Comparação de Métodos de Otimização de Hiperparâmetros Baseados em Meta-Heurísticas e Estratégias Convencionais para Modelos de Detecção de Fraudes

Comparison of Metaheuristic-Based and Conventional Hyperparameter Optimization Methods for Fraud Detection Models.

Caio Emanoel Serpa Lopes[1]
orcid.org/0000-0001-7969-5754

Roberta Andrade de Araújo Fagundes[2]
orcid.org/0000-0002-7172-4183

[1]Escola Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: caio.eslopes@upe.poli.br

[1]Escola Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: roberta.fagundes@upe.poli.br

DOI: 10.25286/repa.v11i3.3971

RESUMO

A otimização de hiperparâmetros é importante na detecção de fraudes em cartões de crédito porque é um problema raro, dinâmico e altamente sensível a erros, e pequenos ajustes nos modelos podem gerar grandes diferenças nos resultados. Com isso, a otmização é crucial porque eleva o desempenho, a robustez e a capacidade de capturar padrões raros, evitando tanto perdas financeiras quanto alarmes falsos. Assim, este estudo avalia diferentes estratégias de otimização de hiperparâmetros aplicadas à detecção de fraudes em cartões de crédito, um problema marcado por forte desbalanceamento e elevada complexidade. Técnicas tradicionais são comparadas à otimização bayesiana e a meta-heurísticas bioinspiradas. Para isso, foi utilizado o conjunto de dados Credit Card Fraud Detection e modelos de machine learning. O desempenho foi mensurado por métricas adequadas a eventos raros, como F1-Score, Recall e AUC-ROC, e a significância estatística foi analisada pelo teste de Wilcoxon. A comparação entre métodos revelou diferenças relevantes e padrões que motivam uma análise mais aprofundada apresentada ao longo do estudo.

PALAVRAS-CHAVE: Aprendizagem de Máquina; Algoritmos Bio-inspirados; Detecção de anomalias; Otimização de hiperparâmetros

ABSTACT

Hyperparameter optimization is crucial in credit card fraud detection because it is a rare, dynamic, and highly error-sensitive problem, where small adjustments in the models can lead to significant differences in results. As such, optimization is essential because it enhances performance, robustness, and the ability to capture rare patterns, preventing both financial losses and false alarms. Therefore, this study evaluates different hyperparameter optimization strategies applied to credit card fraud detection, a problem characterized by severe class imbalance and high complexity. Traditional techniques are compared with Bayesian optimization and bio-inspired metaheuristics. For this purpose, the Credit Card Fraud Detection dataset and machine learning models were employed. Performance was measured using metrics suitable for rare events, such as F1-Score, Recall, and AUC-ROC, and statistical significance was analyzed using the Wilcoxon test. The comparison between methods revealed relevant differences and patterns that motivate a more in-depth analysis presented throughout the study.

KEY-WORDS: Machine Learning; Bio-inspired Algorithms; Anomaly Detection; Hyperparameter Optimization

REVISTA DE Engenharia e Pesquisa Aplicada

1 INTRODUÇÃO

A crescente digitalização dos serviços financeiros tem impulsionado de forma significativa o volume de transações eletrônicas, ampliando também a incidência e a complexidade das fraudes envolvendo cartões de crédito **[1]**. A literatura recente destaca que tais ataques tornaram-se mais sofisticados e difíceis de detectar, exigindo mecanismos de análise cada vez mais robustos e escaláveis para lidar com ambientes de alta demanda e grande fluxo de dados. Nesse contexto, a aplicação de modelos de aprendizado de máquina (Machine Learning — ML) tornou-se fundamental **[2]**. No entanto, a eficácia desses modelos está fortemente condicionada à adequada configuração de seus hiperparâmetros, que influenciam diretamente sua capacidade de generalização e sensibilidade à classe fraudulenta. Embora técnicas convencionais como *Grid Search* (Grade de busca), *Random Search* (Busca aleatória) e mesmo otimização Bayesiana sejam amplamente adotadas, desafios persistem — especialmente devido ao severo desbalanceamento dos dados, no qual a classe fraudulenta representa uma fração mínima do total de transações **[3]**.

Métodos como *Grid Search* e *Random Search* (e variações) e abordagens bayesianas clássicas têm sido utilizados para o ajuste de hiperparâmetros. Entretanto, essas estratégias apresentam limitações reconhecidas: Grid Search é computacionalmente custosa e pouco eficiente em espaços de busca extensos; Random Search, embora menos custoso, não garante uma exploração sistemática ou inteligente do espaço; e mesmo abordagens bayesianas podem encontrar dificuldades em cenários altamente desbalanceados e de alta dimensionalidade **[4]**. Como alternativa, frameworks modernos de otimização Bayesiana, como o Optuna têm ganhado destaque. Optuna permite definir o espaço de busca dinamicamente e oferece estratégias eficientes de amostragem, bem como poda de tentativas não promissoras, tornando o processo de busca mais escalável e adaptável a diferentes contextos **[5]**.

