CS675 - Machine Learning - Project Report
Wine Quality Prediction using Machine Learning

1 Introduction

In this project we analyzed the various factors that influenced the quality of wine. We are using Machine Learning model on the wine data set which has various attributes that effect the quality of wine. We performed different Exploratory Data Analysis on different features which told us the importance of it. Later we have used Random Forest Classifier model to predict whether the wine is good or bad. The result we got predicted the quality of wine with 76 percent accuracy.

2 Description of the project

A simple yet challenging project, to anticipate the quality of wine. The complexity arises due to the fact that the data set has fewer samples is highly imbalanced. Can you overcome these obstacles build a good predictive model to classify them? Our features in the data set are continuous but all have different ranges, so we normalized them by dividing with the respective maximum of each feature which reduces the range between 0 and 1. Now that we have a normalized data set we have used Machine Learning model, in this case random forest, which is a ensemble model which creates decision trees. Random forest is a versatile, convenient machine learning technique that, in most cases, gives good results even without hyper-parameter adjustment. Due to its simplicity and adaptability, it is also one of the most widely used algorithms it can be used for both classification and regression tasks, so we have used the random forest classifier for our project.

3 Observations

1.Volatile acidity has a correlation value of 0.407.
2.Looks like higher levels of total sulfur dioxide means higher values of free sulfur dioxide.
3.Low levels of free sulfur dioxide and total sulfur dioxide usually mean a better quality.

4.Citric acid has a correlation value of 0.241.
5.Wines with high levels of citric acid usually fall into the quality category of 0 and 5.
6.Alcohol has the greatest value of correlation, a correlation value of 0.485.

4 Experiments

First we tried to use Support Vector Machine (SVM) model but due to the different ranges of the features accuracy turned out to be just 50 percent which was not good.
Later we tried decision tree classifier but because of lots of data the overfitting has occured and the pruning method made the problem more complicated which gave only 58
We also tried XGBoost but there were so many features that meant lots of unstructured data so even with this model we got only 66 percentage of accuracy.
As there were different ranges in the features we tried to classify them with KNN but even with the KNN we could not find ideal k-value for the data set and the best accuracy we got is 61 percent.
Finally we have used Random Forest Classifier which gave us the highest accuracy of 76 percent.
We have trained the model on previously normalized data set, we used the similar data set to predict the model by splitting 25 percentage of the data set for testing and 75 percentage for training. To do this we have used sklearn's train test split method by setting the shuffle attribute to true and random state to 0.33

5 Conclusion

Using the ML models, we have analyzed which of the features were contributing more to the result and which were not and we have predicted quality of the wine which turned out be 76 percent. We introduced the RF and AdaBoost model as a machine learning classifiers to predict wine quality after evaluating its performance based on the accuracy,precision, recall, F1 scores, the ROC-AUC score. According to the results, AdaBoost predicted wine quality with higher accuracy during without feature selection, with feature selection (XGB) and with essential variables. Overall, performance of all classifiers (except KNN)improved when model trained and tested using essential variables. The usefulness of data generation algorithms and importance of feature selection is the key feature in this study. We are in progress of developing a machine learning-based web application that wine researchers and wine growers can use to predict wine quality based on the important available chemical and physio-chemical compounds in their wines, one that has the capability to tune various variable quantities.

6 Contributions

Shravani Aedla - Understood how classifiers work, normalised the data and trained the model and used Random Forest ML model.
Harini Reddy Gade - Introduced XGBoost model and further classified data using KNN.
Nandini Palwai - Learned about SVM and implemented it to our data set.