**Wine Classification**

**Omkumar Patel   
 Northwood University**

**Solv Probs W/ Machine Learning- MGT-665  
 Professor: - Dr. Itauma   
 June 15, 2025**

# Abstract

In this paper, I investigated the classification performance of three well known classifiers Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree on the Wine dataset. All models were trained and evaluated with the main performance measures accuracy, precision, recall and F1-score. The experiments results are visualized in confusion matrices, a correlation heatmap, as well as a bar chart which compares macro- average scores. Results The winning algorithms are Logistic Regression and Decision Tree with k-NN having a little less precision. The investigation highlights the importance of the model choice when tailored to the data.

# Introduction

One of the most important supervised machine learning problems is classification that has important applications such as health, marketing, and finance industries where making predictions with high accuracy is essential. The objectives of these jobs are to fit input data into a few pre-defined classes learned from past data. Owing to the exponential expansion of data and computational capacity, humans now increasingly rely on ML models to facilitate and optimize classification tasks. In this study, the Wine dataset from the UCI Machine Learning Repository was applied for evaluation, and three popular classification models, e.g., Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree were employed side by side. These algorithms have gained great popularity in research and industry due to their interpretability, the simplicity of the algorithm and their excellent performance across a wide range of applications (Han et al., 2011). The objective of this publication is to compare these methods in a more realistic environment, i.e., on a non-binary problem with a multiclass structure and to provide the reader with an insight into which task these algorithms are better suited for when it comes to classification.

# Related Work

Logistic Regression is perhaps the most widely used algorithm for binary and multi class classification problems (Hosmer et al., 2013). k-NN, a distance-based classifier, is often the first choice for problems with non-linear patterns and takes advantage of feature scaling (Altman, 1992). Decision Trees have been used for their interpretability and the possibility to efficiently deal with non-linear relationships without feature scaling (Quinlan, 1986). Comparing these models provides insights into algorithm selection strategies.

# Methodology

The Wine dataset, consisting of thirteen chemical features of wines from three cultivars, was divided into training and testing sets. Features were scaled for Logistic Regression and k-NN. The training of each model was done using scikit-learn, and classification metrics were employed for evaluation. Besides, confusion matrices and a correlation heatmap were created for visualization. The comparison was based on the macro averaged precision, recall and F1 scores.

# Results

Figures 1, 2, and 3 depict confusion matrices of each classifier. Figure 4 is the correlation heatmap of the dataset features, and Figure 5 is a comparative bar chart of macro average scores.

Figure 1: A graph of a logistic regression

AI-generated content may be incorrect.

Figure 2: A graph with green squares and numbers

AI-generated content may be incorrect.

Figure 3: A diagram of a decision tree

AI-generated content may be incorrect.

Figure 4: A diagram of wine data

AI-generated content may be incorrect.

Figure 5: A graph of different colors

AI-generated content may be incorrect.

# Discussion

Three different algorithms, which is to say all three models, have been found to be quite accurate when it comes to classification tasks with the Wine dataset. Logistic regression showed both high precision and recall figures and was thus very suitable for balanced data. Though intuitive, k-NN suffered a slight decrease in F1-score, thus it was less effective probably because it was sensitive to the choice of k and the way features were scaled. Decision trees could be a proper choice as they combine the features of interpretability and performance, however, they are liable to overfitting if no pruning techniques are applied.

# Conclusion

This article empirically explores the performances of Logistic Regression, k-NN and Decision Trees which align well with the multi classification task. However, the different classification algorithms were found to be quite similar in their behavior with Logistic Regression and Decision Trees performing more consistently across various evaluation metrics. The choice of the model should be based not only on the data characteristics but also on how much interpretability is needed and the cost of computation.

# References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175–185.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1, 81–106.