Paralelamente, cresce o interesse por meta-heurísticas bioinspiradas, como Algorítmo Genético (GA), Particle Swarm Optimization (PSO) e Grey Wolf Optimizer (GWO). Essas técnicas demonstram capacidade de explorar espaços de busca não convexos, multimodais e de alta dimensionalidade, o que permite escapar de mínimos locais e buscar soluções mais globalmente ótimas. Em particular, em tarefas de detecção de fraude — caracterizadas por forte desbalanceamento, elevada dimensionalidade e padrões dinâmicos — tais meta-heurísticas podem melhorar de modo significativo o desempenho preditivo quando combinadas a modelos supervisionados **[6]**. Diante deste panorama, o presente estudo propõe uma análise comparativa abrangente entre três paradigmas distintos de otimização de hiperparâmetros para detecção de fraudes em cartões de crédito: (i) métodos tradicionais; (ii) otimização Bayesiana via Optuna; e (iii) meta-heurísticas bioinspiradas (Algoritmos Genéticos - GA, Particle Swarm Optimization - PSO, Grey Wolf Optimization - GWO). Essas abordagens serão aplicadas ao ajuste de hiperparâmetros de modelos consolidados na literatura — como *LightGBM*, *XGBoost*, Floresta Aleatória e regressão logística — treinados sobre o conjunto de dados *Credit Card Fraud Detection* contendo mais de 280 mil transações, das quais aproximadamente 1% corresponde a fraudes **[7][8]**.

O objetivo central deste trabalho é investigar em que medida os métodos bioinspirados — incluindo Algoritmos Genéticos (GA), Particle Swarm Optimization (PSO) e Grey Wolf Optimizer (GWO) — , os métodos tradicionais de busca, como Random Search e Grid Search, e a abordagem bayesiana de otimização, representada pelo Optuna, diferem significativamente entre si no processo de otimização de hiperparâmetros. A análise é conduzida em cenários de forte desbalanceamento entre classes, nos quais a escolha adequada dos hiperparâmetros exerce papel determinante no desempenho dos modelos de aprendizado de máquina no desempenho dos modelos de ML.

A adoção de métricas robustas e sensíveis ao comportamento da classe minoritária torna-se essencial nesse contexto. Entre as medidas de desempenho mais relevantes, destacam-se o Recall, o F1-Score e a AUC-ROC, amplamente empregadas em tarefas de detecção de anomalias e fraudes no setor financeiro **[9]**.

Evidências recentes sugerem que meta-heurísticas e otimização bayesiana tendem a melhorar a detecção da classe minoritária e aumentar a estabilidade dos modelos; no entanto, comparações diretas entre essas abordagens ainda são escassas. Dessa forma, este estudo busca

preencher essa lacuna, realizando testes estatísticos para validação dos resultados e proporcionando uma análise comparativa quantitativa e transparente entre paradigmas distintos de otimização, contribuindo para o avanço de técnicas aplicadas à prevenção de fraudes financeiras.

2 TRABALHOS RELACIONADOS

A literatura sobre detecção de fraudes em cartões de crédito aponta avanços consistentes na aplicação de técnicas de aprendizado de máquina, otimização de hiperparâmetros e engenharia de atributos. No entanto, também revela limitações persistentes relacionadas ao desbalanceamento extremo dos dados, à baixa capacidade de generalização e à dificuldade de validar modelos em cenários realistas. Os trabalhos a seguir representam contribuições recentes e relevantes, cada um apresentando ganhos específicos, mas também desafios que justificam a proposta deste estudo.

Souza e Bordin Jr. **[10]** analisam a influência sistemática do ajuste de hiperparâmetros por meio de um planejamento de experimentos aplicado a modelos supervisionados de detecção de fraudes. Os autores demonstram que a performance dos modelos varia significativamente de acordo com a combinação de parâmetros, evidenciando ganhos de até 12% em F1-score quando a calibração é feita adequadamente. A principal limitação do estudo reside no uso de um conjunto reduzido de algoritmos — concentrando-se majoritariamente em árvores de decisão — o que impede conclusões mais amplas sobre modelos de maior complexidade, como XGBoost e LightGBM.

Santos et al. **[11]**, por sua vez, exploram a geração de dados sintéticos como estratégia para melhorar a avaliação de classificadores e detectores de anomalias. O estudo mostra que técnicas como SMOTE e variações avançadas podem ampliar o desempenho dos modelos em cenários desbalanceados, aumentando a sensibilidade sem prejudicar drasticamente a especificidade. Entretanto, os autores destacam que datasets sintéticos podem introduzir viés adicional, levando a modelos superajustados ao conjunto artificial, o que reduz a capacidade de generalização quando aplicados a bases reais de transações financeiras.

Complementando essa perspectiva, Cavalcanti, Brandão e Bezerra **[12]** conduzem uma avaliação comparativa de estratégias de balanceamento e seleção de atributos aplicadas à bases desbalanceadas. Os autores concluem que métodos combinados — como SMOTE + Relief-F — resultam em melhorias consistentes no desempenho de modelos como Random Forest e SVM, especialmente nas métricas de AUC e recall da classe minoritária. Apesar disso, o estudo apresenta como limitação a ausência de técnicas modernas de otimização, restringindo o *tuning* dos hiperparâmetros a valores empíricos, o que pode ter reduzido o desempenho máximo alcançável pelos modelos analisados.

