UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA QUÍMICA

MATEUS VITOR VIEIRA

# UTILIZAÇÃO DE APRENDIZADO SUPERVISIONADO NA PREDIÇÃO DA DEMANDA DE ENERGIA NO PROCESSO DE PRODUÇÃO DE CUMENO

SÃO CARLOS -SP
2021

MATEUS VITOR VIEIRA

# UTILIZAÇÃO DE APRENDIZADO SUPERVISIONADO NA PREDIÇÃO DA DEMANDA DE ENERGIA NO PROCESSO DE PRODUÇÃO DE CUMENO

Trabalho de conclusão de curso apresentado ao Departamento de Engenharia Química da Universidade Federal de São Carlos, para obtenção do título de bacharel em Engenharia Química

Orientadora: Prof. Dra. Alice Medeiros de Lima

São Carlos-SP

2021

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciência Exatas e de Tecnologia
Departamento de Engenharia Química

**Folha de aprovação**

Assinatura dos membros da comissão examinadora que avaliou e aprovou o Trabalho de Conclusão de Curso do candidato Mateus Vitor Vieira, realizada em 29/06/2021:

_____

Prof. Dra. Alice Medeiros de Lima
UFSCar

_____

Prof. Dr. José Maria Corrêa Bueno
UFSCar

_____

Ícaro Augusto Maccari Zelioli
UNICAMP

## BANCA EXAMINADORA

Trabalho de Conclusão de Curso apresentado de forma virtual no dia 29 de junho de 2021 perante à seguinte banca examinadora:

Orientadora: Profa. Dra. Alice Medeiros de Lima, Departamento de Engenharia Química, Universidade Federal de São Carlos (DEQ/UFSCar).

Convidado: Ícaro Augusto Maccari Zelioli, Faculdade de Engenharia Química, Universidade Estadual de Campinas (FEQ/UNICAMP).

Professor da disciplina: Prof. Dr. José Maria Corrêa Bueno, Departamento de Engenharia Química, Universidade Federal de São Carlos (DEQ/UFSCar).

**DEDICATÓRIA**

Dedico este trabalho à minha mãe, Soraya. Sem ela, a caminhada até a conclusão deste trabalho seria muito mais difícil.

## AGRADECIMENTO

Agradeço a Deus, pela minha vida, por toda força, paciência e coragem para superar todos os obstáculos para a conclusão deste trabalho.

À minha mãe e à minha irmã, pelo amor incondicional que me faz querer superar todas as coisas.

Aos meus amigos, em especial Talita, Iara, Nádia e Ana Maria, por estarem comigo nos momentos bons e nos momentos nem tão bons assim. Por toda troca de experiência que me permitiram crescer não apenas como pessoa, mas também como um formando.

À minha orientadora, Alice, por todas as reuniões, conselhos e direções dadas que levaram à conclusão deste trabalho.

Aos meus familiares, em especial aos meus avós, Isabel e Antônio e aos meus tios, Maria José e José Carlos, que me apoiaram durante toda a graduação.

Aos meus mentores, Ícaro e Rafael, que me ajudaram a enxergar os caminhos e soluções para a conclusão deste trabalho.

**RESUMO**

Ciência de dados (*Data Science*) é uma área do conhecimento muito requisitada atualmente, sobretudo nas empresas digitais e de tecnologia. Apesar do aumento de investimento nessa área também por parte de indústrias químicas, alguns elementos da ciência de dados, como inteligência artificial e machine learning são utilizados nessa indústria desde os anos 80, sendo as redes neurais artificiais (ANN) os tipos de algoritmos mais usados. Estes algoritmos, apesar na capacidade de resolver problemas não lineares complexos, apresentam algumas desvantagens como baixa compreensibilidade e alta dependência de *hardwares.* Sendo assim, este trabalho trata-se da criação e da validação de três modelos de aprendizado supervisionado (regressão linear múltipla, árvore de decisão e floresta aleatória) de alta compreensibilidade e baixa dependência de *hardwares* para a previsão da energia total e da energia específica (energia por massa de cumeno produzido) de uma planta produtora de cumeno através alquilação do benzeno com propeno. As variáveis de entrada dos modelos foram vazão molar de propileno, vazão molar de benzeno, temperatura e pressão do processo. Os dados foram gerados por meio de simulações no software Aspen Plus® e posteriormente tratados e modelados utilizando o Pyhton e MS Excel*.* Os modelos conseguiram prever melhor a energia total da planta do que a energia específica. Em relação à energia total da planta, mesmo com toda a complexidade de um processo químico, os modelos conseguiram prever a energia requerida com erros de aproximadamente ±10%.

Palavras-chaves: Data Science. Machine Learning. Cumeno. Energia. Aprendizado supervisionado. Regressão linear. Árvore de decisão. Floresta aleatória.

**ABSTRACT**

Data Science is a field of knowledge very requested nowadays, specially for digital and technology industries. Despite the increasing investment in this area, also by chemical industries, some elements of data science, such as Artificial Intelligence and machine learning have been used by these industries since 1980s, with artifical neuralnetworks (ANN) being the most used type of algorithms. ANN are great at solving complex non-linear problems but this type of algorithms have some disadvantages such as poor comprehensibility and high hardware dependence. Therefore, this study is about the use of three supervised learning models (linear regression, decision tree and random forest) of high comprehensibility and low hardware dependence for the prediction of total energy and specific energy (energy per mass of cumene) of a cumene producing plant by benzene with propene alkylation. The features are molar flow of propene, molar flow of benzene, temperature and pressure. Data were generated by simulations in Aspen Plus® software, treated and modeled using Python and MS Excel. The models were better at prediction of total energy than specific energy. Even with all the complexity of a chemical process, the models were able to predict the energy required with approximately ±10% errors.

Keywords: Data Science. Machine learning. Cumene. Energy. Supervised learning. Linear regression. Decision Tree. Random Forest.

## LISTA DE ILUSTRAÇÕES

**Figura 1** - Exemplos de variância e viés nos modelos. ..... 9
**Figura 2** - Exemplo de árvore de decisão. ..... 12
**Figura 3** - Fluxograma simplificado do processo de produção de cumeno. ..... 16
**Figura 4** - Fluxograma do processo de produção de cumeno. ..... 17
**Figura 5** - Distribuição da energia específica na seção de separação. ..... 25
**Figura 6** - Distribuição do ln(energia específica) na seção de separação ..... 26
**Figura 7** - Energia específica em relação às variáveis de entrada. ..... 27
**Figura 8** - Distribuição dos resíduos absolutos e percentuais de regressão linear, árvore de decisão e floresta aleatória. ..... 29
**Figura 9** - Distribuição dos resíduos absolutos e percentuais de ln(energia específica). ..... 30
**Figura 10** - Distribuição da energia específica na seção de reação ..... 31
**Figura 11** - Distribuição do ln(energia específica) na seção de reação. ..... 31
**Figura 12** - Distribuição do resíduo absoluto e em porcentagem dos modelos de regressão linear, árvore de decisão e floresta aleatória para a energia específica na seção de reação. ..... 33
**Figura 13** - Distribuição do resíduo absoluto e em porcentagem dos modelos de regressão linear, árvore de decisão e floresta aleatória de decisão para o ln(energia específica) na seção de reação. ..... 34
**Figura 14** - Distribuição da energia específica no processo completo. ..... 35
**Figura 15** - Distribuição do ln (energia específica) no processo. ..... 35
**Figura 16** - Distribuição dos resíduos absolutos e em porcentagem dos modelos de regressão linear, árvore de decisão e floresta aleatória para a energia específica no processo todo. ..... 38
**Figura 17** - Distribuição dos resíduos absolutos e em porcentagem de regressão linear para o ln (energia específica) no processo todo. ..... 39
**Figura 18** - Distribuição da energia na seção de separação. ..... 42
**Figura 19** - Distribuição dos resíduos absolutos e em porcentagem dos modelos de regressão linear, árvore de decisão e floresta aleatória múltipla na seção de separação ..... 44
**Figura 20** - Distribuição da energia total da seção de reação ..... 45

**Figura 21** - Distribuição dos resíduos absolutos e em porcentagem dos modelos de regressão linear, árvore de decisão e floresta aleatória para a energia na seção de reação. ..... 46
**Figura 22** - Distribuição da energia no processo completo. ..... 47
**Figura 23** - Distribuição dos resíduos absolutos e em porcentagem dos modelos de regressão linear, árvore de decisão e floresta aleatória para a energia no processo completo. ..... 49

## LISTA DE TABELAS

**Tabela 1** - Dados cinéticos das reações na formação do cumeno..... 15
**Tabela 2** - Variáveis de entrada do processo..... 17
**Tabela 3** - Configuração do reator. ..... 18
**Tabela 4 -** Configurações para os trocadores de calor ..... 20
**Tabela 5 -** Configurações das colunas de destilação C-1, C-2 e C-3. ..... 21
**Tabela 6** - Coeficientes de determinação de treino e de teste dos modelos preditivos para a energia específica na seção de separação..... 24
**Tabela 7** - Coeficientes de validação cruzada dos modelos preditivos para a energia específica na seção de separação. ..... 25
**Tabela 8** - Coeficientes de determinação de treino e de teste dos modelos preditivos para o ln(energia específica) na seção de separação..... 26
**Tabela 9** - coeficientes de validação cruzada dos modelos preditivos para o ln(energia específica) na seção de separação. ..... 26
**Tabela 10** - Coeficientes de determinação de treino e de teste dos modelos preditivos para a energia específica na seção de reação. ..... 31
**Tabela 11** - Coeficientes de validação cruzada dos modelos preditivos para a energia específica na seção de reação..... 32
**Tabela 12** - Coeficientes de determinação dos modelos preditivos para o ln(energia específica) na seção de reação..... 32
**Tabela 13** - Coeficientes de determinação dos modelos preditivos para o ln(energia específica) na seção de reação..... 32
**Tabela 14** - Coeficientes de determinação de treino e de teste dos modelos preditivos para a Energia específica na planta toda..... 36
**Tabela 15** - Coeficientes de determinação de treino e de teste dos modelos preditivos para o ln (energia específica) na planta toda. ..... 36
**Tabela 16** - Coeficientes de validação cruzada dos modelos preditivos para a energia específica no processo..... 36
**Tabela 17** - Coeficientes de validação cruzada dos modelos preditivos para o ln(energia específica) no processo..... 37
**Tabela 18** - Erro percentual médio (em módulo) dos modelos utilizados para previsão da energia específica nas seções..... 40
**Tabela 19** - Erro percentual médio (em módulo) dos modelos utilizados para a previsão de ln(energia específica) nas seções..... 41

**Tabela 20** - Coeficientes de correlação de treino e de teste dos modelos preditivos para a energia na seção de separação. ..... 42
**Tabela 21** - Coeficientes de validação cruzada dos modelos preditivos para a energia na seção de separação. ..... 43
**Tabela 22** - Erro percentual (em módulo) dos modelos utilizados para a previsão da energia total na seção de separação. ..... 43
**Tabela 23** - Coeficientes de correlação de treino e de teste dos modelos preditivos para a energia na seção de reação..... 45
**Tabela 24** - Coeficientes de validação cruzada dos modelos preditivos para a energia na seção de reação. ..... 45
**Tabela 25** - Erro percentual (em módulo) dos modelos utilizados para a previsão da energia total na seção de reação. ..... 47
**Tabela 26** - Coeficientes de determinação de treino e de teste dos modelos preditivos para a energia no processo todo. ..... 47
**Tabela 27** - Coeficientes de validação cruzada dos modelos preditivos para a energia no processo. ..... 48
**Tabela 28** - Erro percentual (em módulo) dos modelos utilizados para a previsão da energia total no processo. ..... 48

**SUMÁRIO**

1. INTRODUÇÃO ..... 1

2. REVISÃO BIBLIOGRÁFICA ..... 3

2.1 Machine Learning ..... 3

2.2 Machine learning na indústria química ..... 4

2.3 Aprendizado supervisionado ..... 5

2.4 Etapas do aprendizado supervisionado ..... 5

2.5 Produção de cumeno ..... 13

3 METODOLOGIA ..... 17

3.1 Configurações do processo ..... 18

3.2 Tratamento dos dados ..... 21

3.3 Modelos de machine learning ..... 21

3.4 Validação do modelo ..... 22

4 RESULTADOS E DISCUSSÕES ..... 24

4.1 Energia específica ..... 24

4.2 Energia ..... 41

*5* CONCLUSÃO ..... 52

REFERÊNCIAS BIBLIOGRÁFICAS ..... 53

# 1. INTRODUÇÃO

Em 2012, A *Harvard Business Review*, uma publicação da *Harvard Business Publishing,* que tem como objetivo levantar discussões sobre melhores práticas de gestão de negócios, definiu a ciência de dados como “A profissão mais sexy do século 21”, devido a previsão da popularidade dessa profissão nos anos seguintes (DAVENPORT & PATIL, 2012). Segundo a McKinsey, as estimativas eram de que os estados unidos teriam de 140.000 a 190.000 cientistas de dados até 2019 (MCKINSEY & COMPANY, n.d.). Em meio a um mundo cada vez mais orientado a dados, o cientista de dados (profissional da área de ciência de dados) é alguém capaz de analisar, interpretar, modelar e até transformar dados incialmente desestruturados em arte visual (GHIRALDINI, 2020)

Ciência de dados é um termo amplo e inclui diferentes disciplinas (SASIKUMAR, 2021). Essa área utiliza estatística na análise de dados, além de envolver também a captura e armazenamento de informações, e utilização de algoritmos avançados de inteligência artificial, *machine learning* e *deep learning* (*CIÊNCIA DE DADOS*, 2019)

Diversas indústrias têm explorado ciência de dados para encontrar *insights* úteis e entender os dados a fim de aumentar a eficiência e o desempenho de processos. Algumas companhias incluindo BP, ExxonMobil e GE têm investido quantias significativas na inteligência artificial para aplicação em engenharia química em melhorias de segurança e confiabilidade de processos, identificação de novas conexões para fluxos de trabalho, na aceleração de projetos e possibilidade do rápido uso de dados de processos (YAN et al., 2020).

Recentemente, a AspenTech, empresa fornecedora de softwares e serviços para indústrias de processos como têm lançado soluções para análise de dados e inteligência artificial industriais. como *Aspen Industrial Workbench*$^{TM}$, Aspen IoT Analytics Suite$^{TM}$, *Aspen Data Science Studio*$^{TM}$ e *Aspen Enterprise Insights*$^{TM}$ (ARTIFICIAL INTELLIGENCE OF THINGS, 2021).Na engenharia química, Inteligência Artificial e *Machine Learning* têm sido utilizados na modelagem e simulação de processos químicos, no gerenciamento de energias renováveis, em controles de processos e na predição de propriedades físico-químicas de fluidos e catalisadores.

Apesar de ciência de dados já ser utilizada na engenharia química desde 1980, o tipo de algoritmo mais utilizado em problemas de predição são as redes neurais artificiais (YAN et al., 2020). Esse tipo de algoritmo possui algumas vantagens como a habilidade de trabalhar com informações incompletas, armazenamento da informação em toda a rede, tolerância a falhas, existência de memória distribuída, capacidade de processamento paralelo (MIJWEL, 2018). No entanto, são algoritmos complexos, altamente dependentes de hardwares, baixa compreensibilidade e baixa escalabilidade. Esses critérios são importantes quando as informações obtidas pelo modelo precisam ser acessadas e interpretadas por não especialistas em *machine learning*, como por exemplo um executivo de negócios ou um operador (WUJEK et al., 2016).

Diante disso, o objetivo desse trabalho é exemplificar o uso de modelos mais simples de *machine learningutilizar* como o modelo linear, árvore de decisão e floresta aleatória de decisão na predição da demanda de energia de uma planta produtora de um dos cinco compostos mais produzidos mundialmente: o cumeno (CHUDINOVA et al., 2015). Em 2016, o mercado global de cumeno valia US$ 18,8 bilhões, e sua demanda tem crescido desde então em várias aplicações como laminados, compósitos e plásticos. Em 2021, mesmo durante a pandemia de coronavírus, projeta-se uma taxa de crescimento anual composta para o cumeno de 4% para o período de 2021 a 2026. O mercado global de cumeno é beneficiado pelo aumento na demanda por fenol e acetona, compostos utilizados para a produção de resinas e solventes utilizados na fabricação de adesivos, laminados elétricos, revestimentos, pavimentação e na produção de copolímeros e homopolímeros (CUMENE, 2020; CUMENE, 2017).

# 2. REVISÃO BIBLIOGRÁFICA

Nesse capítulo é apresentado uma revisão sobre os principais tópicos referentes à aplicação da ciência de dados em processos químicos. Nesse trabalho, adotou-se especificamente o processo de produção de cumeno. Este capítulo é dividido em duas partes principais: a primeira parte destaca o conceito de *machine learning,* focando no aprendizado supervisionado e nos modelos mais utilizados. A segunda parte é focada na produção de cumeno, descrevendo um pouco dos processos utilizados, das reações envolvidas e do contexto comercial do cumeno.

## 2.1 Machine Learning

Utiliza-se *machine learning* (ML) para extrair *insights* sobre os dados, utilizando técnicas de estatística, matemática e computação para utilizar dados gerados no passado para prever comportamentos futuro (MÜLLER & GUIDO, 2017). De acordo com Ayodele (2010), machine learning pode ser definido como o processo de construção de sistemas computacionais que automaticamente melhoram com a experiência e implementam um processo de aprendizagem. ML é um subcampo da Inteligênica Artificial (IA), que é uma forma mais ampla de pensar sobre os avanços da inteligência de computadores (REESE, 2017).

*Machine Learning* pode ser categorizada em três abordagens: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado reforçado (SHIN et al., 2019). No aprendizado supervisionado é necessário pares de entrada-saída e o modelo tenta descobrir as relações entre as variáveis de entrada (chamadas de variáveis independentes) e as variáveis de saída (chamadas de variáveis dependentes) existentes (MÜLLER & GUIDO, 2017). Como este trabalho usará métodos de aprendizado supervisionado na simulação de uma planta de cumeno, discorreremos sobre eles em termos de principais métodos, aplicações e principais etapas de um aprendizado supervisionado.

Diferentemente do aprendizado supervisionado, no aprendizado não-supervisionado não são necessários pares de entrada e saída, e o próprio algoritmo encontra relações entre os dados. Isso faz a abordagem não-supervisionada ser atrativa em aplicações onde a obtenção dos dados é barata, mas a identificação desses dados é cara ou inexistente. Devido a esta característica, este tipo de

aprendizagem é utilizado na detecção de falhas e anomalias ao longo do tempo (WITTEK, 2014)

No aprendizado reforçado, o modelo não é movido pelos dados e sim pelo problema a ser resolvido. O modelo não recebe uma ordem do que fazer, como em outros tipos de aprendizagens, mas sim de maximizar o sinal de recompensa. As duas principais características desse modelo são tentativas e erros e atraso na recompensa. Em processos químicos, esse tipo de aprendizado pode ser utilizado no controle adaptativo de refinarias de petróleo (SUTTON & BARTO, 1998).

## 2.2 Machine learning na indústria química

Técnicas de *machine learning* têm sido utilizadas para diversas aplicações na indústria química como na modelagem, simulação, controle, estimativa e predição de processos químicos (YAN et al., 2020). Tunckaya & Koklukaya (2015), por exemplo, utilizou algoritmos de Redes Neurais Artificiais, Regressão Linear Múltipla e Modelos Autorregressivos de média na predição e modelagem do efluente gasoso de uma planta térmica de carvão como formar de reduzir a emissão de gases poluentes. Nourani et al. (2018), por sua vez, empregou técnicas de inteligência artificial na predição da performance de uma planta de tratamento de água em termos de demanda biológica de oxigênio, demanda química de oxigênio e nitrogênio total no efluente.

