Wine Quality Prediction using Machine Learning

Nisha Indapure
Department of Electronics Engineering
Vishwakarma Institute of Technology
Pune,India
nisha.indapure18@vit.edu

Vaishnavi Ganji
Department of Electronics Engineering
Vishwakarma Institute of Technology
Pune,India
vaishnavi.ganji18@vit.edu

Abstract:**Wine Quality testing is difficult task as it needs to undergo various test parameters before sending it to shops.Since it is produced in large quantity,testing every bottle becomes a tedious job.An automative predcitive system can be used for testing which gives the quality of wine based on previous standard meansures of various wine componentsAlso,the quality of wine can be increased by adding or removing specific components in required amount.Since the quality are claasifed based on different features,the classification models used are :1)Multiple Linear Regression 2)Logistic Regression 3)Decision Tree 4)Random Forest.**

Keywords-WineQuality,MultipleLinear Regression

I.INTRODUCTION

All types of industries are improving by adopting new technologies.These technologies are helpful in increasing the production ans quality of product.This project mainly is build to support wine industry to increase their quality and provide user-friendly,easy and fast testing of wine.The main aim of this project is to build models to predict the rating on a scale of 1-10 of a wine.We have used samples of red wine as our dataset.The quality of wine depends on various factors like density, pH value,alcohol and other acids.These factors act as an input to the model and the rating of quality predicted is our output.The machine learning algorithms used are Multiple Linear Regression,Logistic Regression,Decsion Tree and Random Forest.

II.DATASET AND FEATURES

The data used in this project was obtained from UCI Machine Learning Repository.The dataset consists of 1599 samples of red wine. Each sample of red wine consists of 12 physiochemical variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality rating. The quality rating is based on a sensory test carried out by at least three sommeliers and scaled in 11 quality classes from 0 - very bad to 10 –excellent

TABLE 1:

Rating	**Quality**
0-2	Very bad
3-5	Bad
6-8	Good
9-10	Excellent

III.Methods

1.Multiple Linear Regression:We began with simple algorithm of regression.The Equation of MLR is:y=a0+a1x1+a2x2..anxn.Here,x are the multiple inputs which are various features of wine.Y is the output i.e the rating.After performing MLR,we got follwing predictions (fig.1).

The problem encountered in MLR was we couldn't predict the quality as good or bad because of continous data.Hence,we first converted the ratings into two categories:bad as '0' and good as '1'.

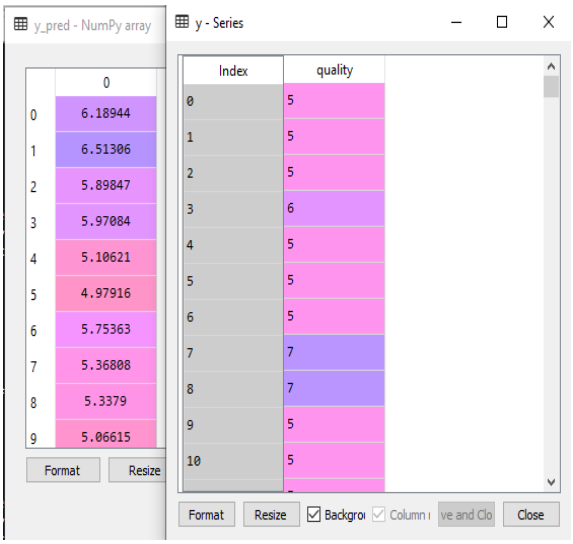

Fig.1:MLR Output

2.Logistic Regression (LR):

LR is used mostly where the output is either logic 0 or logic 1. As we classified the rating of wine into bad and good,we applied LR algorithm to our dataset. We then made the confusion matrix(fig:2) and found the accuracy of LR algorithm.(fig3).We decided to increase the accuracy of prediction by using next algorithm which is decision tree.

