

Data Mining Techniques ITE2006

PROJECT REPORT

On

Prediction of wine Quality using Data minig techniques

Under the guidance of

LAKSHMI PRIYA GG

Submitted by

LakshyaPurohit	15BIT0041
Mohit Kewalramani	15BIT0048
Sai Srivastav Rajput	15BIT0181



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Winter Semester 2017-18

CERTIFICATE

This is to certify that the project work entitled “**Prediction of wine Quality using Data minig techniques**” for the course DATA MINING(ITE2006) is a bonafide work done under mysupervision.

Place: Vellore

Date: 21/03/2018

Signature of Students:

LAKSHYA PUROHIT

MOHIT KEWALRAMANI

SAI SRIVASTAV RAJPUT

Signature of Faculty: Prof. Lakshmi Priya GG

ACKNOWLEDGEMENT

We sincerely thank our Prof. LAKSHMI PRIYA GG who supported us throughout our journey. We would also like to acknowledge VIT University for giving the candidates an opportunity to carry out our studies at the university.

We have dedicated a great amount of time and energy to properly research and document this project. Still, implementation would not have been possible if we did not have a support of many individuals and lab staffs. Therefore, we would like to extend our sincere thanks to all of them who made this project possible.

TABLE OF CONTENTS

- I. ABSTRACT
- II. PROBLEM STATEMENT
- III. MOTIVATION
- IV. INTRODUCTION
- V. MODEL DESCRIPTION
- VI. DATASET DESCRIPTION
- VII. EVALUATION MEASURES
- VIII. RESULTS
- IX. CONCLUSION

ABSTRACT

Certification and quality assessment are crucial issues within the wine industry. Currently, wine quality is mostly assessed by physicochemical (e.g. alcohol levels) and sensory (e.g. human expert evaluation) tests. In this paper, we propose a data mining approach to predict wine preferences that is based on easily available analytical tests at the certification step. A large dataset is considered with white vinho verde samples from the Minho region of Portugal. Wine quality is modeled under a regression approach, which preserves the order of the grades. Explanatory knowledge is given in terms of a sensitivity analysis, which measures the response changes when a given input variable is varied through its domain

PROBLEM STATEMENT

“Prediction of wine Quality using Data mining techniques”

Profound Question: Can we predict the quality of wine by applying a data mining model on the analytical dataset that we have from physicochemical tests of Vinho Verde wines?

Goal: The goal of this project is to derive rules to predict the quality of wines based on data mining algorithms.

Application: The predictions from this model can be used by manufacturers to improve the quality of wines, certification agencies for better understanding of aspects important for quality and lastly by customers to be make informed decisions while buying wines.

MOTIVATION

The wine industry is a multi-billion dollar industry that continues to grow each year. It is a global industry with major producers in Europe, Africa, Australia, and both of the American continents. A lot of time and money is put into making the best product to gain advantage on international market. The traditional method of determining the grade of a batch of wine is to hire a group of people with trusted distinguishing tastes to determine the quality of wine. This is expensive to do, especially with a lot of taste testers to get a good average. Personal bias can also affect the taster's grade (e.g. the relationship with the vineyard, the known location of the vineyard, etc.), which is undesirable. Wine already goes through