Métodos de inteligência artificial também foram utilizadas para gerenciamento de energia. Zahraee et al. (2016) usou métodos de inteligência artificial na otimização de sistemas híbridos de energia renovável.

De acordo com Yan et al. (2020), o uso de IA na engenharia química não é novo. No começo (1980), sistemas especialistas e algoritmos genéticos foram utilizados na indústria química, no desenvolvimento de catalisadores e na predição de propriedades termofísicas de fluidos. Após um tempo, os algoritmos existentes passaram a ser substituídos por redes neurais artificiais, que hoje são os métodos de ML mais utilizados na engenharia química em estimativas.

Como todos os métodos de ML, as redes neurais artificiais têm suas vantagens e desvantagens. Inspiradas em neurônios biológicos, elas são capazes de aprender relação não-lineares e multivariadas complexas (YAN et al., 2020). No entanto, como desvantagens, elas possuem uma grande dependência de *hardwares,* seu

comportamento é inexplicado, uma vez que sua solução é produzida sem uma dica ou um porquê, não possuem estrutura definida, apresentam dificuldades em entender problemas que não envolvem números, possuem duração desconhecida (MIJWEL, 2018).

## 2.3 Aprendizado supervisionado

Aprendizado supervisionado é um dos mais usados e bem sucedidos tipos de *machine learning* (MÜLLER & GUIDO, 2017). Esse tipo de ML tenta descobrir as relações entre as variáveis de entrada (chamadas de variáveis independentes) e as variáveis de saída (chamadas de variáveis dependentes). A relação descoberta entre essas variáveis é representada numa estrutura chamada de modelo (MAINMON & ROKACH, 2005). No aprendizado supervisionado, as variáveis podem ser contínuas, categóricas ou binárias, e as variáveis de saída associadas às variáveis de entrada são conhecidas (KOTSIANTIS, 2007).

## 2.4 Etapas do aprendizado supervisionado

A resolução de um problema de aprendizado supervisionado apresenta as seguintes etapas: definição do problema, identificação e coleta dos dados requeridos, pré-processamento dos dados, separação em dados de treino e dados de teste, seleção do algoritmo, treino do algoritmo/sintonização dos parâmetros, validação do modelo com o grupo de teste (KOTSIANTIS, 2007).

### 2.4.1 Identificação do problema

Os problemas envolvendo aprendizado supervisionado são divididos em duas categorias: problemas de classificação e problemas de regressão (MAINMON & ROKACH, 2005). Na classificação, o objetivo é predizer uma *class label*, uma classe à qual os dados pertencem dentro de uma lista de possibilidades pré-definidas, enquanto na regressão o objetivo é prever um número contínuo, ou um número de ponto flutuante (ou número real, em termos matemáticos) (MÜLLER & GUIDO, 2017).

### 2.4.2 Coleta e identificação dos dados

Após a definição do problema, a primeira etapa é a identificação e coleta dos dados. Esta etapa é chamada de *Feature* Engineering e é dividida em quatro etapas: *brainstorming,* criação, seleção e avaliação.

Na fase de brainstorming, entende-se o domínio do problema e reune-se informações para descobrir quais poderiam ser os atributos (*features*) que melhor representam os dados (BIANCHI, 2020). Aqui, se um especialista não estiver disponível, recomenda-se coletar informações sobre o maior número de *features* (atributos) possível na esperança de conseguir isolar as variáveis mais relevantes (KOTSIANTIS, 2007).

Na fase de criação as *features* são produzidas. Na etapa de seleção pode-se remover ou ainda expandir algumas *features* e na fase de avaliação realiza-se uma estimativa da qualidade do modelo, utilizando as *features* selecionadas (Bianchi, 2020).

A etapa de coleta de dados contém, na maioria dos casos, valores faltantes e ruídos, o que requer uma etapa posterior de pré-processamento dos dados (Kotsiantis, 2007).

### 2.4.3 Pré-processamento

De acordo com Zhang et al. (2003), a etapa de pré-processamento dos dados pode ser mais demorada e apresentar até mais dificuldade do que a etapa de coleta e identificação dos dados. Zhang et. al (2003) também selecionou alguns motivos pelos quais essa etapa é importante:

- Dados reais podem ser incompletos, cheiros de ruídos e inconsistentes o que pode disfarçar padrões úteis;

- Preparação dos dados gera um conjunto de dados menor do que o original, o que pode melhorar significativamente a eficiência da mineração dos dados;

- Preparação dos dados gera dados de qualidade, o que pode ajudar na identificação de padrões de qualidade nos dados.

Mainmon & Rokach ( 2005) citou alguns métodos genéricos de preparação dos dados:

- Estatístico: identificação de *outliers* usando valores como média, desvio padrão e *range* baseados no teorema de Chebyshev e considerando o intervalo de confiança em cada campo;

- Agrupamento: identificação de *outliers* utilizando técnicas de agrupamento baseados na distância Euclidiana (ou outros tipos);

- Padrão de comportamento: identificação de *outliers* que não estão em conformidade com os padrões dos dados existentes;

- Regras de associação: quando uma regra de associação está bem definida, pontos que não se encaixam nessa regra são considerados *outliers.*

### 2.4.4 Separação entre grupos de treino e grupo de teste

No aprendizado supervisionado, o objetivo é construir um modelo utilizando os dados do grupo de treino e então fazer predições acuradas em novos, e ainda não vistos, dados chamados de grupo de teste (MÜLLER & GUIDO, 2017). Usualmente assume-se que o grupo de treino é gerado aleatoriamente e independentemente de acordo com a alguma distribuição probabilística (MAINMON & ROKACH, 2005). Essa separação dos dados em grupos de treino e grupo de teste e chamado de *Resampling* (MOURA, 2016).

Moura (2016), enumerou quatro tipos de *resampling*: *hold-out validation, bootstrap* e *cross validation*. Hold-out é a forma mais simples de separar os dados, nela uma porcentagem dos dados é definida como sendo de treino e de teste. É utilizada quando o número de dados é muito grande e a amostra é considerada como tendo significância para representar a população.

O método Bootstrap é utilizado para estimar a habilidade do modelo em generalizar informações por meio de valores estatísticos como média e desvio padrão (BROWNLEE, 2019). Nele, diferentes amostras de tamanho N de dados não vistos pelo modelo são testadas para medir a habilidade do modelo em prever dados que não estão contidos no grupo de treino (BROWNLEE, 2019; YEN, 2019).

Existem diferentes tipos de cross validation (validação cruzada), porém a técnica mais comum é conhecida com *k-fold cross validation.* Neste método, um número K de amostras é criado, sendo que cada uma delas é deixada de lado enquanto o modelo treina com o restante delas. Os valores comuns de K de amostras são entre 5 e 10 (MOURA, 2016).

Utiliza-se o grupo de treino para a construções de modelos que produzam predições acuradas para esse grupo. Se o grupo de treino e o grupo de teste tiverem uma quantidade suficientes de características em comum, então espera-se que o modelo também seja acurado para o grupo de teste. No entanto, se os dois grupos

forem muito diferentes, o modelo não será capaz de generalizar o conhecimento aprendido com os dados de treino. O mesmo acontece na construção de modelos muito complexos. Quanto mais complexo o modelo, mais ele se ajustará aos dados de treino, no entanto a capacidade de generalização da aprendizagem pode ficar prejudicada (MÜLLER & GUIDO, 2017). Modelos que se ajustam totalmente aos dados de treino, mas tem uma capacidade de generalização pobre é dito apresentar *overfitting* (sobreajuste). Modelos que não se ajustam bem aos dados é dito apresentar *underfitting.*

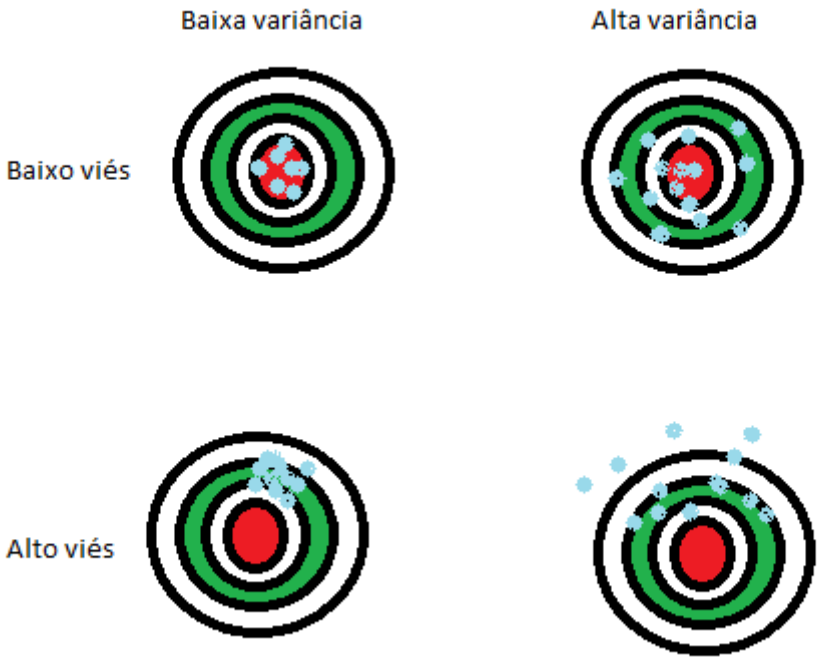
### 2.4.5 Escolha dos modelos de aprendizagem supervisionada

A escolha do algoritmo a ser usado é uma etapa crítica (KOTSIANTIS, 2007). Naturalmente, os modelos que produzem resultados mais acurados são considerados os melhores para serem utilizados, no entanto outros critérios podem ser tão importantes quanto, como por exemplo a complexidade computacional, além da compreensibilidade e escalabilidade (MAINMON & ROKACH, 2005).

A escolha do algoritmo a ser utilizado geralmente leva a um balanço entre duas principais característica de: variância (*variance*) e viés (*bias*). A variância diz respeito à consistência dos resultados obtidos pelo modelo adotado, enquanto o viés diz respeito à acurácia do modelo adotado. O modelo preditivo ideal é aquele que consegue unir baixa variância e baixo viés, fazendo com que os resultados obtidos sejam consistentes e acurados (BONFIM, 2020). A Figura 1 exemplifica alguns casos em que a variância e o viés são observados. Os modelos de alta variância e baixo viés são, em média, acurados, porém inconsistentes. Modelos com alto viés e baixa variância são, consistente, mas em média não são acurados. Modelos com alta variância e alto viés não são acurados nem consistentes.

De acordo com Müller (2017), ao trabalhar com um novo conjunto de dados em geral é uma boa ideia começar com modelos simples e ver o quão longe é possível chegar. Depois de entender mais sobre os dados, pode-se mover para modelos mais complexos. Ohri (2021) listou os cinco algoritmos de aprendizado supervisionado mais utilizados na regressão, sendo eles modelo linear, regressão *ridge*, redes neurais artificiais, regressão *lasso* e árvore de decisão. Esses modelos serão descritos a seguir, e acrescentaremos a floresta aleatória, que é um modelo baseado em árvore de decisão.

**Figura 1** - Exemplos de variância e viés nos modelos.




Adaptado de Bonfim (2020).

MODELO LINEAR

Modelos lineares são uma classe de modelos amplamente usados na prática. Esse tipo de modelo faz a predição das saídas usando uma função linear das variáveis de entrada (MÜLLER & GUIDO, 2017). Para regressão, a fórmula geral de predição usando modelo linear é dada pela equação (1).

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + \ldots + w[p] * x[p] + b \quad (1)$$

Onde x[0] até x[p] denotam as variáveis de entrada de um único dado, w e b são parâmetros do modelo linear que são aprendidos e $\hat{y}$ é a predição do modelo (MÜLLER & GUIDO, 2017). Quando existe apenas uma variável de entrada, uma regressão utilizando o modelo linear é chamada de regressão linear simples. Com duas ou mais variáveis de entrada, o nome é regressão linear múltipla.

Esse método assume linearidade entre as variáveis de entrada e as variáveis de saída. No entanto, nem sempre essa relação é linear. Desse modo, é necessário

a utilização de métodos estatísticos que consigam tornar o modelo mais robusto e menos enviesado (KAYRI et al., 2017). O modelo de regressão linear utilizado nesse trabalho é o *ordinary least squares* (Mínimos quadrados ordinários). Nesse modelo, os parâmetros w e b são calculados minimizando o Erro Quadrado Médio (EQM) entre as predições do modelo e o valor real de saída, isto é, minimizando a soma do quadrado das diferenças entre a saída prevista e a saída real (MÜLLER & GUIDO, 2017).

Como vantagens, modelos lineares são rápidos no treinamento e na predição dos dados, eles escalam bem para grandes conjuntos de dados e trabalham bem com dados escassos. Modelos lineares também apresenta boa compreensibilidade. No entanto, eles apresentam melhor desempenho quando o número de *features* (variáveis de entrada) é grande comparado ao número de amostras, em espaços com poucas dimensões outros modelos podem generalizar melhor que os modelos lineares (MÜLLER & GUIDO, 2017).

## REGRESSÃO LASSO E RIDGE

*Ridge* e *Lasso* são algumas técnicas utilizadas na redução da complexidade do modelo e prevenção de *overfitting* resultado de uma regressão linear simples (BHATTACHARYYA, 2018). Ridge e Lasso também são modelos lineares de regressão (MÜLLER & GUIDO, 2017).

Na regressão *Ridge* os coeficientes w são escolhidos não apenas para se ajustarem bem aos dados de treinamento, mas também para se ajustar a uma limitação a mais (MÜLLER & GUIDO, 2017). A regressão *Ridge* diminui os coeficientes do modelo linear e ajuda a reduzir a complexidade e multicolinearidade. Quanto menor a restrição nas variáveis de entrada, mais o modelo se assemelhará a uma regressão linear (BHATTACHARYYA, 2018).

Na regressão *Lasso*, restrições também são adicionadas ao modelo. A diferença entre essas duas formas de regressão é que na *Lasso* os coeficientes do modelo linear não só são menores, como podem ser exatamente zero. O que significa que algumas variáveis são completamente ignoradas pelo modelo (MÜLLER & GUIDO, 2017)

Por também serem modelos lineares, tanto a regressão *Lasso* quanto a *Ridge* possuem as mesmas forças e fraquezas desse tipo de modelo.

## REDES NEURAIS ARTIFICIAIS

Inspiradas pelas redes neurais biológicas, as redes neurais artificiais (ANN) são sistemas de computadores maciçamente paralelos que consistem em um número extremamente grande de processadores simples com muitas conexões (GUPTA, 2013). Essa família de algoritmos tem sido referenciada como *deep learning*. Algoritmos de *deep learning* são frequentemente feitos sob medida para usos específicos (MÜLLER & GUIDO, 2017).

Uma ANN consiste de uma camada de entrada de neurônios, uma, duas ou três camadas de neurônios escondidos e uma camada final de saída (WANG, 2003). As entradas são multiplicadas por *weights*. Esses weights são então computados por uma função matemática que determina a ativação dos neurônios. Outra função computa as saídas dos neurônios artificiais. Os neurônios dessa rede somam suas entradas. Como cada neurônio de entrada tem apenas uma entrada, suas saídas são a multiplicação da entrada pelo *weights* (GUPTA, 2013).

Mijwel (2018) apresenta algumas vantagens e desvantagens desse tipo método. Como vantagens: habilidade de trabalhar com informações incompletas, armazenamento da informação em toda a rede, tolerância a falhas, existência de memória distribuída, capacidade de processamento paralelo. Entre as desvantagens: uma grande dependência de *hardwares,* seu comportamento é inexplicado, uma vez que sua solução é produzida sem uma dica ou um porquê, não possuem estrutura definida e apresentam dificuldades em entender problemas que não envolvem números.

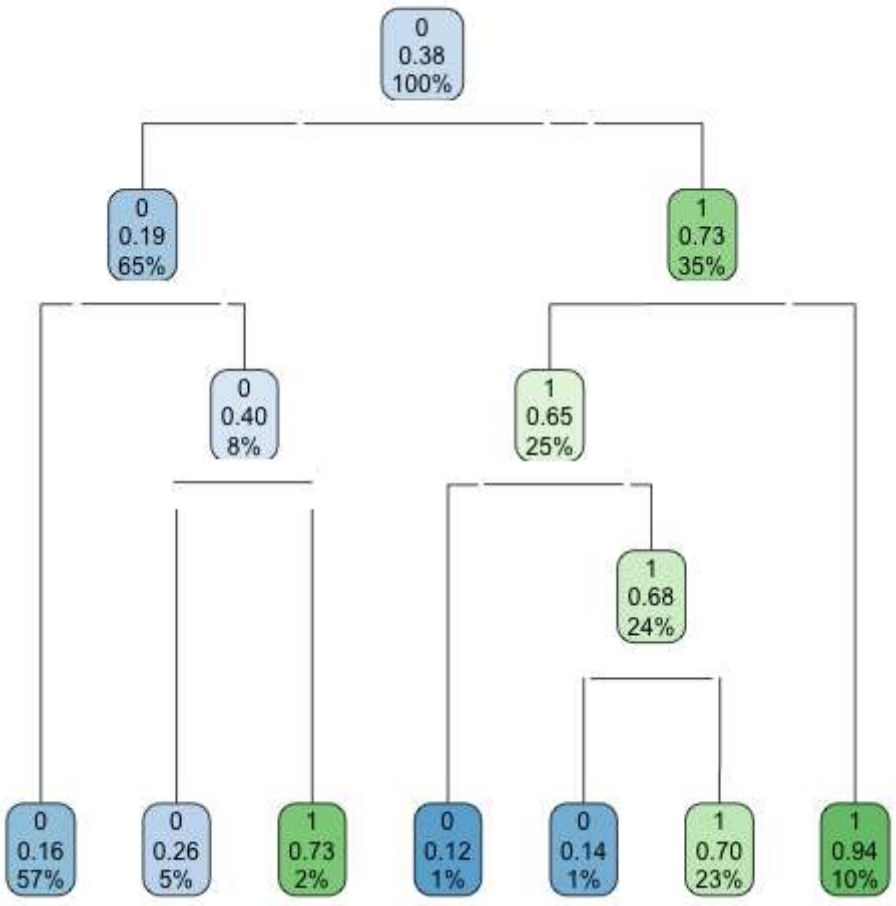
## ÁRVORE DE DECISÕES

Na modelagem por árvore de decisão, uma árvore empírica representa uma segmentação dos dados e é criada pela aplicação de uma série de regras simples (TSO & YAU, 2007). O objetivo é alcançar a resposta certa (número real, quando usada para regressão, ou categoria quando usada para classificação), por uma série de perguntas de se/senão (MÜLLER & GUIDO, 2017).

O algoritmo de árvore de decisão funciona da seguinte forma: um primeiro teste é feito para tentar representar o modelo. Os dados são avaliados e recebem como resposta “verdadeiro” ou “falso”. Assim uma segunda rodada de testes é realizada

para grupo verdadeiro e para o grupo falso. Isso continua até que os dados sejam totalmente representados pelo conjunto de testes propostos (MÜLLER & GUIDO, 2017). A Figura 2 apresenta um exemplo de árvore de decisão.

**Figura 2** - Exemplo de árvore de decisão.




Adaptado de Prates (2018).

