ANÁLISE DA ROBUSTEZ DE ALGORITMOS DE APRENDIZADO DE MÁQUINA FRENTE A DADOS RUIDOSOS: UM ESTUDO COM O DATASET IRIS

Gabriel Nunes de Lara (UEPG) E-mail: gabrielnlara@gmail.com
Luis Roberto Molotto (UEPG) E-mail: molotto.luis@gmail.com
Victor Angelo Legat Cerqueira (UEPG) E-mail: victor.legat.cerqueira@gmail.com

Resumo: Este estudo analisa a robustez dos algoritmos *K-Nearest Neighbors* (*KNN*), *Decision Tree, Random Forest* e *Support Vector Machine* (*SVM*) frente à inserção progressiva de ruído gaussiano nos atributos do *dataset* Iris. O protocolo experimental utilizou validação cruzada estratificada em cinco dobras e avaliou acurácia, *recall* e *F1-score*. Para verificar se as diferenças observadas entre os modelos eram estatisticamente sustentáveis, aplicou-se ANOVA de uma via em cada nível de ruído. Os resultados evidenciam degradação gradual de desempenho em todos os classificadores. Em termos práticos, *SVM* e *Random Forest* apresentaram menor perda relativa, enquanto *KNN* e *Decision Tree* foram mais sensíveis. Contudo, não foram identificadas diferenças estatisticamente significativas entre os modelos ($p > 0,62$). Conclui-se que a baixa dimensionalidade, o pequeno tamanho amostral e a elevada separabilidade do Iris comprimem as diferenças entre classificadores, reforçando a importância de combinar métricas preditivas com validação estatística em estudos de *benchmarking*.

Palavras-chave: Aprendizado de máquina, robustez, ruído gaussiano, análise estatística, dataset Iris.

Abstract: *This study investigates the robustness of K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM) classifiers under progressively injected Gaussian noise in the Iris dataset. The experimental protocol employed stratified 5-fold cross-validation and evaluated accuracy, macro recall, and macro F1-score. One-way ANOVA was used to test whether the observed differences among the models were statistically significant at each noise level. The results show a gradual performance decline for all classifiers as noise increases. In practical terms, SVM and Random Forest displayed lower degradation, whereas KNN and Decision Tree were more sensitive. However, no statistically significant differences were found among the models ($p > 0.62$). These findings suggest that the small size, low dimensionality, and high class separability of the Iris dataset compress performance differences among classifiers, highlighting the importance of combining predictive metrics with inferential statistical analysis in benchmarking studies.*

Keywords: *Machine Learning, robustness, gaussian noise, Iris dataset, statistical analysis*.

1. Introdução

A classificação é uma das tarefas mais tradicionais do aprendizado de máquina, com aplicações em reconhecimento de padrões, diagnóstico, triagem e apoio à decisão. Entre os conjuntos de dados mais utilizados para fins didáticos e comparativos está o *dataset* Iris, proposto por Ronald Fisher em 1936, composto por 150 amostras balanceadas de três espécies de flores e descritas por quatro atributos numéricos (Fisher, 1936). Pela sua simplicidade e ampla adoção no ensino e na experimentação inicial, o Iris permanece como referência recorrente em estudos de classificação supervisionada (Dani; Artanta Ginting, 2024).

A Figura 1 apresenta as três subespécies tradicionalmente associadas ao *dataset* Iris, frequentemente utilizadas como exemplo introdutório em tarefas de classificação (Mithy et al., 2022).

