UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

CURSO DE ENGENHARIA QUÍMICA

LARISSA SOUZA PINHEIRO

AVALIAÇÃO DE EFICIÊNCIA ENERGÉTICA EM SISTEMAS DE ÁGUA GELADA PARA REFINARIAS DE ÓLEO VEGETAL UTILIZANDO MINERAÇÃO DE DADOS

FORTALEZA

2022

LARISSA SOUZA PINHEIRO

AVALIAÇÃO DE EFICIÊNCIA ENERGÉTICA EM SISTEMAS DE ÁGUA GELADA PARA REFINARIAS DE ÓLEO VEGETAL UTILIZANDO MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia Química da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de Bacharel em Engenharia Química.

Orientadora: Prof.ª Dr.ª Andréa da Silva Pereira.

FORTALEZA

2022

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

P72a Pinheiro, Larissa Souza.
Avaliação de eficiência energética em sistemas de água gelada para refinarias de óleo vegetal utilizando mineração de dados / Larissa Souza Pinheiro. – 2022.
96 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Química, Fortaleza, 2022.
Orientação: Profa. Dra. Andréa da Silva Pereira.

1. Eficiência energética. 2. Refrigeração. 3. COP. 4. Mineração de dados. 5. Aprendizado de Máquina. I. Título.

CDD 540

LARISSA SOUZA PINHEIRO

AVALIAÇÃO DE EFICIÊNCIA ENERGÉTICA EM SISTEMAS DE ÁGUA GELADA PARA REFINARIAS DE ÓLEO VEGETAL UTILIZANDO MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia Química da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de Bacharel em Engenharia Química.

Aprovada em: 01/12/2022

BANCA EXAMINADORA

_____

Prof.ª Dr.ª Andréa da Silva Pereira (Orientadora)

Universidade Federal do Ceará (UFC)

_____

Eng. Ismael Francisco Miranda Freire

_____

Prof.ª Dr.ª Maria Valderez Ponte Rocha

Universidade Federal do Ceará (UFC)

À minha avó, Maria

In memoriam

AGRADECIMENTOS

À minha mãe, Sandra, primeiramente, por todo o amor, incentivo e apoio incondicionais durante toda a minha vida. Obrigada por toda a educação, formal e informal, e por acreditar, que mesmo com todas as limitações impostas pela vida, eu chegaria lá.

Ao meu pai, José, que mesmo distante, também sempre procurou me incentivar.

Ao meu companheiro, Leonardo, que sempre enxergou muito em mim, e fez desses anos os mais leves, embora difíceis. Obrigada pelo cuidado e pela preocupação.

À minha melhor amiga, Amanda, que mesmo de longe, sempre se manteve presente. Do jardim de infância ao ensino médio juntas, e acredito que só não seguimos pela mesma faculdade por ela ser péssima em exatas, e porque eu jamais faria Direito.

Aos amigos que ganhei na UFC, em especial, aos componentes do grupo de estudos mais memorável de todos: Ana Luisa, Jaryson, Keila, Lívia, Luiza, Lucas, M. Luiza, Pedro, Vitor, V. Natasha e V. Nunes. Obrigada por encontrarem os mais sinceros momentos de risadas em meio ao caos, e por sempre acreditarem que conseguiríamos.

A todos os professores, orientadores e colegas de iniciação científica do GPSA, bem como aos tutores da Escola Piloto de Engenharia Química (EPEQ-UFC), meus mais sinceros agradecimentos pelas oportunidades de desenvolvimento.

Um agradecimento, em especial, à Bella, por ter sido minha maior dupla acadêmica, e também uma amiga que espero levar para a vida. Obrigada por me ensinar – e ajudar – tanto, e por ter sido a pessoa com quem pude compartilhar uma das felicidades mais genuínas até hoje, que foi a publicação do nosso primeiro artigo.

Aos amigos que ganhei na vida profissional, em especial: Pedro e Gabryel, por me mostrarem o lado simples da vida. Ao meu chefe, Renan, pela ideia do projeto, ao Mateus por me ensinar e escutar tanto, e ao Henrique, por ter sido meu "padrinho" de estágio.

À minha orientadora Andréa, que é alguém que sempre admirei desde o primeiro ano de graduação, pela inteligência e destaque acadêmico. Obrigada por confiar no meu trabalho.

Ao Ismael, que foi uma das grandes pessoas que pude conhecer na indústria, e que me ensinou, e continua ensinando tanto. Obrigada pela paciência e pelos puxões de orelha.

À professora M. Valderez, por ser uma verdadeira "mãe" de seus alunos na Engenharia Química, e excelente profissional que tive o prazer de ser aluna.

À Universidade Federal do Ceará, por tudo.

“O seu melhor e mais sábio refúgio de todos os problemas está na sua ciência.”

Ada Lovelace

RESUMO

Eficiência energética é um tema atemporal quando se pensa em plantas industriais, sobretudo em sistemas de refrigeração. Em geral, tais sistemas funcionam por meio da compressão de fluidos refrigerantes. Consequentemente, para essa operação, é exigido trabalho de eixo do motor - fonte elétrica - para que o fluido seja comprimido, e depois, respectivamente, condensado, expandido e evaporado, retornando ao seu estado inicial. Para que esse ciclo termodinâmico seja minimamente eficiente, é preciso que a capacidade de refrigeração, em kW, seja maior do que a energia elétrica fornecida. Contudo, esse tipo de indicador não é estático, e depende de diversas variáveis de processo. Neste ponto, o uso cada vez maior de tecnologias para a digitalização de instrumentos de medição em equipamentos torna-se um fator competitivo para a análise e predição de um grande volume de dados em tempo real. Dessa forma, o principal objetivo deste trabalho foi o de minerar dados de um circuito de resfriamento de água gelada (*Chiller*) a fim de extrair conhecimento útil do sistema. Por tratar-se de uma planta em operação, a metodologia utilizada foi o CRISP-DM, na qual foram listadas, coletadas e tratadas as variáveis que supostamente impactam a capacidade de refrigeração do circuito. A variável alvo calculada foi o COP (Coeficiente Operacional de Performance), o qual mede a relação da carga térmica de refrigeração pela energia fornecida ao ciclo. A partir daí, separou-se os dados do COP em duas regiões de operação denominadas "Ruim" ou "Regular", nas quais as outras variáveis de operação foram classificadas. Para isso, foram testados três modelos de aprendizado de máquina: Árvore de Decisão, Floresta Aleatória e XGBoost, com os três possuindo mais de 80% de desempenho. O XGBoost mostrou-se ser o melhor modelo, o qual apontou que as variáveis mais significativas para o COP são: temperatura de entrada e saída do evaporador, perda de carga no evaporador, temperatura de saída do condensador, perda de carga no condensador, temperatura de superaquecimento de descarga, temperatura de entrada do condensador, corrente do motor do compressor e volume da câmara de compressão. O projeto visa trazer um tema real de melhoria industrial atrelado a uma abordagem moderna de resolução, utilizando linguagens computacionais, como Python e SQL, e ilustrando como é o perfil cada vez mais exigido de um engenheiro químico hoje na indústria.

Palavras-chave: Eficiência energética; Refrigeração; COP; Mineração de dados; Aprendizado de Máquina.

ABSTRACT

Energy efficiency is an ageless subject in industrial plants, mainly in refrigeration systems. Overall, these systems work by refrigerant fluid compression. Hence, it is requested power from the engine – an electric source – to compress, condense, expand and evaporate the fluid, returning to its outset state. For this thermodynamic cycle occurs in its minimal efficiency, it is necessary that the cooling capacity (kW) be higher than the provided electrical power. Nevertheless, this type of indicator is not static, depending on several process variables. Thus, the increasing use of digital technologies for equipment measurement instruments has become a competitive condition for the analyse and prediction of a huge volume of data in current time. Thereby, the main objective of this work was mining data from a chill water refrigeration circuit (Chiller) in order to take out knowledge from the system. Once the plant was fully operating, the methodology used was CRISP-DM, in which the seeming variables that influence the cooling capacity were listed, collected and treated. The target variable calculated was OCP (Operational Coefficient of Performance), which measure the ratio between heat load and the supplied energy to the cycle. Then, the OCP data was split in two operational ranges labelled “Bad” and “Regular”, in which the other variables were classified. For that, three machine learning based models were tested: Decision Tree, Random Forest and XGBoost, with all three models owning more than 80% in score metrics. XGBoost was the best model tested, which pointed out the most feature importance for OCP: inlet and outlet temperatures from evaporator, head loss in the evaporator, outlet condenser temperature, head loss in the condenser, superheat discharge temperature, inlet condenser temperature, electrical current in the engine and compression camara volume. The project aims to bring a real industrial improvement case connected with a modern resolution approach, using computational languages, like Python and SQL, highlighting which is the chemical engineer profile searched in industries nowadays.

Keywords: Energy Efficiency. Refrigeration. OCP. Data Mining. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BEN	Balanço Energético Nacional
CLP	Controlador Lógico Programável
COP	Coeficiente Operacional de Performance
CRISP-DM	Processo Padrão de Indústria Cruzada para Mineração de Dados
HVAC	Aquecimento, ventilação, ar-condicionado
IA	Inteligência Artificial
IQR	Intervalo Interquartílico
PSE	Engenharia de Processos de Sistemas
R-717	Refrigerante Amônia
SQL	Linguagem de Consulta Estruturada

LISTA DE SÍMBOLOS

η	Rendimento de um Ciclo de Potência
P_i	Pressão de fluido refrigerante
T_i	Temperatura de fluido refrigerante
V_i	Volume de fluido refrigerante
h_i	Entalpia de fluido refrigerante
s_i	Entropia de fluido refrigerante
Q_L	Calor absorvido da fonte quente em um Ciclo de Refrigeração
Q_H	Calor rejeitado para fonte fria em um Ciclo de Refrigeração
Q_{frio}	Carga térmica no evaporador do *Chiller*
Q_{quente}	Carga térmica no condensador do *Chiller*
t_f	Temperatura teórica no evaporador
t_q	Temperatura teórica no condensador
$T_{ent-evap}$	Temperatura de entrada no evaporador do *Chiller*
$T_{ent-cond}$	Temperatura de entrada no condensador do *Chiller*
$T_{saída-evap}$	Temperatura de saída no evaporador do *Chiller*
$T_{saída-cond}$	Temperatura de saída no condensador do *Chiller*
$P_{ent-evap}$	Pressão de entrada no evaporador do *Chiller*
$P_{ent-cond}$	Pressão de entrada no condensador do *Chiller*
$P_{saída-evap}$	Pressão de saída no evaporador do *Chiller*
$P_{saída-cond}$	Pressão de saída no condensador do *Chiller*
ΔP_{evap}	Perda de carga no evaporador
ΔP_{cond}	Perda de carga no condensador
$\dot{m}$	Vazão mássica de água gelada
$\overline{cp}$	Calor específico médio da água
ω	Trabalho de eixo
K	Constante de Henry
x	Fração molar do componente na solução
P	Pressão de vapor do componente na solução
μ	Média da população
σ	Desvio-padrão da população

k	Parâmetro utilizada no Método Normal
Q_1	Quartil inferior (25% dos dados)
Q_3	Quartil superior (75% dos dados)

SUMÁRIO

1 INTRODUÇÃO

Segundo o último Balanço Energético Nacional (BEN), referente ao ano base de 2021, o setor industrial e o de transportes responderam por aproximadamente 65% do consumo de energia no país. Em números absolutos, pensando em energia elétrica, de uma oferta de 679,2 TWh, somente a indústria utilizou 213,3 TWh, o que reitera a constante atenção aos temas sobre o uso racional de recursos, sobretudo, em larga escala (BEN, 2022).

Em termos de processos, a área de utilidades industriais é um ponto chave nesse cenário, visto ser a responsável por fornecer energias – elétrica e térmica – para que sistemas produtivos funcionem.

Termodinamicamente, essas energias são obtidas por meio de ciclos de potência ou de refrigeração. O primeiro, também chamado de Ciclo de *Rankine* (1859), tem por objetivo converter energia térmica em elétrica. Nesse mecanismo, uma fonte de combustível sofre queima, gerando calor suficiente para aquecer e mudar de fase um fluido de trabalho, dentro de uma caldeira. Por sua vez, esse vapor movimenta uma turbina, gerando trabalho de eixo, o qual é convertido à energia elétrica.

De forma oposta, para se refrigerar sistemas, é preciso que trabalho, na forma de energia elétrica, seja fornecido ao ciclo. Nesse caso, fluidos refrigerantes são utilizados a fim de que facilmente mudem de fase, gerando carga térmica suficiente para remover calor da fonte quente, e rejeitá-lo para uma fonte fria. O rendimento de um ciclo de refrigeração é mensurado por seu Coeficiente Operacional de Performance (COP).

Carnot (1824) descreve um ciclo termodinâmico ideal, no qual o rendimento (η) atinge 100%, graças à sua reversibilidade. Contudo, no mundo real, essas etapas são irreversíveis, isto é, há perdas de energia para fora do sistema. Assim, a demanda por eletricidade e altas e/ou baixas temperaturas de processo ditam o custo da conversão energética, sendo imprescindível que essa seja, portanto, a mais eficiente possível, dentro de suas limitações.

Varbanov et al. (2004) preconiza que a eficiência energética industrial pode ser melhorada por meio de: *retrofit* de processos obsoletos, sobretudo de trocadores de calor, que são abastecidos pelas utilidades internas; melhoria dos próprios sistemas de utilidades industriais, como o reaproveitamento energético de correntes de processo; e manutenção das condições básicas a nível operacional de plantas já existentes. Entretanto, isso não é uma tarefa fácil. Cada uma dessas ações requer sólido conhecimento para, por exemplo, estimar reais *savings* econômicos, e justificar possíveis aquisições e mudanças no processo.

A fim de contornar tal desafio, a modelagem matemática tradicional, também conhecida como “caixa branca”, é uma ferramenta que tem se mostrado útil para descrever tais processos, pelo menos desde a década de 1960 (BIRD et al, 1960).

Ahamed et al (2011) revisa o uso de formulações teóricas para obter melhorias em ciclos de refrigeração por compressão de vapor. Balanços de energia e exergia – máxima quantidade de trabalho útil que pode ser obtida em um ciclo – são feitos, a partir de hipóteses simplificadoras. E, embora bastante robustos, tais balanços não são perfeitos, uma vez que não são capazes de capturar todas as particularidades físicas do sistema.

Porém, com o advento da massiva digitalização industrial, e consequentemente, a imensa disponibilidade de dados de processo, modelagens híbridas (“caixa-cinza”) ou orientada a dados (“caixa-preta”) tornaram-se cada vez mais comuns, a fim de contornar as incertezas impostas pela anterior (SANSANA et al, 2021; BRADLEY et al, 2022).

Na literatura recente, trabalhos na área de refrigeração industrial buscaram criar modelos de predição operacionais, mapeando a eficiência do sistema de acordo com a carga térmica demandada (CARDOSO et al, 2020), bem como otimizar indicadores de performance, maximizando o COP (CHANG et al, 2021).

Assim, o presente trabalho busca analisar dados de processo de um sistema industrial de resfriamento de água (*Chiller*) de uma refinaria de óleos vegetais do estado do Ceará, a fim de avaliar a sua eficiência energética, por meio de modelagem orientada a dados (“caixa-preta”). Os dados extraídos são oriundos de sistemas de transmissão de instrumentos de medida em campo. Uma vez transmitidas, as informações são armazenadas em um ou banco de dados em nuvem (PostgreSQL), chamado também de historiador. Para extração, tratamento, visualização e modelagem foi utilizada a linguagem de programação (Python). Por meio desta monografia, pode-se ainda aliar o estudo de um sistema termodinâmico clássico de refrigeração a abordagens modernas de interpretação, mostrando – de fato – os desafios da Engenharia Química, e do seu profissional, para as próximas décadas.

1.1 Objetivos

1.1.1 Objetivo geral

O objetivo geral deste trabalho é extrair informação de um sistema de resfriamento de água industrial (*Chiller*) utilizado em uma refinaria de óleo vegetal cearense, a fim de transformá-lo em conhecimento útil, utilizando mineração de dados.

1.1.2 Objetivos específicos

1. Escolha e definição do problema de negócios (Sistema de água gelada);
2. Listagem das variáveis de processo que serão mineradas;
3. Planejamento da extração dos dados (período de coleta e agrupamento da série temporal);
4. Limpeza geral dos dados (tratamento de dados nulos, ruídos e *outliers*, sobretudo de fundo de escala em instrumentos de medição);
5. Cálculo de indicadores (cargas térmicas, perdas de carga, diferenciação de totalizadores de consumo);
6. Análise exploratória e visualização dos dados (encontrar padrões, tendências e correlações por meio de distribuições e teste de hipóteses);
7. Definição de modelos orientados a dados que permitam classificar regiões de operação do *Chiller* a fim de entender a importância das variáveis de processo em seu Coeficiente Operacional de Performance (COP);
8. Avaliação dos modelos por meio de métricas de desempenho, de acordo com o tipo de classificação proposta.

1.2 Estrutura da monografia

Esta monografia divide-se em sete partes principais: (I) Introdução, contendo a motivação do projeto e informações preliminares; (II) Revisão Bibliográfica, na qual buscou-se na literatura trabalhos relacionados e referências autorais; (III) Descrição do Processo ilustrando e explicando o funcionamento físico do sistema em questão; (IV) Metodologia baseada no CRISP-DM, contendo o desenvolvimento do trabalho de acordo com cada etapa proposta; (V) Resultados e Discussão, trazendo tabelas, gráficos e métricas a fim de argumentar sobre aquilo que foi proposto nas etapas anteriores; (VI) Conclusão, com o encerramento do projeto e reflexões aprendidas; e (VII) com trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

2.1 Ciclos de refrigeração

O conceito de refrigeração artificial é contraintuitivo, uma vez que na natureza, pela segunda lei da termodinâmica, a direcionalidade da energia térmica é fluir do corpo mais quente para o mais frio, até que o equilíbrio seja atingido.

Historicamente, a geração de frio era exclusivamente natural. Em 220 A.C, sistemas subterrâneos rudimentares de armazenamento de alimentos e bebidas já eram utilizados na Ásia. Porém, ao longo dos séculos, a forma mais comum de se gerar frio era com a estocagem de gelo nos meses de inverno. "*Ice Harvesting"* ou colheita de gelo era o processo de extração de grandes blocos de gelo do rio Hudson, em Nova Iorque. Esses eram transportados de navio para diferentes localidades, a fim de serem armazenados em depósitos com isolamento térmico conhecidos como *"Ice Houses"* ou casas de gelo (GANTZ, 2015). É evidente que esse tipo de estratégia não era eficaz do ponto de vista energético, uma vez que os blocos de gelo, invariavelmente, começavam a derreter em algum momento.

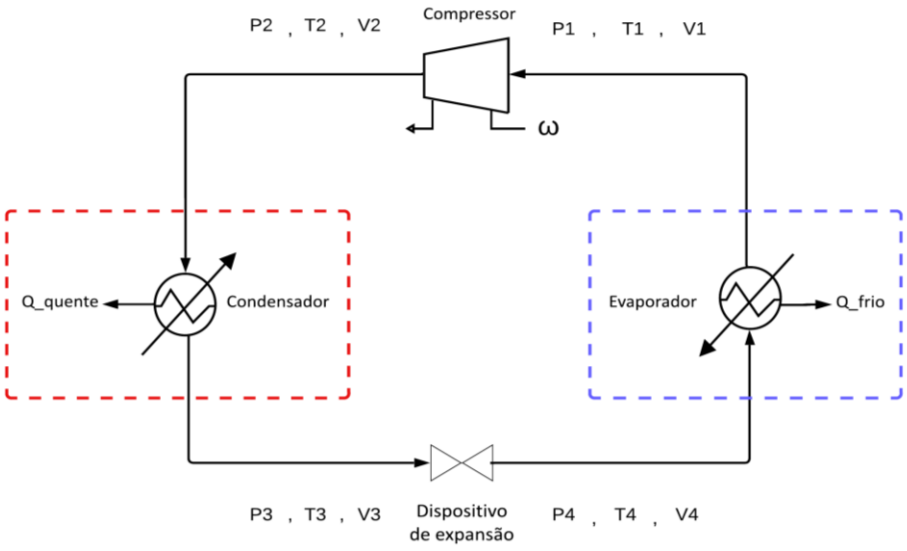
William Cullen (1755) foi o primeiro cientista a demonstrar um sistema de refrigeração industrial propriamente dito, produzindo gelo ao criar vácuo em um recipiente contendo um fluido volátil. Jacob Perkins (1834), contudo, foi o pioneiro a descrever o ciclo completo de refrigeração por compressão de vapor, contendo as quatro etapas básicas (compressão, condensação, expansão e evaporação), gerando a primeira patente do que viria ser o pai dos refrigeradores modernos. A partir daí, James Harrison (1857) criou o primeiro dispositivo de refrigeração para a conservação de alimentos em uma fábrica de cerveja. Finalmente, em 1913 a "Domelre" (acrônimo de *Domestic Electric Refrigerator*) foi criada, sendo a primeira geladeira doméstica. Em 1927, a "Monitor-Top", um modelo posterior da *General Electrics* ficou muito conhecida, sendo que seu compressor se localizava no topo do refrigerador, visto produzir altíssimas quantidades de calor.

Outra aplicação da refrigeração envolve a climatização em edifícios, sejam residenciais, comerciais ou industriais. Conhecidos como sistemas HVAC (*Heating, ventilation and air conditioning*) ou, em português, AVAC (Aquecimento, ventilação e ar-condicionado), eles aquecem os espaços em dias frios, ventilam e renovam o ar no ambiente, retirando odores e impurezas, e refrigeram o local em dias quentes. No comércio e na indústria, esses sistemas são usados também em câmaras frias ou frigoríficas.

Assim, em termos técnicos, o conceito de refrigeração envolve a manutenção de uma temperatura inferior à temperatura da vizinhança. Isso pode ser alcançado com a contínua absorção de calor em um nível baixo de temperatura, usualmente efetuado por meio da evaporação de um fluido refrigerante em um processo contínuo – com escoamento – em estado estacionário. O vapor formado é retornado ao seu estado líquido original para uma nova evaporação por dois caminhos: compressão seguida de condensação, ou absorção por um líquido de baixa volatilidade, a partir do qual ele é posteriormente evaporador a uma pressão superior (SMITH; VAN NESS; ABBOTT, 2007, p. 236).

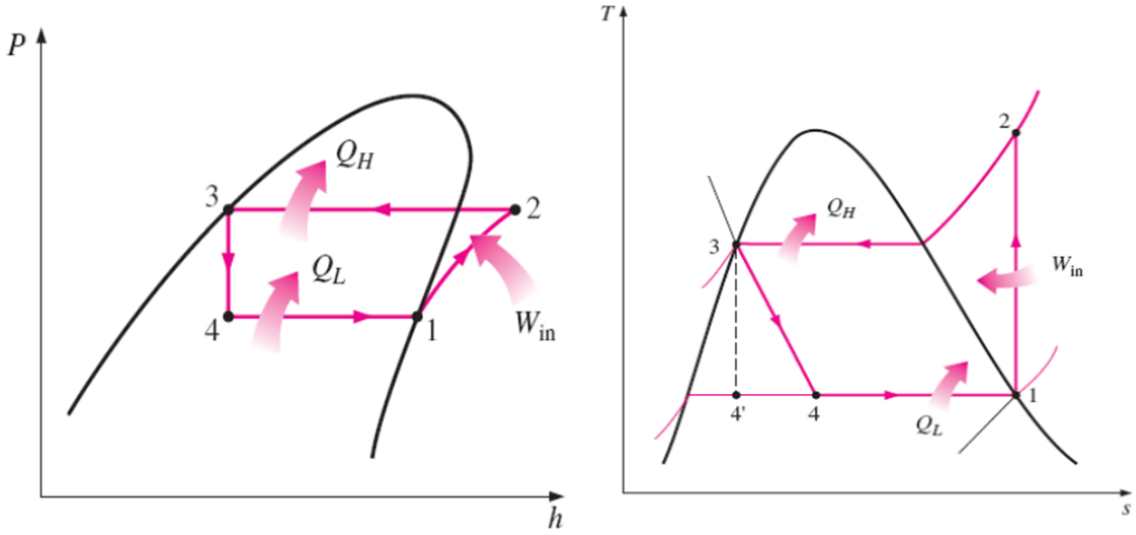
O ciclo mais utilizado é o de compressão de vapor (Figura 1). Nesse, um fluido refrigerante na forma de vapor saturado (P_1, T_1, V_1) adentra a um compressor. Após comprimido, sua pressão e temperatura aumentam de forma que o mesmo passa a um estado de vapor superaquecido (P_2, T_2, V_2). Com sua carga térmica em valor máximo, esse calor é rejeitado à vizinhança em um condensador. Assim, o vapor é liquefeito em uma operação isobárica à líquido saturado (P_3, T_3, V_3). Em geral, esse líquido é submetido à expansão em uma válvula ou sistema de boia. Uma vez expandido, sua pressão e temperatura diminuem chegando ao estado de mistura de líquido e vapor (P_4, T_4, V_4). Nessa fase, tem-se a menor carga térmica do sistema. Finalmente, pela segunda lei da termodinâmica, o líquido "rouba" calor da vizinhança, que está a uma temperatura mais alta. A partir daí, esse se transforma em vapor saturado, encaminhando-se novamente ao compressor. Diagramas de *Mollier* (1923) P-h (pressão x entalpia) e T-s (temperatura x entropia) são mostrados na Figura 2.

Figura 1 – Ciclo básico de refrigeração por compressão



Fonte: Elaborado pela autora.

Figura 2 – Diagrama P-h e T-s

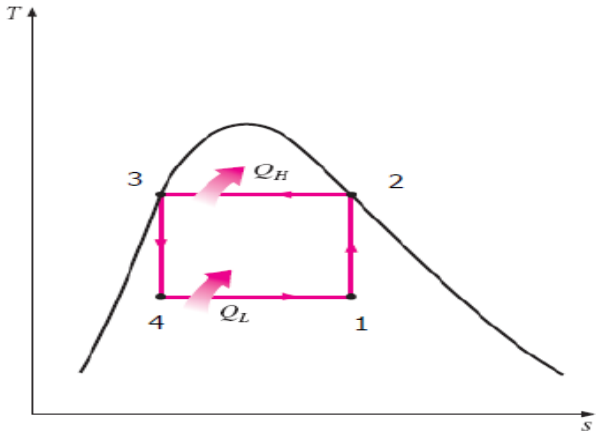


Fonte: Adaptado de Escola Politécnica da Universidade de São Paulo.

Diagramas de *Mollier* permitem a visualização das quatro etapas do ciclo a partir de correlações termodinâmicas. As linhas cheias pretas representam o domo de saturação. O espaço abaixo do domo representa a região de mistura de líquido e vapor. A região à esquerda, o líquido subresfriado, e a região à direita, o vapor superaquecido.

Em um processo de refrigeração ideal de Carnot (Figura 3), o ciclo termodinâmico é executado no domo de saturação do fluido refrigerante. Assim, de (2) para (3), e de (4) para (1), considera-se a condensação e evaporação, respectivamente, isobáricas. Consequentemente, a mudança de fase ocorre à temperatura constante também. Logo, os processos seriam reversíveis. Contudo, isso obrigaria a compressão, (1) para (2), a ocorrer dentro do domo de saturação, o que levaria à compressão de uma mistura líquido-vapor, gerando golpes de líquido no compressor, o que é inviável.

Figura 3 – Ciclo ideal de refrigeração (Carnot)



Fonte: Escola Politécnica da Universidade Federal de São Paulo.

A eficiência do ciclo é mensurada pelo Coeficiente Operacional de Performance (COP). Matematicamente, o COP é definido como o calor absorvido do reservatório frio (efeito de refrigeração ou capacidade frigorífica) pelo trabalho requerido imputado ao sistema (KORETSKY, 2007, p.129).

$$COP = \frac{Q_{frio}}{W} \quad (1)$$

Também, o COP pode ser escrito como a diferença de entalpia do fluido refrigerante no evaporador sobre a diferença de entalpia na etapa de compressão:

$$COP = \frac{h4-h1}{h2-h1} \quad (2)$$

aonde, h_4 e h_1 são as entalpias do fluido antes de entrar e após sair, respectivamente, do evaporador e h_1 e h_2 são as entalpias do fluido antes e após saírem do compressor. Em um refrigerador de Carnot, o COP pode ser calculado como:

$$COP = \frac{Tf}{Tq-Tf} \quad (3)$$

em que T_f é a temperatura atingida no evaporador e T_q no condensador. Nota-se que isso só é possível já que essas operações são a temperaturas constantes no ciclo ideal. Dessa forma, ela fornece o valor máximo possível de eficiência para qualquer refrigerador operando em valores especificados de T_q e T_f. Em sistemas reais bem projetados, o COP situa-se em torno de 2 e 5 (KORETSKY, 2007, p. 130).

O compressor, em geral, é considerado o coração do sistema de refrigeração, uma vez que ele é o componente que recebe a maior carga de energia, na forma de trabalho de eixo (W) através do motor (BELINI, 2019). Em termos de eficiência energética, para um determinado valor de trabalho (kW) inserido no motor, é imprescindível que a capacidade frigorífica (kW) gerada seja ao menos um ($COP = 1$). Do contrário, se estará desperdiçando energia elétrica sem obter resultados térmicos.

Dessa forma, uma das maneiras de se avaliar a eficiência energética, e consequentemente, a saúde de sistemas de refrigeração, é observar o impacto das variáveis de processo em um indicador-chave, usualmente o COP. No compressor, algumas dessas variáveis são a pressão e temperatura de sucção, pressão e temperatura de descarga,

temperatura de superaquecimento, bem como corrente e velocidade do motor do compressor, que podem indicar sobrecarga do mesmo. Ainda, em compressores de deslocamento positivo, a relação de volumes de compressão e descarga (Vi) no bloco de capacidade, bem como a pressão e temperaturas do óleo geram impactos no efeito de refrigeração (BELINI, 2019).

A avaliação da eficiência de processos pode se dar a partir de um direcionamento de modelagem puramente teórica ou baseada em dados obtidos em sistemas reais.

2.2 Modelagem orientada a dados

A indústria e o estudo da engenharia química têm utilizado a modelagem de processos para monitorar, controlar, otimizar e projetar sistemas, pelo menos desde a Terceira Revolução Industrial. Todavia, com a Quarta Revolução Industrial, a massiva captação, digitalização e armazenamento de variáveis de processo, bem como o poder computacional, propiciaram uma nova era para a extração de conhecimento a partir da modelagem orientada a dados (*Data-Driven model*) (SANSANA et al., 2020).

Pelo menos desde a década de 1960, o estudo da Engenharia de Processos de Sistemas, ou em inglês *Process Engineering System* (PSE) tem dominado a área da chamada modelagem caixa-branca (*White-Box*). Em geral, eles buscam predizer o comportamento dos sistemas a partir de modelos mecanicistas ou fenomenológicos. Como característica, utilizam-se de métodos analíticos e/ou numéricos para resolver algum problema muito bem definido e pré determinado. O conhecimento sobre a cinética, fenômenos de transporte (*momentum*, calor e massa), propriedades termodinâmicas e físico-químicas, e propriedades dos materiais é fundamental para descrever minimamente o processo. A resolução se dá pelo uso de ferramentas de balanço de massa e energia aplicados à volumes de controle, a partir de hipóteses simplificadoras e condições de contorno impostas (SANSANA et al., 2020).