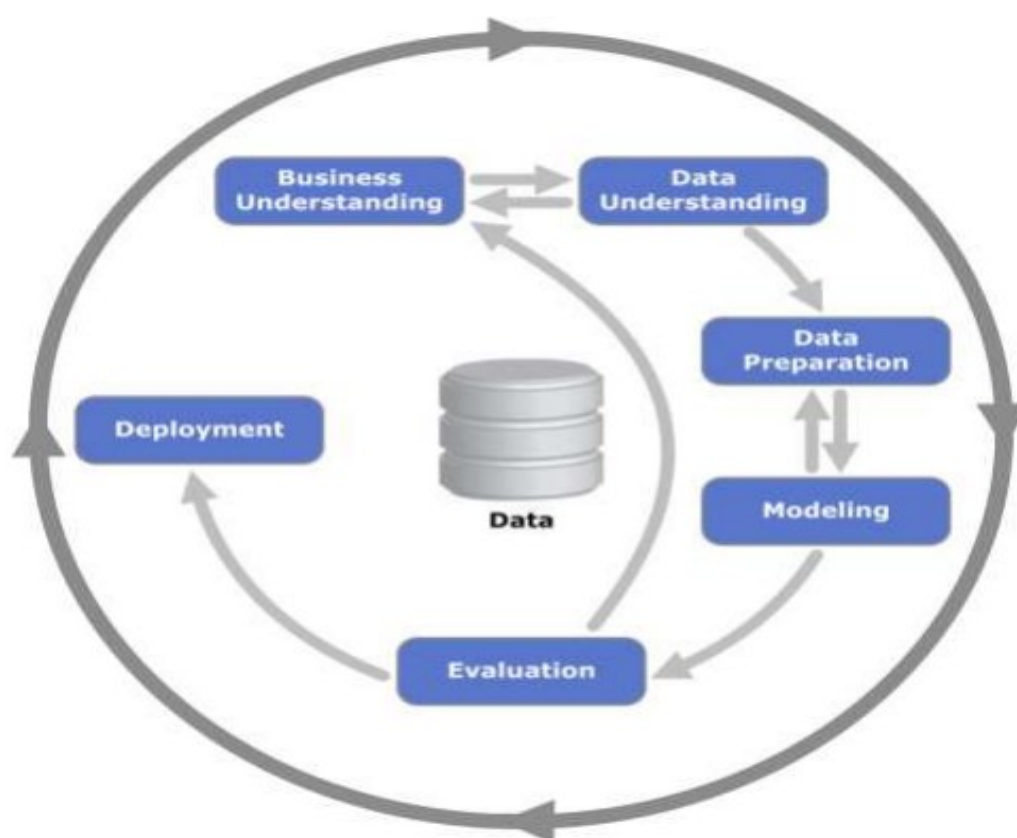
physicochemical tests to determine the safety of the product, to make sure that the levels of chemicals in each batch are normal. Using only the amount of chemicals in the sample, it is desirable to try to use a model to generate a wine grade for the sample. This saves time and money that would normally be used gathering and paying taste testers. It also eliminates any human bias that might come up, and requires a much smaller quantity of wine to test, which saves more of the product to sell. Once a proper wine grade is determined, branding can be done to maximize product. Wine that is determined to be of a high quality can be branded to a specific brand to build brand recognition and brand loyalty, while the wine of a lower quality can still be sold under a different label to still make a profit without tarnishing brand image. Physicochemical properties have traditionally not been used to attempt to grade wine because there has never been a database large enough. Recently computing power and space have allowed people to gather and store the data that might allow for classification of wine using the physicochemical properties. This same method, if successful, can be applied to other specific industries within the greater food industry. The development labs of food companies will be able to use the classification method for a preliminary indicator of how well a new product will be received, and can be used to greatly reduce the cost of launching a doomed product.

INTRODUCTION

Wine which was once viewed as a luxury product is increasingly enjoyed by a wider variety of customers today. Wine certification and assessment are essential elements in the wine industry in Portugal that prevent adulteration and are important for quality assurance. Wine certification includes physiochemical tests like determination of density, pH, alcohol quantity, fixed and volatile acidity etc. We have a large datasets having the physiochemical tests results and quality on the scale of 1 to 10 of wines, We are presenting a case study for modeling taste preferences based on the analytical dataset. Such a model can be used not only by the certification bodies but also by the wine producers to improve quality based on the physiochemical properties and by the consumers to predict the quality of wines This study can also be used by the several wine magazines to develop a guide for their readers to choose the best wines. Wine certification is often assessed by physicochemical and sensory tests [9]. Physicochemical laboratory tests routinely used to characterize wine include determination of density, alcohol or pH values, while sensory tests rely mainly on human experts. It should be stressed that taste is the least understood of the human senses , thus wine classification is a difficult task. Moreover, the relationships between the physicochemical and sensory analysis

are complex and still not fully understood. On the other hand, advances in information technologies have made it possible to collect, store and process massive, often highly complex datasets. All this data hold valuable information such as trends and patterns, which can be used to improve decision making and optimize chances of success . Data mining (DM) techniques aim at extracting high-level knowledge from raw data. There are several DM algorithms, each one with its own advantages. When modeling continuous data, the linear/multiple regression (MR) is the classic approach. Neural networks (NNs) have become increasingly used since the introduction of the backpropagation algorithm . More recently, support vector machines (SVMs) have also been proposed . Due to their higher flexibility and nonlinear learning capabilities, both NNs and SVMs are gaining an attention within the DM field, often attaining high predictive performances . SVMs present theoretical advantages over NNs, such as the absence of local minima in the learning phase. When applying these methods, performance highly depends on a correct variable and model selection, since simple models may fail in mapping

MODEL DESCRIPTION



1. Business understanding

This phase involves clearly defining the project objectives and goals, and translating these goals into a problem statement.