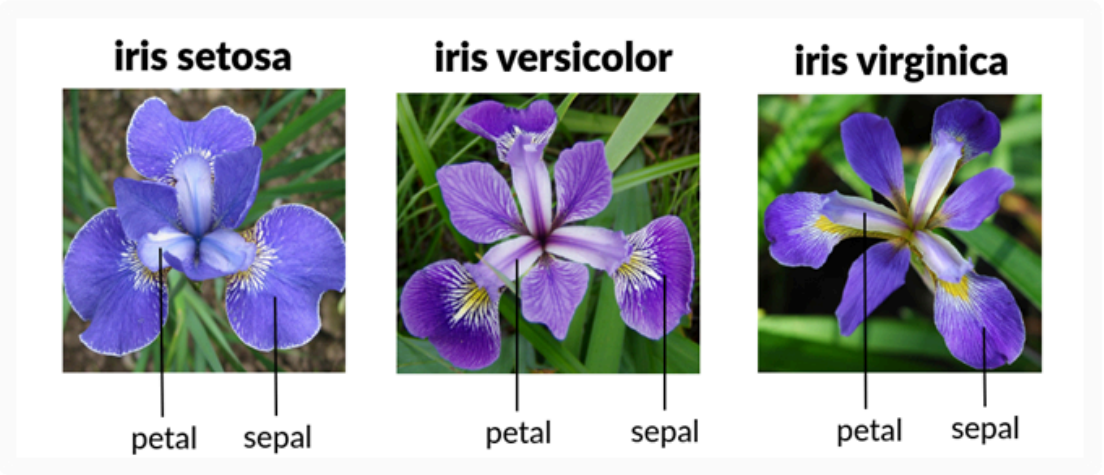

Figura 1 - Três subespécies da flor IRIS (Fonte: Mithy et al., 2022)

Embora benchmarks controlados sejam úteis para comparar modelos, a literatura recente mostra que o desempenho observado em condições ideais nem sempre se mantém quando os dados apresentam imperfeições. Em aplicações reais, sensores podem introduzir erros de medição, processos de coleta podem gerar inconsistências e atributos numéricos podem sofrer perturbações que comprometem a estabilidade das predições. Nesse contexto, robustez pode ser entendida como a capacidade de um modelo manter comportamento estável diante de alterações não adversariais nos dados de entrada (Padró-Ferragut; Martínez-Plumed; Ramírez-Quintana, 2026; Giudici; Raffinetti; Riani, 2025).

A literatura sobre robustez em aprendizado de máquina tem avançado tanto em avaliações amplas envolvendo múltiplos datasets quanto em discussões sobre confiabilidade de modelos em ambientes ruidosos. Ainda assim, permanece relevante realizar estudos controlados e reprodutíveis que examinem, em um benchmark amplamente conhecido, se diferenças aparentes entre classificadores clássicos permanecem quando submetidas a perturbações graduais e à validação estatística. Essa abordagem é particularmente importante em estudos de *benchmarking*, nos quais pequenas diferenças visuais entre métricas podem ser interpretadas de forma excessiva.

Neste contexto, o presente estudo analisa a robustez dos algoritmos *K-Nearest Neighbors*, *Decision Tree*, *Random Forest* e *Support Vector Machine* diante da inserção progressiva de ruído gaussiano nos atributos do dataset Iris. Esses modelos foram escolhidos por representarem estratégias distintas de classificação: o *KNN* depende de relações locais de distância; a *Decision Tree* realiza partições hierárquicas do espaço de atributos; o *Random Forest* combina múltiplas árvores para reduzir variância (Breiman, 2001); e o *SVM* busca hiperplanos de margem máxima entre classes (Cortes; Vapnik, 1995). O diferencial do trabalho reside na combinação entre inserção controlada de ruído, comparação entre modelos clássicos e uso de testes inferenciais para distinguir diferenças práticas de diferenças estatisticamente sustentáveis.

2. Metodologia

Os experimentos foram implementados em *Python 3.12*, com apoio do ecossistema científico formado por *scikit-learn*, *pandas*, *numpy*, *scipy*, *statsmodels*, *matplotlib* e *seaborn*. A biblioteca *scikit-learn* 1.5.0 foi empregada tanto para o

carregamento do *dataset* quanto para a modelagem e a validação cruzada, devido à sua ampla adoção em experimentos reprodutíveis de aprendizado de máquina (Pedregosa et al., 2011).