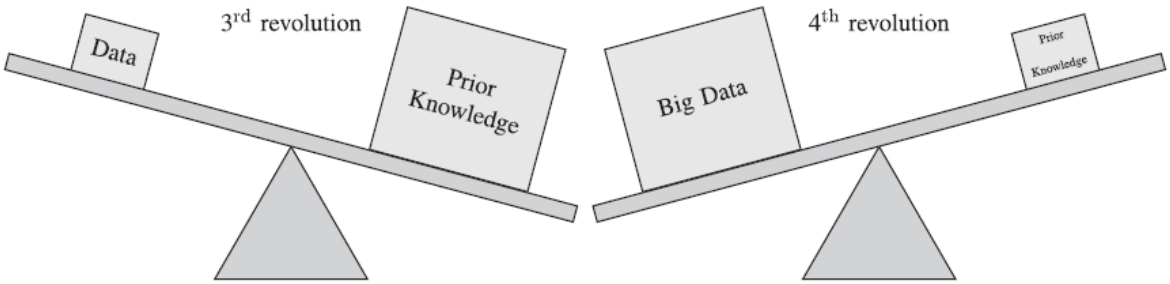
Já a modelagem orientada a dados, também conhecida como caixa-preta ("*Black-Box*"), tem por base modelos estatísticos e empíricos. Assim, o que importa são os dados coletados em si, em termos de quantidade, e principalmente, qualidade. A partir desses, é possível inferir padrões e extrair conhecimento útil baseado nos dados, com o mínimo de suposições possíveis sobre o fenômeno (SANSANA et al., 2020).

Apesar das ferramentas para a extração, limpeza, análise e visualização de dados serem recentes, o início do uso da modelagem orientada a dados no campo da Engenharia Química remonta a 1980 com a Quimiometria (GELADI, 1988). A Quimiometria é uma disciplina que se utiliza da matemática e estatística para projetar e otimizar procedimentos e

experimentos, a fim de reter apenas o conhecimento relevante em uma pesquisa, baseada em dados experimentais (HÉBERGER, 2008).

A Figura 4 ilustra a diferença da modelagem entre a Terceira Revolução Industrial e a Quarta Revolução Industrial, em termos de importância de dados.

Figura 4 – Representação da diferença de peso imposta no tipo de informação em cada modelagem



Fonte: Sansana et al. (2020).

Neste trabalho, o termo *Big-Data*, sem tradução para o português, é utilizado para caracterizar o atual cenário, no qual não apenas há uma massiva quantidade de dados disponíveis, mas também as ferramentas necessárias a isso, como robustos recursos computacionais, e ferramentas de análises de dados. Dessa maneira, tem-se visto surgir novas soluções para problemas antigos, com o uso da ciência de dados. Contudo, há limitações no uso da modelagem orientada a dados. A depender da fonte, e de como esses dados são extraídos, as variabilidades podem ser limitadas, e consequentemente, a análise também. Assim, hoje, além de *Big-data*, outro termo bastante utilizado é *Smart-Data* (sem tradução para o português), isto é, dados com qualidade. É importante salientar que a modelagem orientada a dados tem contribuído não apenas no campo da Engenharia Química, mas também em produtos aplicados ao setor da indústria, comércio e serviços como um todo, englobando o que se caracteriza por Indústria 4.0.

Entre a modelagem caixa-branca e a caixa-preta, existe ainda a modelagem caixa-cinza (*"Grey-Box"*), conhecida como modelagem híbrida. Esse tipo de análise surgiu da dificuldade em se capturar todas as particularidades físicas de um sistema utilizando apenas a caixa-branca, visto que sempre é preciso assumir hipóteses simplificadoras nessa abordagem. Por outro lado, embora a modelagem caixa-preta consiga explicar os fenômenos baseada em estatística, muitos sentem falta da correlação entre ambas as partes. Assim, a modelagem híbrida busca chegar em modelos que trazem a fenomenologia aliada aos dados, a fim de validar hipóteses e responder perguntas. Neste trabalho, o foco será na modelagem caixa-

preta, mais precisamente no Aprendizado de Máquina. Um resumo ilustrativo dos três tipos de modelagem é mostrado na Figura 5.

Figura 5 – Representação da modelagem orientada a dados, caixa-branca e híbrida

Fonte: Elaboração própria.

2.3 Aprendizado de Máquina

Aprendizado de Máquina ou *Machine Learning* é um ramo da inteligência artificial (IA) que utiliza algoritmos similares ao processo de aprendizagem humano. Computacionalmente, um sistema de aprendizagem pode ser definido como um algoritmo que toma decisões baseado em experiências acumuladas através da solução bem sucedida de problemas anteriores. Se o cérebro cria cognição por meio de experiências vividas, a máquina melhora gradualmente sua performance a cada iteração com novos dados.

Historicamente, o termo foi cunhado por Arthur Samuel, em 1959, com sua pesquisa sobre o jogo de damas (SAMUEL, 1959). Trata-se de um dos primeiros trabalhos na área, na qual um algoritmo de aprendizagem baseado em decisão foi utilizado. Em 1962, Robert Nealey, considerado o melhor jogador de damas à época, perdeu a partida para a máquina. O estudo de Arthur é considerado ainda hoje um grande marco para a ciência da computação.

O algoritmo de Aprendizado de Máquina pode ser dividido em três partes: (I) processo de tomada de decisão; (II) função de erro; e (III) processo de otimização do modelo.

O processo de tomada de decisão tem como objetivo predizer ou classificar dados de entrada, produzindo um valor baseado em uma estimativa de padrão dos dados. Já a função erro tem a finalidade de avaliar a tomada de decisão. Em geral, as métricas de avaliação de desempenho dos modelos são baseadas na função erro. A otimização do modelo consiste em

reduzir a discrepância entre o dado real e o ponto estimado. Essa redução baseia-se na atualização de pesos de maneira autônoma até que se atinja um limite de precisão.

Já os métodos de Aprendizado de Máquina podem ser (I) supervisionados; (II) não supervisionados; e (III) por reforço.

O aprendizado supervisionado é definido pelo uso de conjuntos de dados rotulados para treinar algoritmos que classificam dados ou preveem resultados com precisão. Tecnicamente, conforme os dados são alimentados ao modelo, seus pesos são ajustados até a otimização. Os métodos mais utilizados de aprendizado supervisionado são: Redes Neurais, Naive-Bayes, Regressão Linear, Regressão Logística, Árvore de Decisão, Máquinas de Vetores Suporte (SVM), dentre outras.

O aprendizado não supervisionado é definido por agrupar conjuntos de dados não rotulados. Esse método é bastante útil em etapas de análise exploratória de mineração de dados e tratamento de imagens por reconhecer padrões ocultos sem a necessidade de intervenção humana. Em um processo de modelagem, pode ser usado na redução de dimensionalidades, e consequentemente, recursos. Os métodos mais utilizados de aprendizado não supervisionado são: Redes Neurais, *clusterização* com K-Means, PCA (Principal *Component Analysis*), dentre outros.

O aprendizado por reforço difere do aprendizado supervisionado por não ser treinado com dados de amostra e depois de teste. O modelo simplesmente aprende por tentativa e erro. Uma sequência de resultados bem-sucedidos serve de reforço para a máquina aprender.

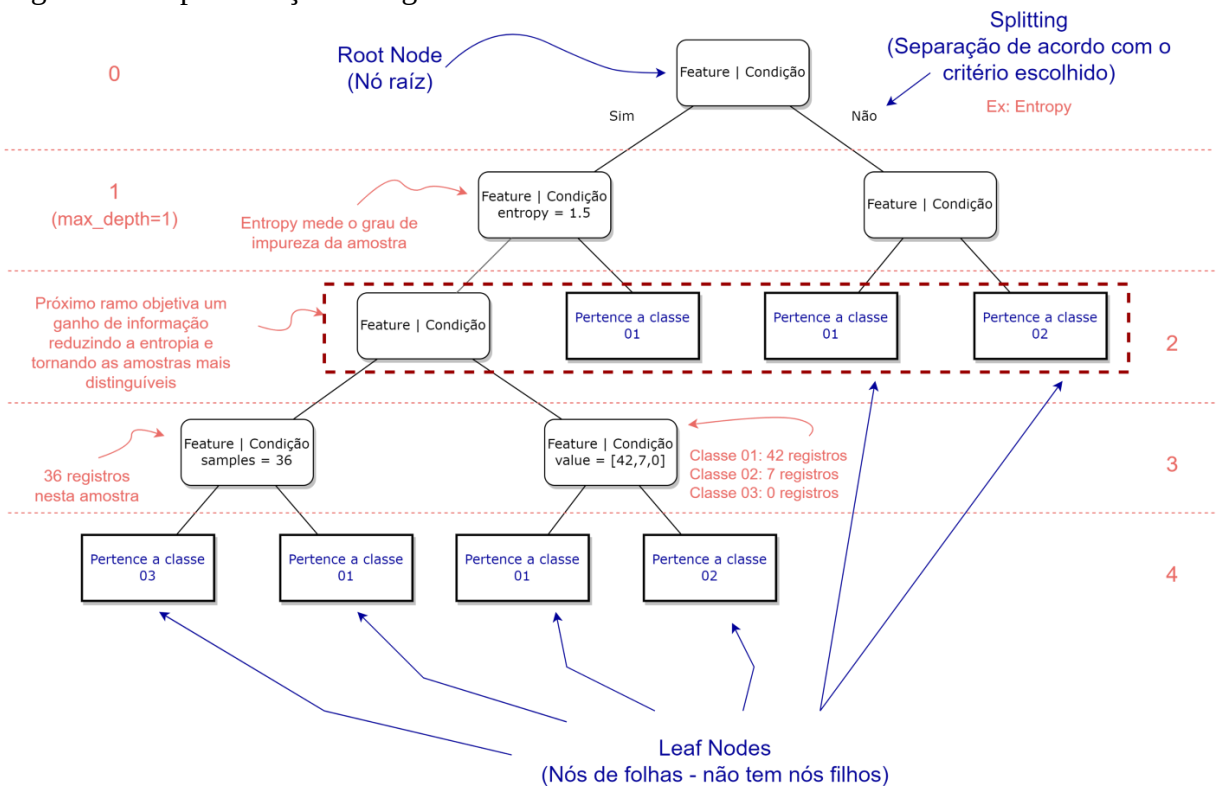
2.4 Modelos de Classificação

Modelos de Classificação são utilizados para designar um rótulo ou classe a novas instâncias que lhe são apresentadas como entrada. Dados históricos rotulados, nos quais a resposta objetivada é previamente conhecida – método supervisionado – são treinados, e posteriormente, quando entrarem em contato com novos dados, cujos rótulos são desconhecidos, os classificarão corretamente.

Árvores de Decisão (*Decision Tree*) (SWAIN; HAUSKA, 1977) são uma das ferramentas mais populares e práticas a serem utilizadas em tarefas de classificação. Sua estrutura é a de um fluxograma, na qual cada nó representa um teste sobre um atributo, cada galho representa um resultado, que pode gerar um nó-filho, e assim sucessivamente, até chegar-se a um nó terminal ou residual, com o rótulo final. Em um nó, tem-se a condição

imposta para uma determinada variável, o número de registros na amostra e uma lista com o número de registros discriminado para cada rótulo ou classe (Figura 6).

Figura 6 – Representação do algoritmo de árvore de decisão



Fonte: Sigmoidal (2022).

A forma como é decidido o critério de divisão de cada nó, e qual variável será testada é melhor explicado na seção 4.5.2 sobre a Otimização de Hiperparâmetros. Porém, em geral, dadas as características matemáticas de como uma árvore é construída, é provável que esta possua maior potencial a sofrer um sobreajuste, e criar um viés de classificação. Assim, pode-se usar conjuntos (*Ensembles*) de modelos.

Ensembles são combinações de diferentes modelos base. As três principais formas de agrupamento são os: (I) Ensacadores (*Bagging*); (II) Impulsionadores (*Boosting*) e (III) Empilhadores (*Stacking*).

Modelos ensacadores funcionam particionando pequenas amostras do conjunto de dados a fim de treiná-las em paralelo. O critério matemático utilizado é o de que essas devem ser semelhantes entre si, a fim de reduzir a variância, já que o resultado entregue por cada micro modelo contribuirá para a média do *bagging*. Em classificações, a classe que mais aparece, vence. Floresta Aleatória (BREIMAN, 2001) é o *ensemble* clássico de uma Árvore de Decisão.

Modelos impulsionadores funcionam particionando também pequenas amostras do conjunto de dados, porém o treino é sequencial. Os erros dos primeiros micromodelos treinados são utilizados para fazer ajustes nos pesos dos próximos micro modelos. Assim, ao invés da variância, é o viés que é reduzido. AdaBoost (FREUND; SCHAPIRE, 1995), GradientBoost e XGBoost (CHEN; GUESTRIN, 2016) são exemplos de impulsionadores.

Modelos empilhadores funcionam a partir da criação de meta-modelos. O conjunto de dados de treino é particionado. Uma parte utiliza um modelo base, isto é, uma Árvore de Decisão única, e a outra parte um modelo mais robusto, como uma Rede Neural. Contudo, os valores preditos pelo modelo fraco também servem de entrada ao modelo robusto. Daí o nome empilhador.

Assim, este trabalho buscou testar um modelo-base (Árvore de Decisão), um ensacador paralelo (Floresta Aleatória) e um impulsionador (XGBoost), utilizando como fonte os dados de um sistema de refrigeração que abastece um processo de refino de óleo vegetal. A descrição do sistema físico encontra-se no próximo capítulo.

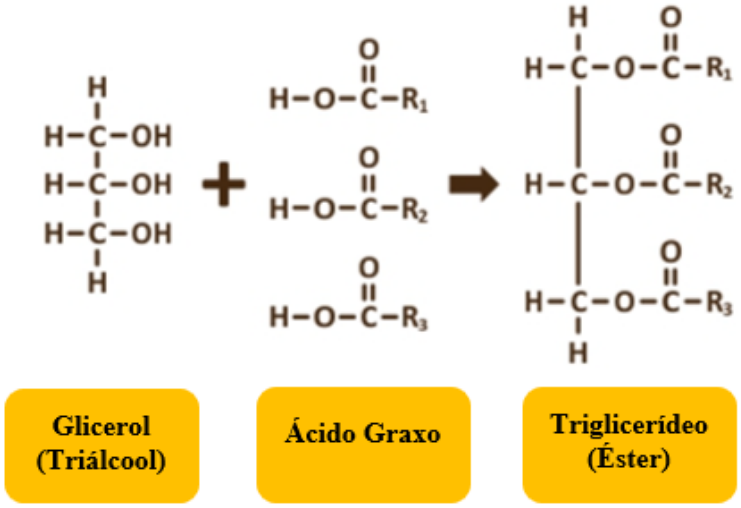
3 DESCRIÇÃO DO PROCESSO

3.1 Refino de óleos vegetais

Refinarias de óleo, seja ele mineral ou vegetal, possuem um objetivo principal: transformar a matéria prima bruta em um produto acabado ou semiacabado, livre de impurezas. Para atingir esse objetivo, um considerável número de operações unitárias é empregado. No caso da indústria de refino de óleo vegetal, pode-se citar quatro grandes blocos: (1) neutralização, (2) branqueamento ou clarificação, (3) hidrogenação e/ou interesterificação e (4) desodorização.

Quimicamente, óleos vegetais pertencem à classe dos lipídeos, sendo essencialmente formados por triglicerídeos de cadeia longa, saturados ou insaturados. Os triglicerídeos ou triacilgliceróis, por sua vez, são formados estequiometricamente a partir de três moléculas de ácido graxo e uma de glicerol, a partir de uma reação de esterificação (Figura 7).

Figura 7 – Formação da estrutura química de óleos vegetais



Fonte: Elaborado pela autora.

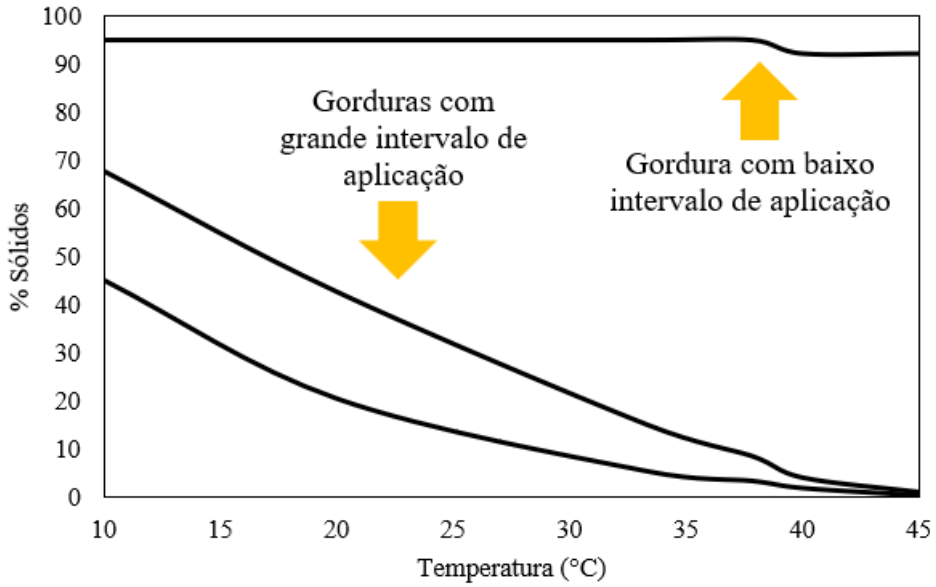
A neutralização consiste de um refino químico a fim de reduzir principalmente o teor de fosfatídeos, umidade e acidez. Em geral, esse processo ocorre no óleo de soja, sendo caracterizado pela acidificação (degomagem ácida) seguida de reação de saponificação (hidrólise alcalina). Para óleos, como o de palma, é feita apenas a degomagem ácida.

O branqueamento é a segunda etapa do processo. Nesse, são removidos os sabões residuais da etapa anterior, traços de metais (complexos ferrosos), óxidos e pigmentos de clorofila, por meio de adsorção com terra clarificante.

Como dito anteriormente, as cadeias de triglicerídeos podem ser saturadas ou insaturadas. Essas ligações simples, duplas ou triplas influenciam no ponto de fusão da substância. Convencionalmente, moléculas cujo intervalo de fusão está abaixo da temperatura ambiente são chamadas de óleos, e aquelas que são sólidas ou semissólidas a temperatura ambiente são denominadas gorduras. Além disso, o grau e tipo de insaturações também influenciam na estabilidade oxidativa do óleo. Em termos sensoriais, essa estabilidade é importante para aplicações alimentícias (margarinas e gorduras especiais) uma vez que o sabor e o odor podem ser revertidos a partir delas.

Graficamente, a classificação desses *blends* é expressa em curvas de sólidos SFC (*Solid Fat Content*), gerando a partir daí, diferentes aplicações alimentícias para cada produto (Figura 8).

Figura 8 – Curva de sólidos de óleos vegetais



Fonte: Elaborado pela autora.

A operação de hidrogenação permite, portanto, alterar as características de fusão dessas moléculas, à medida que o hidrogênio elimina parcialmente as insaturações. Já a interesterificação modifica o comportamento cristalino dos ácidos graxos presentes na matéria prima, aumentando a diversidade de triglicerídeos não presentes no óleo original. Essa modificação pode ser randômica (interesterificação por catálise química) ou direcionada

(interesterificação enzimática). Em resumo, consegue-se enriquecer a formulação de óleo bruto original sem a adição de outros componentes adicionais.

Por fim, a última etapa de refino é a desodorização. Trata-se de uma operação de destilação baseada na diferença de volatilidade. Seu objetivo é remover resíduos de ácido graxo livres, responsáveis pela cor (β-carotenos) e odor (aldeídos, cetonas, álcoois e hidrocarbonetos). O mecanismo de volatilização se baseia na lei de Henry:

$$P = K * x \tag{4}$$

onde,

P: pressão de vapor do componente em questão na solução;

K: constante de Henry;

x: fração molar do componente na solução.

Como os componentes voláteis estão presentes em concentrações muito baixas no óleo, por consequência, suas pressões de vapor na solução também o são. Dessa forma, para que a volatilização ocorra, pode-se utilizar duas opções:

(1) Aumentar a pressão de vapor do componente na fase líquida com o aumento da temperatura do sistema;

(2) Reduzir a pressão de vapor do componente na fase gasosa por meio de injeção direta de algum composto inerte, por exemplo, vapor na coluna. O vapor ocupará o volume gasoso original, diminuindo a disponibilidade de pressão parcial dos ácidos graxos na fase gasosa.

O presente sistema de estudo utiliza a segunda opção.

3.2 Sistema de vácuo

O desodorizador opera com pressões absolutas inferiores a 5 mBar. Para que essas pressões sejam atingidas, um sistema de vácuo fechado via termo compressão é implementado (Figura 9).

O entendimento do sistema de vácuo começa na injeção de vapor motriz no ejetor ou *Booster* (1). Ejetores são equipamentos de compressão de gases. No caso, vapor motriz a alta pressão é direcionado através de um difusor, sendo comprimido pela diminuição da área da seção transversal da tubulação (tubo de Venturi). Como a área é inversamente proporcional a velocidade, esse vapor atinge velocidades supersônicas na tubulação, criando o arraste

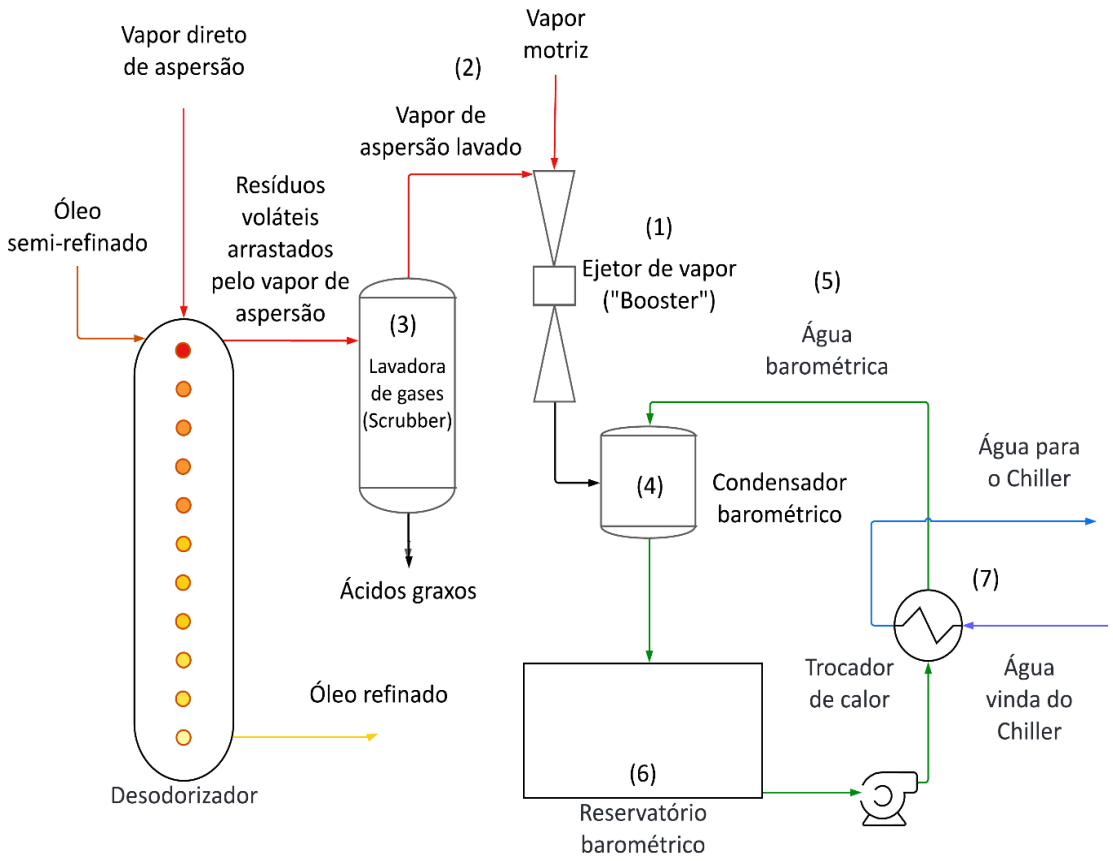
necessário à sucção do vapor de aspersão (2), já separado dos ácidos graxos pela lavadora de gases (3), oriundo da desodorização. Essa etapa de compressão é dita intermediária, não sendo atingido completamente o vácuo demandado.

Para que a etapa seja completa, utiliza-se um condensador de contato direto (4) com água a temperaturas baixas como meio condensante (5). Assim, a mistura de vapores é descarregada nesse condensador que promove a compressão final por mudança de fase. O condensado resultante é direcionado para um reservatório barométrico (6), que continuamente retroalimenta o sistema. A depender, pode-se ter *n* estágios de compressão seguidos de condensação desses.

Finalmente, pode-se entender, portanto, a relação das principais variáveis que influenciam o vácuo em uma torre de desodorização: vapor de aspersão, vapor motriz e taxa de condensação.

A taxa de condensação, por sua vez, é influenciada diretamente pela temperatura da água do reservatório barométrico. Por se tratar de um sistema fechado, a água é enviada em torno de 6°C retornando o condensado à 12°C. Dessa maneira, para que a água se mantenha adequada para o envio à 6°C, ela precisa ser continuamente resfriada no reservatório barométrico. Esse sistema de resfriamento é via trocador de calor (7) com água gelada proveniente de um *Chiller*, que é o principal objeto de estudo deste trabalho.

Figura 9 – Sistema de vácuo do desodorizador

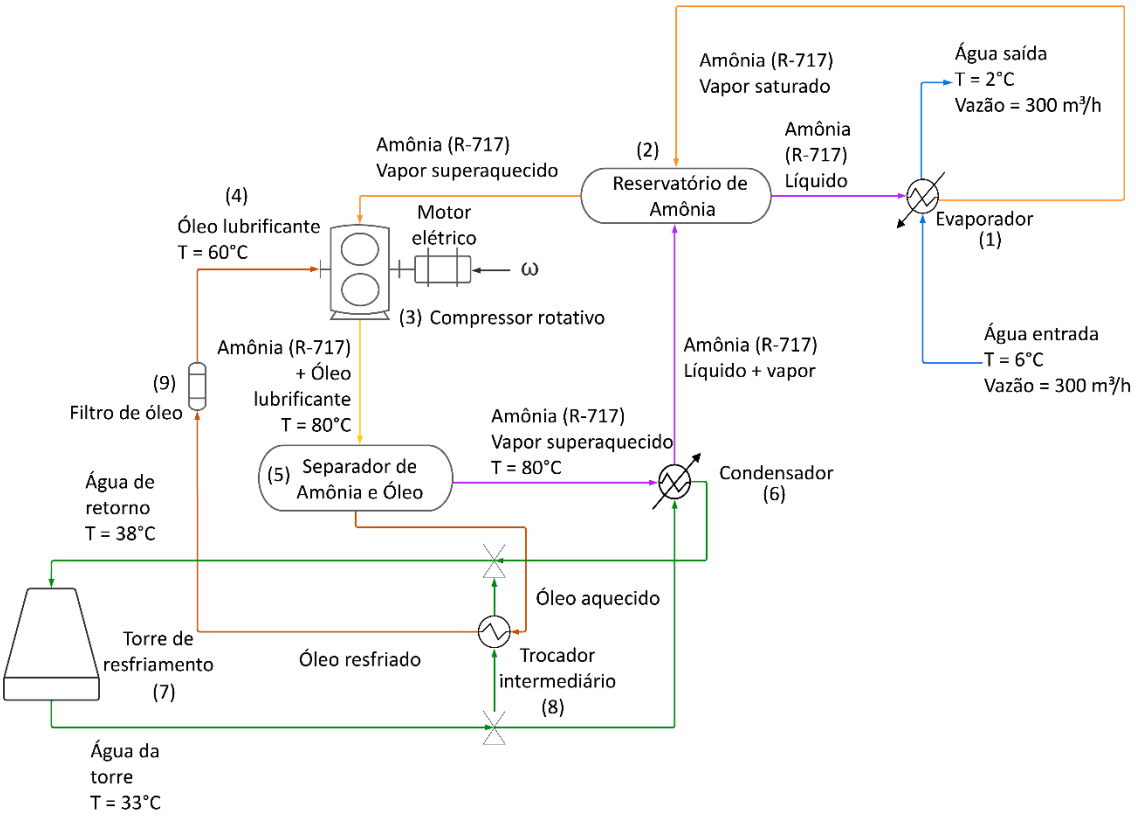


Fonte: Elaborado pela autora.

3.3 Sistema de água gelada (*Chiller*)

O *Chiller* (Figura 10) é o responsável por fornecer água gelada para trocar calor com a água barométrica do sistema de vácuo da desodorização. O sistema utiliza amônia (R-717) como fluido refrigerante, e assim como todo ciclo termodinâmico de refrigeração por compressão de vapor, possui quatro etapas básicas: evaporação, compressão, condensação e expansão.

Figura 10 – Sistema de resfriamento de água (Chiller)



Fonte: Elaborado pela autora.

Amônia (R-717) no estado líquido entra no evaporador (1) a fim de remover calor da água que está retornando do trocador de calor do sistema de vácuo. A água passa, assim, de uma temperatura de entrada de 6°C à uma de saída à 2°C. Quando removido esse calor da água, a amônia passa para o estado de vapor saturado.

A próxima etapa é a de compressão. Como compressores não toleram a sucção de líquidos, há um separador intermediário de líquido-vapor (2) entre as duas etapas, que também serve como um vaso equalizador de pressão. Assim, a fração de vapor entra em um compressor de deslocamento positivo do tipo rotativo (3). Esse compressor funciona a partir da rotação de um par de parafusos (macho e fêmea), no qual para haver a compressão de fato, é preciso que óleo lubrificante (4) entre no sistema.

É bastante comum o uso de óleos minerais lubrificantes em compressores, sejam eles rotativos, alternativos, de membranas ou de turbo compressão, a fim de somente lubrificar as suas partes móveis, sem entrar em contato direto com o gás refrigerante. Contudo, em compressores rotativos de parafuso, o óleo faz parte do elemento de compressão, uma vez que ajuda a “selar” a cavidade na qual o gás refrigerante está adentrando. Como ambos entram em contato, é preciso que na descarga do compressor, amônia seja separada

novamente do óleo lubrificante. Para isso, um separador de amônia e óleo (5) é utilizado, aonde internamente há um filtro coalescente, o qual permite níveis de purificação da ordem de 0,01 μm para a amônia retornar ao ciclo.

Na descarga do separador, tanto amônia quanto o óleo mineral lubrificante estão quentes. Dessa forma, como em um ciclo comum, a amônia é condensada à uma mistura de líquido e vapor em um condensador (6) utilizando para isso água proveniente de uma torre de resfriamento (7). Porém, o óleo lubrificante também precisa ser resfriado, a fim de retornar ao compressor. Assim, em um trocador intermediário (8) o óleo é resfriado com a mesma água da torre. A água sai da torre à 33°C e retorna para a mesma à 38°C. O óleo resfriado passa por um último filtro externo (9), e assim é reciclado de volta para o compressor.

Indo para a última etapa do ciclo, a mistura de amônia líquida com vapor é enviada ao reservatório (2) novamente, onde é expandida por um sistema de boia, permitindo apenas a passagem de líquido para o evaporador, reiniciando o ciclo novamente.

Ademais, pode-se observar que o compressor (3) é o coração do sistema, uma vez que o balanço energético entre o trabalho de eixo (ω), na forma de potência elétrica do motor, e a carga térmica por compressão produzida, concentram-se nele.

4 METODOLOGIA

4.1 CRISP-DM

A metodologia utilizada para avaliar a eficiência energética do sistema de água gelada (*Chiller*), a partir de dados de processo, foi o CRISP-DM (*Cross Industry Standard Process for Data Mining*) ou Processo Padrão de Indústria Cruzada para Exploração de Dados, em tradução para o português (WIRTH; HIPP, 2000).

O CRISP-DM norteia os passos a serem seguidos na análise a partir da estruturação do problema, permitindo consistência, repetitividade e objetividade razoáveis. A Figura 11 ilustra o diagrama padrão:

Figura 11 – Metodologia CRISP-DM

Fonte: Medium (2022).

A primeira parte concentra-se no "*Business Understanding"* ou "Compreensão do Negócio". Essa etapa é primordial, visto que envolve as perguntas que devem ser feitas, e posteriormente respondidas, utilizando os dados e suas ferramentas. Se o conhecimento acerca desse problema de negócio não for claro o suficiente, as perguntas – e muito menos – as respostas, o serão.

