PREDICTING QUALITY OF THE WINE USING MACHINE LEARNING TECHNIQUES

Nishitha Yendapally
Department of Computer Science and Electrical Engineering
Texas A&M University – Kingsville
Kingsville, Texas
nishitha.yendapally@students.tamuk.edu

Yogitha Reddy Koppolu
Department of Computer Science and Electrical Engineering
Texas A&M University – Kingsville
Kingsville, Texas
Yogitha_reddy.koppolu@students.tamuk.edu

Abstract

These days, manufacturing companies use product quality records to promote their goods. It is a tedious process that necessitates manual expert evaluation, making it extremely expensive. Hence, in this project we utilize the machine learning technique such as Random Forest for predicting the better quality of wine. This is used for identifying the dependency of target variable over independent variables.

KEYWORDS: Random Forest; wine quality; Heat map

INTRODUCTION

Industries are improving by incorporating new techniques and implementing these in all areas. Identifying the quality of wine is by experimental methods is tedious, time consuming and costly. The Features like alcohol, PH values, acidity, density, and residual sugar contents are all factors to consider when evaluating the standard of Wine.

Machine learning techniques helps to create models based on information from known category labels in order to anticipate a wine's standard. The Wine used to be thought of as a magnificence product. But now it is commonplace and adored by a huge number of people. Qualified wine research offers insight into wines that are mass-produced on a yearly basis. By utilizing the Machine learning (ML) approaches the wine quality is predicted easily with less cost.

1. METHODOLOGY

This chapter explains the project dataset, project Description and evaluation metrics and prediction model in more detail.

1.1 DATASET

The wine dataset [1] is segregated into two, wine_train dataset and wine_test dataset. Whereas 70% of the data is for training and 30% of the data is for independent testing of the model. The Data pre-processing should be established by comprising independent variables and measured variables, handling missing data, data cleaning and the data is splitted to train data and test data, as well as data encoding, are among the features.

1.2 DATA REPRESENTATION

The wine dataset collected in this project is .csv file. Firstly, we implemented data cleaning technique where the null values are identified and removed from the columns of both training and the test datasets. Here we used sk-learn, matplotlib and seaborn libraries. For the cleansed data correlation is determined. The below heat map represents the correlation between the factors in identifying the quality of wine.

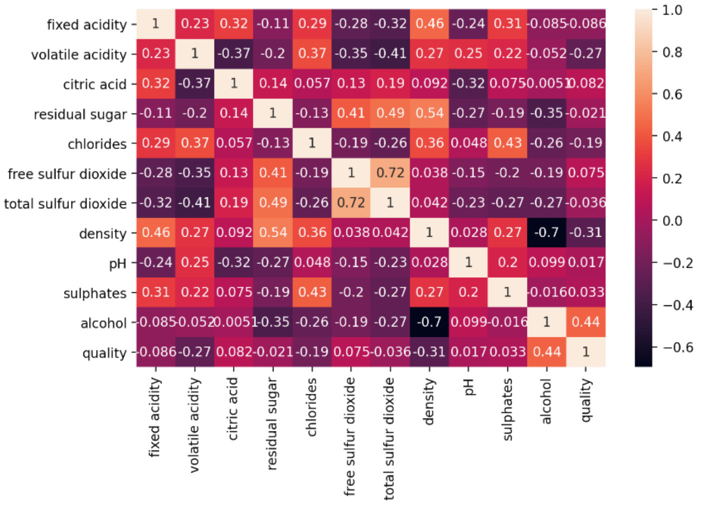


Figure1: Heat map representing the correlation of factors for identifying the better quality of wine.

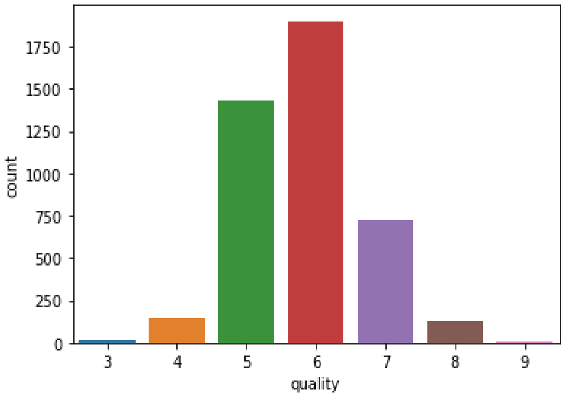


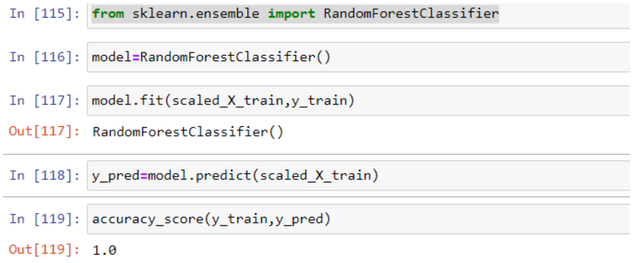
Figure2: Graph describing the quality of the wine.

