UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

# Aprimoramento do controle operacional em estação de tratamento de efluentes: Avaliação comparativa de algoritmos de aprendizado de máquina

**Thiago da Silva Ribeiro**

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
**Instituto de Ciências Matemáticas e de Computação**

# Aprimoramento do controle operacional em estação de tratamento de efluentes: Avaliação comparativa de algoritmos de aprendizado de máquina

***Thiago da Silva Ribeiro***

USP - São Carlos

2023

Thiago da Silva Ribeiro

# Aprimoramento do controle operacional em estação de tratamento de efluentes: Avaliação comparativa de algoritmos de aprendizado de máquina

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Aprendizado de máquina

Orientador: Prof. Dr. Ricardo Araújo Rios

USP - São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

R484a

Ribeiro, Thiago
Aprimoramento do controle operacional em estação de tratamento de efluentes: Avaliação comparativa de algoritmos de aprendizado de máquina / Thiago Ribeiro; orientador Ricardo Rios. -- São Carlos, 2023.
47 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Detecção de Falha. 2. Estação de Tratamento de Efluente. 3. PyCaret. 4. Aprendizado de Máquina. I. Rios, Ricardo, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

# DEDICATÓRIA

*Adão Ribeiro (in memoriam)*

# AGRADECIMENTOS

Expresso minha profunda gratidão a Deus pela dádiva da vida. Este projeto só se concretizou devido ao inestimável apoio que recebi ao longo dessa jornada.

Quero estender meu reconhecimento ao meu orientador, Professor Ricardo, cujo suporte foi essencial para a conclusão deste trabalho.

Gostaria também de abranger os meus agradecimentos a todos os professores do curso pelo compartilhamento de conhecimento, o qual me tornou um profissional mais capacitado e preparado para os desafios futuros. Em particular, expresso minha gratidão às Professoras Roseli e Solange por terem depositado confiança em mim.

Agradeço às amizades cultivadas durante meus estudos por terem tornado esta jornada muito mais enriquecedora.

# EPÍGRAFE

*Essencialmente, todos os modelos estão errados, mas alguns são úteis.*

George Box

# RESUMO

RIBEIRO, T. S. **Aprimoramento do controle operacional em estação de tratamento de efluentes: Avaliação comparativa de algoritmos de aprendizado de máquina**. 2023. 47 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A eletrocoagulação é um método emergente de tratamento de efluentes que combina os benefícios da coagulação, flotação e eletroquímica. Devido à complexidade inerente dos processos associados às estações de tratamento de efluentes, é difícil responder rapidamente e corretamente às circunstâncias dinâmicas necessárias para garantir a qualidade do efluente. Portanto, este trabalho tem como objetivo identificar a condição operacional de uma estação de tratamento de efluentes que adotou a eletrocoagulação como método de tratamento. Duas condições operacionais, baseadas na clarificação do efluente e no lodo de reação, foram a variável-alvo. Foram monitoradas onze variáveis, como condutividade, pH, voltagem, corrente, polaridade e potencial de oxidação-redução. Diversos algoritmos de aprendizado de máquina foram testados utilizando a biblioteca PyCaret 3.2.0 no Google Colaboratory para analisar seu desempenho. Os modelos que obtiveram os maiores valores de F1-score no treinamento foram regressão logística, floresta aleatória, XGBoost e SVM com kernel radial, todos com pontuações médias acima de 0,93. Destes, o XGBoost se destacou ao mostrar uma baixa taxa de erro do tipo I e um nível aceitável de erros do tipo II, evidenciando sua capacidade de minimizar falsos positivos. O modelo de floresta aleatória demonstrou a maior precisão, seguido de perto pelo XGBoost. Este último apresentou o segundo melhor desempenho em recall, sendo superado apenas pela regressão logística. Além disso, o XGBoost exibiu o mais alto valor de acurácia. Assim, o estudo conclui que o XGBoost e a floresta aleatória se destacaram como modelos promissores para prever a eficácia operacional, com o XGBoost mostrando um ótimo equilíbrio.

Palavras-chave: Aprendizado de Máquina; Detecção de Falha; Estação de Tratamento de Efluente; PyCaret; Regressão Logística; Floresta Aleatória; XGBoost; SVM

# ABSTRACT

RIBEIRO, T. S. **Enhancement of operational control in wastewater treatment plant: Comparative evaluation of machine learning algorithms**. 2023. 47 f. Course completion work (MBA in Artificial Intelligence and Big Data) - Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, 2023.

Electrocoagulation is an emerging method for treating wastewater, combining the advantages of coagulation, flotation, and electrochemistry. Due to the inherent complexity of processes associated with wastewater treatment plants, promptly and accurately responding to the dynamic circumstances necessary to ensure effluent quality is challenging. Therefore, this study aims to identify the operational condition of a wastewater treatment plant that has adopted electrocoagulation as a treatment method. Two operational conditions, based on effluent clarification and reaction sludge, were the target variable. Eleven variables were monitored, including conductivity, pH, voltage, current, polarity, and oxidation-reduction potential. Various machine learning algorithms were tested using the PyCaret 3.2.0 library in Google Colaboratory to assess their performance. The models achieving the highest F1-score values during training were logistic regression, random forest, XGBoost, and radial kernel SVM, all with mean scores above 0.93. Among these, XGBoost stood out by displaying a low rate of Type I errors and an acceptable level of Type II errors. The random forest model exhibited the highest precision, closely followed by XGBoost, which showed the second-best recall performance, surpassed only by logistic regression. Additionally, XGBoost displayed the highest accuracy value. Consequently, the study concludes that XGBoost and random forest emerged as promising models for predicting the operational efficiency of the wastewater treatment plant, with XGBoost demonstrating a remarkable balance and achieving good overall precision.

Keywords: Machine Learning; Fault Detection; Wastewater Treatment Plant; PyCaret;

Logistic Regression; Random Forest; XGBoost; SVM

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS E SIGLAS

| | | |
|---|---|---|
| EC | – | Eletrocoagulação |
| ETE | – | Estação de Tratamento de Efluente |
| ETEs | – | Estações de Tratamento de Efluentes |
| ORP | – | Potencial de Oxidação-Redução |
| SVM | – | Máquina de Vetores de Suporte |
| XGBoost | – | Extreme Gradient Boosting |
| DAF | – | Flotação por Ar Dissolvido |
| SMOTE | – | Técnica de Oversampling Sintético de Minoria |
| AUC | – | Área Sob a Curva ROC |
| MCC | – | Coeficiente de Correlação de Matthews |

# SUMÁRIO

## 1. INTRODUÇÃO

Não apenas nossa existência depende da água, mas também nosso bem-estar econômico. A água desempenha um papel em todos os nossos processos de produção. Não há substitutos, e embora seja renovável, seu suprimento é limitado. Hoje, todos estão preocupados com a possibilidade de escassez de água diante das crescentes necessidades, em grande parte impulsionadas pela população, bem como as repercussões que isso pode ter em nossa produção de energia e alimentos (Diaz-Elsayed et al., 2019).

Múltiplos fatores contribuem para a poluição da água, incluindo resíduos industriais, operações de mineração, esgoto, fertilizantes químicos, uso de energia, entre outros (Mousazadeh et al., 2021). Esforços constantes devem ser empreendidos para proteger os suprimentos de água nesta situação.

Em geral, os desafios enfrentados durante o tratamento de águas residuais são bastante complicados, uma vez que o efluente compreende vários tipos de contaminantes com base em sua fonte. Consequentemente, existem vários tipos de efluentes a serem tratados, cada um com suas próprias propriedades que necessitam de técnicas de tratamento exclusivas (Crittenden et al., 2012).

Dentre as técnicas de tratamento existentes, a eletrocoagulação (EC) é uma abordagem potencial para o tratamento efluentes devido aos seus custos operacionais mais baixos, design simples, sedimentação rápida, pouca ou nenhuma adição de produtos químicos e baixa produção de lodo (Das; Sharma; Purkait, 2022). Hoje, o principal objetivo de engenheiros químicos e ambientais é projetar uma estação de tratamento de efluente (ETE) que possa permitir o tratamento descentralizado (Alabi; Telukdarie; Van Rensburg, 2019). Nesse sentido, a EC é uma técnica descentralizada eficaz.

Nos últimos anos, as estações de tratamento de efluentes (ETEs) têm sido expostas a uma quantidade sem precedentes de dados, resultantes da queda nos custos dos sensores, da crescente prevalência da conectividade sem fio e da proliferação de dispositivos móveis capazes de coletar continuamente dados e realizar cálculos complexos (Kijak, 2021).

A crescente automação das ETEs permite o acesso a dados em massa e o desenvolvimento de soluções orientadas por dados (Newhart et al., 2019). A maioria dos sistemas de detecção de falhas é baseada em dados, pois podem identificar rapidamente circunstâncias anormais, são mais simples de implementar e exigem menos conhecimento prévio (Md Nor; Che Hassan; Hussain, 2020).

Os recentes avanços no monitoramento de processos e desempenho baseados em dados podem oferecer à essa indústria uma oportunidade de reduzir custos e melhorar as operações. Devido à não linearidade e à crescente complexidade do processo, as técnicas de detecção de falhas baseadas em dados estão em alta demanda (Dairi et al., 2019).

Este estudo tem como objetivo identificar a condição operacional de uma ETE que adotou a EC como método de tratamento. Duas condições operacionais baseadas na clarificação do efluente e no lodo de reação foram consideradas como variável-alvo. Onze features, incluindo condutividade, pH, voltagem, corrente, polaridade e potencial de oxidação-redução (ORP), foram monitoradas. Diferentes algoritmos foram empregados e o desempenho dos modelos foi comparado.

A Figura 1 representa a visualização em rede que apareceu nas publicações obtidas pelo banco de dados Scopus usando as palavras-chave EC e modelagem, e que foram publicadas entre 2004 e 2022. O mapa de nuvem exibe a frequência das palavras-chave nos artigos, bem como suas relações. Cada cor reflete um conjunto de termos que foram agrupados em clusters. Como pode ser observado, os métodos estatísticos tradicionais desempenham um papel significativo na modelagem de EC. Métodos que empregam algoritmos de aprendizado de máquina ainda são escassos na literatura avaliada; portanto, existe uma lacuna de conhecimento.

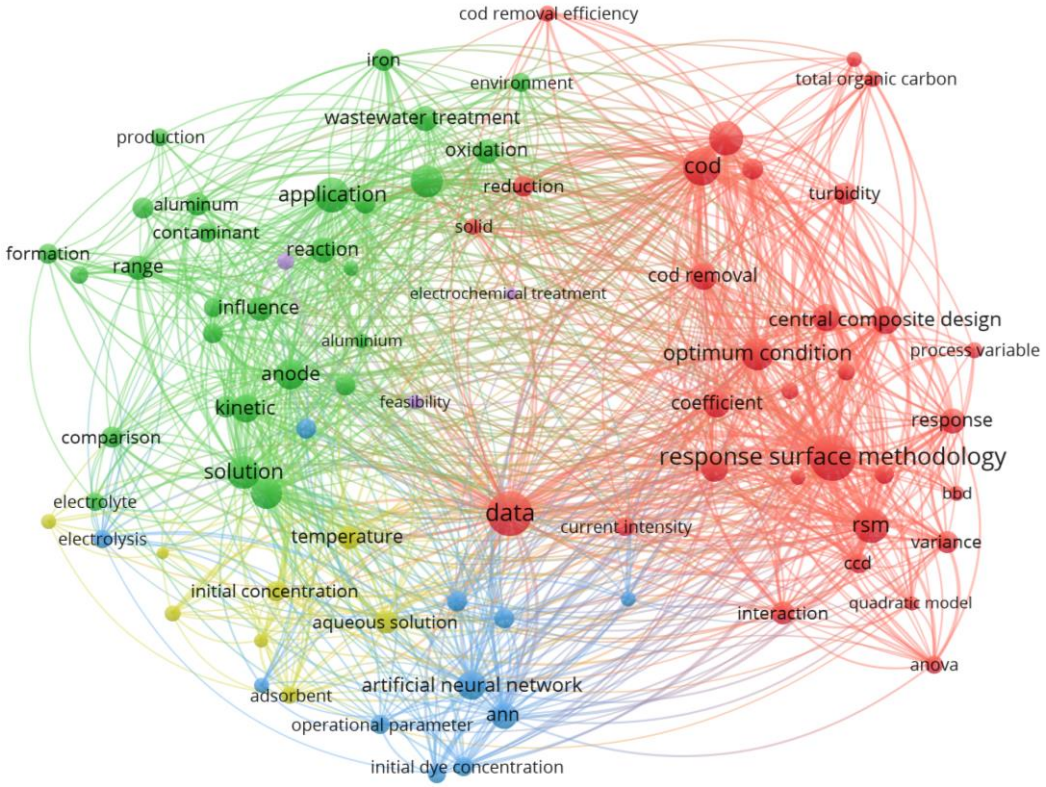

Figura 1 – Visualização em rede da pesquisa de modelagem de EC (gerada usando o software *VOSviewer*) (Fonte: Autoral, 2023).

