# Application of Supervised Learning Algorithms

**Mark Merten** mmerten@ucsd.edu
Prepared for COGS 118A: Supervised Machine Learning Algorithms, Professor Zhuowen Tu, Fall 2023

## Abstract

This study conducts a comparative analysis of three machine learning classifiers on three datasets from the UCI machine learning repository, practicing some of the methodologies outlined in Caruana and Niculescu-Mizil's empirical study. The classifiers evaluated include a Decision Tree, K-Nearest Neighbors (KNN), and a Support Vector Machine (SVM) with a linear kernel. The performance of these classifiers is evaluated based on classification accuracy across three different dataset partitions (20/80, 50/50, 80/20) using cross-validation for hyper-parameter optimization. The study aims to validate the general trends observed by Caruana and Niculescu-Mizil using the classifiers.

## 1. Introduction

The study of machine learning offers several classification algorithms, each with its unique strengths and suitability for different types of data. Reviewing the study by Caruana and Niculescu-Mizil provides an extensive comparison of several classifier algorithms, offering insights into their relative performance across various metrics. This project replicates a part of their study by focusing on various classifiers and three UCI datasets, specifically for two-class classification problems. This project aims to observe the consistencies of results with those reported in the original study and acknowledge potential variations due to implementation differences and dataset characteristics.

The three datasets used from the UCI machine learning repository include the Auto MPG dataset (Quinlan, 1993), the Car Evaluation dataset (Bohanec, 1997), and the Liver Disorders dataset (Forsyth 1990). The datasets were tested with the implementation of supervised learning algorithms using Python programming language (https://www.python.org) and Jupyter notebook (https://jupyter.org). Python libraries such as Pandas (https://pandas.pydata.org) and Numpy (https://numpy.org) were also used to complete this project. The three datasets selected from the UCI repository represent diverse domains and complexities, ensuring a comprehensive evaluation of the classifiers. Each with its unique strengths and suitability for different types of data.

## 2. Methodology

### 2.1. Objective Approach

The primary objective of this study is to assess the performance of various machine learning classifiers on three distinct datasets: Auto MPG, Car Evaluations, and Liver Disorders. These datasets were chosen for their diversity in characteristics and the challenges they present in machine learning classification tasks. The variation in data partitioning was essential to assess the robustness of each classifier under different training and testing scenarios. A key aspect of the methodology used was cross-validation to determine the best hyper-parameters for each classifier. This approach ensured that the models were not overfitting and could be generalized. The final selection of hyper-parameters was based on their performance in cross-validation.

### 2.2. Auto MPG Dataset

The Auto MPG dataset is used to predict whether a car has high fuel efficiency based on its characteristics. It consists of several features including cylinders, displacement, horsepower, weight, acceleration, model year, and origin. The original dataset contains continuous and categorical data, with a total of 398 data points. The primary target feature, 'mpg' (miles per gallon), is continuous. To convert this into a binary classification problem, a new label called 'mpg_high' was created. This label categorizes a car as having high fuel efficiency ('mpg_high' = 1) if its mpg value is above the median of all cars in the dataset, and low fuel efficiency ('mpg_high' = 0) otherwise. The 'mpg' and 'car name' features were then dropped from the dataset. The 'horsepower' feature, which initially contained some missing values denoted as '?', was cleaned by replacing these placeholders with NaN and subsequently dropping rows with NaN values. The

cleaned 'horsepower' data was then converted to a numeric type.

The features were standardized using StandardScaler to have zero mean and unit variance. The performance of each classifier was evaluated based on the average training and testing accuracies across all trials and data splits. The best hyperparameters for each classifier were also recorded and reported. This approach provided a comprehensive evaluation of each classifier's ability to predict high fuel efficiency in automobiles based on a range of features. The use of different data splits and repeated trials helped in assessing the models' robustness and generalizability.

### 2.3. Car Evaluation Dataset

The Car Evaluation dataset is utilized to predict the acceptability of cars based on various categorical attributes. This dataset comprises features such as buying price, maintenance cost, number of doors, persons capacity, luggage boot size, and safety ratings. It contains 1,728 data points, each providing insight into different aspects of car evaluation.