No contexto de seleção de atributos, Alkurdi et al. **[13]** investigam o uso da correlação ponto-bisserial para identificar variáveis mais relevantes na detecção de fraude. Os resultados mostram que essa técnica reduz significativamente a dimensionalidade sem perdas substanciais de desempenho, além de melhorar a interpretabilidade do modelo. No entanto, o método apresenta limitações quando utilizado em bases com relações não lineares ou interações complexas, uma vez que a correlação ponto-bisserial é essencialmente linear.

Já Alzahrani et al. **[14]** aplicam meta-heurísticas — incluindo algoritmos genéticos e PSO — ao processo de seleção de atributos combinado com classificadores Random Forest e SVM. O estudo evidencia que essas técnicas podem gerar subconjuntos de atributos mais eficientes e melhorar o desempenho em até 15% em bases altamente desbalanceadas. Contudo, o trabalho destaca que tais abordagens apresentam elevado custo computacional e podem sofrer variações consideráveis entre execuções, dada sua natureza estocástica.

Por fim, Marazqah et al. **[15]** apresentam uma revisão sistemática de abordagens de machine learning aplicadas à detecção de fraudes. A análise aponta que modelos baseados em boosting e redes neurais tendem a superar técnicas lineares, mas reforça que a falta de padronização nos protocolos experimentais dificulta comparações diretas entre estudos. Além disso, os autores destacam a escassez de análises sobre técnicas avançadas de otimização de hiperparâmetros — incluindo métodos bayesianos e bioinspirados — revelando uma lacuna importante para pesquisas futuras.

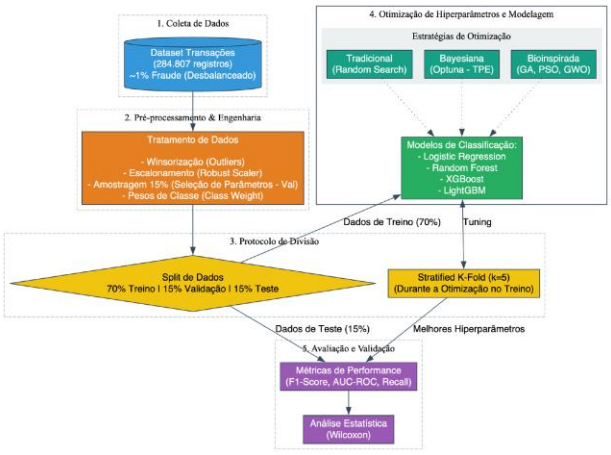
DOI: 10.25286/repa.v11i3.3971

De forma geral, os trabalhos existentes mostram avanços importantes, mas também deixam claro que ainda há espaço para uma avaliação comparativa sistemática entre métodos tradicionais, bayesianos e meta-heurísticos de otimização aplicados à detecção de fraude. O presente estudo busca justamente preencher essa lacuna, oferecendo uma análise integrada e padronizada desses paradigmas sobre modelos amplamente utilizados no setor financeiro.

3 METODOLOGIA

A metodologia, baseada no CRISP-DM **[16]** e organizada em cinco etapas (Figura 1), iniciou-se pela coleta e definição do problema (1), estabelecendo o foco na detecção de fraudes e nas métricas de interesse. Em seguida, foram realizados o pré-processamento e a exploração dos dados (2), incluindo análise das variáveis e seleção de atributos. O protocolo de divisão (3) estruturou o conjunto em treino, validação e teste. Na etapa de modelagem e otimização (4), aplicaram-se Regressão Logística, Random Forest, XGBoost e LightGBM, combinados a diferentes estratégias de ajuste de hiperparâmetros — *Grid Search*, *Random Search*, otimização bayesiana e meta-heurísticas bioinspiradas. Por fim, a avaliação (5) utilizou métricas adequadas a eventos raros, como F1-Score, Sensibilidade e AUC-ROC, além do teste de Wilcoxon para verificar significância estatística entre os métodos.

Figura 1 – Fluxograma metodológico do projeto



Fonte: Os Autores.

3.1 CONJUNTO DE DADOS

Para a condução dos experimentos, foi utilizado o conjunto de dados "*Credit Card Fraud Detection*" **[8]**, disponível publicamente no repositório Kaggle. Este *dataset* compreende transações financeiras com cartão de crédito, registradas ao longo de um período de 48 horas. Com um total de 284.807 instâncias, a distribuição das classes é altamente assimétrica, contendo apenas 492 transações fraudulentas, o que corresponde a aproximadamente 0.172% do total. Esta característica define um cenário de desbalanceamento severo, que representa um dos principais desafios para a construção de modelos preditivos robustos neste domínio.

3.1.1 TRATAMENTO DE DADOS

O processo de preparação dos dados foi conduzido com o objetivo de reduzir ruídos, mitigar a influência de valores extremos e padronizar as variáveis antes da aplicação dos modelos de aprendizado de máquina, etapa reconhecida como crítica para a qualidade das inferências estatísticas e preditivas **[17]**. Inicialmente, a variável alvo foi convertida para o tipo inteiro, garantindo consistência semântica com o problema de classificação supervisionada e compatibilidade com os algoritmos utilizados.

