aprendizagem supervisionada



Fonte: Elaborado pelo autor, 2021.

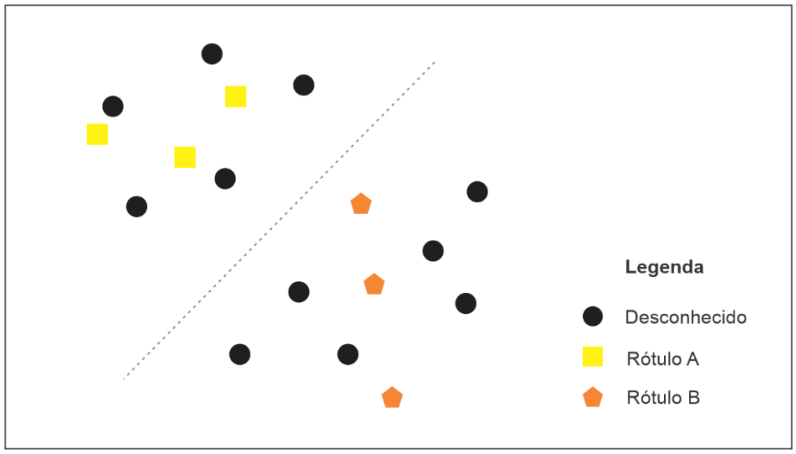
Portanto, conforme Fontana (2020), utiliza-se esta abordagem para problemas com amostras rotuladas, no qual os rótulos de saída podem assumir um conjunto de rótulos pré-definidos ou de qualquer valor real, envolvendo algoritmos de classificação e regressão respectivamente.

2.2.2 Aprendizagem semi-supervisionada

Na aprendizagem semi-supervisionada, uma parte dos dados utilizados no treinamento é rotulada, enquanto a outra consiste em dados não-rotulados. Nessa classe de problemas, não se pode assegurar que os padrões rotulados representem adequadamente o sistema a ser aprendido, restringindo o uso do paradigma supervisionado. Considerando essa abordagem, as amostras sem rótulo podem melhorar o desempenho de modelos de AM. Assim, utiliza-se padrões não rotulados como fonte de informação, garantindo maior capacidade de generalização (LELIS, 2007).

Portanto, a aprendizagem semi-supervisionada é quando se utilizam dados com e sem rótulo em um projeto para a criação de um modelo. A Figura 3 ilustra o funcionamento de um classificador utilizando a abordagem semi-supervisionada.

Figura 3 - Classificação por meio de aprendizagem semi-supervisionada



Fonte: Elaborado pelo autor, 2021.

Segundo Lelis (2007), esta abordagem geralmente é utilizada em problemas que as amostras rotuladas são difíceis de serem obtidas, e as sem rótulo, por sua vez,

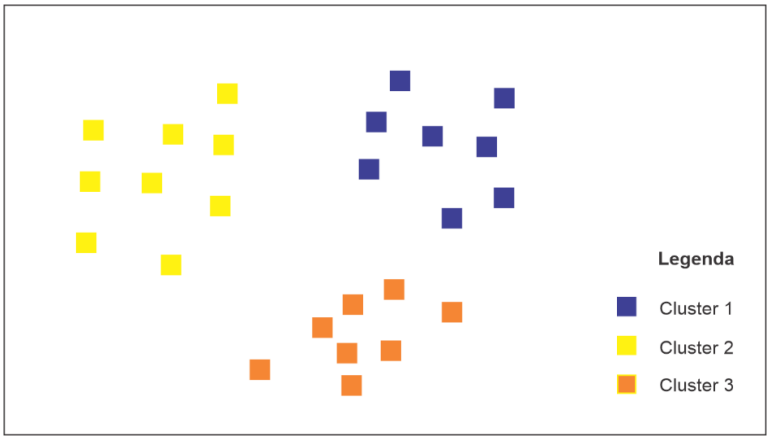
são abundantes e facilmente coletadas. O algoritmo de treinamento semi-supervisionado busca, então, gerar a superfície de separação entre as classes com base também nas amostras sem rótulo.

2.2.3 Aprendizagem não supervisionada

Os algoritmos de aprendizagem não supervisionada buscam determinar como os dados estão organizados. Esses dados de treinamento consistem apenas de exemplos de entrada, sem rótulos ou valores de saída. Nesta abordagem, o objetivo é encontrar padrões no espaço de entradas, observando quais são as regiões com maior e menor densidade de dados, por exemplo (BREVE, 2010).

Dessa forma, na aprendizagem não supervisionada os algoritmos partem dos dados de entrada, procurando estabelecer relacionamentos entre eles, sem existir o rótulo da saída desejada (BREVE, 2010). A Figura 4 representa o funcionamento de um *cluster* utilizando a abordagem não supervisionada.

Figura 4 - Clusterização por meio de aprendizagem não supervisionada



Fonte: Elaborado pelo autor, 2021.

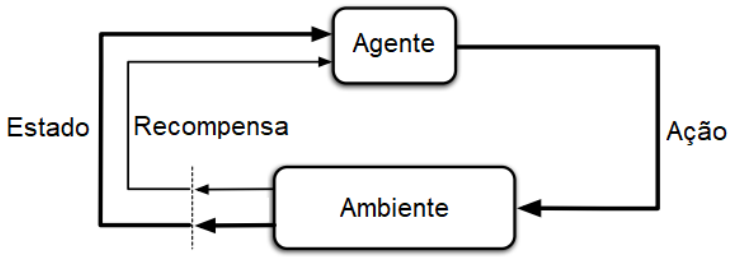
Nessa abordagem, algumas das aplicações são para tarefas que envolvem segmentação e associação de dados com tendências semelhantes.

2.2.4 Aprendizagem por reforço

Os algoritmos de aprendizagem por reforço envolvem assimilar o que fazer de modo a maximizar um valor de recompensa, considerando que para tal um agente de aprendizagem por reforço deve preferir ações que tentou no passado e descobriu ser eficaz na produção desta recompensa. Contudo, a fim de descobrir estas ações, ele deve tentar ações que não selecionou anteriormente (SUTTON; BARTO, 2015).

Portanto, esta abordagem, diferente das abordagens apresentadas anteriormente, não se caracteriza por ter ou não um rótulo de saída, mas sim pelo objetivo de aprender a se comportar em um ambiente por meio de tentativa e erro, maximizando a sua recompensa. A Figura 5 mostra o funcionamento de um modelo baseado em aprendizagem por reforço.

Figura 5 - Interação agente-ambiente na aprendizagem por reforço



Fonte: Sutton & Barto, 2015 (tradução livre).

Algumas das aplicações de aprendizagem por reforço são na robótica, jogos e gerenciamento de recursos em *clusters* de computadores (DATA SCIENCE ACADEMY, 2020).

Assim, uma vez que a abordagem de aprendizagem e a tarefa a ser aplicada são estabelecidas, deve-se considerar diferentes técnicas para a criação dos modelos, de forma a avaliar e definir o que melhor se adapta aos dados.

2.3 Técnicas de regressão

A tarefa de regressão compreende a busca por um modelo capaz de prever valores numéricos com base em dados históricos.

Algumas das técnicas que podem ser aplicadas para resolver problemas de regressão são as técnicas LASSO, *Random Forest*, Máquina de Vetores de Suporte e Redes Neurais. Estas serão discutidas a seguir.

2.3.1 LASSO

A técnica LASSO (operador de seleção e de encolhimento de mínimos absolutos, tradução livre, do inglês, *last absolute shrinkage and selection operator*) é caracterizada pela regularização, cujo objetivo é sanar problemas de multicolinearidade. Esse problema acontece quando o modelo possui coeficientes redundantes, ou seja, com alta correlação. Nessa perspectiva, o cenário pode levar a um modelo com alta variância e com possibilidade de *overfitting*, no qual o modelo se adapta muito bem aos dados de treinamento, no entanto, ao ser submetido a um conjunto desconhecido de dados, há uma perda considerável de desempenho (RASCHKA, MIRJALILI, 2017). Assim, para resolver este problema, as técnicas de regularização aplicam uma penalização aos coeficientes considerados redundantes, equilibrando a variância e o viés (PENNA, 2021).

Faz-se necessário o conhecimento da função de custo da soma residual dos quadrados a fim de compreender o funcionamento da técnica. Essa função é utilizada para ajustar a melhor curva a uma distribuição de pontos de um conjunto de dados, escolhendo os coeficientes w que minimizam a função com base nos dados de treinamento. A função de custo da soma residual dos quadrados, ou RSS (do inglês, *Residual Sum of Squares)*, é dada pela Equação 1, onde y_i é a variável dependente, w o coeficiente angular, x_i a variável independente e b o coeficiente linear.

$$RSS = \sum_{i=1}^{n} [y_i - (w \cdot x_i + b)]^2 \quad (1)$$

Ao utilizar este método em algumas bases de dados, especialmente em bases que têm muitos atributos, podem ocorrer erros de generalização elevados. A técnica LASSO diminui este problema adicionando um termo a função de custo, regularizando os coeficientes w, ou seja, restringindo seu tamanho (ALCÂNTARA, 2021).

A equação da técnica LASSO, ou RSS_{L1}, mostrada na equação 2, é parecida com a RSS, porém há a adição de um novo termo, no qual, caso este termo seja nulo, retorna-se a regressão linear múltipla tradicional com os estimadores de mínimos quadrados ordinários. Na Equação 2, nota-se a adição de um termo à função de custo, regularizando o coeficiente w. Assim, ao reduzir a função de custo, os coeficientes são automaticamente minimizados.

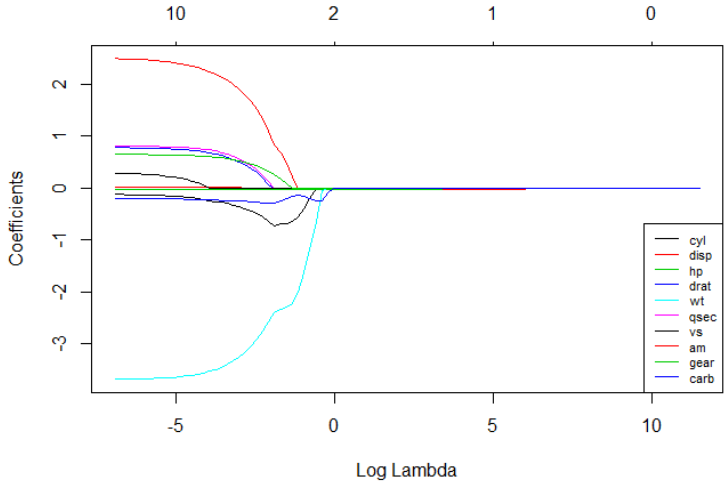
$$RSS_{L1} = \sum_{i=1}^{n} [y_i - (w \cdot x_i + b)]^2 + \lambda \sum_{j=1}^{p} |w_j| \tag{2}$$

A técnica LASSO pode ser formulada também incorporando uma restrição de penalização, conforme Equação 3.

$$\sum_{i=1}^{n} [y_i - (w \cdot x_i + b)]^2, \quad sujeito a: \lambda \sum_{j=1}^{p} |w_j| \leq s \tag{3}$$

Conforme ilustrado na Figura 6, proveniente da base de dados “mtcars”, que contém dados da revista Motor Trend US sobre características automóveis, existe um *trade-off* a ser considerado na variação do parâmetro *Lambda* ao aplicar a técnica LASSO. Observa-se que, ao aumentar o *Lambda*, o número de coeficientes diminui, o que aumenta a possibilidade de viés. Em contrapartida, ao diminuir o *Lambda*, o número de coeficientes é maior e há menor encolhimento, aumentando a variância. Sendo assim, o cenário ideal é onde existe um equilíbrio entre a variância e o viés, fazendo-se necessário a aplicação de testes com diferentes valores de *Lambda* para o modelo.

