

# Avaliação Comparativa de Algoritmos de Aprendizado Supervisionado em Tarefas de Regressão e Classificação Multiclasse

1º Matheus Coelho

Departamento de Ciência da Computação  
Centro Federal de Educação Tecnológica de Minas Gerais  
Belo Horizonte, Brasil  
matheuscoelho060@gmail.com

2º Rodrigo Silva

Departamento de Ciência da Computação  
Centro Federal de Educação Tecnológica de Minas Gerais  
Belo Horizonte, Brasil  
rodrigopiece27@gmail.com

**Abstract**—Este trabalho apresenta uma avaliação comparativa de algoritmos de aprendizado supervisionado aplicados a tarefas de regressão e classificação multiclasse. Foram utilizados os conjuntos de dados Energy Efficiency, California Housing, Wine Quality (Red) e Iris, com diferentes técnicas de pré-processamento e validação. Seis algoritmos de regressão e cinco de classificação foram analisados quanto ao desempenho, robustez e interpretabilidade, utilizando métricas como RMSE, MAE, acurácia e F1-score, com 30 repetições para garantir robustez estatística nos algoritmos estocásticos e de classificação. Os resultados evidenciam as vantagens e limitações de cada abordagem, destacando a importância do pré-processamento e da escolha adequada do modelo para cada tarefa.

**Index Terms**—Regressão, Classificação Multiclasse, Avaliação de Algoritmos, Validação Estatística

## I. INTRODUÇÃO

O aprendizado supervisionado tem se destacado como uma das principais abordagens em ciência de dados, permitindo a resolução de problemas complexos em diversas áreas, como energia, saúde e indústria [1]. A escolha do algoritmo e das técnicas de pré-processamento pode impactar significativamente o desempenho dos modelos, especialmente em tarefas de regressão e classificação multiclasse [2]. Este trabalho tem como objetivo comparar diferentes algoritmos supervisionados em cenários distintos, analisando não apenas o desempenho, mas também aspectos como robustez e interpretabilidade.

## II. METODOLOGIA

Foram conduzidos dois experimentos principais: um de regressão e outro de classificação multiclasse. Em ambos, buscou-se seguir boas práticas de ciência de dados, incluindo a divisão adequada dos dados, repetição dos experimentos com diferentes sementes e o uso de métricas apropriadas para cada tarefa.

### A. Bases de Dados

Foram utilizados os seguintes conjuntos de dados:

- **Energy Efficiency** [3]: tarefa de regressão.
- **California Housing** [4]: tarefa de regressão.
- **Wine Quality (Red)** [5]: classificação multiclasse.
- **Iris** [6]: classificação multiclasse.

### B. Pré-processamento

Para os dados numéricos, foi aplicada padronização (*StandardScaler*) para garantir que todas as variáveis tivessem média zero e desvio padrão um, fundamental para algoritmos sensíveis à escala [1]. Para dados categóricos, utilizou-se *LabelEncoder* quando necessário [2].

### C. Configuração Experimental

Na regressão, os dados foram divididos em 70% treino e 30% teste; na classificação, 80% treino e 20% teste, com estratificação das classes. Para algoritmos estocásticos, foram realizadas 30 repetições variando a semente aleatória, conforme recomendado para avaliação robusta [7]. Os hiperparâmetros seguiram os valores padrão do *scikit-learn*, exceto quando indicado.

### D. Algoritmos supervisionados utilizados

**Regressão Linear (e Ridge):** A Regressão Linear modela relações lineares entre variáveis, sendo um ponto de partida simples e interpretável. A variante Ridge adiciona uma penalidade para evitar o sobreajuste [11].