Devido à característica de aumentar o número de testes até que todos os dados sejam contemplados, os modelos de árvore de decisão tendem a ter um viés muito alto, ou seja, os valores de coeficiente de determinação acabam sendo 1 ou muito próximo disso. Um viés muito alto (*overfitting*) significa que o modelo se ajustou bem ao grupo de treino, mas pode não generalizar tão bem esse conhecimento para o grupo de testes. Além do *overfitting* esse modelo apresenta outras desvantagens como o fato de não funcionarem bem com variáveis contínuas, uma pequena diferença nos dados tem a causar grandes diferenças na estrutura da árvore, levando a instabilidade, os cálculos envolvidos podem se tornar complexos comparados a

outros algoritmos e levar mais tempo par treinar o modelo e fica relativamente caro à medida que o tempo e a complexidade aumentam. Entre suas vantagens estão a alta compreensibilidade do modelo para não especialistas, além de o modelo não sofrer variações para o escalonamento dos dados (MÜLLER & GUIDO, 2017).

FLORESTA ALEATÓRIA

A Floresta aleatória é um modelo baseado na árvore de decisão que surge como uma forma de evitar o overfitting por esse modelo. Nesse modelo, várias árvores de decisões diferentes e aleatórias são construídas. O coeficiente de determinação é então calculado pela média dessas árvores (MÜLLER & GUIDO, 2017). Existem diferentes variantes de florestas aleatórias as quais são caracterizadas pelo modo como cada árvore individual é construída, pelo procedimento usado na geração dos dados que darão origem às diferentes árvores individuais, o modo como as predições das árvores individuais são agregadas para gerar uma predição final (BOULESTEIX et al., 2012). Na engenharia química, existem alguns estudos desse algoritmo na predição de solubilidade aquosa (PALMER et al., 2007), na predição biossorção de tintas utilizando resíduos agrícolas (SOARES et al., 2020) e na análise de modelos microcinéticos (PARTOPOUR et al., 2018).

Em relação às vantagens e desvantagens desse algoritmo, as vantagens da floresta aleatória são a redução do *overfitting* nas árvores de decisão o que torna o modelo mais acurado, é flexível para problemas de regressão e de classificação, trabalha bem com valores contínuos e categóricos e não requer normalização dos dados. No entanto, esse método requer muito recurso computacional para a construção de muitas árvores, requer muito tempo para treinamento para classificação, não é tão compreensível quanto outros métodos(FOREST, 2020). Os modelos baseados em árvore também não possuem uma capacidade de extrapolação (MÜLLER & GUIDO, 2017).

## 2.5 Produção de cumeno

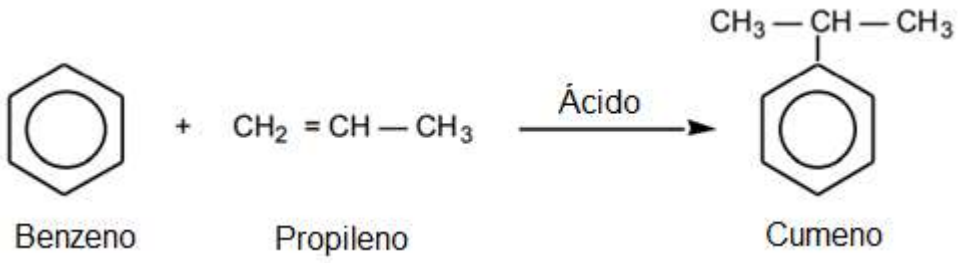
O cumeno é usado amplamente como matéria prima nas plantas de fenol onde é convertido para fenol e acetona. O principal uso final do cumeno é para a produção de vários produtos como bisfenol A, policarbonato e resinas epóxi, nylon 6 entre outros (JUNQUEIRA et al., 2018). Em termos de volume de produção, o cumeno é um dos

cinco compostos mais produzidos em larga-escala, dividindo espaço com o etileno, propileno, benzeno e etilbenzeno (CHUDINOVA et al., 2015). Em 2018, a Ásia liderava o mercado de cumeno participando de 47% da receita gerada por esse composto (JUNQUEIRA et al., 2018).

A única rota comercial de produção de cumeno utilizada atualmente é através da alquilação do benzeno com propeno em catalisador ácido (CUMENE, 1999). Os catalisadores utilizados são: ácido fosfórico sólido, cloreto de alumínio anidro e zeólitas (AL-KINANY et al., 2001). A reação é exotérmica (–23.4 kcal/mol a 250° C).

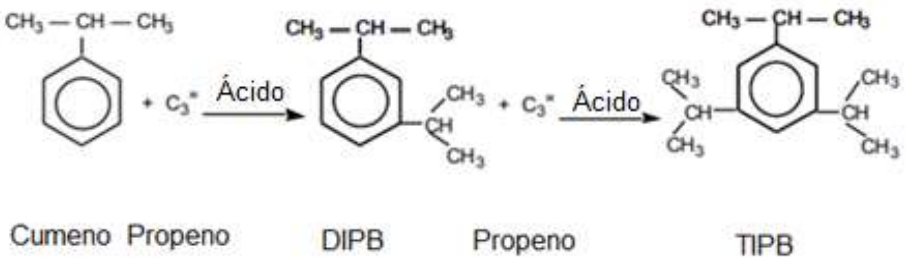
No processo de alquilação ocorrem cinco reações.

1) O benzeno reage com o propileno em catalisador ácido para formar o cumeno, de acordo com a equação (2).




(2)

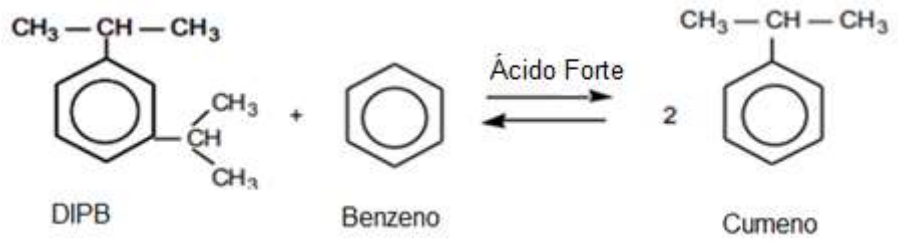
2) Uma pequena porção do cumeno reage com o propileno para formar diisopropilbenzeno (DIPB) que por sua vez reage com o propeno para formar triisopropilbenzeno (TIPB) de acordo com a equação (3).



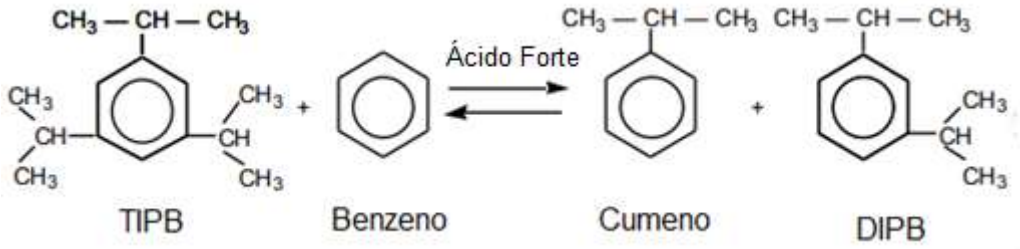

(3)

3) Na presença de ácidos fortes, DIPB e TIPB são convertidos novamente a cumeno. Essa reação é chamada de transalquilação e envolve a transferência

de grupos isopropil entre as diferentes espécies de acordo com as equações (4) e (5) apresentam as reações de transalquilação.




(4)




(5)

A Tabela 1 apresenta os dados da cinética das reações envolvidas.

**Tabela 1** - Dados cinéticos das reações na formação do cumeno.

| | Reação 1 | Reação 2 e 3 | Reação 4 e 5 |
|---|---|---|---|
| Equações | (2) | (3) | (4) e (5) |
| K | $2{,}8 \times 10^{7}$ | $2{,}32 \times 10^{9}$ | $6{,}52 \times 10^{-3}$ |
| E (kJ/kmol) | 104.174 | 146.742 | 27.240 |
| Taxa (kmol $s^{-1}$ $m^{-3}$) | $kc_p c_{b^*}$ | $kc_p c_{c^*}$ | - |

Fonte: Elaborado pelo autor com base no Aspen plus, s.d.

Notas:

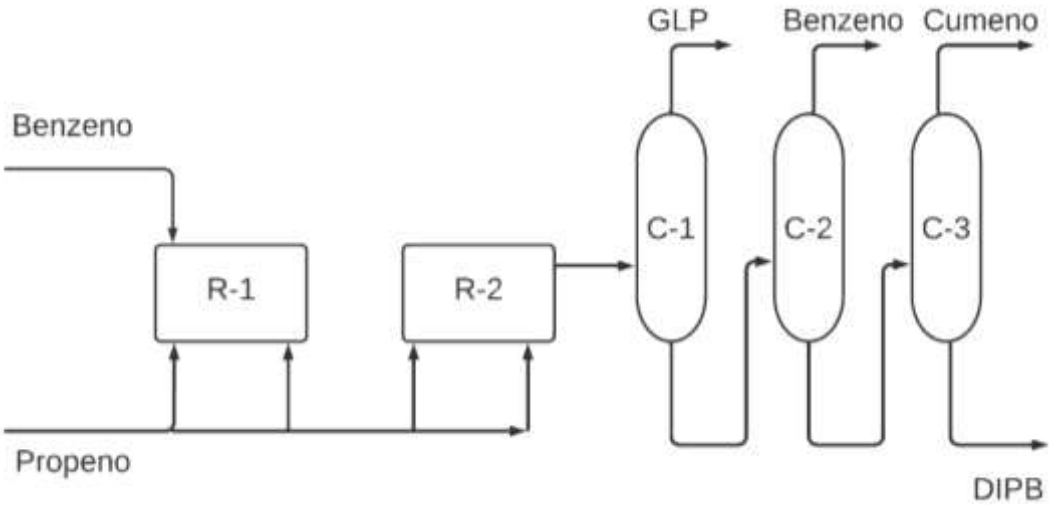
* onde $c_p$ é a concentração molar de propileno, $C_b$ é a concentração molar de benzeno e $C_c$ é a concentração molar de cumeno.

O processo de produção de cumeno é ilustrado na Figura 3. Ele é composto por duas etapas principais: alquilação e recuperação do cumeno. Na etapa de alquilação, benzeno é colocado em contado com o propeno em dois reatores em série, R-1 e R-2 contendo dois leitos de catalisador cada. A alimentação de aromáticos é carregada no topo do primeiro reator a 125 °C, enquanto o propileno fresco é dividido em quatro leitos catalíticos. A alquilação acontece a 33 bar. O principal subproduto da

reação é o diisopropilbenzeno (DIPB), que é em sequência separado e convertido a cumeno por transalquilação.

Na recuperação do cumeno, o efluente do reator passa por três colunas de destilação. A primeira coluna (C-1) tem o produto de topo separado em uma fase propano, formando o gás liquefeito de petróleo (GLP), que pode ser utilizado como combustível. O produto de fundo da primeira coluna é enviado para a segunda coluna (C-2), onde praticamente todo o benzeno não reagido é recuperado e utilizado novamente da reação de alquilação. A terceira e última coluna (C-3) separa o cumeno (produto de interesse) do DIPB. O DIPB pode ser utilizado novamente na reação de alquilação.

**Figura 3** - Fluxograma simplificado do processo de produção de cumeno.



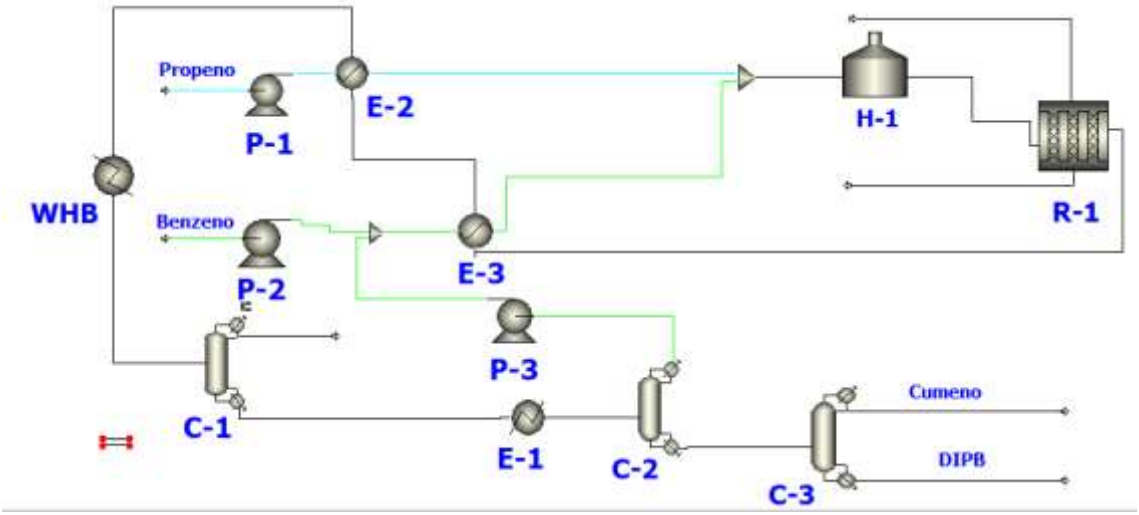

Elaborado pelo autor com base em Production of Cumene, 1999.

Nandi et al. (2004) realizou a modelagem e otimização do processo de produção de cumeno, utilizando redes neurais, regressão por vetores de suporte e algoritmos genéticos. No presente estudo, o foco será a predição de energia total e energia específica do processo de produção de cumeno, cujo fluxograma é disponibilizado na biblioteca de modelos de processos do *sotware Aspen Plus*®.

## 3 METODOLOGIA

Os dados utilizados nesse trabalho foram gerados por meio de simulações no software de simulação de processos *Aspen Plus*® V10 do processo de produção do cumeno, parte da biblioteca de processos do software. A Figura 4 apresenta o fluxograma utilizado nas simulações.

**Figura 4** - Fluxograma do processo de produção de cumeno.




Elaborado pelo autor com base em Aspen Plus®, n.d.

Identificação dos equipamentos: C-1, C-2, C-3: colunas de destilação; E-1, E-2, E-3: trocadores de calor; H-1: aquecedor; P-1, P-2, P-3: bombas; R-1: reator; WHB: caldeira recuperadora de calor

Em cada simulação variou-se os valores da alimentação fresca de propeno e benzeno, bem como a temperatura e pressão do processo. A Tabela 2 contém os valores utilizados nas simulações.

**Tabela 2** - Variáveis de entrada do processo.

| Benzeno (kmol/h) | Propeno (kmol/h) | Temperatura (ºC) | Pressão (bar) |
|---|---|---|---|
| 75,00 | 75,00 | 300 | 15 |
| 88,75 | 88,75 | 325 | 20 |
| 102,50 | 102,50 | 350 | 25 |
| 116,25 | 116,25 | 375 | 30 |

| | | | |
|---|---|---|---|
| 130 | 130 | 400 | - |

Elaborado pelo autor com base em Aspen Plus®, n.d.

As simulações foram feitas da seguinte forma: (1) fixou-se um valor de temperatura no aquecedor H-1 e no reator R-1 dentre os valores contidos na Tabela 2, (2) fixou-se o mesmo valor de pressão em todas as bombas P-1, P-2 e P-3 e (3) realizou-se uma análise de sensibilidade onde os valores de alimentação de benzeno e propeno variavam de acordo com a Tabela 2 e obteve-se os dados referentes à quantidade de energia requerida ou liberada em cada um dos equipamentos, bem como as vazões de cumeno no efluente do reator e no efluente da coluna de destilação C-3.

## 3.1 Configurações do processo

O processo utilizado é composto por 14 equipamentos, sendo eles: 3 bombas, dois pontos de mistura, um forno, um reator, 4 trocadores de calor e 3 colunas de destilação.

### 3.1.1 Reator (R-1)

O reator é do tipo multitubular com fluido térmico concorrente. A Tabela 3 apresenta as configurações do reator.

**Tabela 3** - Configuração do reator.

| **Especificação** | **Valores** |
|---|---|
| Especificação do parâmetro de transferência de calor | 65 Watt/sqm-K |
| Número de tubos | 2500 |
| Comprimento do tubo | 6 metros |
| Diâmetro do tubo | 0,0763 metro |
| Fase da corrente de processo | Vapor |
| Fase da corrente de fluido térmico | Vapor |
| Reações | Equações (2) a (5) |
| Vazio no leito | 0,5 |
| Densidade do catalisador | 2000 Kg/m³ |

Elaborado pelo autor com base em Aspen Plus®, n.d.

Para manter a temperatura do reator constante, variou-se os valores de vazão do fluido de resfriamento utilizando a ferramenta *Design especs* do software Aspen Plus® V10. Essa ferramenta funciona como um controle feedback, em que varia-se o valor de uma variável para deixar outra variável constante. A vazão do fluido de resfriamento (água) variou entre 0,5 e 3 toneladas/hora.

### 3.1.2 Aquecedor (H-1)

Para o aquecedor, especificou-se a temperatura de saída de acordo com os valores de temperatura contidos na Tabela 2 e considerou-se zero a queda de pressão.

### 3.1.3 Bombas (P-1, P-2 e P-3)

Para as bombas foram especificadas a pressão de descarga de acordo com os valores de pressão descritos na Tabela 2, a eficiência da bomba (0,75) e a eficiência do motor (0,98). A utilidade utilizada é a eletricidade.

### 3.1.4 Trocadores de calor (E-1, E-2, E-3, WHB)

O processo em questão possui quatro trocadores de calor: E-1, E-2 e E-3 e uma caldeira de recuperação de calor (WHB). O trocador de calor E-1 é responsável por resfriar O trocador E-1 é responsável por resfriar a corrente de entrada da coluna de benzeno. Para ele foram especificados os valores de temperatura (90 ºC) e de pressão (2,5 kg/cm²).

O trocador E-2 é responsável por pré-aquecer a corrente de propeno antes de entrar no forno. Nele, o fluido quente é a corrente de produto do reator depois de trocar calor com as correntes de benzeno fresco e reciclado no trocador de calor E-3. Diferentemente do trocador de calor E-1, aqui foi especificado a temperatura de aproximação entre o fluido frio e o fluido quente em 15 ºC. Os coeficientes globais de transferência de calor foram mantidos em 850 W/m²-K.

O trocador de calor E-3 é responsável pela primeira etapa de integração energética dessa planta. Esse equipamento é responsável por utilizar a corrente de produto do reator como fluido quente para pré-aquecer as correntes de benzeno reciclado e benzeno fresco depois de misturadas. Os coeficientes globais de transferência de calor são os mesmos utilizados no trocador E-2 (850 W/m²-K).

O WHB é utilizado para gerar vapor de baixa pressão ao resfriar o efluente do reator antes da entrada na primeira coluna de destilação. As especificações são temperatura (150 °C) e pressão (20,3943 kg/cm²). As fases envolvidas são vapor e líquido. A Tabela 4 contém a configuração dos trocadores de calor.

**Tabela 4 -** Configurações para os trocadores de calor.

| **Especificação** | **E-1** | **E-2** | **E-3** | **WHB** |
|---|---|---|---|---|
| Tipo de bloco | *Heater* | *HeatX* | *HeatX* | *Heater* |
| Fidelidade do modelo | - | *Shortcut* | *Shortcut* | - |
| Direção da corrente | - | Contracorrente | Contracorrente | - |
| Modo de cálculo | - | *Design* | *Design* | - |
| Especificação | Temperatura e pressão | Temperatura de aproximação | Calor trocado | Temperatura e Pressão |
| Valor da especificação | 90 °C<br>2,5 bar | 15 °C | 4 Gcal/hr | 150 °C<br>20 bar |
| Temperatura mínima de aproximação | - | 15 °C | 10 °C | - |

Elaborado pelo autor com base em Aspen Plus®, n.d.