Figura 6 - Comportamento dos coeficientes em relação ao lambda



Fonte: Oleszak, 2019.

Portanto, uma das principais vantagens da utilização da técnica LASSO é que, além de realizar o encolhimento dos coeficientes, ele também faz a seleção das variáveis mais importantes, eliminando os coeficientes que possuem correlação alta com outros que foram mantidos (PEREIRA, 2021).

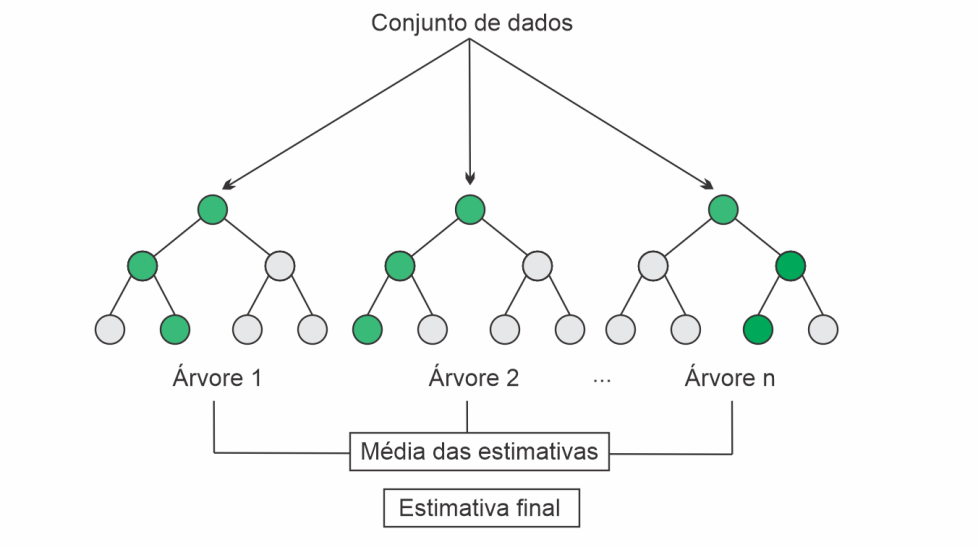
2.3.2 *Random Forest*

No AM é esperado que a combinação dos resultados de vários estimadores apresente melhor desempenho se comparada à utilização de um único estimador. Esta combinação é obtida por meio de métodos *ensemble*, o qual o *Random Forest* (RF) está incluído. O modelo gera várias árvores de decisão cujas previsões são combinadas pela média das estimativas.

O algoritmo RF faz a seleção randômica de variáveis explicativas no processo de indução da árvore. Essa seleção se trata de um sorteio feito a cada nó da árvore, apurando aleatoriamente algumas variáveis candidatas para dividir este nó. Com a utilização desta técnica, diferentes conjuntos de variáveis poderão aparecer em níveis distintos na formação de cada uma das árvores (BREIMAN, 2001).

Conforme citado, o resultado da regressão é a média da estimação feita pelas diversas árvores de decisão. O funcionamento macro do algoritmo RF para regressão é mostrado na Figura 7.

Figura 7 - Funcionamento do algoritmo *Random Forest Regressor*



Fonte: Elaborado pelo autor, 2021.

Portanto, essa técnica é utilizada eficientemente em grandes bases de dados e possibilita o processamento de milhares de variáveis, dispensando a necessidade de exclusões, pois o algoritmo realiza uma seleção de variáveis, removendo aquelas que são redundantes ou indesejáveis. Assim, o desempenho da estimação é otimizado (DANTAS; DONADIA, 2013).

2.3.3 Máquina de Vetores de Suporte

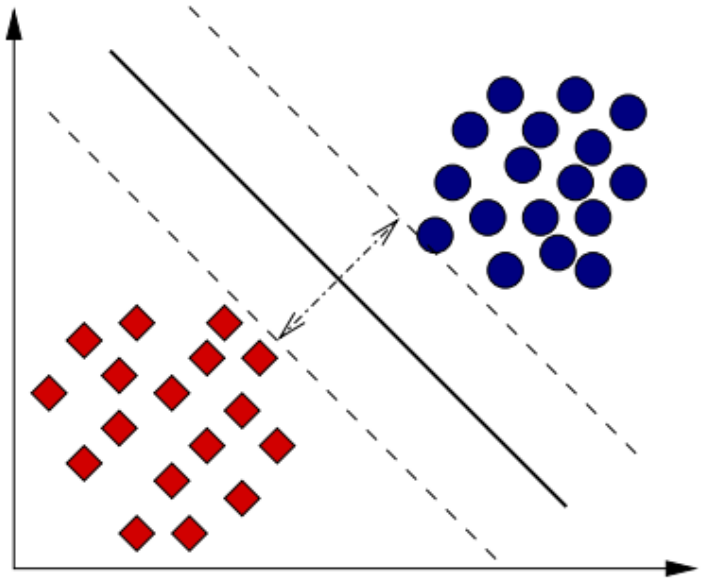
As Máquinas de Vetores de Suporte, do inglês *Support Vector Machines* (SVM) constituem em uma técnica de AM que se baseia na estratégia de dividir um espaço de características em regiões por meio de hiperplanos (SCHÖLKOPF, 1997).

O modelo mais simples de SVM trabalha apenas com dados linearmente separáveis, sendo restrito a poucas aplicações, porém apresentando propriedades importantes para formulação de SVMs mais sofisticadas. O objetivo é encontrar um

hiperplano que divide as classes, maximizando a margem de separação entre elas (SANTOS, 2002).

A Figura 8 representa a separação de classes por meio de um hiperplano de margem máxima.

Figura 8 - Hiperplano de separação para duas classes



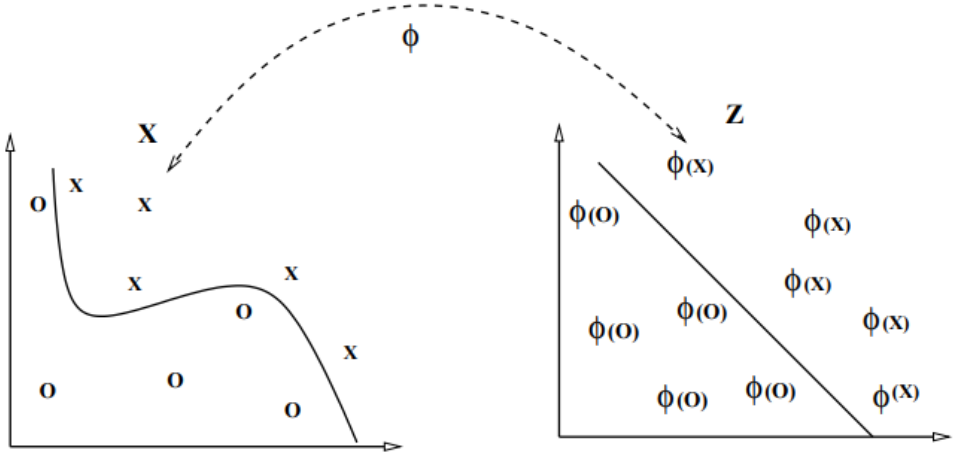
Fonte: Santos, 2002.

Para o caso linearmente separável, o algoritmo de SVM tem como objetivo encontrar este hiperplano. No entanto, quando aplicado a dados não separáveis linearmente, o classificador de margem máxima não encontra a solução desejada. Dessa forma, a fim de tornar este método capaz de manipular dados não linearmente separáveis, é necessário que seja feita uma transformação em um novo espaço de características, nos quais os padrões têm alta probabilidade de se tornarem linearmente separáveis (SANTOS, 2002). Isto pode ser feito por meio da aplicação de funções *kernel*.

As funções *kernel* projetam os dados em um espaço de características com alta dimensão para permitir a classificação em espaços não-linearmente separáveis. A Figura 9 ilustra um mapeamento de um espaço de entrada linearmente inseparável,

para um espaço de características de maior dimensão, nas quais as informações podem ser separadas linearmente.

Figura 9 - Mapeamento do espaço de entrada por meio de funções kernel



Fonte: Santos, 2002.

Conforme Haykin (1999), algumas das funções *kernel* que podem ser aplicadas são as descritas na Tabela 1.

Tabela 1 - Funções *kernel*

Tipo de *kernel*	**Função** (x_i, x_j)	**Tipo do classificador**
Polinomial	$\left((x_i . x_j) + 1\right)^p$	Máquina de aprendizagem polinomial
Gaussiano	$exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right)$	Rede RFB
Sigmoidal	$tanhtanh\left(\beta_0 \langle x_i . x_j \rangle\right) + \beta_1$	Perceptron de duas camadas

Fonte: Haykin, 1999.

Algumas das vantagens do SVM são o poder de generalização e o pequeno número de parâmetros a serem ajustados, além de ser um algoritmo útil quando há

muitas entradas, já que sua complexidade está atrelada aos vetores e não a dimensão do espaço de entrada do problema (SANTOS, 2002).

2.3.4 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) apresentam um modelo inspirado na estrutura neural do cérebro humano que adquire conhecimento por meio da experiência. Os primeiros trabalhos de Inteligência Artificial (IA) objetivaram criar RNAs baseados na hipótese de que a atividade mental consiste, principalmente, de atividade eletroquímica em redes de células cerebrais intituladas neurônios (RUSSEL; NORVIG, 1995).

Faz-se necessário compreender o funcionamento básico do sistema nervoso a fim de facilitar o entendimento de uma RNA. No sistema nervoso, o cérebro é representado por uma rede de neurônios que recebem informações e tomam decisões, por meio de receptores que convertem estímulos em impulsos elétricos que transmitem informação para a rede neural, convertendo-os em respostas para a saída do sistema. Os neurônios são conectados entre si por intermédio de sinapses, formando a chamada rede neural biológica. Nos neurônios, a comunicação é realizada por meio de impulsos que produzem uma substância neurotransmissora o qual flui do corpo celular para o axônio (CINTRA, 2018). Em resumo, conforme explica Cintra (2018), os principais constituintes de um neurônio são:

• Dendritos: São incumbidos de receber estímulos transmitidos por outros neurônios;

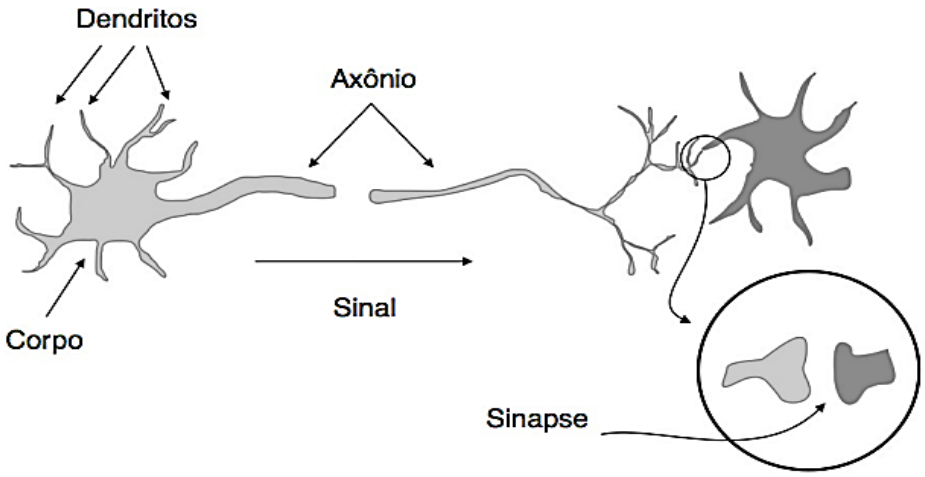
• Corpo celular: É responsável pela coleta e combinação de informações vindas de outros neurônios;

• Axônio: É responsável por transmitir estímulos entre as células;

• Sinapses: São responsáveis por transmitir as informações entre as células.

