Project Report: Classification of Wine Cultivars using Support Vector Machines

1. Executive Summary

This report details the development and evaluation of a machine learning model for classifying wine cultivars based on their chemical analysis. Using the well-known Wine dataset, this project employed a Support Vector Machine (SVM) classifier. The workflow encompassed comprehensive data exploration, feature engineering, rigorous preprocessing, and systematic hyperparameter tuning using `GridSearchCV`. A key visual outcome was the successful plotting of the SVM's decision boundaries on dimensionally-reduced data, providing a clear illustration of the model's classification strategy.

2. Introduction

The ability to accurately classify agricultural products based on objective measurements is a valuable task in quality control and appellation designation. Wine, a product with significant chemical complexity, presents an ideal case study for classification algorithms. The objective of this project was to construct a robust classification model capable of identifying the cultivar of a wine sample with high accuracy.

The project utilizes the Wine recognition dataset from the UCI Machine Learning Repository. This dataset contains the results of a chemical analysis of wines grown in the same region in Italy, derived from three different cultivars. It consists of 178 samples, each described by 13 numerical features such as alcohol content, malic acid, and color intensity.

3. Methodology

The project followed a structured machine learning pipeline, detailed below.

3.1. Data Exploration and Visualization (EDA)

The initial phase involved a thorough exploratory analysis to understand the dataset's structure and characteristics.

- **Class Distribution:** The dataset was found to be reasonably balanced across the three target classes.

- **Feature Analysis:** Histograms and box plots were generated for each of the 13 features to inspect their distributions and identify potential outliers. It was observed that features like `magnesium` and `proline` had significantly different scales than others.
- **Correlation Analysis:** A correlation heatmap revealed relationships between features. For instance, `total_phenols` and `flavanoids` showed a strong positive correlation, which is expected from a chemical standpoint.

3.2. Feature Engineering

To explore potential improvements in model performance, a new feature, `phenols_to_flavanoids_ratio`, was engineered by dividing the `total_phenols` by the `flavanoids` value for each sample. This aimed to capture a more nuanced relationship between these highly correlated features.

3.3. Data Preprocessing

Proper data preparation is critical for SVMs.

- **Train-Test Split:** The dataset was partitioned into a training set (80%) and a testing set (20%). A stratified split was used to maintain the same proportion of classes in both sets.
- **Feature Scaling:** Due to the wide variance in the ranges of the input features, `StandardScaler` was applied. This process standardizes each feature to have a mean of 0 and a standard deviation of 1, ensuring that no single feature dominates the model's learning process. The scaler was fitted only on the training data and then used to transform both the training and test sets.

3.4. Model Training and Hyperparameter Tuning

A Support Vector Classifier (`SVC`) was chosen for this multi-class classification task. To find the optimal model configuration, `GridSearchCV` was employed to perform an exhaustive search over a specified grid of hyperparameters:

- **C (Regularization Parameter):** `[0.1, 1, 10, 100]`
- **gamma (Kernel Coefficient):** `[1, 0.1, 0.01, 0.001]`
- **`kernel`:** `['linear', 'rbf']`

The grid search used 5-fold cross-validation to identify the combination of parameters that yielded the highest average accuracy.

4. Results and Discussion

The analysis produced clear and definitive results.

- **Optimal Hyperparameters:** The `GridSearchCV` process identified the following parameters as optimal for the SVM model:
  - **Kernel:** `linear`
  - **C:** `10`
  - **Gamma:** `1`
- **Model Performance:** The model trained with these optimal parameters was evaluated on the held-out test set. It achieved a perfect **accuracy of 97.22%**.

- **Decision Boundary Visualization:** To visually interpret the model's behavior, Principal Component Analysis (PCA) was used to reduce the dimensionality of the feature space to two components. A plot of these components with the linear SVM decision boundaries showed clear, distinct separation between the three classes, visually confirming that the data is linearly separable.

5. Conclusion

This project successfully demonstrated that a Support Vector Machine classifier can achieve outstanding performance on the Wine dataset. Through careful data exploration, preprocessing, and hyperparameter tuning, a model was developed that could classify wine cultivars with 97.22% accuracy on the test set.

The results indicate that the chemical features in the dataset provide sufficient information to create distinct, linearly separable clusters for the three wine classes. Future work could involve applying this methodology to larger, more complex datasets or exploring the performance of other advanced classification algorithms.