# Cogs 118A Final Project

**Timothy Lue** thlue@ucsd.edu

Cognitive Science 118A, University of California San Diego

## Abstract

Currently there are numerous online resources for supervised learning methods that are easily accessible to the general public for learning and use. Even beginners with no experience at all are able to pick up the skills necessary to use them. As such there are articles and papers that provide a compare and contrast analysis of the many supervised learning methods. In this paper, I present a small-scale replication of one of the many comparison papers, *An Empirical Comparison of Supervised Learning Algorithms* by Rich Caruana and Alexandru Niculescu-Mizil. Looking mainly at the supervised learning methods Support Vector Machine, Random Forests, and Decision Trees. The main point of comparison in my study is the performance metric of threshold, accuracy.

## Introduction

With the emergence of Big Data and the prevalence of data based technology, supervised learning methods are now used in many domains of study, research, and industry. We can now take data from thousands or even millions of people and apply supervised learning methods to solve and figure out answers to problems that we never would have been able to solve. Such examples of the use of supervised learning methods can be seen in big tech industry companies like Facebook, Uber, Google and more who use data driven analysis and prediction to better tailor their services to their millions of users.

Today anyone can go and learn the skills to utilize supervised learning methods from the many resources that are provided online. With the many different and unique supervised learning methods, there have been many compare and contrast studies on their performance in different situations with differing datasets. This paper is a small-scale replication of one those many supervised learning method comparison studies. More specifically, this is a replication of *An Empirical Comparison of Supervised Learning Algorithms* written by Rich Caruana and Alexandru Niculescu-Mizil, focusing mainly on the supervised learning methods Support Vector Machine, Random Forests, and Decision Trees with the main point of comparison being the performance metric of threshold, accuracy.

## Methods

**Learning Algorithms:**

SVM: (Support Vector Machine) I used the Support Vector Classifier with a linear kernel that underwent Grid Search Cross Validation in order to determine the best or most accurate C value from the list [0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1]. Which was later used to calculate the weighted average accuracy of the two-class classification on the data.

Random Forest: I used the Random Forest Classifier with 10 estimators and an entropy criterion that underwent Grid Search Cross Validation in order to determine the best or most accurate maximum depth value from the list [1, 2, 3, 4, 5]. Which was later used to calculate the weighted average accuracy of the two-class classification on the data.

Decision Trees: I used the Decision Tree Classifier that, similarly to Random Forest, used an entropy criterion that underwent Grid Search Cross Validation in order to determine the best or

most accurate maximum depth value from the list [1, 2, 3, 4, 5]. Which was later used to calculate the weighted average accuracy of the two-class classification on the data.

**Performance Metrics:**

Unlike *An Empirical Comparison of Supervised Learning Algorithms* by Rich Caruana and Alexandru Niculescu-Mizil which this paper is trying to replicate, the performance metric this study focuses on is only the threshold metric, accuracy. More specifically the accuracy I am looking at in this study consists of the precision value of the weighted average accuracy between the two class classification categories that have been defined based on the data set. Using that accuracy of each supervised learning algorithm, I will be able to compare and contrast their accuracies to determine their strengths and weaknesses.

**Data Sets:**

The data sets used in this study consist of three data sets: the Wine Quality: Red Wine Data Set, the Wine Quality: White Wine Data Set, and the Adult Data Set all from the UCI Machine learning repository.

The Wine Quality: Red Wine Data Set and the Wine Quality: White Wine Data Set consists of data based on red and white variants of the Portuguese "Vinho Verde" wine respectively. In the data set there are input variables that were based on physicochemical tests such as: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol with the output variable that was based on sensory data being quality that was scored between 0 and 10. Furthermore, I divided each wine into one of two sub categories based on the median quality score where wines that were above the median were labeled “good”, set to 1, and wines that were below the median were labeled “bad”, set to 0.