A Figura 10 representa os principais constituintes de um neurônio presente em uma rede neural biológica.

Figura 10 - Representação simplificada de um neurônio biológico



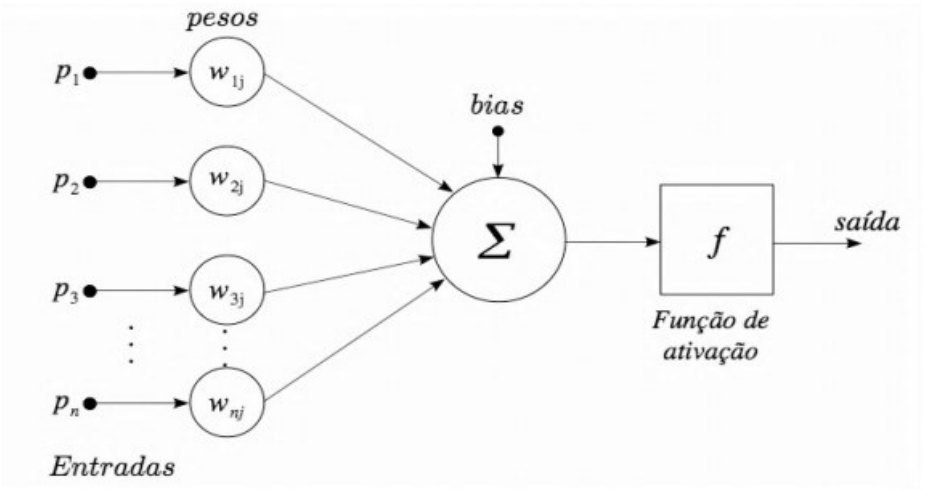
Fonte: Faceli *et al*., 2011.

Essa rede proporciona uma alta capacidade de processamento e armazenamento de informação, tornando-se um interessante objeto de estudo para o desenvolvimento de redes baseadas em modelos computacionais (FACELI *et al*., 2011).

Baseadas no modelo biológico, as RNAs são sistemas compostos por unidades de processamento alocados em uma ou mais camadas, sendo interligados por várias conexões que estão associadas a pesos que armazenam o conhecimento adquirido e servem para mensurar a entrada recebida por cada neurônio na rede (BRAGA; LUDERMIR; CARVALHO, 2000).

A Figura 11 representa o funcionamento de um neurônio artificial, onde p_i corresponde a um conjunto de padrões e w_i os pesos associados.

Figura 11 - Estrutura de um neurônio artificial



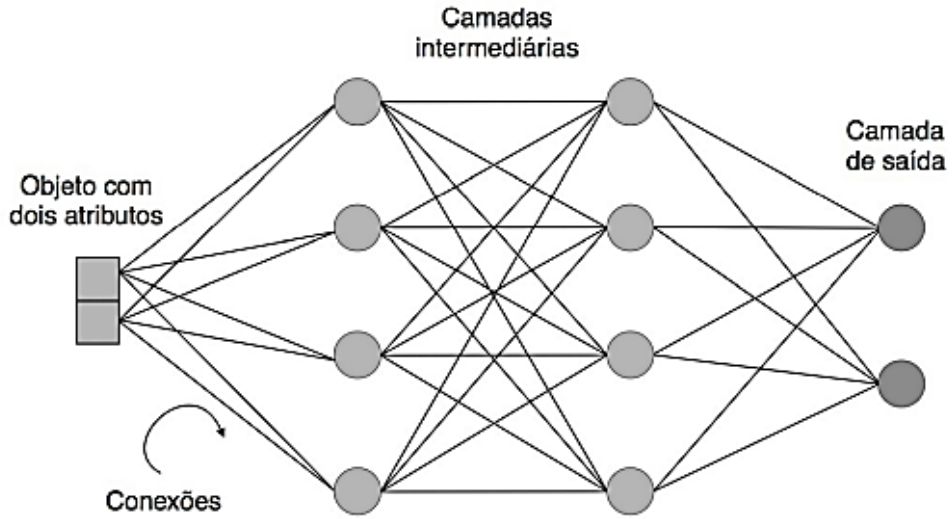
Fonte: Cintra, 2018.

Em uma RNA, é aplicado um valor a cada nó de entrada, que representa os dendritos. O valor é então passado por cada nó às conexões, multiplicando pelo peso associado. Na camada seguinte os nós recebem o valor correspondente a soma dos valores produzidos pelas conexões anteriores, e, um cálculo por meio de uma função de ativação é realizado sobre o eles, podendo ser por exemplo uma função linear ou sigmoide. Este processo é realizado nas camadas subsequentes de nós até que os nós de saída sejam alcançados. Portanto, em uma RNA, dado um conjunto de padrões p_i e de saídas desejadas, o objetivo é encontrar o conjunto de pesos ótimos w_i (CINTRA, 2018).

Assim como em uma rede neural biológica, em uma RNA, os neurônios podem estar dispostos em mais de uma camada. Nessas circunstâncias, além das camadas de entrada e de saída, tem-se as camadas ocultas, podendo um neurônio receber em seus terminais de entrada valores de saída de neurônios da camada anterior ou enviar seu valor de saída para terminais de entrada de neurônios da camada posterior (FACELI *et al*., 2011).

A Figura 12 representa uma RNA multicamadas, ilustrando a utilização de duas camadas ocultas ou intermediárias.

Figura 12 - RNA multicamadas



Fonte: Faceli *et al*., 2011.

Conforme Faceli *et al*. (2011), as RNAs são muito utilizadas para resolver problemas complexos, principalmente por terem como vantagens a tolerância a dados com ruído e por serem naturalmente paralelizáveis, o que pode acelerar o processo computacional.

2.4 Métricas de avaliação

A avaliação das estimações feitas pelas tarefas AM é realizada por meio de métricas de avaliação. Essas métricas permitem averiguar se o modelo se adaptou bem aos dados e se poderá ser utilizado nas tomadas de decisão.

Em modelos de regressão, algumas das métricas mais conhecidas são o *Mean absolute error* (MAE), o *Mean absolute percentage error* (MAPE) e o R-quadrado. As métricas MAE e MAPE apresentam valores não negativos, nas quais quanto menor o valor, melhor o modelo será (SAMPAIO *et al*., 2019). Já o R-quadrado estará sempre entre 0 e 1, considerando que quanto maior o R-quadrado, melhor é o ajuste do modelo aos dados (NASCIMENTO; ARAÚJO, 2009).

2.4.1 *Mean absolute error*

O MAE mede o afastamento entre os valores previstos e observados, constituindo na média dos erros da previsão. Dessa forma, quanto menor o valor, melhor é a previsão feita pelo modelo (MENTZER; BIENSTOCK, 1998). O MAE é definido na Equação 4, onde $\hat{y}_i$ é o valor previsto, y_i o valor real e n o tamanho da amostra.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{4}$$

2.4.2 *Mean absolute percentage error*

O MAPE trata-se do erro absoluto dividido pelo valor real, obtendo-se o afastamento do valor predito ao observado em percentual, sendo constituído pela média desses percentuais (SAMPAIO *et al*., 2019). O MAPE é definido na equação 5, onde $\hat{y}_i$ é o valor previsto, y_i o valor real e n o tamanho da amostra.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{(\hat{y}_i - y_i)}{y_i}\right| \tag{5}$$

2.4.3 R-quadrado

O R-quadrado é uma medida de qualidade do ajuste do modelo selecionado e uma medida de precisão na predição, indicando o quão próximos os dados estão da linha de regressão ajustada (NASCIMENTO; ARAÚJO, 2009).

Conforme citado, o R-quadrado, varia de 0 a 1, onde 1 indica a melhor pontuação possível e 0 que o modelo não se ajustou aos dados. O R-quadrado é definido na Equação 6, onde $\hat{y}$ é o valor estimado, $\overline{y}$ a média amostral e y o valor real.

$$R^2 = \frac{VarianciaExplicada}{VarianciaTotal} = \frac{\sum(\hat{y}-\overline{y})^2}{\sum(y-\overline{y})^2} \tag{6}$$

Observa-se que as métricas apresentadas são de interpretação simples, e, se analisadas em conjunto, podem indicar a eficiência do modelo de regressão.

2.5 Energias alternativas

O desenvolvimento de energias alternativas é considerado como uma das formas de reduzir a poluição ambiental. A busca por essas matrizes de energia aumentou com a constatação de que a humanidade enfrenta uma crise global, em face da finitude dos recursos naturais (MAFACIOLLI, 2012).

A concentração de dióxido de carbono (CO2) na atmosfera atingiu, em 2021, níveis 47% maiores se comparado às concentrações pré-industriais (NOAA, 2021). Antes da revolução industrial, a concentração deste gás era de 280 ppm (partes por milhão) (IPCC, 2007), aumentando para 390 ppm em 2010 e 414 ppm atualmente (NOAA, 2021). Algumas das consequências deste cenário são o aumento da temperatura global por conta do efeito estufa, e a falta de energia, reforçando a necessidade do equilíbrio entre a demanda e o consumo, o que pode ser alcançado com a utilização de energias alternativas (HINRICHS, 2008 apud MAFACIOLLI, 2012).

A preocupação com este cenário levou a acordos entre países, no qual foram estabelecidas as necessidades de controle sobre as intervenções humanas que levam a mudanças climáticas, devendo os mesmos reduzir emissões de Gases de Efeito Estufa (GEE). Por exemplo, na Conferência do Clima da Organização das Nações Unidas de 2021 (COP 26), aprovou-se o acordo para a redução dos combustíveis fósseis no planeta, destacando a atual crise climática (BBC, 2021).

Analisando a matriz energética mundial, percebe-se que a oferta de energia renovável do Brasil é de 46,1%, o que é muito superior ao resto do mundo, no qual somente 14,5% das fontes são renováveis (MME, 2020). Todavia, a participação do biogás na oferta interna de energia corresponde apenas a 0,09% da matriz energética brasileira (ABIOGÁS, 2020), tendo este, potencial de crescimento e, consequentemente, sendo uma importante matriz de energia sustentável para os próximos anos.

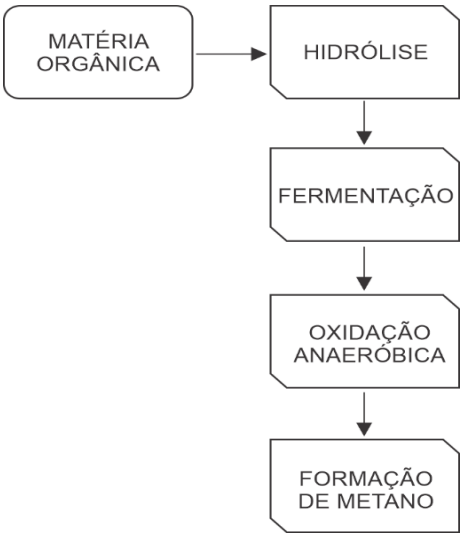
É nesse contexto que se coloca o biogás. Como resultados indiretos, a biodigestão reduz a carga orgânica da biomassa residual em tratamento sanitário e, ao gerar energia, proporciona a diminuição de emissões de GEE, pois retira dos aterros as fontes de metano (CH4) e dióxido de carbono (CO2), os dois gases mais importantes causadores do efeito estufa (BLEY JÚNIOR, 2020). Vale citar que a combustão do biogás produz CO2, porém, ainda que o processo de obtenção seja poluente, é menor se comparado com os combustíveis fósseis (OBAIDEEN *et al*., 2018).

