Empirical Comparison of Classification and Regression Algorithms

Ramon Calderon McDargh-Mitchell

Abstract

This paper presents an empirical comparison of machine learning algorithms—gradient-boosted trees (XGBoost), Random Forests, and Neural Networks—following the methodology of Caruana and Niculescu-Mizil. Three classification models were evaluated on three datasets, and three regression models were evaluated on two datasets, using 20/80, 50/50, and 80/20 train–test splits averaged over three independent trials. For classification, XGBoost achieved the strongest performance on two of the three datasets, while Random Forest performed best on the remaining dataset. Performance was evaluated using F1 score as the primary metric with ROC–AUC as the secondary metric. For regression tasks, XGBoost achieved the lowest error on one dataset, whereas Random Forest performed best on the other. Regression performance was primarily evaluated using test RMSE, with R^2 as the secondary metric.

1 Introduction

Richard Caruana and Alexandru Niculescu-Mizil, in their paper *An Empirical Comparison of Supervised Learning Algorithms*, conducted a thorough investigation into the performance of machine learning methods across a wide range of models and evaluation metrics. However, an updated empirical comparison of the best-performing predictive models in the modern AI era would be valuable, given the substantial advances in computational power and the availability of modern libraries that enable extensive hyperparameter tuning during training.

2 Methodology

2.1 Datasets

We evaluate our models on five publicly available datasets from the UCI Machine Learning Repository: the Bank Marketing dataset, the Differentiated Thyroid Cancer Recurrence dataset, and the Wine dataset. These datasets were used for classification tasks, whereas the Infrared Thermography Temperature dataset and the Parkinson's Telemonitoring dataset were used for regression tasks. All five datasets were tabular.

2.2 Experimental Pipeline

All datasets were cleaned and preprocessed prior to model training. For each dataset, we evaluated three train–test splits (20/80, 50/50, and 80/20), and repeated each split for three independent trials using different random seeds to reduce variance due to sampling.

Within each trial, model hyperparameters were selected using 5-fold cross-validation on the training portion only. Stratified 5-fold cross-validation was used for classification tasks to preserve class proportions, while standard 5-fold cross-validation was used for regression tasks. Hyperparameter selection was performed via grid search, and the best configuration was refit on the full training set before final evaluation on the held-out test set.

We compared three model families: gradient-boosted trees (XGBoost), Random Forests, and Neural Networks (multilayer perceptrons). For regression experiments, models were tuned using negative mean squared error during cross-validation, and final performance was summarized using test RMSE, with MAE and R^2 reported as additional context where appropriate. For classification experiments, models were tuned using F1-based metrics to emphasize performance across classes, and final performance was summarized using test F1 and ROC–AUC when applicable.

3 Experiments

3.1 Classification

Table 1: Classification Performance (Test F1 / ROC–AUC, 80/20 split averaged across 3 trials)

Algorithm	Bank Marketing	Thyroid Cancer Recurrence	Wine
XGBoost	0.516 / 0.803	0.929 / 0.992	1.000 / 1.000
Random Forest	0.507 / 0.797	0.935 / 0.987	0.991 / 1.000
Neural Network	0.351 / 0.792	0.901 / 0.979	0.972 / 1.000

3.2 Regression

Table 2: Regression Performance (Test RMSE / R^2, 80/20 split averaged across 3 trials)

Algorithm	Infrared Thermography	Parkinson's Telemonitoring
XGBoost	0.232 / 0.736	1.430 / 0.981
Random Forest	0.231 / 0.735	5.029 / 0.768
Neural Network	0.300 / 0.546	3.780 / 0.869

4 Discussion

Overall, the experiments show that model performance depends on both dataset characteristics and task type. While tree-based models were generally stable and effective, their behavior varied across datasets, making it difficult to identify a single best-performing algorithm across all scenarios.

Neural networks exhibited consistent overfitting across both classification and regression tasks. This aligns with prior findings that standard multilayer perceptrons often struggle on tabular datasets with limited sample sizes. Future work could explore smaller network architectures, stronger regularization, or alternative neural approaches better suited to tabular data.

Random Forests also showed overfitting on the Parkinson's Telemonitoring regression task under the selected hyperparameter grid, leading to reduced test performance despite strong training results. More targeted hyperparameter tuning or alternative regularization strategies could likely improve generalization for this dataset.

Support Vector Machines (SVMs) were implemented as part of the experimental pipeline but were not included in the final results. Due to the large hyperparameter search space required to make SVMs competitive—especially on larger datasets—and time constraints, they were excluded to maintain consistency with the experimental methodology.

The full experimental pipeline, code, and additional visualizations are available at:
`https://github.com/RamonsArchive/empirical-ml-comparison`

5 Conclusion

This study presented an empirical comparison of three widely used machine learning model families—gradient-boosted trees, Random Forests, and neural networks—across multiple classification and regression datasets. Following the experimental framework of Caruana and Niculescu-Mizil, extensive cross-validation and multiple train–test splits were used to evaluate model performance under varying data availability.

For classification tasks, tree-based models consistently outperformed neural networks across all datasets. XGBoost achieved the strongest performance on the Bank Marketing and Wine datasets, while Random Forest achieved the highest F1 score on the Thyroid Cancer Recurrence dataset. These results suggest that while boosting methods often provide strong performance, ensemble methods such as Random Forests can be competitive or superior depending on dataset characteristics, particularly in structured tabular settings.

For regression tasks, no single model dominated across datasets. XGBoost achieved the lowest test RMSE on the Parkinson's Telemonitoring dataset, whereas Random Forest performed best on the Infrared Thermography dataset. Neural networks generally underperformed relative to tree-based methods, likely due to sensitivity to hyperparameter choices and overfitting in smaller tabular datasets.

Overall, these findings reinforce conclusions from prior empirical studies: ensemble tree-based methods remain highly effective for tabular data, and model performance depends strongly on dataset properties and training data size. Future work could explore more extensive hyperparameter tuning for neural networks, alternative regularization strategies, and additional datasets to further characterize model behavior across domains.

6 References

References

[1] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 161–168, 2006.

[2] D. Dua and C. Graff. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml`, 2019.