A segunda etapa é de "*Data Understanding*" ou "Compreensão dos Dados". Se a compreensão dos negócios é o objetivo, os dados em si são a matéria prima a partir do qual a solução será construída (PROVOST; FAWCETT, 2016, p.28). Essa etapa consiste no entendimento dos dados, sobretudo, para certificar-se de suas limitações, visto que cada atributo (*features*) do processo não tem correlação exata com o problema, isso porque, em geral, dados históricos não são coletados para fins específicos. Dessa forma, é comum a revisitação da primeira etapa a fim de se fazer mais perguntas, explorar mais o processo, incluir ou excluir variáveis, e reestabelecer a forma de extração dos dados. Além disso, a partir dessa etapa, é possível até que se crie a intenção de projetar novas formas de coletar ou incluir variáveis que antes nem mensuradas eram.

Após isso, tem-se a terceira etapa, conhecida como "*Data Preparation*" ou "Preparação dos Dados". Em geral, é a etapa que mais demanda tempo na metodologia. Para que uma modelagem seja feita, é preciso que a informação esteja pronta para uso, isto é, limpa. Assim, os dados que foram coletados precisam ser manipulados e convertidos de sua forma bruta para o formato final desejado. As tarefas mais comuns são: conversão para o formato tabular, renomeação de atributos, arranjo dos tipos de variáveis (tempo, objeto, texto ou booleano), agrupamentos de dados segundo critérios estabelecidos nas etapas anteriores, tratamento de valores nulos ou que não tenham formato específico ("*NaN*", do inglês "*Not a number*", em português, "Não é um número"), remoção de pontos fora da curva ("*outlier's*") e ruídos em geral, fora do padrão. É nessa etapa que as visualizações e estatísticas consolidadas (análise exploratória) são feitas também, a fim de se observar como os dados estão distribuídos, podendo-se detectar padrões e anomalias no conjunto.

Uma vez limpos, chega-se à parte da Modelagem ("*Modelling*"). Um modelo, como discutido na seção 2.2, é uma representação simplificada da realidade, criada para servir a um propósito específico. Seja como "caixa-branca", "caixa-cinza" ou "caixa-preta", cada abordagem tem suas vantagens e limitações, sendo papel de quem irá modelar, selecionar detalhes que são relevantes, e discriminar aqueles que não o são.

A partir daí, pode-se seguir por uma metodologia supervisionada, quando há um alvo específico definido na pergunta inicial, e as sub tarefas de mineração são projetadas pensando-se nesse objetivo; ou não supervisionada, quando não se tem um alvo específico. Usualmente, processos químicos são orientados por supervisão. Os modelos supervisionados podem ser separados em preditivos ou descritivos. Modelos preditivos estimam um valor desconhecido de interesse, que é a variável alvo, podendo ser divididos em modelos de classificação, quando o alvo é um valor categórico, ou regressão, quando é um valor

numérico. Já a modelagem descritiva tem por objetivo principal obter informações sobre o fenômeno, e não estimar um valor.

Em ambas as análises, o principal conjunto de tarefas da modelagem se divide em: pré-processamento dos dados, seleção de atributos, escolha do modelo, otimização de parâmetros e implementação. O pré-processamento pode consistir no fatiamento dos dados (*data binning*), a fim de substituir os valores contidos em um intervalo de tempo por um valor representativo desse todo; padronização, que pode ser a normalização, escalonamento ou a codificação de rótulos (*label encoding*); ou mesmo alguma engenharia de dados manual. A escolha do modelo se dá de acordo com o objetivo inicial e o conhecimento obtido a partir das etapas anteriores. Caso seja um modelo de classificação, como os utilizados neste trabalho, em geral, as classes podem estar desbalanceadas quanto ao tamanho. Como consequência, a distorção em direção às classes mais populares pode transparecer no modelo. Para lidar com isso, pode-se utilizar técnicas de aumento da amostra minoritária (*upsampling)* ou diminuição da majoritária (*downsampling*). Outras formas é a de se buscar algoritmos que minimizem isso, como as Árvores de Decisão e conjuntos (ensembles) como ensacadores (Florestas Aleatórias) e impulsionadores (XGBoost e AdaBoost). O tipo de métrica utilizada na etapa de avaliação também pode ajudar no balanceamento das classes.

Já a seleção de atributos tem como objetivo reduzir as dimensionalidades do problema. À medida que é aumentada a dimensão dos dados, esses tendem a se tornar esparsos, dificultando a obtenção de sinais úteis. Também, o próprio tempo de execução do modelo é função do número de atributos. Além disso, atributos considerados irrelevantes podem causar efeitos negativos na modelagem, assim como aqueles com baixa variabilidade ou hiper correlacionados. Esses deixam os coeficientes de uma regressão ou classificação instáveis ou difíceis de se interpretar (HARRISON, 2019, p.76-88).

Por fim, para o modelo de aprendizado de máquina ser implementado, é preciso que os dados sejam separados em dados de treino e dados de teste. Isso é um passo essencial, visto que, se um modelo for testado com os dados de treino, isto é, os mesmos dados em que fora construído, seu resultado não poderá ser generalizado para o ambiente real, já que não se saberá qual é o seu comportamento em dados nunca antes vistos anteriormente. Em geral, essa separação pode ser feita manual, com a porcentagem de amostra escolhida pelo usuário, ou por validação cruzada (*cross-validation*), uma técnica de particionamento de dados randômica. Na validação cruzada, se um conjunto de dados for particionado em *n* amostras, cada uma será treinada e testada por outra, *n* vezes. Já se for separada manualmente, por

exemplo, em 30% para treino e 70% para teste, essa porcentagem não será alterada. Os 30% de treino sempre serão os mesmos para aqueles 70% de teste.

Na implementação, além da separação da amostra de dados, é importante realizar a otimização de hiperparâmetros (*hyperparameters tuning*). No campo da própria engenharia química e da física, em alguns modelos e/ou correlações, quando o ajuste é feito sobre um conjunto de dados, alguns parâmetros são obtidos, tendo sentido físico ou não. No campo da aprendizagem de máquina, o mesmo ocorre, porém com um número mais complexo de parâmetros. A depender dos valores imputados para esses, é provável que o modelo sofra um sobreajuste (*overfitting*), predizendo de 99% à 100% corretamente os dados de treino, porém com um ajuste ruim dos dados de teste. Na prática, é como dizer que o modelo está enviesado. Assim, a fim de contornar isso, faz-se quantas iterações sejam possíveis de combinações dos hiperparâmetros, até se obter um modelo confiável. Qualitativa e quantitativamente, essa confiança é obtida na próxima etapa, que é a avaliação.

A etapa de "*Evaluation*" ou "Avaliação" tem como objetivo estimar os resultados de mineração de dados de maneira qualitativa e quantitativa, a fim de obter confiança de que aqueles padrões extraídos são válidos, e não apenas coincidência. Em termos de aprendizado de máquina ou aprendizado profundo ("*Deep Learning*"), é nessa etapa que os dados de teste são colocados à prova fora do ambiente controlado de treino, a fim de se observar se o modelo, de fato, capturou as particularidades reais do sistema. Para isso, métricas de avaliação são utilizadas. Em regressões, é comum o uso do coeficiente de determinação (r^2), entre 0 e 1, representando o percentual da variância do alvo com o qual os atributos contribuem, bem como a plotagem dos resíduos. Em classificações, as métricas mais utilizadas são a acurácia, que é a porcentagem de classificações corretas; o *recall* (revocação) ou sensibilidade, que dita a porcentagem de valores positivos classificados corretamente; precisão que é a porcentagem de predições positivas que estavam corretas; e F1, que trata da média harmônica de revocação e precisão. Matrizes de confusão (*Confusion Matrix*) e a importância de atributos (*Feature Importances*) são outras métricas bastante utilizadas na etapa.

Uma observação importante é a de que o rendimento do modelo também é criticamente dependente dos dados que foram inseridos. Logo, caso uma mudança seja feita no sistema físico, aquele modelo pode não atender mais a aquele fenômeno (RAEDER et al., 2012).

Por fim, a última etapa é a de implantação ("*Deployment*"). Nessa, não apenas o modelo, mas as próprias técnicas utilizadas em toda a mineração são implantadas, de forma a construir e testar automaticamente novos dados inseridos. Além disso, a implementação da

mineração em si é até mais preferível, visto que algumas tendências produtivas podem se alterar rapidamente, e assim toda modelagem teria de ser depurada e refeita toda vez, o que é inviável (PROVOST; FAWCETT, 2016, p.33).

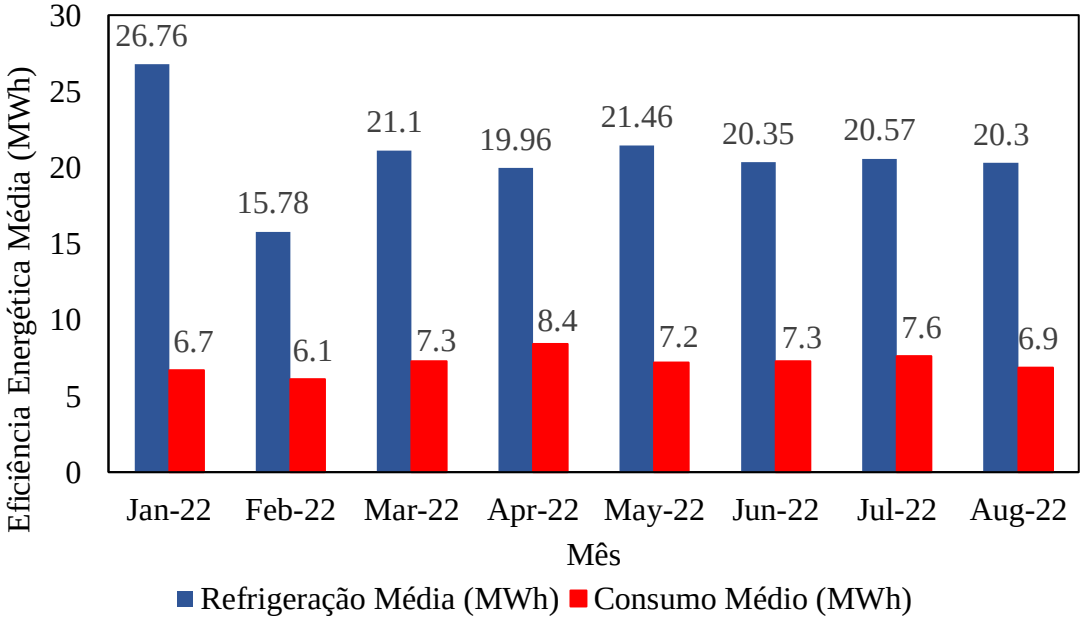
Após a implantação, mesmo bem-sucedida, é comum que se retorne a fase de compreensão do negócio. O processo de mineração de dados produz, por si só, uma grande quantidade de conhecimento sobre o problema que se quer resolver, e suas dificuldades associadas. Assim, o CRISP-DM é uma metodologia caracteristicamente cíclica e iterativa, por isso os atalhos de retorno. A repetição é a regra e não a exceção. Assim, passar por uma etapa sem ter resolvido o problema não é considerado algo negativo, mas sim ganho de conhecimento.

Nesta monografia, a metodologia foi implementada até a fase de avaliação do modelo.

4.2 Compreensão do problema

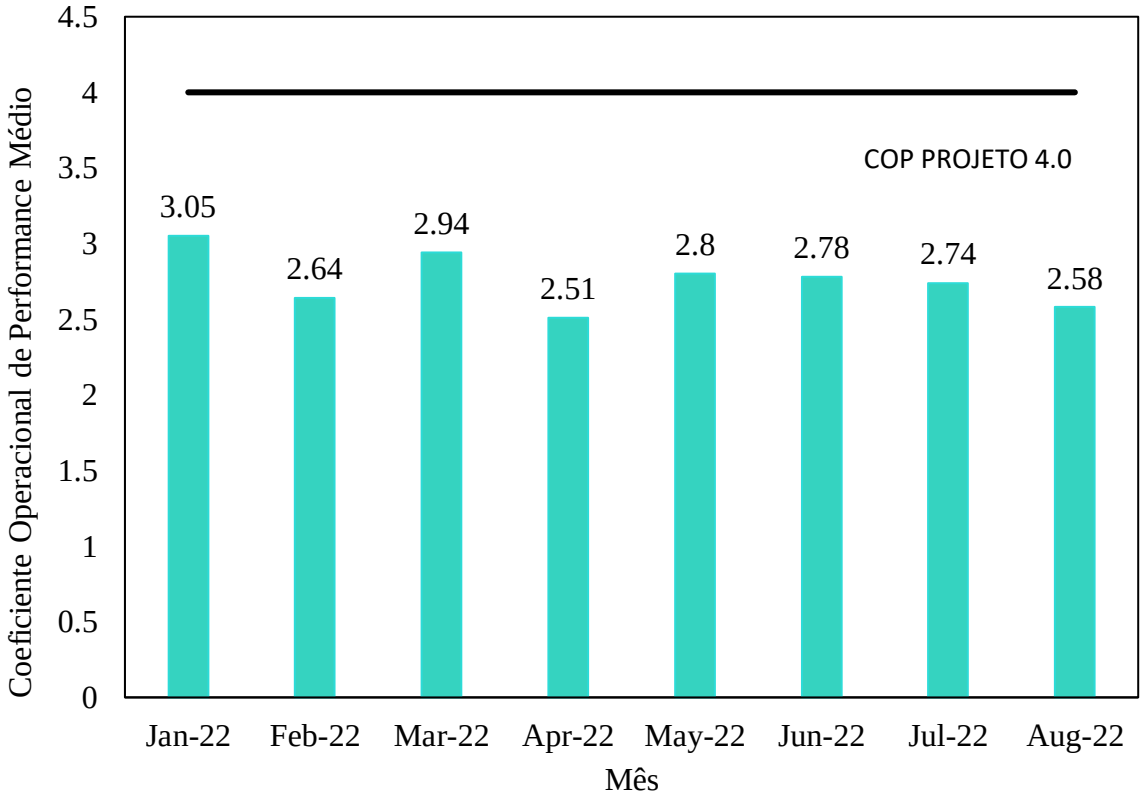
O planejamento da extração dos dados foi feito a partir da análise mensal de indicadores de desempenho do *Chiller* (Figuras 12 e 13).

Figura 12 – Capacidade de refrigeração e consumo energético médio mensal do Chiller



Fonte: Elaborado pela autora a partir de banco de dados privado.

Figura 13 – Coeficiente Operacional de Performance (COP) médio mensal do Chiller



Fonte: Adaptado de banco de dados privado.

Foram analisados o COP, a capacidade de refrigeração e o consumo de energia elétrica do motor do compressor. Observando o primeiro semestre do ano corrente de 2022, escolheu-se extrair o mês com o melhor e o pior desempenho, sendo janeiro e abril, respectivamente. Ademais, a partir de conhecimento de rotinas de problemas da planta, extraiu-se três períodos a mais: quebra do elemento rotativo do compressor do *Chiller* em fevereiro, inserções de amônia no sistema em julho e defeito na boia de expansão em agosto. Pode-se perceber a influência dessas anomalias, por exemplo, pelos valores mais baixos de COP em agosto (2,58) e fevereiro (2,64).

Como se observa também, o coeficiente operacional de performance de projeto é 4.0. No entanto, o sistema tem trabalhado em torno de 70% da sua capacidade. A partir disso, tem-se a motivação inicial para avaliar os ofensores que impedem o *Chiller* de trabalhar sob um desempenho melhor. Também, uma adaptação do *datasheet* de projeto do fabricante permite ganhar noção sobre o quanto as variáveis do sistema podem estar distantes de seus valores originais, e consequentemente, impactando no processo (Tabela 1). Os dados foram alterados do catálogo original do fabricante em um erro percentual de 0.07% para mais ou para menos.

Tabela 1 - Dados nominais de operação do Chiller

DADOS NOMINAIS DE OPERAÇÃO					
MOTOR	Corrente (A)	Velocidade (rpm)	Consumo (kW)	COP	-
	630	3323	343	4.0	
EVAPORADOR	Temperatura de entrada (°C)	Temperatura de saída (°C)	Vazão (m^3/h)	Capacidade Frigorífica (kW)	Perda de Carga (bar)
	5.6	1.9	286.0	1249.0	0.6
CONDENSADOR	Temperatura de entrada (°C)	Temperatura de saída (°C)	Vazão (m^3/h)	Calor rejeitado do sistema (kW)	Perda de Carga (bar)
	31.0	35.0	288	-	0.84

Fonte: Adaptado de banco de dados privado.

4.3 Extração dos dados

Inicialmente, após o estudo do processo (seções 3.2 e 3.3), listou-se 34 variáveis de interesse a serem extraídas (Tabela 2).

Tabela 2 – Lista de atributos (variáveis de processo) extraídas

ATRIBUTO	**UNIDADE**
Capacidade de compressão	%
Corrente elétrica do motor do compressor	A
Pressão de descarga	bar
Pressão de sucção de óleo	bar
Pressão de sucção de amônia	bar
Temperatura de superaquecimento na descarga do compressor	°C
Temperatura de superaquecimento na sucção do compressor	°C
Rendimento do compressor	%
Status do compressor	-
Status do CLP	-
Temperatura de descarga da amônia	°C
Temperatura de óleo na sucção	°C

Temperatura de sucção da amônia	°C
Velocidade do motor	rpm
Volume da câmara de compressão	%
Alarme de fluxo de água no Chiller	Bool
Consumo em kWh do Chiller	kWh
Pressão de entrada no evaporador	bar
Pressão de entrada no condensador	bar
Pressão de saída no evaporador	bar
Pressão de saída no condensador	bar
Temperatura de entrada no evaporador	°C
Temperatura de entrada no condensador	°C
Temperatura de saída no condensador	°C
Temperatura de saída no condensador	°C
Vácuo no topo do desodorizador	mbar
Vácuo no fundo do desodorizador	mbar
Vazão de vapor no booster	kg/h
Pressão de vapor no booster	bar
Vazão de vapor direto	kg/h
Pressão de vapor direto	bar
Temperatura de entrada da água na caixa barométrica	°C
Temperatura de saída da água na caixa barométrica	°C
Consumo em kWh da Torre de Resfriamento	kWh

Fonte: Elaborado pela autora.

Os sinais dessas variáveis são coletados por meio de instrumentos de medida digitais em campo. Por sua vez, esses sinais são enviados à controladores lógico programáveis (CLP's) que fazem a automação, e consequentemente, o controle dos processos. Uma cópia desse sinal é enviada a um servidor remoto, em nuvem (PostgreSQL), o qual armazena esses dados em uma tabela principal, denominada "Evento". As três colunas são: *tag*, valor e tempo. A primeira refere-se a variável de processo em si. Nesta monografia, a palavra *tag* foi substituída por atributo ou mesmo variável de processo. Já o valor é a medida assumida por essa variável em um instante de tempo. E a tabela tempo, por sua vez, registra esse instante de coleta. Assim, uma consulta nesse banco de dados relacional é feita solicitando-se o tempo, ou algum agrupamento deste, a variável de interesse e o valor que ela assume naquele instante.

As consultas são feitas em linguagem SQL, "*Structured Query Language*", em tradução para o português, Linguagem de Consulta Estruturada.

Para a extração dessas, fez-se uma consulta para cada variável nos períodos analisados (janeiro, fevereiro, abril, julho e agosto), utilizando um agrupamento de tempo por segundo. Uma vez coletados, o formato extraído é de um arquivo .csv ("*comma-separated-values*"), em português, valores separados por vírgula, contendo o tempo, a *tag* e o valor (evento).

Separou-se a extração por grupos pré-definidos: variáveis do compressor, variáveis dos trocadores de calor (evaporadores e condensadores) e variáveis do sistema de vácuo de desodorização. O critério foi a memória computacional, visto que apenas um arquivo tornava inviável qualquer tratamento posterior.

4.4 Tratamento dos dados

4.4.1 Tratamento geral

Uma vez extraídos, os dados foram carregados para o Python. Todos os códigos foram escritos no editor VS Code (*Microsoft Visual Studio Code*). Como tratamento geral, passou-se o conjunto de dados em .csv para um formato tabular, renomeando-se as colunas de atributos e utilizando a coluna tempo como índice. Os tipos de dados foram padronizados também (data e hora, contínuas e categóricas). Em especial, os dados provenientes do CLP do compressor vêm multiplicados por um fator de 10. Assim, esses atributos foram divididos por esse fator, e corrigidos.

4.4.2 Tratamento de dados faltantes

O próximo passo foi observar o tamanho do conjunto de dados brutos, e a porcentagem de dados nulos ou sem formato específico. Como regra geral, atributos numéricos em que mais de 90% dos dados eram nulos, foram excluídos. Esse critério foi adotado baseado na redução de dimensionalidades, preconizado no CRISP-DM. Uma coluna de atributos em que apenas 10% dos dados possuem valor, terá pouca, ou quase nenhuma influência, em qualquer tipo de modelagem ou avaliação do processo.

Após a remoção desses atributos, as colunas restantes, em que ainda havia dados faltantes, foram preenchidas a partir do último valor numérico, para cima ou para baixo. Esse

procedimento foi validado pela forma com que a tabela evento originalmente é construída no historiador. Se não houve evento, o valor registrado no instante atual é o mesmo do anterior.

Uma vez preenchidos, os conjuntos de dados foram agregados em uma base única, utilizando o índice (tempo) como o agregador em comum.

4.4.3 Tratamento das variáveis categóricas

Por fim, tratou-se as variáveis categóricas, mapeando seus sinais numéricos para o significado do respectivo status do compressor (Tabela 3):

Tabela 3 – Mapeamento do status do compressor

BIT DO SENSOR	**CORRESPONDÊNCIA**
1	PRONTO
2	EM FUNCIONAMENTO
3	PARTINDO
4	INTERROMPIDO
5	PAUSA
6	PRÉ LUBRIFICAÇÃO
7	BAIXA CAPACIDADE
8	EM SOBRECARGA
9	LIMITE DE DESCARGA
10	LIMITE DE SUCÇÃO
11	PARADO

Fonte: Elaborado pela autora.

4.4.4 Tratamento de outlier's

Como esperado, dados históricos possuem ruídos. Estatisticamente, esses são pontos de coleta que não obedecem à tendência global observada. No caso de dados extraídos de equipamentos industriais em funcionamento, o ruído pôde ser separado em alguns grupos característicos: períodos de intervenções, como manutenções, nas quais o equipamento está parado, falhas ou quebras, oscilações ou quedas de energia elétrica, e fundos de escala dos instrumentos.

Por definição, fundo de escala é o valor máximo que o instrumento de medida pode mensurar sem ser danificado. Nos atributos advindos do CLP do compressor, foram observados, especialmente, valores de fundo de escala bastante característicos, em torno de 6000. Fisicamente, para variáveis como temperatura, pressão, corrente elétrica e volume de compressão (0-100%), a detecção desses fundos de escala foi nítida, uma vez que não têm sentido algum.

Para a retirada desses outlier's, plotou-se a distribuição dos dados, e aplicou-se dois métodos: intervalo entre quartis ou IQR, e o da normal (MONTGOMERY; RUNGER, 2004).

O método IQR segue a seguinte equação:

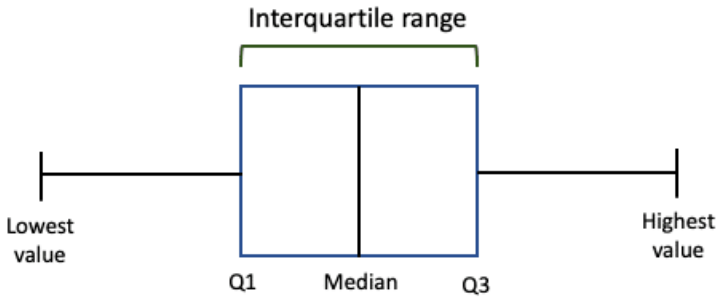
$$IQR = Q_3 - Q_1 \tag{5}$$

$$Limite\ inferior = Q_1 - IQR * 1{,}5 \tag{6}$$

$$Limite\ superior = Q_3 + IQR * 1{,}5 \tag{7}$$

no qual, Q_3 é o quartil superior (correspondente à 75% dos dados) e Q_1 é o quartil inferior (correspondente à 25% dos dados). A seleção de dados sem *outlier* é restrita à apenas o que restar entre os limites, como ilustrado na Figura 14:

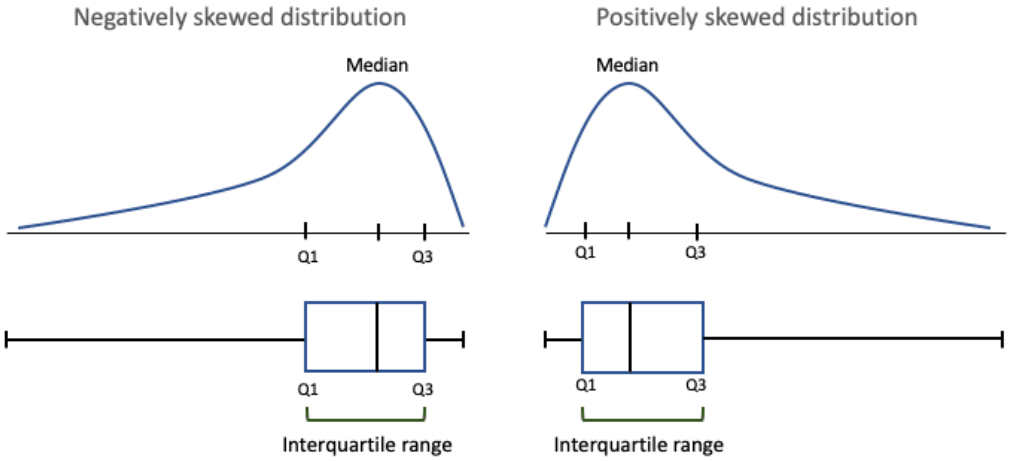
Figura 14 – Intervalo interquartílico em *Box Plot*



Fonte: Scribbr (2022).

A vantagem de utilizar o método IQR é que este se aplica a conjuntos de dados que não seguem uma distribuição normal, e possuem "caldas" de dados fora (Figura 15):

Figura 15 – Intervalo interquartílico aplicado a distribuições não normais



Fonte: Scribbr (2022).

Já o método da normal, segue a seguinte equação:

$$Limite\ inferior = \mu - k * \sigma \quad (8)$$

$$Limite\ superior = \mu + k * \sigma \quad (9)$$

no qual μ é a média, σ é o desvio padrão e k é um parâmetro que, em geral, varia de 1 à 3. Neste trabalho, foi utilizado k = 3 no teste.

Os métodos foram aplicados para cada conjunto de dados distribuídos pelo *status* de funcionamento do compressor. Essa abordagem foi utilizada como uma forma de entender melhor onde estariam localizados esses *outlier's*, e a partir daí excluir tais períodos de forma consciente.

4.4.5 Cálculo de indicadores

Com a base de dados reduzida e limpa, calculou-se os indicadores de performance do sistema (Tabela 4):

Tabela 4 – Atributos calculados

VARIÁVEL CALCULADA	**UNIDADE**
Carga térmica no evaporador	kW
Carga térmica no condensador	kW
Perda de carga no evaporador	bar
Perda de carga no condensador	bar
COP	Adimensional

Fonte: Elaborado pela autora.

A carga térmica produzida no evaporador, também conhecida como capacidade frigorífica do sistema, foi calculada por:

$$Q_{frio} = \dot{m} * \overline{cp} * (T_{ent-evap} - T_{saída-evap}) \quad (10)$$

aonde $\dot{m}$ é a vazão de água gelada, em $\frac{kg}{h}$, $\overline{cp}$ é o calor específico médio no intervalo de temperatura considerado, em $\frac{kJ}{kg\ K}$, $T_{ent-evap}\ (°C)$ é a temperatura de entrada de água no trocador que retorna do sistema de vácuo e $T_{saída-evap}\ (°C)$ é a temperatura de saída de água gelada que será enviada ao sistema de vácuo. Fisicamente, Q_{frio} assume sinal positivo, visto que a água de retorno possui maior temperatura do que a que está sendo conduzida à refinaria.

Já a carga térmica produzida no condensador é o calor que foi removido da amônia após sua saída do compressor, e que será rejeitado ao ambiente na torre de resfriamento:

$$Q_{quente} = \dot{m} * \overline{cp} * (T_{ent-cond} - T_{saída-cond}) \quad (11)$$

em que $\dot{m}$ é a vazão de água da torre de resfriamento, em $\frac{kg}{h}$, $\overline{cp}$ é o calor específico médio no intervalo de temperatura analisado, em $\frac{kJ}{kg\ K}$, $T_{ent-cond}\ (°C)$ é a corrente de água que vem da torre, e $T_{saída-cond}\ (°C)$ é a água com maior energia térmica que retorna à torre para ser resfriada pelo ar. Q_{quente} assume sinal negativo, visto que a temperatura de saída do condensador é maior do que a de entrada.

A vazão de água gelada e de água de resfriamento utilizadas foram as teóricas da Tabela 1, visto que em campo seus instrumentos de medição estavam descalibrados. O valor de calor específico médio utilizado foi de 4,18 kJ/kg K.

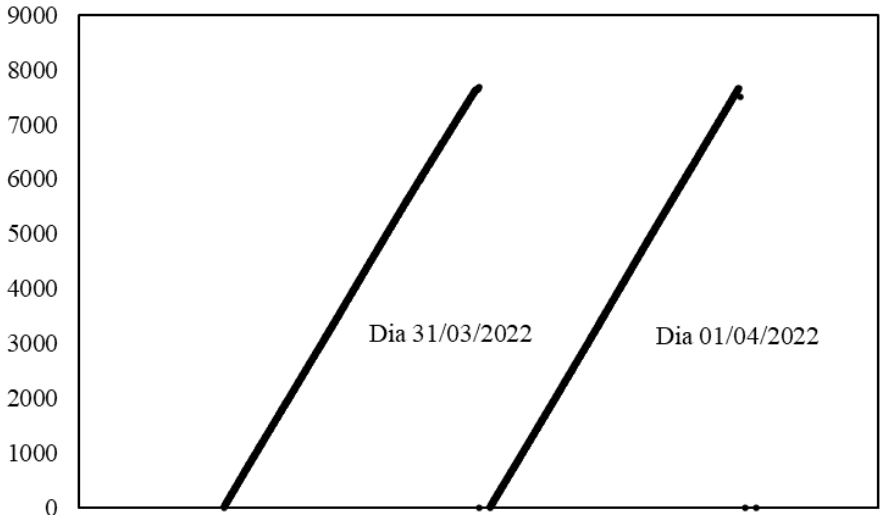
O COP foi calculado sendo a razão entre a capacidade frigorífica e o consumo do motor do compressor do *Chiller*:

$$COP = \frac{Q_{frio}}{Consumo} \quad (12)$$

O atributo de consumo de energia elétrica do motor foi extraído como um totalizador diário, em kWh (Figura 16). Assim, os dados obtidos eram valores crescentes

linearmente ao longo de um período de 24h, caindo logo depois para zero, e reiniciando a contagem. Tratou-se a variável, portanto, fazendo-se a sua derivada, a fim de obter o consumo instantâneo. Contudo, como observado, os incrementos eram constantes (aproximadamente um fator de 8kW por minuto), logo, a variabilidade era nula. Também, isso foi constatado na etapa da seção 4.4.2, na qual mais de 95% dos dados eram NaN. Dessa forma, percebeu-se que o COP variaria apenas pelo seu numerador, isto é, pela capacidade frigorífica.

Figura 16 – Totalizador de consumo do motor em kWh por dia



Fonte: Elaborado pela autora.

Por fim, as perdas de carga nos trocadores de calor (evaporador e condensador) foram calculadas, respectivamente, por:

$$\Delta P_{evap} = P_{ent-evap} - P_{saída-evap} \quad (13)$$

$$\Delta P_{cond} = P_{ent-cond} - P_{saída-cond} \quad (14)$$

nos quais $P_{ent-evap}$ e $P_{ent-cond}$ são as pressões de entrada no evaporador e condensador, respectivamente, e $P_{saída-evap}$ e $P_{ent-cond}$ são as de saída.