2.5.1 Produção de biogás

O biogás é um composto gasoso, constituído, principalmente, de metano (CH4) e dióxido de carbono (CO2), além de uma pequena quantidade de outros gases, resultado da degradação anaeróbia (em ausência de oxigênio) da matéria orgânica realizada por microrganismos. Ele é considerado um recurso renovável, uma vez que faz parte do ciclo biogeoquímico do carbono. Assim, no processo de conversão da matéria orgânica até a geração do biogás, toda a matéria orgânica que estava exposta ao meio ambiente é atacada por microrganismos detritívoros (BLEY JUNIOR, 2020).

A Figura 13, representa as etapas do processo de conversão da matéria orgânica em biogás.

Figura 13 - Processo de conversão da matéria orgânica



Fonte: Adaptado de Karlsson *et al*., 2014.

Os substratos, excedentes da indústria sucroalcooleira, contêm matéria orgânica e são dosados nos biodigestores, onde são submetidos a quatro etapas da digestão anaeróbia, até a produção de biogás.

Karlsson *et al*. (2014) descreve de forma resumida as quatro etapas da digestão anaeróbia:

• Hidrólise: O material orgânico é quebrado em pequenas moléculas para que os microrganismos consigam se alimentar delas;

• Fermentação: Os ácidos são formados por meio das reações e dividem-se em ácidos orgânicos, álcoois e amoníaco (NH3), além de hidrogênio (H) e CO2. Os produtos formados dependem dos microrganismos disponíveis e de fatores ambientais;

• Oxidação anaeróbica (acetogênese): As bactérias acetogênicas convertem o material degradado em ácido acético (C2H4O2), H e CO2;

• Formação de metano (metanogênese): Esta é a etapa final do processo global de degradação anaeróbia, onde tem-se a fase de formação de CH4, o principal componente do biogás.

É importante que o processo seja acompanhado e adaptado a fim de garantir que as bactérias metanogênicas possam se sentir da melhor maneira possível, pois é o CH4, presente no biogás, que gera a maior rentabilidade (KARLSSON *et al*., 2014).

Conforme descrito na Figura 1, o biogás é produzido por meio de uma série de fases que ocorrem na ausência de oxigênio. A fim de obter um melhoramento na produção de biogás, tem-se utilizado a codigestão anaeróbia no processo de biodigestão (SILVEIRA, 2017; ALVES, 2016).

A codigestão anaeróbia é a união de diferentes tipos de resíduos passíveis de fermentação que são misturados no biodigestor, com o objetivo de melhorar o rendimento no processo de biodigestão, podendo promover o equilíbrio de nutrientes e o aumento da quantidade de matéria orgânica, por exemplo (LEITE *et al*., 2017). Portanto, quando mais de um tipo de resíduo é utilizado em conjunto, geralmente há maior desempenho do processo, favorecendo o aumento na taxa de produção de biogás (JINGURA; MATENGAIFA, 2009).

Um dos setores promissores na disponibilização de matéria prima para a geração de biogás por meio da codigestão anaeróbia é o sucroalcooleiro. Essa área gera diariamente toneladas de excedentes da produção que podem ser aproveitados na codigestão anaeróbia para produção de biogás.

Segundo Bley Junior (2020), somente o setor sucroalcooleiro poderia garantir a matéria-prima para uma produção de 20 bilhões de metros cúbicos de biogás por ano. Este setor é a principal referência em agroenergia, e os excedentes da produção, como o bagaço da cana, são utilizados como biomassa para a geração de energia por meio da queima. Porém, nesta área é possível incorporar a codigestão anaeróbia utilizando outros excedentes.

Esses excedentes, se dosados de forma consistente e equalizada, junto ao controle adequado do processo, são potenciais substratos para produção de biogás, utilizado, por exemplo, para geração de energia elétrica (KARLSSON, 2014).

O poder energético do biogás pode ser transformado convertendo a energia química em energia mecânica por processos de combustão controlada, em motores estacionários que movem geradores, promovendo a conversão direta em energia elétrica. Além disso, há a utilização do biogás, após purificado, como gás veicular, evidenciando seus benefícios econômicos e no combate ao descontrole do efeito estufa (BLEY JUNIOR, 2020).

3. TRABALHOS CORRELATOS

Esta sessão descreve alguns trabalhos publicados que tratam da aplicação de modelos de AM para a análise de fatores que influenciam a produção de biogás ou de seus compostos.

Clercq *et al.* (2019) aplicaram algoritmos preditivos a dados diários de produção de duas grandes instalações de biogás chinesas para estimar qual o valor de produção de biogás, classificar essa produção em “baixa”, “média” ou “alta” e entender quais eram os insumos mais importantes que afetavam a produção. Os modelos de AM utilizados incluíram as técnicas SVM, RF, *Extreme Gradient Boosting* (XGBoost) e KNN. O melhor resultado da estimação numérica foi obtido por meio do KNN, com R-quadrado de 0,87. Nesse estudo, os autores concluíram que os resíduos fecais municipais, restos de comida de cozinha, chorume e cama de frango foram insumos que maximizaram a produção de biogás. Os autores desenvolveram também uma ferramenta *web* baseada no modelo de AM de melhor desempenho, com o objetivo de aprimorar a capacidade analítica dos operadores de usinas de biogás.

Wang *et al.* (2021) utilizaram *Automated Machine Learning* (AutoML) por meio do *Tree-based Pipeline Optimization Tool* (TPOT) para entender como as diferentes entradas de resíduos e condições operacionais afetam o rendimento de biogás. Para esse estudo, os autores utilizaram uma base de dados com 31 atributos de resíduos e cinco atributos de parâmetros operacionais, cujas instâncias referem-se a uma base de dados de frequência diária para um período de oito anos, coletados de uma usina de codigestão anaeróbia. Para a avaliação do modelo, os autores utilizaram as métricas *Root Mean Square Error* (RMSE) e R-quadrado, além de fazer uma comparação com uma aplicação de RNA, que mostrou resultado inferior comparado ao modelo criado por meio do TPOT.

Yang *et al.* (2021) utilizaram as técnicas *Adaptive Neuro Fuzzy Inference System* (ANFIS) e *Least-Squares Support-Vector Machine* (LSSVM), baseadas em RNA e SVM, respectivamente, para comparar qual o melhor modelo para estimar a produção de biogás a partir de dados de resíduos alimentícios, frutas e vegetais. Para avaliar os modelos, os autores utilizaram as métricas MAPE, MSE e R-quadrado. O modelo criado por meio do LSSVM estimou a produção de biogás com maior precisão, se comparado ao modelo ANFIS na base de dados utilizada.

Seo *et al.* (2021) aplicaram um modelo *black box Recurrent Neural Network* (RNN) para estimar a taxa de produção de biogás a partir da digestão anaeróbia de resíduos orgânicos alimentícios. O objetivo do estudo foi construir um modelo efetivo, alternativo aos modelos baseados em processo, que, segundo os autores, são de difícil aplicação, devido à extensa caracterização do substrato e o grande número de parâmetros, que, além da quantidade, sofrem alterações com o tempo. Nesse estudo, a utilização de atributos químicos foram fatores importantes para estimar a taxa de produção de biogás.

Xiao *et al.* (2021) propuseram a utilização de um algoritmo de AM de dois estágios, a rede neural híbrida NARX-BP, para estimar a produção de metano (principal composto do biogás) via transferência direta de elétrons. A avaliação do modelo foi feita por meio das métricas R-quadrado e MSE, onde o modelo NARX-BP apresentou melhores resultados se comparado com outros modelos tradicionais de RNAs. O estudo sugeriu que o modelo de AM desenvolvido mostrou potencial no auxílio ao controle da digestão anaeróbia de águas residuais.

Neto *et al.* (2021) aplicaram um modelo de RNA para estimar a produção de biogás provindo de resíduos de frutas e legumes, alimentos sólidos e a mistura de ambos. O modelo desenvolvido apresentou valores aceitáveis de coeficiente de determinação. Observou-se também que, para o estudo realizado, a biodigestão de frutas e legumes levou à maior produção de biogás.

Shahsavar *et al.* (2021) utilizaram técnicas de AM para compor sua pesquisa relacionada ao fornecimento de energia de biogás em edifícios verdes, com o objetivo de propor um *framework* de estrutura inteligente para o fornecimento de energia de biogás integrando IA com outras metodologias. Nesse estudo foram utilizadas algumas técnicas como RF, RNA e ANFIS. Após os modelos serem avaliados com as métricas R-quadrado, MAE, MSE, *Relative Absolute Error* (RAE) e *Relative Squared Error* (RSE), a técnica ANFIS apresentou melhores resultados na previsão da produção acumulada de biogás.

Olatunji *et al.* (2022) utilizaram a técnica ANFIS para estimar a produção de biogás e metano. Para avaliação do desempenho do modelo desenvolvido foram aplicadas as métricas RMSE, MAPE, *Median Absolute Deviation* (MAD) e R-quadrado. Além disso, os autores compararam o modelo de AM com um modelo gerado a partir

do *Response Surface Methodological* (RSM). Os resultados revelaram melhor desempenho do ANFIS em relação ao RSM, com menor erro de predição e maior precisão.

Gaida *et al.* (2012) aplicaram uma combinação de RF*, Generalized Discriminant Analysis* (GerDA*)* e *Linear Discriminant Analysis* (LDA) para desenvolver um modelo preditor do vetor de estado que compõe o *Anaerobic Digestion Model nº 1* (ADM1). O ADM1 é um modelo baseado em processo que inclui múltiplos passos para modelar a digestão anaeróbia, descrevendo tanto os processos bioquímicos, como os processos físico-químicos por meio de uma série de equações diferenciais ordinárias e equações algébricas. Os resultados mostraram que o pH, a produção de biogás, teores de metano e gás carbônico e medições de conteúdo e alimentação de substrato, foram importantes atributos para a previsão do vetor de estados do ADM1.

4. METODOLOGIA

Esta sessão descreve o conjunto de dados e a sequência da pesquisa, bem como as ferramentas utilizadas em cada uma das etapas.

4.1 Conjunto de dados

Para esta pesquisa foi utilizada a base de dados privada de uma usina produtora de biogás proveniente de resíduos orgânicos da indústria sucroalcooleira. A base de dados é composta por atributos relacionados ao processo, qualidade e matéria-prima, coletadas por meio de sensores, análises laboratoriais e apontamentos manuais.

Por razões de confidencialidade dos dados, os atributos foram nomeados de acordo com o padrão utilizado comumente em usinas de biogás. Os atributos selecionados estão descritos na Tabela 2.

Tabela 2 - Atributos selecionados

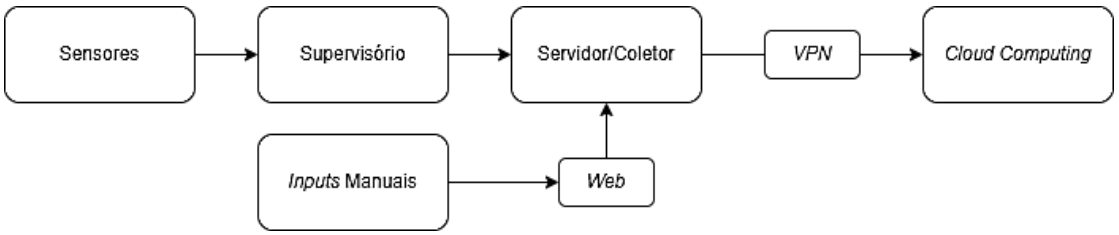
Atributo	Descrição	Unidade de medida
AG_1	Corrente do Agitador 1	A
AG_2	Corrente do Agitador 2	A
AG_3	Corrente do Agitador 3	A
PRESSURE	Pressão interna do biodigestor	mbar
PROD_LEVEL	Nível de produto no biodigestor	%
TEMPERATURE	Temperatura interna do biodigestor	ºC
FOS	Ácidos orgânicos voláteis	mg/L
TAC	Carbonato inorgânico total	mg/L
BV_DM	Matéria seca do substrato	%
BV_OM	Matéria orgânica do substrato	%
BV_PH	Potencial hidrogeniônico do substrato	pH
VIN_VOL	Volume de resíduo líquido dosado	m³
VIN_DM	Matéria seca do resíduo líquido dosado	%

VIN_OM	Matéria orgânica do resíduo líquido dosado	%
VIN_PH	Potencial hidrogeniônico do resíduo líquido dosado	pH
DW	Toneladas de resíduo seco dosado	t
DW_DM	Matéria seca do resíduo seco dosado	%
DW_OM	Matéria orgânica do resíduo seco dosado	%
DW_PH	Potencial hidrogeniônico do resíduo seco dosado	pH
BIOGAS_PROD	Produção de Biogás	Nm³

Fonte: Elaborado pelo autor, 2022.