2. Data understanding

This phase involves collection of data and performing a preliminary analysis on the data to evaluate the data quality. Data understanding phase may also contain making subsets of data that may have any actionable patterns.

3. Data preparation

This phase is the most time taking one in the data mining process. It involves cleaning the data, performing certain transformations on the data to get the final dataset

4. Modeling

This phase involves selecting the appropriate modeling technique

5. Evaluation

The models generated during the modeling phase are evaluated for quality and also its determined whether the business objective is which means whether the problem statement is solved or not

6. Deployment

In this phase the effective models are finally put to use. It may be making a simple report or using the insights in the daily functioning of a company

Code:

Random Forest:

```
from sklearn.ensemble import RandomForestClassifier  
  
from sklearn.model_selection import train_test_split  
  
from sklearn.metrics import mean_squared_error  
  
from sklearn.metrics import confusion_matrix
```



```

import pandas as pd

df = pd.read_csv('dataset/winequality-white.csv', header=0,
sep=';')

X = df[list(df.columns)[:-1]]

y = df['quality']

X_train, X_test, y_train, y_test = train_test_split(X, y)

forest = RandomForestClassifier(n_estimators = 150)

forest.fit(X_train, y_train)

y_predict = forest.predict(X_test)

forest.score(X_test, y_test)

acc=forest.score(X_test, y_test)*100

print ('\nAccuracy ', acc )

print ('\nRMSE:', mean_squared_error(y_predict, y_test) **
0.5)

print ('\nConfusion Matrix
:\n',confusion_matrix(y_test,y_predict))

```

KNN:

```

from sklearn.neighbors import KNeighborsClassifier

from sklearn.model_selection import train_test_split

from sklearn.metrics import mean_squared_error

from sklearn.model_selection import GridSearchCV

from sklearn.metrics import confusion_matrix

```

```

import pandas as pd

df = pd.read_csv('dataset/winequality-red.csv', header=0,
sep=';')

X = df[list(df.columns)[:-1]]

y = df['quality']

X_train, X_test, y_train, y_test = train_test_split(X, y)

model = KNeighborsClassifier()

model.fit(X_train, y_train)

y_predict = model.predict(X_test)

print ('\nAccuracy:', model.score(X_test, y_test)*100)

print ('\nRMSE:', mean_squared_error(y_predict, y_test) **
0.5)

print ('\nConfusion Matrix :')

print (confusion_matrix(y_test,y_predict))

```

Logistic Regression:

```

from sklearn.linear_model import LogisticRegression

from sklearn.model_selection import train_test_split

from sklearn.metrics import mean_squared_error

from sklearn.metrics import confusion_matrix

import pandas as pd

df = pd.read_csv('dataset/winequality-red.csv', header=0,
sep=';')

```

```
X = df[list(df.columns)[: -1]]

y = df['quality']

X_train, X_test, y_train, y_test = train_test_split(X, y)

model = LogisticRegression()

model.fit(X_train, y_train)

y_predict = model.predict(X_test)

print ('\nAccuracy:', model.score(X_test, y_test)*100)

print ('\nRMSE:', mean_squared_error(y_predict, y_test) **
0.5)

print ('\nConfusion Matrix :')

print (confusion_matrix(y_test,y_predict))
```

Quality Histogram:

```
import pandas as pd

import matplotlib.pyplot as plt


df = pd.read_csv('dataset/winequality-white.csv', header=0,
sep=';')


df['quality'].plot(kind='hist', bins=7, color='black')


plt.xlabel('Quality')

plt.ylabel('Count')

plt.title('White wines')

plt.axis([2, 10, 0, 2500])

plt.show()
```

```
df = pd.read_csv('dataset/winequality-red.csv', header=0,  
sep=';')
```

```
df['quality'].plot(kind='hist', bins=6, color='red')
```

```
plt.xlabel('Quality')
```

```
plt.ylabel('Count')
```

```
plt.title('Red wines')
```

```
plt.axis([2, 9, 0, 800])
```

```
plt.show()
```

DATASET DESCRIPTION

Two datasets are included, related to red and white Vinho Verde wine samples, from the north of Portugal. The project application aims is to predict wine quality based on physicochemical tests using different chemical attributes and permutations.