Assim, a principal contribuição apresentada neste trabalho é abordar uma lacuna de conhecimento na detecção de falhas orientada por dados em ETEs baseadas em EC.

## 2. REVISÃO BIBLIOGRÁFICA

### 2.1. Aprendizagem de máquina em ETEs

Inteligência artificial e aprendizado de máquina são tecnologias/métodos utilizados no setor de águas para proporcionar vantagens nos processos de tratamento de água, identificando antecipadamente problemas e notificando em tempo real o engenheiro responsável. A Internet das Coisas é uma tecnologia potencial para a indústria da água, pois permite o monitoramento remoto em tempo real (Yasin et al., 2021). Espera-se que a tecnologia de big data e análise transforme essa indústria em um setor totalmente orientado por dados (Alabi; Telukdarie; Van Rensburg, 2019).

A inteligência artificial emergente e o aprendizado de máquina, em conjunto com tecnologias inteligentes, estão preenchendo uma lacuna em aplicações que anteriormente era negligenciada pelos métodos convencionais e formas de pensar (Lowe; Qin; Mao, 2022). Antecipa-se que a inteligência artificial, aprendizado de máquina e tecnologias inteligentes possam modelar e resolver desafios complexos nos processos de tratamento de água, devido à sua generalização, robustez e relativa simplicidade de projeto, visando a redução de custos e a melhoria das operações. Aplicações que têm feito uso significativo de aprendizado de máquina incluem tratamento de água e efluentes, monitoramento de sistemas naturais e agricultura de precisão (Zhao et al., 2020).

A detecção de falhas é um uso proeminente de algoritmos de aprendizado de máquina em ETEs. A seção a seguir destacará alguns dos conceitos mais relevantes desse campo.

### 2.2. Detecção de falhas em ETEs

Uma falha é a divergência indesejada de pelo menos uma característica distintiva de um sistema de seu estado normal, aceitável ou padrão. Há um crescente interesse em criar soluções para lidar com falhas que ocorrem em processos industriais, garantindo assim saídas seguras e

eficientes (Li et al., 2020). A detecção de falhas é a técnica essencial para resolver esse problema.

A detecção de falhas pode ser realizada por métodos de primeiros princípios, baseados em dados ou baseados em conhecimento. Métodos de primeiros princípios exigem o desenvolvimento de um modelo matemático com base em entendimento teórico. Este método é frequentemente ineficaz devido à complexidade do modelo matemático gerado. O método baseado em conhecimento, por outro lado, requer compreensão prévia ou conhecimento das conexões entre falhas e parâmetros ou estados do modelo. Também é difícil aplicar este método a sistemas em grande escala devido ao tempo e expertise necessários para criar esses modelos complexos de falhas (Venkatasubramanian et al., 2003).

Devido à não linearidade e à crescente complexidade da indústria de processos contemporânea, técnicas orientadas por dados estão em demanda. As tecnologias de banco de dados e mineração de dados fornecem suporte tecnológico confiável para o desenvolvimento de técnicas de modelagem orientadas por dados em processos industriais (Md Nor; Che Hassan; Hussain, 2020).

Abordagens analíticas orientadas por dados dependem substancialmente do tipo de dados obtidos. É essencial compreender a estrutura e a qualidade dos dados para decidir sobre a organização e uso dos mesmos. Uma instalação de ETE coleta dados de várias fontes, incluindo análises laboratoriais, leituras de sensores online, gerenciamento de operações e manutenção, dados de clientes e fabricantes de tecnologia (Newhart et al., 2019).

Historicamente, as ETEs careciam de gestão de processos orientada por dados, com as decisões operacionais diárias sendo consideradas mais uma arte do que uma ciência. Apesar dos desafios específicos apresentados, a automação de sistemas orientada por dados e o controle em tempo real são essenciais para o funcionamento de uma ETE contemporânea. Para diminuir o impacto de uma falha na qualidade da água do efluente, os operadores da planta devem estar prontos para reagir rapidamente a uma falha no sistema a fim de evitar danos aos equipamentos ou falhas no sistema (Mamandipoor et al., 2020).

### 2.3. Algoritmos de aprendizado de máquina

A aprendizagem de máquina é um subcampo da abordagem de inteligência artificial que permite que sistemas adquiram conhecimento automaticamente, sem programação explícita. Esse processo começa com a análise dos dados e a busca por padrões para tomar decisões melhores (Alom et al., 2019).

Com o uso de rótulos, algoritmos supervisionados de aprendizado de máquina utilizam as informações adquiridas de dados passados e atuais para prever ocorrências futuras. Essa estratégia inicia-se com o treinamento de um conjunto de dados, após o qual o algoritmo desenvolve uma função inferida para prever os valores de saída. Com um procedimento de treinamento suficiente, o sistema é capaz de fornecer resultados com base nos dados de entrada. O método para aprendizado de máquina compara os resultados gerados com os resultados reais e previstos para encontrar erros e ajustar o modelo conforme necessário (Sarker, 2021).

Uma versão clássica da tarefa de aprendizado de máquina supervisionado é o problema de classificação, no qual o modelo deve estimar o comportamento de uma função que mapeia um vetor em uma das classes ao observar várias amostras de entrada-saída da função (Nasteski, 2017). Na próxima seção, serão revisados alguns dos algoritmos de aprendizado de máquina supervisionado.

### 2.3.1. Máquina de Vetores de Suporte (SVM)

A noção do SVM foi herdada da rede neural artificial, ou pode-se argumentar que o SVM é a extensão matemática da mesma. O SVM realiza classificação ao traduzir os dados iniciais de treinamento para um espaço multidimensional e gerar um hiperplano com dimensões maiores. O SVM é uma estratégia eficaz de aprendizado matemático baseada em hiperplanos. O algoritmo busca por pontos vetoriais, conhecidos como vetores de suporte, que definem o limite de decisão e proporcionam uma separação marginal significativa entre as classes (Figura 2). No plano de decisão, o SVM distingue classes com a maior distância marginal (Wang, 2005).

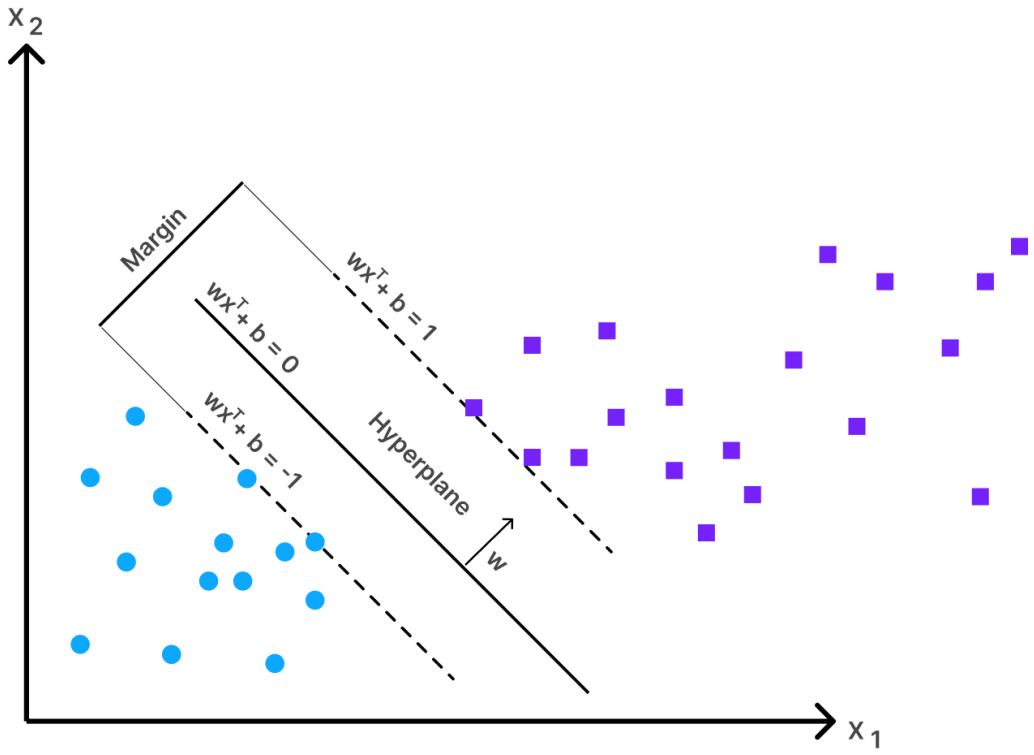


Figura 2 – Modelo Linear SVM (Fonte: Autoral, 2023).

A separação (margem) entre as fronteiras de decisão é maximizada em um espaço altamente dimensional. Ao determinar funções de decisão diretamente a partir dos dados de treinamento, esta abordagem de classificação reduz as imprecisões de classificação dos dados de treinamento e melhora a capacidade de generalização (Chauhan; Dahiya; Sharma, 2019).

A seguir está a equação geral para o hiperplano adicional (Equação 1):

$$y_i(w \cdot x_i - b) \geq 1, \forall\ 1 \leq i \leq n \qquad (1)$$

onde $w$ representa o vetor normal, $b$ representa o viés, $\cdot$ representa o produto escalar, e $x_i$ representa o vetor dimensional que deve ser categorizado em $y_i$.

No modelo SVM linear, dois tipos de parâmetros devem ser otimizados: o fator de penalidade C (C > 0) e os parâmetros da função do kernel, que podem ser lineares, polinomiais ou funções de base radial. C é um parâmetro fixo e ajustável que determina a severidade da penalização em caso de amostras incorretas (Yang; Li; Yang, 2015).

O SVM está ganhando popularidade porque tem uma base matemática sólida e parece ter um bom desempenho em várias aplicações do mundo real (Chauhan; Dahiya; Sharma, 2019).

### 2.3.2. Extreme Gradient Boosting (XGBoost)

O XGBoost é uma ferramenta poderosa no campo do aprendizado supervisionado, permitindo, em muitas situações, um ótimo desempenho de classificação. Em sua base, o XGBoost é um algoritmo de boosting de árvore de decisão. O boosting é uma estratégia de aprendizado em conjunto que envolve o desenvolvimento de muitos modelos em uma ordem sequencial, sendo que cada novo modelo visa corrigir erros do modelo anterior. Cada modelo adicional adicionado ao conjunto é uma árvore de decisão (Figura 3). Uma técnica de gradiente descendente é utilizada para minimizar a perda neste tipo de procedimento de boosting (Ferreira; Figueiredo, 2012).

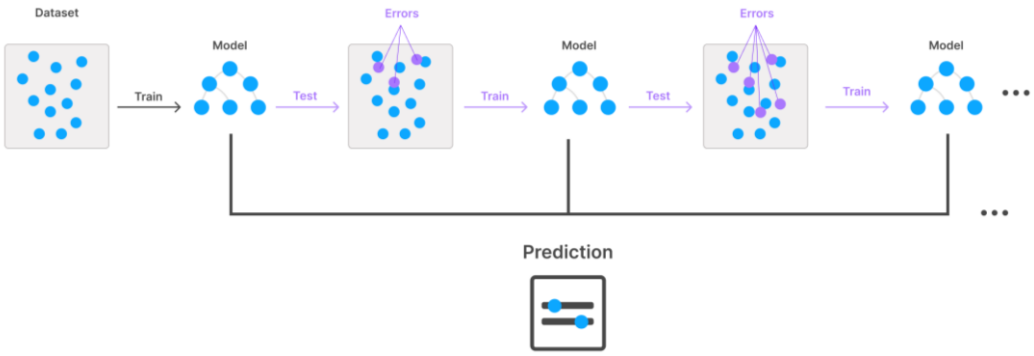


Figura 3 – Disposição da arquitetura do XGBoost (Fonte: Autoral, 2023).

O algoritmo soma todos os resultados das K árvores para obter o valor previsto final, $\hat{y}_i$, representado como (Equação 2–3):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \qquad (2)$$

$$F = \{f(x) = w_{q(x)}\}, \quad \left(q: R^m \to T, w \in R^T\right) \qquad (3)$$

onde F representa o conjunto de árvores de decisão, m representa o número de features, f(x) representa uma das árvores e $w_{q(x)}$ representa o peso do nó. O número de nós é representado por T, e a estrutura de cada árvore é representada por q, que mapeia a amostra para o nó correspondente.

O valor previsto do XGBoost é a soma dos valores dos nós de cada árvore. O objetivo do modelo é aprender essas k árvores de forma que a função objetivo possa ser minimizada. A penalização da árvore de decisão é ajustada usando a regularização Ω, que pode prevenir o sobreajuste, e γ é um hiperparâmetro que determina a complexidade do modelo.