**Random Forest (Regressão e Classificação):** Random Forest é um algoritmo robusto que combina múltiplas árvores de decisão para alta precisão e flexibilidade, além de identificar a importância das variáveis. [8]

**SVR e SVC (Máquinas de Vetores de Suporte):** SVMs (SVR para regressão, SVC para classificação) buscam um hiperplano ideal com a maior margem entre os dados, oferecendo uma abordagem geométrica poderosa. [12]

**MLP (Perceptron de Múltiplas Camadas):** MLP é uma rede neural artificial que aprende padrões complexos e não lineares, representando a abordagem de deep learning para comparar com modelos tradicionais. [13]

**KNN (K-Vizinhos Mais Próximos):** KNN é um algoritmo simples e intuitivo que prevê com base nos "K" vizinhos mais próximos, sendo não paramétrico e útil para dados sem distribuição específica. [14]

**KNN com Ponderação por Distância (Fuzzy-like KNN):** Esta variação do KNN pondera a influência dos vizinhos pela

distância, dando mais peso aos mais próximos, e foi incluída como um sistema de classificação com inspiração fuzzy. [15]

### III. EXPERIMENTOS COMPUTACIONAIS

#### A. Experimento 1: Regressão

Foram avaliados seis algoritmos: Regressão Linear, Ridge Regression, Random Forest Regressor, SVR, MLP Regressor e KNN Regressor. Os experimentos foram realizados tanto no Energy Efficiency quanto no California Housing. A principal métrica utilizada foi o RMSE, complementada por MAE e  $R^2$ . Os resultados médios e desvios padrão são apresentados nas Tabelas I e II.

TABLE I  
RESULTADOS DE RMSE (MÉDIA  $\pm$  DESVIO PADRÃO, QUANDO APLICÁVEL) PARA CADA ALGORITMO DE REGRESSÃO

Algoritmo	Energy Efficiency	California Housing
Regressão Linear	$2.9731 \pm 0.000$	$0.7284 \pm 0.000$
Ridge Regression	$2.9786 \pm 0.000$	$0.7284 \pm 0.000$
Random Forest	$0.4894 \pm 0.0035$	$0.5050 \pm 0.0014$
SVR	$2.8285 \pm 0.000$	$0.5891 \pm 0.000$
MLP Regressor	$1.6279 \pm 0.2649$	$0.5152 \pm 0.0056$
KNN Regressor	$2.4309 \pm 0.000$	$0.6507 \pm 0.000$

TABLE II  
RESULTADOS DE MAE PARA CADA ALGORITMO DE REGRESSÃO (MÉDIA  $\pm$  DESVIO PADRÃO, QUANDO APLICÁVEL)

Algoritmo	Energy Efficiency	California Housing
Regressão Linear	$2.1551 \pm 0.000$	$0.5272 \pm 0.000$
Ridge Regression	$2.1529 \pm 0.000$	$0.5272 \pm 0.000$
Random Forest	$0.3379 \pm 0.0031$	$0.3315 \pm 0.0011$
SVR	$1.9373 \pm 0.0000$	$0.3963 \pm 0.000$
MLP Regressor	$1.1387 \pm 0.1655$	$0.3507 \pm 0.0080$
KNN Regressor	$1.6060 \pm 0.000$	$0.4431 \pm 0.000$

