CS675 - Machine Learning - Project Report

Wine Quality Prediction using Machine Learning

Table of contents

1 Team Members

Saksham Jain

Vijay Muni Reddy

Sriram Gottipati

2 Description of the project

A simple yet challenging project, to anticipate the quality of the wine. The complexity arises because the data set with fewer samples is highly imbalanced. Can you overcome these obstacles and 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 them by 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 the Machine Learning model, in this case, random forest, which is an ensemble model which creates decision trees. Random forest is a versatile, convenient machine learning technique that usually 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. Higher levels of total sulfur dioxide mean higher values of free sulfur dioxide.
3. Low levels of free sulfur dioxide and total sulfur dioxide usually mean 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 the decision tree classifier but because of lots of data overfitting occurred and the pruning method made the problem more complicated which gave only 58. We also tried Random forests and decision trees but there were so many features that meant lots of unstructured data so even with this model we got only 66 percent 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 the ideal k-value for the data set and the best accuracy we got is 61 percent. Finally, we used the Random Forest Classifier which gave us the highest accuracy of 76 percent. We have trained the model on previously normalized data sets, we used a similar data set to predict the

model by splitting 25 percent of the data set for testing and 75 percent 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 analyzed which of the features were contributing more to the result and which were not and we predicted the quality of the wine which turned out to be 76 percent. We introduced the RF as a machine learning classifier to predict wine quality after evaluating its performance based on accuracy, precision, recall, F1 scores, and the ROC-AUC score. According to the results, AdaBoost predicted wine quality with higher accuracy without feature selection, with feature selection, and with essential variables. Overall, the performance of all classifiers (except KNN) improved when the model was trained and tested using essential variables. The usefulness of data generation algorithms and the importance of feature selection is the key feature of this study. We are in the 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 physiochemical compounds in their wines, one that has the capability to tune various variable quantities.

6 Contributions

Saksham Jain - Understood how classifiers work, normalized the data, and trained the model. Performed data visualization Bivariate analysis.

Vijay Muni Reddy – Introduced Logistic Regression model, stochastic gradient descent classifier, and Random Forest.

Sriram Gottipati - Learned about SVM and implemented it to our data set.

7 Project Links

Code Link:

https://colab.research.google.com/drive/1chYbLPITch7bNWbdwzDswlfrfTSElf64?usp = sharing