O estudo utilizou o *dataset* Iris, disponibilizado pela própria *scikit-learn*, contendo 150 amostras distribuídas igualmente entre três classes e descritas por quatro atributos numéricos. Para cada nível de ruído, foi gerada inicialmente uma versão perturbada do conjunto completo segundo a expressão $X_{ruidoso} = X + \varepsilon$, em que ε segue distribuição normal com média zero e desvio padrão igual ao nível de ruído adotado. Foram analisados cinco níveis de perturbação, com σ igual a 0,0; 0,1; 0,2; 0,3; e 0,4. A semente aleatória foi fixada em 42 para garantir reprodutibilidade.

Os algoritmos avaliados foram *K-Nearest Neighbors*, *Decision Tree*, *Random Forest* e *Support Vector Machine*. Para os modelos sensíveis à escala dos atributos, *KNN* e *SVM*, aplicou-se o *StandardScaler* dentro de pipelines de treinamento. Os hiperparâmetros adotados foram: *KNN* com *n_neighbors* = 5, *weights* = *uniform* e distância de Minkowski com p = 2; *Decision Tree* com *criterion* = *gini* e *splitter* = *best*; *Random Forest* com *n_estimators* = 100, *criterion* = *gini* e *max_features* = *sqrt*; e *SVM* com *kernel* = *rbf*, C = 1,0 e *gamma* = *scale*. Nos modelos baseados em árvore foi utilizado *random_state* = 42. Optou-se por hiperparâmetros padrão, com exceção da fixação das sementes aleatórias, para manter uma linha de base reprodutível e homogênea entre os classificadores. Essa escolha favorece comparabilidade, mas pode limitar o desempenho absoluto de alguns modelos.

A avaliação foi conduzida por validação cruzada estratificada com cinco dobras, embaralhamento dos dados e preservação da proporção entre classes em cada partição, por meio de *StratifiedKFold*(*n_splits* = 5, *shuffle* = *True*, *random_state* = 42). As métricas analisadas foram acurácia, *recall* macro e *F1-score* macro. A opção por métricas macro se justifica pelo caráter multiclasse do problema e pela necessidade de preservar a contribuição de cada classe na avaliação média (Nti; Nyarko-Boateng; Aning, 2021).

Para verificar se as diferenças observadas entre os modelos excediam a variabilidade esperada entre as dobras da validação cruzada, aplicou-se o teste ANOVA de uma via para cada métrica e para cada nível de ruído. O teste de Tukey foi previsto como procedimento pós-hoc apenas nos casos em que a ANOVA indicasse significância estatística, conforme prática recomendada em comparações múltiplas de médias (Juarros-Basterretxea, 2024). Como nenhum dos cenários apresentou significância, a interpretação inferencial concentrou-se nos resultados da ANOVA.

Embora o *dataset* Iris seja adequado como *benchmark* introdutório, sua baixa dimensionalidade, o pequeno tamanho amostral e a elevada separabilidade entre classes restringem a generalização dos achados para cenários mais complexos. Além disso, como o estudo adotou hiperparâmetros padrão e avaliou apenas ruído gaussiano aditivo, os resultados devem ser interpretados como evidência controlada em um cenário didático, e não como uma hierarquia universal de robustez entre classificadores.

3. Análise dos resultados

Os experimentos realizados com o *dataset* Iris revelaram degradação progressiva de desempenho em todos os algoritmos à medida que o nível de ruído aumentou. As Figuras 2 e 3 sintetizam a evolução da acurácia e do *F1-score* ao longo dos cinco cenários analisados. Com dados sem ruído, o *KNN* apresentou a maior acurácia média

(0,9733), seguido por *SVM* (0,9600), *Decision Tree* (0,9533) e *Random Forest* (0,9467). No cenário mais severo, com 40% de ruído, o *SVM* passou a apresentar o melhor desempenho médio (0,8867), seguido por *Random Forest* (0,8667), *KNN* (0,8600) e *Decision Tree* (0,8400).

