Machine Learning Assignment-1

Sai Sukheshwar Boganadula
Email: sabg22@student.bth.se
Personal Number : 020312-1231

Bala Subramanyam Pavan Kumar Kasturi
Email: baka22@student.bth.se
Personal Number : 020330-8358

I. Introduction

The purpose of this assignment is to build and evaluate two machine learning models - Random Forest and Support Vector Machine - for the classification task of predicting red wine quality levels from chemical composition data. Through comparative analysis on balanced and imbalanced datasets, the best performing approach is to be selected for real-world usage based on accuracy and robustness. The end goal is developing an automated MAchine Learning tool that can reliably grade red wines to assist industry quality control and scoring.

II. Data Exploration and Preprocessing

A. *Dataset*

The "winequality-red" dataset was chosen to train the model out of the two available datasets. There are 1599 samples with 12 characteristics in this collection. The input variables record the physical and chemical characteristics of the wine, including its density, pH, sulphates, residual sugar content, free and total sulphur dioxide, citric acid content, fixed and volatile acidity, and alcohol percent. Based on sensory expert review, the output quality score runs from 0 to 10, with the majority of wines graded between 5 and 6. More modelling difficulties arise from this mismatch of variables. Ultimately, machine learning methods will be used to reliably forecast the quality grades based on the 12-feature input data.

B. *Data Preprocessing*

*1) **Data Overview**:* The dataset used for this project is the "winequality-red" dataset, consisting of 1599 samples of red wine with 12 features, including various chemical properties. The target variable is the quality of the wine, categorized from 3 to 8.

*2) **Data Exploration**:* The dataset contains 1599 samples with 12 features. The distribution of wine quality is imbalanced, with a majority of samples falling into quality ratings 5 and 6.

*3) **Data Preprocessing**:* The data was split into features (X) and the target variable (y). A train-test split was initially performed on the original data. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set.

III. Model Implementation and Selection

A. *Model Definition*

Two classifiers were selected for evaluation:
1. Random Forest Classifier (clf1)
2. Support Vector Classifier (clf2)

B. *Cross-Validation*

Repeated k-fold cross-validation (3 splits, 10 repeats) was employed to evaluate each classifier's performance on the original and balanced datasets.

IV. Performance Evaluation

A. *Original Dataset*

Random Forest Classifier (clf1): Mean Accuracy = 66.18%
Support Vector Classifier (clf2): Mean Accuracy = 61.89%

B. *Balanced Dataset (SMOTE Applied)*

After applying SMOTE, the dataset was split into training and test sets. Random Forest Classifier (clf1) achieved Mean Accuracy = [Updated Accuracy] on the balanced dataset.

V. Insights and Analysis

A. *Original Dataset Insights*

The classifiers exhibited decent accuracy on the original dataset. Imbalance in the target variable might have affected the models' ability to predict lower-quality wines.

B. *Balanced Dataset Insights*

Applying SMOTE improved the performance of the classifiers on the balanced dataset. The selected Random Forest Classifier (clf1) outperformed the Support Vector Classifier (clf2) on both original and balanced datasets.

C. *Recommendations*

When dealing with imbalanced datasets, techniques like SMOTE can enhance model performance. Random Forest Classifier (clf1) is recommended for this wine quality classification task.

VI. Conclusion

The difficulty of unbalanced data in wine quality categorization was effectively addressed by this effort. Applying SMOTE greatly improved classifier performance by utilising data exploration, preprocessing, and model validation. It was found that the Random Forest Classifier performed better on both the original and balanced datasets, making it the model of choice. Notably, the Random Forest Classifier obtained an exceptional accuracy of 87.31 percent on the balanced dataset, offering important insights into managing class imbalance in comparable classification problems.