A Comparative Study of Ensemble Learning and Neural Networks on the Abalone and Contraceptive Method Choice Datasets

Xudong Yang
z5713177
z5713177@ad.unsw.edu.au

Yuzheng Liu
z5598864
z5598864@ad.unsw.edu.au

Abstract—This project takes a closer look at how several common machine-learning models behave on two different classification problems. We compare decision trees, random forests, gradient boosting, XGBoost, and a small neural network. The first dataset (Abalone) tries to infer age from physical measurements.[1] The second dataset (Contraceptive Method Choice) looks at demographic factors related to which contraceptive method a woman chooses.[2] We try a bit of everything along the way: tuning hyperparameters, pruning overly complex decision trees, using PCA to shrink the feature space, and testing different optimization strategies for neural networks—mainly Adam, SGD, dropout, and regularization. [9][10] Across both datasets, XGBoost ends up performing the best, clearly ahead of standard trees and the neural networks. We also find that pruning helps trees behave much better on unseen data, while the neural networks react strongly to the exact optimization setup and how much regularization we apply. Overall, the project offers a complete, reproducible workflow that compares several ensemble models on tabular data, highlights how their generalization differs, and provides all code for anyone who wants to try the experiments themselves.

Keywords—ensemble learning, decision trees, pruning, random forests, gradient boosting, XGBoost, neural networks, optimization algorithms, PCA, tabular data classification, abalone dataset, contraceptive method choice

INTRODUCTION

Machine learning models—whether simple decision trees, ensemble learning models, or neural networks—have become fundamental tools for processing structured data. They excel in tasks such as classification and prediction because they capture details that are difficult for humans to notice.

In recent years, methods such as random forests and ensemble learning (e.g., Boosting) have become particularly popular, combining multiple weak learners to build more robust and stable ensemble models. [6] Neural networks, supported by optimizers such as Adam and SGD, continue to be widely used in a wide range of fields, from computer vision to more traditional tabular datasets.

However, neural networks typically require large amounts of data and require manual adjustment of the learning rate and appropriate regularization—otherwise they can overfit.

On the other hand, tree-based models, when faced with the possibility of multiple splits or the construction of large ensemble models, may result in excessively many layers in the decision tree if no intervention pruning is performed, splitting every small detail of the training model and ultimately leading to severe overfitting.

Meanwhile, real-world datasets also introduce complexity: for example, chaotic categorical variables and imbalanced classes, so special preprocessing steps are required to ensure that the model learns the feature distribution correctly.

Although ensemble learning methods are widely used, there is currently a lack of sufficient comparative studies to examine their performance across different hyperparameters and datasets.

This project aims to compare various machine learning models and determine which models perform best on tabular mixed-type datasets.

So, this project combines ensemble learning techniques, classic decision trees, and neural networks to better understand their performance on two representative tabular datasets. The abalone dataset involves predicting biological age groups based on continuous measurements, while the contraceptive method selection dataset focuses on sociodemographic characteristics, most of which are categorical variables. These contrasting characteristics allow us to evaluate the generalization ability of models across different scenarios.

METHODOLOGY

The Abalone dataset contains 4177 samples with 8 continuous input features, and the target variable “Rings” is converted into 4 age classes (0–7, 8–10, 11–15, >15 years).
The Contraceptive Method Choice dataset contains 1473 samples and 9 categorical features. The target variable represents one of three contraceptive choices.

Preprocessing of our project:

1. Continuous features standardized using StandardScaler.

2. Categorical features encoded using OneHotEncoder.

3. PCA applied to Abalone dataset to obtain 2D scatter plots and to retain 95% and 98% variance in reduced-dimension experiments.

4. Stratified train/test split used for all experiments.

Our Models Overview:

1. Decision Trees with multiple depths. [3]

2. Cost-complexity pruning to reduce overfitting.[4]

3. Random Forests with varying number of estimators. [5]

4. Gradient Boosting and XGBoost for ensemble learning.

5. Simple neural networks (MLP) using ReLU activation.

6. Neural network experiments using SGD vs Adam and dropout.

All experiments were implemented using Python, scikit-learn, XGBoost, PyTorch, and matplotlib/seaborn for visualisation. Code is fully reproducible.