1. FIXED ACIDITY:

Acidity is a fundamental property of wine, imparting sourness and resistance to microbial infection. Fixed acidity is the no. of grams of tartaric acid per dm³

2. VOLATILE ACIDITY

Wine spoilage is legally defined by volatile acidity which is calculated as no. of grams of acetic acid per dm³ of wine

3. CITRIC ACID

It is the no. of grams of citric acid per dm³ of wine

4. RESIDUAL SUGAR

Residual sugar refers to the sugar remaining after fermentation stops. Given as no. of grams per dm³

5. CHLORIDES

It is the no. of grams of sodium chloride per dm³ of wine

6. FREE SULPHUR DI OXIDE

It is the no. of grams of free sulfites per dm³ of wine

7. TOTAL SULPHUR DI OXIDE

It is the no. of grams of total sulfite (free sulfite+ bound sulfite) in per dm³ of wine

8 .DENSITY

It gives the density of the wine in grams per cm³

9. PH

It gives the pH of the wines. pH is used to measure ripeness in relation to acidity

10. SULPHATES

It gives the no. of grams of potassium sulphate per dm³ of wine

11. ALCOHOL

It gives the volume of alcohol in percentage

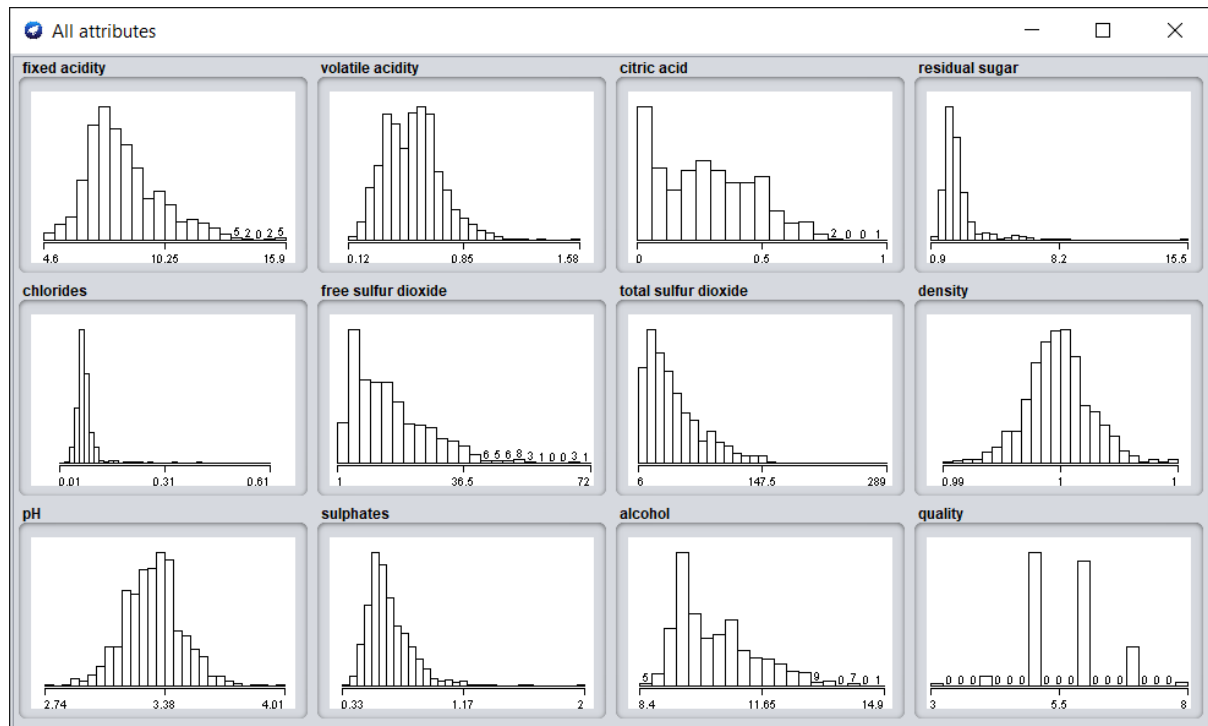
12. QUALITY

This is the target variable. Here the wine is rated from 1- 10 based on the quality

