ST4248 Statistical Learning II

2020/2021 Sem 2: Term Paper Report

Title: Wine Quality Classification

A0176820A

Summary:

In this analysis, the main aim is to build the best classification model that could accurately classify wine quality based on chemical variables and sensory variables. The data used in this project is the Red Wine Quality dataset obtained from Kaggle. Algorithms such as Linear Discriminant Analysis, Quadratic Discriminant Analysis, Support Vector Classifier, Decision Trees and Random Forest will be compared, and the best performing model will be selected for this problem.

Table of content:

1. Introduction

Wine is an alcoholic drink that is normally made up of fermented grape. Common types of wine include red wine, white wine and many other types of varieties. There are many variables that could affect wine quality such as *pH*, *acidity*, *sugar* and others. The goal of this analysis is to classify wine quality based on chemical variables (e.g. *citric acid*, *fixed acidity*) and sensory variables (e.g. *pH*, *density*) into 3 different wine quality categories: “bad”, “normal” and “good”.

Dataset

The dataset used in this project is the Red Wine Quality dataset obtained from Kaggle[1] which contains 1599 rows and 12 variables. Some variables include *citric acid* which contributes to the freshness of wine, *residual sugar* which is leftover natural sugar from grapes after the fermentation process has stopped, *alcohol* which is the percentage of alcohol in wine. This dataset had no missing values.

2. Data Processing

There are a total of 6 wine quality ranging from 3 to 8 out of 10. We will segment wine quality into 3 different categories. Firstly, “bad” wine quality with ratings 3 and 4. Secondly, “normal” wine quality with ratings 5 and 6. Lastly, “good” wine quality with ratings 7 and 8.

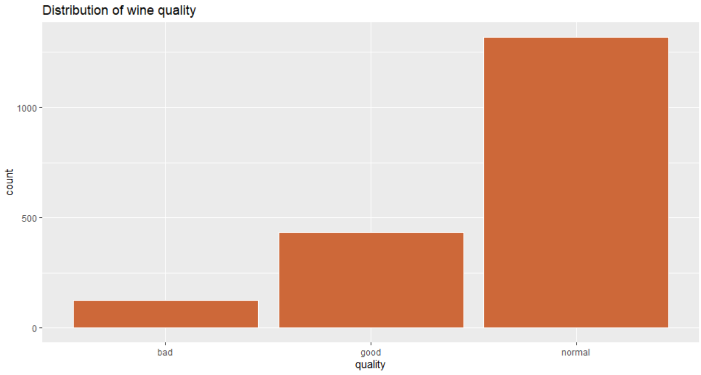

Figure 2.1: Wine Quality

From figure 2.1, we noticed a presence of an imbalanced data in the “bad” and “good” quality wine. As a result, we will proceed to up-sample our dataset by resampling with

[1] https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

replacement such that all classes have the same number of samples. We will then split our dataset into 80% training set and 20% test set for our analysis.

3. Exploratory Data Analysis

Correlation Plot

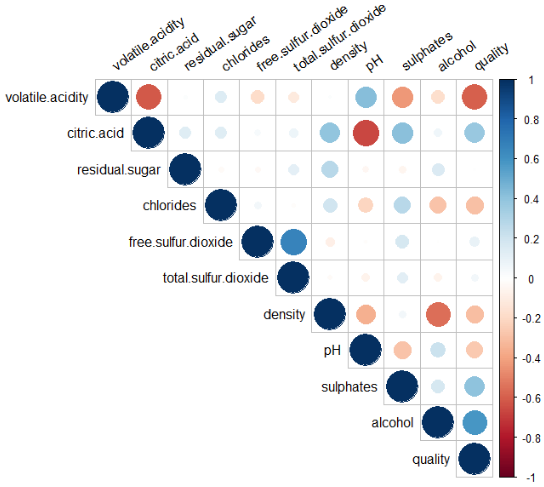

Figure 3.1: Correlation Plot

Pairwise correlation was checked to identify any highly correlated variables in the dataset. From figure 3.1, no variables are extremely correlated with each other.