Na sequência, realizou-se uma seleção preliminar de atributos com base na correlação em relação à variável de saída, a partir do cálculo da matriz de correlação do conjunto de dados **[18]**. Com isso, foram escolhidas as 15 variáveis com maior correlação absoluta com a classe, excluindo-se o próprio atributo alvo, de modo a reduzir a dimensionalidade e concentrar o treinamento em características potencialmente mais informativas para a tarefa de classificação.

Considerando que bases de dados de fraude tipicamente apresentam ruídos e valores extremos, empregou-se a técnica de "winsorização", limitando os valores observados nos percentis inferiores e superiores (5% em cada extremidade) **[19]**. Essa abordagem é descrita na literatura como uma forma de atenuar o impacto de outliers sobre estimadores sem descartar observações, preservando o tamanho amostral e a estrutura geral dos dados.

Posteriormente, os atributos numéricos foram escalonados utilizando o "*RobustScaler*", cujo procedimento se baseia na mediana e no intervalo interquartil (IQR) para transformar as variáveis **[20]**. Por empregar medidas robustas de posição e dispersão, esse tipo de escalonamento mostra-se

mais resistente à presença de valores discrepantes do que métodos baseados em média e desvio padrão, sendo particularmente adequado para dados financeiros assimétricos **[21]**.

O pipeline completo de tratamento de dados contemplou: (i) conversão da variável alvo para o tipo inteiro; (ii) seleção das 15 variáveis mais correlacionadas com a variável alvo; (iii) "winsorização" dos atributos numéricos nos percentis de 5% inferiores e superiores; e (iv) escalonamento robusto com base no IQR. Esse conjunto de pré-processamentos visou estabilizar a distribuição dos preditores e favorecer melhor desempenho e robustez dos modelos de aprendizado de máquina aplicados ao problema de detecção de fraude.

3.2 MODELOS AVALIADOS

O Para a avaliação do desempenho preditivo na detecção de inadimplência, foram selecionados quatro modelos amplamente discutidos na literatura e reconhecidos por sua eficácia em tarefas de classificação supervisionada: Regressão Logística, *Random Forest*, XGBoost e LightGBM. A escolha desses algoritmos se justifica por representarem distintas famílias de modelos — lineares, baseados em árvores e métodos de *gradient boosting* — possibilitando uma análise comparativa entre abordagens tradicionais e técnicas mais avançadas.

A Regressão Logística, enquanto método estatístico consolidado para problemas de classificação binária, é recorrentemente empregada em análises de risco de crédito devido à sua interpretabilidade, simplicidade e baixo custo computacional **[22]**. O *Random Forest*, por sua vez, fundamenta-se na combinação de múltiplas árvores de decisão via *bagging*, proporcionando maior robustez, redução de sobreajuste e boa capacidade de generalização, mesmo em cenários com dados ruidosos ou de alta variabilidade **[23]**. Já os algoritmos de *boosting*, como XGBoost e LightGBM **[24]**, destacam-se por alcançarem desempenho superior em diversos *benchmarks*, impulsionados pela otimização baseada em gradiente, estratégias avançadas de regularização e uso eficiente dos recursos computacionais.

Considerando essa diversidade metodológica, este estudo busca comparar modelos tradicionais, *ensembles* e técnicas de *boosting*, avaliando seu desempenho por meio de métricas amplamente aceitas na literatura, tais como sensibilidade, F1-score e AUC, especialmente relevantes em cenários de desbalanceamento.

No processo de otimização dos modelos, foram definidos hiperparâmetros específicos para cada algoritmo. Para o LightGBM, foram ajustados n_estimators (50–150), learning_rate (0,01–0,1), num_leaves (31–63) e class_weight (balanced ou None) **[25]**. No XGBoost, analisaram-se n_estimators (50–150), learning_rate (0,01–0,1), max_depth (3–6) e scale_pos_weight, configurado como 1 ou como a razão entre instâncias das classes positiva e negativa **[26]**. Para o Random Forest, os hiperparâmetros considerados foram n_estimators (50–150), max_depth (None, 10 ou 20) e class_weight (balanced ou None) **[27]**. Por fim, na Regressão Logística, foram avaliados C (0,1–10), penalty (l1 ou l2), solver (liblinear) e class_weight (balanced ou None) **[28]**.

3.3 ESTRATÉGIAS DE OTIMIZAÇÃO DE HIPERPARÂMETROS

A performance de algoritmos de aprendizado de máquina depende fortemente da escolha adequada de seus hiperparâmetros. Para garantir comparabilidade e maximizar o desempenho dos modelos avaliados, foram utilizadas cinco estratégias de otimização: Algoritmos Genéticos (GA), Otimização por Lobos-Cinzentos (GWO), Enxame de Partículas (PSO), *Random Search* e Otimização Bayesiana com Optuna. Estas técnicas permitem explorar de maneira eficiente o espaço de busca, aumentando a probabilidade de encontrar combinações de hiperparâmetros capazes de elevar a qualidade preditiva dos modelos **[29]**.

3.3.1 Otimização por algoritmos Genéticos (GA)

Os Algoritmos Genéticos são metaheurísticas inspiradas na seleção natural, capazes de explorar espaços complexos de hiperparâmetros por meio de operações como seleção, cruzamento e mutação. GA é especialmente útil em cenários onde o espaço de busca é grande e potencialmente irregular, evitando convergência prematura e permitindo maior diversidade de soluções **[30]**.