A arquitetura de obtenção dos dados é formada por um ambiente centralizador em nuvem, conectado por meio de uma rede virtual privada ao servidor de dados da usina. O servidor de dados, por sua vez, recebe dados do sistema supervisório da planta industrial e de *inputs* manuais por meio de uma plataforma *web*, conforme ilustrado na Figura 14.

Figura 14 - Fluxo de obtenção dos dados



Fonte: Elaborado pelo autor, 2022.

A disponibilização dos dados se dá por meio de uma *Application Programming Interface* (API), abstraindo a comunicação com o ambiente centralizador em nuvem.

Utilizou-se a ferramenta *Pentaho Data Integration* para extrair, filtrar e armazenar os dados, de forma a disponibilizar dados semiestruturados para as ferramentas analíticas utilizadas. O armazenamento foi feito utilizando o banco de dados MongoDB. O fluxo de coleta está ilustrado na Figura 15.

Figura 15 - Fluxo de coleta e armazenamento dos dados

Fonte: Elaborado pelo autor, 2022.

O PDI é um *software* de análise de dados que fornece recursos de ETL (*Extract, Transform and Load*) facilitando o processo de captura, limpeza e armazenamento de dados (HITACHI, 2017). Por sua vez, o MongoDB é um banco de dados NoSQL, orientado a documentos, flexível e escalável, que armazena dados em documentos do tipo JSON (*JavaScript Object Notation*) (MongoDB, 2021).

Por fim, foram armazenados 20 atributos e 546 instâncias, utilizados para a análise, geração e implantação dos modelos. As 546 instâncias se referem aos dados diários coletados do processo de dois biodigestores idênticos, entre os meses de agosto de 2021 à abril de 2022, período em que foi implantada a arquitetura demonstrada.

4.2 Sequência da pesquisa

Para desenvolver uma ferramenta funcional, foram necessárias diversas etapas. Essas etapas incluíram a coleta de dados, conforme descrito na sessão 4.1, a construção e avaliação de modelos e o desenvolvimento de uma interface gráfica.

O tratamento e a limpeza do conjunto de dados foram realizados utilizando a biblioteca *Pandas*, uma ferramenta de análise e manipulação de dados de código aberto construída sobre a linguagem de programação *Python*. Além disso, foram utilizadas as bibliotecas *Numpy*, para manipulação dos dados, e *Scikit-learn*, para criação dos modelos. Por fim, para a implantação, utilizou-se as tecnologias *ASP.NET Core*, *Javascript* e *Flask*.

Primeiro, selecionou-se os atributos considerados importantes para o estudo, conforme descrito na Tabela 2 da sessão 4.1. Os atributos e as instâncias com mais de 60% de dados faltantes foram desconsiderados, sendo que para os atributos restantes, os dados ausentes foram substituídos por estimações de um modelo de

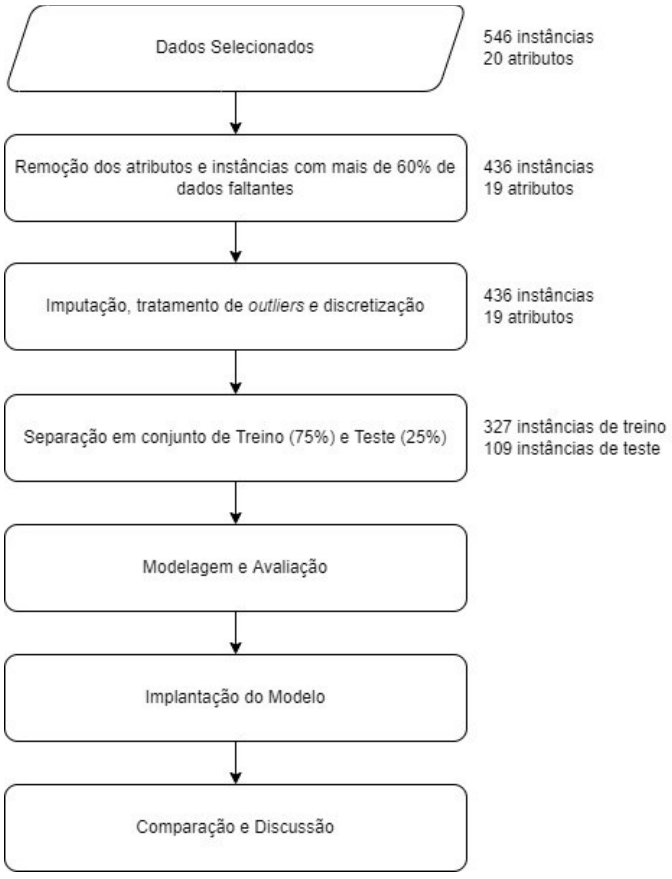
imputação iterativo (*Iterative Imputer*) da biblioteca *Scikit-learn*, no qual cada característica é modelada em função dos outros atributos, e os *outliers* foram substituídos pela média e mediana, para atributos com baixa e alta quantidade de valores discrepantes, respectivamente.

Após a transformação das instâncias, devido à disparidade de escala dos atributos, foi feita a normalização dos dados por meio do método *MinMax*, redimensionando a escala todas as variáveis para o intervalo entre 0 e 1, por meio da equação 7.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{7}$$

Em seguida, o conjunto foi separado em dados de treino e teste, sendo 25% das instâncias para teste. Por fim, foi dada sequência com a construção, avaliação e discussão dos modelos. A Figura 16 ilustra os passos descritos.

Figura 16 - Fluxograma das etapas da pesquisa



Fonte: Elaborado pelo autor, 2022.

Após a avaliação do melhor modelo, foi desenvolvida uma interface gráfica de usuário, permitindo que, por meio de uma página *web*, a estimação produção de biogás fosse obtida com a digitação do conjunto de entradas que compõem o modelo.

5. RESULTADOS E DISCUSSÕES

Nesta sessão foi feita a avaliação dos modelos de AM, a análise dos resultados obtidos e a explicação sobre as etapas para a implantação do modelo em uma plataforma *web*.

Durante a análise, verificou-se que havia excesso de dados faltantes e *outliers,* o que fez com que fosse necessária a redução de instâncias. Os principais motivos foram a falta de calibração dos sensores ou seu mal funcionamento, problemas constantemente observados em usinas de biogás. Além disso, a falta de constância no apontamento de dados de análises físico-químicas contribuiu para o excesso de dados faltantes. Observou-se ainda que haviam *outliers* referentes a dados gerados a partir de cálculos que necessitavam dos respectivos dados de análises físico-químicas que não foram lançados em alguns períodos. Devido a isso notou-se, por exemplo, uma alta dispersão nos atributos relacionados à matéria seca e matéria orgânica, apresentado em alguns casos, desvio padrão maior que a média. Foi observado ainda que o sensor de vazão de resíduo líquido é influenciado pela dosagem de água, que é feita quando há falta de matéria prima provinda do resíduo líquido gerado pela indústria de etanol. Uma possível melhoria, seria a adoção de um mapeamento, diferenciando os períodos em que o sensor recebe vazão de água e vazão de resíduo líquido, ou ainda a utilização de sensores separados, quando possível. A constância e precisão nas análises, bem como a calibração e monitoramento adequado dos sensores aumentariam a qualidade dos dados e consequentemente o resultado do modelo, já que seu treinamento é com base em dados históricos.

Os modelos de AM foram treinados por meio dos atributos selecionados após o pré-processamento, sendo removido o atributo DW_PH, pois continha mais de 60% de dados faltantes. Os atributos utilizados para o treinamento dos modelos estão descritos na Tabela 3.

Tabela 3 - Atributos utilizados para o treinamento do modelo

Atributo	Média
AG_1	22,85
AG_2	29,39
AG_3	24,95

PRESSURE	1,25
PROD_LEVEL	59,44
TEMPERATURE	44,13
FOS	1056,25
TAC	5042,38
BV_DM	14,45
BV_OM	7,52
BV_PH	7,44
VIN_VOL	73,74
VIN_DM	4,22
VIN_OM	4,85
VIN_PH	4,16
DW	105,76
DW_DM	29,27
DW_OM	18,15

Fonte: Elaborado pelo autor, 2022.

Para facilitar a intepretação das métricas de avaliação, foi considerado o valor médio de produção de biogás (atributo *target*) de aproximadamente 13500 Nm³.

5.1 Avaliação dos modelos

A partir da separação do conjunto de treino (75%) e teste (25%) foi feita a modelagem e avaliação dos modelos. Inicialmente combinou-se manualmente os parâmetros de cada algoritmo treinado, com o intuito de escolher o modelo com os melhores resultados para posteriormente realizar o aprimoramento de hiperparâmetros.

Para avaliar os modelos foram utilizadas as métricas R-quadrado, MAE e MAPE. Os resultados obtidos estão descritos na Tabela 4.

Tabela 4 - Resultado dos modelos

Métricas	***Random Forest***		**SVM**		**RNA**		**LASSO**	
	Treino	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**

R-quadrado	0,95	0,86	0,78	0,74	0,64	0,59	0,66	0,67
MAE	499,55	1075,04	1538,60	1564,84	1842,75	1875,01	1768,02	1674,58
MAPE	0,08	0,10	0,18	0,22	0,23	0,25	0,23	0,21

Fonte: Elaborado pelo autor, 2022.

É possível notar que o algoritmo que apresentou o melhor desempenho foi o *Random Forest*, se sobressaindo entre os algoritmos avaliados, com R-quadrado de 0,95 na base de treinamento e 0,86 na base de teste.

Para avaliar a possibilidade de obter melhores resultados por meio da utilização de outros algoritmos, foi utilizada a biblioteca *Lazy Predict*. O *Lazy Predict* ajuda a entender quais modelos tendem a ter um melhor resultado, por meio do treinamento, avaliação e ranqueamento de vários modelos. A Tabela 5 mostra os melhores algoritmos ranqueados pelo *Lazy Predict*.

Tabela 5 - Resultado dos modelos com a biblioteca Lazy Predict

Modelo	**R-quadrado**
RandomForestRegressor	0,87
Hist Gradient Boosting Regressor	0,86
LGBMRegressor	0,86
ExtraTreesRegressor	0,86
XGBRegressor	0,85
Bagging Regressor	0,85
GradientBoostingRegressor	0,84
AdaBoostRegressor	0,82
NuSVR	0,72
SVR	0,69
SGDRegressor	0,68
RidgeCV	0,68
BayesianRidge	0,67
LassoLarsIC	0,67
Ridge	0,65
TransformedTargetRegressor	0,64
LinearRegression	0,64

KNeighborsRegressor	0,64
ElasticNetCV	0,62
LassoCV	0,61

Fonte: Elaborado pelo autor, 2022.