A escalabilidade do XGBoost é resultado de inúmeras melhorias significativas nos sistemas e algoritmos. Entre esses avanços, há uma técnica de aprendizado de árvore exclusiva para lidar com dados esparsos e um processo para lidar com pesos no aprendizado. A computação paralela e distribuída acelera o aprendizado (Chen; Guestrin, 2016).

### 2.3.3. Floresta Aleatória

A floresta aleatória é um método de aprendizado de máquina desenvolvido por Breiman (2001) que combina a abordagem de amostragem por bagging com a seleção aleatória de features. Esse método constrói uma coleção de árvores de decisão controlando a variação entre elas.

No processo de construção das árvores individuais na floresta aleatória, a aleatorização é aplicada ao escolher o melhor nó para dividir. Tipicamente, isso é determinado como √F, onde F é o número de features no conjunto de dados (Probst; Wright; Boulesteix, 2019).

O índice utilizado é uma função que mede a impureza dos dados, refletindo a incerteza em relação à ocorrência de um evento, como a determinação do rótulo de classe. No contexto da classificação, esse índice é representado pela equação:

$$\mathrm{Gini(t)} = 1 - \sum_{i=1}^{N} P(C_i|t)^2 \qquad (4)$$

onde $t$ é uma condição, $N$ é o número de classes no conjunto de dados e $C_i$ é o i-ésimo rótulo de classe.

Breiman (2001) demonstrou que a taxa de erro da floresta aleatória está relacionada à correlação e à força entre as árvores. Aumentar a correlação entre as árvores aumenta a taxa de erro da floresta, enquanto árvores com baixa taxa de erro representam classificadores fortes. Reduzir a correlação e aumentar a força individual das árvores contribui para diminuir o erro na classificação.

A construção de uma floresta aleatória envolve o uso de um grande número de árvores de decisão não podadas, onde suas saídas são combinadas por uma votação majoritária de classes. Para criar árvores precisas, são introduzidos processos de randomização no algoritmo de indução das árvores (Probst; Wright; Boulesteix, 2019):

1. Amostragem bootstrap das instâncias dos dados de treinamento, permitindo que cada árvore seja treinada em diferentes amostras retiradas com reposição do conjunto original;

2. Em vez de escolher a melhor divisão entre todos os atributos, o algoritmo seleciona aleatoriamente um subconjunto de atributos para determinar a melhor divisão entre eles. O tamanho do subconjunto é recomendado ser aproximadamente $n = \log_2 N + 1$, onde N é o número de atributos.

O classificador da floresta aleatória é um conjunto de árvores de classificação que são treinadas em subconjuntos aleatórios de amostras de treinamento (in-bag samples). Uma porção das amostras restantes (out-of-bag samples) é usada para uma técnica de validação cruzada interna, estimando o desempenho do modelo da floresta aleatória por meio do erro out-of-bag (Biau; Scornet, 2016).

Cada árvore de decisão é desenvolvida sem poda, e a decisão final da classificação é obtida pela média das probabilidades de atribuição de classe calculadas por todas as árvores geradas. Dessa forma, uma nova entrada de dados não rotulada é avaliada por todas as árvores no conjunto, votando em uma classe, e a classe com mais votos é selecionada como resultado final.

Embora dois parâmetros precisem ser ajustados para construir as árvores da floresta aleatória - o número de árvores de decisão a serem geradas (Ntree) e o número de variáveis para a melhor divisão (Mtry) - a sensibilidade da precisão da classificação ao Ntree é menor do que ao parâmetro Mtry. Assim, Ntree pode ser configurado para um valor elevado, e comumente, é definido como 500 em muitos estudos, devido à estabilização dos erros antes desse número de árvores ser alcançado (Biau; Scornet, 2016).

### 2.3.4. Regressão Logística

A análise de regressão logística tornou-se uma ferramenta estatística cada vez mais utilizada, embora suas origens remontem ao século XIX. É amplamente considerada a estatística de escolha para situações em que se pretende prever a ocorrência de um resultado binário (dicotômico) a partir de uma ou mais variáveis independentes (preditoras).

A regressão logística não pressupõe uma relação linear entre a variável dependente e as variáveis independentes, mas entre o logito do resultado e os valores dos preditores (Healy, 2006). A variável dependente deve ser categórica; as variáveis independentes não precisam ser intervalares, nem distribuídas normalmente, nem relacionadas linearmente, nem de variância igual dentro de cada grupo, e por fim, as categorias (grupos) devem ser mutuamente exclusivas

e exaustivas. Um caso pode estar apenas em um grupo e cada caso deve ser membro de um dos grupos. A regressão logística tem o poder de acomodar tanto variáveis independentes categóricas quanto contínuas. A inspeção dessas suposições mostra que essa técnica pode ser empregada de forma um pouco mais flexível do que as técnicas de regressão tradicionais, tornando-a adequada para muitas situações (Spitznagel, 2007).

Para qualquer caso dado, a regressão logística calcula a probabilidade de que um caso com um conjunto específico de valores para as variáveis independentes seja membro da categoria modelada. A regressão logística atribui a cada preditor um coeficiente que mede sua contribuição independente para a variação na variável dependente. A variável dependente Y assume o valor 1 se a resposta for "Sim" e o valor 0 se a resposta for "Não" (Healy, 2006).

Na Equação (5), o modelo de regressão logística relaciona diretamente a probabilidade de Y com as variáveis preditoras. O objetivo da regressão logística é estimar os k + 1 parâmetros desconhecidos β na Equação (5). Isso é feito com a estimativa de máxima verossimilhança, que envolve encontrar o conjunto de parâmetros para o qual a probabilidade dos dados observados é maior. Os coeficientes de regressão indicam o grau de associação entre cada variável independente e o resultado. Cada coeficiente representa a quantidade de mudança que esperaríamos na variável de resposta se houvesse uma mudança de uma unidade na variável preditora. O objetivo da regressão logística é prever corretamente a categoria do resultado para casos individuais usando o melhor modelo (Spitznagel, 2007). Para alcançar esse objetivo, um modelo é criado que inclui todas as variáveis preditoras que são úteis na previsão da variável de resposta.

$$P(Y) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}} \qquad (5)$$

Y é a variável dicotômica do resultado; $X_1,X_2,...,X_k$ são as variáveis preditoras, $\beta_0,\beta_1,\beta_2,...,\beta_k$ são os coeficientes de regressão (modelo) e $\beta_0$ é o intercepto.

A variável dependente binária tem os valores 0 e 1 e o valor previsto (probabilidade) deve estar limitado a cair na mesma faixa. Para definir uma relação limitada entre 0 e 1, a regressão logística usa a curva logística para representar a relação entre a variável independente e dependente. Em níveis muito baixos da variável independente, a probabilidade se aproxima de 0, mas nunca atinge 0. Da mesma forma, se a variável independente aumenta, os valores previstos aumentam ao longo da curva e se aproximam de 1, mas nunca igualam a 1.

## 3. MATERIAIS E MÉTODOS

Nesta seção, delinearemos os procedimentos adotados para a comparação de algoritmos, abrangendo desde a estruturação do conjunto de dados até a obtenção das métricas de avaliação. A Figura 4 ilustra os passos executados, os quais serão detalhados a seguir.

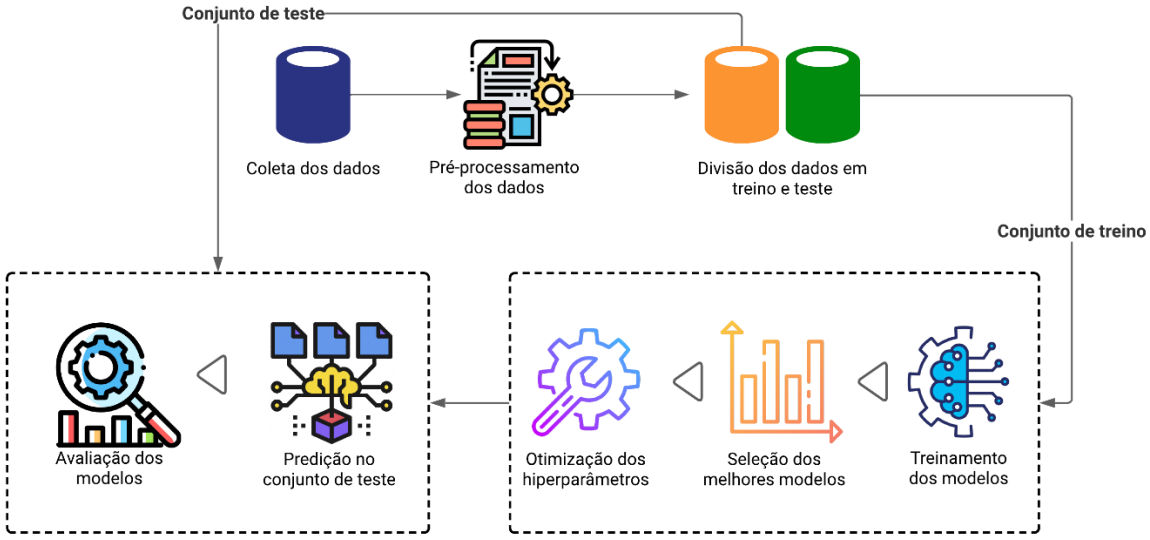


Figura 4 – A metodologia experimental empregada na comparação dos algoritmos (Fonte: Autoral, 2023).

O treinamento dos modelos de aprendizado de máquina foi conduzido utilizando a versão 3.2.0 da biblioteca PyCaret. Esta ferramenta de código aberto em Python foi escolhida devido à sua compatibilidade com o ambiente colaborativo Google Colaboratory. Este estudo empregou os algoritmos de classificação disponíveis na biblioteca, os quais abrangem a regressão logística, k-vizinhos mais próximos, floresta aleatória, máquinas de vetores de suporte (SVM), naive Bayes e XGBoost. Esta diversidade de algoritmos representou uma abordagem abrangente e significativa para explorar diferentes técnicas de aprendizado de máquina pertinentes ao escopo dessa investigação.

### 3.1. Descrição do conjunto de dados

Os dados foram obtidos durante a pesquisa realizada para o meu doutoramento (Ribeiro, 2022) e originam-se de uma ETE descentralizada, a qual empregou o método de eletrocoagulação. Essa unidade foi comissionada pela VentilAQUA em uma fábrica de panificação na Eslovênia.

A tecnologia VABEC® da VentilAQUA consiste em um sistema de fluxo contínuo de EC, com célula multi-eletrodo composta por eletrodos feitos de materiais adequados para oxidação e coagulação, com uma configuração modular e uma geometria interna projetada para máxima eficiência. Após a fase de reação química, é empregado um procedimento de flotação para separação sólido-líquido. A unidade de flotação por ar dissolvido (DAF) é um sistema compacto pré-montado construído com a tecnologia VAMEF® da VentilAQUA.

Este sistema contém uma caixa elétrica dedicada com um retificador de energia para fornecer corrente elétrica aos eletrodos, ajustar a amperagem para atender aos objetivos de tratamento, além de realizar um deslocamento automático e programado de potência como procedimento antipassivação. Um painel de controle elétrico equipado com um controlador lógico programável gerencia a operação de toda a unidade.

As variáveis monitoradas para determinar a qualidade e eficiência do processo foram as seguintes (Figura 5): condutividade (antes e depois do processo de EC); pH (antes e depois do processo de EC, e dentro do sistema DAF); fluxo (nos sistemas de EC e DAF); voltagem; corrente; polaridade; e ORP (dentro do sistema DAF).

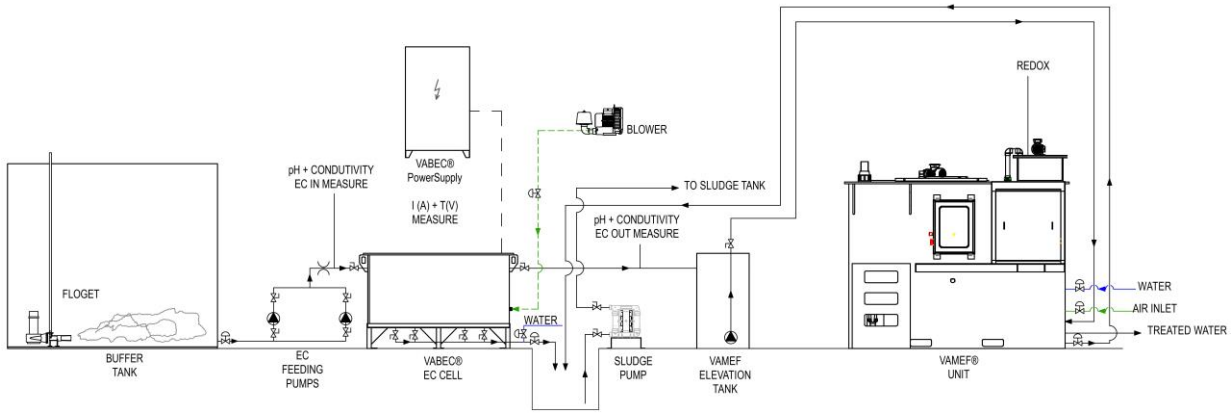


Figura 5 – As posições relativas dos sensores na ETE (Fonte: Autoral, 2023).