3.3.2 Otimização por Lobos-Cinzentos (GWO)

DOI: 10.25286/repa.v11i3.3971

O algoritmo dos lobos-cinzentos (GWO) baseia-se na estrutura social e comportamentos de caça dos lobos-cinzentos, sendo amplamente reconhecido por sua simplicidade e boa capacidade de balancear exploração e explotação no processo de otimização **[31]**. Seu desempenho tem sido demonstrado em problemas de classificação, agrupamento e ajuste de hiperparâmetros, oferecendo convergência estável e competitiva.

3.3.3 Otimização por Enxame de Partículas (PSO)

O Particle Swarm Optimization (PSO), inspirado no comportamento coletivo de bandos de pássaros e cardumes de peixes, realiza a otimização através da atualização iterativa das posições e velocidades das partículas em busca da melhor solução **[32]**. PSO é amplamente utilizado por sua rápida convergência e facilidade de implementação, sendo eficaz no ajuste de hiperparâmetros de modelos supervisionados.

3.3.4 Otimização Grade Aleatória (*Random Search*)

O *Random Search* foi proposto como alternativa mais eficiente que o *grid search* tradicional, ao amostrar aleatoriamente combinações de hiperparâmetros, permitindo explorar regiões mais amplas do espaço e aumentando as chances de encontrar configurações superiores. Sua simplicidade e baixo custo computacional tornam-no uma técnica amplamente utilizada na prática **[33]**.

3.3.5 Otimização Bayesiana (OPTUNA)

A otimização bayesiana busca modelar o desempenho esperado de diferentes hiperparâmetros por meio de funções de aquisição e modelos probabilísticos, concentrando as buscas em regiões promissoras. No presente estudo, utilizou-se a biblioteca Optuna, reconhecida por sua eficiência, automação e suporte a *pruning* dinâmico. Essa abordagem permite encontrar configurações otimizadas com menor número de avaliações, sendo altamente adequada para modelos complexos como *boosting* **[34]**.

3.4 MÉTODOS DE AVALIAÇÃO

A avaliação dos modelos otimizados foi realizada por meio de validação cruzada e comparação baseada em métricas amplamente utilizadas em cenários com classes desbalanceadas:

- F1-score, adequado para cenários onde precisão e recall precisam ser equilibrados **[35]**;

$$F_1 = 2 \times \frac{Precisão \times Sensbilidade}{Precisão + Sensibilidade} \quad \textbf{(1)}$$

- AUC-ROC, apropriado para medir a capacidade discriminativa do modelo independentemente do ponto de corte **[36]**;

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN} \quad AUC = \int_0^1 ROC(x)dx \quad \textbf{(2)}$$

- A Sensibilidade (Recall), essencial dado o baixo volume de fraudes no conjunto de dados **[37]**.

$$Sensibilidade = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos} \quad \textbf{(3)}$$

3.5 TESTE ESTATÍSTICOS

A comparação entre diferentes seletores de hiperparâmetros exige métodos estatísticos adequados, especialmente em cenários onde os modelos são avaliados sob validação cruzada. Nessa configuração, as observações são pareadas, não independentes e as métricas utilizadas geralmente não seguem distribuição normal. Para lidar com essas características, este estudo emprega o *Wilcoxon Signed-Rank Test*, amplamente recomendado na literatura recente de avaliação de modelos de *machine learning* **[38]**.

3.5.1 Justificativa para Uso do Teste de Wilcoxon

As métricas utilizadas — AUC, F1-score, Sensibilidade e Precisão — possuem limites fixos e distribuições assimétricas, sendo inadequadas para testes paramétricos que assumem normalidade **[39]**. Trabalhos recentes demonstram que métricas de desempenho em aprendizado de máquina dificilmente seguem distribuição normal, reforçando a necessidade de testes não paramétricos.

Assim, o Wilcoxon *Signed-Rank* é apropriado por não exigir normalidade, nem homogeneidade de variâncias **[40]**.

A validação cruzada gera dependência entre *folds* devido ao compartilhamento de parte do conjunto de dados. Conforme discutido por Raschka, essa

dependência viola premissas fundamentais de testes paramétricos tradicionais **[41]**.

O teste de Wilcoxon, ao operar sobre pares correspondentes de resultados *fold* a *fold*, respeita essa estrutura de dependência.

Ao considerar sinais e *rankings* das diferenças, o teste verifica se um método supera o outro de maneira consistente, e não apenas em média **[42]**.