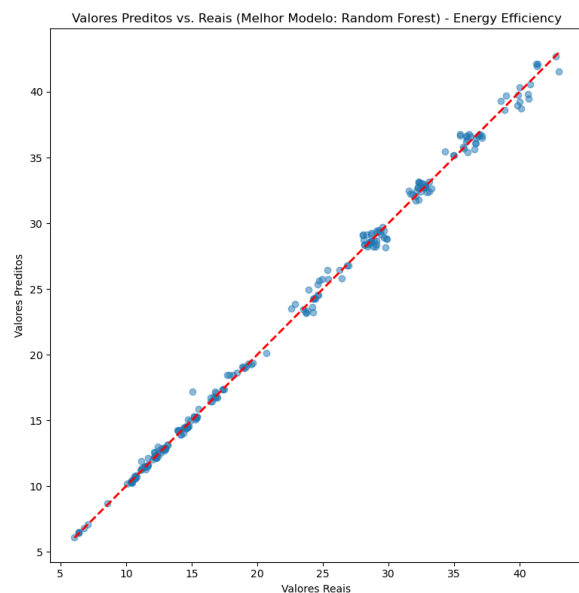


Fig. 2. Random Forest (melhor desempenho) para o dataset Energy Efficient

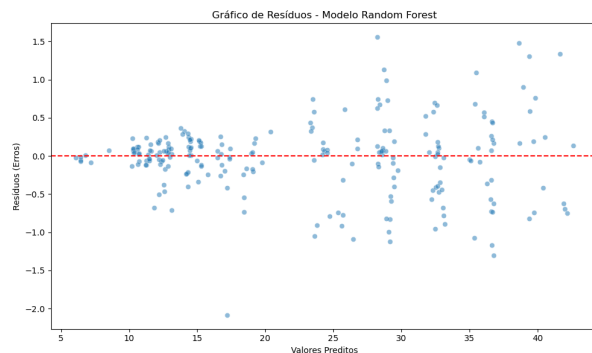


Fig. 3. Gráfico de resíduos para o modelo Random Forest no dataset Energy Efficient

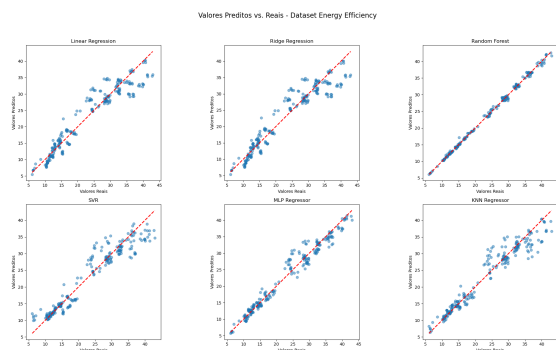


Fig. 1. Todos os gráficos de dispersão dos modelos utilizados no dataset Energy Efficient

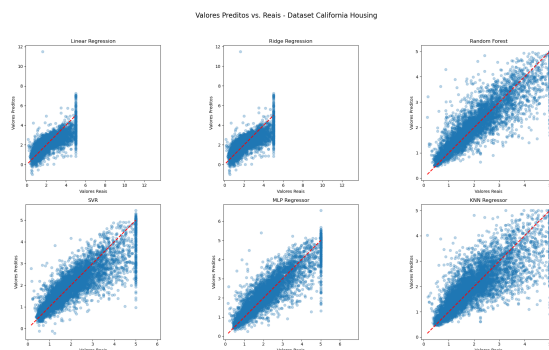


Fig. 4. Todos os gráficos de dispersão dos modelos utilizados no dataset California Housing