### 3.1.5 Colunas de destilação

O fluxograma de produção de cumeno contém três colunas de destilação. A coluna de destilação C-1 tem como função retirar o propano remanescente no efluente do reator. Logo após, coluna de destilação C-2 tem como objetivo retirar o benzeno não reagido para ser reutilizado no processo. Por último temos a coluna C-3 cuja função é a obtenção do cumeno puro no topo e o DIPB no fundo, que pode ser reutilizado para produção de cumeno. A Tabela 5 contém as configurações das três colunas citadas.

**Tabela 5 -** Configurações das colunas de destilação C-1, C-2 e C-3.

| **Especificações** | **C-1** | **C-2** | **C-3** |
|---|---|---|---|
| Tipo de cálculo | Equilíbrio | Equilíbrio | Equilíbrio |
| Número de estágios | 16 | 20 | 30 |
| Condensador | Parcial-Vapor | Total | Total |
| Refervedor | *Kettle* | *Kettle* | *Kettle* |
| Fases | Vapor-Líquido | Vapor-Líquido | Vapor-Líquido |
| Convergência | Padrão | Padrão | Padrão |
| *Reflux ratio* | 4,13 | 0,66 | 0,90 |
| *Boilup ratio* | 0,38 | 0,85 | 0,99 |
| Estágio de alimentação | 5 | 6 | 15 |
| Correntes de produtos | 1 e 16 | 1 e 20 | 1 e 30 |
| Pressão no primeiro estágio | 12 bar | 1,75 bar | 0,2 bar |
| Pressão no segundo estágio | 12,1 bar | 1,85 bar | 0,21 bar |
| Queda de pressão no resto da coluna | 0 | 0,2 bar | 0,007 bar |

Elaborado pelo autor com base em Aspen Plus®, n.d.

## 3.2 Tratamento dos dados

Após serem gerados no software Aspen Plus® V10, os dados foram armazenados em planilha no Microsoft Excel e tratados no Python utilizando as bibliotecas pandas, numpy, sklearn, matplotlib para criação de gráficos e dos modelos de *Machine Learning.* Foram gerados no total 510 dados. Todos os gráficos e modelos preditivos foram construídos utilizando o *Google Colaboratory* (*COLABORATORY*, n.d.).

## 3.3 Modelos de machine learning

Como dito anteriormente, três modelos de aprendizado supervisionado foram utilizados para prever o comportamento das variáveis de interesse energia específica e energia total da planta, sendo eles: Regressão Linear Múltipla, Árvore de Decisão (regressão) e Floresta Aleatória de Decisão (regressão). Para realizar um estudo mais

aprofundado dividiu-se a análise de cada uma dessas duas variáveis em três partes: (i) análise da seção de reação, (ii) análise da seção de separação e (iii) análise do processo completo.

A seção de reação compreende os blocos (P-1, P-2, E-2, E-3, H-1 e R-1), enquanto a seção de separação compreende os blocos (WHB, E-1, C-1, C-2, C-3 e P-3). A vazão de cumeno utilizada como base para o cálculo da energia específica na seção de reação foi o cumeno efluente do reator, enquanto a vazão de cumeno utilizada como base para o cálculo da energia específica na seção de separação é o cumeno efluente da coluna C-3.

Os dados foram separados em dados de treino e dados de teste. Os dados de treino, como o próprio nome diz, são usados para treinar o modelo, ou seja, para calcular os parâmetros do modelo. O grupo de teste tem como objetivo medir a capacidade do modelo em generalizar a informação. Ou seja, o quão bem o modelo criado com os dados de treino pode ser utilizado para prever dados nunca vistos pelo algoritmo. Após essa etapa, os modelos foram validados.

## 3.4 Validação do modelo

A etapa de validação do modelo envolve estudar a acurácia do modelo tanto para o ajuste dos dados de treino, quanto para a generalização da aprendizagem para os dados de teste. Para isso, observou-se os valores do coeficiente de determinação ($R^2$) do conjunto treino e do conjunto teste para todos os três modelos. Os valores de coeficiente de correlação variam de 0 a 1. Quanto mais próximo de 1, mais bem ajustado o modelo está em relação aos dados. Nesses casos é preciso entender se o modelo não está enviesado.

Como trata-se de um conjunto de dados relativamente pequeno, foi empregada a técnica de validação cruzada. A validação cruzada é uma técnica para dividir os dados em grupos de treino e de testes de forma mais eficiente. No modelo de divisão discutido anteriormente, os dados foram separados somente em duas categorias: grupo de treino (aproximadamente 75% dos dados) e grupo de teste (aproximadamente 25% dos dados). Na validação cruzada, os dados são divididos em K grupos com aproximadamente a mesma quantidade de dados. Desses K grupos, um grupo é escolhido para ser o grupo teste, enquanto os outros K-1 grupos são escolhidos para treinar o modelo. Depois, um segundo grupo é escolhido como grupo

de teste e os outros K-1 grupos restantes (incluindo o primeiro grupo) são escolhidos como grupo de treinamento do modelo. O processo se repete até que todos os K grupos sejam escolhidos como grupo de teste (MÜLLER, 2017).

Neste método, ao invés de se obter um único valor de coeficiente de correlação, obtém-se K valores de coeficiente dos quais se obterá a média.

O método de validação cruzada permite um uso mais eficiente dos dados. Ao escolher, por exemplo, 5 grupos, temos que em cada uma das 5 validações 80% dos dados serão utilizados para o treinamento do modelo e 20% para o teste. Ao escolher 10 grupos, esses números vão para 90% e 10%. Outra vantagem da utilização da validação cruzada é conseguir enxergar a sensibilidade do seu modelo por meio dos K diferentes valores de coeficientes de correlação que serão encontrados (MÜLLER, 2017). Nesse trabalho, utilizou-se o valor de K igual a 5 para os três modelos utilizados.

Outra forma de validar o modelo é por meio do estudo dos resíduos (ou erros). Os resíduos são a diferença entre o valor real da variável de saída e o valor previsto utilizando os modelos ajustados. Idealmente, espera-se que os resíduos tenham as seguintes características de média zero e homocedasticidade, ou seja, variância constante.

Nesse trabalho, realizou-se uma análise exploratória do comportamento dos resíduos utilizando histogramas e gráficos de dispersão do resíduo em relação às variáveis de interesse.

# 4 RESULTADOS E DISCUSSÕES

Essa seção apresentará as análises feitas, bem como as discussões acerca dos resultados discutidos.

## 4.1 Energia específica

Aqui serão apresentadas as análises da energia específica que serão divididas em três partes: (i) energia específica na seção de separação, (ii) energia específica na seção de reação, e (iii) energia específica total.

### 4.1.1 Análise da seção de separação

Nesta seção foram avaliados os modelos utilizados na predição da energia específica na seção de separação. A Tabela 6 apresenta os coeficientes de determinação de treino e de teste dos modelos preditivos para a energia específica na seção de separação. A Tabela 7 apresenta os coeficientes de determinação encontrados na validação cruzada.

**Tabela 6** - Coeficientes de determinação de treino e de teste dos modelos preditivos para a energia específica na seção de separação.

| Coeficientes de determinação | Regressão Linear | Árvore de decisão | Floresta aleatória |
|---|---|---|---|
| Treino | 0,54 | 1,00 | 0,99 |
| Teste | 0,52 | 0,91 | 0,96 |

Elaborado pelo autor.

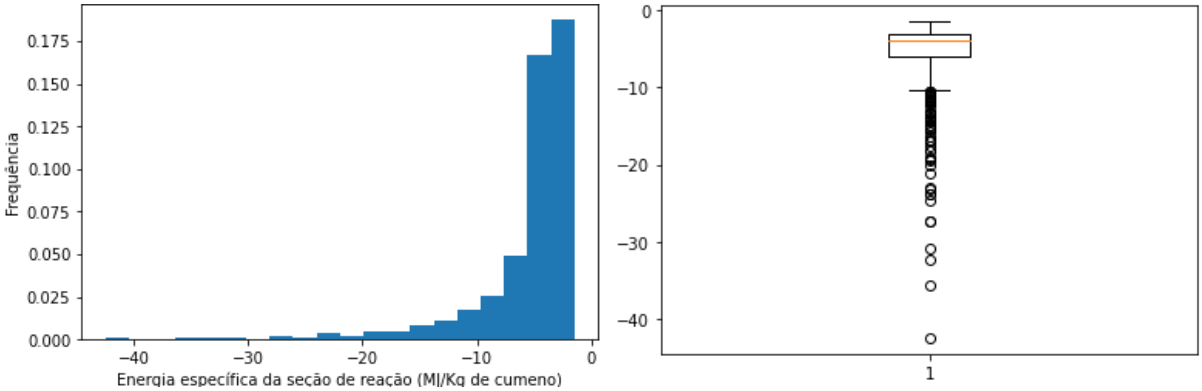
**Tabela 7** - Coeficientes de validação cruzada dos modelos preditivos para a energia específica na seção de separação.

| Coeficiente de determinação | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| 1 | 0,53 | 0,86 | 0,97 |
| 2 | 0,52 | 0,89 | 0,96 |
| 3 | 0,54 | 0,88 | 0,95 |
| 4 | 0,29 | 0,95 | 0,97 |
| 5 | 0,53 | 0,91 | 0,98 |
| **Médio** | **0,48** | **0,90** | **0,97** |

Elaborado pelo autor.

Realizou-se também um estudo da energia específica na seção de separação. A Figura 5 apresenta o comportamento da energia específica nesta seção.

**Figura 5** - Distribuição da energia específica na seção de separação.



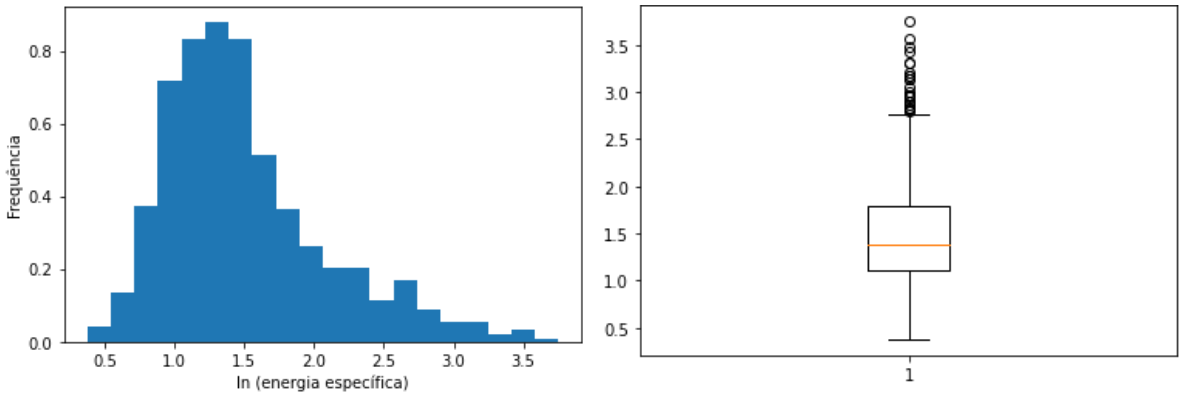

Elaborado pelo autor.

Na Figura 5 é possível ver que a distribuição da energia específica no processo é assimétrica. A assimetria é uma característica ruim para a construção de modelos de regressão linear, uma vez que o modelo precisa se ajustar para compensar valores que não aparecem tão frequentemente. Nesse caso, algumas ações podem ser tomadas: transformação da variável de interesse, aplicando-se um logaritmo, por exemplo, ou dividindo o modelo em múltiplas partes que sejam mais simétricas.

Outra característica da energia específica na seção de separação (Figura 5) é que seus valores são negativos, indicando que nessa seção há mais energia sendo retirada (principalmente nos condensadores) do que fornecida (nos refervedores).

Dessa forma, utilizou-se o módulo dos valores de energia específica para o cálculo da transformação logarítmica. A Figura 6 apresenta a distribuição da variável de interesse depois da aplicação do logaritmo natural. A Tabela 8 apresenta os coeficientes de determinação de treino e teste para a transformação logarítmica da energia específica, enquanto a Tabela 9 apresenta os coeficientes de determinação de validação cruzada.

**Figura 6** - Distribuição do ln(energia específica) na seção de separação




Elaborado pelo autor.

**Tabela 8** - Coeficientes de determinação de treino e de teste dos modelos preditivos para o ln(energia específica) na seção de separação.

| Coeficientes de determinação | Regressão Linear | Árvore de decisão | Floresta aleatória |
|---|---|---|---|
| Treino | 0,70 | 1,00 | 0,99 |
| Teste | 0,68 | 0,93 | 0,96 |

Elaborado pelo autor.

**Tabela 9** - coeficientes de validação cruzada dos modelos preditivos para o ln(energia específica) na seção de separação.

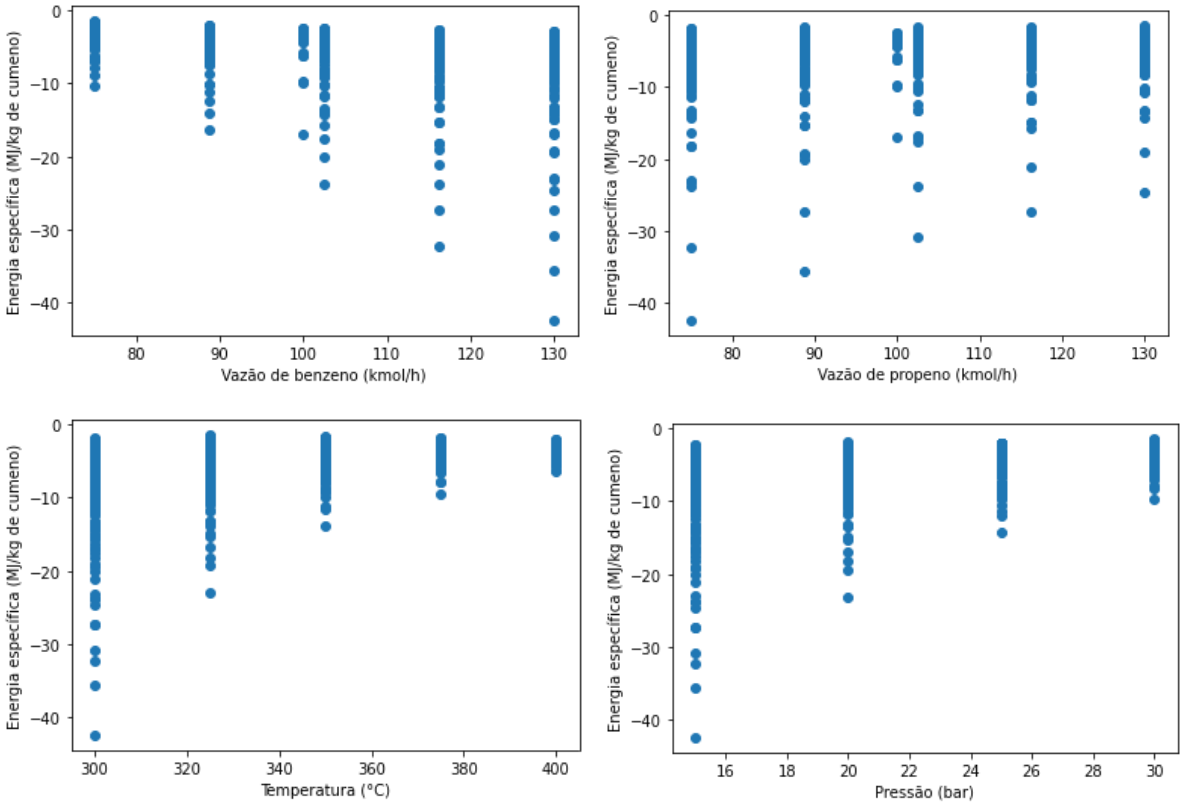
| Coeficientes de determinação | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| 1 | 0,65 | 0,89 | 0,97 |
| 2 | 0,69 | 0,89 | 0,95 |
| 3 | 0,69 | 0,93 | 0,96 |
| 4 | 0,70 | 0,87 | 0,94 |

| | | | |
|---|---|---|---|
| 5 | 0,74 | 0,94 | 0,96 |
| **Médio** | **0,69** | **0,90** | **0,96** |

Elaborado pelo autor.

Pela validação cruzada, é possível perceber que o valor de coeficiente de determinação médio do modelo de regressão linear é 0,48, o que significa que o modelo não representa bem os dados coletados. A Figura 7 apresenta os gráficos da energia específica em relação às variáveis de entrada.

**Figura 7** - Energia específica em relação às variáveis de entrada.




Elaborado pelo autor.

Dentre as quatro variáveis, a pressão e a temperatura são as que mais aparentam ter uma relação não linear com a energia específica. Os gráficos de temperatura e pressão em relação à energia específica tem uma forma semelhante a uma função logarítmica. Isso pode ser explicado ao verificar que a energia específica é calculada utilizando a vazão de cumeno. Enquanto a vazão de cumeno, por sua vez, é calculada utilizando a expressão de Arrhenius, em que a constante de velocidade

(k) da reação varia de forma exponencial com a temperatura. A equação de Arrhenius é descrita em (5)

$$K = A.e^{-\frac{Ea}{RT}} \quad (5)$$

Onde A é o fator pré-exponencial, Ea é a energia de ativação, R é a constante universal dos gases e T a temperatura.

A relação não linear entre a pressão e energia específica também é relacionada à vazão de cumeno. Reações em estado gasoso, como o caso da reação de alquilação descrita nesse trabalho, são mais afetadas pela pressão do que as reações em estados líquidos e sólidos. Smith e Van Ness (2018), propuseram a seguinte equação para o equilíbrio em estado gasoso:

$$\prod_i (y_i \hat{\phi}_i)^{\nu_i} = \left(\frac{P}{P^\circ}\right)^{-\nu} K \quad (6)$$

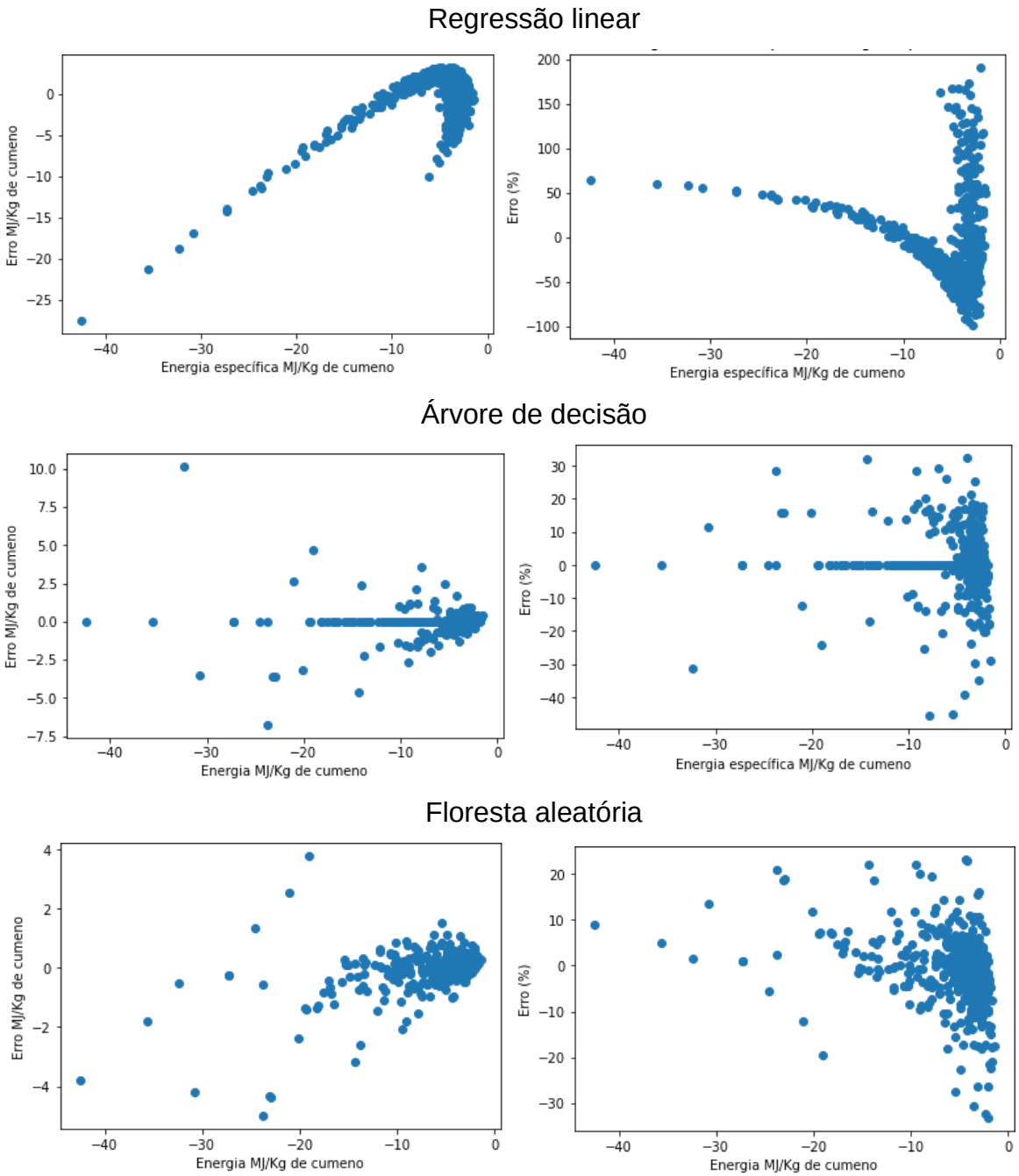
Onde K é a constante de equilíbrio, $y_i$ é a composição molar da espécie i, ϕ é o coeficiente de fugacidade da espécie, v é o somatório dos coeficientes, e vi os coeficientes das espécies. Essa mesma discussão acerca da não linearidade pode ser generalizada para a energia específica nas seções de reação e no processo completo.