4 RESULTADOS E CONCLUSÕES

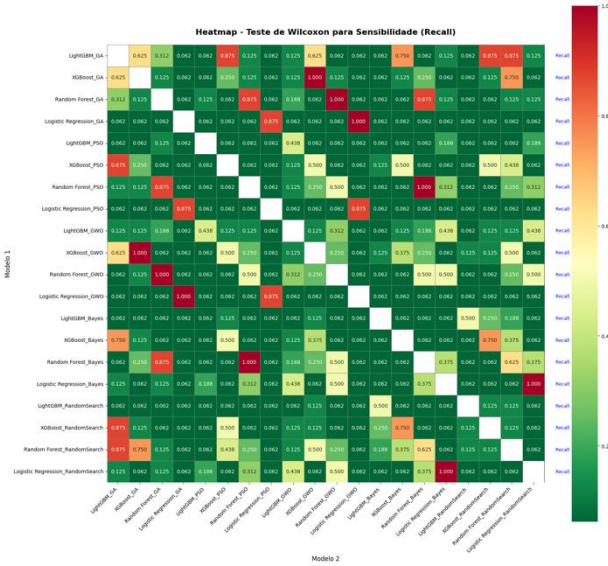
A partir dos experimentos realizados, observou-se que os modelos baseados em árvores de decisão — especialmente XGBoost, LightGBM e *Random Forest* — apresentaram desempenho superior nas métricas F1-Score, AUC-ROC e Sensibilidade. O XGBoost otimizado por Algoritmo Genético (GA) destacou-se com F1-Score de 0,7832 e AUC-ROC de 0,973, enquanto o *Random Forest* otimizado por GA atingiu o maior F1-Score do conjunto (0,8028). Esses resultados refletem a capacidade desses modelos de capturar relações não lineares e padrões complexos típicos de cenários de fraude, demonstrando maior robustez e eficiência em comparação aos modelos lineares.

A análise dos otimizadores revelou que o GA produziu, de forma consistente, as melhores combinações de hiperparâmetros para modelos baseados em árvores, enquanto PSO e GWO apresentaram desempenho mais instável. A abordagem bayesiana Optuna mostrou-se uma alternativa competitiva, fornecendo resultados estáveis e superiores ao *Random Search* na maioria dos cenários. Além disso, verificou-se que a escolha dos hiperparâmetros é crucial para a previsão de fraude, pois influencia a capacidade do modelo de detectar classes raras, equilibrar falsos positivos e negativos, captar padrões atípicos e manter viabilidade computacional. Assim, a otimização de hiperparâmetros configura-se como etapa essencial para construir sistemas robustos, confiáveis e escaláveis.

As métricas utilizadas forneceram uma visão completa do desempenho dos modelos. O F1-Score permitiu avaliar o equilíbrio entre precisão e sensibilidade, essencial para evitar tanto alarmes falsos quanto fraudes não detectadas. A Sensibilidade destacou a habilidade dos modelos em identificar eventos fraudulentos em um cenário de classe rara, enquanto a AUC-ROC avaliou a robustez das decisões sob diferentes limiares, reduzindo o risco de interpretações enganosas. Em conjunto, essas métricas garantiram uma avaliação abrangente e adequada à natureza do problema de fraude.

Por fim, os testes estatísticos de *Wilcoxon Signed-Rank* ($\alpha = 0.10$) **[43]** indicaram que, embora os modelos otimizados por GA apresentem desempenho numericamente superior, especialmente em *F1-Score* e AUC, a maior parte das comparações apresentou valores de p acima de 0.10, sugerindo ausência de diferenças estatisticamente significativas na maioria dos casos. Ainda assim, observou-se uma tendência consistente de superioridade dos modelos baseados em árvores combinados com GA, indicando que, embora nem sempre estatisticamente significativas, as melhorias obtidas são sistemáticas e relevantes do ponto de vista prático. Dessa forma, conclui-se que modelos de *boosting* otimizados por GA constituem a alternativa mais promissora para detecção de fraude, equilibrando desempenho, estabilidade e capacidade de generalização.

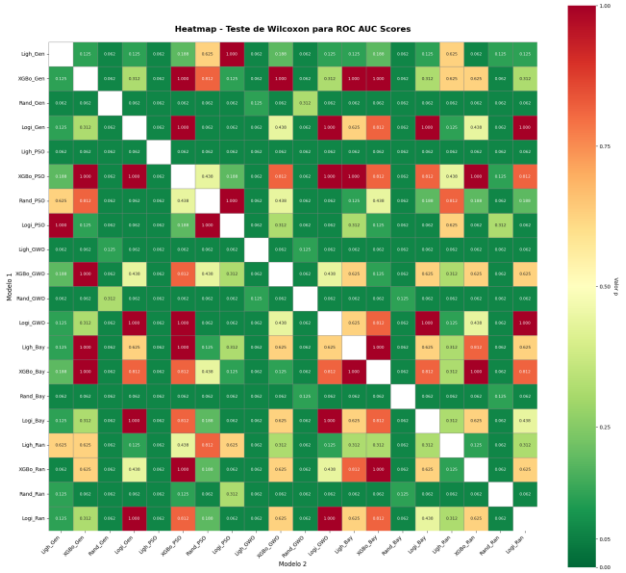
Figura 2 – Mapa de calor do teste de Wilcoxon para a métrica sensibilidade.



Fonte: Os Autores.