A variável-alvo era dois modos operacionais baseados na clarificação do efluente e no lodo de reação. Tratamos o problema como uma tarefa de classificação, com duas classes com base no conhecimento especializado da seguinte forma:

- Classe 0: Não clarificado, exibindo turbidez;
- Classe 1: Clarificado, exibindo baixa turbidez.

### 3.2. Análise descritiva estatística

Examinar a dinâmica populacional dos dados representa um aspecto crucial e substancial na pesquisa, visto que desempenha um papel fundamental na compreensão do sistema em questão. Tal análise possibilita a formulação de conjecturas hipotéticas e a avaliação mais rigorosa dos resultados, contribuindo significativamente para a profundidade da análise e interpretação dos dados.

A base de dados não apresenta ausência de valores em nenhuma das variáveis analisadas. A Tabela 1 e a Figura 6 fornecem uma visão geral das variáveis e características do conjunto de dados, respectivamente.

Tabela 1 – Visão geral do número de variáveis e observações no conjunto de dados.

| | |
|---|---|
| **Número de variáveis de entrada** | 11 |
| **Número de variáveis de saída** | 1 |
| **Número de observações** | 1207 |

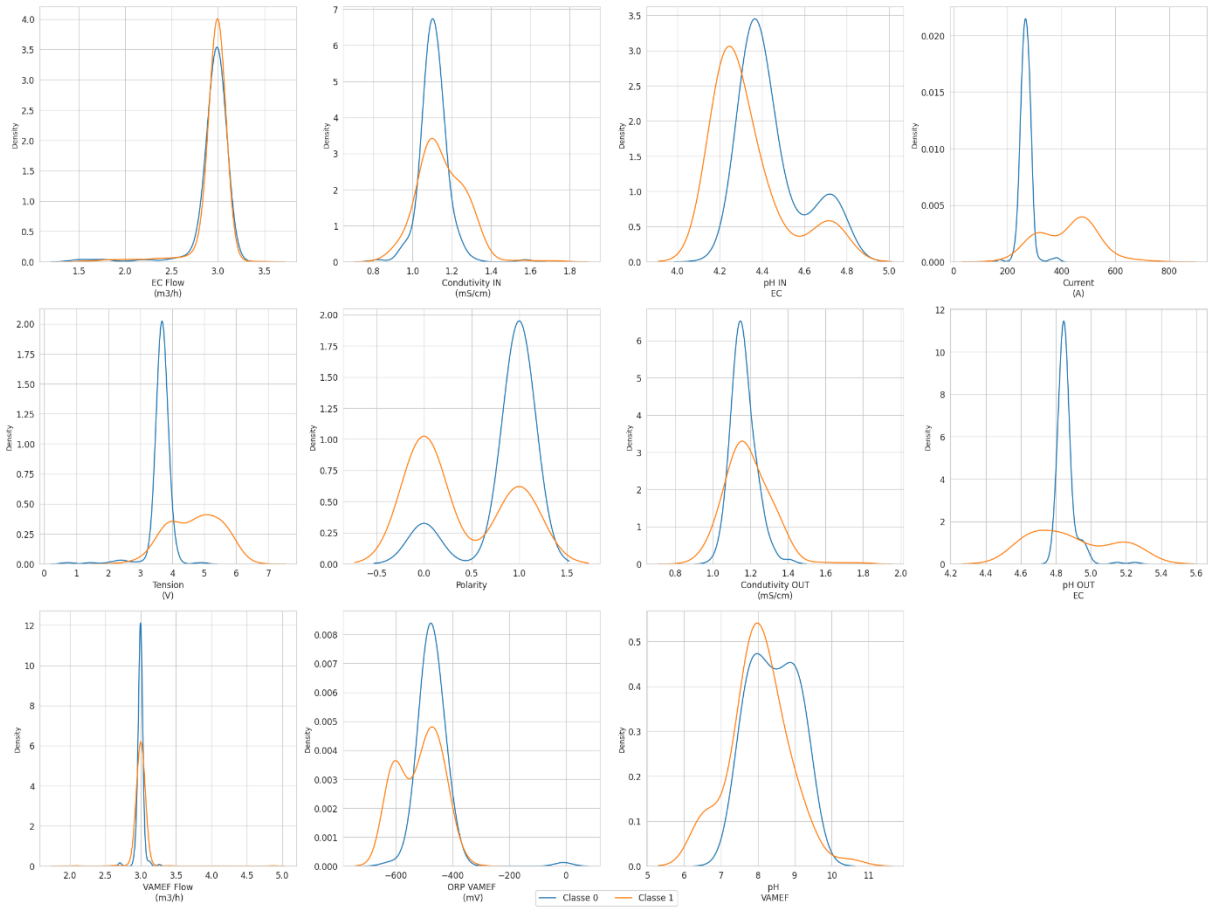

Figura 6 – Densidades das variáveis por estado (Fonte: Autoral, 2023).

A análise da Figura 6 revela a ausência de uma distinção substancial entre as classes em relação à maioria das variáveis estudadas. Não obstante, é pertinente ressaltar que as variáveis corrente, tensão e pH de saída exibem padrões de comportamento mais distintos entre as duas densidades observadas por estado.

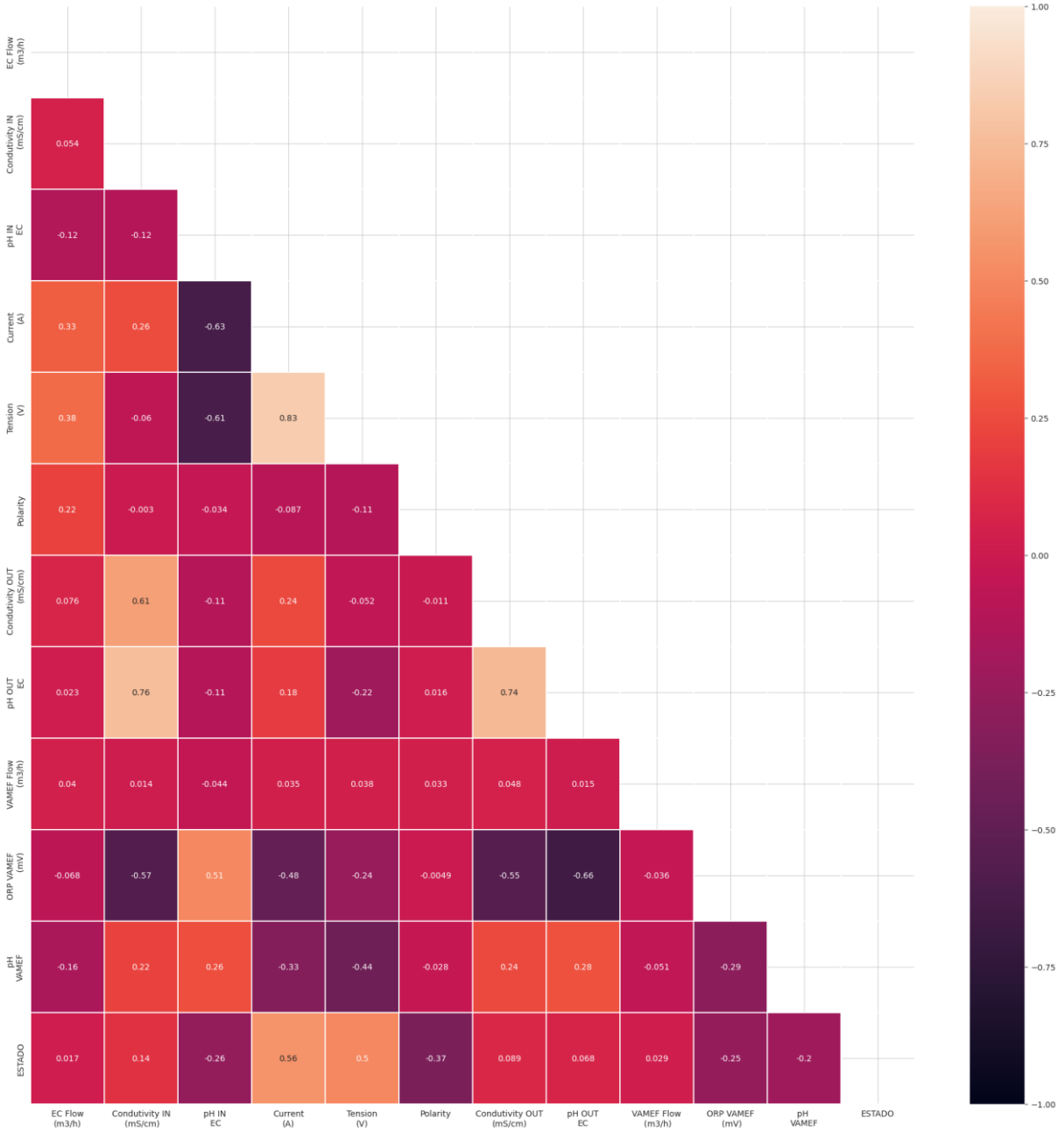


Figura 7 – Matriz de correlação (Pearson) das variáveis de entrada e de saída (Fonte: Autoral, 2023).

A Figura 7 ilustra as correlações existentes entre as variáveis consideradas, inclusive a variável de saída, revelando diferentes relações fundamentadas nos fenômenos subjacentes ao processo de EC. Notavelmente, a correlação altamente positiva entre corrente e tensão é coerente com os preceitos da Lei de Ohm, refletindo aspectos físicos do processo. Adicionalmente, ambas as variáveis demonstram uma forte correlação inversa com a variável de pH de entrada, o que se justifica pelo fato de que, no processo de eletrocoagulação, fatores como tempo de eletrólise, intensidade de corrente e pH inicial do meio guardam relação proporcional com a concentração do coagulante e o pH da solução, fundamentados na Lei de Faraday.

Adicionalmente, a Figura 7 evidencia que tanto a tensão quanto a corrente exibem uma correlação significativa e positiva com o estado do processo, representado pela variável de saída. Tal associação é plausível, dado que essas variáveis exercem influência sobre diversos aspectos termodinâmicos e cinéticos inerentes à EC. Este resultado complementa as observações obtidas na Figura 6, reforçando as constatações previamente mencionadas.

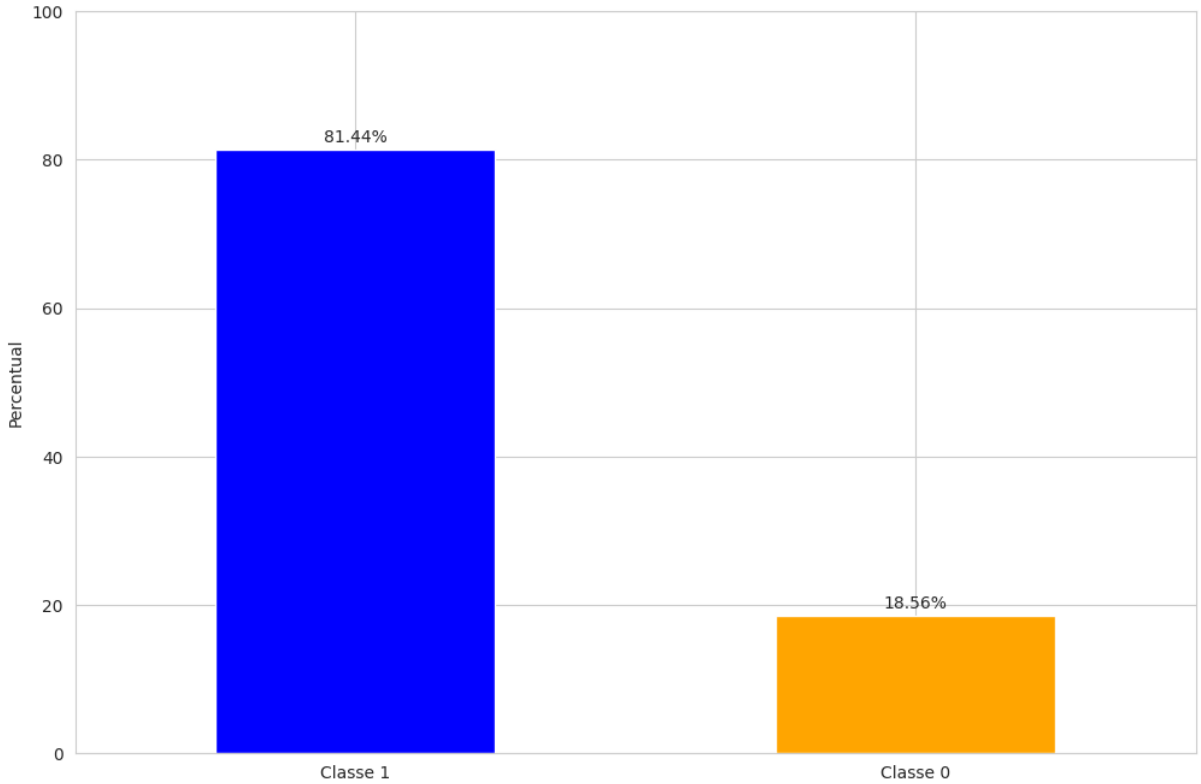


Figura 8 – Distribuição das classes da variável de saída (Fonte: Autoral, 2023).