Ao comparar as Figuras 8 e 9 contendo a distribuição de resíduos em relação à energia específica e resíduos em relação à transformação logarítmica, percebe-se que após a transformação logarítmica os resíduos ficaram menos concentrados. Após a análise das Tabelas 8 e 9, percebeu-se que após a transformação da variável energia específica os valores de coeficientes de determinação aumentaram, sobretudo para o modelo de regressão linear múltipla.

Apesar da melhoria do comportamento dos resíduos dos modelos após a transformação logarítmica, ainda é possível perceber a existência de pelo menos duas regiões com variâncias diferentes. Uma região cobrindo os valores de energia específica de 0 a -10 KJ/Kg de cumeno e outra região cobrindo os valores menores que -10 KJ/Kg de cumeno. Essa mudança de comportamento de resíduos pode ser explicada pela assimetria dos dados. Uma vez que os modelos foram construídos minimizando o erro quadrado médio, a assimetria faz com o modelo tenha que aumentar o erro em uma seção para compensar na predição da outra seção. O modelo

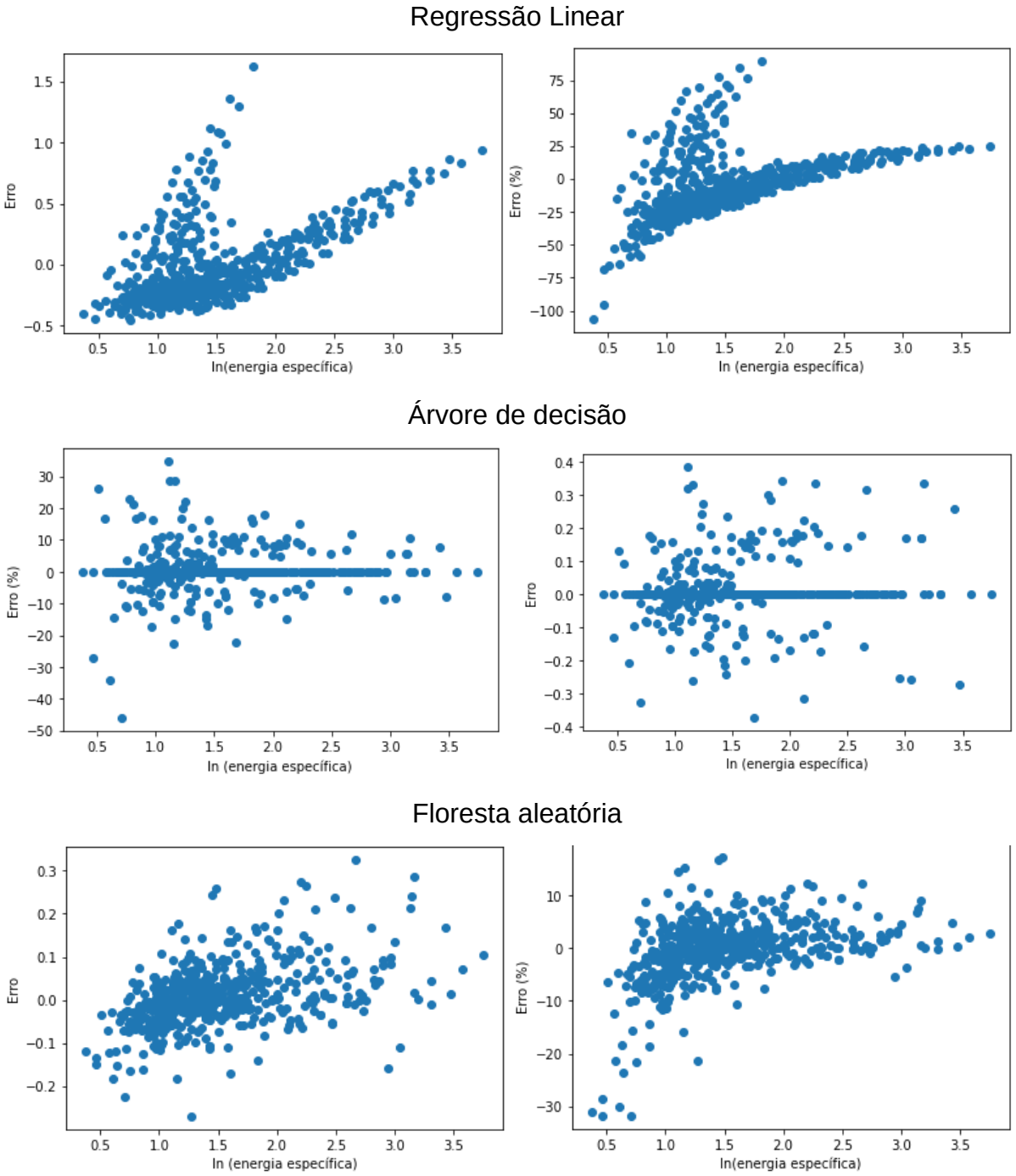
precisa levar em conta nos cálculos o erro de valores que não aparecem tão frequentemente. Uma sugestão para lidar com esse problema seria escolher outros tipos de métrica de desempenho de modelos que não sejam o $R^2$. Ao observar a Figura 7 é possível também notar o aumento da variabilidade dos valores de energia específica abaixo dos valores de -10 KJ/Kg de cumeno.

**Figura 8** - Distribuição dos resíduos absolutos e percentuais de regressão linear, árvore de decisão e floresta aleatória.




Elaborado pelo autor.

**Figura 9** - Distribuição dos resíduos absolutos e percentuais de ln(energia específica).



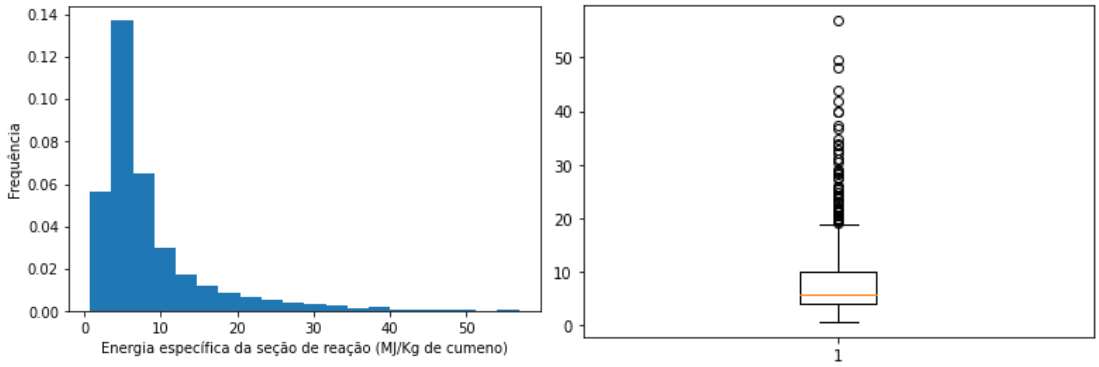

Elaborado pelo autor.

### 4.1.2 Análise da seção de reação

Assim como na análise a seção de separação, realizou-se um estudo da variável de interesse. A energia específica na seção de reação também apresentou assimetria, como mostra a Figura 10, e dessa forma, também se realizou uma transformação logarítmica na variável de interesse. A Figura 11 mostra a distribuição

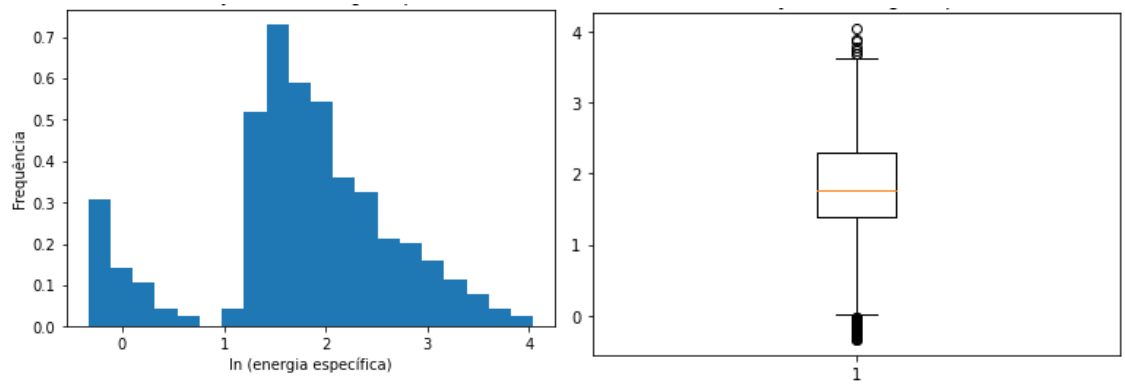
da variável de interesse depois da transformação. As Tabelas 10 a 13 contém os valores de coeficientes de determinação para a variável de interesse e para a variável de interesse transformada, bem como os coeficientes de validação cruzada.

**Figura 10** - Distribuição da energia específica na seção de reação




Elaborado pelo autor.

**Figura 11** - Distribuição do ln(energia específica) na seção de reação.




Elaborado pelo autor.

**Tabela 10** - Coeficientes de determinação de treino e de teste dos modelos preditivos para a energia específica na seção de reação.

| Coeficiente de determinação | Regressão Linear | Árvore de decisão | Floresta aleatória |
|---|---|---|---|
| Treino | 0,70 | 1,00 | 0,99 |
| Teste | 0,67 | 0,94 | 0,97 |

Elaborado pelo autor.

**Tabela 11** - Coeficientes de validação cruzada dos modelos preditivos para a energia específica na seção de reação.

| Coeficiente de determinação | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| 1 | 0,69 | 0,96 | 0,97 |
| 2 | 0,67 | 0,92 | 0,96 |
| 3 | 0,75 | 0,95 | 0,95 |
| 4 | 0,60 | 0,95 | 0,95 |
| 5 | 0,71 | 0,93 | 0,96 |
| Médio | 0,68 | 0,94 | 0,96 |

Elaborado pelo autor.

**Tabela 12** - Coeficientes de determinação dos modelos preditivos para o ln(energia específica) na seção de reação.

| Coeficientes de determinação | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| Treino | 0,86 | 1,00 | 0,99 |
| Teste | 0,87 | 0,93 | 0,97 |

Elaborado pelo autor.

**Tabela 13** - Coeficientes de determinação dos modelos preditivos para o ln(energia específica) na seção de reação.

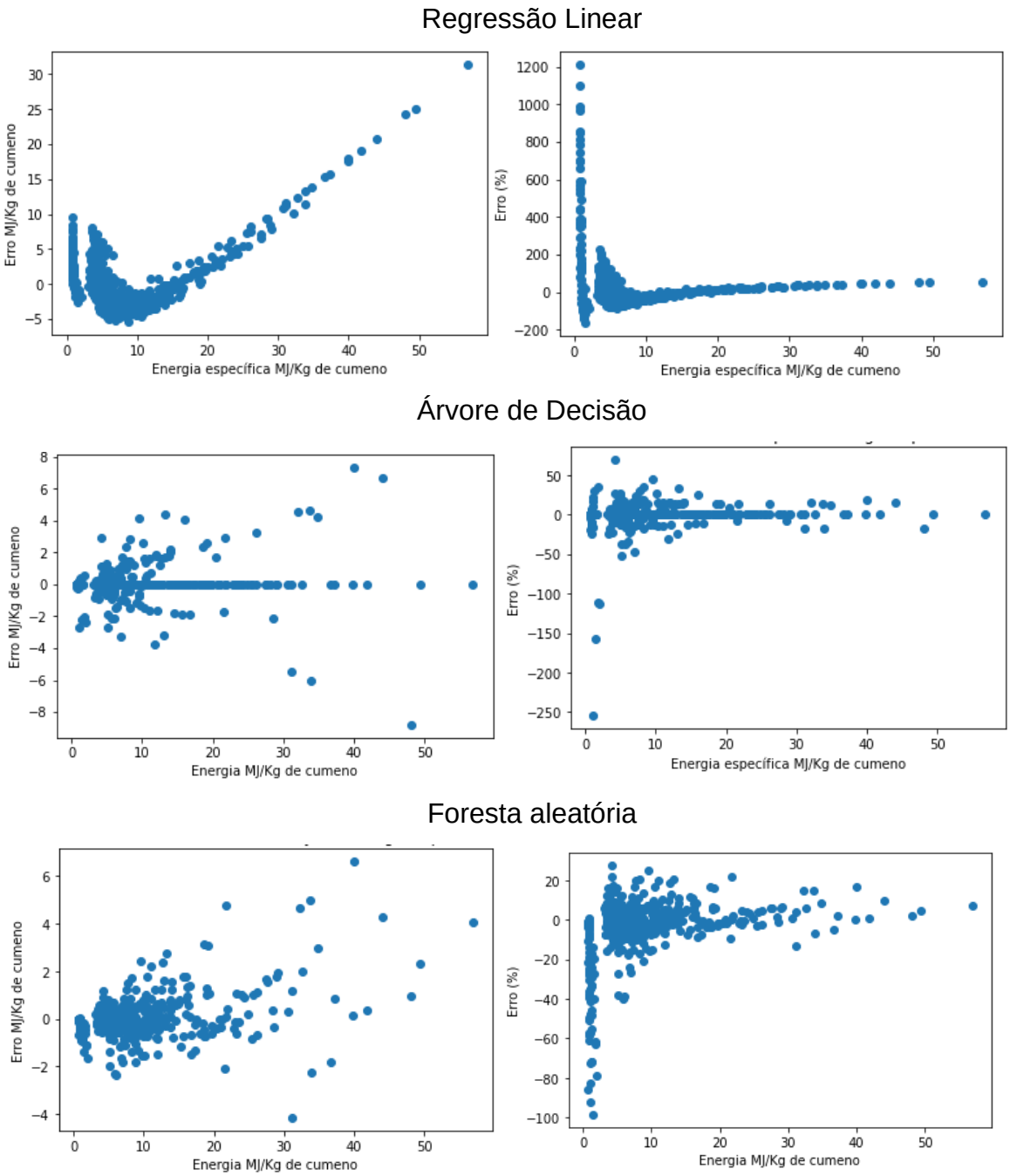
| Coeficientes de determinação | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| 1 | 0,86 | 0,95 | 0,97 |
| 2 | 0,85 | 0,90 | 0,95 |
| 3 | 0,87 | 0,89 | 0,96 |
| 4 | 0,83 | 0,91 | 0,97 |
| 5 | 0,87 | 0,91 | 0,93 |
| Médio | 0,86 | 0,90 | 0,95 |

Elaborado pelo autor.

Houve um aumento nos valores dos coeficientes de determinação após a transformação da variável de interesse. Ao analisar os coeficientes de validação

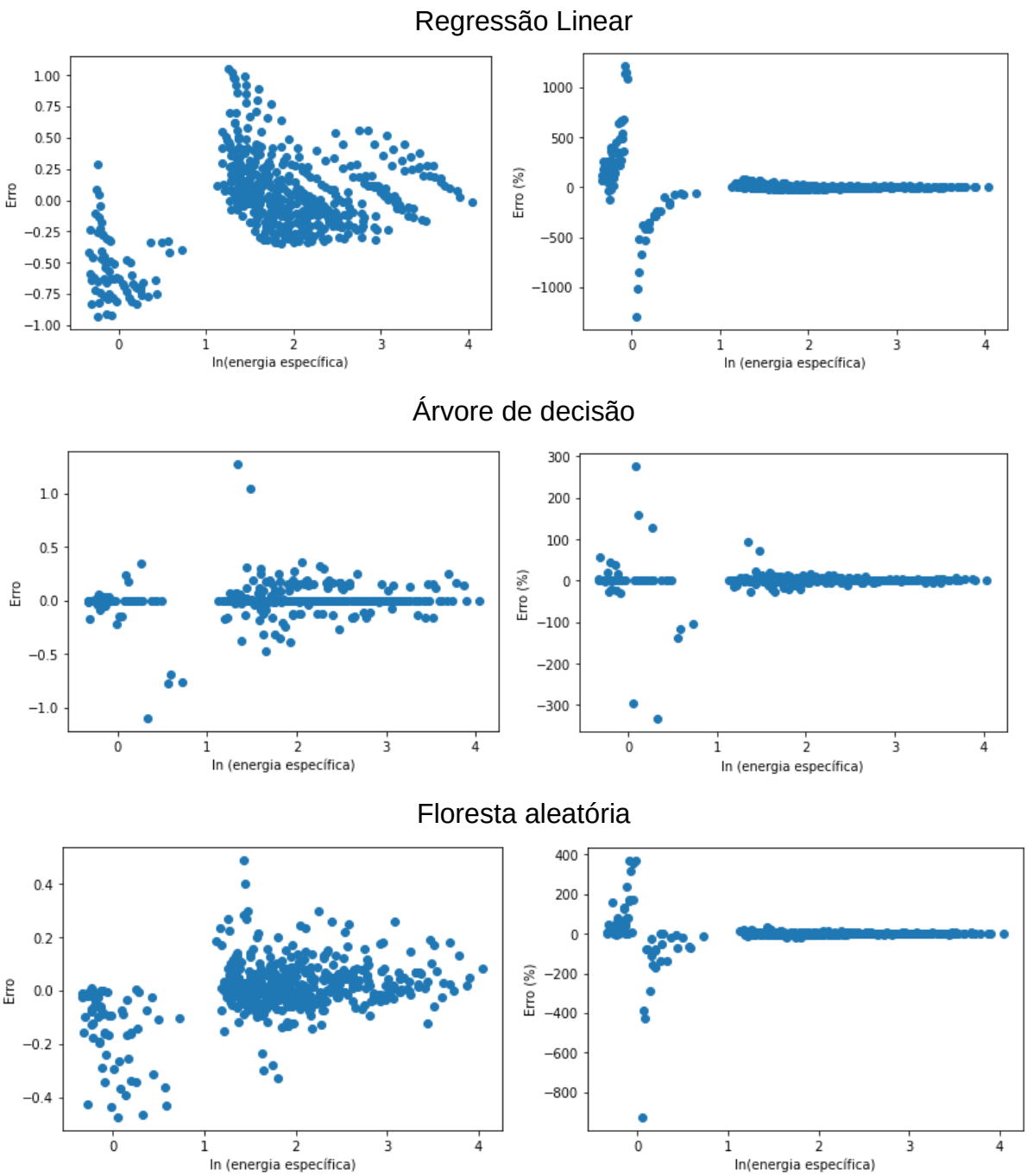
cruzada, antes da transformação a diferença entre o menor e o maior coeficiente de regressão linear era de 0,15, ou seja, 22% em relação ao valor médio (0,68). Depois da transformação, a diferença passou a ser 0,04, representando 4,6% em relação ao valor médio (0,86). As Figuras 12 e 13 apresentam a distribuição dos resíduos dos modelos sem transformação da variável, com transformação e retirando outliers.