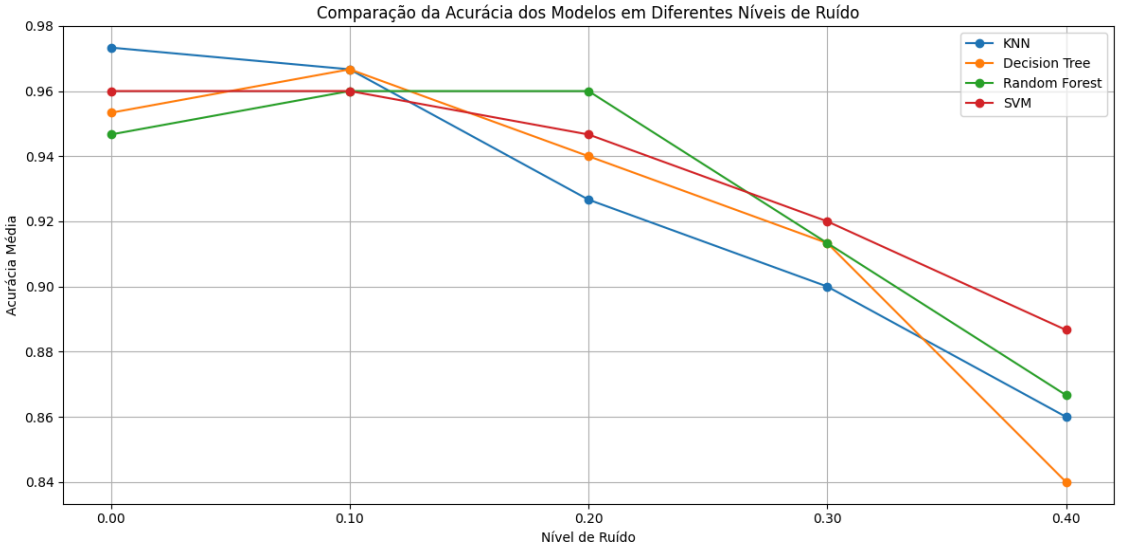

Figura 2 - Comparação da Acurácia dos Modelos em Diferentes Níveis de Ruído (Fonte: Os Autores.)

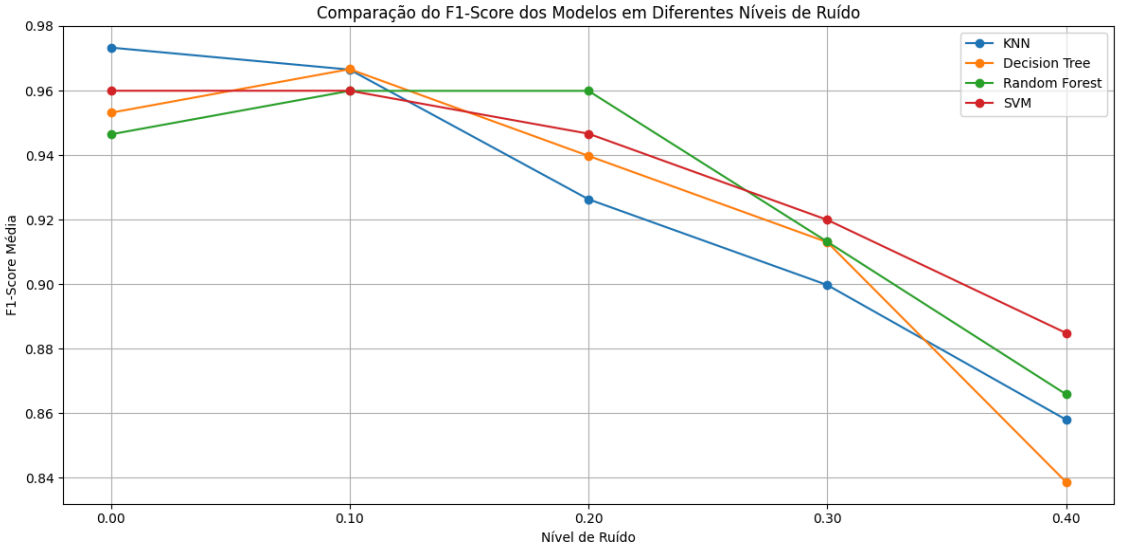

Figura 3 - Comparação do F1 dos Modelos em Diferentes Níveis de Ruído (Fonte: Os Autores.)