Summary statistics

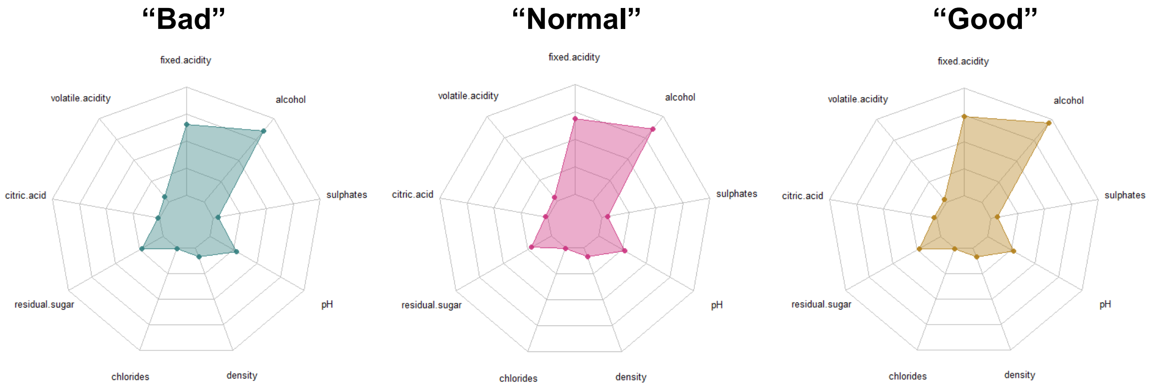

Figure 3.2 : Radar plot of different wine category

From the above radar plots of the different wine categories (figure 3.2) , all three categories of wine seem to have very similar overall statistics for the input variables. The “good” wine quality on the far right has the highest alcohol content and citric acid, while having the lowest volatile acidity as compared to the other 2 categories. Variables such as pH and density are rather similar for all 3 categories.

4. Models

Discriminant Analysis

We will first use Linear Discriminant Analysis (LDA) for our classification. LDA assumes that all classes share a common covariance, leading to lower variance in the bias-variance trade-off. This can help to improve predictions. The accuracy score obtained from LDA is low at 0.698232. Following LDA, we will next use Quadratic Discriminant Analysis (QDA). QDA removes the assumption of common covariance, which leads to lower bias in the bias-variance trade-off. QDA performed slightly better than LDA with an accuracy score of 0.719697.

Support Vector Classifier

Following LDA and QDA, we move onto support vector classifiers as it allows for greater robustness to individual observations and allows for better classification of most of the training observations. We performed tuning on parameters “cost” and “kernel” using a 5-fold cross-validation. We obtained the best accuracy score of 0.666667 with the chosen parameters of a linear kernel and cost of 1.4.

Decision Trees

Next, we explore decision trees due to its easy interpretation and graphical ability. Similar to support vector classifiers, we will perform tuning via 5-fold cross-validation on criterion, number of features to use, and the maximum depth of the decision tree. Parameters chosen were: entropy and maximum depth of each tree as 18. Decision tree managed to obtain a high accuracy score of 0.957071.

Random Forest

Random forest classifier was the next model considered as it could reduce high variance of a single classification tree and also reduce highly correlated trees when averaging the trees. To reduce highly correlated trees, only a random subset of $\sqrt{p}$ predictors out of all *p* predictors are chosen as split candidates at each tree split. We perform tuning via 5-fold cross-validation on criterion, maximum depth of tree and the number of trees to use. Parameters chosen were: entropy, maximum depth of each tree as 14, maximum number of features used for each split as 3 and number of trees as 200. Random Forest obtained a high accuracy score of 0.968434.

