Avaliação de Algoritmos de Classificação em Dados Desbalanceados sob o Trade-off entre Precisão e *Recall*

Caio Serpa[1], **Roberta Fagundes**[1]

[1]Universidade de Pernambuco (UPE) – Recife – PE – Brazil

{cesl@ecomp.poli.br, roberta.fagundes@upe.br}

Resumo. *O desbalanceamento de classes é um desafio central na classificação supervisionada. Este estudo analisa quatro algoritmos (SVM, Árvore de Decisão, BRF e XGBoost) em oito cenários de desbalanceamento (IR 3,91–129,53). Sob protocolo estatístico de Wilcoxon com correção de Bonferroni-Holm, os resultados comprovam a superioridade do XGBoost em Recall ($p_{adj} \leq 0{,}047$, rank 1,06), tornando-o ideal para aplicações onde falsos negativos são críticos. Quanto à Precisão, XGBoost, BRF e SVM apresentam equivalência estatística, mas o XGBoost destaca-se por sua alta eficiência computacional frente aos demais. O estudo conclui que a escolha do classificador deve ser rigorosamente orientada pelo custo do erro e pela viabilidade temporal do domínio.*

1. Introdução

Em domínios críticos como diagnósticos médicos, detecção de fraudes e controle industrial, a sub-representação da classe minoritária impõe um viés sistemático aos classificadores tradicionais, comprometendo sua utilidade prática [Khedmati et al. 2024, Liaw et al. 2025]. Nesse cenário desbalanceado, a seleção de métricas como Precisão e *Recall* é determinante, pois reflete diretamente o custo assimétrico dos erros de predição na aplicação final [Khedmati et al. 2024]. Por exemplo, no contexto médico, um falso negativo gera risco à vida (priorizando o *Recall*); já em detecção de fraudes, o excesso de falsos positivos eleva custos operacionais (exigindo otimização da Precisão).

Este trabalho apresenta uma análise comparativa rigorosa do *trade-off* entre eficácia preditiva e eficiência computacional. Avaliam-se dois modelos baseados em algoritmo isolado (SVM com Penalização e Árvore de Decisão) e dois em *ensemble* (Florestas Aleatórias Balanceadas - BRF e XGBoost) sobre oito conjuntos de dados com Razão de Desbalanceamento (*Imbalance Ratio* - IR) entre 3,91 e 129,53. As divergências de desempenho são validadas pelo teste de Wilcoxon com correção de Bonferroni-Holm [Demšar 2006], fornecendo um arcabouço estruturado para guiar a seleção de modelos com base no impacto financeiro ou operacional do erro e nas restrições de tempo de execução.

2. Fundamentação Teórica

A Razão de Desbalanceamento (IR) quantifica a disparidade entre as classes majoritária e minoritária, gerando fronteiras de decisão enviesadas quando tratadas por algoritmos convencionais [Chamlal et al. 2024]. Estratégias em nível de algoritmo mitigam esse efeito penalizando assimetricamente os erros da minoria, ao passo que abordagens de *ensemble* utilizam subamostragem interna ou reforço adaptativo para diversificar as hipóteses avaliadas [Zhang et al. 2022].

Nesse cenário, ocorre o paradoxo da acurácia: ao prever apenas a classe majoritária, o modelo atinge alto acerto global, mas falha completamente na detecção da classe rara. Assim, a acurácia torna-se uma métrica ilusória, exigindo o monitoramento isolado de Precisão e *Recall* para quantificar as taxas reais de falsos positivos e negativos [Liaw et al. 2025]. Para garantir a validade científica das comparações, adota-se o protocolo estatístico não paramétrico de Demšar [Demšar 2006]: aplicação do teste emparelhado de postos sinalizados de Wilcoxon seguida pela correção de Bonferroni-Holm para controle da taxa de erro familiar (α=0,05).

3. Metodologia

3.1. Conjuntos de Dados

A avaliação experimental baseia-se em oito conjuntos de dados heterogêneos extraídos do repositório UCI [Dai et al. 2024], acrescidos de um cenário sintético regulado. A Tabela 1 detalha os ambientes de teste ordenados por suas características estruturais e níveis de raridade da classe minoritária.

Tabela 1. Conjuntos de dados utilizados na avaliação.

#	Dataset	N	Attr.	IR
1	Sintético (95-5)	1.000	10	19,00
2	Abalone-19	4.177	10	129,53
3	Yeast-ME2	1.484	8	28,10
4	Mammography	11.183	6	42,01
5	Oil Spill	937	49	21,85
6	Thyroid Sick	3.772	28	15,33
7	Car Evaluation	1.728	6	3,91
8	Wine Quality	6.497	12	25,77

IR: Razão de Desbalanceamento; N: número de amostras.