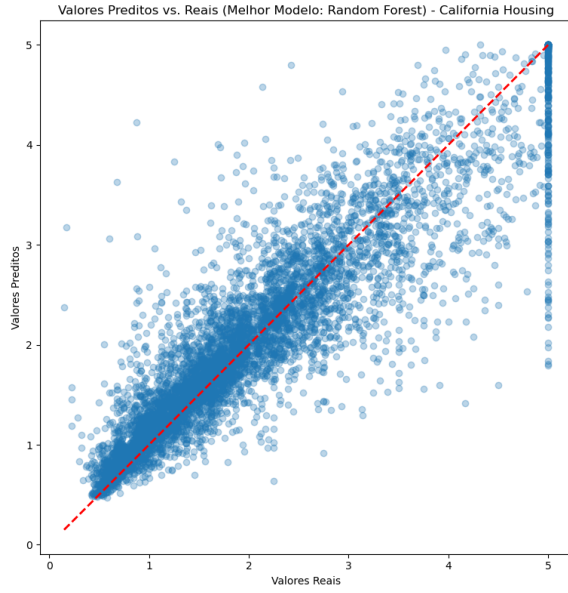


Fig. 5. Random Forest (melhor desempenho) para o dataset California Housing

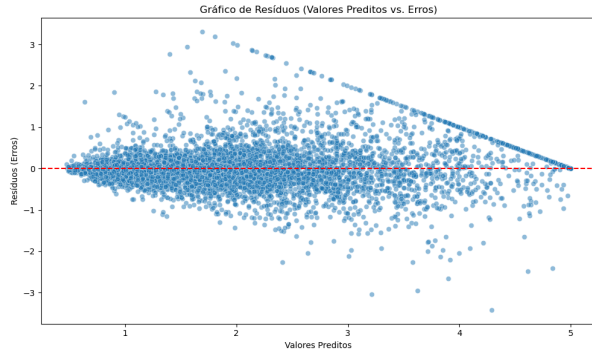


Fig. 6. Gráfico de resíduos para o modelo Random Forest no dataset California Housing

### B. Discussão dos Resultados de Regressão

Os resultados apresentados nas Tabelas I (RMSE) e II (MAE) mostram que o **Random Forest** foi o algoritmo com melhor desempenho em ambas as bases, apresentando os menores valores de RMSE e MAE, além de baixa variabilidade entre execuções. Isso indica não só maior precisão, mas também menor erro absoluto médio, o que é relevante para aplicações onde grandes desvios individuais são penalizados.

O **MLP Regressor** também se destacou, especialmente no California Housing, com valores competitivos de RMSE e MAE, embora com maior desvio padrão, indicando sensibilidade à inicialização dos pesos. Os modelos lineares (Regressão Linear e Ridge) e o KNN apresentaram desempenho inferior em ambas as métricas, sugerindo que as relações não-lineares presentes nos dados são melhor capturadas por modelos mais complexos. O SVR teve desempenho intermediário, com resultados melhores no California Housing em relação ao Energy Efficiency.

A análise conjunta de RMSE e MAE é importante, pois, enquanto o RMSE penaliza mais fortemente grandes erros, o MAE fornece uma visão mais robusta ao impacto de outliers. A consistência dos resultados do Random Forest em ambas as métricas reforça sua robustez para os conjuntos avaliados.

### C. Experimento 2: Classificação Multiclasse

Foram avaliados cinco algoritmos: Random Forest, SVM, MLP Classifier, KNN Classifier e Fuzzy-like KNN. Os experimentos foram realizados tanto no Wine quanto no Iris. A principal métrica utilizada foi a acurácia, complementada por F1-score e  $R^2$ . Os resultados médios e desvios padrão são apresentados nas Tabelas III e IV.

TABLE III  
RESULTADOS DE ACURÁCIA (MÉDIA  $\pm$  DESVIO PADRÃO, QUANDO APLICÁVEL) PARA CADA ALGORITMO DE CLASSIFICAÇÃO

Algoritmo	Wine Quality (Red)	Iris
Random Forest	0.6866 $\pm$ 0.0114	0.6741 $\pm$ 0.0115
SVM	0.5725 $\pm$ 0.0097	0.5349 $\pm$ 0.0103
MLP Classifier	0.5966 $\pm$ 0.0142	0.5904 $\pm$ 0.0138
KNN Classifier	0.5600 $\pm$ 0.0124	0.5485 $\pm$ 0.0125
Fuzzy-like KNN	0.6633 $\pm$ 0.0121	0.6549 $\pm$ 0.0121

TABLE IV  
RESULTADOS DE F1-SCORE (MÉDIA  $\pm$  DESVIO PADRÃO, QUANDO APLICÁVEL) PARA CADA ALGORITMO DE CLASSIFICAÇÃO

Algoritmo	Wine Quality (Red)	Iris
Random Forest	0.6741 $\pm$ 0.0115	0.9474 $\pm$ 0.0375
SVM	0.5349 $\pm$ 0.0103	0.9666 $\pm$ 0.0356
MLP Classifier	0.5904 $\pm$ 0.0138	0.9520 $\pm$ 0.0383
KNN Classifier	0.5485 $\pm$ 0.0125	0.9509 $\pm$ 0.0333
Fuzzy-like KNN	0.6549 $\pm$ 0.0121	0.9631 $\pm$ 0.0351