**Figura 12** - Distribuição do resíduo absoluto e em porcentagem dos três modelos para a energia específica na seção de reação.




Elaborado pelo autor.

**Figura 13** - Distribuição do resíduo absoluto e em porcentagem dos três modelos para o ln(energia específica) na seção de reação.




Elaborado pelo autor.

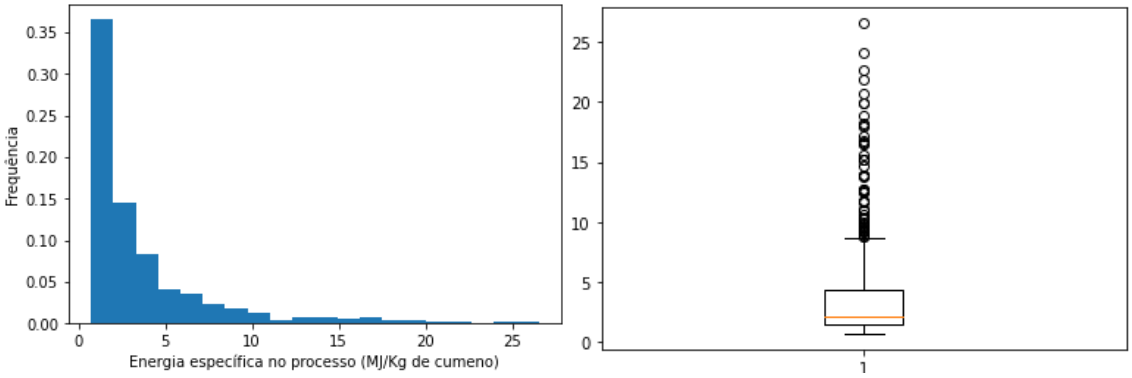
Assim como na seção de separação, foi observado a presença de duas regiões com variabilidade diferentes na energia específica na seção de reação. Uma região

para os valores entre 0 e 20 MJ/Kg de cumeno e outra região para os valores acima de 20 MJ/Kg de cumeno.

### 4.1.3 Análise do processo completo

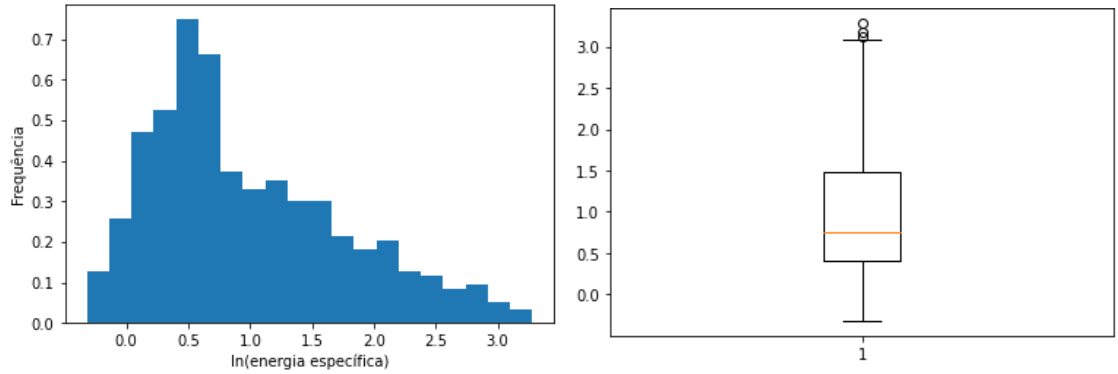
Realizou-se um estudo da variável de interesse, energia específica. A Figura 14 contém a distribuição da energia específica no processo completo. Enquanto a Figura 15 contém a distribuição do ln (energia específica).

**Figura 14** - Distribuição da energia específica no processo completo.




Elaborado pelo autor.

**Figura 15** - Distribuição do ln (energia específica) no processo.




Elaborado pelo autor.

Percebe-se uma melhor simetria dos dados, o que pode fazer com que os coeficientes de determinação do modelo de regressão linear melhores. As Tabelas 14 e 15 apresentam os valores de coeficiente de determinação de treino e teste para a variável de interesse sem e com transformação logarítmica.

**Tabela 14** - Coeficientes de determinação de treino e de teste dos modelos preditivos para a Energia específica na planta toda.

| Coeficientes de determinação | Regressão linear múltipla | Árvore de decisão | Floresta aleatória de decisão |
|---|---|---|---|
| Treino | 0,63 | 1,00 | 0,99 |
| Teste | 0,59 | 0,94 | 0,96 |

Elaborado pelo autor.

**Tabela 15** - Coeficientes de determinação de treino e de teste dos modelos preditivos para o ln (energia específica) na planta toda.

| Coeficientes de determinação | Regressão linear múltipla | Árvore de decisão | Floresta aleatória de decisão |
|---|---|---|---|
| Treino | 0,84 | 1,00 | 1,00 |
| Teste | 0,83 | 0,96 | 0,98 |

Elaborado pelo autor.

Ao comparar os dados da Tabela 14 com os da Tabela 15, percebe-se um aumento dos coeficientes de determinação de treino e teste para todos os três modelos utilizados. O aumento mais expressivo foi para os coeficientes de determinação de regressão linear múltipla. As Tabelas 16 e 17 apresentam, respectivamente, os valores de coeficiente de determinação de validação cruzada para energia específica e para o logaritmo natural da energia específica.

**Tabela 16** - Coeficientes de validação cruzada dos modelos preditivos para a energia específica no processo.

| Coeficientes de determinação | Regressão linear múltipla | Árvore de decisão | Floresta aleatória de decisão |
|---|---|---|---|
| 1 | 0,56 | 0,93 | 0,95 |
| 2 | 0,55 | 0,93 | 0,94 |
| 3 | 0,62 | 0,93 | 0,95 |
| 4 | 0,64 | 0,93 | 0,95 |
| 5 | 0,71 | 0,95 | 0,97 |

| | | | |
|---|---|---|---|
| Médio | 0,62 | 0,93 | 0,95 |

Elaborado pelo autor.

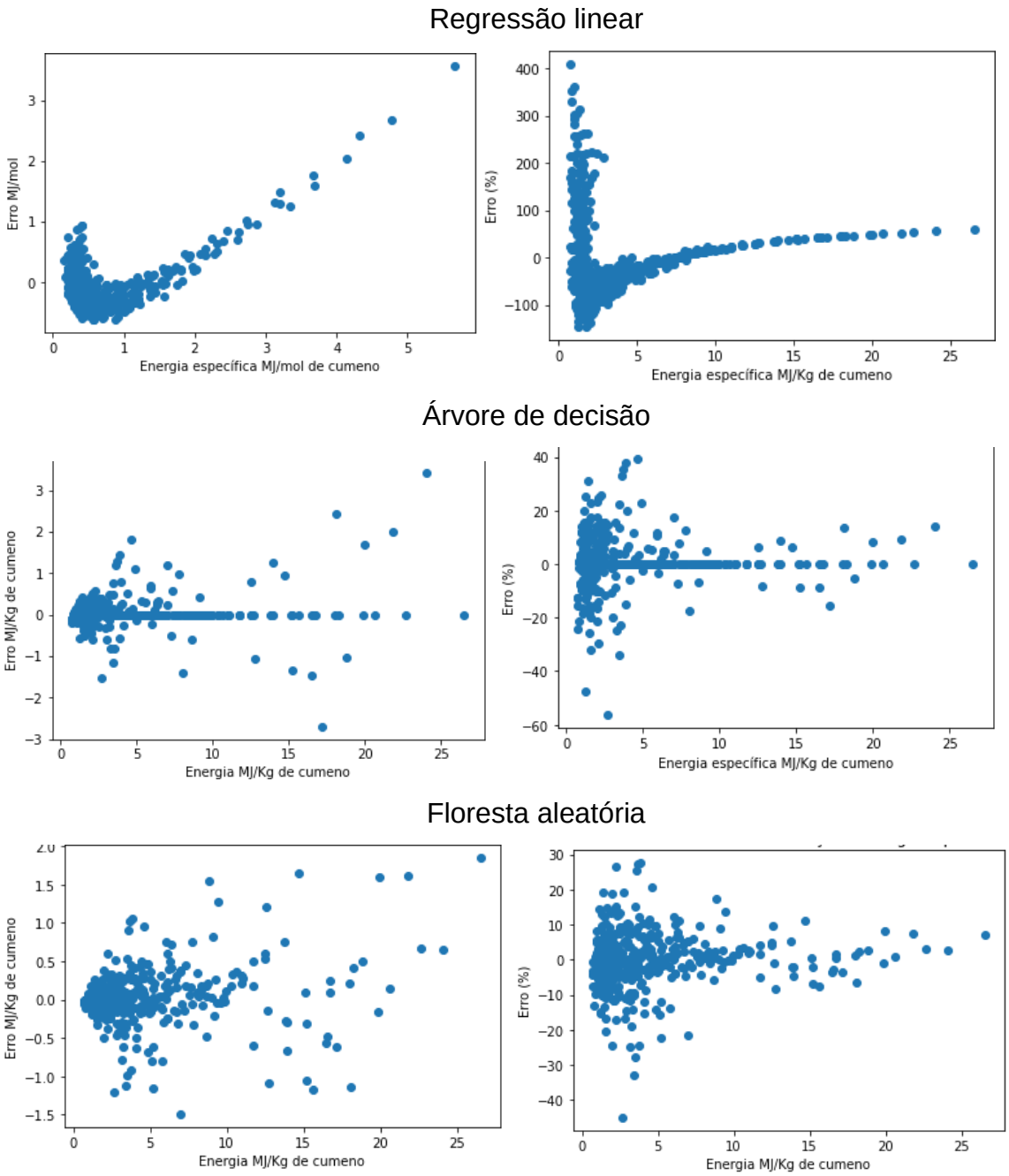
Na Tabela 16, a diferença entre o menor e o maior valor de coeficiente de determinação para a regressão linear múltipla foi de 0,15, representando um desvio de quase 25% quando comparado com o valor médio. Na Tabela 17, esse valor baixou para 0,05, o que equivale a 6% de desvio em relação ao valor médio. Com esses resultados podemos verificar que a transformação da variável de interesse melhorou significativamente o modelo de regressão linear. O comportamento dos resíduos absolutos e percentuais dos modelos de regressão linear, árvore de decisão e floresta aleatória estão sendo representados nas Figuras 16 e 17.

**Tabela 17** - Coeficientes de validação cruzada dos modelos preditivos para o ln(energia específica) no processo.

| Coeficientes de determinação | Regressão linear múltipla | Árvore de decisão | Floresta aleatória de decisão |
|---|---|---|---|
| 1 | 0,84 | 0,95 | 0,97 |
| 2 | 0,80 | 0,95 | 0,98 |
| 3 | 0,83 | 0,97 | 0,97 |
| 4 | 0,85 | 0,95 | 0,97 |
| 5 | 0,83 | 0,96 | 0,97 |
| Médio | 0,83 | 0,96 | 0,97 |

Elaborado pelo autor.

**Figura 16** - Distribuição dos resíduos absolutos e em porcentagem dos três modelos para a energia específica no processo todo.

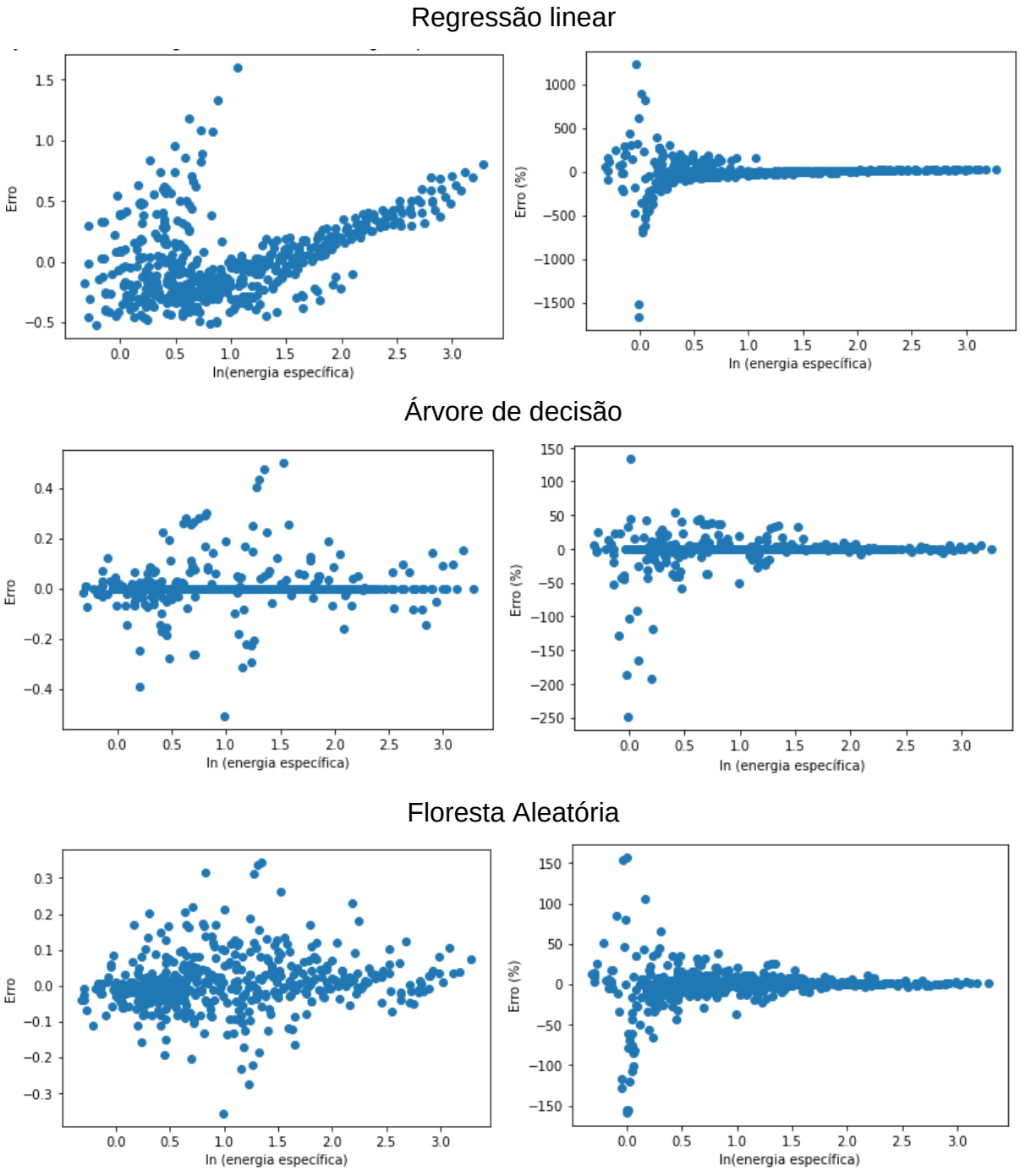



Elaborado pelo autor.

Ao analisar os resíduos nas Figuras 16 e 17, é possível perceber duas regiões com comportamentos variabilidade diferentes. Isso pode ser um indicativo de que a separação dos dados em dois grupos pode melhorar o modelo. Além disso, mesmo com a transformação da variável de interesse ainda é observado a tendência de que quanto menor o valor de energia específica, maior o erro percentual em todos os três modelos utilizados. Uma alternativa para resolver esse problema, seria aumentar o

número de dados coletados para então dividir o modelo em duas seções, de acordo com o valor de energia específica. A Tabela 18 mostra os dados referentes aos erros dos modelos preditivos para a energia específica.

**Figura 17** - Distribuição dos resíduos absolutos e em porcentagem de regressão linear para o ln (energia específica) no processo todo.




Elaborado pelo autor.

**Tabela 18** - Erro percentual médio (em módulo) dos modelos utilizados para previsão da energia específica nas seções.

| | Regressão linear | Árvore de decisão | Floresta aleatória |
|---|---|---|---|
| **Processo completo** | | | |
| Média | 67,60% | 4,28% | 5,01% |
| Desvio padrão | 61,90% | 7,46% | 5,46% |
| Mínimo | 0,02% | 0,00% | 0,00% |
| Máximo | 410% | 56,50% | 44,97% |
| 75%* | 86% | 5,84% | 6,78% |
| **Seção de reação** | | | |
| Média | 74,82% | 5,48% | 8,67% |
| Desvio padrão | 148,33% | 16,75% | 13,93% |
| Mínimo | 0,45% | 0,00% | 0,01% |
| Máximo | 57,83% | 254,89% | 8,64% |
| 75%* | 1214,28% | 5,52% | 98,58% |
| **Seção de separação** | | | |
| Média | 44,02% | 4,98% | 5,49% |
| Desvio padrão | 32,35% | 7,50% | 5,33% |
| Mínimo | 0,41% | 0,00% | 0,00% |
| Máximo | 191,10% | 45,76% | 33,24% |
| 75%* | 56,70% | 8,17% | 7,09% |

Elaborado pelo autor. *75% dos dados apresentam erros menores que os valores indicados.

O modelo de regressão linear apresentou erros percentuais maiores que os erros apresentados nos modelos baseado em árvore de decisão. Dividir o processo em seções serviu para diminuir o erro percentual médio do modelo de regressão linear para a seção de separação, mas não para a seção de reação. Em relação aos modelos baseados em árvore, tanto a seção de reação quanto a seção de separação apresentaram erros percentuais médios maiores do que o modelo aplicado para a energia específica do processo todo. Os valores de desvio padrão também aumentaram ao dividir o processo em seções. A análise do erro percentual também foi feita para os modelos que utilizaram a variável de interesse transformada. A Tabela 19 contém os dados sobre erros dos modelos para o logaritmo natural da energia específica.