3.2. Modelos, Hiperparâmetros e Protocolo

Os modelos foram construídos em linguagem Python via `scikit-learn` e `XGBoost`, executados em ambiente de múltiplos núcleos. Aplicou-se `StandardScaler` nos métodos lineares e `LabelEncoder` na variável alvo. O ajuste de custos (`class_weight='balanced'`) remediou o desbalanceamento no SVM (kernel RBF) e na Árvore de Decisão; no XGBoost (100 estimadores), o parâmetro `scale_pos_weight` foi definido como a razão entre as classes de cada base. O BRF utilizou sua subamostragem nativa. A semente aleatória foi fixada (*random_state*=108 para modelos; 42 para dados).

O desempenho foi avaliado via validação cruzada estratificada (k=10), extraindo-se a média e o desvio padrão das partições para aferir a robustez. Os ensaios utilizaram processamento paralelo (`n_jobs=-1`), alinhando-se a princípios de Computação de Alto Desempenho (HPC) para a aferição realista do custo temporal (*fit time* e *score time*). Diferenças estatísticas sistemáticas foram confrontadas pelo *rank* médio e testes de Wilcoxon com correção de Bonferroni-Holm (α=0,05).

4. Resultados e Discussão

4.1. *Rank* Médio e Significância Estatística

A Tabela 2 consolida a consistência relativa dos classificadores ao longo dos experimentos através de seu *rank* médio geral.

Tabela 2. *Rank* médio por métrica (1 = melhor posição relativa).

Modelo	Recall	Precisão
XGBoost	**1,06**	**1,62**
Decision Tree	2,06	3,12
SVM Penalization	3,00	2,88
BRF	3,88	2,38

O XGBoost demonstrou dominância nas duas vertentes avaliadas, estabelecendo vantagem expressiva em *Recall* (rank 1,06). Esse comportamento decorre de seu gradiente iterativo que, parametrizado pelo `scale_pos_weight`, eleva o peso dos erros sobre a minoria a cada rodada de *boosting*. O BRF destacou-se exclusivamente em Precisão (rank 2,38), demonstrando que a subamostragem randômica por árvore delimita bem a classe majoritária, reduzindo falsos positivos. A Árvore de Decisão registrou bom *Recall* relativo (2,06) penalizado pelo pior escore de Precisão (3,12), fruto do sobreajuste e memorização de ruídos locais da classe rara.

A Tabela 3 restringe a análise empírica aos pares que exibiram divergências estatisticamente significativas após a aplicação da correção de Bonferroni-Holm.

Tabela 3. Pares com diferença significativa – Wilcoxon + Bonferroni-Holm (α=0,05).

Métrica	Par	p bruto	p adj.
Recall	SVM vs. XGBoost	0,0078	0,0312
	DT vs. BRF	0,0078	0,0312
	BRF vs. XGBoost	0,0078	0,0312
	DT vs. XGBoost	0,0156	0,0469
Precisão	DT vs. XGBoost	0,0078	0,0469

DT: *Decision Tree*; BRF: *Balanced Random Forest*. Par listado: $p_{\text{adj}}<0{,}05$.

Em *Recall*, o XGBoost confirmou-se estatisticamente superior a todos os demais algoritmos ($p_{\text{adj}}\leq 0{,}047$), validando sua robustez adaptativa. A Árvore de Decisão individual superou significativamente o BRF ($p_{\text{adj}}=0{,}031$), evidenciando que o processo de subamostragem agressivo do BRF pode descartar instâncias limítrofes valiosas da minoria. Em Precisão, apenas a Árvore de Decisão divergiu negativamente do líder ($p_{\text{adj}}=0{,}047$), mantendo o BRF e o SVM em patamar de equivalência estatística ao XGBoost ($p_{\text{adj}}\geq 0{,}39$).

4.2. Desempenho Absoluto e Impacto do IR

A Tabela 4 reporta as métricas absolutas médias obtidas por modelo em cada cenário computado. Os baixos índices de desvio padrão monitorados nas partições de validação atestaram a estabilidade das médias apresentadas.

Tabela 4. Precisão e *Recall* médios por *dataset*.

	XGBoost		BRF		SVM		DT	
Dataset	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**
Sint. (95-5)	0,976	0,977	0,955	0,888	0,978	0,976	0,938	0,941
Abalone-19	0,986	0,989	0,991	0,703	0,988	0,805	0,985	0,985
Yeast-ME2	0,961	0,960	0,965	0,819	0,960	0,876	0,960	0,960
Mammography	0,986	0,986	0,979	0,926	0,981	0,952	0,981	0,982
Oil Spill	0,964	0,964	0,959	0,835	0,957	0,923	0,945	0,945
Thyroid Sick	0,987	0,987	0,977	0,968	0,953	0,899	0,987	0,987
Car Evaluation	1,000	1,000	0,986	0,977	0,994	0,991	0,999	0,999
Wine Quality	0,956	0,962	0,960	0,816	0,952	0,864	0,952	0,955

BRF: *Balanced Random Forest*; DT: *Decision Tree*.