Os resultados demonstram que os modelos que obtiveram os melhores desempenhos são os baseados em métodos *ensemble*, como é o caso do RF. As possíveis razões para tal resultado podem ser explicadas pelas vantagens que algoritmos baseados em árvores de decisão tem sobre outros algoritmos. A natureza aleatória da construção das árvores minimiza o sobreajuste, selecionando os atributos mais importantes de acordo com uma métrica de avaliação interna. Além disso, o algoritmo RF possibilita um método de treinamento efetivo em bases de dados que contêm valores discrepantes (DANTAS; DONADIA, 2013).

Assim, como tentativa de melhorar o resultado dos modelos escolhidos inicialmente para o estudo (SVM, RNA e LASSO), selecionou-se parte dos atributos para o treinamento dos modelos, sendo eles: PROD_LEVEL, TEMPERATURE, TAC, BV_DM, BV_OM, BV_PH, VIN_VOL, VIN_OM. Por fim, foi feito o treinamento com diferentes combinações de parâmetros utilizando a biblioteca *GridSearchCV*. Os resultados estão descritos na Tabela 5.

Tabela 6 - Avaliação de hiperparâmetros SVM, RNA e LASSO

Modelo	Parâmetros testados	Valor Selecionado	R-quadrado
SVM	tol: 0,001; 0,0001; 0,00001 C: 1, 1,5; 2 kernel: rbf, poly, sigmoid	tol: 0,001 C: 1 kernel: rbf	0,67
RNA	activation: relu, logistic solver: adam, sgd batch_size: 10; 56	activation: relu solver: adam batch_size: 10	0,63
LASSO	alpha: 0,0001; 0,001; 0,01; 1; 10; 100	0,0001	0,56

Fonte: Elaborado pelo autor, 2022.

Os resultados demonstraram que o modelo RNA obteve melhora, aumentando o R-quadrado de 0,59 para 0,63. Os modelos SVM e LASSO obtiveram queda com os novos testes, indicando que a utilização dos atributos e parâmetros testados não foram suficientes para melhorar os modelos.

Como os testes iniciais indicaram uma melhor performance do modelo gerado por meio do algoritmo RF, o mesmo foi utilizado para continuidade do estudo. Inicialmente também foram analisados os diferentes parâmetros do RF que maximizam os resultados do modelo, sendo estes mostrados na Tabela 7.

Tabela 7 - Avaliação de hiperparâmetros RF

Parâmetro	Valor padrão	Valores testados	Valor selecionado
n_estimators	100	50, 150, 250 e 500	150
min_samples_leaf	1	1, 5 e 10	1
min_samples_split	2	2, 5 e 10	2

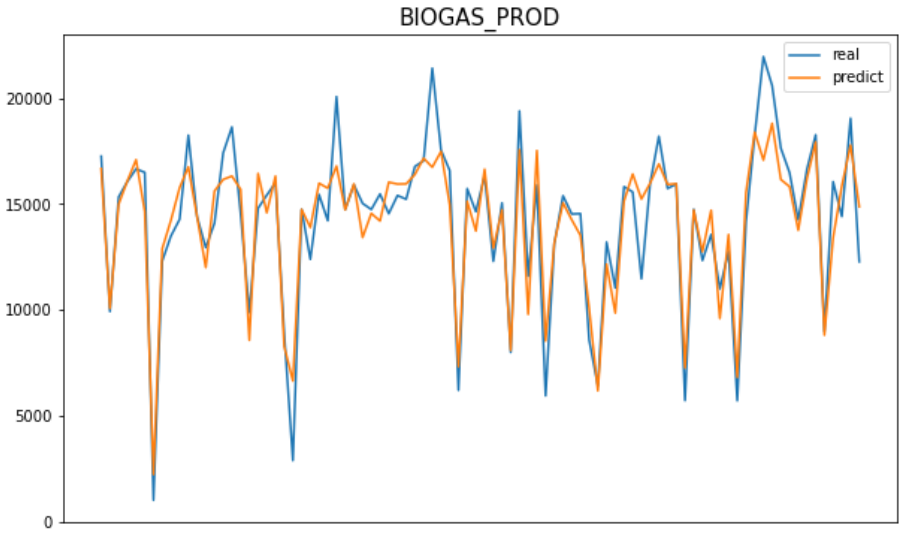
Fonte: Elaborado pelo autor, 2022.

Visto que cada execução do algoritmo gera métricas de valores diferentes (devido aos diferentes conjuntos de treinamento criados em cada análise), para a obtenção de um número que melhor o avalie foi realizada uma validação cruzada k-*fold*, com o parâmetro k igual a 5. Dessa maneira, foram feitos 30 testes, de forma que, em cada um deles, o conjunto de dados é dividido em 5 e o R-quadrado é calculado individualmente em cada conjunto para a obtenção de um número médio. O R-quadrado foi de 0,83, tendo uma pequena redução em relação ao resultado obtido com os parâmetros combinados manualmente e antes da validação cruzada.

5.2 Análise dos resultados

Para analisar o comportamento da estimação, foi plotado o gráfico comparando os valores preditos na base de teste com os valores reais, conforme ilustrado na Figura 17.

Figura 17 - Comparação entre estimação do modelo e valores reais

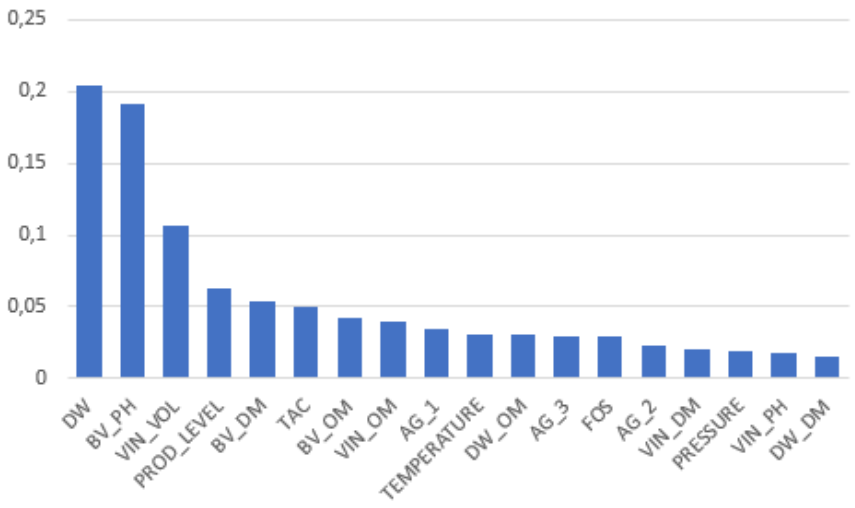


Fonte: Elaborado pelo autor, 2022.

Observa-se que o modelo obteve comportamento semelhante ao real nos dados intermediários e em picos de baixa produção, porém o modelo não se adaptou bem aos dados nos picos de maior produção.

Os resultados obtidos pelo modelo RF permitem ainda que seja possível conhecer a importância de cada atributo para a estimação. A partir do modelo gerado foi extraída a importância das variáveis, conforme ilustrado na Figura 18.

Figura 18 - Importância dos atributos para a estimação do modelo



Fonte: Elaborado pelo autor, 2022.

Os atributos mais relevantes para o modelo foram a quantidade de resíduo seco dosado, o pH do substrato e a quantidade de resíduo líquido dosado, respectivamente.

Analisando o coeficiente de correlação entre os atributos de dosagem de resíduos e a produção de biogás, observa-se que há correlação forte (0,76) para o resíduo seco e moderada (0,6) para o resíduo líquido, indicando que, para o conjunto de dados estudado, houve maior geração de biogás nos dias em que houve maior dosagem de resíduos. Porém, é importante ressaltar que nem sempre o aumento da dosagem e resíduos é ideal para o processo de biodigestão. Conforme Karlsson *et al*. (2014), o mais importante para a quantificação do biogás é a composição do substrato, o que está diretamente ligada à quantidade de nutrientes e contaminantes potenciais. Assim, o ideal é que haja acompanhamento dos parâmetros de controle e constância na alimentação, sem alterações bruscas, mantendo um ambiente favorável a produção de biogás e evitando a inibição das bactérias.

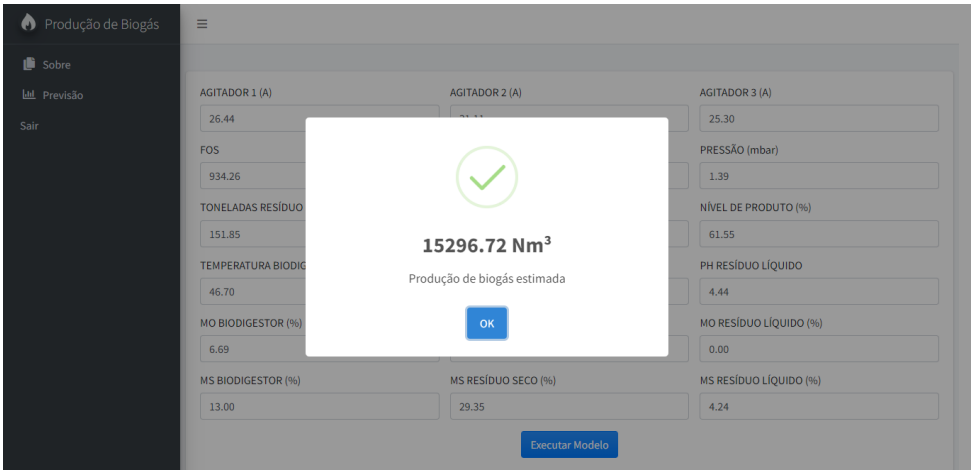
Com relação ao pH, conforme Chernicharo (2007), nas etapas de acetogênese e metanogênese é importante o controle do pH na faixa de 6,5 a 8, pois pode ocorrer inibição das bactérias metanogênicas pelo acúmulo de amônia dissolvida com o aumento de pH, ou uma maior produção de ácidos orgânicos voláteis com a diminuição do pH. Dessa forma, pequenas alterações, para menores ou maiores valores de pH, podem influenciar a produção de biogás e metano.

5.3 Implantação do modelo

Como o modelo RF obteve o melhor desempenho para o conjunto de dados estudado, o mesmo foi utilizado para realizar as estimações por meio de uma aplicação *web*. O objetivo desse desenvolvimento foi sugerir um método de implantação capaz de disponibilizar de forma simples o modelo para ser consumido na maioria das linguagens de programação, facilitando a integração com diferentes plataformas. Vale citar que, para soluções mais robustas, existem ferramentas específicas para esse tipo de tarefa, o que, dependendo do cenário, pode ser viável em termos de segurança, disponibilidade e tempo de desenvolvimento.

Com base no modelo gerado, a plataforma *web* permite que os usuários especifiquem entradas de atributos para o modelo, que retorna uma estimação da produção de biogás, conforme ilustrado na Figura 19.

Figura 19 - Estimação na plataforma *web*



Fonte: Elaborado pelo autor, 2022.

Para testar o modelo integrado a plataforma, foram coletados os dados reais da usina estudada referente ao período de 1 mês após a coleta dos utilizados no treinamento. Os resultados estão descritos na Tabela 8.

Tabela 8 - Comparação da estimação com valores reais

Valor de produção real (Nm³)	Valor estimado pelo modelo (Nm³)	Erro percentual (%)
15271	15296	0,16
17941	17656	1,59
7500	8033	7,11

Fonte: Elaborado pelo autor, 2022.