Uma vez calculados, foram excluídos os atributos anteriores dos quais estes foram originados, a fim de evitar, na etapa de modelagem, efeitos de colinearidade. Isso porque as cargas térmicas e perdas de carga teriam correlações em torno de 100% com os atributos dos quais foram originados, como as temperaturas e pressões.

4.4.6 Análise exploratória dos dados

Nessa etapa, estatísticas consolidadas foram feitas (contagem, média, moda, mediana, mínimos, máximos, quartis e desvios-padrão), a fim de se observar como os dados estavam distribuídos. Visualização das séries temporais e testes de hipótese pela média entre períodos foram considerados também para observar tendências de comportamento e responder questionamentos prévios. Os resultados encontram-se nas seções *5.1.5* e *5.1.6*.

4.5 Modelagem

Para a modelagem do problema, escolheu-se testar três modelos de classificação: Árvore de Decisão, Floresta Aleatória e XGBoost. Esse tipo de modelagem foi escolhido, uma vez que um dos objetivos do trabalho era entender a influência, e importância, das variáveis de processo do *Chiller* em relação à variável alvo (*target*), que é o COP.

4.5.1 Pré processamento dos dados

Fez-se a segmentação e a rotulagem dos dados do COP em duas regiões de operação: “ruim” e “regular”. O intervalo “ruim” varia de 1,0 à 2,5 e o “regular” de 2,5 à 4,0. Assim, tratou-se o problema como uma classificação binária. A classe de interesse foi a “Ruim”.

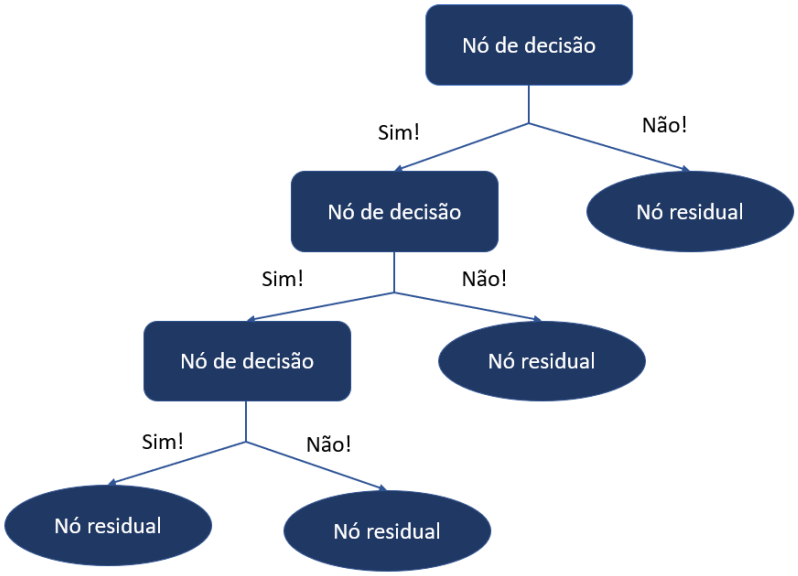
Foi inviável uma modelagem com mais classes, uma vez que se fosse considerada uma possível classe “ótima”, com dados de COP acima de 3,0, a classe “ruim” entre 1,0 e 2,0 e a “regular” entre 2,0 e 3,0, a distorção devido ao desbalanceamento de classes seria muito alta. Em porcentagem, teríamos 0,2% dos dados na classe “ruim”, 3% na classe “boa” e 97% na “regular”. Com a configuração binária, obteve-se 85% dos dados na classe “regular” e 15% na “ruim”. Os resultados encontram-se na seção *5.2.1*.

4.5.2 Otimização de hiperparâmetros

Para compreender os valores imputados nos hiperparâmetros, é importante revisar os principais conceitos dentro do modelo de classificação envolvendo árvores de decisão (Figura 17). Um algoritmo de árvore de decisão deve ser lido como uma árvore invertida. A raiz é o primeiro nó de decisão, o qual contém os dados a serem treinados. A partir daí, esses dados são divididos em ramificações distintas, seguindo uma estrutura de “se” / “se não” (*if/else loop*). O critério para haver mais uma divisão, ou não, é função de parâmetros de

divisão (*splitting parameters*), que por sua vez, baseiam-se em índices probabilísticos de ganho, como será discutido a seguir.

Figura 17 – Algoritmo de Árvore de Decisão



Fonte: Elaborado pela autora.

Para decidir qual atributo será usado para dividir o nó, e em qual ponto da coluna deverá haver a divisão, são usadas funções de perda, também chamadas de funções de custo. O objetivo dessas funções é o de minimizar a distância entre o valor real e o valor predito. Em classificações, a probabilidade da classe predita é usada, sendo o Índice de Gini ou a Entropia esses indicadores. Quanto menor o Índice de Gini ou Entropia, menor o custo, e melhor a divisão do nó.

Já a Entropia mede a aleatoriedade ou desordem de um sistema. Em termos de dados, a aleatoriedade pode ser definida como a informação que está sendo processada. Quanto maior a entropia, mais difícil de processar a informação. O Ganho de Informação (IG) mede a quantidade de informação fornecida por um determinado atributo em relação à classe alvo. Conforme a árvore é construída, o objetivo é encontrar o atributo que possui o maior ganho de informação, e por consequência a menor entropia.

Assim, a qualidade da divisão de um nó é medida por esses parâmetros. Se os atributos da classificação são categóricos, o Ganho de Informação é o critério. Já se forem numéricos, o Índice de Gini é utilizado. Neste trabalho, o critério utilizado foi o Índice de Gini nos três modelos.

Quando o algoritmo é percorrido livremente a primeira vez, é comum que essas divisões dos nós tenham a melhor qualidade possível. Ou seja, para cada pergunta no nó de

decisão, o Índice de Gini gerado será tão pequeno ao ponto de cada atributo ser testado e dividido em amostras tão difusas, que a árvore se torna extremamente extensa e treinada naquele conjunto. Assim, para evitar isso, é preciso calibrar hiperparâmetros.

Hiperparâmetros são argumentos que não se modificam ao longo do processo de treino, mas podem ser modificados pelo usuário no início, a fim de reduzir o sobreajuste dos dados, e permitir, portanto, uma melhor generalização para quando dados novos forem inseridos. Já parâmetros são iteráveis. Exemplos de hiperparâmetros que foram calibrados neste trabalho, e seus respectivos significados, estão sintetizados na tabela abaixo (Tabela 5):

Tabela 5 – Hiperparâmetros otimizados em Árvores de Decisão

HIPERPARÂMETRO	**NOMEAÇÃO TÉCNICA**	**SIGNIFICADO**	**PADRÃO**
Máxima profundidade	max_depth	Profundidade máxima da árvore. Caso não seja limitada, a árvore se expande até o último nó residual conter um único valor. Ou seja, ela treina todas as amostras, gerando o sobreajuste.	None
Máximos nós residuais	max_leaf_nodes	É o número de nós residuais da árvore. Também uma forma de controlar a complexidade do modelo. Caso não seja limitada, ela será controlada apenas pelo Índice de Gini.	None
Mínimo de amostras no nó residual	min_samples_leaf	Número mínimo de amostras necessário para restar em um nó residual.	1

Mínimo de amostras para o nó ser dividido	min_samples_split	Número mínimo de amostras para um nó ser dividido. Como padrão, a árvore tenta dividir o nó até restar apenas uma linha. Isso pode gerar sobreajuste.	1
Decréscimo mínimo de impureza	min_impurity_decrease	Usado para limitar a divisão de um nó residual baseado na Impureza de Gini. Um nó somente será dividido se o Índice de Gini for reduzido a um valor maior ou igual a esse hiperparâmetro.	0
Peso das classes	class_weight	Peso associado a cada classe. No caso de classes desbalanceadas, o ideal é utilizar '*balanced*'. Caso seja usado o 'None', é suposto que as classes têm o mesmo peso.	None

Fonte: Elaborado pela autora.

Quando se trabalha com a otimização desses hiperparâmetros, a avaliação da classificação na árvore tende a ser mais realista. Contudo, uma forma de calibração mais robusta é otimizá-los ao mesmo tempo em mais de uma árvore. Daí o uso de uma floresta aleatória. Na floresta aleatória, os hiperparâmetros mais comuns otimizados são mostrados na tabela abaixo (Tabela 6):

Tabela 6 – Hiperparâmetros otimizados em Floresta Aleatória

HIPERPARÂMETRO	**NOMEAÇÃO TÉCNICA**	**SIGNIFICADO**	**PADRÃO**
Número de estimadores	n_estimators	Trata-se do número de árvores classificadoras no modelo. Quanto maior o número de árvores, melhor. Porém, maior tempo computacional será utilizado. Também, há um número crítico de árvores no qual o aumento não fará diferença no resultado.	100
Número de processadores	n_jobs	Delimita o número de processadores utilizados na máquina para construir a floresta. Caso seja -1, ele utiliza todos os processadores de forma paralela.	1
Inicialização	bootstrap	Delimita o número de dados a ser usado para construir cada árvore. Em caso de 'False', todo o conjunto de dados é utilizado.	True

Fonte: Elaborado pela autora.

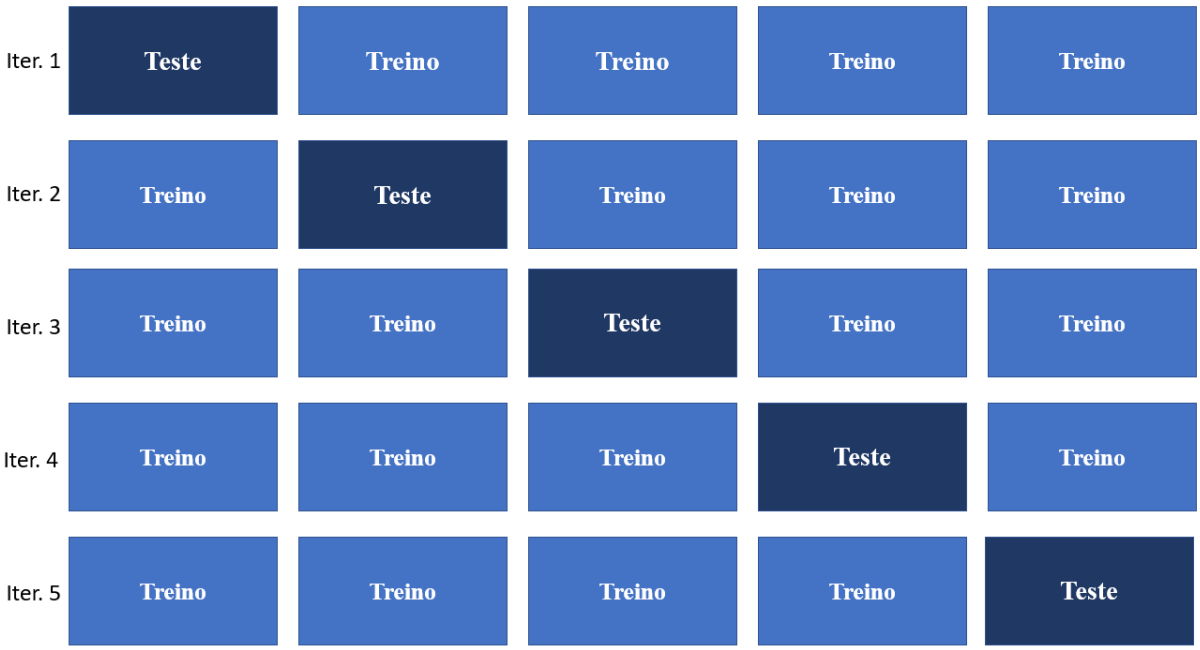
Como pode ser visto, há uma quantidade significativa de hiperparâmetros para cada modelo, podendo ser gerada uma infinidade de combinações, e uma tarefa inviável caso a calibração seja por tentativa e erro do usuário. Assim, há ferramentas (classes da biblioteca *sickit-learn*) que ajudam na calibração: o GridsearchCV e o RandomizedsearchCV.

No GridsearchCV, para cada hiperparâmetro, coloca-se os possíveis valores a serem testados em uma lista. Dessa maneira, percorre-se cada combinação no espaço até se obter o conjunto com as melhores. Já no RandomizedsearchCV, utiliza-se um intervalo. Por exemplo, se no GridsearchCV, o número de árvores (n_estimadores) testado foi [50, 100, 200], e o melhor valor foi 100, nada é garantido que 99 ou 101 não seriam melhores. Assim, ao invés de três valores, o método percorre um intervalo no espaço entre dois valores: [50:100]. A desvantagem nesse caso é de tempo computacional, já que há uma infinidade de

pontos a serem percorridos nesse espaço, e testados paralelamente entre os outros hiperparâmetros. Neste trabalho, o uso do GridsearchCV já reproduziu bons resultados.

Ademais, usou-se validação cruzada nos três modelos, com CV=5, a fim de se obter modelos mais generalistas. Ou seja, o conjunto de dados foi particionado em 5 partes iguais (20%). Para cada iteração, um conjunto foi usado para teste, enquanto os outros quatro foram usados para treino (Figura 18). Os resultados encontram-se nas seções *5.2.2*, *5.2.3* e *5.2.4*.

Figura 18 – Algoritmo de Validação Cruzada



Fonte: Elaborado pela autora.

4.5.3 Avaliação do modelo

Após a implementação de cada modelo, é preciso avaliá-los. Por tratar-se de uma classificação binária (COP "Regular" e COP "Ruim"), pôde-se comparar a performance de um modelo a outro por uma Matriz de Confusão.

A Matriz de Confusão é uma tabela 2x2 que ilustra as frequências de classificação para cada classe do modelo. Para isso, é importante entender o conceito de verdadeiro positivo (*True Positive ou tp*), falso positivo (*False Positive ou fp*), verdadeiro negativo (*True Negative ou tn*), e falso negativo (*False Negative ou fn*).

Os verdadeiros positivos ocorrem quando, no conjunto de teste, a classe que se está buscando prever foi prevista corretamente. Já os falsos positivos ocorrem quando a classe buscada é classificada de maneira errada. De forma contrária, os falsos verdadeiros ocorrem

quando a classe que **não** se está buscando prever, é prevista corretamente. E, por fim, os falsos negativos, quando a classe que **não** se está buscando prever é prevista incorretamente. Em suma, as classificações corretas, isto é, verdadeiros positivos e verdadeiros negativos são as que, de fato, interessam na avaliação. Além disso, é possível gerar relatórios de classificação, com a pontuação de cada métrica. Abaixo, tem-se o cálculo de cada uma delas. Como dito na seção 4.5.1, as classes do COP estão desbalanceadas. Nesses casos, a métrica indicada é a precisão e a revocação (HARRISON, 2019, p.97).

$$Acurácia = \frac{tp+tn}{tp+tn+fp+fn} \quad (15)$$

$$Revocação = \frac{tp}{tp+fn} \quad (16)$$

$$Precisão = \frac{tp}{tp+fp} \quad (17)$$

$$F1 = \frac{2*Precisão* Revocação}{Precisão+Revocação} \quad (18)$$

Para concluir a metodologia, e consequentemente, o desenvolvimento do ciclo, é preciso entender como a importância dos atributos foi calculada no modelo. Para cada nó residual, tem-se a probabilidade característica deste, que é o número de amostras restantes no nó dividido pelo número de amostras totais. Quanto maior esse valor, mais importante o atributo. Esses resultados encontram-se também nas seções *5.2.2*, *5.2.3* e *5.2.4*.

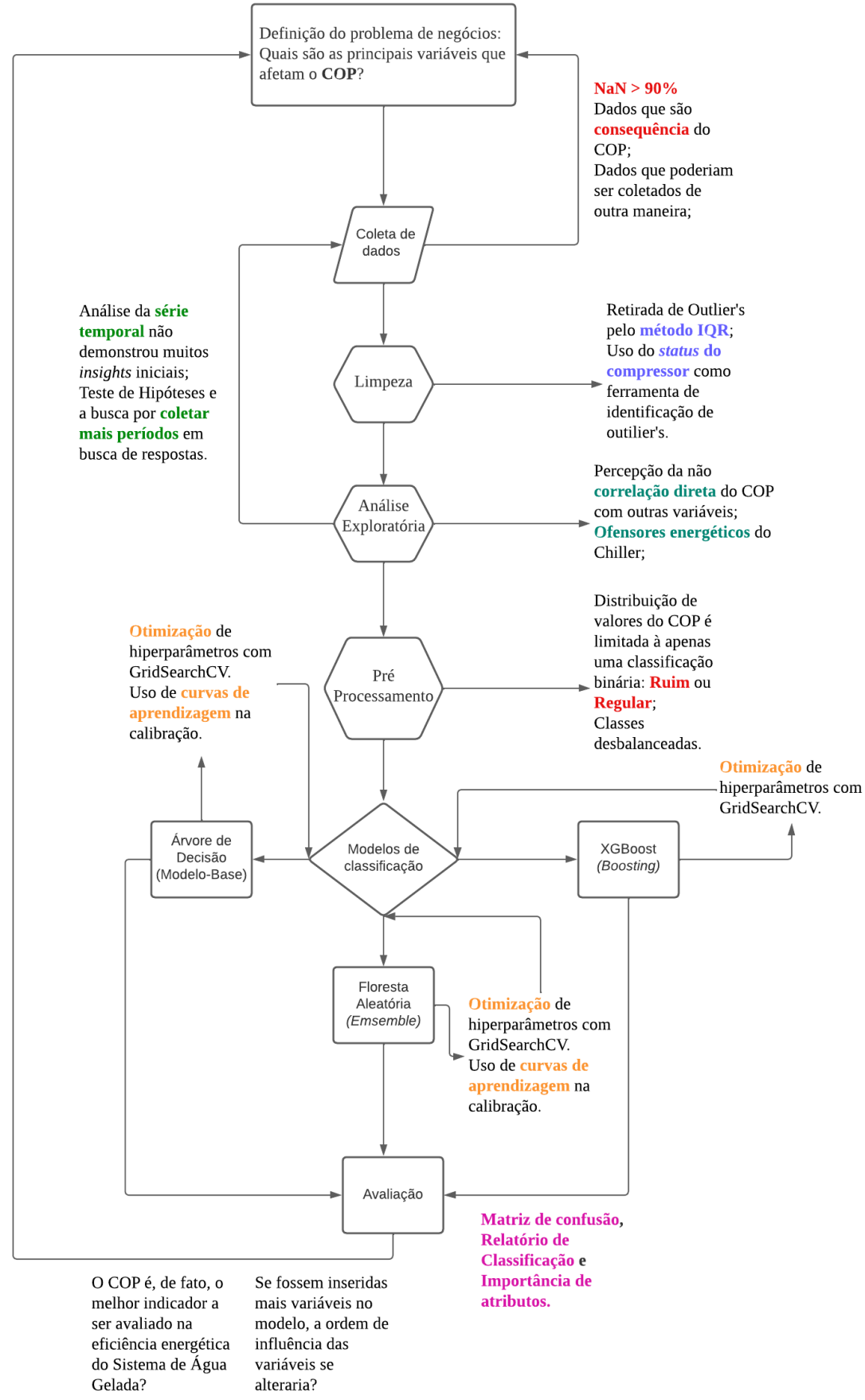
Por fim, uma tabela contendo as bibliotecas do Python utilizadas no desenvolvimento deste trabalho pode ser encontrada abaixo. Também, um fluxograma sumarizando as etapas seguidas na metodologia se encontra na Figura 19.

Tabela 7 – Bibliotecas utilizadas

BIBLIOTECA	**USABILIDADE**	**DOCUMENTAÇÃO OFICIAL**
pandas	Análise de dados	https://pandas.pydata.org/docs/
numpy	Manipulação de vetores	https://numpy.org/doc/
matplotlib	Visualização de dados na forma de gráficos	https://matplotlib.org/stable/index.html
seaborn	Visualização de estatísticas na forma de gráficos	https://seaborn.pydata.org/index.html
scikit-learn	Implementar algoritmos de aprendizado de máquina	https://scikit-learn.org/stable/#
yellowbrick	Implementar e visualizar métricas de avaliação de modelos	https://www.scikit-yb.org/en/latest/#

Fonte: Elaborado pela autora.

Figura 19 – Fluxograma de Mineração de Dados para o Sistema de Água Gelada



Fonte: Elaborado pela autora.

5 RESULTADOS E DISCUSSÃO

5.1 Tratamento dos dados

5.1.1 Seleção de atributos

Para cada conjunto de dados extraído (períodos de janeiro, fevereiro, abril, julho e agosto), excluiu-se os atributos com mais de 90% de variáveis nulas (ver seção 4.3). De um total de 34, apenas 14, excluindo-se a variável "Status do compressor", que é categórica, foram selecionadas para a análise.

Embora apresentassem variabilidade superior à 90%, as variáveis do sistema de vácuo da refinaria foram descartadas da análise. Isso porque as temperaturas da água da caixa barométrica são uma consequência da condição de troca térmica com a água do *Chiller*, mas também dependem fortemente da recirculação da bomba do tanque, do trocador de calor intermediário, do funcionamento adequado dos condensadores barométricos e do abastecimento de vapor direto e do ejetor. Logo, saem do escopo de análise do sistema de controle do *Chiller*. O mesmo raciocínio vale para as próprias vazões e pressões de vapor. Foi mantido, excepcionalmente, apenas o vácuo para análises de correlação, porém, da mesma forma, o mesmo não foi utilizado no modelo, por ser uma consequência, e não causa do COP.

A Tabela 8 descreve as variáveis selecionadas:

Tabela 8 – Atributos com variabilidade maior que 90%

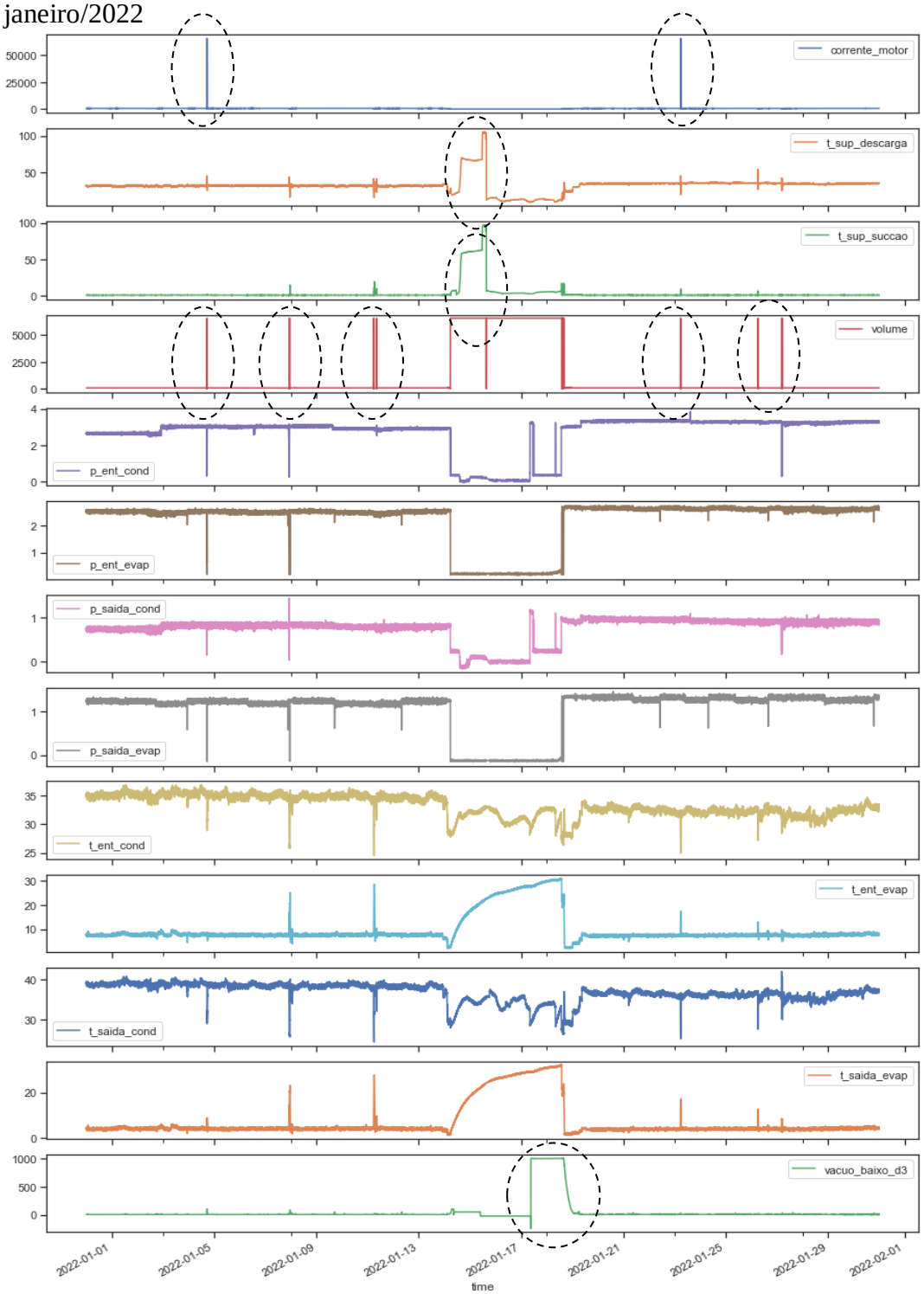
ATRIBUTO	**UNIDADE**
Corrente elétrica do motor do compressor	A
Temperatura de superaquecimento na descarga do compressor	°C
Temperatura de superaquecimento na sucção do compressor	°C
Status do compressor	-
Volume da câmara de compressão	%
Pressão de entrada no evaporador	bar
Pressão de entrada no condensador	bar
Pressão de saída no evaporador	bar
Pressão de saída no condensador	bar
Temperatura de entrada no evaporador	°C
Temperatura de entrada no condensador	°C
Temperatura de saída no condensador	°C
Temperatura de saída no condensador	°C
Vácuo no fundo do desodorizador	mbar

Fonte: Elaborado pela autora.

5.1.2 Análise da série temporal bruta

A Figura 20 mostra a visualização das variáveis brutas ao longo do tempo no período de janeiro. A visualização dos demais períodos encontra-se no APÊNDICE A.

Figura 20 – Comportamento das variáveis de processo (brutas) do *Chiller* em janeiro/2022



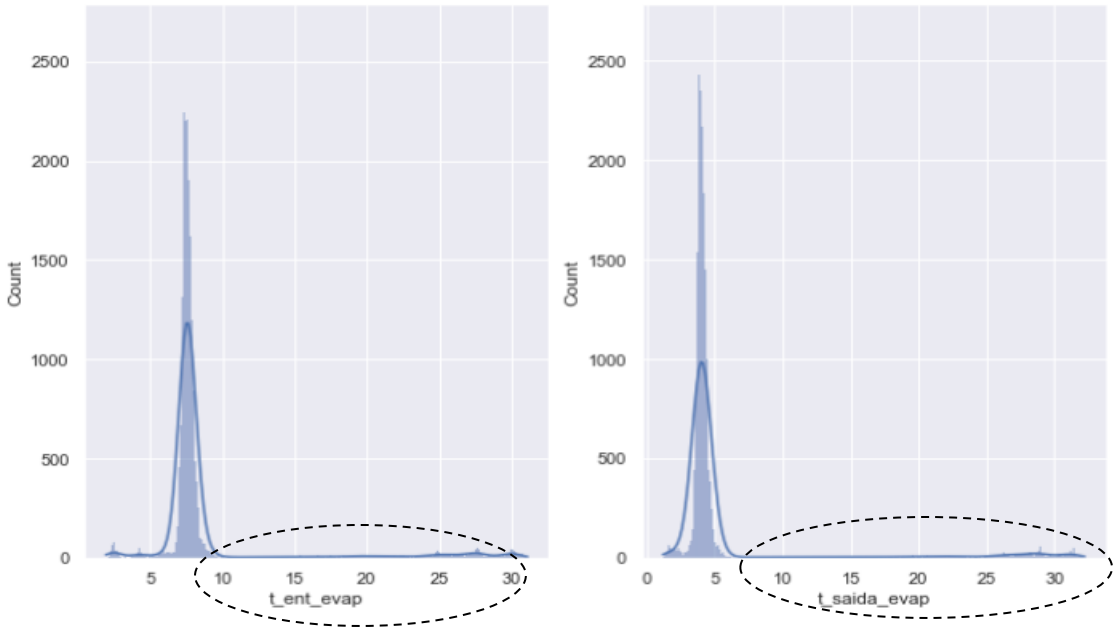
Fonte: Elaborado pela autora.

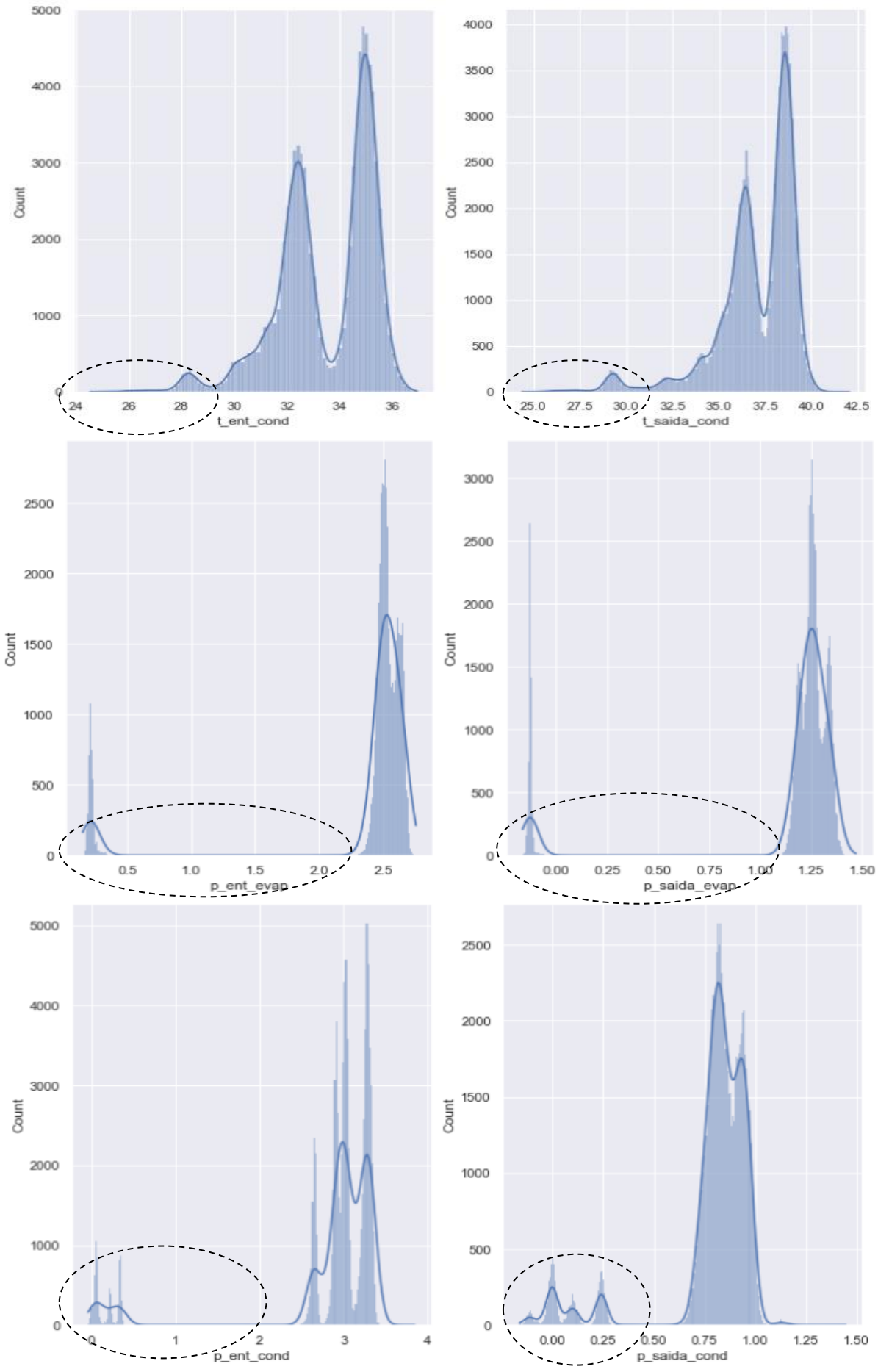
Um comportamento anômalo, como pode ser observado nos círculos hachurados, são picos fora da tendência global observada no tempo. Na corrente do motor do compressor, esses valores são em torno de 50000 A, no volume da câmara de compressão do compressor são de 5000%, nas temperaturas de superaquecimento de sucção e descarga, em torno de 100°C e para o vácuo em torno de 1000 mbar. Esses picos trata-se de fundos de escala característicos de medição, visto que, fisicamente, essas variáveis não chegariam a tais valores. Também, há períodos com comportamentos anômalos, os quais trata-se de manutenção (13/01 – 18/01), quebra do elemento rotativo do compressor (13/02 – 18/02) e intervenções corretivas (12/08 –15/08).