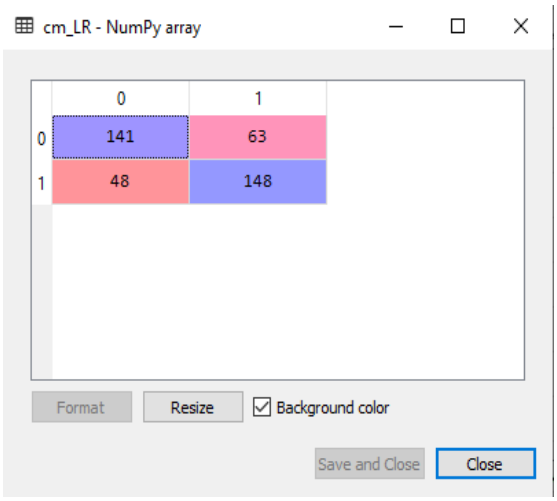

Fig.2: Confusion Matrix of LR

```
...: cm_LR=confusion_matrix(y_test,y_pred)
Accuracy of LR :  0.7325
```

Fig.3:Accuracy of LR

3.Decision Tree (DT):

A **decision tree** is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (**decision** taken after computing all attributes).After applying dataset to decision tree we made confusion matrix (fig.4) and accuracy(fig.5).The accuracy of Decision Tree is 77% which is greater than that of Logistic Regression.For further improvement we applied Random Forest Algorithm.

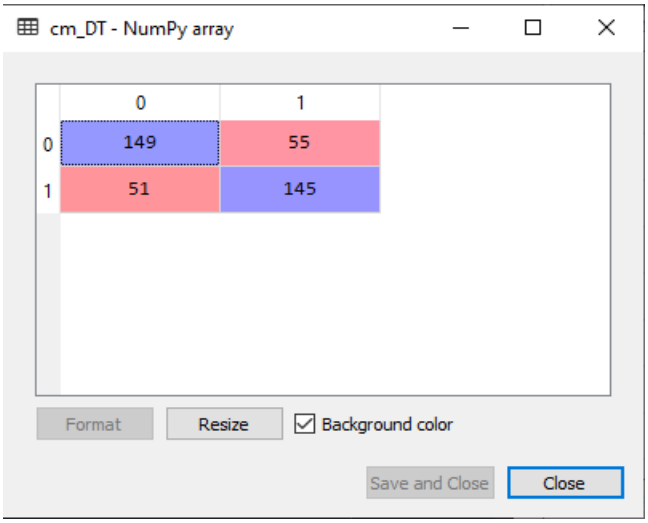

Fig.4:Confusion Matrix of DT

```
...: cm_DT=confusion_matrix(y_test,y_pred)
Accuracy of DT :  0.77
```

Fig.4:Accuracy of DT

4.Random Forest (RF):

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It predicts output with high accuracy, even for the large dataset it runs efficiently. After applying dataset to random forest we made confusion matrix (fig.5) and accuracy(fig.6).The accuracy of Random Forest is 80.75% which is greater than that of Decision Tree.

Fig.5: Confusion Matrix of RF

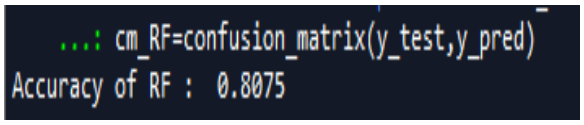

Fig.6: Accuracy of RF

IV. CONCLUSION & FUTURE WORK

We explored various machine learning models and concluded that random forest has best accuracy out of all explored models.Random Forest combines the output of multiple decision trees and calculates the majority and gives the final output.Hence,in wine quality prediction we use Random Forest Algorithm to get the quality rating.

In Future,we can explore different models like k-means clustering,Neural Network or Support Vector Machine.

V. References:

[1] Wine quality dataset: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[2] Predicting Wine Quality: web.stanford.edu/~ilker/doc/wine_Stats315A.pdf

[3]Amelia Lemionet, Yi Liu, Zhenxiang Zhou. Predicting quality of wine based on chemical attributes. *CS 229 project*, 2015. http: //cs229.stanford.edu/proj2015/245 report.pdf

[4]Fan Chao, Pengbo Li, Renxiang Yan. Predicting Review Rating for Wine Recommendation. https://cseweb.ucsd.edu/ ~jmcauley/ cse190/reports/fa15/020.pdf