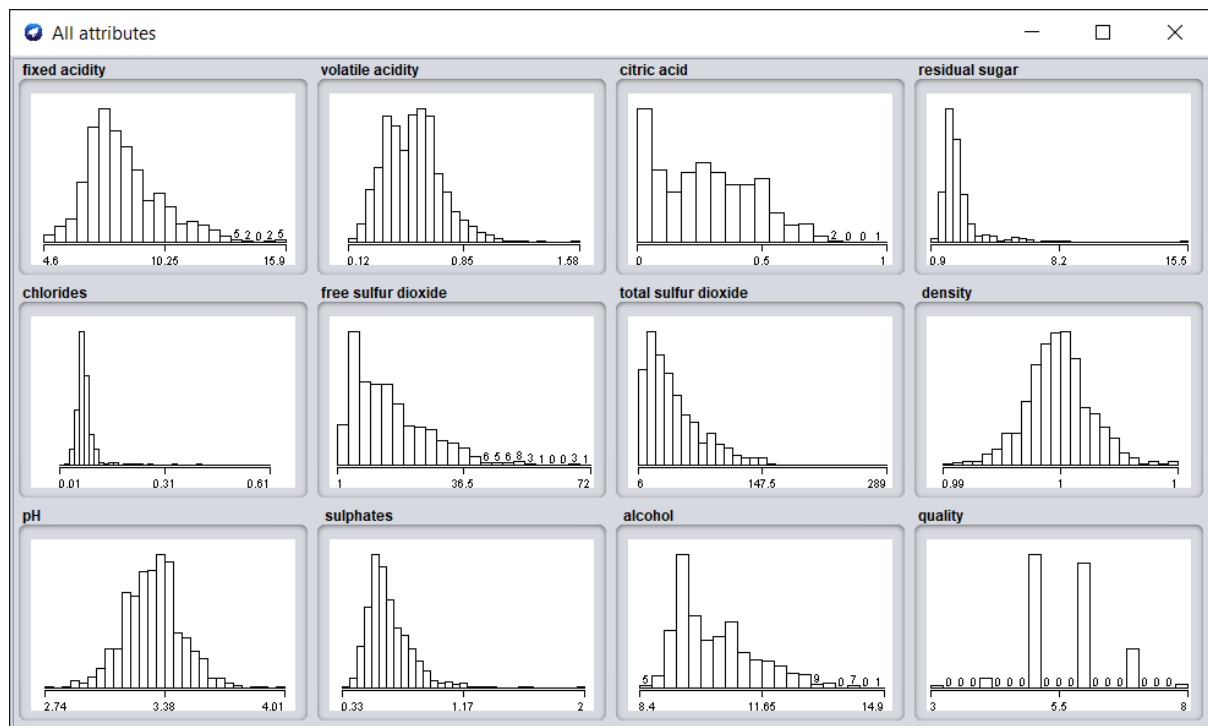
No. of Instances: 4898

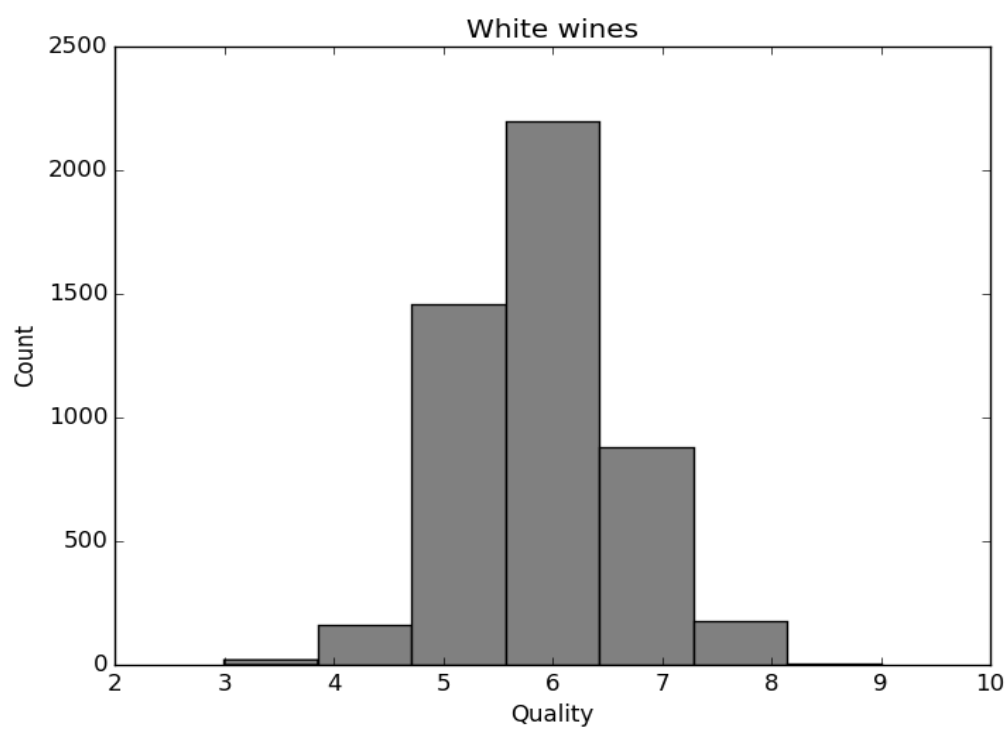
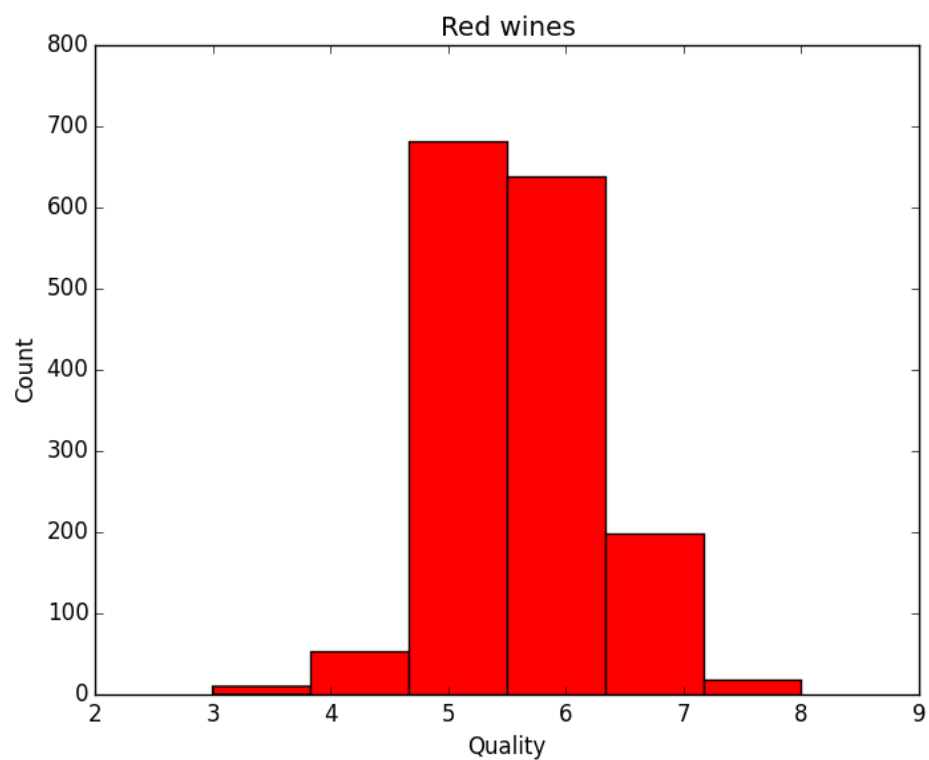
Dataset link: <https://archive.ics.uci.edu/ml/datasets/wine+quality>

White Wine



Red Wine





RESULTS

Random Forest:

```
Windows PowerShell
PS V:\vit\SEM-6\DM\project> python .\forest.py

Accuracy 66.61224489795919

RMSE: 0.6700593942604899

Confusion Matrix :
[[ 0  0  3  2  0  0  0]
 [ 0  5 22 13  0  0  0]
 [ 0  2 266 97  3  0  0]
 [ 0  0  95 411 40  0  0]
 [ 0  0  5  96 120  2  0]
 [ 0  0  0  15  13 14  0]
 [ 0  0  0  1  0  0  0]]
PS V:\vit\SEM-6\DM\project>
```