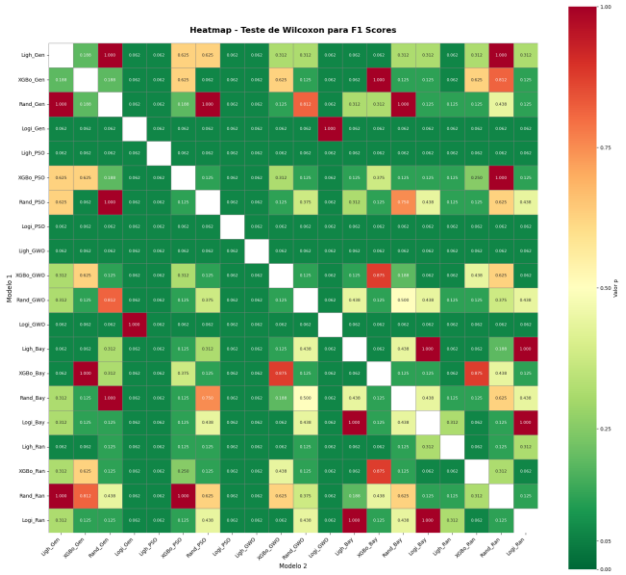
Figura 3 – Mapa de calor do teste de Wilcoxon para a métrica AUC-ROC.

DOI: 10.25286/repa.v11i3.3971



Fonte: Os Autores.

Figura 4 – Mapa de calor do teste de Wilcoxon para a métrica F1-Score.



Fonte: Os Autores.

Assim, conclui-se que a combinação entre modelos baseados em árvores e otimizadores bio-inspirados, em especial o algoritmo genético, constitui a estratégia mais eficiente para o contexto analisado, tanto em termos de desempenho preditivo quanto de significância estatística. Esses achados contribuem para o avanço das aplicações de métodos bio-inspirados em problemas de crédito e inadimplência, fornecendo evidências empíricas e estatísticas robustas que podem orientar futuras pesquisas e aplicações práticas no setor financeiro.

REFERÊNCIAS

[1] CAVALCANTI, A.; BRANDÃO, D.; BEZERRA, E.; COUTINHO, R. **Avaliação de Técnicas de Balanceamento de Dados na Detecção de Fraude em Transações Online de Cartão de Crédito.** In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD), 39., 2024, Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2024. DOI: 10.5753/sbbd.2024.243462.

[2] LOU, Y.; LIU, J.; SHENG, Y.; WANG, J.; ZHANG, Y.; REN, Y. **Addressing Class Imbalance with Probabilistic Graphical Models and Variational Inference.** arXiv, [S. l.], 2025. Disponível em: https://doi.org/10.48550/arXiv.2504.05758. Acesso em:20/11/2025.

[3] BTOUSH EAL, et al. **A systematic review of literature on credit card cyber fraud detection using machine and deep learning.** 2023 doi:10.7717/peerj-cs.1278

[4] PATHAK, A. K. et al. **Randomized-Grid Search for Hyperparameter Tuning in Decision Tree Model to Improve Performance of Cardiovascular Disease Classification.** S. l., 2024. Disponível em: arxiv:2411.18234

[5] OAKIBA, T.; SANO, S.; YANASE, T.; OHTA, T.; KOYAMA, M. **Optuna: a next-generation hyperparameter optimization framework.** In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 2019. 2019. DOI: 10.1145/3292500.3330701.

[6] AYOUB, M. et al. Granular computing **framework for credit card fraud detection**. Alexandria Engineering Journal, 2025. Disponível em: https://www.sciencedirect.com/science/article/pii/S1110016825001863.

[7] SURE, P.; PANDEY, S.; PARNAMI, T.; SAXENA, A. **Feature selection techniques for enhancing credit card fraud detection performance: a hybrid metaheuristic approach using nature-inspired algorithms.** IJRASET Journal for Research in Applied Science and Engineering Technology, 2024. DOI: 10.22214/ijraset.2024.58549.

[8] ULB, Machine Learning Group. **Credit Card Fraud Detection.** Kaggle, 2023. Disponível em: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. Acesso em: 14/10/2025

[9] BHATTACHARYA, S. et al. **Fraud detection using optimized machine learning tools under imbalanced classes.** Engineering Applications of Artificial Intelligence, v. 120, p. 106013, 2023.

[10] SOUZA, A. M.; BORDIN JR., L. **Planejamento de Experimentos Aplicado à Detecção de Fraude em Cartão de Crédito Utilizando Aprendizado Supervisionado: Uma Abordagem Metodológica.** Revista Brasileira de Computação Aplicada, v. 15, n. 3, 2023.

[11] SANTOS, A. et al. **Geração de Dados Sintéticos para Avaliação de Modelos de Classificação e Detecção de Anomalias Voltados à Detecção de Fraude em Cartão de Crédito.** Anais do Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg), 2024. Sociedade Brasileira de Computação.

[12] CAVALCANTI, M.; BRANDÃO, A.; BEZERRA, E. **Avaliação de Estratégias de Balanceamento e Seleção de Atributos com Classificadores em Bases Desbalanceadas.** Anais do Simpósio Brasileiro de Banco de Dados (SBBD), 2024. Sociedade Brasileira de Computação.

[13] ALKURDI, M. et al. **Credit Card Fraud Detection using Point-Biserial Correlation Feature Selection.** Revista Gestão & Tecnologia, v. 24, n. 2, 2024.

[14] ALZAHRANI, A. et al. **Meta-Heuristic Optimization for Feature Selection in Credit Card Fraud Detection Using Random Forest and SVM.** Mathematics, v. 12, n. 14, 2250, 2024. MDPI. DOI: 10.3390/math12142250.