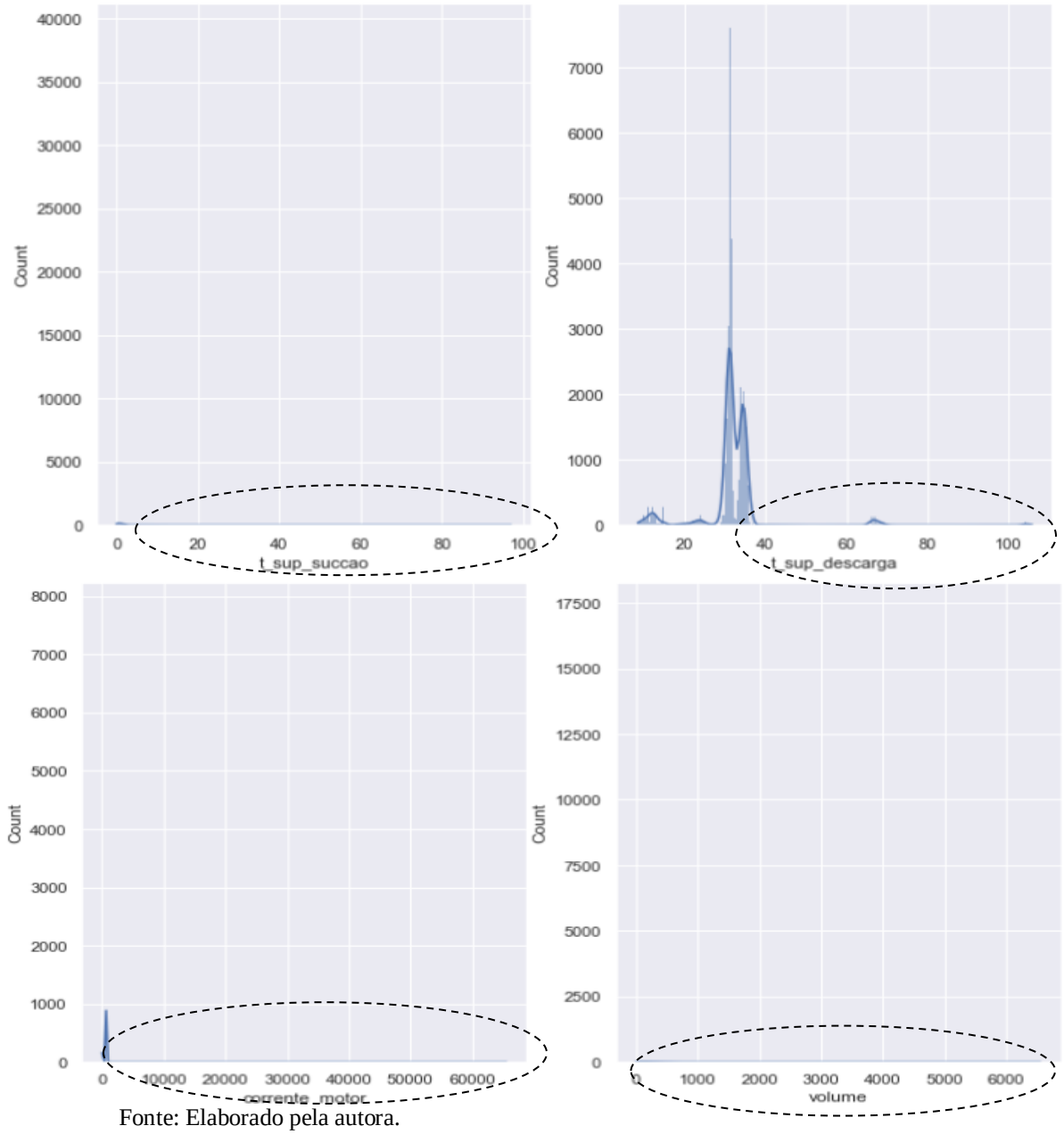
5.1.3 Distribuição dos dados brutos

A Figura 21 mostra a distribuição dessas variáveis brutas no período de janeiro. A visualização dos demais períodos encontra-se no APÊNDICE B.

Figura 21 – Distribuição das variáveis de processo (brutas) do Chiller em janeiro/2022




Count
t_ent_cond
t_saida_cond
p_ent_evap
p_saida_evap
p_ent_cond
p_saida_cond



Fonte: Elaborado pela autora.

Como pode-se observar, os dados brutos não seguem uma distribuição normal perfeita. Para as temperaturas do evaporador, tem-se desvios positivos, que variam de 10°C à 30°C. Para as temperaturas do condensador, o inverso ocorre, tendo-se desvios negativos, que variam de 25°C à 30°C, com a ocorrência de binormais. Essa característica de faixa de outlier com intervalos próximos para cada trocador, reflete uma condição de equilíbrio térmico. Ou seja, momentos em que o *Chiller* está parado, e a água apenas está circulando no evaporador e no condensador. Já para as pressões, tem-se regiões de desvios próximas de zero, e até negativas, o que indicam fundo de escala.

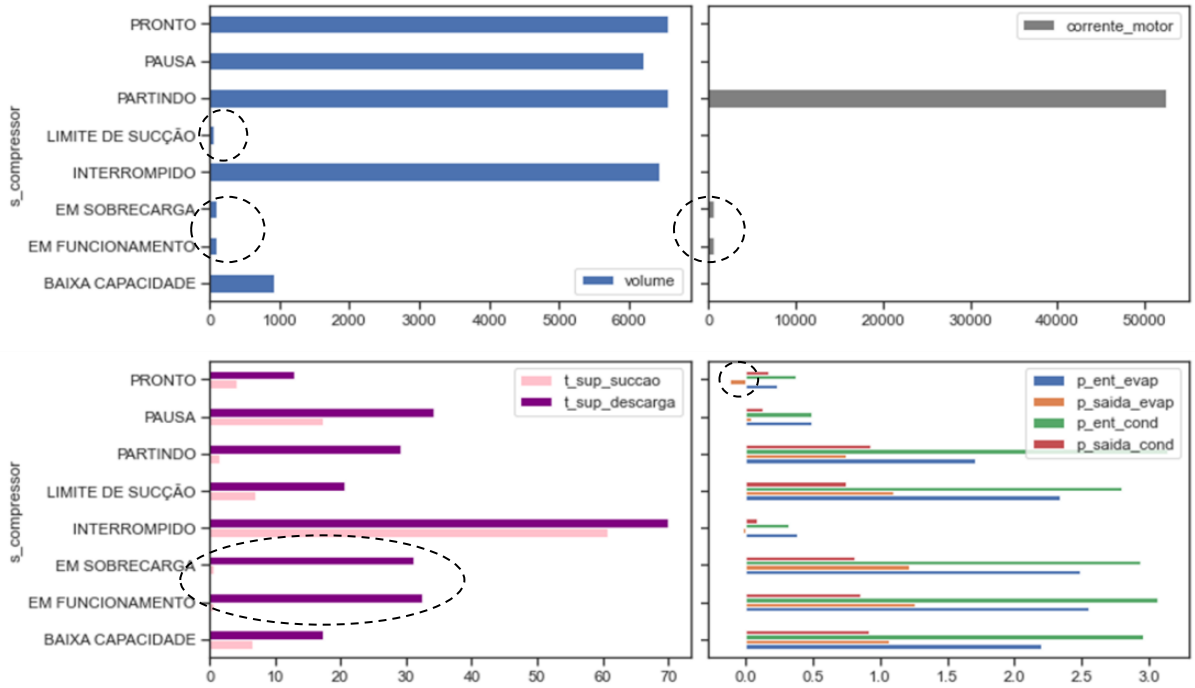
Para as variáveis do compressor, nota-se o desvio extremo para as temperaturas de superaquecimento. Em média, o valor para a sucção deve se manter em 1°C e a descarga em

40°C. Devido a isso, percebe-se como é difícil até a visualização dos pontos próximos de 1°C.Por fim, o comportamento é parecido para a corrente do motor, que em média opera em 630 A, e o volume da câmara que deve ir no máximo à 100%, porém possuem desvios positivos de fundo de escala.

5.1.4 Outlier's pelo status do compressor

A variável categórica do status do compressor serviu de auxílio para a retirada desses outlier's. Primeiramente, agrupou-se a média de algumas variáveis de processo de acordo com o status do CLP do compressor, a fim de se verificar o comportamento das mesmas em diferentes regiões de operação. Os resultados seguem abaixo para o mês de janeiro.

Figura 22 – Comportamento da média das variáveis de acordo com o *status* do compressor



Fonte: Elaborado pela autora.

Pode-se notar que no volume da câmara de compressão, o qual deve ter uma variação de 0-100%, os outlier's surgem principalmente fora das regiões de funcionamento efetivo do compressor, as quais são: "Em funcionamento", "Em sobrecarga" e "Limite de sucção". Ou seja, no transiente do equipamento ("Partindo"), ou quando há desarmes ("Interrompido"), retomadas ("Pronto"), ou mesmo pausas feitas pelo operador ("Pausa"), o instrumento sobe para valores de fundo de escala. Uma outra variável que segue esse comportamento é a própria corrente do motor. Seus valores sobem para fundos de escala

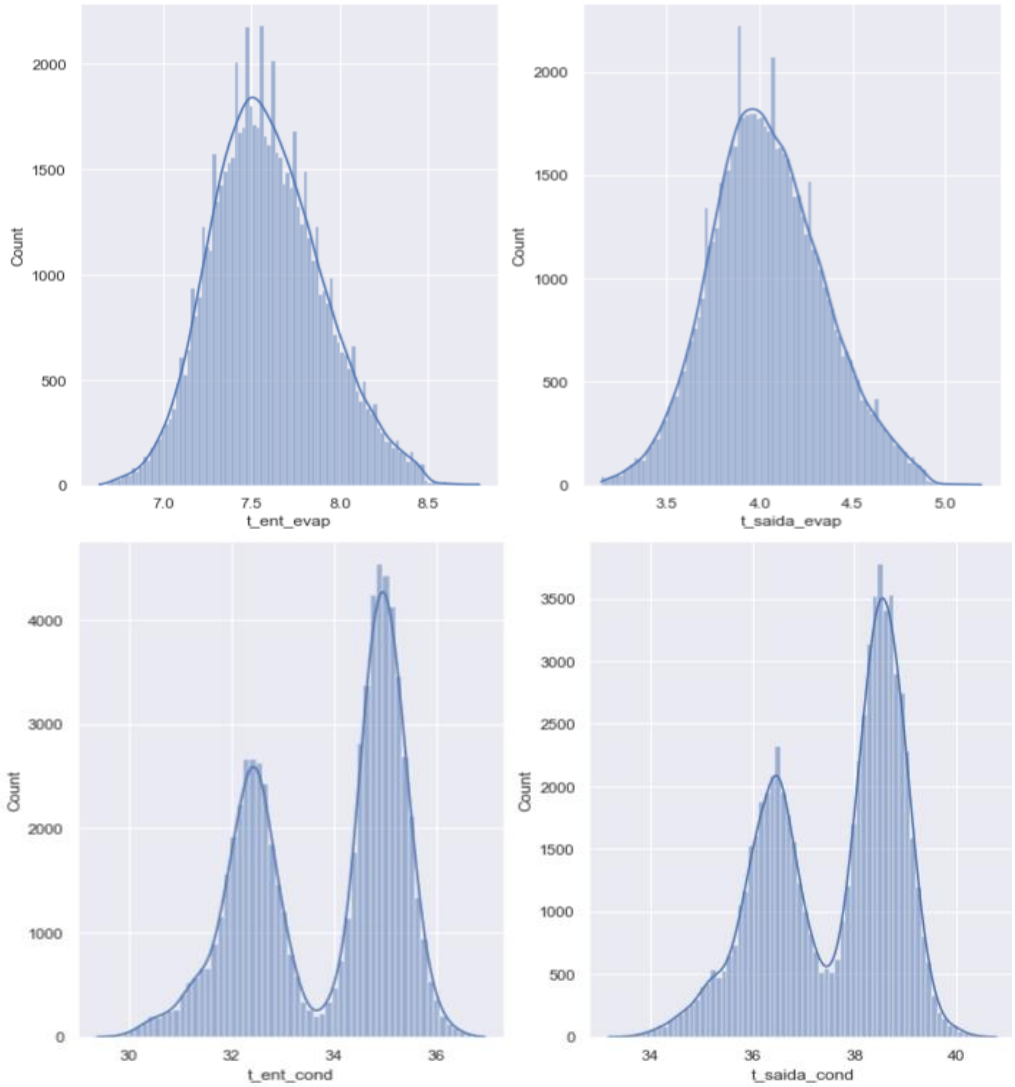
sempre nas partidas, mesmo com o motor contendo inversor e a partida não sendo direta. Em momentos de desarmes e pausas, os valores zeram, visto o compressor estar parado.

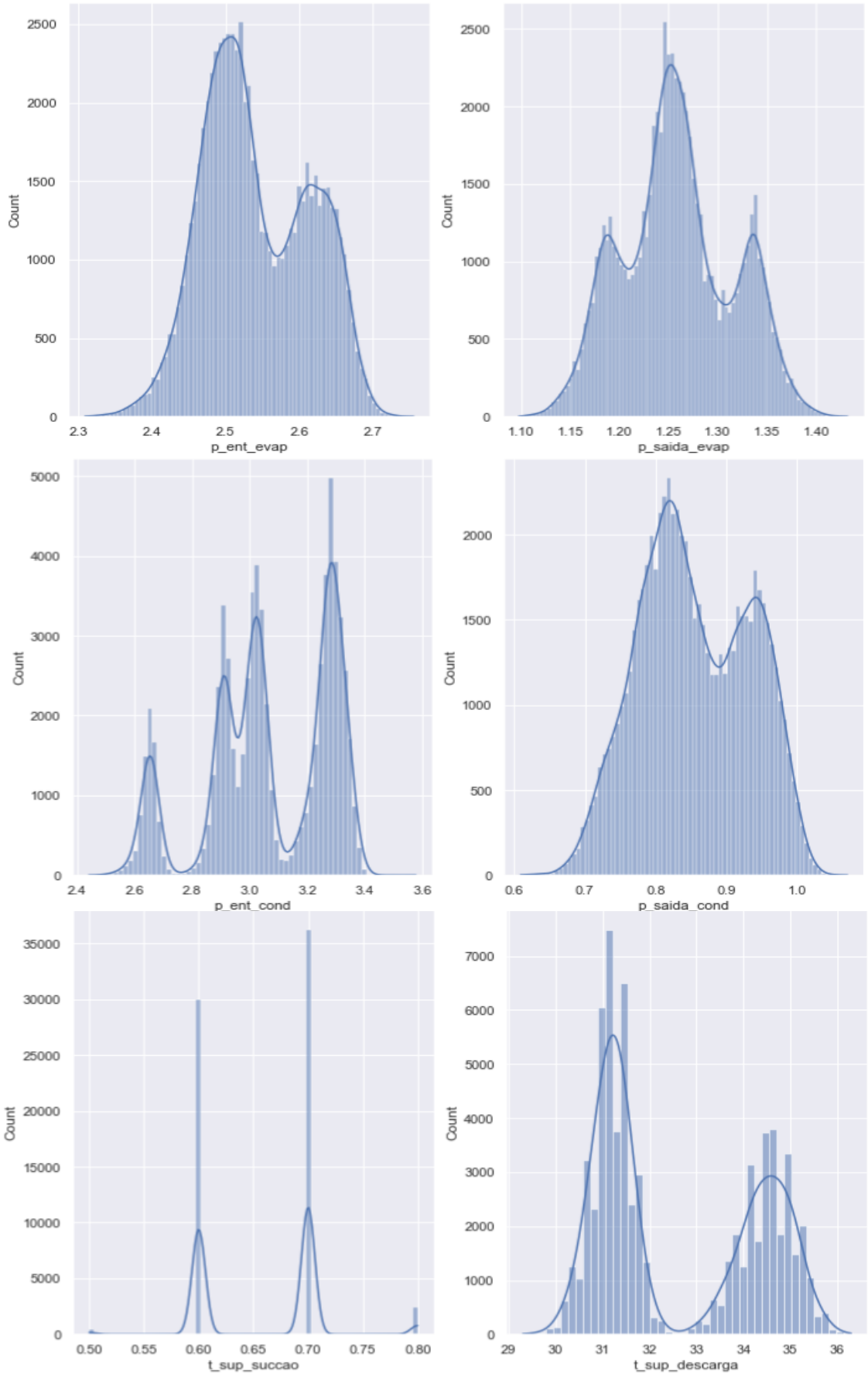
Além disso, as faixas normais de temperaturas de superaquecimento também se concentram apenas no status de “Em funcionamento” e “Em sobrecarga”. Pressões negativas e com valor zero encontram-se também fora dessas regiões.

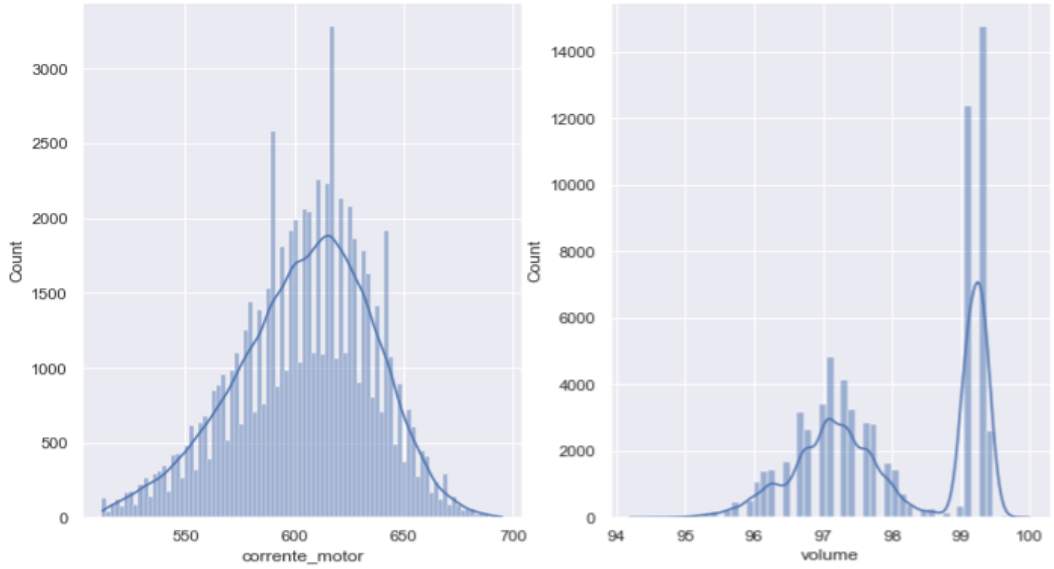
5.1.5 Distribuição dos dados tratados

Tratou-se os dados pelo método do Intervalo Interquartílico (IQR) e excluiu-se as regiões de operação fora de “Em funcionamento” e “Em sobrecarga”. A distribuição dos dados limpos para o período de janeiro segue abaixo. Para os demais períodos, consultar o APÊNDICE C.

Figura 23 – Distribuição das variáveis de processo tratadas do Chiller em janeiro/2022




Count
p_ent_evap
p_saida_evap
p_ent_cond
p_saida_cond
t_sup_succao
t_sup_descarga



Fonte: Elaborado pela autora.

Uma vez tratados, pode-se perceber uma melhor homogeneidade dos dados. Percorrendo as variáveis de processo, algumas percepções podem ser tomadas. Em relação ao evaporador, sua temperatura de entrada, em tese, deve ser em torno de **6°C**, contudo, neste período, a média esteve 1,5°C acima, ou seja **7,5°C**. Já a temperatura de água gelada de saída, a qual deve ser entregue, em torno de **2°C**, esteve o dobro, **4°C**.

Para o condensador, há ainda duas regiões de distribuição. Fisicamente, isso pode ser explicado pelo fato de a água que circula no mesmo ser oriunda da torre de resfriamento. A torre se mantém exposta ao ambiente externo, e a depender da hora do dia, pode estar mais quente ou mais fria. Além disso, teoricamente, a corrente de entrada no condensador deve ser em torno de **31°C**, sendo que a média da primeira curva normal se mantém próxima disso (**32°C**). Para a corrente de saída, pós remover calor da amônia, a temperatura teórica deve ser em torno de **35°C**, tendo ficado a média da primeira normal em **36°C**, ou seja, bastante factível.

A temperatura de superaquecimento de sucção de amônia se manteve dentro de seus limites, até **1°C**, embora para isso a variação seja muito pequena, bem como o superaquecimento de descarga ficou dentro de limites aceitáveis, de **30°C** até **40°C**. Para o compressor, a corrente elétrica possui variações dentro da normalidade, de **500 A** até **700 A**, bem como o volume da câmara de compressão tem seu limite superior em 100%. Uma observação importante é a de que em todos os períodos analisados, o volume da câmara de compressão praticamente não variou. Em números, sempre se manteve entre **90%** e **100%**. Em termos técnicos, é possível afirmar que a mesma não modula. Isso explica, indiretamente, o fato de o consumo de energia elétrica do motor do compressor ser aproximadamente

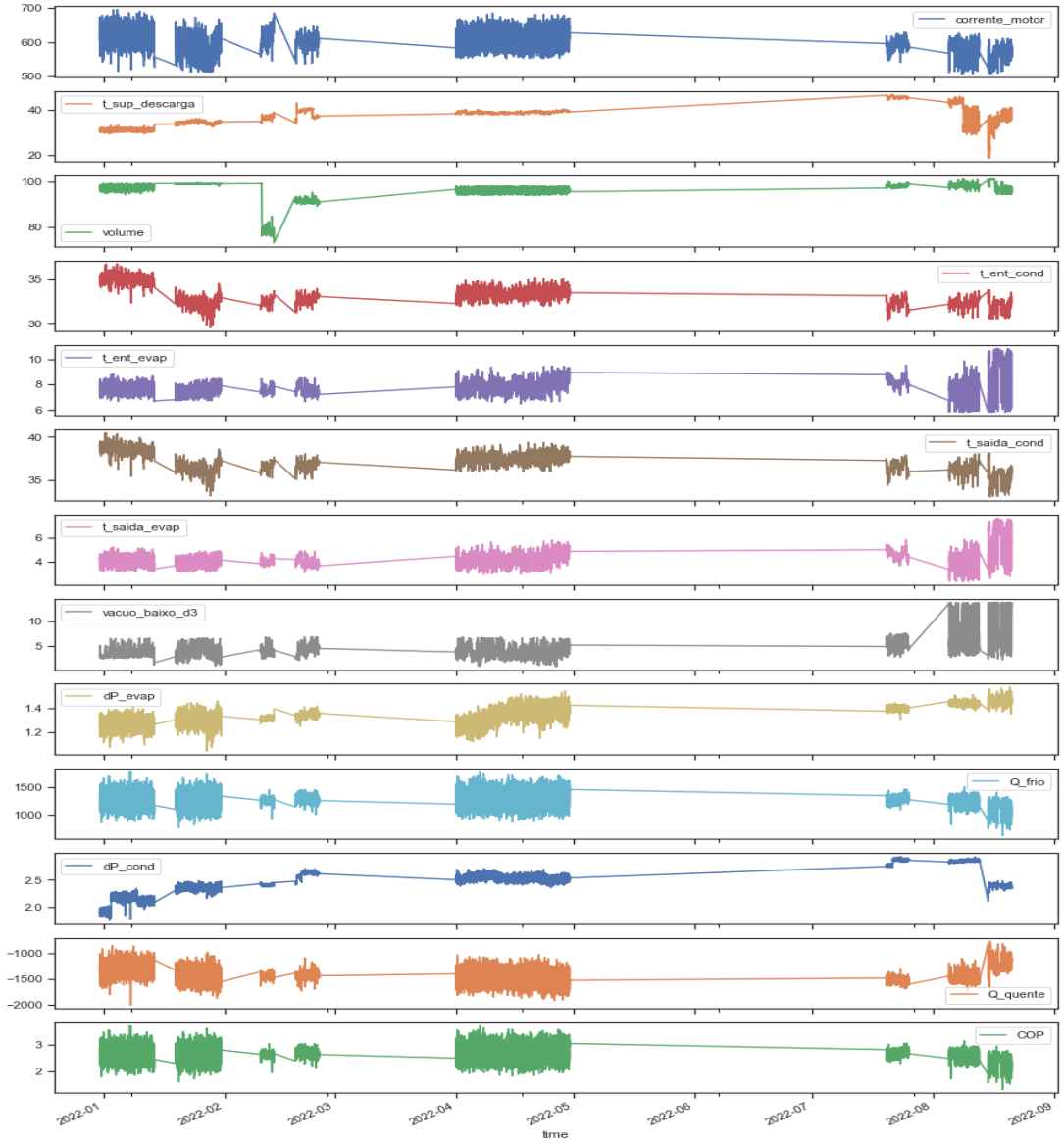
constante. Em termos de projeto, é como se o motor e o compressor houvessem sido subdimensionados para as cargas térmicas demandadas do ciclo de refrigeração imposto, e sempre operassem em seu máximo. Ou seja, um indício prático de ofensor energético.

Após a retirada dos ruídos em geral do conjunto de dados, foi possível calcular os indicadores ($Q_{frio}, Q_{quente}, \Delta_{P-evap}, \Delta_{P-cond,}\ COP$), bem como concatenar os períodos em uma única base de dados para a análise exploratória geral.

5.1.6 Teste de hipótese entre períodos

Foi plotado um gráfico da série temporal de janeiro até agosto após a retirada dos outlier's (Figura 24).

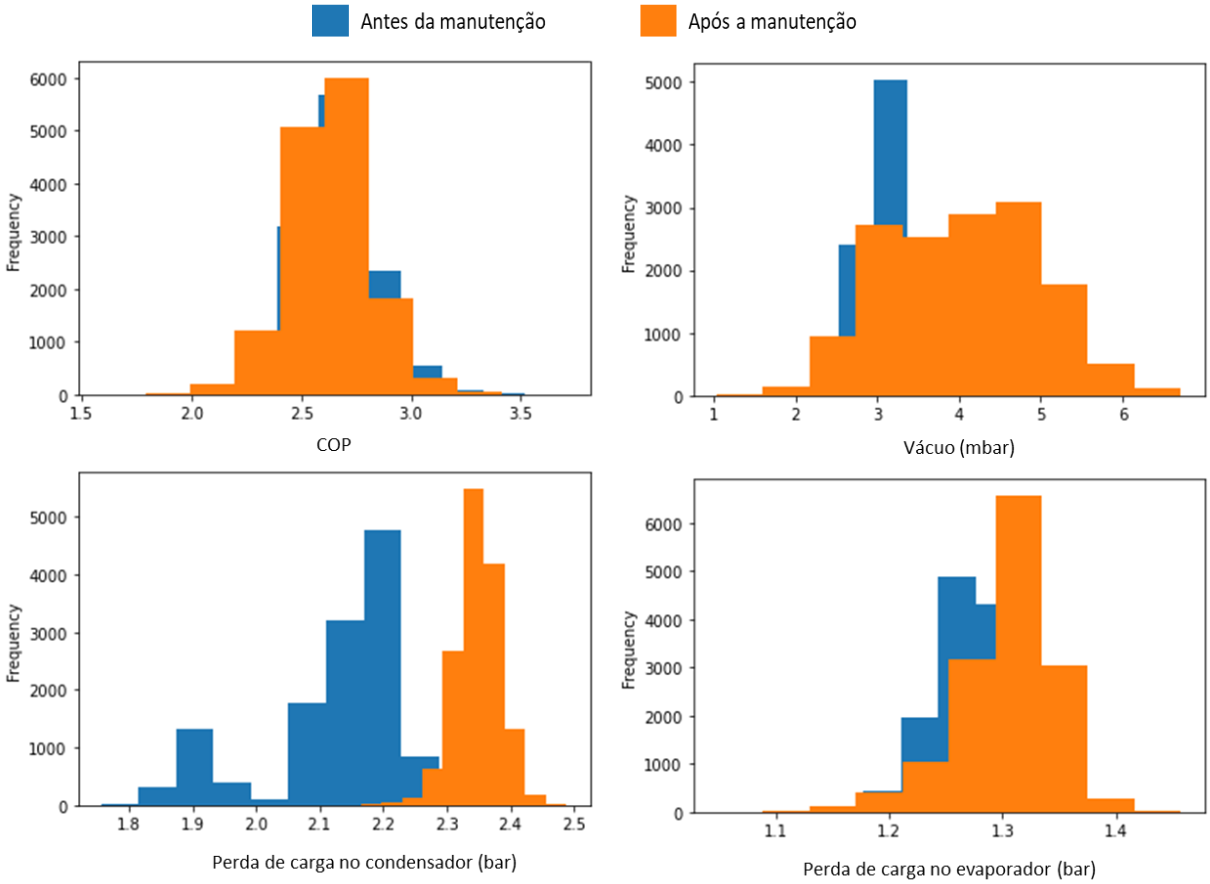
Figura 24 – Comportamento das variáveis de processo tratadas do Chiller



Fonte: Elaborado pela autora.

Embora a Figura 24 projete uma visão ampla de possíveis tendências nas variáveis de processo ao longo do primeiro semestre de 2022, sua leitura não é tão prática assim. Dessa forma, para comparar os períodos, foram feitos testes de hipóteses entre os principais marcos de mudança nos períodos. Os resultados seguem abaixo:

Figura 25 – Teste de hipótese no período de janeiro

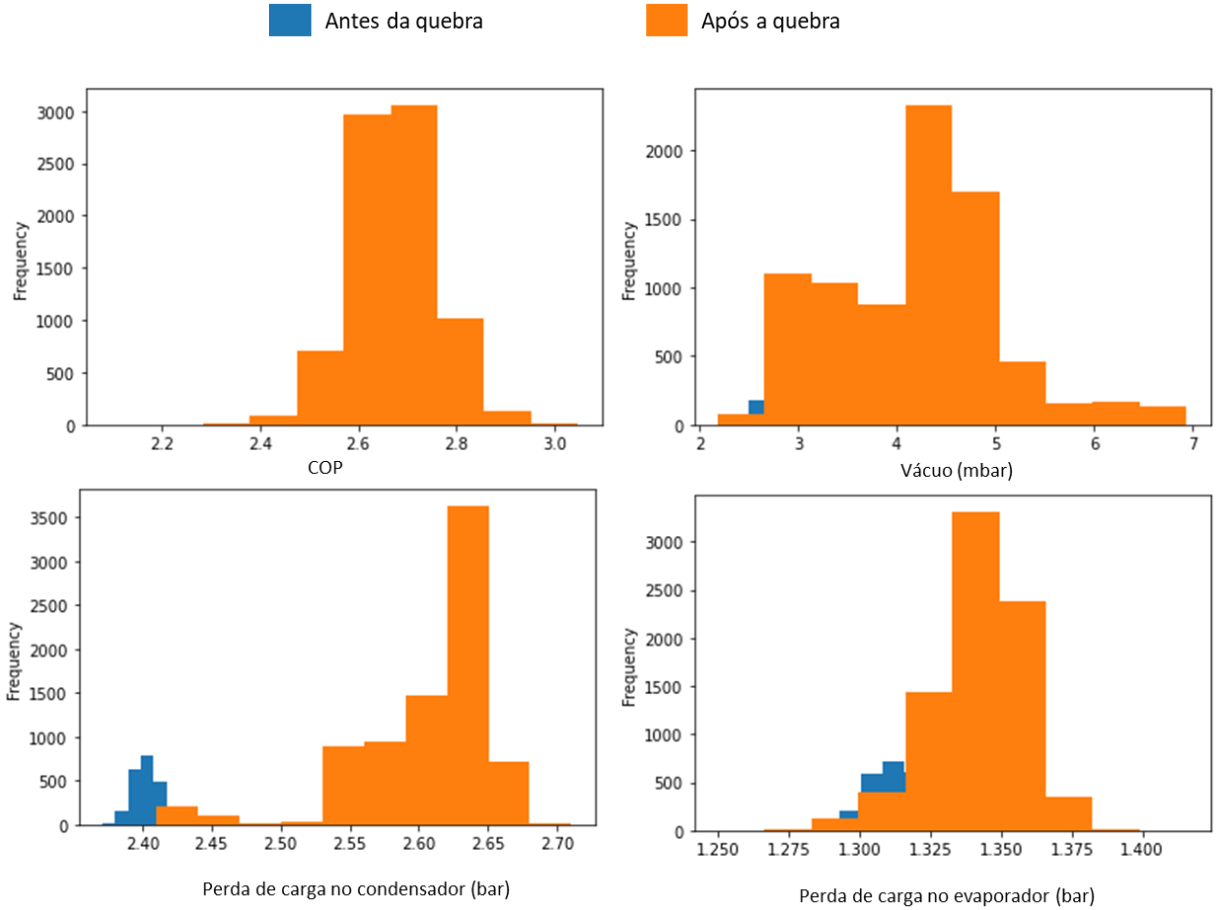


Fonte: Elaborado pela autora.

