COIS 4400H

Lab 3

Winter 2025

Dikshith Reddy Macherla

Student Id : 0789055

Bachelor Of Computer Science, Trent University

COIS 4400H- Data Mining

Professor : Sabine McConnell

3rd March, 2025

Lab Report: Additional Classifiers Using Scikit-Learn

Introduction

In this lab, I explored four different classification models—Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine (SVM)—using the Wine dataset from Scikit-Learn. This dataset contains three classes of wine, each with 13 attributes, such as alcohol content and flavonoids. My goal was to evaluate the performance of these classifiers using 10-fold cross-validation and determine which model performs best.

Methodology

I used Stratified 10-Fold Cross-Validation to ensure that each fold contained an equal proportion of classes. The classifiers I tested were:

1. Decision Tree – A simple, interpretable model, but prone to overfitting.
2. Random Forest – An ensemble method that improves generalization by averaging multiple decision trees.
3. Naïve Bayes – A probabilistic model based on Bayes' Theorem, which assumes feature independence.
4. Support Vector Machine (SVM) – A powerful model that works well with high-dimensional data.

After performing cross-validation, I identified Random Forest as the best-performing classifier. I then trained it on 80% of the data and tested it on 20%, generating a confusion matrix to analyze its predictions.

Results & Discussion

The Random Forest classifier achieved the highest accuracy of 98.89%, outperforming the other models. Naïve Bayes (97.78%) and SVM (96.11%) also performed well, while the Decision Tree had the lowest accuracy (89.97%). The confusion matrix showed that Random Forest correctly classified all test samples with no misclassifications.

This makes sense because Random Forest reduces overfitting by averaging multiple decision trees. On the other hand, the Decision Tree overfitted the data, leading to lower accuracy. SVM and Naïve Bayes performed well, but Naïve Bayes assumes independence between features, which may not always hold in real-world datasets.

Conclusion

Through this lab, I learned that Random Forest is an excellent classifier due to its ability to reduce overfitting and improve generalization. This experiment highlighted the importance of using cross-validation to assess model performance before making a final selection. I also gained a deeper understanding of how different classifiers work and their advantages and limitations.