[15] MARAZQAH BTOUSH, EYAD ABEL LATIF et al. **A systematic review of literature on credit card cyber fraud detection using machine and deep learning.** PeerJ. Computer science vol. 9 e1278. 17 Apr. 2023, doi:10.7717/peerj-cs.1278

[16] SHIMAOKA, A. M.; FERREIRA, R. C.; GOLDMAN, A. **The evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks.** SBC Computing Reviews, v. 4, n. 1, p. 28–43, 2024. DOI: https://doi.org/10.5753/reviews.2024.3757.

[17] WILCOX, R. R. **Winsorize**. Encyclopedia of Research Design, 2024.

[18] ZHANG, Y. et al. **Hybrid feature selection framework for enhanced credit card fraud detection**. PLOS ONE, 2024.

[19] HAWKINS, D. **Outlier Impact and Accommodation Methods: Multiple Comparisons.** Wayne State University, 2024.

[20] AMORIM, L. B. V. et al. **The choice of scaling technique matters for classification performance.** Applied Soft Computing, v. 133, p. 109924, jan. 2023. Sociedade Brasileira de Computação.

[21] HAMPEL, F. R. et al. **Robust Statistics: The Approach Based on Influence Functions.** John Wiley & Sons, 1986.

[22] HOSMER, D.; LEMESHOW, S.; STURDIVANT, R. **Applied Logistic Regression.** Wiley, 2013.

[23] BREIMAN, L. **Random Forests.** Machine Learning, 2001. Springer.

[24] CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System.** Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2016. ACM.

[25] KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T. **LightGBM: A Highly Efficient Gradient Boosting Decision Tree.** Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS), 2017. MIT Press.

[26] BERGSTRA, J.; BENGIO, Y. **Random Search for Hyper-Parameter Optimization.** Journal of Machine Learning Research (JMLR), 2012. Microtome Publishing.

[27] ZHANG, L. et al. **Improved LightGBM for extremely imbalanced data and application to credit card fraud detection.** Expert Systems with Applications, v. 242, p. 122987, 2025.

[28] WANG, Y.; NI, X. S. **A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization.** arXiv preprint, arXiv:1901.08433, 2019.

[29] ABURBEIAN, A. H.; ASHQAR, H. I. **Credit card fraud detection using enhanced Random Forest classifier for imbalanced data.** arXiv preprint, arXiv:2303.06514, 2023.

[30] KHEKARE, G.; SUNDA, S.; BOTHRA, Y. **A comprehensive performance comparison of traditional and ensemble machine learning models for online fraud detection.** arXiv preprint, arXiv:2509.17176, 2025.

[31] MITCHELL, M. An **Introduction to Genetic Algorithms.** MIT Press, 1998.

[32] MIRJALILI, S.; MIRJALILI, S. M.; LEWIS, A. **Grey Wolf Optimizer.** Advances in Engineering Software, 2014. Elsevier.

[33] KENNEDY, J.; EBERHART, R. **Particle Swarm Optimization.** Proceedings of the IEEE International Conference on Neural Networks (ICNN), 1995. IEEE.

[34] AKIBA, T.; SANO, S.; YANASE, T.; OHTA, T.; KOHN, K. **Optuna: A Next-Generation Hyperparameter Optimization Framework.** Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2019. ACM.

[35] GARCÍA, V.; SÁEZ, J. A.; HERRERA, F. **An Updated Review on Imbalanced Data Learning: Taxonomy, Challenges, and Solutions.** ACM Computing Surveys, v. 54, n. 2, p. 1–36, 2021.

[36] SOKOLOVA, M.; LAPLANTE, F. **On the Use of ROC Analysis for Binary Classification under Class Imbalance.** Information Processing & Management, v. 58, n. 3, p. 102–152, 2021.

[37] KRAWIEC, K. et al. **Handling Imbalanced Data in Classification: A Review**. Expert Systems with Applications, v. 206, p. 117–240, 2022.

[38] BENAVOLI, A.; CORANI, G.; MANGILI, F. **A New Look at Statistical Comparison of Machine Learning Algorithms: Bayesian Perspective.** Machine Learning, v. 109, p. 199–223, 2020.

[39] DERRAC, J.; GARCÍA, S.; MOLINA, D.; HERRERA, F. **A practical tutorial on the use of nonparametric statistical tests for evaluating machine learning algorithms.** Information Sciences, v. 512, p. 248-284, 2020.

[40] GARCÍA, S.; FERNÁNDEZ, A.; LUENGO, J.; HERRERA, F. **A tutorial on the use and a ssessment of statistical tests in machine learning.** Journal of Systems and Software, v. 162, p. 110516, 2019.

[41] BRASCHKA, S. **Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning.** arXiv preprint arXiv:1811.12808, 2020.

[42] ZHANG, Y.; YANG, Q.; LI, X. **Cross-validation and statistical testing in machine learning experiments.** ACM Computing Surveys, v. 54, n. 6, p. 1-36, 2021.

[43] GARCÍA, V.; SÁNCHEZ, J. S.; MARTÍN-FÉLEZ, R.; MOLLINEDA, R. A. **Surrounding neighborhood-based SMOTE for learning from imbalanced data sets.** Progress in Artificial Intelligence, Springer, 2012.