A análise da Figura 8 indica a presença de desbalanceamento entre as classes. Este desequilíbrio é justificável do ponto de vista operacional, considerando que a Classe 1,

caracterizada pela ausência de falhas no processo, é esperada ter uma ocorrência superior em comparação com a Classe 0, que representa situações onde o processo apresenta falhas.

Desta forma, um método de oversampling foi aplicado utilizando a técnica de oversampling sintético de minoria (SMOTE), visando mitigar desequilíbrios nos dados, mantendo resultados aceitáveis e evitando a introdução de ruídos desnecessários. Através da síntese de novas instâncias para a classe minoritária, o desequilíbrio entre as classes majoritárias e minoritárias é reduzido, tornando as classes mais equiparáveis para previsões futuras. A metodologia para gerar novas instâncias sintéticas da classe minoritária baseia-se na identificação dos k-vizinhos mais próximos utilizando a métrica de distância euclidiana entre os dados (Chawla et al., 2002).

A principal vantagem do uso da técnica de amostragem SMOTE em comparação com outros métodos tradicionais reside na criação de observações sintéticas, em vez de recorrer à reutilização de observações já existentes. Isso resulta em um classificador menos suscetível a overfitting (Chawla et al., 2002). É relevante notar que o oversampling é aplicado exclusivamente nos dados de treinamento, sem utilizar informações dos dados de validação para gerar observações sintéticas. Portanto, os resultados obtidos são generalizáveis, tornando essa abordagem particularmente precisa e aplicável para a implementação de um modelo em um ambiente de produção.

O PyCaret apresenta um parâmetro denominado "fix_imbalance" no procedimento de configuração (setup); quando este parâmetro é estabelecido como True, o método de oversampling SMOTE é aplicado durante a fase de preparação dos dados.

Uma etapa crucial realizada no pré-processamento dos dados é a normalização dos dados númericos. Essa técnica desempenha um papel fundamental na padronização dos dados, visando garantir que diferentes variáveis possuam uma escala similar, evitando assim que uma variável com grande amplitude possua um impacto desproporcional durante o treinamento do modelo. Desta forma, o presente trabalho adotou a normalização do tipo minmax.

É relevante mencionar que o pipeline de transformação foi adequadamente ajustado empregando exclusivamente o conjunto de treinamento. Essa abordagem é essencial para evitar vazamento de informações do conjunto de teste para o conjunto de treinamento, mantendo a integridade e a validade dos resultados. Uma vez ajustadas, as transformações foram aplicadas de forma consistente ao conjunto de dados completo, garantindo, por exemplo, uniformidade na escala das variáveis ao longo de todo o conjunto de dados analisado.

Os dados foram particionados em duas partes, sendo alocados 70% para o conjunto de treinamento e 30% para o conjunto de teste. Essa divisão foi realizada através de uma

amostragem estratificada, garantindo que as proporções das classes presentes na amostra original fossem mantidas nos conjuntos de treinamento e teste.

A Figura 9 fornece evidências que sustentam a similaridade das correlações entre as features presentes nos conjuntos de treinamento e teste. A menor diferença média absoluta observada indica um grau de alinhamento na interação das features, reforçando a confiança na consistência dos dados entre os conjuntos.

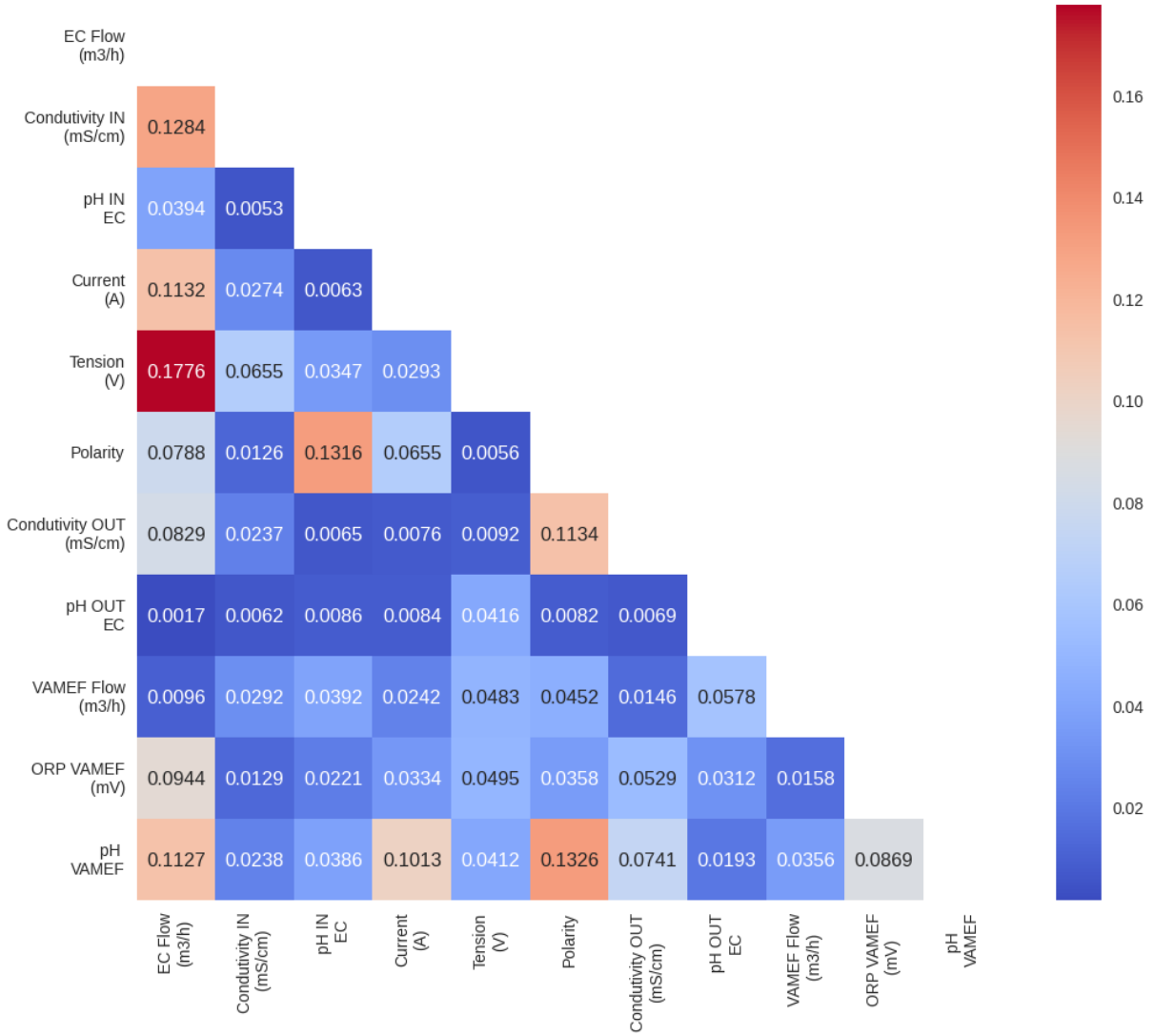


Figura 9 – Diferença média absoluta nas matrizes de correlação (Pearson) entre conjunto de treinamento e teste (Fonte: Autoral, 2023).

### 3.3. Modelo

Através da utilização da função "compare_models" disponível na biblioteca PyCaret, os modelos foram treinados empregando os hiperparâmetros padrão da biblioteca scikit-learn. A técnica de validação cruzada com 10 folds estratificados foi utilizada.

No método de validação cruzada k-fold estratificado o conjunto de dados é dividido em k subconjuntos (ou folds) de tamanho semelhante, mantendo a proporção entre as classes em cada fold, preservando assim a distribuição original das classes. O modelo é treinado k vezes, cada vez usando k-1 folds como conjunto de treinamento e o fold restante como conjunto de teste. Esse processo é repetido k vezes, de forma que cada fold seja usado exatamente uma vez como conjunto de teste. Ao final, são obtidas k métricas de desempenho que são então combinadas (por exemplo, calculando a média) para fornecer uma estimativa geral do desempenho do modelo.

O resultado fornecido pela função "compare_models" é uma grade de pontuação que demonstra as médias das métricas de precisão, AUC (área sob a curva ROC), recall, precisão, F1-score, kappa e MCC (coeficiente de correlação de matthews) ao longo dos 10 folds da validação cruzada. Além disso, são fornecidos os tempos de treinamento de cada modelo, oferecendo assim uma visão comparativa do desempenho e eficiência temporal dos diferentes algoritmos de aprendizado de máquina utilizados. Essa abordagem facilita a identificação dos modelos mais promissores para análises subsequentes, permitindo uma seleção mais embasada dos melhores modelos para o problema em questão.

Os modelos foram avaliados utilizando a métrica F1-score, e os quatro modelos que obtiveram as pontuações mais altas nessa métrica foram selecionados. Os modelos selecionados foram submetidos à função "tune_models", visando a otimização dos hiperparâmetros por meio do método de busca em grade aleatória. Essa abordagem depende do número de iterações para alcançar uma melhoria no ajuste do modelo. Assim, para aprimoramento dos modelos, foram realizadas 50 iterações aleatórias no espaço de busca. A técnica de validação cruzada com 10 folds estratificados também foi utilizada.

No contexto da otimização por grade, cada ponto da grade representa uma combinação específica de valores para os hiperparâmetros. Para cada combinação, o modelo é treinado e avaliado utilizando uma métrica de desempenho específica.

A métrica escolhida durante esse processo foi o F1-score. O F1-score mensura a eficácia de um modelo de classificação por meio do cálculo da média harmônica entre a precisão e o recall do classificador (Wardhani et al., 2019). A fórmula do F1-score pode ser interpretada como uma média ponderada entre precisão e recall, variando de 0 a 1, onde 0 representa a

pontuação mais baixa e 1 representa a mais alta. Precisão e recall contribuem igualmente para o F1-score, permitindo encontrar um equilíbrio ideal entre essas duas métricas.

Por fim, os modelos, cujos hiperparâmetros foram otimizados, foram submetidos a uma avaliação no conjunto de teste. Esse estágio permitiu uma validação independente do desempenho dos modelos, fornecendo uma avaliação de sua capacidade de generalização para novos conjuntos de dados não utilizados durante o treinamento. A análise no conjunto de teste é essencial para verificar a robustez e a eficácia dos modelos em situações do mundo real, contribuindo para a validação externa e confiabilidade dos resultados obtidos.

## 4. RESULTADOS E DISCUSSÕES

### 4.1. Comparação e escolha dos modelos

Os modelos que exibiram os valores mais elevados de F1-score no conjunto de treinamento, conforme indicado na Tabela 2, compreenderam a regressão logística, a floresta aleatória, o XGBoost e o SVM com kernel radial, com pontuações médias superiores a 0,93 para a métrica considerada.

Esses modelos demonstram capacidades que os tornam vantajosos em diferentes cenários de aplicação. A regressão logística, conhecida por sua interpretabilidade e simplicidade, oferece uma compreensão clara do impacto de cada variável no resultado preditivo (Nasteski, 2017). A floresta aleatória, por sua vez, destaca-se pela robustez em lidar com dados não lineares e alta dimensionalidade (Fawagreh; Gaber; Elyan, 2014) ao passo que o XGBoost é reconhecido pela eficiência em lidar com dados de grandes volumes e pela otimização no processo de treinamento, resultando em modelos de alta performance (Natekin; Knoll, 2013). Por fim, o SVM com kernel radial evidencia sua eficácia na separação de classes em espaços de alta dimensionalidade, sendo especialmente útil quando há presença de dados não lineares (Nasteski, 2017). A combinação desses pontos fortes em cada modelo contribuiu significativamente para sua performance na tarefa de classificação.

Tabela 2 – Ranking dos melhores modelos segundo a métrica F1-score no conjunto de treinamento.

| | Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC | TT (Sec) |
|---|---|---|---|---|---|---|---|---|---|
| **rf** | Random Forest Classifier | 0.9516 | 0.9913 | 0.9508 | 0.9895 | 0.9695 | 0.8524 | 0.8588 | 0.389 |
| **lr** | Logistic Regression | 0.9447 | 0.9755 | 0.9678 | 0.9646 | 0.9660 | 0.8169 | 0.8188 | 0.140 |
| **xgboost** | Extreme Gradient Boosting | 0.9420 | 0.9707 | 0.9390 | 0.9894 | 0.9633 | 0.8251 | 0.8332 | 0.292 |
| **rbfsvm** | SVM - Radial Kernel | 0.8978 | 0.9308 | 0.9712 | 0.9105 | 0.9395 | 0.6088 | 0.6256 | 0.122 |
| **nb** | Naive Bayes | 0.8867 | 0.9526 | 0.8763 | 0.9831 | 0.9260 | 0.6865 | 0.7096 | 0.053 |
| **knn** | K Neighbors Classifier | 0.8728 | 0.8755 | 0.9475 | 0.9018 | 0.9237 | 0.5402 | 0.5542 | 0.082 |
| **svm** | SVM - Linear Kernel | 0.8440 | 0.0000 | 0.9271 | 0.8895 | 0.9056 | 0.4279 | 0.4567 | 0.052 |

As próximas seções apresentarão uma análise dos resultados da otimização dos hiperparâmetros para cada modelo.