O primeiro teste de hipótese compara a média do COP, do vácuo e das perdas de carga antes e depois da manutenção do equipamento em janeiro. O COP e a perda de carga no evaporador praticamente não sofreram alterações, porém o vácuo subiu de **3,51 mbar** à **4,05 mbar**, e a perda de carga no condensador subiu de **2,12 bar** à **2,35 bar**.

O segundo teste de hipótese compara o período de fevereiro, antes do elemento rotativo do compressor ser danificado e após a sua quebra (Figura 26). Como pode-se notar, a hipótese de mudança das quatro variáveis é rejeitada, já que as distribuições se sobrepõem. A perda de carga no condensador é a única que tem um leve aumento de **2,4 bar** para **2,6 bar**, o que demonstra um aumento relativo, inclusive ao mês anterior de janeiro.

Figura 26 – Teste de hipótese no período de fevereiro

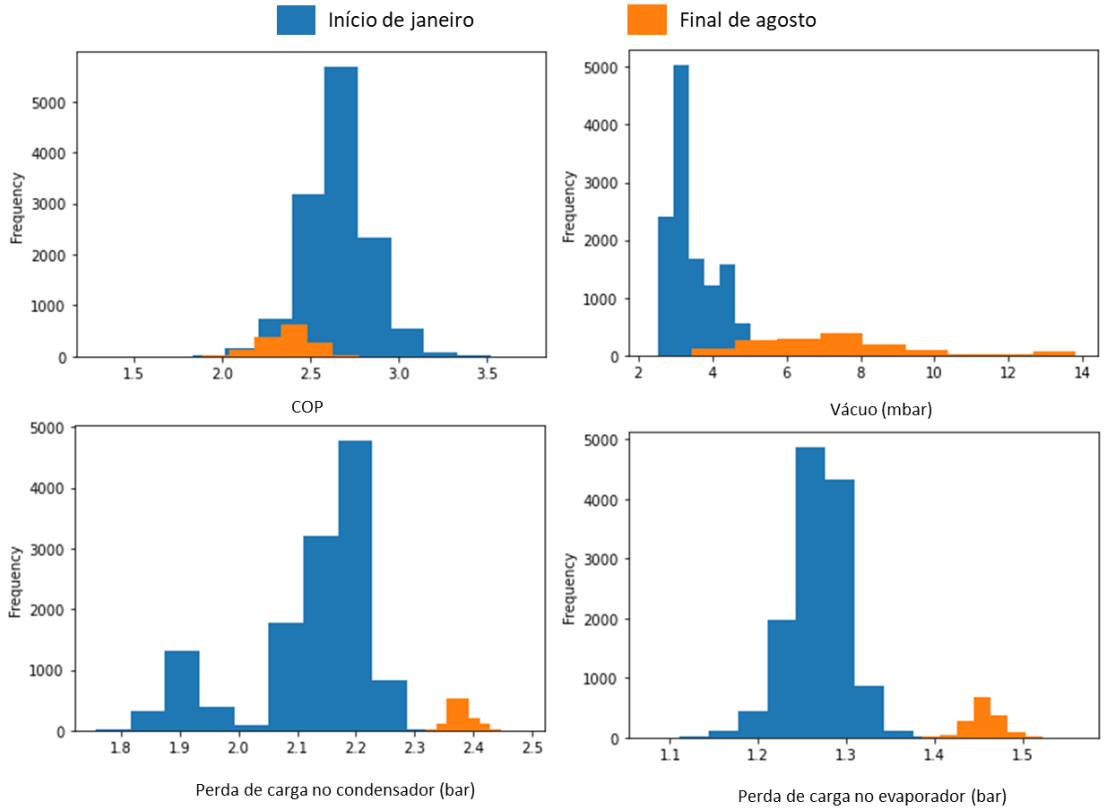


Fonte: Elaborado pela autora.

O terceiro teste comparou o início do período de janeiro com o final do período de agosto, a fim de compreender se haviam, de fato, mudanças significativas ao longo do semestre.

Como pode-se notar na Figura 27, a média do COP cai de **2,7** para **2,3**, o que não é tão significativo. Já o vácuo tem uma piora significativa, de **3,5 mbar** à **7,2 mbar**, mesmo após as intervenções corretivas desse período. A perda de carga no condensador sobe de **2,12 bar** à **2,38 bar**. Comparado com o final do período de fevereiro, há uma queda dessa perda de carga, o que pode ser evidenciado também pela Figura 27. Para o evaporador, há um leve aumento de **1,27 bar** para **1,46 bar**.

Figura 27 – Teste de hipótese no início de janeiro até o fim de agosto



Fonte: Elaborado pela autora.

Após os testes, pode-se perceber que a correlação entre o COP e o vácuo do desodorizador é indireta. Isso porque, mesmo em períodos de piora do vácuo, o COP se altera muito pouco. Teoricamente, isso pode ser explicado pelo fato de que o numerador desse indicador é calculado pelo delta de temperatura do evaporador, ou seja, como observado nas distribuições da seção 5.1.5, caso a temperatura de entrada do evaporador aumente, se a amônia que está circulando no interior do ciclo não for capaz de absorver essa quantidade maior de energia, a temperatura entregue na saída será proporcionalmente maior também, porém o delta ainda será o mesmo. Consequentemente, se terá o mesmo Coeficiente Operacional de Performance.

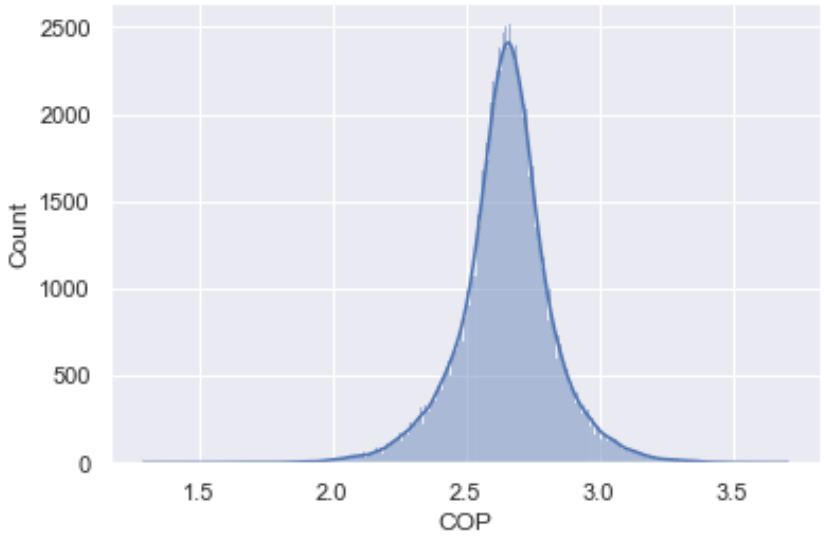
No caso deste *Chiller*, em especial, tem-se o fator contribuinte da modulação irrisória da câmara de compressão, e consequentemente, o trabalho máximo fornecido ao motor do compressor, como discutido ao final da seção 5.1.5 também. Assim, em alguns cenários, o COP pode se manter o mesmo, porém outras variáveis de processo estarão se modificando. A fim de investigar isso, fez-se três modelos de classificação com o objetivo de investigar quais eram as variáveis que influenciariam, de fato, o COP.

5.2 Modelagem

5.2.1 Segmentação dos dados

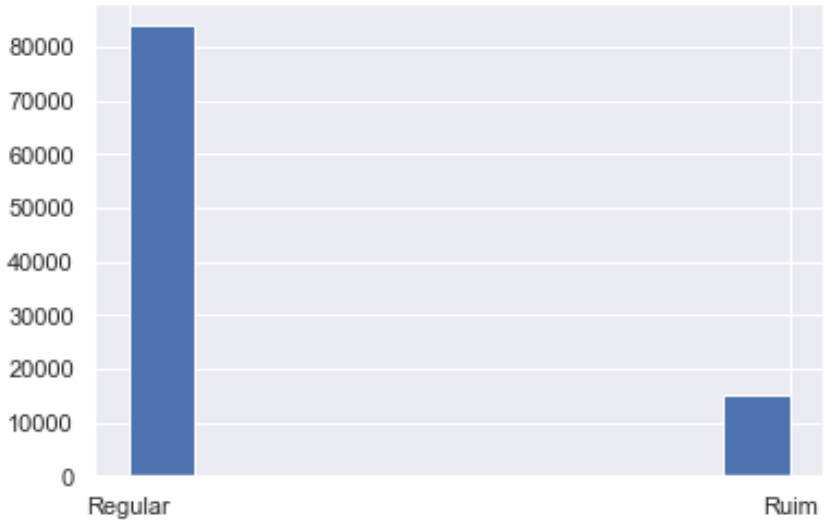
Para a segmentação dos valores contínuos nas classes “Ruim” e “Regular” da variável alvo (COP), foi plotado a distribuição de seus valores (Figura 28). Foram contados 14.920 (15%) como “Ruim” e 83.940 (85%) como regular (Figura 29). A classe “Ruim” foi escolhida como a classe de interesse.

Figura 28 – Distribuição do COP



Fonte: Elaborado pela autora.

Figura 29 – Contagem das classes do COP

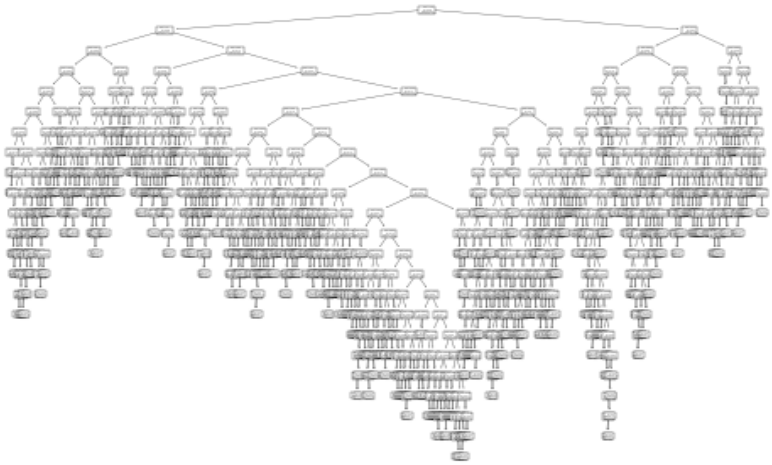


Fonte: Elaborado pela autora.

5.2.2 Árvore de Decisão

Inicialmente, fez-se a primeira implementação do algoritmo, sem calibrar os hiperparâmetros. O resultado, como esperado, foi uma árvore extensa, já que a sua profundidade não foi limitada (Figura 30).

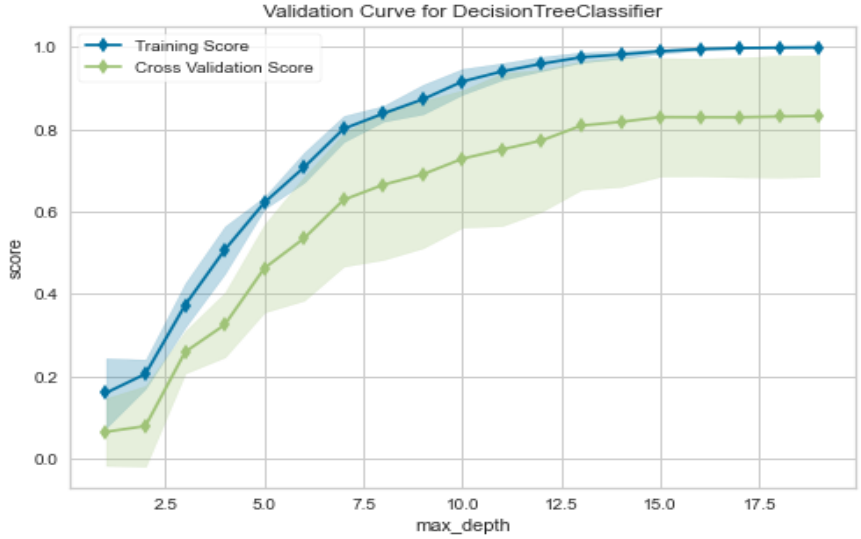
Figura 30 – Visualização parcial da Árvore de Decisão com seus parâmetros padrões



Fonte: Elaborado pela autora.

Assim, o primeiro passo foi calibrar a profundidade (max_depth) de acordo com uma curva de validação. Foi escolhido um intervalo de 1 à 20. A calibração foi feita utilizando validação cruzada (cv=5) e a revocação como pontuação (*score*). O resultado segue abaixo:

Figura 31 – Curva de Validação para Árvore de Decisão



Fonte: Elaborado pela autora.

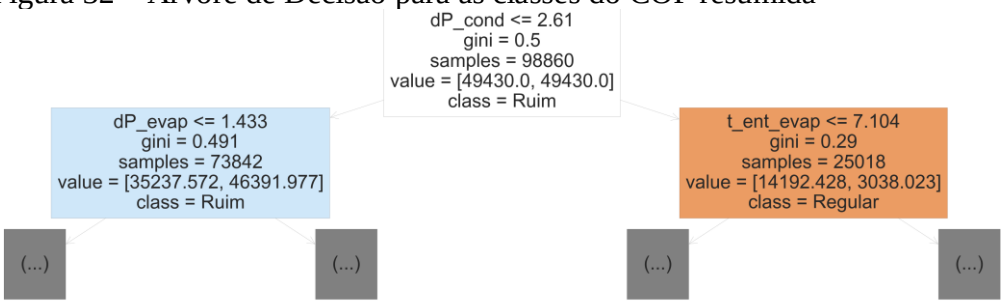
Olhando para a Figura 31, percebe-se que a partir de uma profundidade de árvore de 15, a pontuação do modelo, tanto de treino como de teste não se altera mais. Também, é possível perceber que em valores baixos de profundidade, de até 5, as pontuações são próximas, começando a se distanciar a partir daí, o que, em tese promove o sobreajuste. Assim, testou-se com o GridsearchCV valores finais de profundidade de 7, 11 e 13, bem como valores para os outros hiperparâmetros. Um resumo descritivo dos valores finais otimizados segue na Tabela 9, bem como um resumo da árvore final pode ser encontrado na Figura 32. A árvore completa encontra-se no APÊNDICE D.

Tabela 9 – Resumo dos hiperparâmetros otimizados para Árvore de Decisão

HIPERPARÂMETRO	**NOMEAÇÃO TÉCNICA**	**VALOR FINAL**	**PADRÃO**
Máxima profundidade	max_depth	11	None
Máximos nós residuais	max_leaf_nodes	50	None
Mínimo de amostras no nó residual	min_samples_leaf	3	1
Mínimo de amostras para o nó ser dividido	min_samples_split	5	1
Decréscimo mínimo de impureza	min_impurity_decrease	0	0
Peso das classes	class_weight	*balanced*	None

Fonte: Elaborado pela autora.

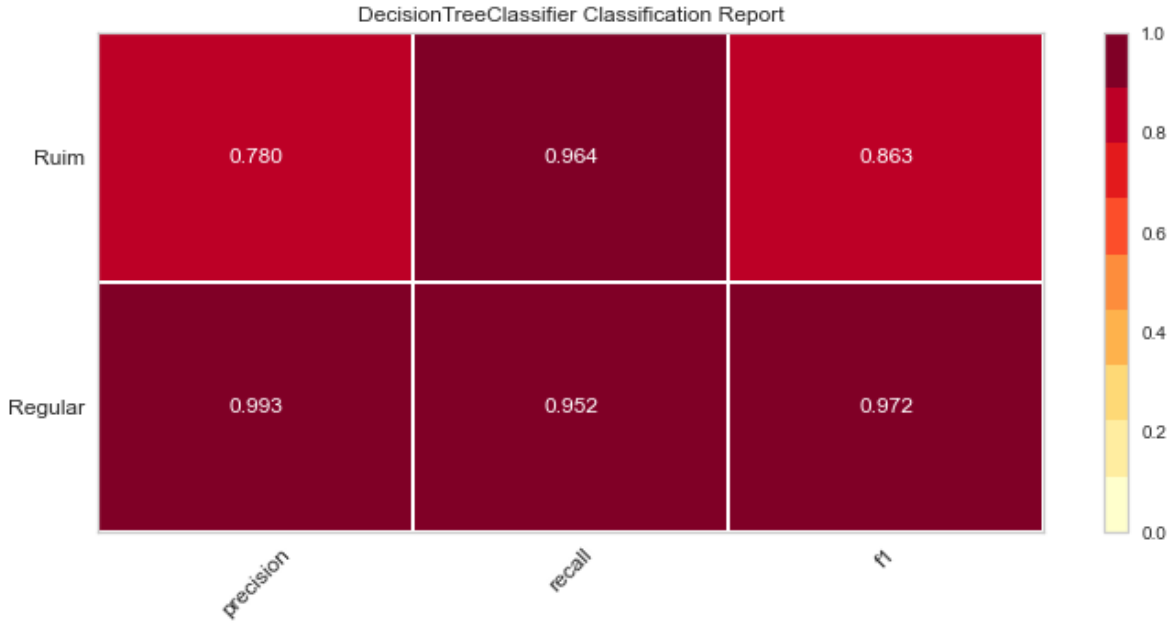
Figura 32 – Árvore de Decisão para as classes do COP resumida



Fonte: Elaborado pela autora

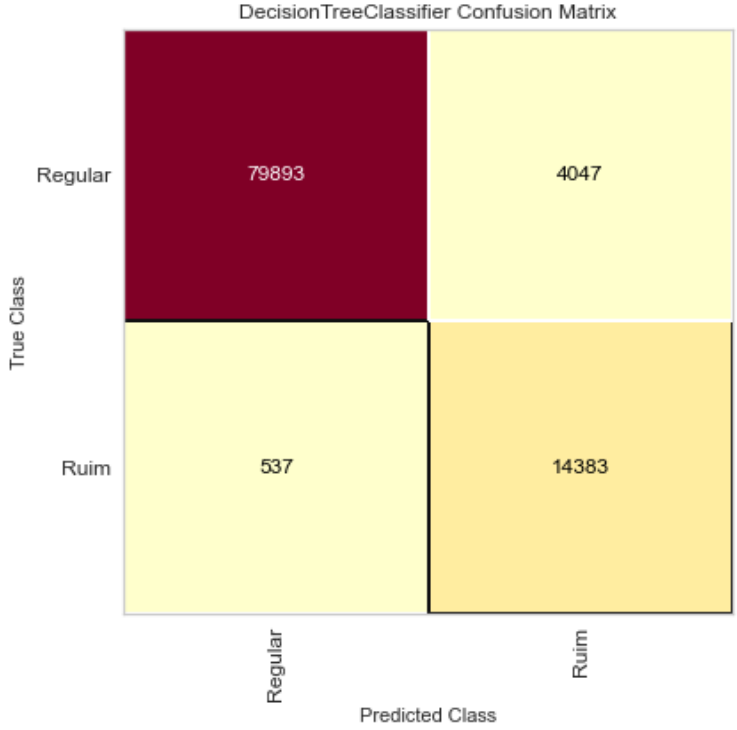
A fim de avaliar o modelo, foram plotados o relatório de classificação com as pontuações de cada métrica, a Matriz de Confusão e um gráfico de Importância dos Atributos, isto é, aquelas variáveis que mais contribuem para a classificação do COP. Os resultados seguem abaixo (Figuras 33, 34, e 35):

Figura 33 – Relatório de Classificação para a Árvore de Decisão



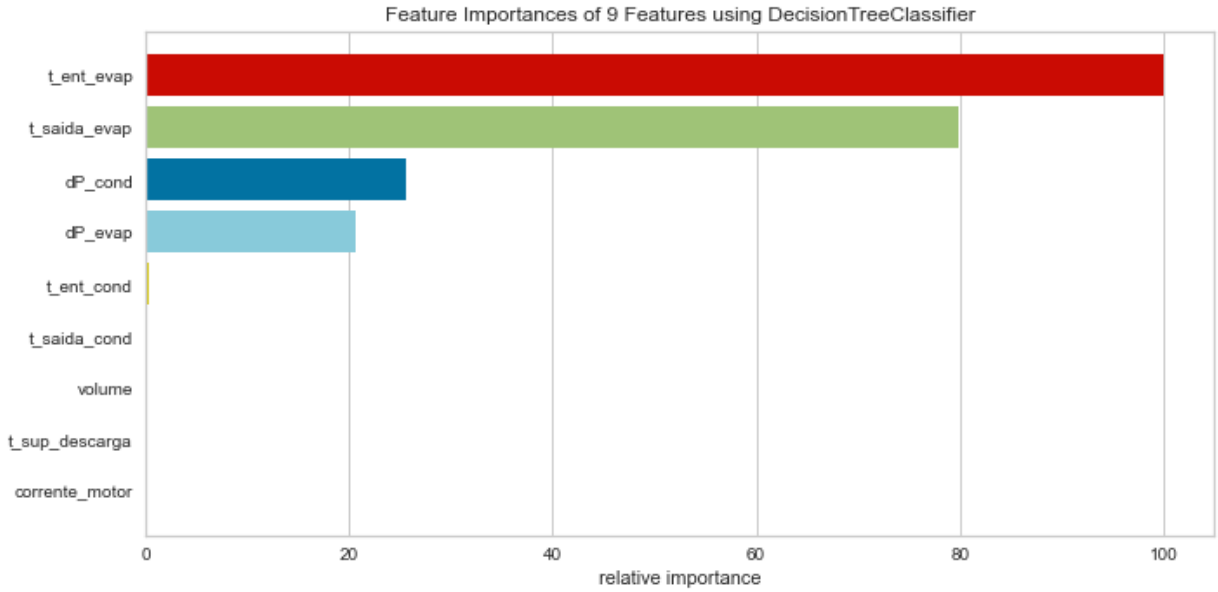
Fonte: Elaborado pela autora

Figura 34 – Matriz de Confusão para a Árvore de Decisão



Fonte: Elaborado pela autora.

Figura 35 – Importância das variáveis classificadas de acordo com o COP para Árvore de Decisão



Fonte: Elaborado pela autora.

Pelo relatório de classificação, para as três métricas, os resultados, no geral, foram bons, estando acima de 80%, com exceção da precisão para a classe Ruim. Além disso, para a revocação, a pontuação foi a mais balanceada para a predição das duas classes, estando acima de 95%. Para a matriz de confusão, o modelo errou apenas 5% das classificações. Por fim, percebe-se a maior importância para as temperaturas do evaporador (entrada e saída, respectivamente), seguida da perda de carga no condensador, perda de carga no evaporador, e uma mínima contribuição da temperatura de entrada do condensador. Para esse algoritmo, as variáveis do compressor possuem 0% de contribuição.

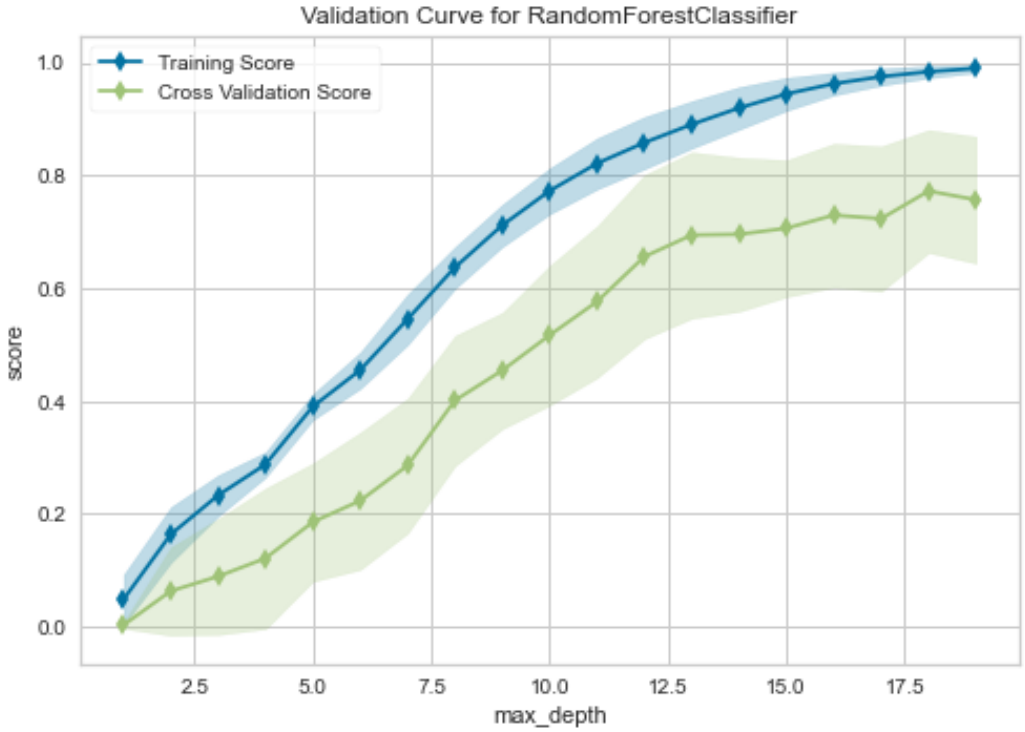
Como explicado na seção 4.5.2, pode-se melhorar o algoritmo, calibrando e percorrendo não apenas uma árvore, mas sim várias, ou seja, uma floresta. Para isso, foi feito um segundo modelo a fim de testar os mesmos dados. Os resultados encontram-se na próxima seção.

5.2.3 Floresta Aleatória

Para a floresta, buscou-se otimizar o modelo testando o número de árvores (*n_estimators*), bem como a profundidade das árvores utilizando uma curva de validação novamente (Figura 36). Testou-se 50, 100, 150 e 200 árvores (n_estimators). O melhor

parâmetros encontrado foi para **100** árvores, com uma profundidade maior: **13**. A curva de validação bem como a tabela com os parâmetros otimizados encontra-se abaixo:

Figura 36 – Curva de Validação para Floresta Aleatória



Fonte: Elaborado pela autora.

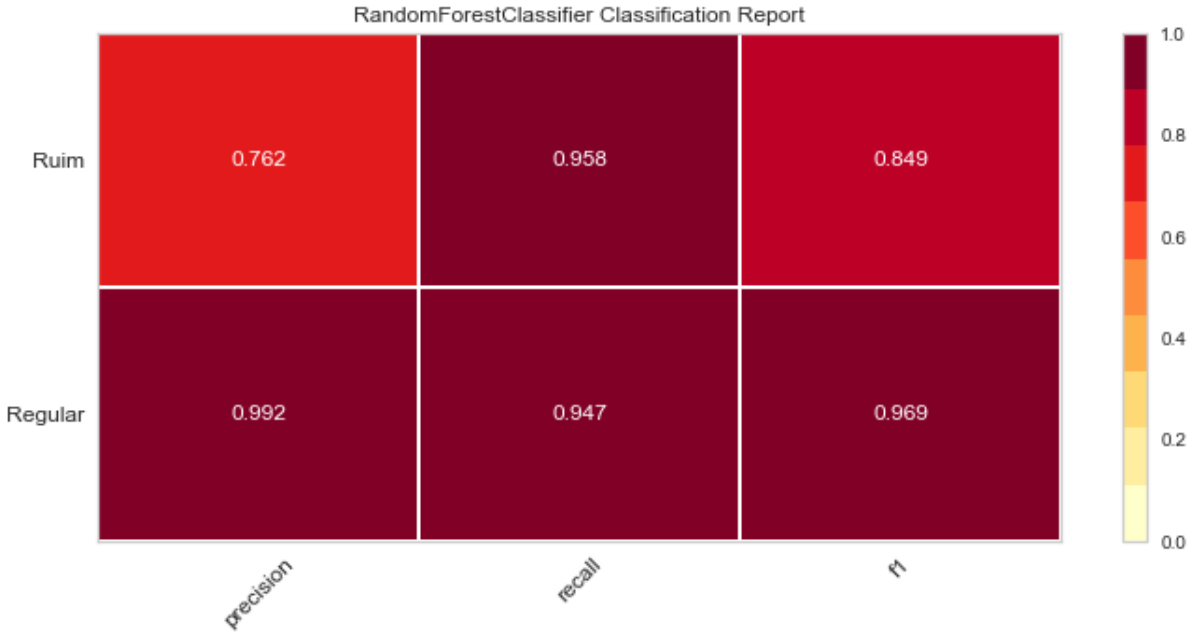
Tabela 10 – Resumo dos hiperparâmetros otimizados para Floresta Aleatória

HIPERPARÂMETRO	**NOMEAÇÃO TÉCNICA**	**VALOR FINAL**	**PADRÃO**
Máxima profundidade	max_depth	13	None
Número de estimadores	n_estimators	100	100
Número de processadores	n_jobs	-1	1
Inicialização	bootstrap	True	True

Fonte: Elaborado pela autora.