Sob desbalanceamento reduzido (Car Evaluation, IR=3,91), nota-se a convergência de todos os classificadores para scores ótimos próximos a 1,00. Contudo, em cenários críticos como Abalone-19 (IR=129,53, contendo apenas 32 instâncias positivas), o BRF colapsou em *Recall* (0,703) devido ao descarte severo de amostras raras na subamostragem de seus estimadores. O SVM também exibiu degradação proeminente (*Recall*=0,805) pela instabilidade geométrica gerada na fixação de suas margens em raridade extrema. O XGBoost, inversamente, manteve robustez elevada com ambas as métricas superiores a 0,98, evidenciando a eficiência do ajuste dinâmico de pesos acoplado à minimização do gradiente iterativo.

4.3. Análise de Custo Computacional

A viabilidade de implantação prática de sistemas preditivos requer a avaliação de sua sobrecarga temporal. A Tabela 5 apresenta a média empírica dos tempos de treinamento (*Fit*) e inferência (*Score*) executados em arquitetura paralela, cujos baixos desvios padrão observados atestam a confiabilidade desta métrica.

Tabela 5. Tempos médios de treinamento e predição por algoritmo (em segundos).

Modelo	**Treinamento (*Fit*) - s**	**Predição (*Score*) - s**
Decision Tree	0,031	0,004
XGBoost	0,095	0,009
Balanced Random Forest	0,353	0,036
SVM Penalization	2,611	0,091

A Árvore de Decisão registrou a menor demanda computacional isolada (0,031 s para ajuste), contudo, seu severo deficit em Precisão inviabiliza sua escolha. O XGBoost despontou como o modelo de maior eficiência prática ao aliar seu elevado ganho preditivo a tempos de processamento altamente competitivos (0,095 s em treino e 0,009 s em predição), impulsionado por suas otimizações em nível de sistema, como computação paralela de nós e uso eficiente de *cache* de *hardware*. Em aplicações modernas, como APIs de inferência em tempo real ou sistemas embarcados de borda (*edge computing*), atrasos na ordem de milissegundos podem degradar a experiência do usuário ou inviabilizar processos industriais síncronos. Nesse aspecto, a eficiência do XGBoost o torna uma escolha pragmática ideal, pois entrega excelência preditiva sem onerar as limitações de latência ou infraestrutura da arquitetura.

O BRF exigiu tempos moderadamente superiores (0,353 s em *fit*), pois a construção paralela de múltiplas árvores independentes anula os ganhos temporais da subamostragem. Por fim, o SVM Penalization revelou-se o modelo mais custoso e de difícil escalabilidade (2,611 s em treino, superando 11 segundos em bases densas como *Mammography*), reflexo direto da complexidade polinomial requerida para solucionar o problema dual de otimização de suas margens separadoras.

5. Conclusão

Os experimentos conduzidos evidenciam que a seleção de classificadores sob desbalanceamento severo deve integrar não apenas o desempenho estatístico, mas também as restrições computacionais do ecossistema de aplicação. Sob essa ótica, o XGBoost qualifica-se como a solução mais robusta, uma vez que combina uma superioridade estatisticamente significativa em *Recall* ($p_{\text{adj}} \leq 0{,}047$) a uma excelente eficiência de processamento (0,095 s). Por sua vez, métodos como o BRF e o SVM atuam com eficácia equivalente na contenção de falsos positivos (Precisão); contudo, o SVM exibe fortes limitações de escalabilidade temporal para volumes maiores de dados (2,611 s). Em contrapartida, as Árvores de Decisão isoladas demonstram-se inviáveis operacionalmente devido ao severo comprometimento de sua Precisão ($p_{\text{adj}} = 0{,}047$). Dessa forma, o mapeamento estruturado deste *trade-off* fornece um arcabouço validado para guiar escolhas algorítmicas alinhadas estritamente ao custo financeiro do erro e à infraestrutura de tempo real disponível. Trabalhos futuros englobarão o aprendizado profundo e métodos de otimização automatizada, investigando se seus ganhos preditivos justificam o incremento de complexidade computacional.

Referências

Chamlal, H., Kamel, H., and Ouaderhman, T. (2024). A hybrid multi-criteria meta-learner based classifier for imbalanced data. *Knowledge-Based Systems*, 285:111367.

Dai, Q. et al. (2024). Class-overlap detection based on heterogeneous clustering ensemble for multi-class imbalance problem. *Expert Systems with Applications*, 255:124558.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Khedmati, M. et al. (2024). A cluster-based smote both-sampling (csbboost) ensemble algorithm for classifying imbalanced data. *Scientific Reports*, 14.

Liaw, L. C. M., Tan, S. C., Goh, P. Y., and Lim, C. P. (2025). A histogram smote-based sampling algorithm with incremental learning for imbalanced data classification. *Information Sciences*, 686:121193.

Zhang, J., Wang, T., Ng, W. W., and Pedrycz, W. (2022). Ensembling perturbation-based oversamplers for imbalanced datasets. *Neurocomputing*, 479:1–11.