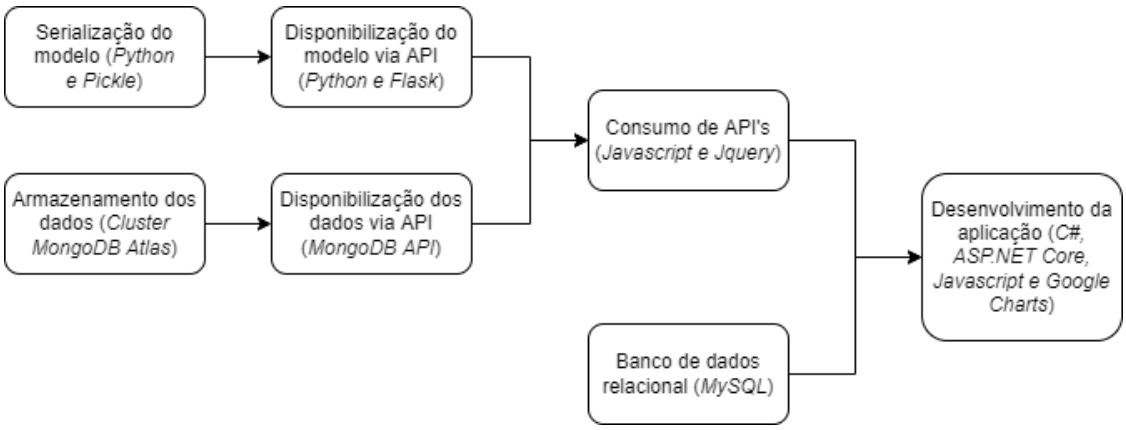
Além disso, foi desenvolvida uma página com uma breve explicação sobre o modelo, além de gráficos de amostra dos dados utilizados no treinamento, conforme ilustrado na Figura 20.

Figura 20 - Gráficos na plataforma *web*

Fonte: Elaborado pelo autor, 2022.

O modelo foi disponibilizado por meio de uma API, desenvolvida utilizando o *framework Flask* e linguagem de programação *Python*. Por meio de uma requisição de método POST, a API recebe os atributos de entrada em formato JSON, processa por meio do modelo de AM, e retorna o valor estimado da produção de biogás. Para consumir a API na aplicação e mostrar o valor retornado na plataforma web, foi utilizada a biblioteca de funções *Javascript Jquery*, que, após receber o valor retornado, envia esse dado para o código HTML a ser processado pelo navegador. O mesmo fluxo se dá para a geração dos gráficos, com a diferença de que o MongoDB disponibiliza API própria, que foi consumida e integrada ao serviço de geração de gráficos *Google Charts*. Por fim, a linguagem C# foi utilizada para o desenvolvimento do *back-end* do sistema, por meio do *framework* ASP.NET CORE. O tema da interface HTML utilizada no *front-end* foi o *AdminLTE 3* baseado em *Bootstrap*. Um banco de dados *MySQL* foi criado e integrado a aplicação para armazenar dados de autenticação da plataforma em uma tela de *login*. A Figura 21 ilustra os passos para a integração do modelo à aplicação web.

Figura 21 - Fluxo de integração do modelo com a plataforma *web*



Fonte: Elaborado pelo autor, 2022.

Para utilizar a plataforma *web* em produção, algumas melhorias podem ser desenvolvidas, como por exemplo aceitar a entrada de dados por meio do *upload* de um arquivo padronizado, diminuindo o tempo de imputação de dados para a predição. Além disso, a plataforma pode ser combinada com outras estruturas de análise de dados, podendo criar sistemas de gerenciamento de geração de biogás mais eficientes.

Vale citar que os algoritmos de AM são baseados em dados históricos, portanto o erro de previsão pode aumentar com a utilização de novos tipos de matéria prima ou alteração das características do ambiente em que o biodigestor está inserido, fazendo-se necessário o treinamento com os novos dados. Isso pode ser feito, por exemplo, através de softwares que possibilitem o treinamento constante ou com base em dados *near real-time*, tornando a ferramenta mais robusta.

6. CONCLUSÃO

Com o objetivo de estimar a produção de biogás foram treinados diversos modelos de AM, capazes de realizar tal estimação a partir de dados de análises e de sensores de uma usina que utiliza como matéria prima resíduos da indústria sucroalcooleira.

Os modelos treinados incluíram os algoritmos RF, SVM, RNA e LASSO. O modelo que obteve o melhor resultado foi o RF, com R-quadrado de 0,83. Também foi desenvolvida uma interface gráfica integrada ao modelo treinado, permitindo realizar a estimação através da entrada de um conjunto de dados referentes aos atributos utilizados.

Assim, conclui-se que o modelo criado pode ser útil principalmente se incorporado a outras ferramentas analíticas, facilitando a tomada de decisão quanto aos procedimentos operacionais do dia-a-dia de uma usina de biogás. No entanto, conforme citado no capítulo 5, devido às constantes mudanças em um ambiente que depende de questões biológicas e que pequenas alterações podem ter grande influência, soluções mais robustas que possibilitem o treinamento constante podem trazer análises mais eficientes, além de gerar ferramentas visuais que auxiliem o acompanhamento em *near real-time*.

REFERÊNCIAS

ABIOGÁS. VII Fórum do Biogás. **Biogás na matriz energética brasileira**. 2020. Disponível em: <https://abiogas.org.br/wp-content/uploads/2021/01/VII-Forum-Biogas-JM-05-11-2020-1.pdf>. Acesso em: 03 out. 2021.

ALCÂNTARA JUNIOR, G. P. **Avaliação do lasso e métodos alternativos em modelos de regressão logística.** 2021. Dissertação. (Mestrado em Estatística) - Universidade Federal de São Carlos, São Carlos, 2021. Disponível em: <https://repositorio.ufscar.br/bitstream/handle/ufscar/14052/disserta%C3%A7%C3%A3o.pdf?sequence=1>. Acesso em: 15 nov. 2021.

ALVES, I. R. F. S. **Avaliação da codigestão na produção de biogás**. 2016. Tese (Doutorado em Ciências da Engenharia Civil) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016. Disponível em: <http://www.coc.ufrj.br/pt/documents2/doutorado/2016-1/2740-ingrid-roberta-de-franca-soares-alves/file>. Acesso em: 10 out. 2021.

AMORIM, T. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados**. 2006. Monografia (Graduação em Ciência da Computação) - Universidade Federal de Pernambuco, Pernambuco, 2006. Disponível em: <https://www.cin.ufpe.br/~tg/2006-2/tmas.pdf>. Acesso em: 30 out. 2021.

ARAÚJO, A. P. C. **Produção de biogás a partir de resíduos orgânicos utilizando biodigestor anaeróbico**. 2017. Trabalho de Conclusão de Curso (Graduação em Engenharia Química) - Universidade Federal de Uberlândia, Minas Gerais, 2017. Disponível em: <https://repositorio.ufu.br/handle/123456789/20292>. Acesso em: 09 out. 2021.

AYSWARRYA, G. Data Sources 101. **KDNuggets**, out. 2019. Disponível em: <https://www.kdnuggets.com/2019/10/data-sources-101.html>. Acesso em: 23 out. 2021.

BLEY JÚNIOR, C. **Biogás**: a energia invisível. 2. ed. São Paulo: CIBiogás, 2015.

BRAGA, A.; LUDERMIR, T. B.; CARVALHO, A. C. P. L. F. **Redes neurais artificiais**: teoria e aplicações. Rio de Janeiro: Editora S.A., 2000.

BREIMAN, L. Random Forest. **Machine Learning**, v. 45, n. 1, p. 5-32, out. 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Acesso em: 06 nov. 2021.

BREVE, F. A. **Aprendizado de máquina em redes complexas**. 2010. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Universidade de São Paulo, São Carlos, 2010. Disponível em: <https://teses.usp.br/teses/disponiveis/55/55134/tde-21092010-104722/pt-br.php>. Acesso em 31 out. 2021.

CAMILO, C O.; SILVA, J. C**. Mineração de Dados**: Conceitos, Tarefas, Métodos e Ferramentas. 2009. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Goiás, Goiás, 2009. Disponível em: <https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 30 out. 2021.

CERRI, R.; CARVALHO, A. C. P. L. F. **Aprendizado de máquina:** breve introdução e aplicações. Cadernos de Ciência & Tecnologia, Brasília, v. 34, n. 3, p. 297-313, set./dez. 2017. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/184785/1/Aprendizado-de-maquina-breve-introducao.pdf>. Acesso em 31 out. 2021.

CHERNICHARO, C. A. L. **Princípios do Tratamento Biológico de Águas Residuárias e Reatores anaeróbicos**. 2. ed. Departamento de Engenharia Sanitária e Ambiental, Universidade Federal de Minas Gerais, Minas Gerais, 2007.

CINTRA, R. **Introdução a neurocomputação**. São Paulo: INPE, 2018. Disponível em: <http://www.inpe.br/elac2018/arquivos/ELAC2018_MC3_apostila.pdf>. Acesso em: 13 nov. 2021.

CLERCQ, D. D.; JALOTA, D.; SHANG, R.; NI, K.; ZHANG, Z.; KHAN, A.; WEN, Z.; CAICEDO, L.; YUAN, K. Machine learning powered software for accurate prediction of biogas production: a case study on industrial-scale Chinese production data. **Elsevier**, v. 218, p. 390-399, 1 maio 2019. Disponível em: <https://doi.org/10.1016/j.jclepro.2019.01.031>. Acesso em: 17 out. 2021.

COP26: how much is spent supporting fossil fuels and green energy? **BBC News**, 15 nov. 2021. Disponível em: <https://www.bbc.com/news/59233799>. Acesso em: 27 nov. 2021.

DANTAS, D.; DONADIA, E. **Comparação entre as técnicas de regressão logística, árvore de decisão, *bagging* e *random forest* aplicadas a um estudo de concessão de crédito.** 2013. Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Federal do Paraná, Curitiba, 2013. Disponível em: <http://www.coordest.ufpr.br/wp-content/uploads/2018/12/TCC_DanielEricson.pdf>. Acesso em: 06 nov. 2021.

DATA SCIENCE ACADEMY. **Deep Learning Book:** Aplicações da Aprendizagem Por Reforço no Mundo Real. 2021. Disponível em: <https://www.deeplearningbook.com.br/aplicacoes-da-aprendizagem-por-reforco-no-mundo-real/>. Acesso em: 02 nov. 2021.

DAUMÉ III, H. **A course in machine learning**. 2017. Disponível em: <http://ciml.info/>. Acesso em: 31 out. 2021.

FACELI, K.; LORENA, A. C.; GAMA, J.; ALMEIDA, T. A.; CARVALHO, A. C. P. L. F. **Inteligência Artificial:** uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **Advances in Knowledge Discovery & Data Mining**.California: American Association for Artificial Intelligence, 1996.

FERREIRA, J. B. **Mineração de dados na retenção de clientes em telefonia celular**. 2005. Dissertação (Mestrado em Engenharia Elétrica) - Faculdade de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2005. Disponível em: <https://www.maxwell.vrac.puc-rio.br/7070/7070_1.PDF>. Acesso em: 24 out. 2021.

FONTANA, E. **Introdução aos algoritmos de aprendizagem supervisionada.** 2020. Universidade Federal do Paraná, Paraná. Disponível em: <https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf>. Acesso em: 31 out. 2021.

FREITAS, A. L.; JÚNIOR SANTANA, O. V. Machine Learning: desafios para um Brasil competitivo. **Revista da Sociedade Brasileira de Computação**, Porto Alegre, v. 39, n. 1, p. 7-10, 2019. Disponível em: < https://www.sbc.org.br/images/flippingbook/computacaobrasil/computa_39/pdf/CompBrasil_39_180.pdf>. Acesso em: 31 out. 2021.

GAIDA, D.; WOLF, C.; MEYER, C.; Stuhlsatz, A.; Lippel, J.; Back, T.; Bongards, M.; McLoone, S. **State estimation for anaerobic digesters using the ADM1**. WATER SCIENCE AND TECHNOLOGY, v. 66, n. 5, p. 1088–1095, 2012.

GÉRON, A. **Mãos à obra**: aprendizado de máquina com Scikit-Learn & TensorFlow. Rio de Janeiro: Alta Books, 2019.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining:** um guia prático. 4. Reimpressão. Elsevier: Rio de Janeiro, 2005.