This created a two class classification problem between “good” wines with higher quality and “bad” wines with lower quality

The Adult data set consists of data based on an extraction of records from a 1994 Census database. In the data set there are many variable based on the many attributed of the individual such as: if they made greater than 50K or less than/equal to 50K, age, work class, education level, education grade, marital status, occupation, relationship status, race, sex, income or capital gain, income or capital loss, work hours per week, native country. Many of the variables like work class, occupation, race, or native country consisted of many different categories so I converted and divided them into numeric categories and equivalents. In order for the learning algorithms to be fully able to analyze the code properly, I also converted the other dataset features that were not numeric to numeric equivalents.

## Experiment

**Process:**

In this study, I chose to compare the threshold metric accuracy of the 3 classifiers: Support Vector Machine, Decision Tree, and Random Forest for the 3 datasets Wine Quality: Red Wine Data Set, Wine Quality: White Wine Data Set, and the Adult Data Set. For each classifier and data set, I went through 3 different trials where I randomized the order of the dataset, alongside using 3 different partitions of 20/80, 50/50, and 80/20 to divide the dataset into training and testing sets. In addition, Grid Search Cross Validation was used each time in order to select the best hyper-parameter for that classifier. In total I looked at 3 trials for 3 classifiers for 3 datasets for 3 partitions which resulted in 81 accuracy metrics.

The coding process was speed up and aided by referencing similar code found on Kaggle written by users Vishal Kumar and Overfitting.

**Results:**

Table 1: Average Weighted Accuracy for each learning algorithm by trial

| | Trial 1 | Trial 2 | Trial 3 |
|---|---|---|---|
| SVM | 0.668 | 0.695 | 0.683 |
| Decision Trees | 0.668 | 0.695 | 0.683 |
| Random Forest | 0.795 | 0.825 | 0.82 |

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

| | Partition 80/20 | Partition 50/50 | Partition 20/80 |
|---|---|---|---|
| SVM | 0.685 | 0.68 | 0.682 |
| Decision Trees | 0.685 | 0.68 | 0.682 |
| Random Forest | 0.828 | 0.815 | 0.797 |

Overall, it appears that the Random Forest classifier did significantly better than the other 2 classifiers, averaging around greater than 10 percent higher than Support Vector Machine and Decision Trees. While both Support Vector Machine and Decision trees tend to have around the same accuracy percentage. This result is consistent with the findings in the paper *An Empirical Comparison of Supervised Learning Algorithms* by Rich Caruana and Alexandru Niculescu-Mizil in which this study is trying to replicate. In Caruana and Niculescu-Mizil's paper, the normal Random Forest learning algorithm also had around a greater than 10 percent higher

accuracy in comparison to normal Support Vector Machine and normal Decision Trees. However, in their paper, normal Support Vector Machine had a greater percent accuracy than normal Decision Trees which this study did not discover. This may have been the case because this study is a small-scale replication of the Caruana and Niculescu-Mizil study where I only used 3 smaller datasets in comparison to their 11 larger datasets. Thus, under larger, more diverse, and well balanced datasets, the accuracies of the classifiers would be more spread out and representative of their individual ability.

## Conclusion

In conclusion, my attempts at a small-scale replication of the paper *An Empirical Comparison of Supervised Learning Algorithms* by Rich Caruana and Alexandru Niculescu-Mizil can be deemed as a success. While there may have been a slight difference with my paper and Caruana & Niculescu-Mizil's paper when it comes to the accuracy differences between the supervised learning methods Support Vector Machine and Decision Trees, the overall findings are consistent where the accuracy of the Random Forest algorithm eclipses both the Support Vector Machine and Decision Trees algorithms by more than 10 percent.

## References

Caruana, Rich., & Niculescu-Mizil , Alexandru. *An Empirical Comparison of Supervised Learning Algorithms*. Department of Computer Science, Cornell University, Ithaca.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Kumar, Vishal. *Prediction of quality of Wine*. Kaggle.

https://www.kaggle.com/vishalyo990/prediction-of-quality-of-wine

Overfitting. *Income prediction on UCI adult dataset*. Kaggle.

https://www.kaggle.com/overload10/adult-census-dataset/downloads/adult.csv/notebook

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.