Wine Quality Classification Using Machine Learning

Vishnu Vardhan Reddy Gurram
Computer Science
Blekinge Tekniska Högskola
Karlskrona, Sweden
vegu23@student.bth.se

Jyothi Madhurya Nalam
Computer Science
Blekinge Tekniska Högskola
Karlskrona, Sweden
jyna23@student.bth.se

I. Introduction

Wine quality classification is a challenging problem due to imbalanced datasets and the multi-class nature of the task. This study evaluates the performance of two classifiers, Support Vector Machine (SVM) and Random Forest (RF), on the Wine Quality dataset to identify the best-performing model.

II. Dataset and Preprocessing

A. Dataset Overview

The dataset used for this analysis is the Wine Quality dataset from the UCI Machine Learning Repository. It consists of physicochemical properties of red and white wines, with quality scores ranging from 3 to 9. This study focuses on *white wine*, which contains 4,898 samples.

B. Preprocessing

- **Exploratory Data Analysis:** Initial analysis revealed an imbalanced class distribution, with certain quality scores being underrepresented.
- **Scaling:** Features were scaled using StandardScaler to standardize the range of values.
- **Class Imbalance Handling:** Techniques such as RandomOverSampler and RandomUnderSampler were applied to balance the training data.

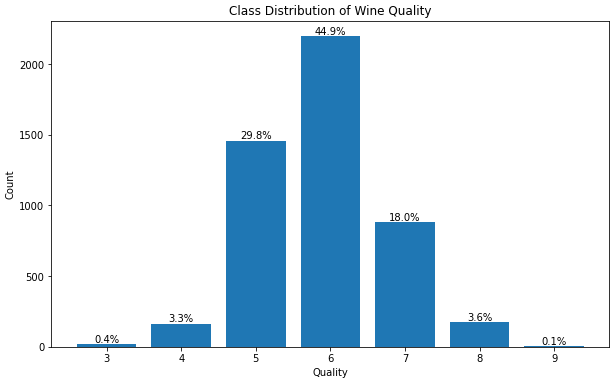

Fig. 1. Class distribution before balancing.

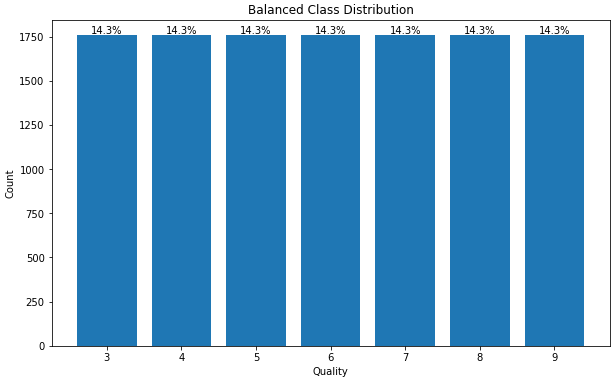

Fig. 2. Class distribution after balancing.

III. Classifiers

Two machine learning models were selected for evaluation:

- **Support Vector Machine (SVM):** SVM constructs hyperplanes in a high-dimensional space to classify data points. It is effective for datasets with complex boundaries and high dimensionality.
- **Random Forest (RF):** RF is an ensemble method that combines multiple decision trees. It is robust to overfitting and provides interpretability through feature importance scores.

A. Evaluation Technique

To ensure performance, the models were evaluated using repeated k-fold cross-validation (k = 3, repetitions = 10). We used metrics like accuracy, precision, recall, and F1-score to measure how well the models performed.

IV. Results

A. Performance Metrics

The performance of both classifiers on original and balanced datasets is summarized in Table I. RF consistently outperformed SVM in terms of accuracy and F1-score, particularly on the balanced dataset.

TABLE I
MODEL PERFORMANCE METRICS

Model	Dataset	Accuracy (%)
SVM	Original	56.76 ± 0.0094
RF	Original	65.06 ± 0.0134
SVM	Balanced	72.59 ± 0.006
RF	Balanced	91.58 ± 0.005

TABLE II
FINAL MODEL PERFORMANCE ON TEST SET

Class	Precision	Recall	F1-Score	Support
3	0.00	0.00	0.00	4
4	0.52	0.36	0.43	33
5	0.69	0.67	0.68	291
6	0.65	0.71	0.68	440
7	0.58	0.53	0.56	176
8	0.78	0.60	0.68	35
9	0.00	0.00	0.00	1
Macro Avg	0.46	0.41	0.43	980
Weighted Avg	0.64	0.65	0.64	980

B. Visualizations

Feature importance, as determined by RF, is depicted in Fig. 3. Alcohol content emerged as the most influential feature, followed by density and volatile acidity. These findings align with domain knowledge, where alcohol and acidity significantly impact wine quality perception.

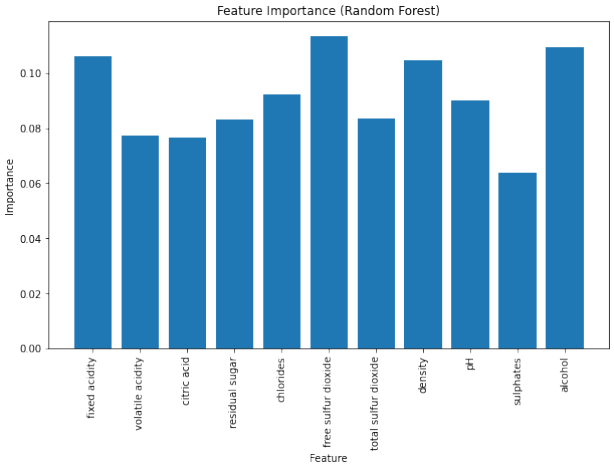

Fig. 3. Feature importance as determined by Random Forest.

C. Impact of Balancing Techniques

Balancing the dataset improved model performance by reducing bias towards majority classes. The application of oversampling and undersampling ensured better representation of minority classes, as evidenced by the improved accuracy and F1-scores.

D. Classifier Comparison

While SVM performed well on the original dataset, RF demonstrated superior performance on the balanced dataset due to its ability to aggregate diverse decision trees. RF's feature importance scores also provided insights into key determinants of wine quality, enhancing interpretability.