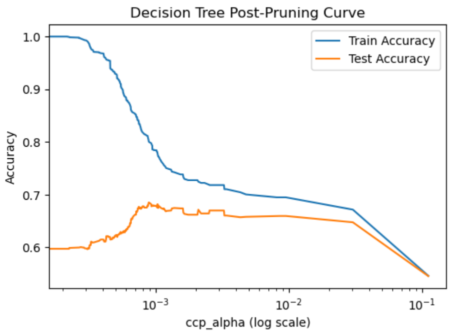
RESULTS

This section compares all models on the Abalone age-class prediction task. Overall, XGBoost delivers the best performance, followed by pruned decision trees and random forests. Neural network results vary depending on the optimiser and regularisation strategy.

3.1 Decision Trees

Across 25 runs with different random seeds and maximum depths, the best performing tree reaches a test accuracy of 0.715 at max_depth = 5.

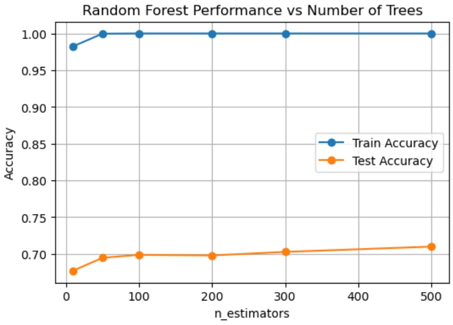
Introducing cost-complexity pruning improves generalisation: the test accuracy peaks at ccp_alpha ≈ 8.7×10^{-4}, after which accuracy drops as the model becomes too shallow.



3.2 Random Forests

Increasing the number of estimators steadily improves test performance.

With 500 trees, the random forest achieves a test accuracy of ≈ 0.71 while maintaining perfect training accuracy due to ensembling.



3.3 Gradient Boosting and XGBoost

Gradient Boosting reaches a test accuracy of 0.704. [8]

XGBoost performs best among all models on the Abalone dataset, achieving a test accuracy of 0.707, with stable training behaviour and minimal overfitting.

3.4 PCA 95% and 98% Experiments

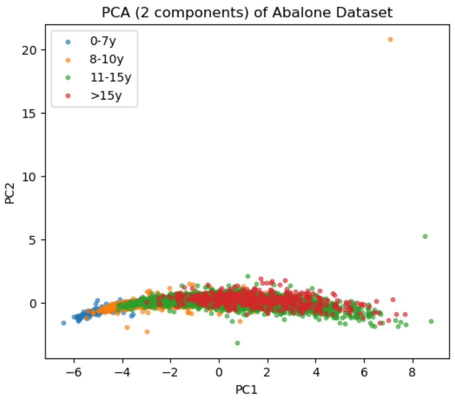
To study dimensionality reduction, PCA was applied to retain 95% (3 components) and 98% (4 components) of the variance.

Using XGBoost on the reduced features yields:

1. 95% variance (3 components): 0.647 accuracy

2. 98% variance (4 components): 0.676 accuracy

Although accuracy decreases relative to full-feature XGBoost (0.707), both PCA variants preserve reasonable predictive performance.



3.5 Neural Network Experiments

The neural network results depend strongly on the optimiser and regularisation configuration.

1. SGD achieves a test accuracy of 0.637.

2. Adam performs substantially better, reaching 0.712.

Dropout experiments show that values between 0 and 0.3 give similar results (test accuracy ≈ 0.70), while strong dropout (0.5) lowers performance, likely due to underfitting on this small dataset.

3.6 Part B: CMC Dataset

The CMC dataset is more challenging due to sparse one-hot encodings and overlapping class boundaries.