A leitura conjunta das curvas indica que todos os modelos são afetados pela deterioração da qualidade dos atributos, porém em magnitudes distintas. Também se observa forte semelhança entre os comportamentos de acurácia, *recall* e *F1-score*, o que é coerente com o balanceamento entre as classes do Iris e com a natureza relativamente estável dos erros produzidos pelos classificadores ao longo dos experimentos.

Para uma análise quantitativa da sensibilidade ao ruído, a Tabela 1 resume a degradação percentual entre os cenários de 0% e 40% de ruído. Em termos práticos, o *SVM* foi o modelo com menor perda relativa, seguido pelo *Random Forest*. *KNN* e

Decision Tree apresentaram quedas mais acentuadas, sugerindo maior vulnerabilidade a perturbações nos atributos de entrada.

Tabela 1 – Análise da Degradação com Aumento do Ruído

MODELO	ACURÁCIA	RECALL	F1-SCORE
KNN	11,64%	11,64%	11,84%
Decision Tree	11,89%	11,89%	12,00%
Random Forest	8,45%	8,45%	8,51%
SVM	7,64%	7,64%	7,82%

Fonte: Os autores.

Os resultados práticos permitem algumas interpretações. O *SVM* tende a ser favorecido por sua estratégia de maximização de margem, o que o torna menos suscetível a pequenas oscilações nos atributos quando as classes já apresentam separabilidade considerável. O *Random Forest*, por sua vez, se beneficia do efeito de ensemble, reduzindo a variância típica de árvores individuais. Em contraste, o *KNN* depende diretamente da geometria local dos dados e pode alterar suas decisões com pequenas mudanças de distância, enquanto a *Decision Tree* é mais sensível a perturbações que afetam pontos de corte em atributos específicos.

Entretanto, essa hierarquia observada na prática precisa ser interpretada com cautela. A Tabela 2 apresenta os p-valores obtidos pela ANOVA em cada nível de ruído e para cada métrica. Em todos os casos, os valores permaneceram muito acima do nível de significância de 5%, variando de 0,6272 a 0,9923. Portanto, não foi possível rejeitar a hipótese nula de igualdade entre as médias dos modelos em nenhum dos cenários analisados.

Tabela 2 – p-valores da ANOVA por nível de ruído e métrica

NÍVEL DE RUÍDO	ACURÁCIA	RECALL	F1-SCORE
0%	0,6755	0,6755	0,6718
10%	0,9921	0,9921	0,9923
20%	0,6860	0,6860	0,6828
30%	0,8542	0,8542	0,8529
40%	0,6272	0,6272	0,6511

Fonte: Os autores.

Esse resultado é central para a interpretação do estudo. Embora existam diferenças visíveis nas curvas e nas taxas de degradação, elas não se mostraram estatisticamente sustentáveis no contexto do Iris. Em outras palavras, o experimento sugere tendências práticas de robustez, mas não evidencia superioridade inferencial entre os classificadores. Tal achado reforça a necessidade de combinar análise descritiva e validação estatística em estudos de benchmarking, especialmente quando se trabalha com conjuntos pequenos e relativamente simples.

Há pelo menos quatro fatores que ajudam a explicar a ausência de significância estatística. Primeiro, o Iris possui apenas 150 amostras, o que reduz o poder estatístico das comparações. Segundo, o problema apresenta alta separabilidade entre classes, comprimindo as diferenças de desempenho entre modelos clássicos. Terceiro, o conjunto tem baixa dimensionalidade, o que reduz a chance de surgirem padrões de sensibilidade mais complexos ao ruído. Quarto, a escolha de hiperparâmetros padrão, embora metodologicamente útil para estabelecer uma linha de base comparável, pode atenuar diferenças que se tornariam mais evidentes após ajuste fino.