HAYKIN, S.; ENGEL, P. M. (Trad). **Redes neurais**: princípios e prática. 2. ed. [S.l.]: Bookman, 2000.

HITACHI. **Pentaho Data Integration**. 2017. Disponível em: < https://help.hitachivantara.com/Documentation/Pentaho/7.1/0D0/Pentaho_Data_Integration>. Acesso em: 15 out. 2022.

IPCC. Summary for Policymakers. **Climate Change 2007:** The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007. Disponível em: <https://www.ipcc.ch/site/assets/uploads/2018/02/ar4-wg1-spm-1.pdf>. Acesso em: 02 out. 2021.

JEONG, K.; ABBAS, A.; SHIN, J.; SON, M.; KIM, Y. M.; CHO, K. H. Prediction of biogas production in anaerobic co-digestion of organic wastes using deep learning models. **Elsevier**, v. 205, out. 2021. Disponível em: <https://doi.org/10.1016/j.watres.2021.117697>. Acesso em: 17 out. 2021.

JINGURA, R. M.; MATENGAIFA, R. Optimization of biogas production by anaerobic digestion for sustainable energy development in Zimbabwe. **Renewable and Sustainable Energy Reviews**, v. 13, n. 5, p. 1116-1120, jun. 2009. Disponível em: <https://doi.org/10.1016/j.rser.2007.06.015>. Acesso em: 17 out. 2021.

KARLSSON, T.; KONRAD, O.; LUMI, M.; SCHMEIER, N. P.; MARDER, M.; CASARIL, C. E.; KOCH, F. F.; PEDROSO, A. G. **Manual básico de biogás**. Lajeado: Editora Univates, 2014.

LEITE, V. D.; BARROS, A. J. M.; MENEZES, J. M. C.; SOUZA, J. T. S.; LOPES, W. S. Codigestão anaeróbia de resíduos orgânicos. **Revista DAE**, v.65, n.208, p.35-46,

out. 2017. Disponível em: <https://doi.org/10.4322/dae.2017.004>. Acesso em: 21 nov. 2021.

LELIS, L. H. S. **Aprendizagem semi-supervisionada aplicada à engenharia financeira.** 2007. Dissertação (Mestrado Engenharia Elétrica) - Universidade Federal de Minas Gerais, Minas Gerais, 2007. Disponível em: <https://www.ppgee.ufmg.br/defesas/392M.PDF>. Acesso em: 31 out. 2021.

MAFACIOLLI, D. **Produção de Biogás através do processo de digestão anaeróbia utilizando dejetos de aves de postura com suplementação de glicerina bruta.** 2012. Monografia (Graduação em Engenharia Ambiental) - Centro Universitário UNIVATES, Tocantins. Disponível em: <https://www.univates.br/bdu/bitstream/10737/424/1/DeboraMafaciolli.pdf>. Acesso em: 03 out. 2021.

MENDES, L. **Data Mining**: estudo de técnicas e aplicações na área bancária. 2011. Monografia (Tecnólogo em Processamento de Dados) - Faculdade de Tecnologia de São Paulo, 2011. Disponível em: <http://www.fatecsp.br/dti/tcc/tcc0031.pdf>. Acesso em: 24 out. 2021.

MENTZER, J. T.; BIENSTOCK, C. C. **Sales Forecasting Management**: understanding the techniques, systems and management of the sales forecasting process. California: Sage, 1998.

MINISTÉRIO DE MINAS E ENERGIA. **Resenha Energética Brasileira**, 2020. Disponível em: http://antigo.mme.gov.br/documents/36208/948169/Resenha+Energ%C3%A9tica+Brasileira+-+edi%C3%A7%C3%A3o+2020/ab9143cc-b702-3700-d83a-65e76dc87a9e. Acesso em: 03 out. 2021.

MongoDB. **O que é o MongoDB?** 2021. Disponível em: <mongodb.com/pt-br/what-is-mongodb >. Acesso em: 15 out. 2022.

NASCIMENTO, G.; ARAÚJO, P. F. **Estudo acerca do coeficiente de determinação nos modelos lineares e algumas generalizações.** 2009. Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Federal do Paraná, Curitiba, 2009. Disponível em: <https://docs.ufpr.br/~lucambio/CE229/TCC_Patricia_e_Gisele.pdf>. Acesso em: 10 nov. 2021.

GONÇALVES NETO, J.; VIDAL OZORIO, L.; CAMPOS DE ABREU, T. C.; FERREIRA DOS SANTOS, B.; PRADELLE, F. **Modeling of biogas production from food, fruits and vegetables wastes using artificial neural network (ANN).** Fuel, v. 285, p. 119081, 2021. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016236120320779>.

NOAA. Earth System Research Laboratories. **Carbon Cycle Greenhouse Gases:** Trends in CO2. Boulder, USA. Disponível em: <https://gml.noaa.gov/ccgg/trends/>. Acesso em: 02 out. 2021.

OBAIDEEN, K.; ABDELKAREEM, M. A.; WILBERFORCE, T.; ELSAID, K.; SAYED, E. T.; MAGHRABIE, H. M.; OLABI, A. G. **Biogas role in achievement of the sustainable development goals:** Evaluation, Challenges, and Guidelines. Journal of the Taiwan Institute of Chemical Engineers. Volume 131, 2022.

OLATUNJI, K. O.; AHMED, N. A.; MADYIRA, D. M.; ADEMAYO, A. O.; OGUNKUNLE, O.; ADELEKE, O. **Performance evaluation of ANFIS and RSM modeling in predicting biogas and methane yields from Arachis hypogea shells pretreated with size reduction.** Renewable Energy, v. 189, p. 288–303, 2022. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0960148122002361>.

OLESZAK M. **Regularization**: ridge, lasso and elastic net. nov. 2019. Disponível em: <https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net>. Acesso em: 15 nov. 2021.

PENNA, I. L. **Seleção de modelos de regressão linear em bases de alta dimensão**. 2021. Trabalho de Conclusão de Curso (Graduação em Estatística) Universidade Federal de Juiz de Fora, Juiz de Fora, 2021. Disponível em: <https://repositorio.ufjf.br/jspui/bitstream/ufjf/12727/1/isabelalopespenna.pdf>. Acesso em: 15 nov. 2021.

PYTHON SOFTWARE FOUNDATION. Python Language Site: Documentation, 2020. Página de documentação. Disponível em: <https://www.python.org/doc/>. Acesso em: 01 de fev. de 2022.

RASCHKA, S.; MIRJALILI, V. **Python machine learning**: machine learning and deep learning with python, scikit-learn, and tensorflow. 2. ed. Birmingham: Packt Publishing, 2017.

REGRESSÃO RÍGIDA E LASSO. Direção e Produção: Departamento de Engenharia Mecânica e de Produção. São João del-Rei: Universidade Federal de São João del-Rei, 2020. Online. Disponível em: <https://youtu.be/CJ2pi1Iw0JI>. Acesso em: 15 nov. 2021.

RUSSELL, S.; NORVIG P. **Artificial Intelligence**: a modern approach. 3. ed. Londres: Pearson Education Limited, 1995.

SAMPAIO, I. G.; BERNARDINI, F.; PAES, A.; ANDRADE, E. O.; VITERBO, J. Avaliação de modelos de predição e previsão construídos por algoritmos de aprendizado de máquina em problemas de cidades inteligentes. In: SANTOS, R.; MARTINOTTO, A. **Tópicos em sistemas de informação**: Minicursos do XVI Simpósio Brasileiro de Sistemas de Informação, Porto Alegre: Sociedade Brasileira da Computação, 2019, pp. 81-113. Disponível em: <https://sol.sbc.org.br/livros/index.php/sbc/catalog/download/33/133/319-1?inline=1>. Acesso em: 10 nov. 2021.

SANTOS, S. M. **Teoria e aplicação de *support vector machines* à aprendizagem e reconhecimento de objetos baseado na aparência.** 2002. Dissertação (Mestrado em Informática) Universidade Federal da Paraíba, Campina Grande, 2002. Disponível em: <http://docs.computacao.ufcg.edu.br/posgraduacao/dissertacoes/2002/Dissertacao_EulandaMirandadosSantos.pdf>. Acesso em: 14 nov. 2021.

SCHMITT, J. **Pré-processamento para a mineração de dados:** uso da análise de componentes principais com escalonamento ótimo. 2005. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Santa Catarina, Florianópolis, 2005. Disponível em: <https://core.ac.uk/download/pdf/30382413.pdf>. Acesso em: 24 out. 2021.

SCHÖLKOPF, B. **Support vector learning.** Tese de Doutorado, Universidade de Berlin, 1997. Disponível em: <https://svms.io/learnability/Scho97.pdf>. Acesso em: 14 nov. 2021.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. **Scikit-learn: Machine Learning in Python**, JMLR 12, pp. 2825-2830, 2011.

SEO, K. W.; SEO, J.; KIM, K.; JI LIM, S.; CHUNG, J. **Prediction of biogas production rate from dry anaerobic digestion of food waste:** Process-based approach vs. recurrent neural network black-box model. Bioresource Technology, v. 341, p. 125829,

2021. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0960852421011706>. Acesso em 13 mar. 2022.

SHAHSAVAR, M. M.; AKRAMI, M.; GHEIBI, M.; KAVIANPOUR, A. M.; FATHOLLAHI-FARD; BEHZADIAN, K. **Constructing a smart framework for supplying the biogas energy in green buildings using an integration of response surface methodology, artificial intelligence and petri net modelling**. ENERGY CONVERSION AND MANAGEMENT, v. 248, 2021.

SILVEIRA, M. R. R**. Potencial de produção de biogás da codigestão anaeróbia termofílica de resíduos de frutas e verduras e lodo de esgoto primário.** 2017. Dissertação (Mestrado em Engenharia Química) - Universidade Federal de Santa Catarina, Florianópolis. Disponível em: <https://repositorio.ufsc.br/handle/123456789/185569>. Acesso em: 09 out. 2021.

SUTTON, R.S.; BARTO, A. G. **Reinforcement learning**: an introduction. London: Bradford Book, 2015.

VIDAL, F. Caderno Setorial ETENE. **Produção e mercado de açúcar.** 2020. Disponível em: <https://www.bnb.gov.br/documents/80223/7600112/2020_CDS_122.pdf/3209edd4-1c0c-ec1d-1519-c32349fa26c0>. Acesso em: 10 out. 2021.

WANG, Y.; HUNTINGTON, T.; SCOWN, C. D. **Tree-Based Automated Machine Learning to Predict Biogas Production for Anaerobic Co-digestion of Organic Waste.** ACS SUSTAINABLE CHEMISTRY & ENGINEERING, v. 9, n. 38, p. 12990–13000, 2021.

WANG, Z.; PENG, X.; XIA, A.; SHAH, A. A.; HUANG, Y.; ZHU, X.; ZHU, X.; LIAO, Q. The role of machine learning to boost the bioenergy and biofuels conversion. **Elsevier**, v. 343, jan. 2022. Disponível em: <https://doi.org/10.1016/j.biortech.2021.126099>. Acesso em: 17 out. 2021.

XIAO, J.; LIU, C.; JU, B.; XU, H.; SUM, D.; DANG, Y. **Estimation of in-situ biogas upgrading in microbial electrolysis cells via direct electron transfer:** Two-stage machine learning modeling based on a NARX-BP hybrid neural network. **BIORESOURCE TECHNOLOGY**, v. 330, 2021.

YANG, Y.; ZHENG, S.; AI, Z.; JAFARI, M. M. M. **On the Prediction of Biogas Production from Vegetables, Fruits, and Food Wastes by ANFIS- and LSSVM-Based Models**. BIOMED RESEARCH INTERNATIONAL, v. 2021, 2021.