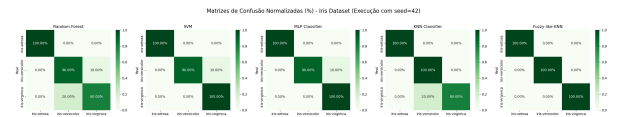


Fig. 7. Matriz de Confusão para todos os modelos no dataset Iris

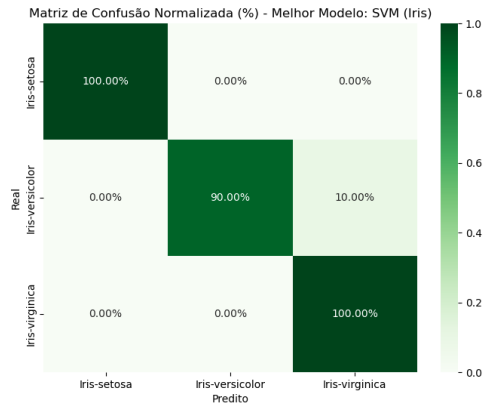


Fig. 8. Matriz de Confusão para o melhor modelo (SVM) no dataset Iris

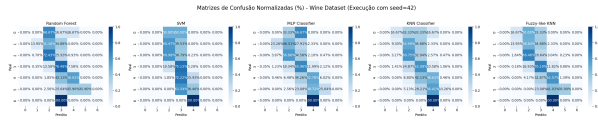


Fig. 9. Matriz de Confusão para todos os modelos no dataset Wine Quality (red)

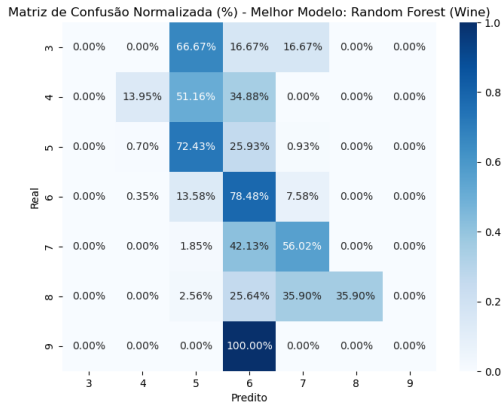


Fig. 10. Matriz de Confusão para o melhor modelo (Random Forest) no dataset Wine Quality (red)

#### D. Discussão dos Resultados de Multiclasse

A análise dos resultados de classificação multiclasse, apresentados nas Tabelas III (Acurácia) e IV (F1-score), revela que o **Random Forest** foi o algoritmo com melhor desempenho geral. Ele obteve as maiores acurácias em ambos os conjuntos de dados (Wine Quality e Iris), com baixa variabilidade, indicando robustez.

O **Fuzzy-like KNN** também se destacou, apresentando resultados muito competitivos e próximos aos do Random Forest, especialmente no Wine Quality, o que sugere a eficácia de sua abordagem para problemas de classificação multiclasse.

Em contraste, o SVM, MLP Classifier e KNN Classifier apresentaram desempenhos inferiores em acurácia, especialmente no conjunto Wine Quality, onde as classes são mais complexas. No conjunto Iris, embora as acurácias fossem menores do que o esperado para um dataset clássico, o F1-score foi consistentemente alto para todos os algoritmos, indicando boa capacidade de identificação das classes, mesmo com acurácia global moderada.

A consistência do Random Forest e do Fuzzy-like KNN em ambas as métricas e conjuntos de dados reforça sua adequação para problemas de classificação multiclasse, superando os demais algoritmos avaliados.

#### IV. DISCUSSÃO GERAL

Os resultados apresentados evidenciam diferenças importantes no desempenho dos algoritmos avaliados, tanto nas tarefas de regressão quanto nas de classificação multiclasse, destacando a superioridade de modelos mais complexos em ambos os contextos.