**Tabela 19** - Erro percentual médio (em módulo) dos modelos utilizados para a previsão de ln(energia específica) nas seções.

| | Regressão linear | Árvore de decisão | Floresta aleatória |
|---|---|---|---|
| **Processo completo** | | | |
| Média | 123,83% | 9,51% | 13,62% |
| Desvio padrão | 618,08% | 54,74% | 36,58% |
| Mínimo | 0,00% | 0,00% | 0,00% |
| Máximo | 11098% | 3,67% | 486,59% |
| 75%* | 68,98% | 1142,02% | 10,27% |
| **Seção de reação** | | | |
| Média | 82,06% | 10,59% | 25,95% |
| Desvio padrão | 363,02% | 86,43% | 172,43% |
| Mínimo | 0,06% | 0,00% | 0,00% |
| Máximo | 5673,95% | 1352,33% | 2620,14% |
| 75%* | 20,73% | 1,15% | 5,70% |
| **Seção de separação** | | | |
| Média | 18,54% | 2,94% | 4,00% |
| Desvio padrão | 16,13% | 5,77% | 4,45% |
| Mínimo | 0,17% | 0,00% | 0,01% |
| Máximo | 106,87% | 46,21% | 31,95% |
| 75%* | 24,57% | 3,60% | 5,24% |

Elaborado pelo autor. *75% dos dados apresentam erros menores que os valores indicados.

Apesar de os coeficientes de determinação terem sido apontados como melhores para os modelos utilizando a variável de interesse transformada, os erros percentuais do modelo foram maiores para todas as seções do processo quando comparados com os modelos que previam a energia específica diretamente. Os modelos transformados também apresentaram maior desvio padrão e maior erro máximo, chegando a um pico de 11098% para a regressão linear do processo completo.

## 4.2 Energia

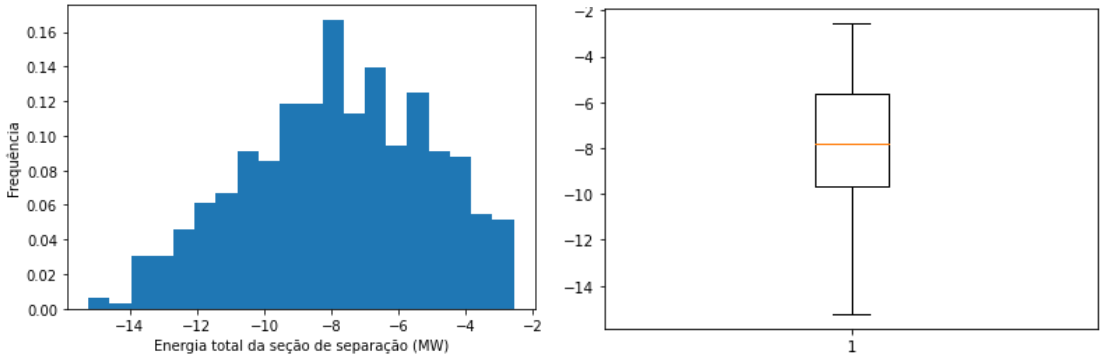
Nesta seção serão avaliados os modelos utilizados para prever a energia líquida da planta de produção de cumeno considerada neste estudo. De uma maneira geral,

a variável de interesse possui uma distribuição mais simétrica do que a distribuição da energia específica, o que contribui para um melhor desempenho dos modelos preditivos. Assim como para a energia específica, realizou-se uma análise (i) apenas na seção de separação, (ii) na seção de reação e (iii) no processo completo.

### 4.2.1 Energia na seção de separação

Realizou-se um da energia na seção de separação. A Figura 18 contém a distribuição da energia no processo. As Tabelas 20 e 21 contém os dados de coeficientes de determinação e coeficientes de determinação da validação cruzada.

**Figura 18** - Distribuição da energia na seção de separação.




Elaborado pelo autor.

**Tabela 20** - Coeficientes de correlação de treino e de teste dos modelos preditivos para a energia na seção de separação.

| Coeficiente de determinação | Regressão linear | Árvore de decisão | Floresta aleatória de decisão |
|---|---|---|---|
| Treino | 0,95 | 1,00 | 1,00 |
| Teste | 0,93 | 0,96 | 0,97 |

Elaborado pelo autor.

**Tabela 21** - Coeficientes de validação cruzada dos modelos preditivos para a energia na seção de separação.

| Coeficiente de correlação | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| 1 | 0,94 | 0,98 | 0,98 |
| 2 | 0,95 | 0,97 | 0,97 |
| 3 | 0,95 | 0,97 | 0,98 |
| 4 | 0,95 | 0,98 | 0,98 |
| 5 | 0,94 | 0,97 | 0,99 |
| Médio | 0,95 | 0,97 | 0,98 |

Elaborado pelo autor.

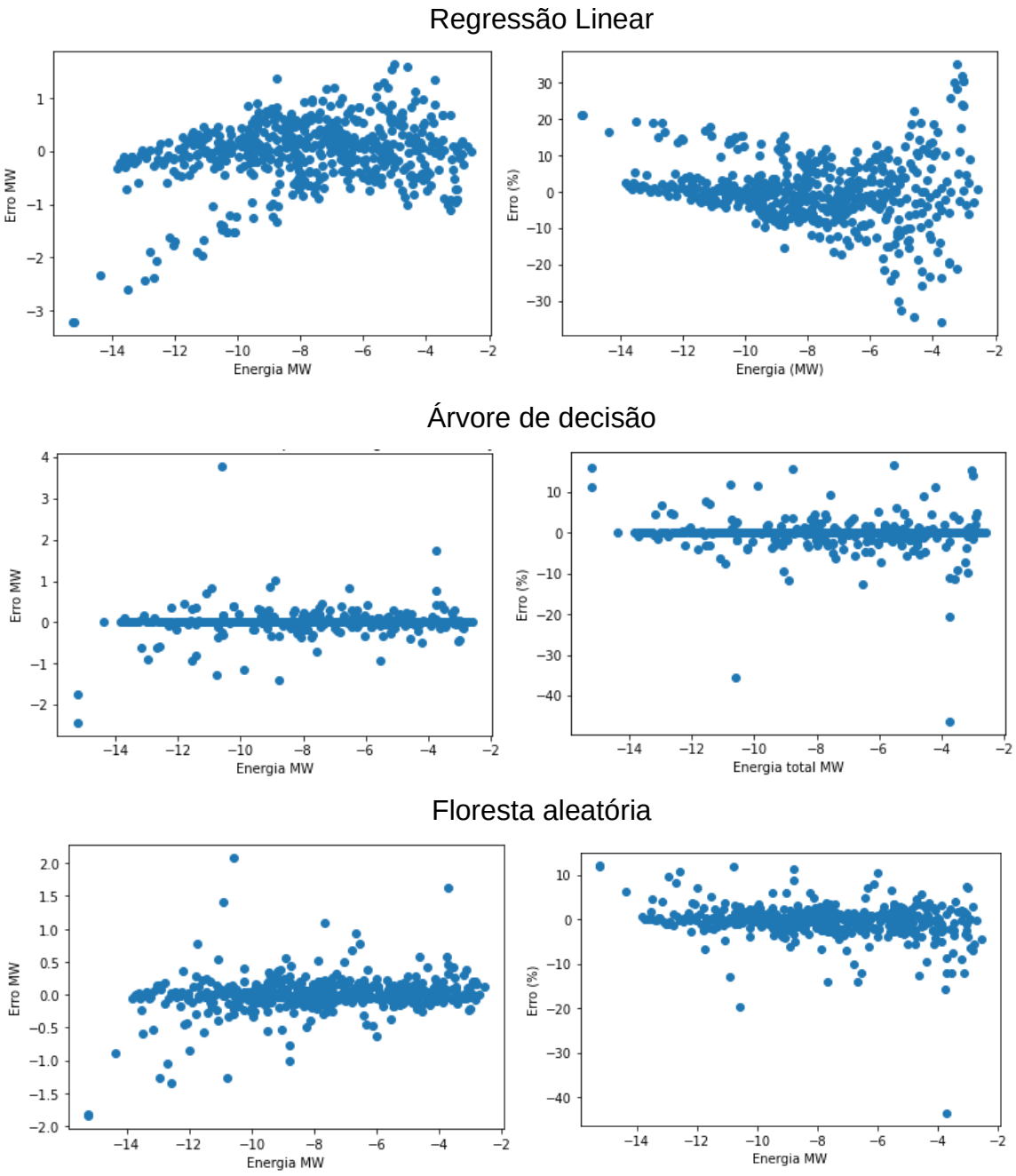
Os coeficientes de determinação para a regressão linear múltipla na predição da energia total mostraram-se superiores aos coeficientes de determinação na predição da energia específica, com uma média de 0,95 na validação cruzada. Essa tendência se manteve ao longo de toda análise de energia total. Dessa forma, não houve a necessidade de uma transformação da variável de interesse em nenhuma das seções do processo. A Tabela 22 apresenta dados sobre o erro percentual dos modelos utilizados. A Figura 19 apresenta a distribuição dos resíduos dos modelos utilizados.

**Tabela 22** - Erro percentual (em módulo) dos modelos utilizados para a previsão da energia total na seção de separação.

| | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| Média | 6,42% | 1,36% | 2,07% |
| Desvio padrão | 6,54% | 3,68% | 3,23% |
| Mínimo | 0,01% | 0,00% | 0,01% |
| Máximo | 35,83% | 46,31% | 43,63% |
| 75%* | 8,97% | 1,32% | 2,44% |

Elaborado pelo autor. *75% dos dados apresentam erros menores que os valores indicados.

**Figura 19** - Distribuição dos resíduos absolutos e em porcentagem dos três modelos na seção de separação.

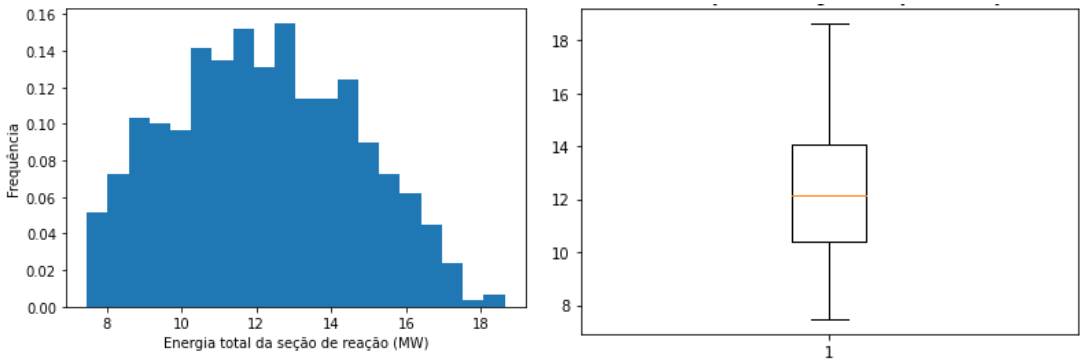



Elaborado pelo autor.

### 4.2.2 Seção de reação

Realizou-se um estudo da energia na seção de reação. A Figura 20 contém a distribuição da energia no processo. As Tabelas 23 e 24 contém os dados de coeficientes de determinação e coeficientes de determinação da validação cruzada.

**Figura 20** - Distribuição da energia total da seção de reação.




Elaborado pelo autor.

**Tabela 23** - Coeficientes de correlação de treino e de teste dos modelos preditivos para a energia na seção de reação.

| Coeficiente de correlação | Regressão linear | Árvore de decisão | Floresta aleatória |
|---|---|---|---|
| Treino | 0,95 | 1,00 | 1,00 |
| Teste | 0,93 | 0,96 | 0,97 |

Elaborado pelo autor.

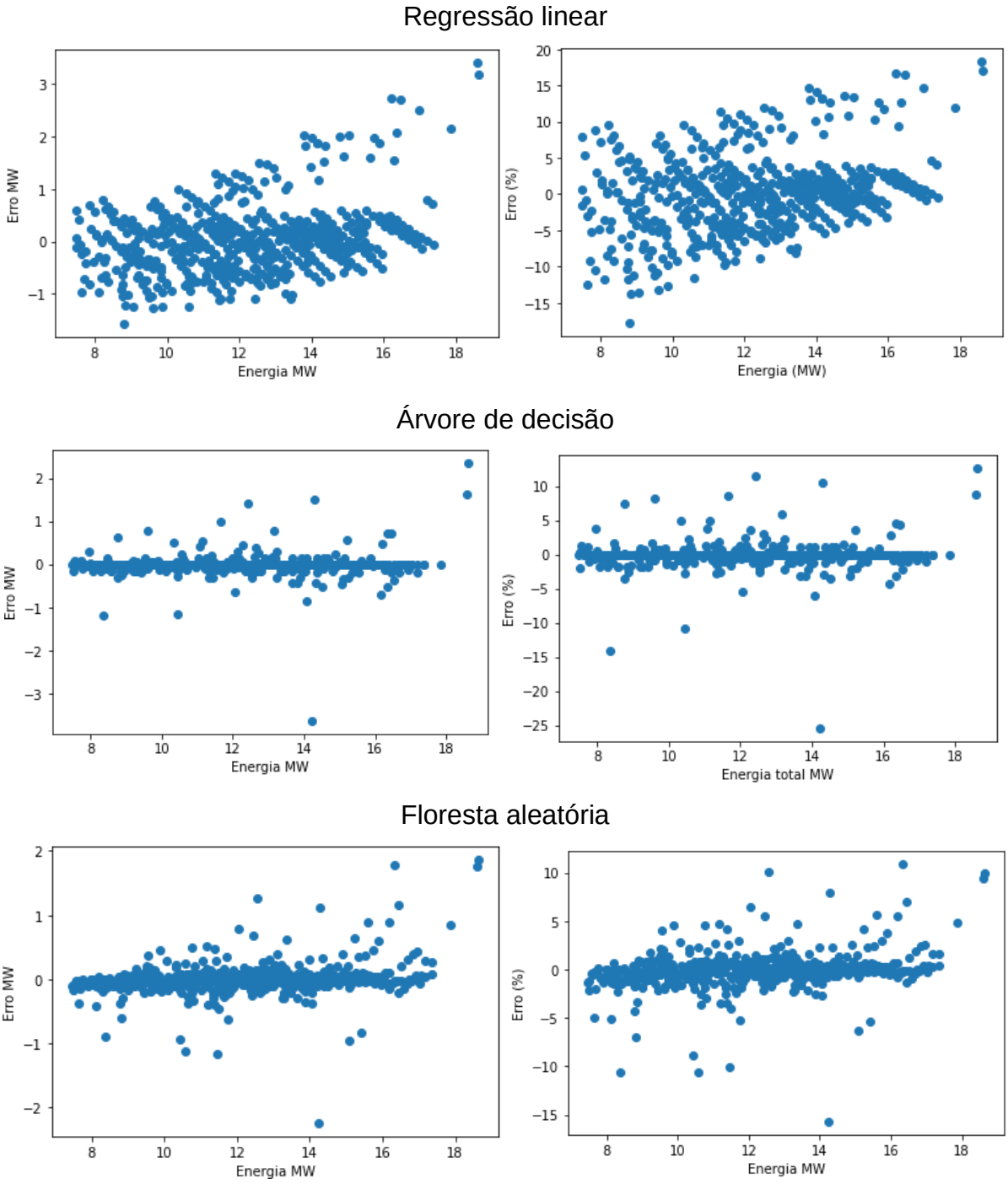
**Tabela 24** - Coeficientes de validação cruzada dos modelos preditivos para a energia na seção de reação.

| Coeficiente de correlação | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| 1 | 0,90 | 0,91 | 0,96 |
| 2 | 0,93 | 0,96 | 0,97 |
| 3 | 0,94 | 0,97 | 0,98 |
| 4 | 0,92 | 0,96 | 0,98 |
| 5 | 0,92 | 0,97 | 0,98 |
| Médio | 0,92 | 0,95 | 0,97 |

Elaborado pelo autor.

Para a validação dos modelos, realizou-se um estudo dos resíduos na seção reação. A Figura 21 apresenta a distribuição dos resíduos para os três modelos, enquanto a Tabela 25 contém os dados referentes aos erros percentuais.

**Figura 21** - Distribuição dos resíduos absolutos e em porcentagem dos modelos de regressão linear, árvore de decisão e floresta aleatória para a energia na seção de reação.




Elaborado pelo autor.

Os resíduos apresentam o comportamento esperado, possuindo média zero e variância constante.

**Tabela 25** - Erro percentual (em módulo) dos modelos utilizados para a previsão da energia total na seção de reação.

| | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| Média | 4,04% | 0,69% | 1,13% |
| Desvio padrão | 3,61% | 1,95% | 1,74% |
| Mínimo | 0,00% | 0,00% | 0,00% |
| Máximo | 18,27% | 25,44% | 15,72% |
| 75%* | 5,71% | 0,74% | 1,24% |

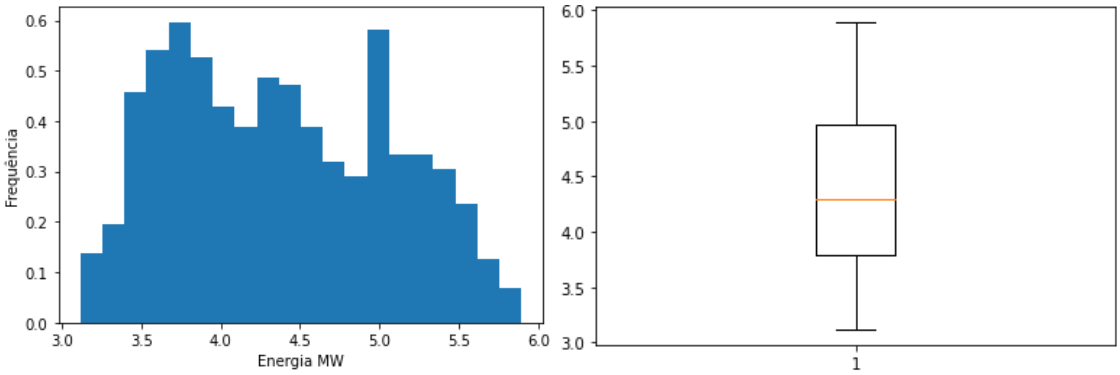
Elaborado pelo autor. *75% dos dados apresentam erros menores que os valores indicados.

Ao analisar a Tabela 25, utilização do modelo preditivo mostrou que 75% dos dados apresentaram erros de regressão linear menores que 6%, enquanto na árvore de decisão, por exemplo, 75% dos erros não chegaram a 1%.

### 4.2.3 Análise do processo completo

Realizou-se um estudo da energia no processo completo. A Figura 22 contém a distribuição da energia no processo completo. As Tabelas 26 e 27 contém os dados de coeficientes de determinação e coeficientes de determinação da validação cruzada.

**Figura 22** - Distribuição da energia no processo completo.




Elaborado pelo autor.

**Tabela 26** - Coeficientes de determinação de treino e de teste dos modelos preditivos para a energia no processo todo.

| Coeficiente de determinação | Regressão linear | Árvore de decisão | Floresta aleatória de decisão |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Treino | 0,93 | 1,00 | 1,00 |
| Teste | 0,91 | 0,96 | 0,97 |

Elaborado pelo autor.

**Tabela 27** - Coeficientes de validação cruzada dos modelos preditivos para a energia no processo.

| Coeficiente de determinação | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| 1 | 0,89 | 0,95 | 0,97 |
| 2 | 0,87 | 0,93 | 0,97 |
| 3 | 0,84 | 0,94 | 0,95 |
| 4 | 0,87 | 0,94 | 0,96 |
| 5 | 0,89 | 0,94 | 0,96 |
| Médio | 0,87 | 0,94 | 0,96 |

Elaborado pelo autor.