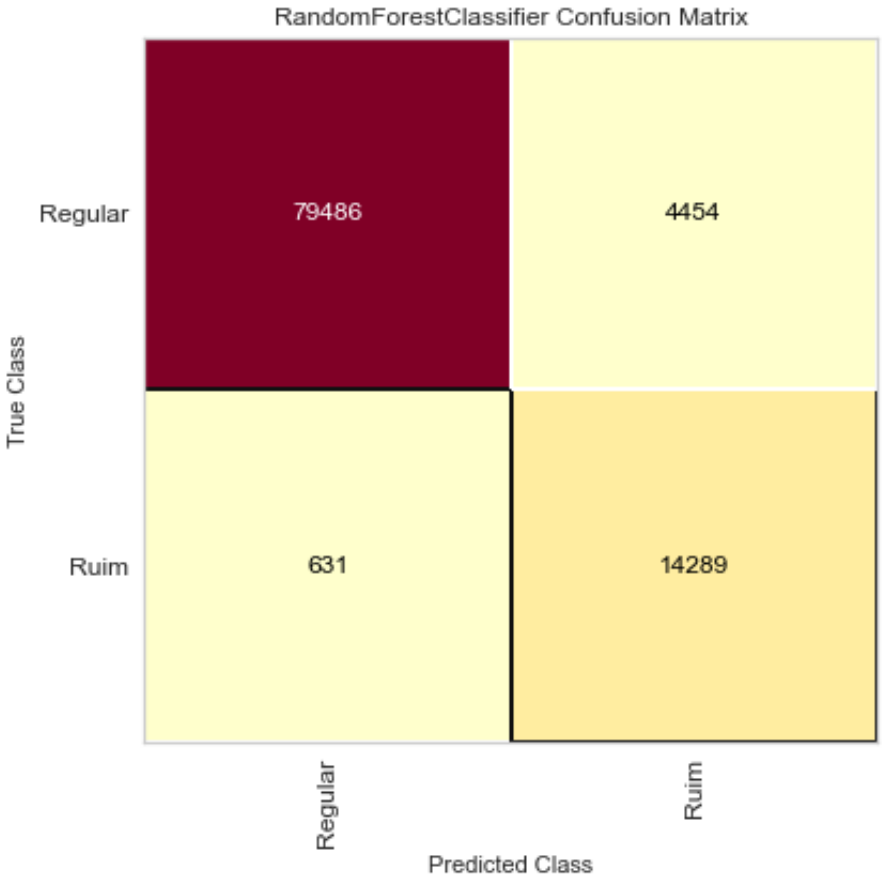
Abaixo, tem-se o relatório de classificação, a Matriz de Confusão e o gráfico de Importância de Atributos para a Floresta Aleatória:

Figura 37 – Relatório de Classificação para a Floresta Aleatória



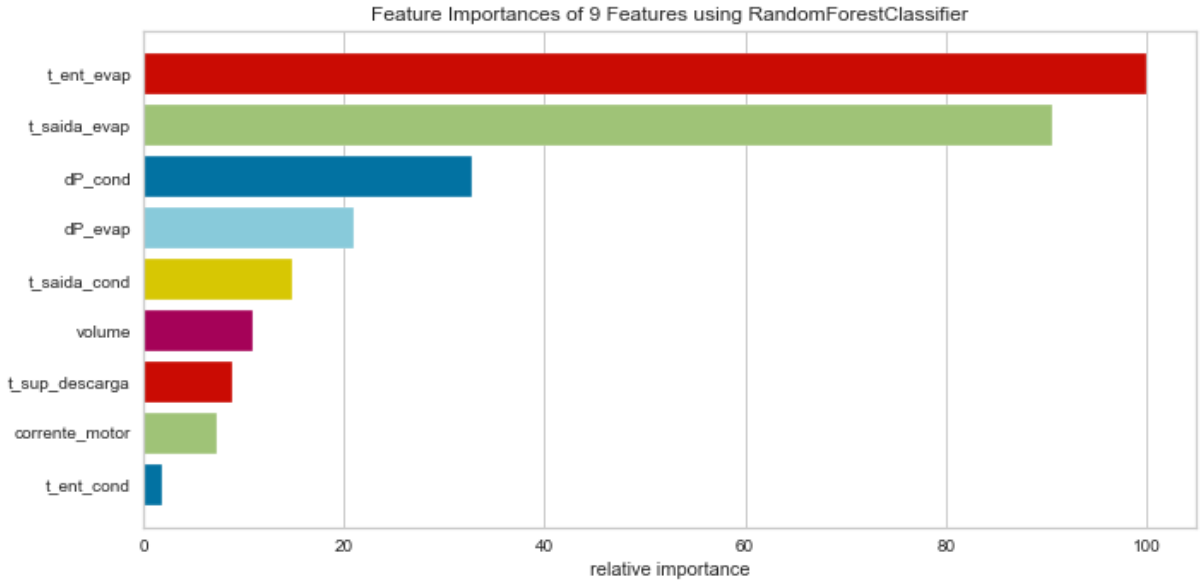
Fonte: Elaborado pela autora.

Figura 38 – Matriz de Confusão para a Floresta Aleatória



Fonte: Elaborado pela autora

Figura 39 – Importância das variáveis classificadas de acordo com o COP para Floresta Aleatória



Fonte: Elaborado pela autora.

Com o segundo modelo, não houve uma melhora significativa nas pontuações, como pode-se perceber pelo relatório de classificação. Pela análise da Matriz de Confusão, o modelo errou também apenas 5% das classificações. Porém, quando se observa o gráfico de importância de atributos, é possível notar que o modelo consegue capturar mais particularidades do sistema, atribuindo importâncias a outras variáveis que a árvore sozinha não previu. Aqui, os primeiro quatro atributos mais importantes se repetem (temperaturas do evaporador, perdas de carga do condensador e evaporador), mas seguidas da temperatura de saída do condensador, volume da câmara de compressão, temperatura de superaquecimento de descarga, corrente elétrica do motor, e por fim, a temperatura de entrada do condensador.

5.2.4 XGBoost

Por fim, testou-se um último modelo de classificação, a fim de comparar com os dois anteriores. Neste caso, a vantagem é a forma como o algoritmo percorre a árvore, minimizando a função de perda. Pela otimização dos hiperparâmetros, conseguiu-se com o dobro de árvores (*n_estimators*) = **200**, porém com uma profundidade (*max_depth*) de apenas **5**, chegar-se à melhores métricas. Especialmente neste algoritmo, não foi possível plotar uma curva de validação devido às limitações computacionais.

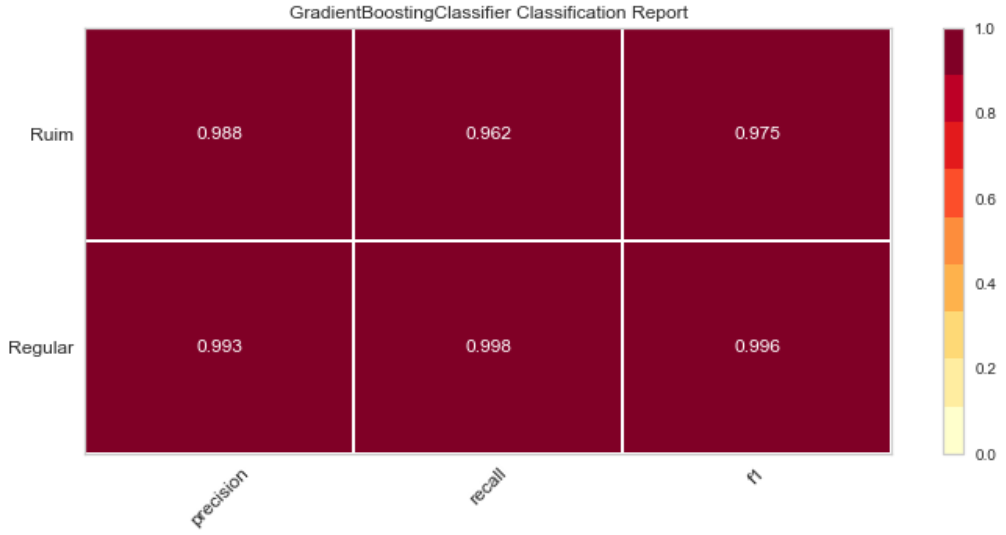
Os resultados seguem abaixo:

Tabela 11 – Resumo dos hiperparâmetros otimizados para o XGBoost

HIPERPARÂMETRO	NOMEAÇÃO TÉCNICA	VALOR FINAL	PADRÃO
Máxima profundidade	max_depth	5	None
Número de estimadores	n_estimators	200	100
Número de processadores	n_jobs	-1	1
Inicialização	bootstrap	True	True

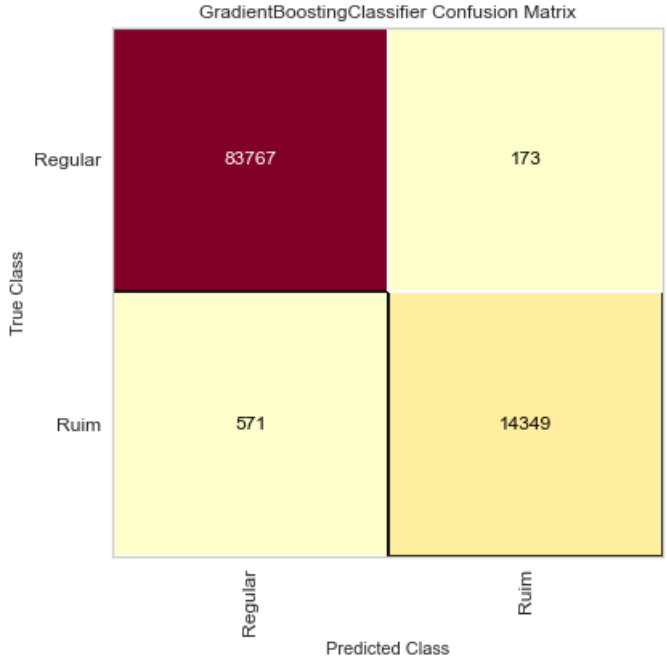
Fonte: Elaborado pela autora.

Figura 40 – Relatório de Classificação para o XGBoost



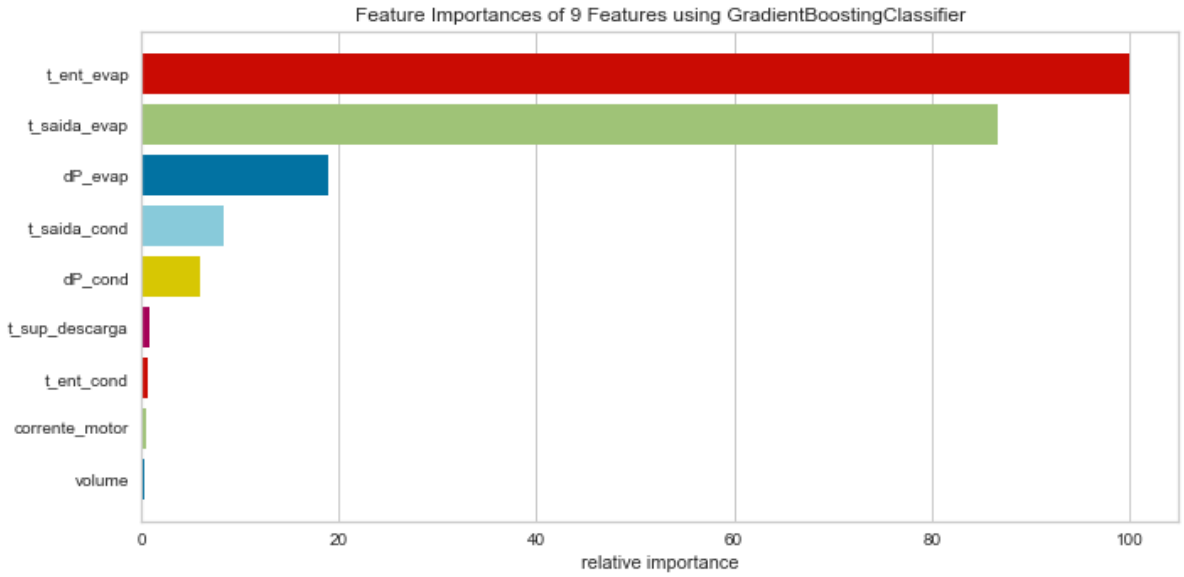
Fonte: Elaborado pela autora.

Figura 41 – Matriz de Confusão para o XGBoost



Fonte: Elaborado pela autora.

Figura 42 – Importância das variáveis classificadas de acordo com o COP para o XGBoost



Fonte: Elaborado pela autora.

Para o último modelo, as três pontuações foram acima de 95%. Além disso, o modelo não classificou corretamente apenas 0,07% dos dados. Para a importância de atributos, o modelo capta as particularidades de forma similar à Árvore de Decisão, no entanto com algumas diferenças. As temperaturas do evaporador mantem-se como os principais atributos de importância, porém seguida da perda de carga no evaporador, temperatura de saída no condensador e perda de carga no condensador. As variáveis do compressor são as que menos contribuem para o modelo, assim como na Árvore de Decisão.

Por fim, uma tabela resumindo a pontuação geral (revocação) para os dados de treino e teste para os três modelos é fornecida abaixo:

Tabela 12 – Comparação dos três modelos

	ÁRVORE DE DECISÃO	**FLORESTA ALEATÓRIA**	**XGBOOST**
DADOS DE TREINO	98%	95%	97%
DADOS DE TESTE	89%	88%	90%

Fonte: Elaborado pela autora.

Os dados de treino possuem uma pontuação bastante próximas, e como esperado, sempre maiores do que a de teste. Para o XGBoost, porém, tem-se a maior pontuação de dados de teste. Já para a floresta aleatória, a pior.

6 CONCLUSÃO

Neste trabalho buscou-se avaliar a eficiência energética de um sistema de água gelada industrial (*Chiller*) utilizando tarefas de mineração de dados. Para isso, foi utilizada uma metodologia iterativa (CRISP-DM), a qual permitiu que ao longo de um ciclo completo na monografia em questão, conclusões fossem obtidas a cada etapa.

Primeiramente, em relação à etapa de tratamento dos dados, de um total de 34 variáveis escolhidas inicialmente, e que se julgavam importantes ao processo no dia-a-dia operacional, apenas 14 foram utilizadas, isto é, menos da metade (41%). Destas, em geral, haviam conjuntos de dados brutos com tamanhos em torno de mais de um milhão de linhas, ao passo que a base final que fora carregada para os três modelos possuía apenas 98.860 linhas, ou seja, apenas 10% do tamanho minerado. Dessa forma, a primeira reflexão obtida é a de que, embora com essa quantidade de dados, ainda se esteja trabalhando com *big data*, em mineração de dados, a qualidade dos dados (*smart data*) deve sempre ser sobreposta a sua quantidade.

Já para essa qualidade ser obtida, passou-se em torno de 90% de tempo do projeto apenas na etapa de tratamento de dados, visto a característica dos dados serem advindas de sinais de instrumentos de medida industrial, e sujeitos a diversos tipos de ruídos. Assim, o retorno à informação de negócio foi feito inúmeras vezes, a fim de se entender a melhor maneira de coletar, agrupar e tratar tais variáveis.

Ainda na fase de tratamento, ao se calcular o Coeficiente Operacional de Performance (COP) do ciclo de refrigeração, notou-se que o consumo energético do motor não variava (NaN>90%), o que posteriormente foi correlacionado ao fato da baixa modulação no volume da câmara do compressor (90-100%) na maioria dos períodos analisados. Ou seja, a variabilidade do principal indicador de eficiência energética dependia apenas da carga térmica do evaporador. No campo prático, esse tipo de análise já indicaria um ofensor energético crônico do sistema, que nesse caso, só poderia ser resolvido com *retrofit*.

Ademais, olhando para a relação de causa e consequência existente entre o sistema de água gelada e a variável do vácuo no processo de desodorização da refinaria, pôde-se concluir pelos testes de hipótese que o COP não possui correlação direta com ela. Isso porque o COP, na prática, mostrou-se ser função apenas da carga térmica do evaporador. Logo, a variabilidade de Q_frio foi dependente exclusivamente das temperaturas de entrada e saída de água do trocador. E, como discutido nos resultados, se o fluido refrigerante (amônia)

não for capaz de absorver energias maiores na entrada, os deltas sempre serão os mesmos, mascarando a real eficiência energética do sistema. Assim, em termos de integração de processos, a temperatura de saída do evaporador é a variável do *Chiller* que possui uma correlação direta com o vácuo, e não o COP.

Em termos de análise exploratória, foi notório apontar também o quanto as variáveis nos períodos analisados estiveram distantes de seus valores nominais. Nos evaporadores e condensadores, as temperaturas de entrada e saída de água estiveram, em média, 2°C acima do valor teórico. Além disso, ao longo dos períodos analisados, houve um aumento das perdas de carga, sobretudo no condensador. Portanto, foi possível abstrair desses sintomas o fato de os trocadores de calor não estarem com sua eficiência de troca térmica plenas, o que implica no aumento gradual das correntes de temperaturas de água, e consequentemente, na baixa eficiência energética do *Chiller*.

Em relação à etapa de modelagem, os três modelos se mostraram satisfatórios para a tarefa de classificar as variáveis de processo dentro das duas regiões operacionais impostas (COP “Regular” e COP “Ruim”), com a precisão e a revocação acima de 80% (com exceção da classe “Ruim” na Árvore de Decisão). Porém, cada um gerou uma ordem de importâncias diferentes para as variáveis de processo. Assim, escolheu-se o modelo com as maiores pontuações de precisão e revocação, no caso o XGBoost. A ordem de importância dos atributos se seguiu com a temperatura de entrada e saída do evaporador, perda de carga no evaporador, temperatura de saída do condensador e perda de carga no condensador. Abaixo dessas, entraram apenas as variáveis de processo do compressor do *Chiller*, as quais desde o início da etapa de entendimento dos dados possuíam uma variabilidade menor, se comparadas com as variáveis dos evaporadores e condensadores.

7 TRABALHOS FUTUROS

Como próximos passos, pretende-se modularizar e tornar os códigos utilizados automáticos, a fim de se inserir novos dados do sistema nos modelos pré implementados, executando a etapa final de implantação do CRISP-DM.

Em termos de eficiência energética, pretende-se, também, utilizar o conhecimento obtido por meio deste projeto para avaliar a performance operacional de outros sistemas de refrigeração da planta. No caso, compressores que modulam de acordo com a carga térmica demandada pela linha de produção, isto é, a eficiência do compressor é medida a partir de uma variável externa, e não o contrário, como ocorre no sistema de água gelada do *Chiller*. Por fim, em termos de ganhos, a partir das informações obtidas com a mineração de dados, se tornará possível melhorar a saúde do processo, estabelecendo ordem de prioridades de manutenção, testes *in loco*, bem como o projeto de novas variáveis a serem mineradas.

8 REFERÊNCIAS

AHAMED, J. U.; SAIDUR, R.; MASJUKI, H. H. A review on exergy analysis of vapor compression refrigeration system. **Elsevier - Renewable and Sustainable Energy Reviews**, v.15, 2011.

BELINI, R. J. **Avaliação de fatores que afetam a eficiência energética em um sistema de refrigeração**. Dissertação (Mestrado em Processos Industriais na área de Desenvolvimento e Otimização em Processos Industriais) – Instituto de Pesquisas Tecnológicas (IPT), São Paulo, 2019.

BIRD, R.B.; STEWART, W.E.; LIGHTFOOT, E.N. **Transport Phenomena.** 1 Ed. New York: John Wiley & Sons, 1960.

BRADLEY, W. et al. Perspectives on the integration between first-principles and data-driven modelling. **Elsevier - Computers & Chemical Engineering**, n.166, 2022.

BREIMAN, L. Random Forests. **Machine Learning**, v.45, 2001.

CARDOSO, E. S.; PRIETO, M. D.; KAMPOUROPOULO, K.; ROMERAL, L. Predictive chiller operation: A data-driven loading and scheduling approach. **Elsevier - Energy and Buildings,** v.208, 2020.

CHANG, K. et al. Optimizing the energy efficiency of chiller systems in the semiconductor industry through big data analytics and an empirical study. **Elsevier - Journal of Manufacturing Systems**, v.60, 2021.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. **ResearchGate**, 2016.

FAWCETT, T.; PROVOST, F. **Data Science para Negócios**. 1. Ed. Rio de Janeiro: Alta Books, 2016.

FREUND, Y.; SCHAPIRE R. E. AdaBoost Algorithm. Disponível em: LNAI 94 (1995).

GANTZ, C. Refrigeration: a history. North Carolina: McFarland and Company, 2015.

HARRISON, M. **Machine Learning:** Guia de Referência Rápida. Trabalhando com dados estruturados em Python. 1. Ed. São Paulo: Novatec Editora Ltda, 2020.

KORETSKY, M. D. **Termodinâmica para Engenharia Química**. 7. Ed. Rio de Janeiro: LTC, 2007.

LI, M; JU, Y. The analysis of the operating performance of a chiller system based on hierarchal cluster method. **Elsevier: Energy and Buildings**, v.138, 2017.

MINISTÉRIO DE MINAS E ENERGIA. **Balanço Energético Nacional 2021:** Ano base 2020. Rio de Janeiro: Empresa de Pesquisa Energética (EPE), 2021.

MONTGOMERY, D. C., RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros.** 2. Ed. Rio de Janeiro: LTC, 2003.

RAEDER, T. et al. Design principles of massive, robust prediction systems. **ResearchGate - Proceedings of the 18th ACM SIGKDD International Conference of on Knowledge Discovery and Data Mining**, 2012.

SAMUEL, A. L.. Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal of Research and Development,** v.3, n.3, 1959.

SANSANA, J. et al. Recent trends on hybrid modeling for Industry 4.0. **ELSEVIER – Computers & Chemical Engineering,** v.151, 2021.

SMITH, J. M; VAN NESS, H. C; ABBOTT, M. M. **Introdução à Termodinâmica da Engenharia Química.** 7. Ed. Rio de Janeiro: LTC, 2007.

STOECKER, W. F. **Industrial refrigeration handbook**. 1. Ed. New York: McGraw-Hill, 1998.

SWAIN, P. H.; HAUSKA, H. The Decision Tree Classifier: Design and Potential. **IEEE Transactions on Geoscience Electronics**, v.3, 1977 .

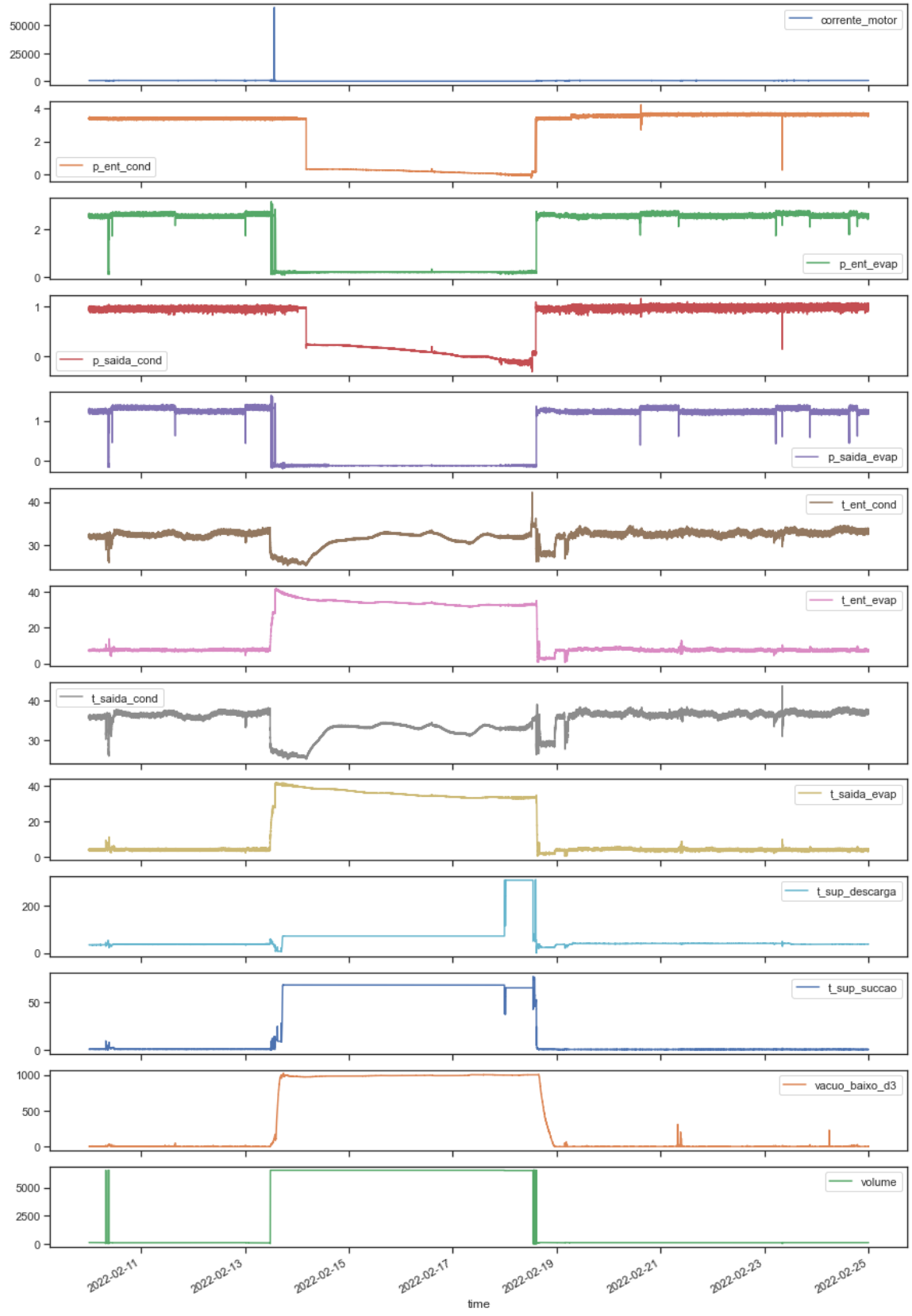
VARBANOV, P. et al. Top-level Analysis of Site Utility Systems. **ELSEVIER - Chemical Engineering Research & Design**, v.82, 2008.

WIRTH, R.; HIPP, J. CRISP-DM: Towards a standard process model for data mining. **ResearchGate - Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining,** 2000.

ZHAO, L.; YOU, F. A data-driven approach for industrial utility systems optimization under uncertainty. **Elsevier – Energy**, v.182, 2019.

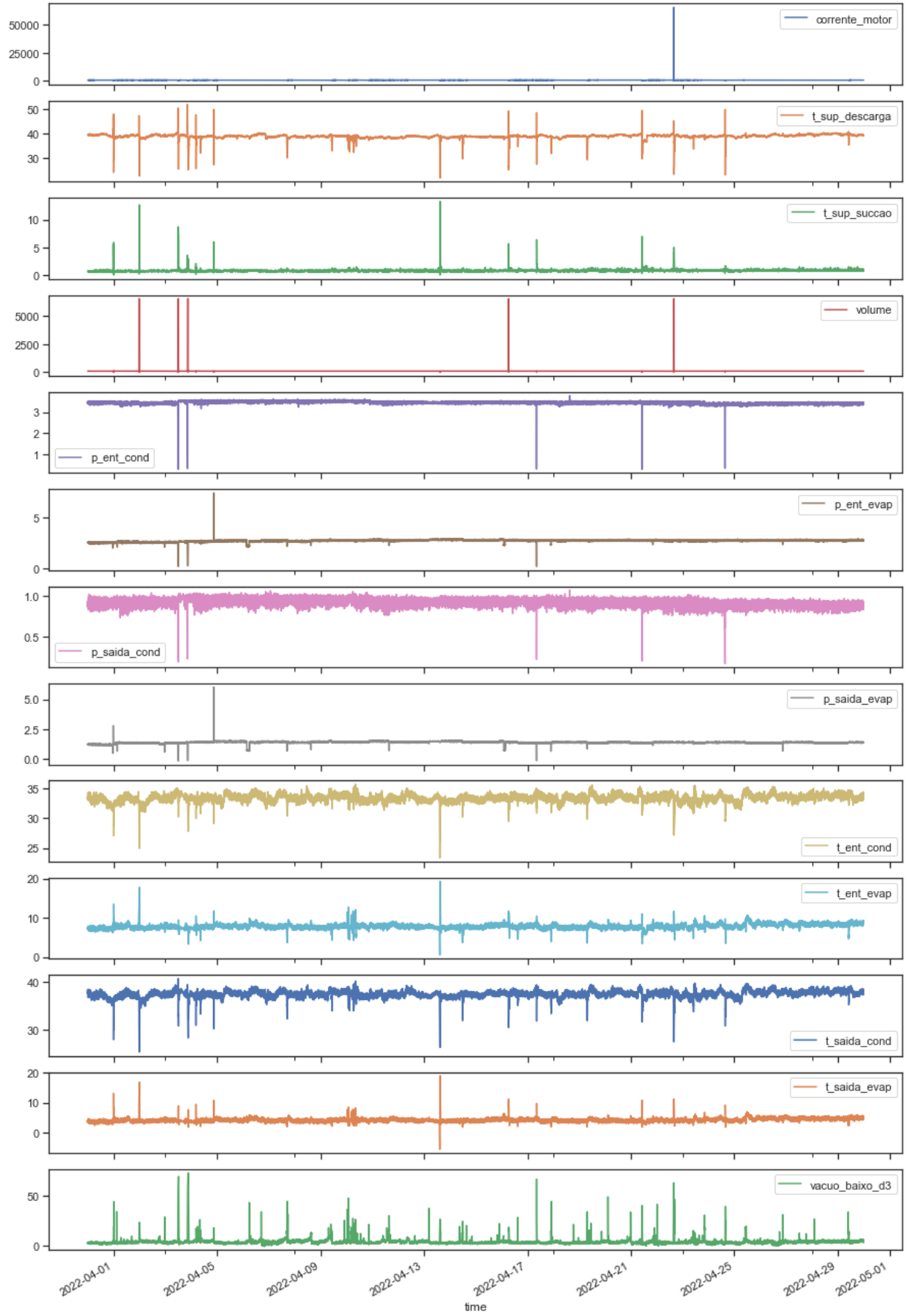
APÊNDICE A – VISUALIZAÇÃO DA SÉRIE TEMPORAL BRUTA

Figura 43 - Comportamento das variáveis de processo (brutas) do Chiller em fevereiro/2022



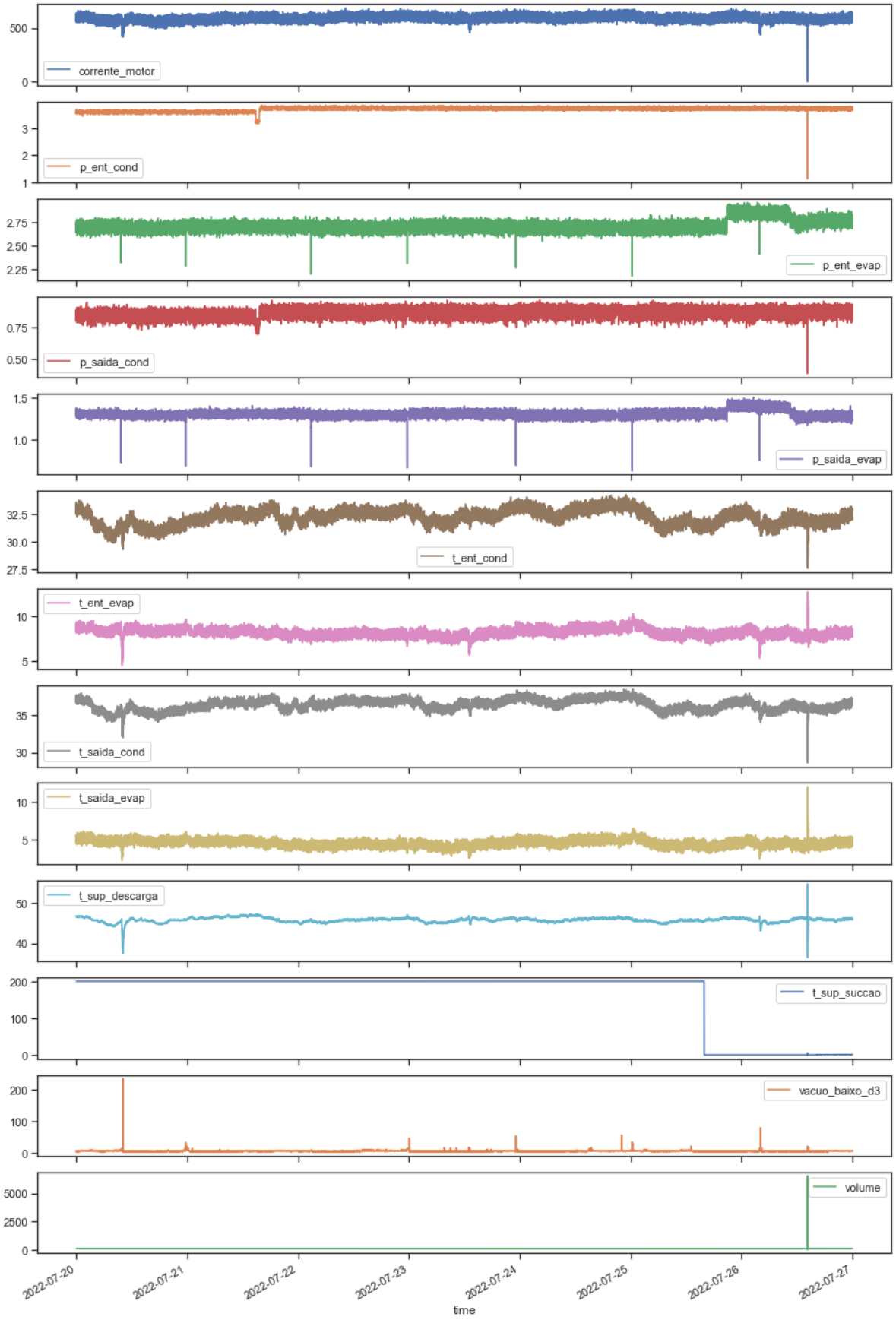
Fonte: Elaborado pela autora.