5. Evaluation

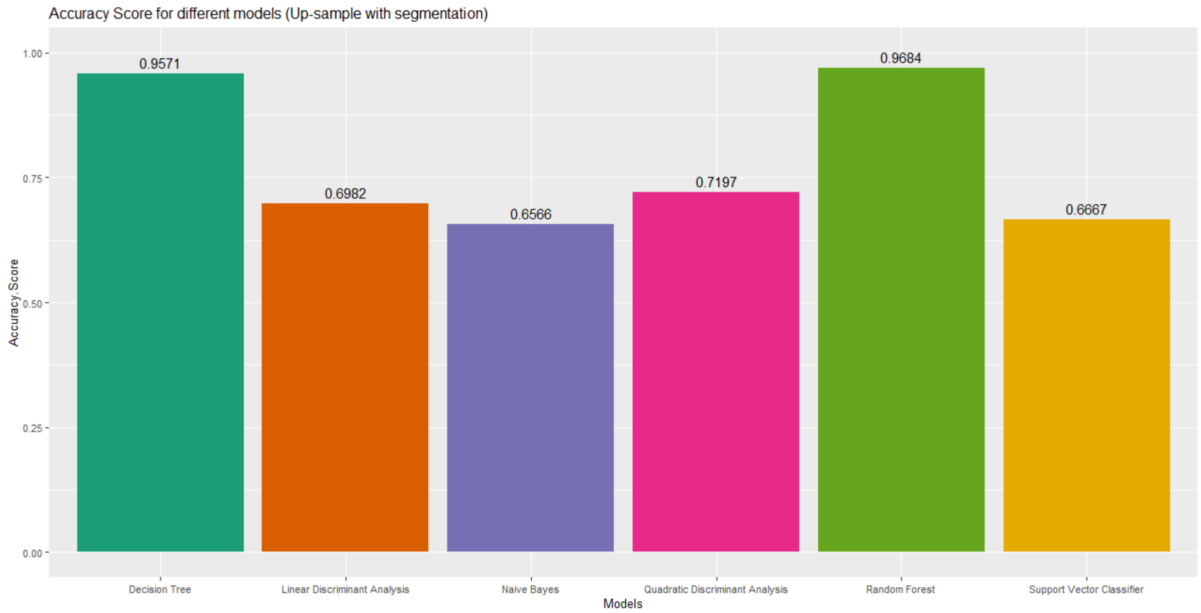

Figure 5.1: Comparison of models

From figure 5.1, comparing the performance of all models, Random Forest performs the best followed by decision trees in terms of accuracy score. Thus, we will pick Random Forest as our wine quality classification model. We obtained the feature importance from Random Forest with the top 3 most important variables: *sulphates*, *volatile acidity* and *alcohol*.

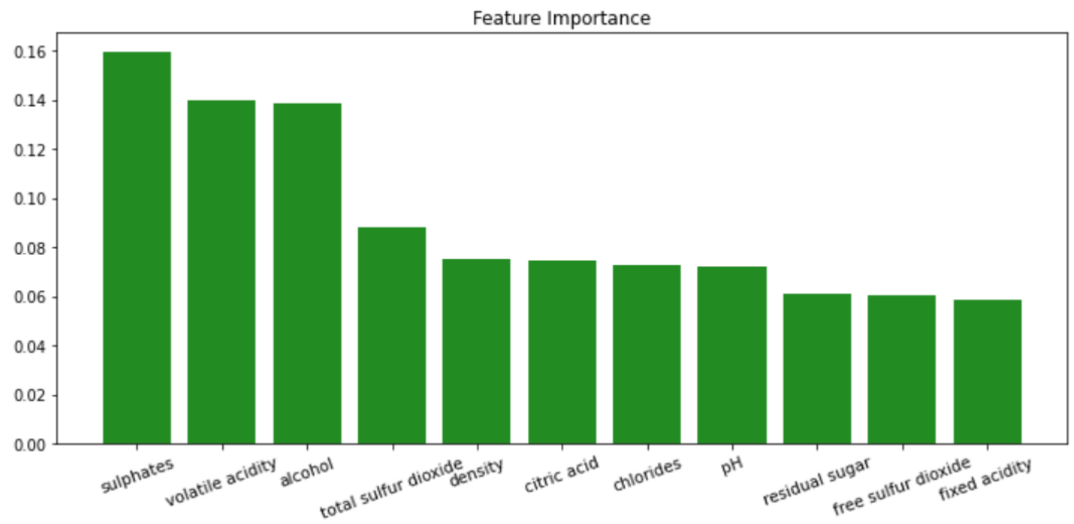

Figure 5.2: Features' importance

Models:	Up-sampling with segmentations	Up-sampling without segmentations	No up-sampling and segmentations
Random Forest	0.968434	0.922983	0.721875
Decision Trees	0.957071	0.909535	0.615625
Support Vector Classifier	0.666667	0.545232	0.609375
Linear Discriminant Analysis	0.698232	0.486553	0.600000
Quadratic Discriminant Analysis	0.719697	0.666259	0.553125

Table 5.3: Comparison of accuracy scores

We also compare the above models with the performance of up-sampling the data without segmenting the wine quality and the original data without up-sampling and segmenting. Table 5.3 shows the performance of all models when the dataset was up-sampled but target variable was not segmented into 3 different categories (table 5.3: "Up-Sampling without segmentations"). The accuracy score of all models are slightly lower compared to when the dataset was up-sampled and segmented, with Random Forest performing the best followed by decision trees.

Table 5.3 also shows the performance of all models when the dataset was not up-sampled and not segmented (table 5.3: "No up-sampling and segmentations"). The models generally performed worse off suggesting that up-sampling and segmenting the target variable into categories improves the performance of this dataset significantly.

6. Conclusion

In conclusion, up-sampling and segmenting the target variables into 3 different wine categories (“bad”, “normal” and “good”) performs the best for this dataset and the best model for classifying wine quality is Random Forest. Thus, wine quality can be classified highly accurately based on the chemical variables and sensory variables available in this dataset.