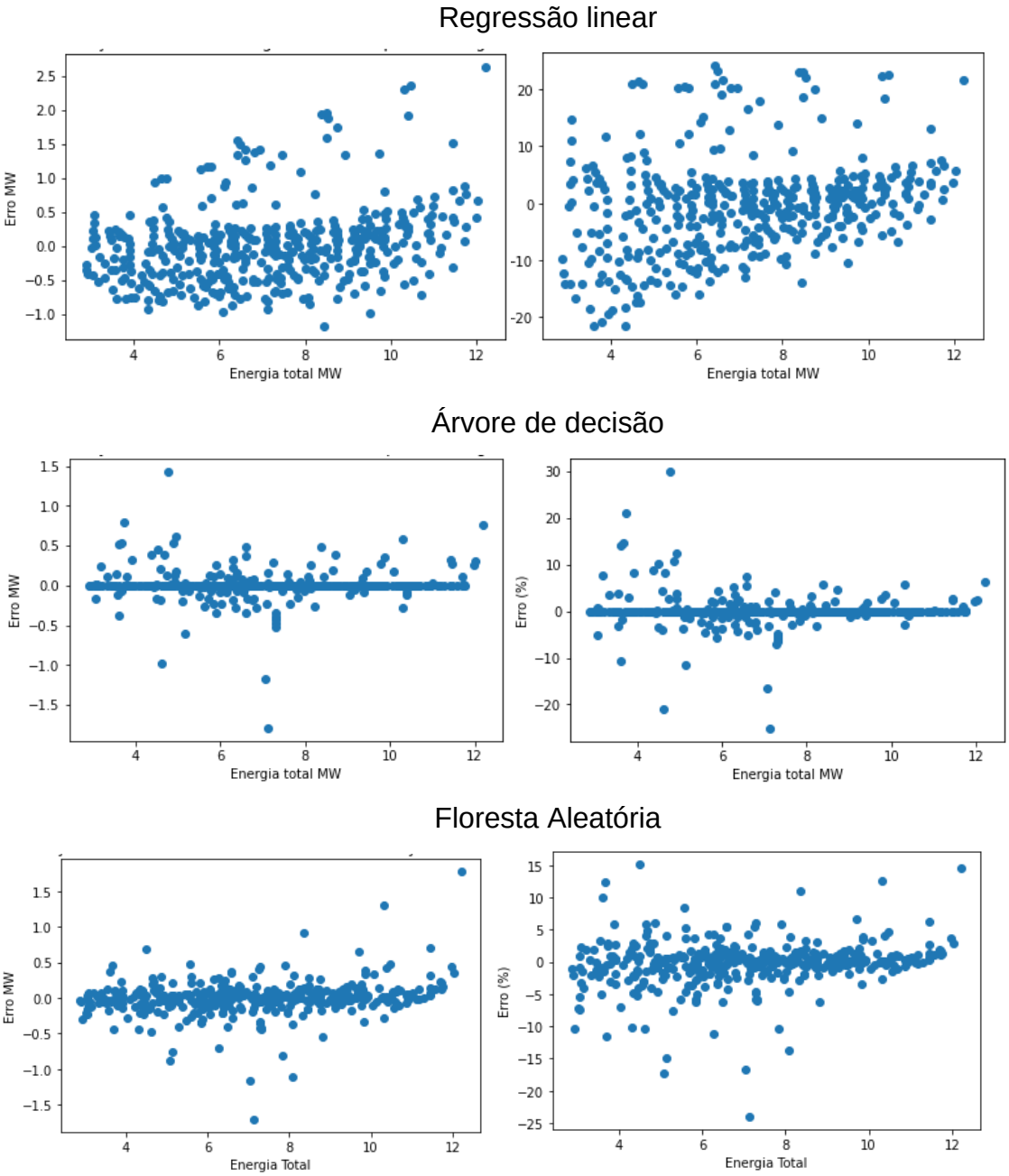
A Figura 23 apresenta a distribuição dos erros para os três modelos preditivos utilizados. Os resíduos dos modelos preditivos para a energia total, sobretudo para os modelos baseados em árvore de decisão, apresentaram um comportamento desejado: média zero e hemocedasticidade. A Tabela 28 apresenta os dados referentes ao erro percentual da predição da energia total do processo.

**Tabela 28** - Erro percentual (em módulo) dos modelos utilizados para a previsão da energia total no processo.

| | Regressão Linear | Árvore de Decisão | Floresta Aleatória |
|---|---|---|---|
| Média | 4,68% | 0,71% | 1,10% |
| Desvio padrão | 3,82% | 1,62% | 1,26% |
| Mínimo | 0,05% | 0,00% | 0,00% |
| Máximo | 23,01% | 11,83% | 11,57% |
| 75%* | 6,25% | 0,63% | 1,46% |

Elaborado pelo autor. *75% dos dados apresentam erros menores que os valores indicados.

**Figura 23** - Distribuição dos resíduos absolutos e em porcentagem dos modelos de regressão linear, árvore de decisão e floresta aleatória para a energia no processo completo.




Elaborado pelo autor.

O maior erro percentual foi o do modelo de regressão linear (23,01%), enquanto o menor erro percentual foi obtido pelo modelo de floresta aleatória (11,57%). Para o modelo de árvore de decisão, 75% dos dados tiveram erros menores que 0,63%. Sendo, assim, o modelo se mostrou bastante acurado na predição da energia do processo.

Comparando os coeficientes de determinação das seções separadas com os valores de coeficientes de determinação do processo completo, verifica-se que a análise das seções separadas apresenta erros percentuais máximos maiores que a análise do processo completo. No entanto, a utilização do modelo preditivo continua promissor, uma vez que 75% dos erros de regressão linear são menores que 9%, enquanto na árvore de decisão, por exemplo, 75% dos erros não chegaram a 2%.

## 4.3 Discussão dos resultados

Como visto nas análises apresentadas, os modelos utilizados foram mais acurados na predição de energia total do que na predição da energia específica. Isso pode ter acontecido pelo aumento da colinearidade entres as variáveis de entrada, uma vez que a energia específica é calculada dividindo-se a energia total pela vazão de cumeno. A vazão de cumeno, por sua vez, está relacionada com as vazões de benzeno e propeno, bem como a temperatura e pressão do processo. O aumento do número de variáveis de entrada, bem como a obtenção de um número maior de dados para permitir a investigação dos dois comportamentos existentes nos gráficos (uma seção para valores de energia específica maiores e outra para valores de energia específica menores).

Para a predição da energia total 75% dos erros percentuais são menores que 10%. Isso mostra o potencial da utilização de *machine learning* em processos químicos. Essas predições podem ser feitas online, sem a necessidade de instalação de um simulador de processos químicos, facilitando a integração com outros *softwares,* utilizando modelos pouco dependentes de *hardware* e alta compreensibilidade, permitindo uma boa comunicação com todos os setores da indústria, e utilizando apenas quatro variáveis de entrada: vazão molar de benzeno, vazão molar de propeno, temperatura e pressão.

Além disso, é importante salientar que para esse projeto os modelos foram validados com base no coeficiente de determinação ($R^2$) que leva em consideração o erro ao quadrado das predições. O $R^2$ pode não ser a melhor alternativa para a validação de variáveis com comportamento assimétrico, como foi a energia específica, assim, para trabalhos futuros seria importante observar outras métricas de validação de métodos de regressão.

Em relação às implicações desses modelos no processo de produção de cumeno, eles podem ser vistos como uma alternativa mais rápida, barata e acessível

a não especialistas do que um simulador de processos químicos, desde que o fluxograma do processo não passe por alterações. Uma vez que esses modelos representaram uma configuração específica do processo, qualquer mudança no projeto do processo, como número de estágios das colunas de destilação, configurações dos trocadores de calor e mudanças nas reações podem alterar significativamente as respostas e um novo modelo precisaria ser treinado para essas condições. Além disso, os erros desses modelos estão atrelados ao quão bem o simulador de processos representou o sistema. Modelos treinados por dados coletados de uma planta real podem gerar predições mais próximas da realidade.

# 5 CONCLUSÃO

No presente trabalho foi possível perceber a presença de ciência de dados na engenharia química. Diversas empresas têm investido nesse setor, além de novas soluções, como a iniciativa da AspenTech de integração com as indústrias 4.0, estarem surgindo nessa área.

Em relação ao uso de *machine* learning na predição do processo de produção de cumeno, os modelos preditivos utilizados foram acurados na predição da energia total do processo, porém não apresentaram um bom desempenho na predição da energia específica. Quanto a separação do processo em seções, essa abordagem não melhorou o melhorou o modelo, então recomenda-se que a análise ocorra para a o processo completo.

O uso de modelos preditivos para a predição de energia na planta se mostrou promissor, uma vez que requer menos requisitos computacionais do que uma simulação em softwares como o *Aspen Plus®*, podendo ser utilizando pelo próprio *browser* de internet, sem a necessidade de uma instalação e com 75% dos erros menores que 10%. Além disso, o modelo requer apenas quatro entradas: vazão molar de benzeno, vazão molar de propeno, temperatura e pressão do processo.

Os modelos baseados em árvore se mostraram superiores ao modelo de regressão linear, no entanto esses modelos não possuem uma capacidade de extrapolação, o que limita a sua utilização dentro da faixa de dados utilizada. Entre o modelo de árvore de decisão e o modelo de floresta aleatória, o primeiro requer menos usos computacionais o que o torno favorito na criação de um modelo mais simples.

Para trabalhos futuros, sugere-se o aumento do número de dados coletados, de forma que seja possível trabalhar com pelo menos duas faixas de energia específica. Para potencializar o uso da regressão linear, sugere-se também, escolher aumentar o número de variáveis independentes além das variáveis propostas. Por último, sugere-se utilizar outras formas de validação de modelos que não seja o $R^2$ proposto nesse trabalho.

## REFERÊNCIAS BIBLIOGRÁFICAS


Al-Kinany, M. C., Al-Khowaiter, S. H., & Al-Malki, F. H. (2001). **Synthesis of cumene (isopropylbenzene) from diisopropylbenzenes in the presence of benzene using triflic acid as catalyst at room temperature**. In *Studies in Surface Science and Catalysis* (Vol. 133). Elsevier Masson SAS. https://doi.org/10.1016/s0167-2991(01)81995-9

**Artificial Intelligence of Things**. (2021). Aspentech. https://www.aspentech.com/en/solutions/artificial-intelligence-of-things

*Aspen Plus*® (No. 10). (n.d.). aspen tech.

Ayodele, T. O. (2010). **Introduction to machine learning**. In Y. Zhang (Ed.), *New Advances in Machine Learning* (pp. 1–8). InTech. http://www.intechopen.com/books/new-advances-in-machine-learning/introduction-to-machine-learning

Bhattacharyya, S. (2018). ***Ridge and Lasso Regression: L1 and L2 Regularization*. Towards Data Scienc**e. https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

Bianchi, A. L. (2020). ***Engenharia de features e dados: como ter um aprendizado de máquina eficiente***. Viceri. https://www.viceri.com.br/insights/engenharia-de-features-e-dados-como-ter-um-aprendizado-de-maquina-eficiente

Bonfim, C. A. (2020). ***O que é bias-variance tradeoff: como isso impacta seu modelo de Machine Learning***. Data Hackers. https://medium.com/data-hackers/o-que-é-bias-variance-tradeoff-a5bc19866e4b

Boulesteix, A. L., Janitza, S., Kruppa, J., & König, I. R. (2012). **Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics**. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(6), 493–507. https://doi.org/10.1002/widm.1072

Brownlee, J. (2019). ***A Gentle Introduction to the Bootstrap Method*. Machine Learning Mastery**. https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/

Chudinova, A., Salischeva, A., Ivashkina, E., Moizes, O., & Gavrikov, A. (2015). **Application of Cumene Technology Mathematical Model**. *Procedia Chemistry*, *15*, 326–334. https://doi.org/10.1016/j.proche.2015.10.052

***Ciência de Dados ou Data Science: O que é, Aplicações e Perfil Profissional***.

(2019). Fundação Instituto de Admnistração. https://fia.com.br/blog/ciencia-de-dados-data-science/

***Colaboratory***. (n.d.). Google. Retrieved July 11, 2021, from https://colab.research.google.com/notebooks/intro.ipynb?utm_source=scs-index#recent=true

***CUMENE MARKET - GROWTH, TRENDS, COVID-19 IMPACT, AND FORECASTS (2021 - 2026)*. (2020)**. Mordor Intelligence. https://www.mordorintelligence.com/industry-reports/cumene-market

***Cumene Market Size, Share & Trends Analysis Report By Production (Zeolite, Solid Phosphoric Acid, Aluminum Chloride), By Application (Phenol, Acetone), By Region, And Segment Forecasts, 2018 - 2025*. (2017)**. Grand View Research. https://www.grandviewresearch.com/industry-analysis/cumene-market

Davenport, T. H., & Patil, D. J. (2012, October). **Data Scientist: The Sexiest Job of the 21st Century**. *HarvardBusiness Review*.

De Miranda Ramos Soares, A. P., De Oliveira Carvalho, F., De Farias Silva, C. E., Da Silva Gonçalves, A. H., & De Souza Abud, A. K. (2020). **Random Forest as a promising application to predict basic-dye biosorption process using orange waste**. *Journal of Environmental Chemical Engineering*, *8*(4), 103952. https://doi.org/10.1016/j.jece.2020.103952

Ghiraldini, M. (2020). ***Porque o cientista de dados ser a profissão do século não é modinha***. ProXXIma. https://www.proxxima.com.br/home/proxxima/how-to/2020/07/31/porque-o-cientista-de-dados-ser-a-profissao-do-seculo-nao-e-modinha.html

Gupta, N. (2013). ***Artificial Neural Network*** (Vol. 3).

Junqueira, P. G., Mangili, P. V., Santos, R. O., Santos, L. S., & Prata, D. M. (2018). **Economic and environmental analysis of the cumene production process using computational simulation**. *Chemical Engineering and Processing - Process Intensification*, *130*, 309–325. https://doi.org/10.1016/j.cep.2018.06.010

Kayri, M., Kayri, I., & Gencoglu, M. T. (2017). **The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data**. *2017 14th International Conference on Engineering of Modern Electric Systems, EMES 2017*, 1–4. https://doi.org/10.1109/EMES.2017.7980368

Kotsiantis, S. B. (2007). **Supervised Machine Learning: A Review of Classification Techniques**. In *Informatica* (Vol. 31, Issue 1, pp. 249–268).

Mainmon, O., & Rokach, L. (2005). ***The Data Mining and Knowledge Discovery Handbook***. Springer.

*McKinsey & Company*. (n.d.). https://www.mckinsey.com.br/

Mijwel, M. M. (2018). **Artificial Neural Networks Advantages and Disadvantages**. *Linkedin*, *March*, 1–2. https://www.researchgate.net/publication/323665827

Moura, W. (2016). ***Resampling: separando os dados entre treino, validação e teste***. Hacking Analytics. https://hackinganalytics.com/2016/09/04/resampling-separando-os-dados-entre-treino-validacao-e-teste/

Müller, A. C., & Guido, S. (2017). ***Introduction to Machine Learning with Python*** (D. Schanafelt & Judy McConville (Eds.); 1st ed.). O'Reilly Media, Inc. https://doi.org/10.1007/978-3-030-36826-5_10

Nandi, S., Badhe, Y., Lonari, J., Sridevi, U., Rao, B. S., Tambe, S. S., & Kulkarni, B. D. (2004). **Hybrid process modeling and optimization strategies integrating neural networks/support vector regression and genetic algorithms: Study of benzene isopropylation on Hbeta catalyst**. *Chemical Engineering Journal*, *97*(2–3), 115–129. https://doi.org/10.1016/S1385-8947(03)00150-5

Nourani, V., Elkiran, G., & Abba, S. I. (2018). **Wastewater treatment plant performance analysis using artificial intelligence - An ensemble approach**. *Water Science and Technology*, *78*(10), 2064–2076. https://doi.org/10.2166/wst.2018.477

Ohri, A. (2021). ***8 Popular Regression Algorithms In Machine Learning Of 2021***. Jigsaw Academy. https://www.jigsawacademy.com/popular-regression-algorithms-ml/

Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell, J. B. O. (2007). **Random forest models to predict aqueous solubility**. *Journal of Chemical Information and Modeling*, *47*(1), 150–158. https://doi.org/10.1021/ci060164k

Partopour, B., Paffenroth, R. C., & Dixon, A. G. (2018). **Random Forests for mapping and analysis of microkinetics models**. *Computers and Chemical Engineering*, *115*, 286–294. https://doi.org/10.1016/j.compchemeng.2018.04.019

Prates, W. R. (2018). ***O que é árvore de decisão (decision tree)? Exemplos em R***. Ciência&Negócios.Com. https://cienciaenegocios.com/o-que-e-arvore-de-decisao-decision-tree-linguagem-r/

***Random Forest Algorithm- An Overview***. (2020). Great Learning. https://www.mygreatlearning.com/blog/random-forest-algorithm/

Reese, H. (2017). ***Understanding the differences between AI , machine learning , and deep learning***. 1–4. https://www.techrepublic.com/article/understanding-the-differences-between-ai-machine-learning-and-deep-learning/

Sasikumar, S. (2021). ***Data Science vs. Data Analytics vs. Machine Learning: Expert Talk***. Simpli Learn. https://www.simplilearn.com/data-science-vs-data-analytics-vs-machine-learning-article

Shin, J., Badgwell, T. A., Liu, K. H., & Lee, J. H. (2019). **Reinforcement Learning – Overview of recent progress and implications for process control**. *Computers and Chemical Engineering*, *127*, 282–294. https://doi.org/10.1016/j.compchemeng.2019.05.029

Sutton, R. S., & Barto, A. G. (1998). ***Reinforcement Learning: an introduction***. The MIT Press.

***THE PRODUCTION OF CUMENE USING Zeolite CATALYST***. (1999). Aspen Model Documentation. http://www.diquima.upm.es/old_diquima/docencia/tqindustrial/docs/cumeno.pdf

Tso, G. K. F., & Yau, K. K. W. (2007). **Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks**. *Energy*, *32*(9), 1761–1768. https://doi.org/10.1016/j.energy.2006.11.010

Tunckaya, Y., & Koklukaya, E. (2015). **Comparative analysis and prediction study for effluent gas emissions in a coal-fired thermal power plant using artificial intelligence and statistical tools**. *Journal of the Energy Institute*, *88*(2), 118–125. https://doi.org/10.1016/j.joei.2014.07.003

Wang, S. C. (2003). **Artificial neural network (ANNs)**. In *Interdisciplinary Computing in Java Programming* (pp. 81–100). Kluwer Academic Publishers. https://doi.org/10.1007/978-3-319-67466-7_4

Wittek, P. (2014). **Unsupervised Learning.** ***Quantum Machine Learning***, 57–62. https://doi.org/10.1016/b978-0-12-800953-6.00005-0

Wujek, B., Hall, P., & Günes, F. (2016). ***Best Practices for Machine Learning Applications*** (No. 2360). https://www.lexjansen.com/wuss/2016/154_Final_Paper_PDF.pdf

Yan, Y., Borhani, T. N., & Clough, P. T. (2020). **Machine Learning Applications in Chemical Engineering**. In *Machine Learning in Chemistry: The Impact of Artifical*

*Intelligence* (CPI Group, pp. 340–371). Royal Society of Chemistry. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2577126%0Ahttps://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=6317305%0Ahttps://doi.org/10.1039/9781839160233%0Ahttp://www.vlebooks.com/vleweb/product/

Yen, L. (2019). ***An Introducvtion to the Bootstrap Method***. Towards Data Science. https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60

Zahraee, S. M., Khalaji Assadi, M., & Saidur, R. (2016). **Application of Artificial Intelligence Methods for Hybrid Energy System Optimization**. *Renewable and Sustainable Energy Reviews*, *66*, 617–630. https://doi.org/10.1016/j.rser.2016.08.028

Zhang, S., Zhang, C., & Yang, Q. (2003). **Data Preparation for Data Mining**. *Appl. Artif. Intel.*, *17(5–6)*, 375–381. https://doi.org/10.1080/08839510390219264