For the trained data we perform Random Forest (RF) technique to find the accuracy score of the model.

Random Forest:

This technique employs a combination of tree indicators, with each tree relying on a random vector. For all trees in the forest, this unpredicted vector has identical circulation. Breiman did a portrayal of it in 2001[2]. RF is a simple method for predicting crucial variables in classification and regression issues.

The data is trained and fitted utilizing RF technique. The below code depicts the utilization of RF technique.



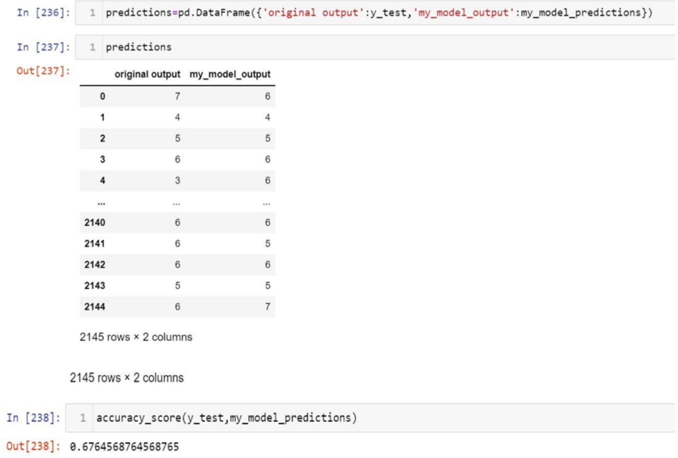
```
In [115]: from sklearn.ensemble import RandomForestClassifier

In [116]: model=RandomForestClassifier()

In [117]: model.fit(scaled_X_train,y_train)
Out[117]: RandomForestClassifier()

In [118]: y_pred=model.predict(scaled_X_train)

In [119]: accuracy_score(y_train,y_pred)
Out[119]: 1.0
```

RESULTS:

The wine dataset consists of unbalanced data after performing the feature scaling and implementing the RF classifier the accuracy of the model is calculated. We must calculate true-positive(TP), true-negative(TF), false-positive(FP), and false-negative(FN) values in order to calculate precision [3,4]. Therefore accuracy is derived as the fraction of correct predictions to the total number values to predict. Precision, also known as positive predict values, represents the likelihood of a positive projection of outcome.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} * 100\%$$

By calculating using the above equation the accuracy that is observed is 67%. (0.67). This analysis will provide a clear picture of the crucial qualities for quality prediction, as well as save the industries a lot of time and money.



```
In [236]: predictions=pd.DataFrame({'original output':y_test,'my_model_output':my_model_predictions})

In [237]: predictions
Out[237]:
```

	original output	my_model_output
0	7	6
1	4	4
2	5	5
3	6	6
4	3	6
...	...	...
2140	6	6
2141	6	5
2142	6	6
2143	5	5
2144	6	7

2145 rows × 2 columns

2145 rows × 2 columns

```
In [238]: accuracy_score(y_test,my_model_predictions)
Out[238]: 0.6764568764568765
```

CONCLUSION:

Wine quality can be classified as "Good" or "Bad" depending on a variety of characteristics. This research focuses on these parameters and uses a variety of machine learning algorithms to predict wine quality. RF used in this project is a robust algorithm for the prediction of quality wine. The accuracy for the random forest algorithm is 67%. In the future, we can experiment with different performance metrics and machine learning techniques to compare results more effectively.

This project will assist the wine industry in predicting the standards of many types of wines based on certain characteristics, as well as assisting them in producing high-quality products in the future.

REFERENCES:

1.V. Preedy, and M. L. R. Mendez, “Wine Applications with Electronic Noses,” in Electronic Noses and Tongues in Food Science, Cambridge,MA, USA: Academic Press, 2016, pp. 137-151.

2.W. L. Martinez, A. R. Martinez, “Supervised Learning” in Computational Statistics Handbook with MATLAB, 2nd ed., Boca Raton, FL, USA: Chapman & Hall/CRC, 2007, pp. 363-431.

3.How Are Precision and Recall Calculated? (2018). Retrieved April 2018, from KDnuggets: https://www.kdnuggets.com/faq/precisionrecall.html

4.Paul, D., Su, R., Romain, M., Sébastien, V., Pierre, V., & Isabelle, G., “Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier,” Computerized Medical Imaging and Graphics, 60 Elsevier, 2017, pp.42-49.