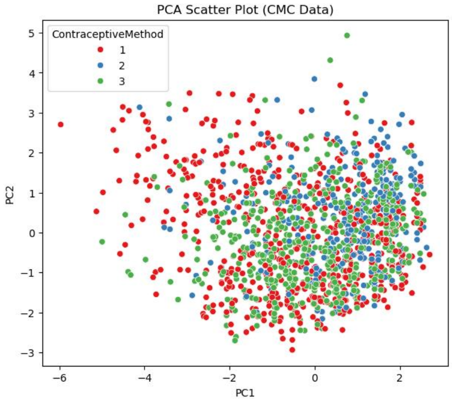
XGBoost:

Accuracy: 0.545, F1-macro: 0.512

MLP (Adam):

Accuracy: 0.475, F1-macro: 0.453

XGBoost outperforms the neural network model, largely because boosted trees handle high-dimensional sparse inputs more effectively than shallow MLPs.



DISCUSSION

Our results show that ensemble learning models consistently outperform neural networks on tabular datasets. XGBoost achieved the best accuracy on both datasets due to its ability to handle nonlinear interactions and sparse feature spaces.

Decision-tree pruning effectively reduced overfitting and improved generalization. Neural networks performed reasonably well on the Abalone dataset but struggled on the CMC dataset, which is small and dominated by categorical variables. Dropout decreased performance, indicating the network capacity was already low.

The PCA experiments reveal that ensemble methods are robust even when dimensionality is reduced, although performance inevitably declines.

CONCLUSION

This project demonstrates that:

1. XGBoost performed well on both datasets because tree models are suitable for handling structured table features, can capture non-linear

relationships, and gradient boosting also brings strong generalization and regularization capabilities. Therefore, on small-scale datasets, XGBoost is more stable than shallow neural networks.

2. Decision tree pruning significantly improves performance compared to the unpruned model, as shown in the experimental results.

3. Neural networks generally perform better with the Adam optimizer. Adam's adaptive learning rate converges quickly in the early training phase, thus finding better solutions more stably on small datasets.

However, network performance is still significantly affected by dropout and regularization strength: moderate regularization can suppress overfitting, while excessive dropout weakens model capacity, resulting in a significant performance drop in the low-dimensional, small-sample scenario of this project. [9][11]

4. For small tabular datasets, ensemble models are generally more reliable than lightweight neural networks. Tree models are better at capturing key nonlinear relationships in structured features, while neural networks often require more data for adequate training. [6][13][14]

5. PCA helps reduce dimensionality but may be counterproductive if the dataset is highly nonlinear. [7]

Reference

[1] W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford. "Abalone," UCI Machine Learning Repository, 1994. [Online]. Available: https://doi.org/10.24432/C55C7W.

[2] T. Lim. "Contraceptive Method Choice," UCI Machine Learning Repository, 1999. Available: https://doi.org/10.24432/C59W2D.

[3] R. Rivera-Lopez, J. Canul-Reich, E. Mezura-Montes, and M. A. Cruz-Chávez,

"Induction of decision trees as classification models through metaheuristics,"

Swarm and Evolutionary Computation, vol. 69, Article 101006, p. 2, 2022, doi: 10.1016/j.swevo.2021.101006.

[4] J. Huyghe, J. Trufin, and M. Denuit, "Boosting cost-complexity pruned trees on Tweedie responses: the ABT machine for insurance ratemaking,"

Scandinavian Actuarial Journal, vol. 2024, no. 5, pp. 417–439, 2024.

[5] L. Breiman, "Random Forests," *Mach. Learning,* vol. 45, *(1),* pp. 5-32, 2001. Available: https://wwwproxy1.library.unsw.edu.au/login?url=https://www.proquest.com/scholarly-journals/random-forests/docview/757027982/se-2. DOI: https://doi.org/10.1023/A:1010933404324.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. KDD, 2016, pp. 785–794.

[7] I. Jolliffe and I. Cadima, "Principal component analysis: A review and recent developments," Philosophical Transactions of the Royal Society A, vol. 374, no. 2065, pp. 1–16, 2016.

[8] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015, pp. 1–15.

[10] H. Robbins and S. Monro, "A stochastic approximation method," Annals of Mathematical Statistics, vol. 22, no. 3, pp. 400–407, 1951.