KNN:

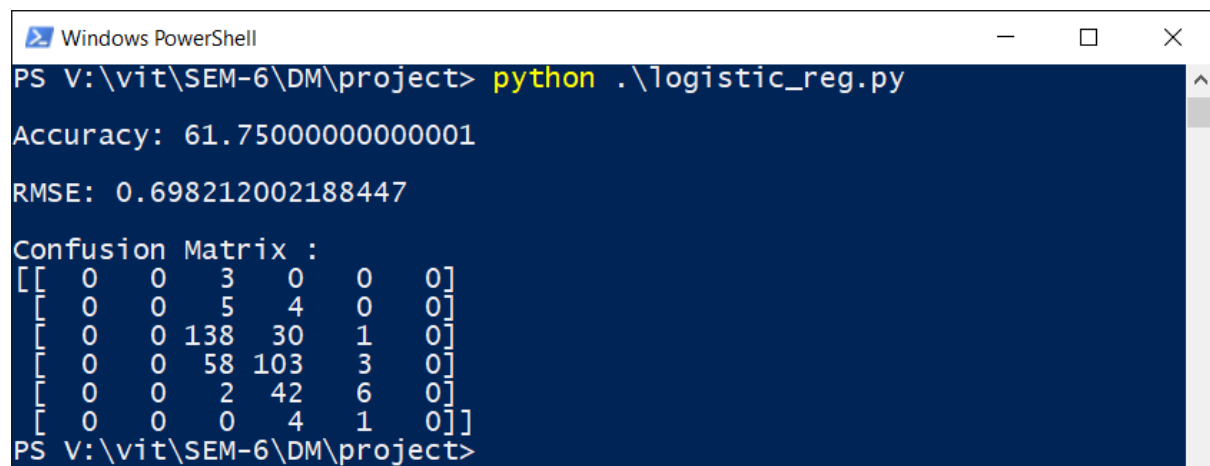
```
Windows PowerShell
PS V:\vit\SEM-6\DM\project> python .\knn.py

Accuracy: 46.75

RMSE: 0.9300537618869137

Confusion Matrix :
[[ 0  0  1  0  0  0]
 [ 0  1 11  4  2  0]
 [ 0  4 109 40  5  0]
 [ 0  1  89 68 10  0]
 [ 0  1  16 22  9  1]
 [ 0  0  2  4  0  0]]
PS V:\vit\SEM-6\DM\project>
```


Logistic Regression:



```
Windows PowerShell
PS V:\vit\SEM-6\DM\project> python .\logistic_reg.py

Accuracy: 61.75000000000001

RMSE: 0.698212002188447

Confusion Matrix :
[[ 0  0  3  0  0  0]
 [ 0  0  5  4  0  0]
 [ 0  0 138 30  1  0]
 [ 0  0  58 103 3  0]
 [ 0  0  2  42 6  0]
 [ 0  0  0  4  1  0]]
PS V:\vit\SEM-6\DM\project>
```

EVALUATION

Data Preparation:

Handling Outliers: The dataset doesn't really have any outliers hence no action is required.

Handling missing value: The dataset again does not have missing values. However there were a couple of rows that were repeated at the end. These rows are eliminated.

We have used 3 algorithms, Random Forest, K-Nearest Neighbours, and Logistic Regression.

First we read the dataset and make dataframes, then divide it into X and y. X being a dataframe of all the attributes and y being the class label.

The data was randomly split into training and testing dataset using the function `train_test_split` from the SciKit learn library.

Now we fit the model with the respective algorithm using the `X_train` and `y_train`. Then we use the `predict` method of the SciKit learn to predict the class label on the test dataset.

Then we compare the `y_predict` (the predicted class labels) and the `y_test` (the class labels from the original dataset).