## 4.2. Otimização dos hiperparâmetros dos modelos

### 4.2.1. Floresta aleatória

A análise dos resultados apresentados na Figura 10, representando a métrica F1-score para cada fold avaliado no modelo de floresta aleatória, revela uma maior variabilidade entre os folds no conjunto de validação, após a otimização dos hiperparâmetros.

Essa variabilidade sugere a possibilidade de que o espaço de busca dos hiperparâmetros adotado durante a otimização poderia não ter sido suficientemente abrangente para capturar de forma abrangente a melhor configuração do modelo. A utilização de uma abordagem de otimização mais ampla ou a exploração de diferentes técnicas além do random grid, como por exemplo, a busca bayesiana, poderia oferecer uma perspectiva mais abrangente do espaço de hiperparâmetros, permitindo uma busca mais eficiente das configurações que melhor se adequam ao conjunto de dados. Essa ampliação estratégica do espaço de busca dos hiperparâmetros ou a exploração de métodos de otimização alternativos pode contribuir significativamente para mitigar a variabilidade observada entre os folds de validação, potencialmente aprimorando a capacidade de generalização e o desempenho preditivo do modelo de floresta aleatória.

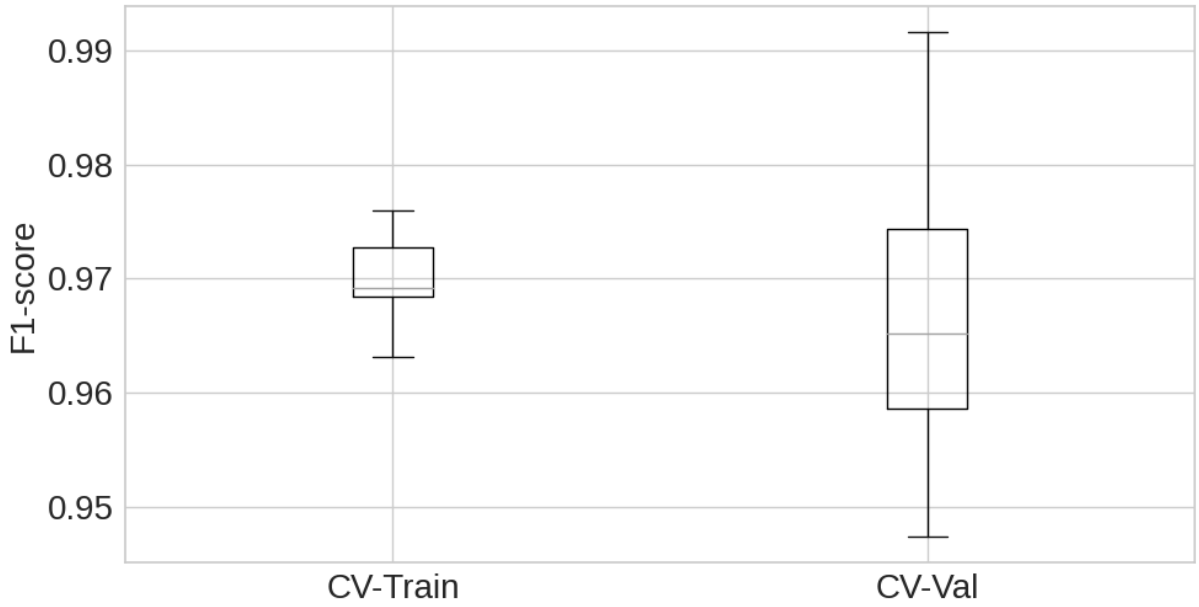


Figura 10 – Desempenho, após otimizar os hiperparâmetros, no conjunto de treinamento e de validação do modelo de floresta aleatória (Fonte: Autoral, 2023).

Os hiperparâmetros definidos durante o processo de tuning foram os seguintes:

- bootstrap=True
- ccp_alpha=0.0
- class_weight=None
- criterion='gini'
- max_depth=None
- max_features='sqrt'
- max_leaf_nodes=None
- max_samples=None
- min_impurity_decrease=0.0
- min_samples_leaf=1
- min_samples_split=2
- min_weight_fraction_leaf=0.0
- n_estimators=100
- oob_score=False
- warm_start=False

### 4.2.2. XGBoost

A análise representada na Figura 11 evidencia uma maior dispersão dos valores de F1-score entre os diferentes folds do conjunto de validação. É importante enfatizar que essa dispersão não necessariamente indica uma limitação na capacidade de generalização do modelo XGBoost, mas sim pode ser um reflexo da complexidade dos dados.

Abordagens como a seleção de features ou a aplicação de técnicas de redução da dimensionalidade surgem como alternativas viáveis para lidar de maneira eficaz com a complexidade dos dados, com potencial para aprimorar a consistência do desempenho do modelo XGBoost. Essas estratégias podem permitir uma representação mais concisa e informativa para o modelo. Ao reduzir a complexidade dos dados sem perder informações cruciais, tais abordagens podem favorecer a capacidade do modelo em capturar padrões.

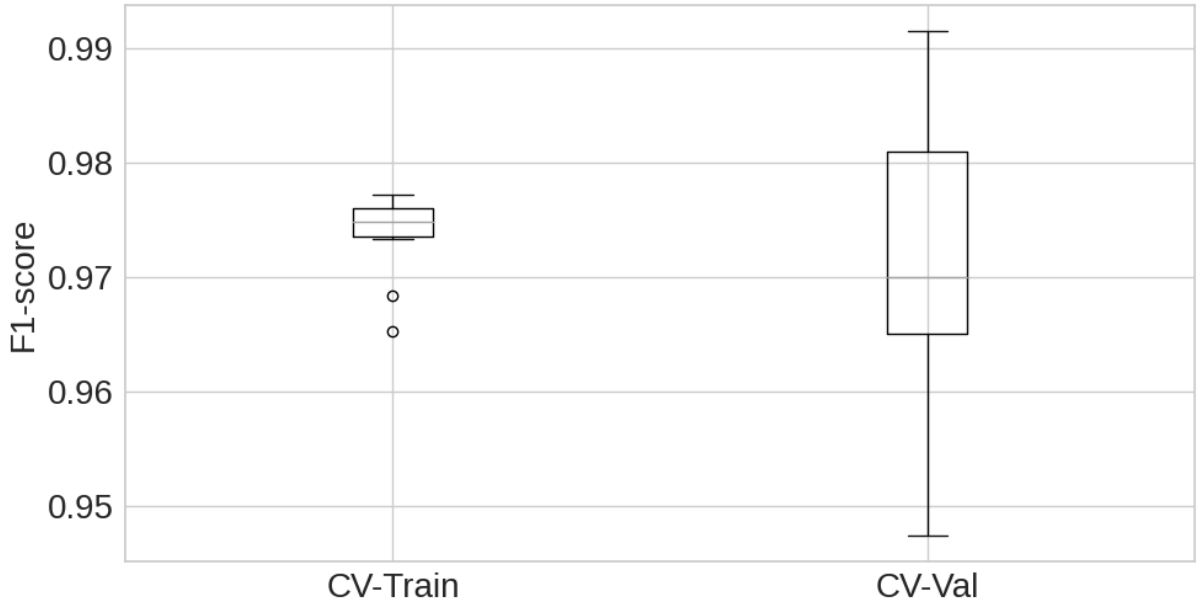


Figura 11 – Desempenho, após otimizar os hiperparâmetros, no conjunto de treinamento e de validação do modelo XGBoost (Fonte: Autoral, 2023).

Os hiperparâmetros definidos durante o processo de tuning foram os seguintes:

- base_score=None
- booster='gbtree'
- colsample_bylevel=None
- colsample_bynode=None
- colsample_bytree=1.0
- early_stopping_rounds=None
- enable_categorical=False
- eval_metric=None

- gamma=None
- grow_policy=None
- importance_type=None
- interaction_constraints=None
- learning_rate=0.49999999999999994
- max_bin=None
- max_cat_threshold=None
- max_cat_to_onehot=None
- max_delta_step=None
- max_depth=11
- max_leaves=None
- min_child_weight=4
- monotone_constraints=None
- multi_strategy=None
- n_estimators=272
- num_parallel_tree=None
- objective='binary:logistic'

#### 4.2.3. Regressão logística

Os resultados referentes à regressão logística são detalhados na Figura 12. Destaca-se uma baixa dispersão nos valores da métrica F1-score ao compararmos o conjunto de treino com o conjunto de validação. Esta consistência nos resultados sugere uma boa capacidade de generalização do modelo para novos dados. No entanto, é importante notar a presença de alguns outliers, indicando valores discrepantes.

Por ser um modelo relativamente simples, a regressão logística tende a ter um baixo risco de overfitting em comparação com modelos mais complexos. Isso significa que o modelo é menos propenso a se ajustar excessivamente aos dados de treinamento e, portanto, mantém uma boa capacidade de generalização.

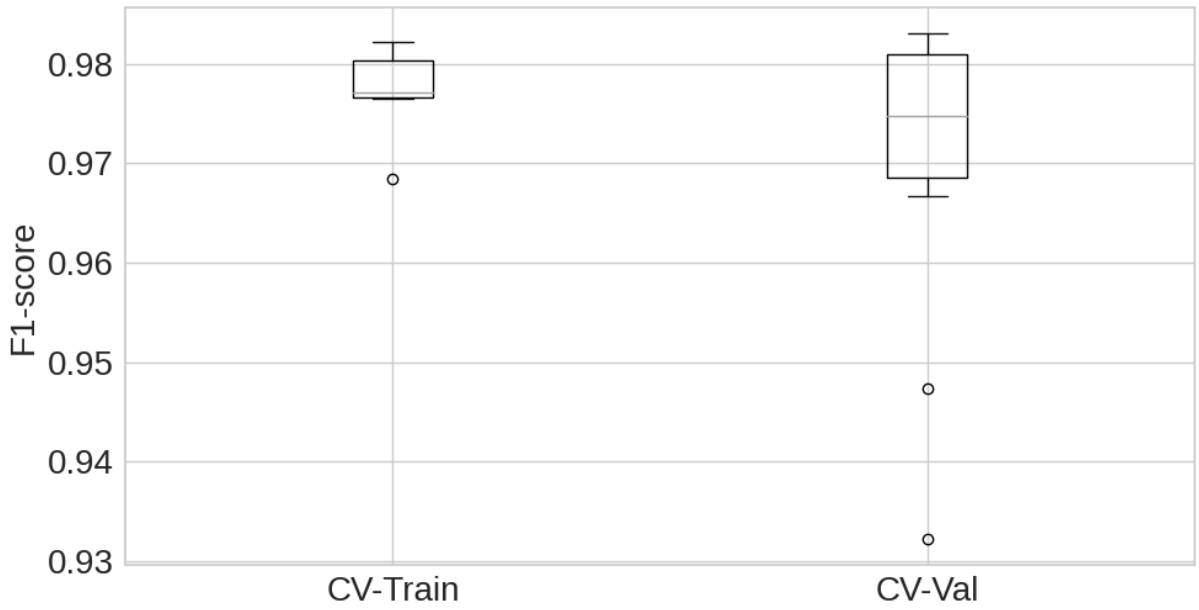


Figura 12 – Desempenho, após otimizar os hiperparâmetros, no conjunto de treinamento e de validação do modelo de regressão logística (Fonte: Autoral, 2023).

Os hiperparâmetros definidos durante o processo de tuning foram os seguintes:

- C=9.925188555887118
- class_weight=None
- dual=False
- fit_intercept=True
- intercept_scaling=1
- l1_ratio=None
- max_iter=1000
- multi_class='auto'
- penalty='l2'
- solver='lbfgs'
- tol=0.0001
- warm_start=False

#### 4.2.4. SVM com kernel radial

A Figura 13 apresenta os resultados para o modelo SVM com kernel radial, onde se observa uma expressiva variação na métrica F1-score entre o conjunto de treinamento e o conjunto de validação nos folds da validação cruzada. Essa variação pode ser atribuída à sensibilidade do SVM com kernel radial aos hiperparâmetros, como o parâmetro de

regularização (C) e a largura da função de kernel (gamma). Pequenas variações nesses parâmetros podem influenciar drasticamente a performance do modelo (Friedrichs; Igel, 2005), levando a uma maior discrepância entre os conjuntos.