A literatura recente corrobora essa leitura. Padró-Ferragut, Martínez-Plumed e Ramírez-Quintana (2026), ao analisarem robustez sob perturbação de atributos em múltiplos datasets e famílias de modelos, observaram que diferenças entre algoritmos tendem a se tornar mais claras quando o cenário experimental é mais diverso e menos didático. De modo convergente, Giudici, Raffinetti e Riani (2025) argumentam que robustez deve ser analisada em conjunto com desempenho e confiabilidade estatística. Assim, o principal mérito deste estudo não está em eleger um vencedor absoluto no Iris, mas em demonstrar que diferenças observacionais nem sempre se convertem em diferenças estatisticamente demonstráveis.

4. Considerações finais

Este estudo analisou, em ambiente controlado, o impacto da inserção de ruído gaussiano sobre o desempenho de quatro algoritmos clássicos de classificação no *dataset* Iris. Como contribuição principal, o trabalho apresenta um protocolo reprodutível que combina inserção gradual de ruído, validação cruzada estratificada, múltiplas métricas de desempenho e verificação inferencial por ANOVA.

Os resultados mostraram que todos os modelos sofreram degradação com o aumento do ruído. Em termos práticos, *SVM* e *Random Forest* apresentaram menor perda relativa, enquanto *KNN* e *Decision Tree* se mostraram mais sensíveis. Ainda assim, a conclusão mais importante do estudo é que essas diferenças não foram estatisticamente significativas no contexto analisado. Dessa forma, os achados não sustentam uma superioridade definitiva entre os classificadores no *dataset* Iris, mas evidenciam a importância de interpretar resultados de *benchmarking* com cautela e suporte inferencial.

Do ponto de vista aplicado, os resultados sugerem que *SVM* e *Random Forest* podem ser escolhas promissoras quando se espera degradação moderada na qualidade dos atributos, embora essa indicação deva ser entendida como tendência prática e não como regra geral. A baixa dimensionalidade do Iris, seu pequeno tamanho amostral, a elevada separabilidade entre classes, o uso de hiperparâmetros padrão e a consideração exclusiva de ruído gaussiano aditivo limitam a generalização dos resultados para problemas reais mais complexos.

Referências

BREIMAN, Leo. *Random forests. Machine Learning*, v. 45, n. 1, p. 5-32, 2001.

CORTES, Corinna; VAPNIK, Vladimir. *Support-vector networks. Machine Learning*, v. 20, n. 3, p. 273-297, 1995.

DANI, Yasi; ARTANTA GINTING, Maria. *Comparison of Iris dataset classification with Gaussian naïve Bayes and decision tree algorithms*. International Journal of Electrical and Computer Engineering (IJECE), v. 14, n. 2, p. 1959, 1 abr. 2024.

FISHER, Ronald A. *The use of multiple measurements in taxonomic problems*. Annals of Eugenics, v. 7, n. 2, p. 179-188, 1936.

GIUDICI, Paolo; RAFFINETTI, Emanuela; RIANI, Marco. *Robust machine learning models: linear and nonlinear*. International Journal of Data Science and Analytics, v. 20, p. 1043-1050, 2025.

JUARROS-BASTERRETXEA, Joel. *Post-hoc tests in one-way ANOVA: The case for normal distribution*. Methodology, v. 20, n. 2, p. 84–99, 28 jun. 2024.

MITHY, S. A.. *Classification of Iris Flower Dataset using Different Algorithms*. v. 9, 2022.

NTI, Isaac Kofi; NYARKO-BOATENG, Owusu; ANING, Justice. *Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation*. International Journal of Information Technology and Computer Science, v. 13, n. 6, p. 61–71, 8 dez. 2021.

PADRÓ-FERRAGUT, Cristina; MARTÍNEZ-PLUMED, Fernando; RAMÍREZ-QUINTANA, María José. *Robustness under noise: assessing the impact of perturbed key attributes on machine learning models*. International Journal of Data Science and Analytics, v. 22, art. 71, 2026.

PEDREGOSA, F. et al. *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.