Na **regressão**, conforme mostrado na Tabela I, o **Random Forest** foi o algoritmo com melhor desempenho, apresentando os menores valores médios de RMSE em ambas as bases (0.4894 no Energy Efficiency e 0.5050 no California Housing), além de baixa variabilidade entre execuções. Isso demonstra não apenas maior precisão, mas também maior consistência, o que é essencial em aplicações práticas. O **MLP Regressor** também apresentou resultados competitivos, especialmente no California Housing (RMSE de 0.5152), mas com maior desvio padrão, indicando sensibilidade à inicialização dos pesos e ao ajuste de hiperparâmetros. Em contrapartida, os modelos lineares (Regressão Linear e Ridge) e o KNN tiveram desempenho inferior, reforçando sua limitação em capturar relações não lineares complexas. O SVR apresentou desempenho intermediário, com melhores resultados no California Housing, sugerindo que sua eficácia depende da estrutura dos dados.

Na **classificação multiclasse**, os resultados das Tabelas III (Acurácia) e IV (F1-score) mostram que o **Random Forest** também foi o algoritmo mais robusto, alcançando as maiores acurácias em ambos os conjuntos de dados (Wine Quality e Iris), com baixa variabilidade. O **Fuzzy-like KNN** apresentou desempenho competitivo, especialmente no Wine Quality, onde obteve resultados próximos aos do Random Forest, sugerindo que sua abordagem fuzzy é eficaz em problemas com maior complexidade de classes. Em contraste, o SVM, MLP Classifier e KNN apresentaram desempenhos inferiores, especialmente no Wine Quality, onde as classes são mais sobrepostas. No conjunto Iris, embora as acurácias tenham sido moderadas, o F1-score foi consistentemente alto para todos os algoritmos, indicando boa capacidade de identificação das classes, mesmo com acurácia global mais baixa.

De forma geral, os resultados reforçam a superioridade de modelos mais complexos, como o **Random Forest**, tanto em tarefas de regressão quanto de classificação multiclasse. A análise conjunta das métricas utilizadas (RMSE, acurácia e F1-score) destaca a importância de considerar não apenas a precisão, mas também a consistência e a capacidade de

generalização dos algoritmos. Além disso, o desempenho competitivo do **Fuzzy-like KNN** em classificação multiclasse sugere que abordagens alternativas, como métodos fuzzy, podem ser úteis em cenários com maior complexidade de classes. Esses achados são relevantes para a escolha de algoritmos em aplicações práticas, onde a robustez e a confiabilidade são fatores críticos.

## V. CONCLUSÃO

Este trabalho apresentou uma análise comparativa de algoritmos supervisionados em tarefas de regressão e classificação multiclasse, utilizando diferentes bases de dados e técnicas de pré-processamento. Os resultados reforçam a importância da escolha adequada do modelo e do pré-processamento, bem como da validação estatística robusta. Como trabalhos futuros, sugere-se a avaliação de técnicas de seleção de atributos e ajuste fino de hiperparâmetros.

## AGRADECIMENTOS

Agradecemos ao CEFET-MG pelo suporte institucional.

## REFERENCES

- [1] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.
- [2] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *JMLR*, vol. 12, pp. 2825–2830, 2011.
- [3] A. Tsanas and A. Xifara, “Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools,” *Energy and Buildings*, vol. 49, pp. 560–567, 2012.
- [4] R. K. Pace and R. Barry, “Sparse spatial autoregressions,” *Statistics & Probability Letters*, vol. 33, no. 3, pp. 291–297, 1997.
- [5] P. Cortez et al., “Modeling wine preferences by data mining from physicochemical properties,” *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.
- [6] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- [7] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.
- [8] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [9] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
- [10] L. I. Kuncheva, “Fuzzy classifier design,” Physica-Verlag HD, 1999.
- [11] A. C. Rencher and G. B. Schaalje, *Linear Models in Statistics*, 2nd ed. New York: John Wiley & Sons, 2008.
- [12] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
- [13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, pp. 533–536, 1986.
- [14] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [15] S. A. Dudani, “The distance-weighted k-nearest-neighbor rule,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 4, pp. 325–327, 1976.