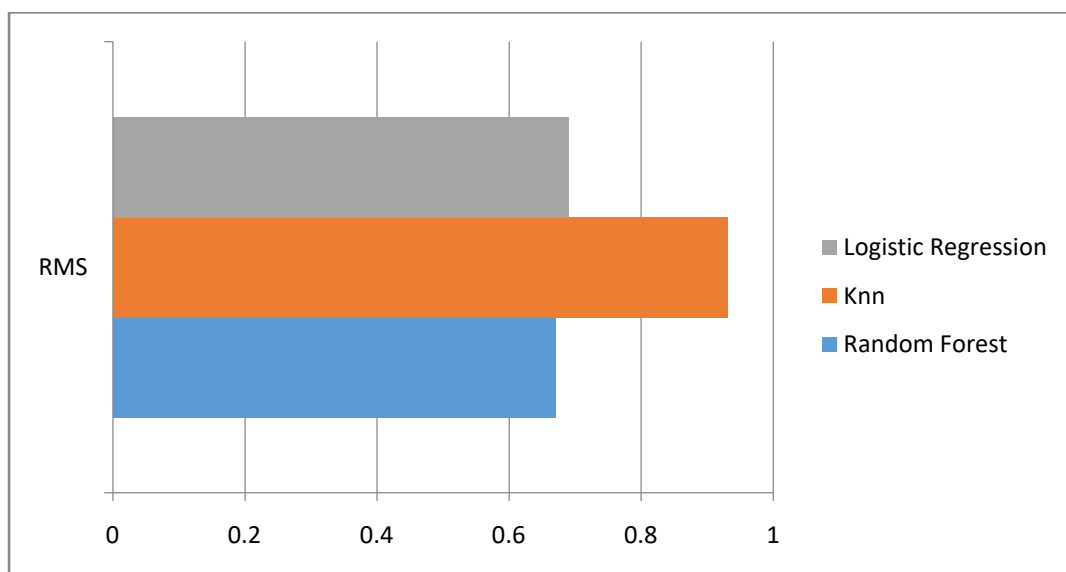
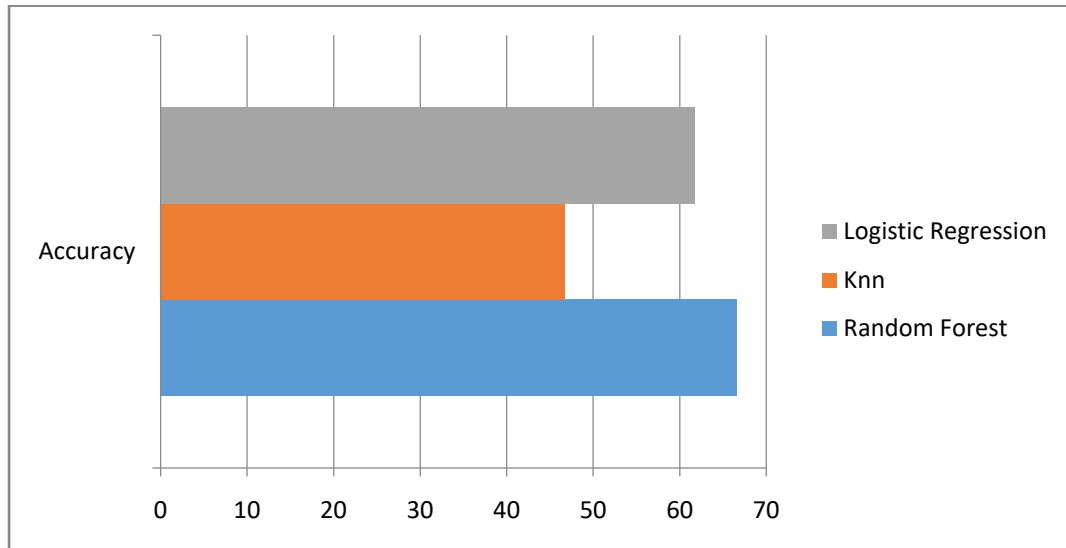
We then used the evaluation metrics from the library `sklearn.metrics` to find out the accuracy, Relative mean squared error and the confusion matrix.

On performing the above process for all the 3 algorithms it has been noted that,

As the data is being shuffled every time, while splitting there had been noted a slight difference in the scores.

Results

It can be stated that Random Forest has been giving the highest accuracy and the least RMS value.



CONCLUSION

As we can clearly see, the SciKit Learn library does a very good job in predicting the wine quality. As the dataset had multiple class labels, we thought its better to use the Random Forest Algorithm, as it had to be reduced into multiple binary classification problems. And we got the highest accuracy on Random Forest.