Figura 44 - Comportamento das variáveis de processo (brutas) do Chiller em abril/2022



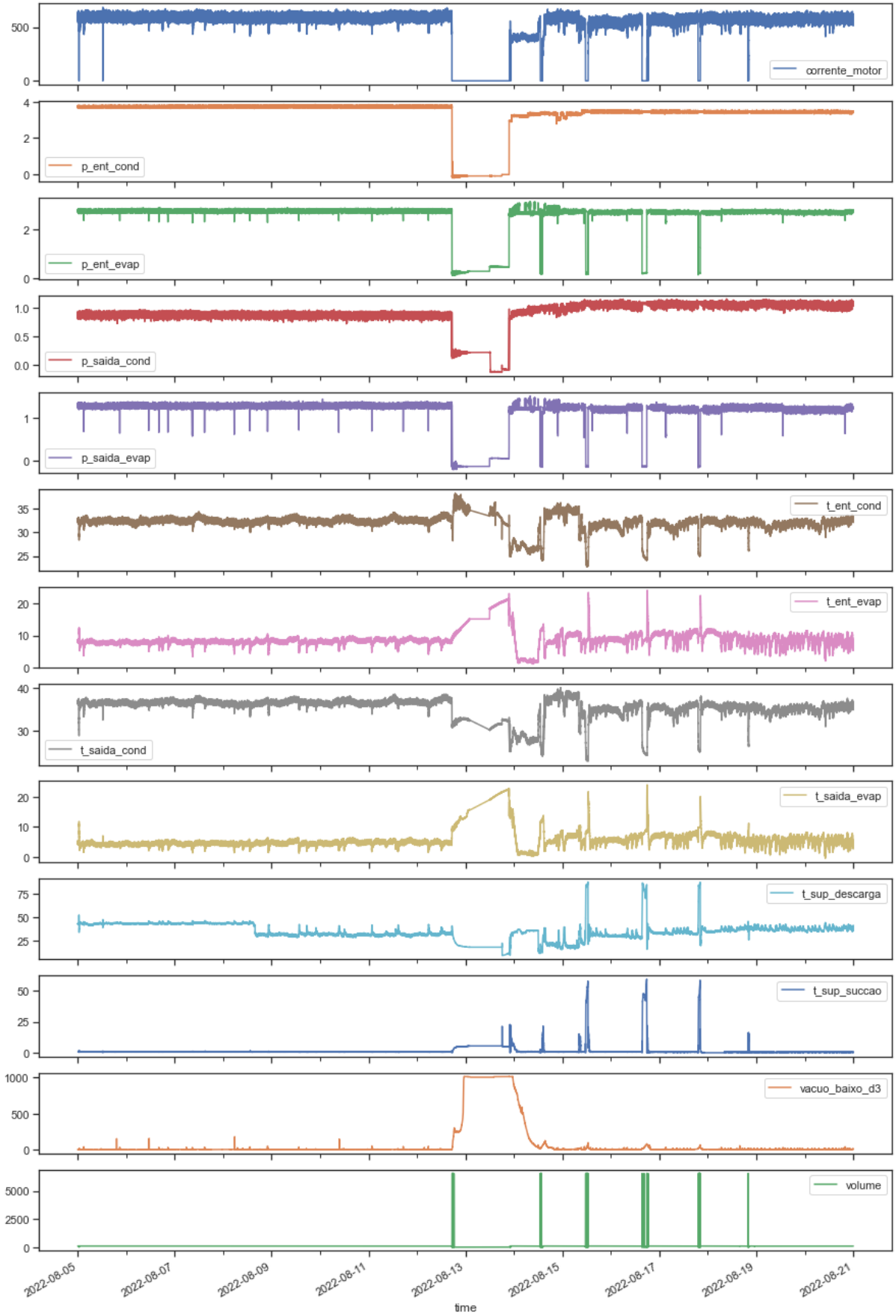
Fonte: Elaborado pela autora.

Figura 45 - Comportamento das variáveis de processo (brutas) do Chiller em julho/2022



Fonte: Elaborado pela autora.

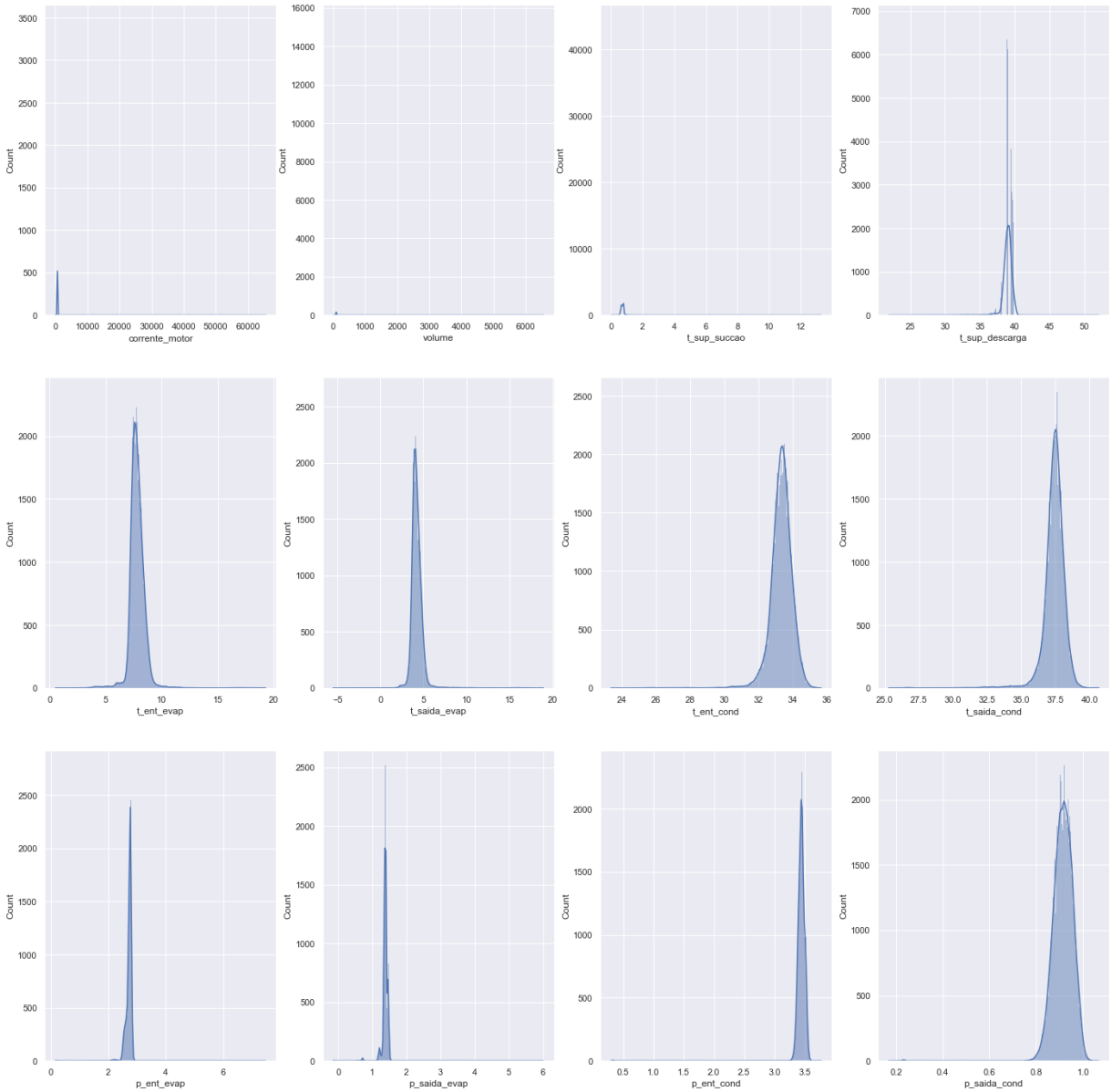
Figura 46 - Comportamento das variáveis de processo (brutas) do Chiller em agosto/2022



Fonte: Elaborado pela autora.

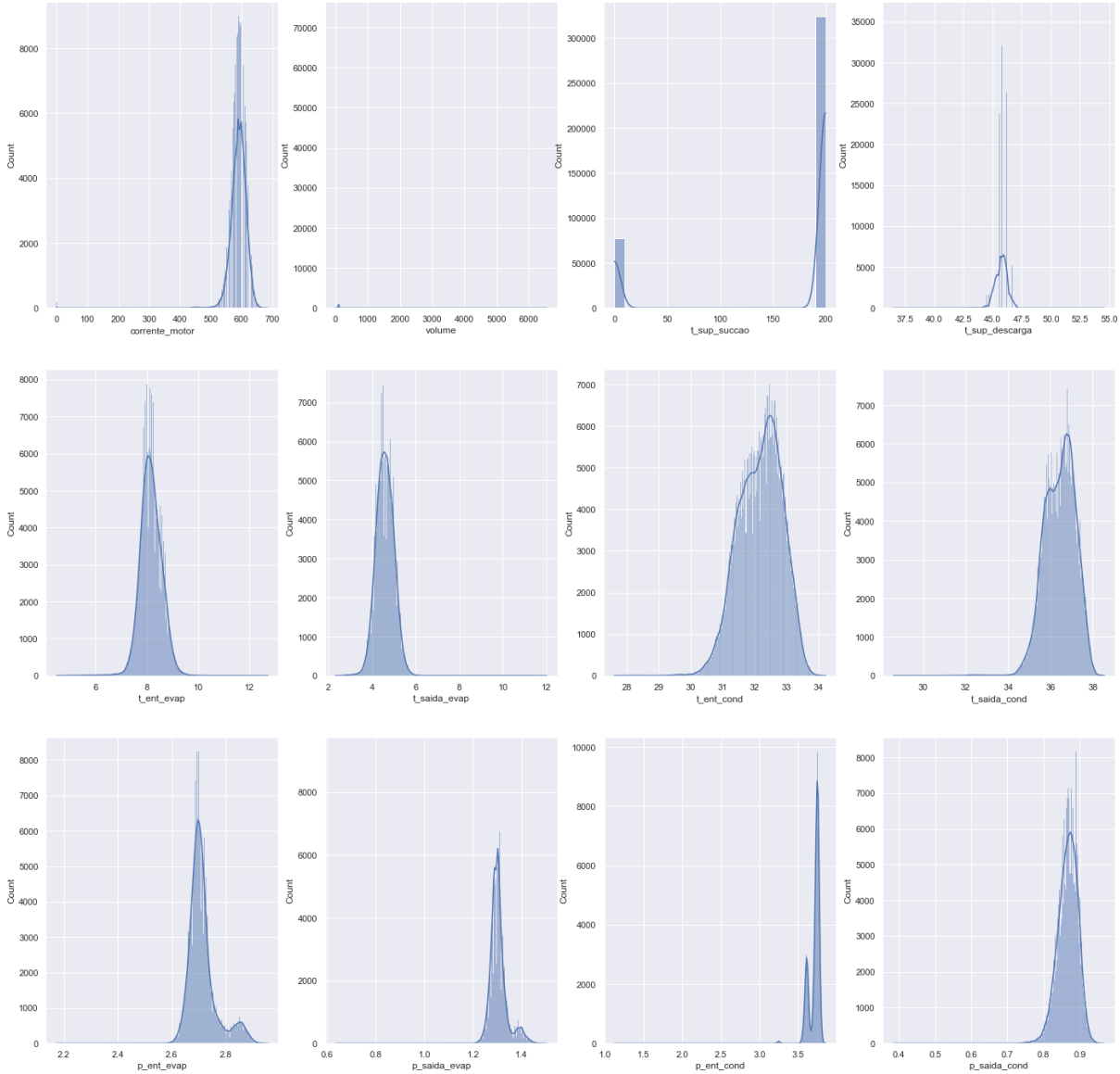
APÊNDICE B – DISTRIBUIÇÃO DOS DADOS BRUTOS

Figura 47 - Distribuição das variáveis de processo (brutas) do Chiller em abril/2022



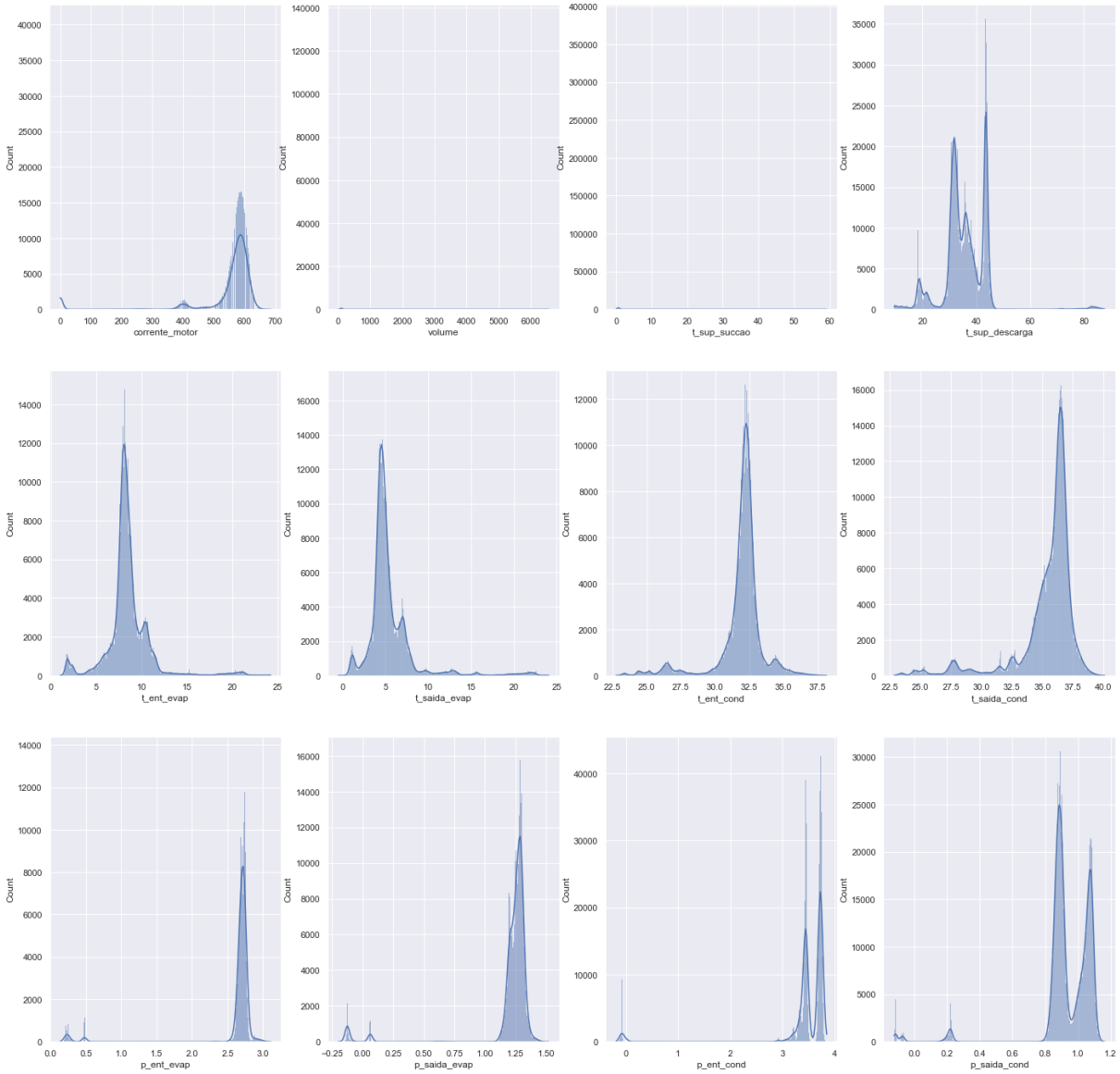
Fonte: Elaborado pela autora.

Figura 48 - Distribuição das variáveis de processo (brutas) do Chiller em julho/2022



Fonte: Elaborado pela autora.

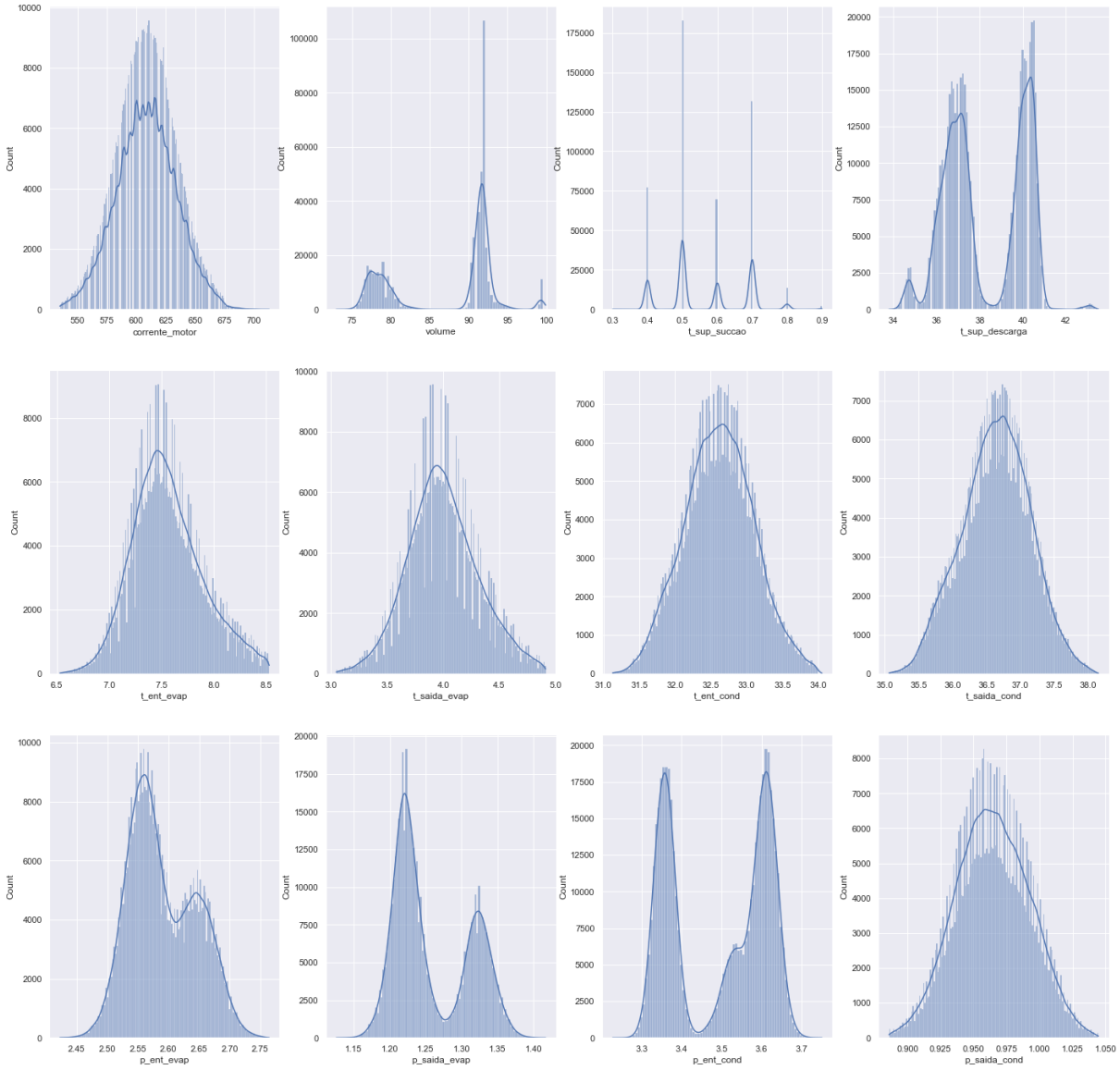
Figura 49 - Distribuição das variáveis de processo (brutas) do Chiller em agosto/2022



Fonte: Elaborado pela autora.

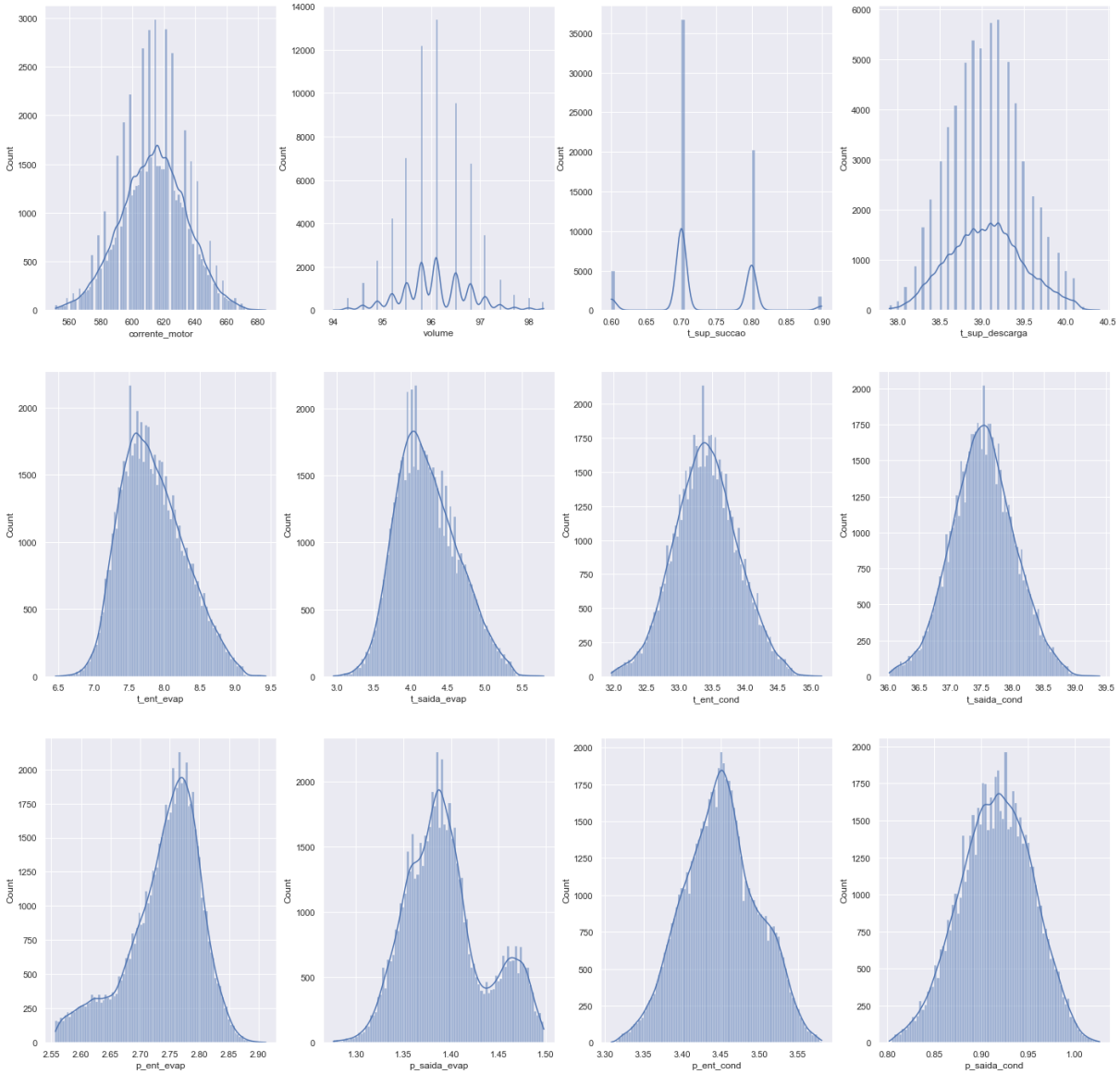
APÊNDICE C – DISTRIBUIÇÃO DOS DADOS TRATADOS

Figura 50 - Distribuição das variáveis de processo (tratadas) do Chiller em fevereiro/2022



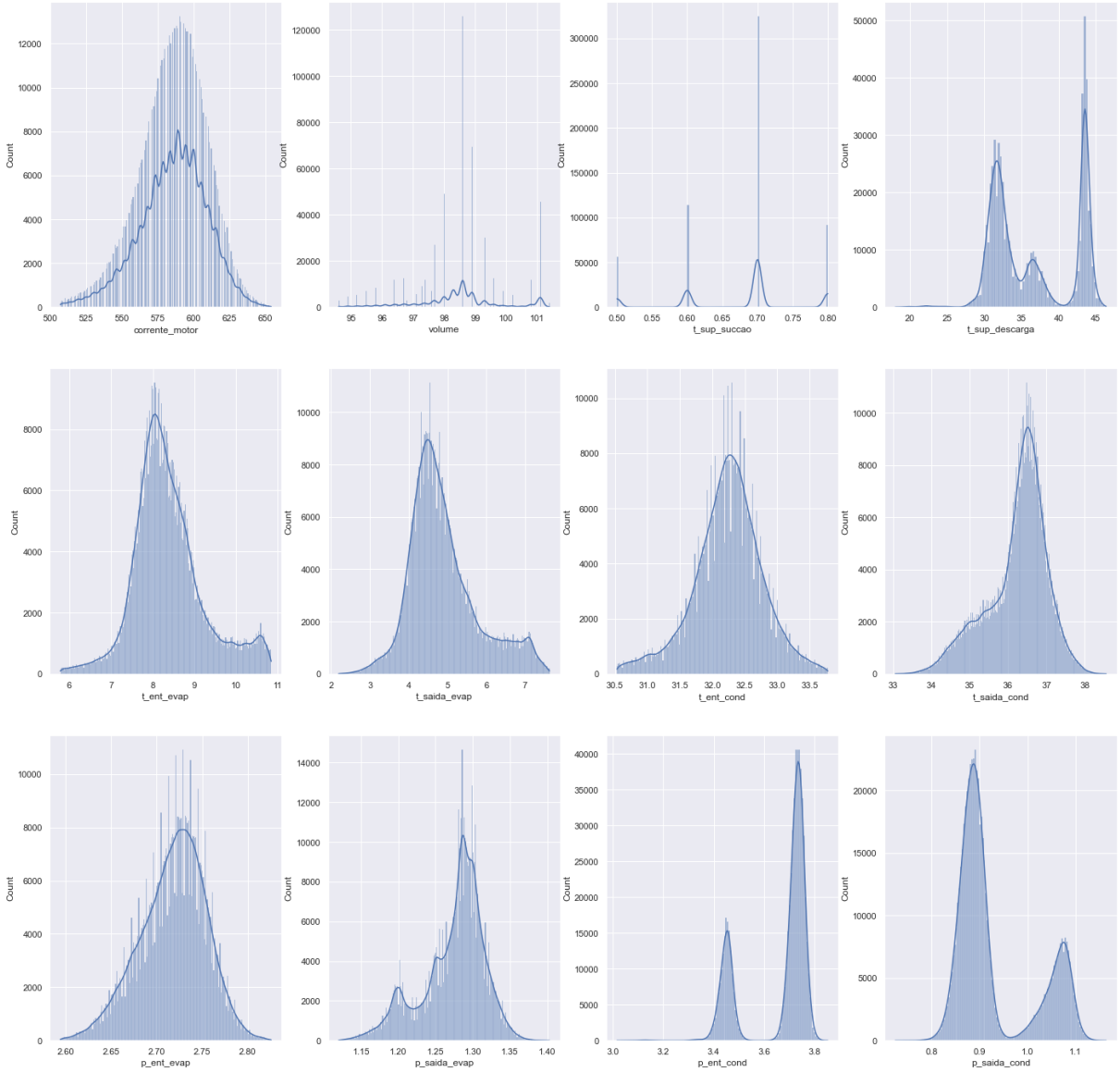
Fonte: Elaborado pela autora.

Figura 51 - Distribuição das variáveis de processo (tratadas) do Chiller em abril/2022



Fonte: Elaborado pela autora.

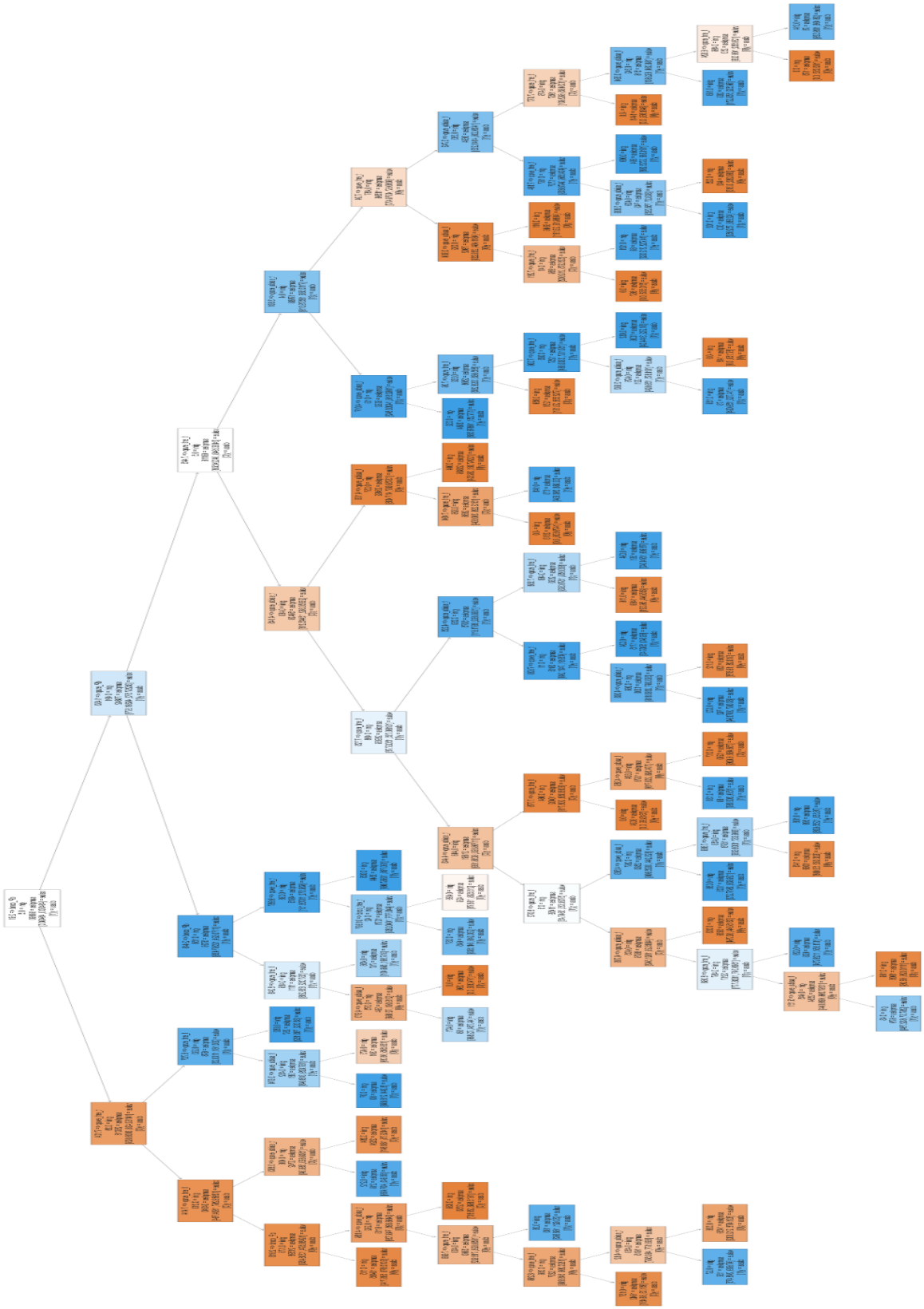
Figura 52 - Distribuição das variáveis de processo (tratadas) do Chiller em agosto/2022



Fonte: Elaborado pela autora.

APÊNDICE D – ÁRVORE DE DECISÃO

Figura 53 – Árvore de Decisão



Fonte: Elaborado pela autora.