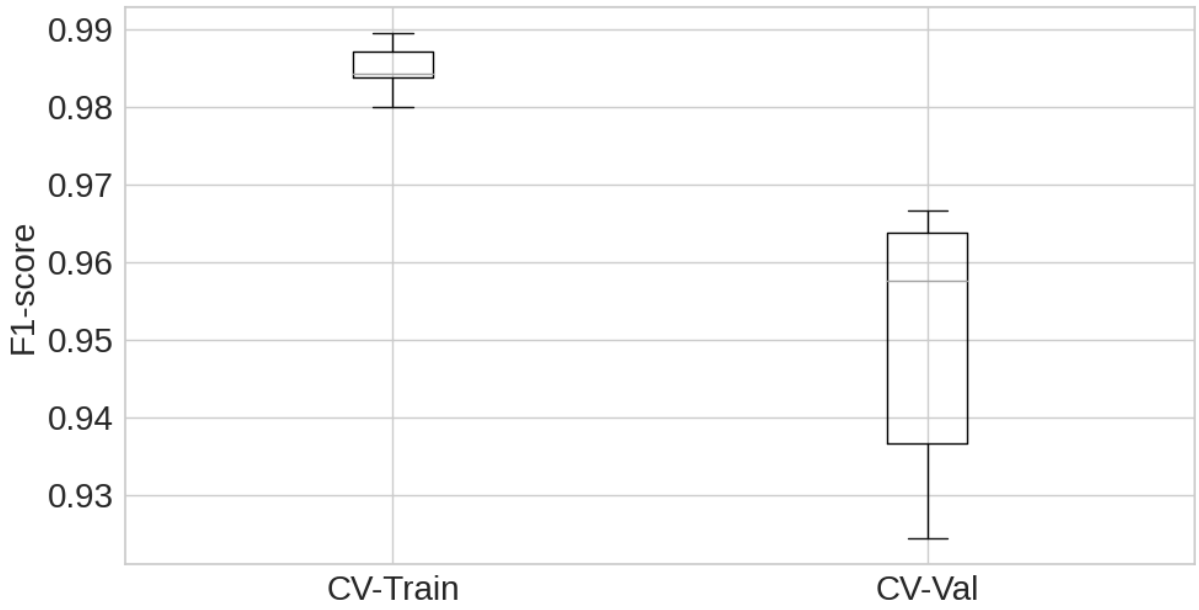


Figura 13 – Desempenho, após otimizar os hiperparâmetros, no conjunto de treinamento e de validação do modelo SVM com kernel radial (Fonte: Autoral, 2023).

Os hiperparâmetros definidos durante o processo de tuning foram os seguintes:

- C=42.93468848796315
- break_ties=False
- class_weight='balanced'
- coef0=0.0
- decision_function_shape='ovr'
- degree=3
- gamma='auto'
- kernel='rbf'
- max_iter=-1
- probability=True
- shrinking=True
- tol=0.001

A próxima seção compreenderá uma análise comparativa dos desempenhos de cada modelo no conjunto de teste.

### 4.3. Comparação de desempenho dos modelos no conjunto de teste

Antes de adentrar na análise comparativa, é crucial salientar um elemento fundamental no contexto do presente estudo. No âmbito dos classificadores, é possível identificar dois tipos de erros, nomeadamente, o erro tipo I e o erro tipo II. O erro tipo I ocorre quando se verifica um falso positivo, enquanto o erro tipo II se manifesta por um falso negativo. O falso positivo, em específico, representa um ponto crítico nesta análise. Ele incorretamente indica um desempenho adequado do sistema, quando na verdade há falhas presentes. Esta condição equivocada pode acarretar prejuízos multifacetados na operação, ressaltando a relevância de sua identificação e mitigação para aprimorar a confiabilidade e a eficiência do sistema em questão.

A análise das matrizes de confusão, conforme ilustrado na Figura 14, revela variações significativas no desempenho dos modelos avaliados. Observa-se que o modelo SVM com kernel radial exibe a maior incidência de erro do tipo I, enquanto o modelo de floresta aleatória apresenta a maior incidência de erro do tipo II. É importante ressaltar que, apesar da predominância da floresta aleatória para o erro do tipo II, a mesma demonstra a menor incidência para o erro do tipo I. É pertinente destacar que o equilíbrio entre os erros do tipo I e II constitui um "trade-off" nesse contexto. Este trade-off refere-se à busca por um ponto ótimo que balanceie a minimização de ambos os tipos de erro, uma vez que a redução de um tipo de erro pode, muitas vezes, levar ao aumento do outro, tornando imperativa a consideração cuidadosa desse equilíbrio.

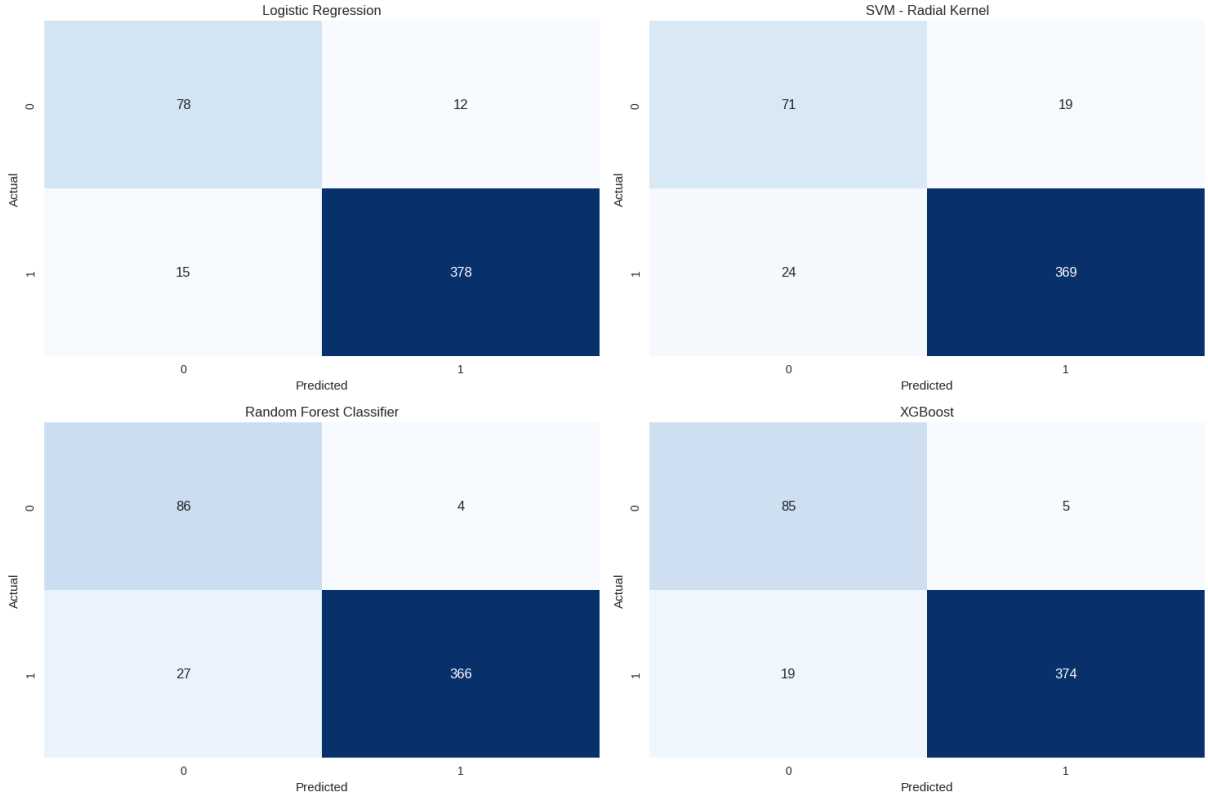


Figura 14 – Matrizes de confusão (conjunto de teste) gerada para cada modelo (Fonte: Autoral, 2023).

Na Figura 14, observa-se que o modelo XGBoost se sobressai como um candidato significativo em comparação aos demais, pois exibe uma baixa incidência de erro do tipo I, além de apresentar um nível aceitável de erros do tipo II. Essa constatação ressalta a atratividade do XGBoost como uma opção viável, pois demonstra a capacidade de minimizar os erros críticos do tipo I sem comprometer excessivamente a ocorrência de erros do tipo II, contribuindo para um desempenho mais equilibrado e confiável do sistema no desafio em questão.

O exame da distribuição das pontuações de predição emerge como um indicativo da confiança subjacente do modelo em suas projeções. Uma considerável sobreposição entre as pontuações associadas aos acertos e aos erros sugere uma menor confiabilidade nas predições do modelo. Em contrapartida, uma notável discrepância entre essas distribuições sugere uma maior confiança do modelo nos casos em que acerta.

Neste contexto, a análise representada na Figura 15 revela que o modelo SVM com kernel radial exibe a maior sobreposição entre as distribuições, corroborando as observações prévias. Os demais modelos, entretanto, demonstram uma maior separação entre as regiões correspondentes aos acertos e erros. Esta figura, portanto, não apenas confirma as considerações

anteriores, mas também oferece insights valiosos para a determinação de limiares de predição ótimos, os quais maximizam a precisão para uma classe específica.

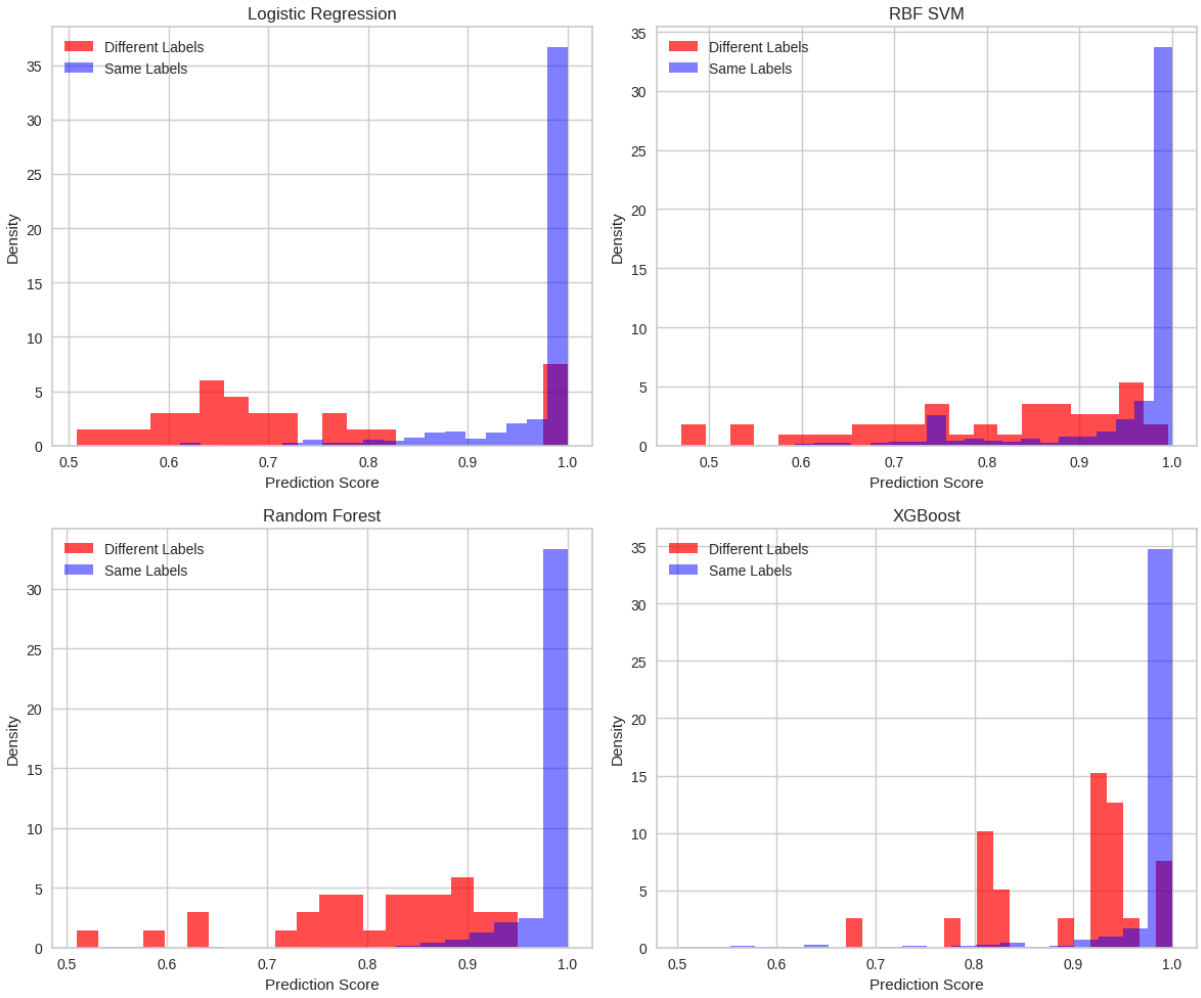


Figura 15 – Comparação das distribuições de pontuações de predição (conjunto de teste) para cada modelo (Fonte: Autoral, 2023).

