COGS 118A Final Project

Anson Wen
anwen@ucsd.edu

Abstract

This project evaluates the performance of three supervised learning classifiers—Support Vector Machine (SVM), Random Forest, and Decision Tree—on three datasets from the UCI repository: Car Evaluation, Wine, and Heart Disease. Each classifier was assessed under three train-test partitions (80/20, 50/50, 20/80) with hyperparameter tuning and cross-validation. Results confirm that Random Forest consistently outperformed other classifiers, while Decision Tree and SVM showed comparable performances. This study provides insights into classifier behavior across diverse datasets and data distributions.

Introduction

Machine learning (ML) techniques offer the potential to leverage existing data to develop predictive models. Among ML classifiers, Support Vector Machines, Random Forests, and Decision Trees are widely studied for their robustness and applicability across various domains. This report replicates a portion of Caruana and Niculescu-Mizil's work by comparing three supervised classifiers: SVM, Random Forest, and Decision Tree. The primary objective is to understand the impact of hyperparameter tuning, data partitioning, and dataset characteristics on classifier performance. Using three binary classification datasets, the report aims to identify trends and provide recommendations for classifier selection based on accuracy.

Method

Classifiers

The three classifiers chosen for this study are:

Support Vector Machine (SVM): Implemented with linear and RBF kernels, tuned over C values.

Random Forest: Ensemble-based model with tuning for the number of estimators and maximum tree depth.

Decision Tree: Single-tree classifier tuned for depth and splitting criteria.

Datasets

I used three datasets that were sourced from the UCI repository, which include the Car Evaluation Dataset, Wine Dataset, and Heart Disease Dataset. Car Evaluation Dataset contained 1,728 samples with 6 features. For the Car Evaluation Dataset, the categorical variables were encoded using label encoding, which converted them into numeric formats that would be suitable for machine learning algorithms. This meant that the target variable was mapped to numerical classes representing the evaluation categories. The Wine Dataset contained 178 samples with 14 features. The target variable for this dataset represents wine classes (1,2,3), which were retained without further transformations. The Heart Disease Dataset contained 920 samples and 14 features. This dataset was built from multiple data sources, so missing values would be addressed by deleting incomplete rows to ensure data integrity. The target variable was binary for two classes that represented presence or absence of heart disease (1,0).

Experiment

Experimental Design

The first step of this study involved data partitioning, where three train-test splits: 80/20, 50/50, and 20/80 will be used. Next we used cross-validation on training sets in order to identify optimal hyperparameters via grid search. Lastly, performance metrics provided a quantitative foundation for assessing and comparing the classifiers, which would give us conclusions about their effectiveness and how well they represent the trends from the literature. Models were evaluated using cross-validation accuracy, test accuracy, and weighted average accuracy, which combined train and test results proportionally.

Results

Table 1: Car Evaluation Dataset: Average Weighted Accuracy for each learning algorithm by partition

Partition	SVM Accuracy	Random Forest Accuracy	Decision Tree Accuracy
80/20	0.723	0.969	0.974
50/50	0.727	0.959	0.964
20/80	0.721	0.905	0.937

Table 2: Wine Dataset: Average Weighted Accuracy for each learning algorithm by partition

Partition	SVM Accuracy	Random Forest Accuracy	Decision Tree Accuracy
80/20	0.944	0.944	0.903
50/50	0.975	0.989	0.910
20/80	1.00	1.00	0.918

Table 3: Heart Disease Dataset: Average Weighted Accuracy for each learning algorithm by partition

Partition	SVM Accuracy	Random Forest Accuracy	Decision Tree Accuracy
80/20	0.88	0.87	0.80
50/50	0.81	0.83	0.71
20/80	0.80	0.80	0.64

Observations

Car Evaluation Dataset

- Decision Tree achieves the highest accuracy (0.974) in the 80/20 split, closely followed by Random Forest (0.969)
- SVM relatively lower accuracy
- Larger training sizes (80/20) improve performance for all models

Wine Dataset

- SVM and Random Forest perform great, achieving top accuracy (1.00) in the 20/80 split
- Decision Tree still has high accuracy, but not as high comparatively in the 50/50 and 80/20 split
- SVM's consistent performance in all partitions highlight its suitability for this dataset

Heart Disease Dataset

- SVM achieves the highest accuracy (0.88) in the 80/20 split, outperforming the others across all partitions
- Random Forest remains reliable with stable accuracy (~0.80-0.87)

- Decision Tree struggles, especially in the 20/80 split (0.64), suggests overfitting to smaller training data and fail to generalize effectively

Overall, Random Forest Classifier performs well consistently across datasets, highlighting its robustness and ability to generalize effectively. On the otherhand, SVM showed strong results for the Wine and Heart Disease Datasets but not for the Car Evaluation Dataset. Lastly, the Decision Tree Dataset performed well in simpler datasets but struggled in complex ones like the Heart Disease Dataset because of overfitting.

Conclusion

This study closely reflects the findings presented in Caruana and Niculescu-Mizil's *An Empirical Comparison of Supervised Learning Algorithms*. While there are some minor differences in accuracy, the results clearly show that the Random Forest Classifier consistently outperforms both the Support Vector Machines and Decision Trees Classifiers, reaffirming its reliability and strong overall performance.

References

1. Caruana, Rich, and Niculescu-Mizil, Alexandru. *An Empirical Comparison of Supervised Learning Algorithms.* Department of Computer Science, Cornell University, Ithaca.
2. Dua, D., and Graff, C. (2019). *UCI Machine Learning Repository.* Irvine, CA: University of California, School of Information and Computer Science. Retrieved from http://archive.ics.uci.edu/ml.