The original 'class' feature in this dataset, which indicates the acceptability of a car, ranges across multiple classes. To simplify the problem into a binary classification task, this feature is transformed. The dataset undergoes preprocessing where all the categorical features are encoded using Label Encoding. This encoding process converts categorical values into a numerical format suitable for machine learning algorithms.
In this dataset, three supervised learning models are employed: Decision Tree (DT), Random Forests (RF), and K-Nearest Neighbors (KNN). Feature standardization is performed using StandardScaler to normalize the data, ensuring that all features contribute equally to the model's training process. The classifiers' performances are evaluated based on their average training and testing accuracies across all trials and test sizes.

This methodology offers a detailed evaluation of each classifier's effectiveness in predicting car acceptability based on categorical attributes. By using different data splits and multiple trials, the approach ensures a comprehensive understanding of the models' performance and generalizability in real-world scenarios.

### 2.4. Liver Disorders Dataset

The Liver Disorders dataset contains 345 data points and is utilized to predict liver disorder presence using biomedical indicators like mean corpuscular volume and alkaline phosphatase. The 'selector' attribute is the target for binary classification. As the features are numerical, no encoding is necessary. Three machine learning models—SVM, KNN, and RF are employed, optimized through GridSearchCV and StratifiedKFold cross-validation. The dataset is split into training and testing sets in three ratios for several trials and standardized using StandardScaler. Model performance is evaluated based on average training and testing accuracies across all splits and trials. This approach aims to assess the models' ability to predict liver disorders using biomedical data.

## 3. Test Results

For the Auto MPG dataset, the application of various classifiers revealed distinct patterns in performance, particularly in the Random Forest and SVM models. The Random Forest classifier achieved an exceptional average training accuracy of 1.0 and a test accuracy of 0.90, with the best parameters being 'max_depth': None and 'n_estimators': 100. The SVM classifier demonstrated a strong average training accuracy of 0.97 and a test accuracy of 0.85, optimized with 'C': 0.1 and 'gamma': 0.1.

In the Car Evaluation dataset the classifiers displayed varying efficacies across different test sizes. The Random Forest classifier consistently outperformed others, particularly with a 20% test size, achieving a test accuracy of 0.97. Decision Tree and KNN classifiers also showed strong results, with their performance fluctuating with changing test sizes.

For the Liver Disorders dataset, a diverse range of accuracies was observed across different partition types. The SVM classifier varied significantly in performance, achieving its highest accuracy of 0.72 at a 20/80 partition. The KNN classifier's best performance was at a 20/80 partition, whereas the Random Forest classifier maintained the most consistent performance across partitions. The optimal parameters for each classifier varied, underlining the importance of hyperparameter tuning in model performance. Furthermore, heatmaps plotted for each dataset in individual trials illustrate these trends vividly.

These results underscore the significance of choosing the right classifier and hyperparameters for different datasets and partition types. The Random Forest classifier emerges as a strong contender across multiple datasets, while the SVM and KNN classifiers show variable performance, indicating the need for careful model selection and tuning based on specific dataset characteristics.

## 4. Conclusion

In summary, this comparative analysis of machine learning classifiers provides insightful findings that align with and expand upon Caruana and Niculescu-Mizil's empirical study. The results reveal that the Random Forest classifier outperformed others across the datasets, showcasing its robustness and high accuracy in classification tasks. This finding supports the efficacy of Random Forests noted by Caruana and Niculescu-Mizil. Meanwhile, the performance of SVM and KNN classifiers varied significantly, emphasizing the importance of context-specific classifier selection and hyperparameter tuning. For example, the SVM classifier showed high accuracy in the Auto MPG dataset but varied performance in the Liver Disorders dataset, indicating its sensitivity to specific dataset characteristics.

Overall, this study not only validates general trends in classifier performance but also provides insights into the behavior of various classifiers under different conditions. It underscores the significance of matching the right machine learning algorithm to the specificities of the dataset at hand, which is a crucial step in the successful application of machine learning techniques in real-world scenarios.

## References

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd International Conference on Machine Learning (pp. 161–168). https://doi.org/10.1145/1143844.1143865

Quinlan, Ross (1993). Auto MPG. UCI Machine Learning Repository. https://doi.org/10.24432/C5859H.

Bohanec, Marko. (1997). Car Evaluation. UCI Machine Learning Repository. https://doi.org/10.24432/C5JP48.

Liver Disorders. (1990). UCI Machine Learning Repository. https://doi.org/10.24432/C54G67.