Por último, a Figura 16 representa todas as métricas avaliadas para cada modelo no conjunto de teste. Ela evidencia que o modelo de floresta aleatória possui a mais alta precisão, seguido de perto pelo modelo XGBoost. O modelo XGBoost demonstra o segundo melhor desempenho em termos de recall, sendo superado apenas pela regressão logística, além de exibir o mais elevado valor de acurácia, juntamente com índices superiores de kappa e MCC.

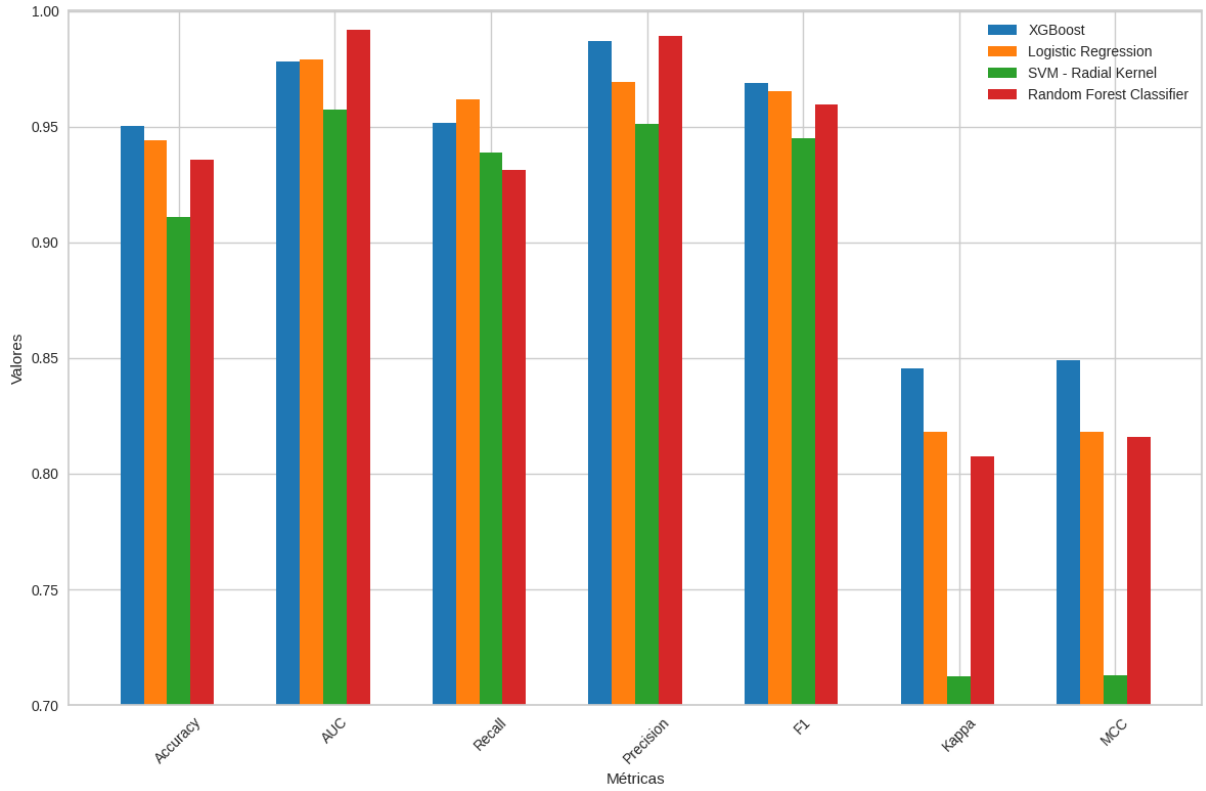


Figura 16 – Comparação das métricas avaliadas (conjunto de teste) para cada modelo (Fonte: Autoral, 2023).

## 5. CONCLUSÃO

Uma estratégia para desenvolver um modelo para detectar comportamentos anômalos em um processo descentralizado de uma ETE por EC foi proposta neste trabalho. Foram utilizados os algoritmos floresta aleatória, regressão logística, SVM com kernel radial e XGBoost para encontrar um modelo adequado para classificar dados em diferentes condições operacionais.

Os resultados no conjunto de teste destacaram que o modelo de floresta aleatória alcançou a mais alta precisão, seguido de perto pelo XGBoost. Este último, além de apresentar altos índices de precisão, também demonstrou desempenho notável em diversas métricas, como acurácia, kappa e MCC.

O modelo XGBoost destacou-se ao minimizar o erro crítico do tipo I (falsos positivos) sem comprometer excessivamente o erro do tipo II (falsos negativos). Essa capacidade de encontrar um equilíbrio entre esses erros críticos é crucial para garantir a confiabilidade do sistema.

A floresta aleatória se destacou não apenas por alcançar uma alta precisão, mas também por uma importante característica: sua notável explicabilidade. Essa qualidade da floresta aleatória proporciona uma compreensão clara e detalhada dos fatores determinantes por trás das previsões. Ao oferecer uma visão transparente sobre quais variáveis e critérios são mais relevantes para a tomada de decisões, a floresta aleatória permite que os operadores das estações de tratamento de efluentes compreendam de maneira mais tangível como as falhas estão sendo identificadas pelo modelo.

Consequentemente, esta pesquisa evidenciou o potencial e a eficácia da abordagem baseada em modelos na detecção de falhas em ETEs, possibilitando a integração com sistemas de apoio à decisão para manter um alto desempenho.

Algumas sugestões são apresentadas para pesquisas futuras:

- Realizar novos experimentos na otimização de hiperparâmetros utilizando novos métodos para aprimorar ainda mais a precisão e a eficiência dos modelos. Essa etapa permitirá explorar configurações mais refinadas de maneira mais eficiente no espaço de busca. Sugere-se a utilização de métodos como otimização bayesiana, otimização por enxame de partículas ou algoritmos genéticos;
- Realizar seleção de features e reavaliar os modelos a fim de ampliar a capacidade dos modelos em identificar padrões relevantes nos dados. Sugere-se, por exemplo, a aplicação de técnicas como análise de importância de variáveis, métodos wrapper como recursive feature elimination ou métodos embedded como LASSO para selecionar as características mais relevantes. Essa abordagem contribuirá não apenas para simplificar e aprimorar o modelo, mas também para aperfeiçoar a precisão e a capacidade de generalização em diversos cenários.
- Aprofundar a análise dos melhores pontos de corte das pontuações de previsão. A identificação precisa desses pontos de corte otimiza a capacidade do modelo em discernir com precisão entre as condições normais e anômalas, refinando a eficácia do sistema de detecção de falhas e a sua habilidade de tomada de decisão.

**REFERÊNCIAS**

ALABI, M. O.; TELUKDARIE, A.; VAN RENSBURG, N. J. Water 4.0: An Integrated Business Model from an Industry 4.0 Approach. In: IEEE INTERNATIONAL CONFERENCE ON INDUSTRIAL ENGINEERING AND ENGINEERING MANAGEMENT (IEEM). Anais… Macao, Macao: IEEE, Dec. 2019. Available: https://ieeexplore.ieee.org/document/8978859/

ALOM, M. Z. et al. A State-of-the-Art Survey on Deep Learning Theory and Architectures. Electronics, v. 8, n. 3, p. 292, 5 Mar. 2019.

BIAU, Gérard; SCORNET, Erwan. A random forest guided tour. Test, v. 25, p. 197-227, 2016.

BREIMAN, Leo. Random forests. Machine learning, v. 45, p. 5-32, 2001.

CHAUHAN, V. K.; DAHIYA, K.; SHARMA, A. Problem formulations and solvers in linear SVM: a review. Artificial Intelligence Review, v. 52, n. 2, p. 803-855, Aug. 2019.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In:: KDD ’16: THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. Anais… San Francisco California USA: ACM, 13 ago. 2016. Available: https://dl.acm.org/doi/10.1145/2939672.2939785

CRITTENDEN, J. C. et al. MWH’s Water Treatment: Principles and Design. New Jersey: Wiley, 2012.

DAS, P. P.; SHARMA, M.; PURKAIT, M. K. Recent progress on electrocoagulation process for wastewater treatment: A review. Separation and Purification Technology, v. 292, p. 121058, July 2022.

DAIRI, A. et al. Deep learning approach for sustainable WWTP operation: A case study on data-driven influent conditions monitoring. Sustainable Cities and Society, v. 50, p. 101670, Oct. 2019.

DIAZ-ELSAYED, N. et al. Wastewater-based resource recovery technologies across scale: A review. Resources, Conservation and Recycling, v. 145, p. 94-112, June 2019.

FAWAGREH, Khaled; GABER, Mohamed Medhat; ELYAN, Eyad. Random forests: from early developments to recent advancements. Systems Science & Control Engineering: An Open Access Journal, v. 2, n. 1, p. 602-609, 2014.

FERREIRA, A. J.; FIGUEIREDO, M. A. T. Boosting Algorithms: A Review of Methods, Theory, and Applications. In: ZHANG, C.; MA, Y. (Eds.). Ensemble Machine Learning. Boston, MA: Springer US, 2012. p. 35-85.

HEALY, Lawrence M. Logistic regression: An overview. Eastern Michighan College of Technology, 2006.

KIJAK, R. Defining Water 4.0. In: KIJAK, R. (Ed.). Water Asset Management in Times of Climate Change and Digital Transformation. Cham: Springer International Publishing, 2021.

LI, W. et al. Process fault diagnosis with model- and knowledge-based approaches: Advances and opportunities. Control Engineering Practice, v. 105, p. 104637, dez. 2020.

LOWE, M.; QIN, R.; MAO, X. A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring. Water, v. 14, n. 9, p. 1384, 24 abr. 2022.

MD NOR, N.; CHE HASSAN, C. R.; HUSSAIN, M. A. A review of data-driven fault detection and diagnosis methods: applications in chemical process systems. Reviews in Chemical Engineering, v. 36, n. 4, p. 513-553, 26 May 2020.

MOUSAZADEH, M. et al. A systematic diagnosis of state of the art in the use of electrocoagulation as a sustainable technology for pollutant treatment: An updated review. Sustainable Energy Technologies and Assessments, v. 47, p. 101353, out. 2021.

MAMANDIPOOR, B. et al. Monitoring and detecting faults in wastewater treatment plants using deep learning. Environmental Monitoring and Assessment, v. 192, n. 2, p. 148, fev. 2020.

NASTESKI, Vladimir. An overview of the supervised machine learning methods. Horizons. b, v. 4, p. 51-62, 2017.

NATEKIN, Alexey; KNOLL, Alois. Gradient boosting machines, a tutorial. Frontiers in neurorobotics, v. 7, p. 21, 2013.

NEWHART, K. B. et al. Data-driven performance analyses of wastewater treatment plants: A review. Water Research, v. 157, p. 498-513, jun. 2019.

PROBST, Philipp; WRIGHT, Marvin N.; BOULESTEIX, Anne-Laure. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, v. 9, n. 3, p. e1301, 2019.

RIBEIRO, Thiago da Silva. MACHINE LEARNING FOR FAILURE DETECTION IN BAKERY INDUSTRIAL EFFLUENTS TREATMENT BY ELECTROCOAGULATION. 2023. Thesis (Scholarly Publication) - Pontifícia Universidade Católica do Rio de Janeiro (PUC-RIO).

SARKER, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, v. 2, n. 3, p. 160, May 2021.

SPITZNAGEL JR, Edward L. 6 logistic regression. Handbook of statistics, v. 27, p. 187-209, 2007.

VENKATASUBRAMANIAN, V. et al. A review of process fault detection and diagnosis Part I: Quantitative model-based methods. Computers and Chemical Engineering, p. 19, 2003.

WARDHANI, Ni Wayan Surya et al. Cross-validation metrics for evaluating classification performance on imbalanced data. In: 2019 international conference on computer, control, informatics and its applications (IC3INA). IEEE, 2019. p. 14-18.

WANG, L. Support vector machines: theory and applications. New York: Springer Science & Business Media, 2005. v. 177

YANG, Y.; LI, J.; YANG, Y. The research of the fast SVM classifier method. In: 12TH INTERNATIONAL COMPUTER CONFERENCE ON WAVELET ACTIVE MEDIA TECHNOLOGY AND INFORMATION PROCESSING (ICCWAMTIP). Anais… Chengdu, China: IEEE, dez. 2015. Available: http://ieeexplore.ieee.org/document/7493959/

YASIN, H. M. et al. IoT and ICT based Smart Water Management, Monitoring and Controlling System: A Review. Asian Journal of Research in Computer Science, p. 42-56, 5 May 2021.

ZHAO, L. et al. Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse. Process Safety and Environmental Protection, v. 133, p. 169-182, jan. 2020.