İstanbul
Bilgi University

by

Berk Özpınar

116200036

CMPE 407 Machine Learning

Red Wine Quality Classification

Istanbul Bilgi University

Spring 2021

I Introduction

Red wine quality was always a mystery for wine lovers in terms of its components. In this project the quality of red wine has been investigated depending on its eleven different components, they are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxides, total sulfur dioxides, density, pH, sulphates and alcohol. A dataset has been used related to the variants of Portuguese Vinho Verde wine. Using classification techniques a wine is considered good quality if its quality score is equal or higher than 7 otherwise it is considered normal.

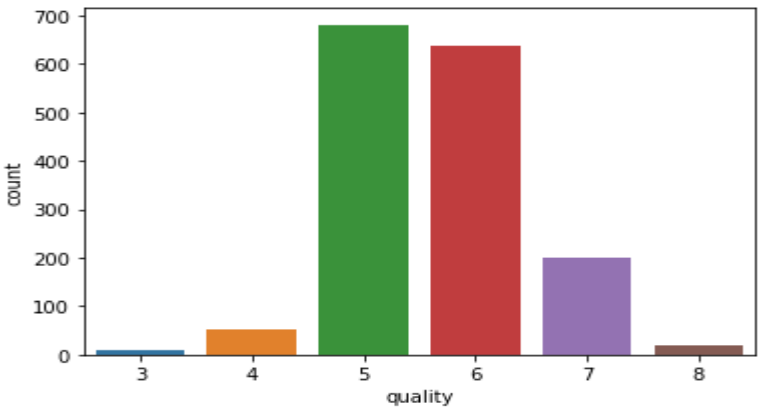
II Dataset

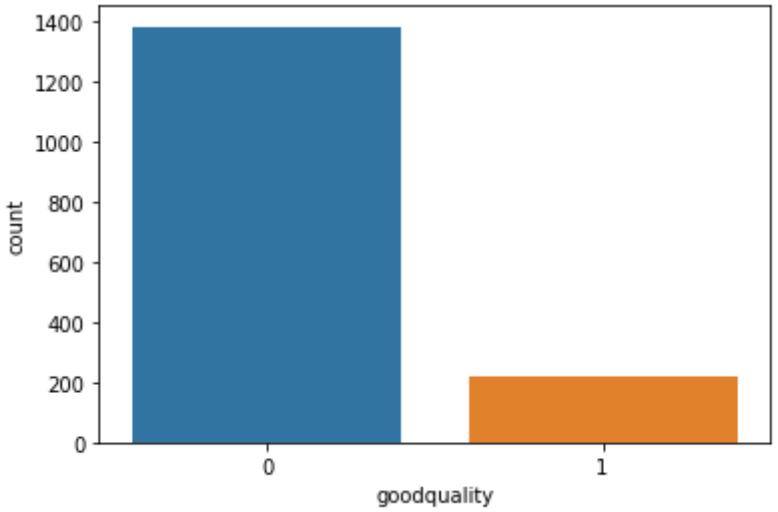
The dataset consists of 1599 different wine variants with different qualities. Quality score is between 3 and 8. The mean quality is 5.6.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

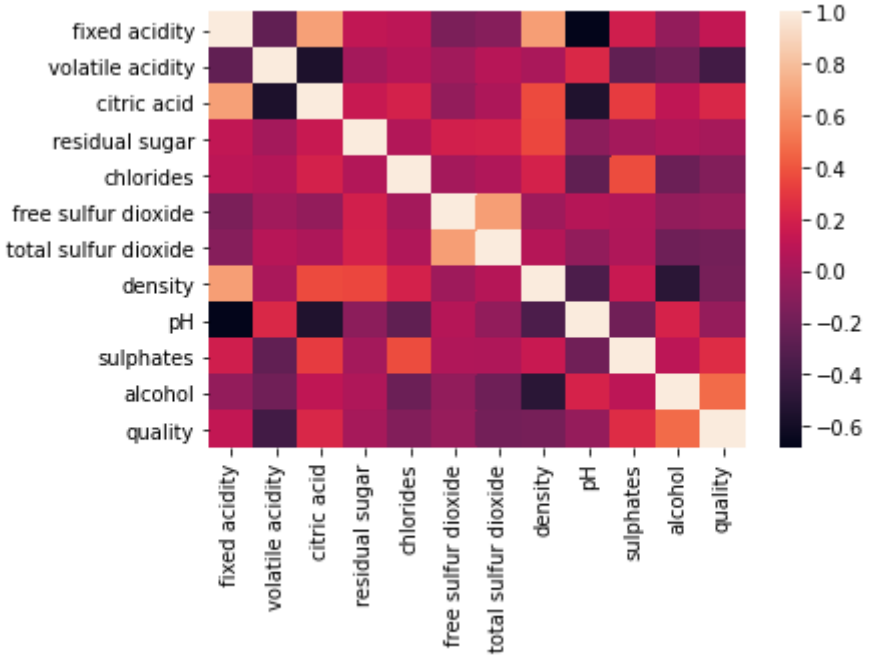
Description of whole dataset

The dataset is not balanced, because of that the majority of wines are between quality 5 and 6. This makes the majority of wines normal quality.

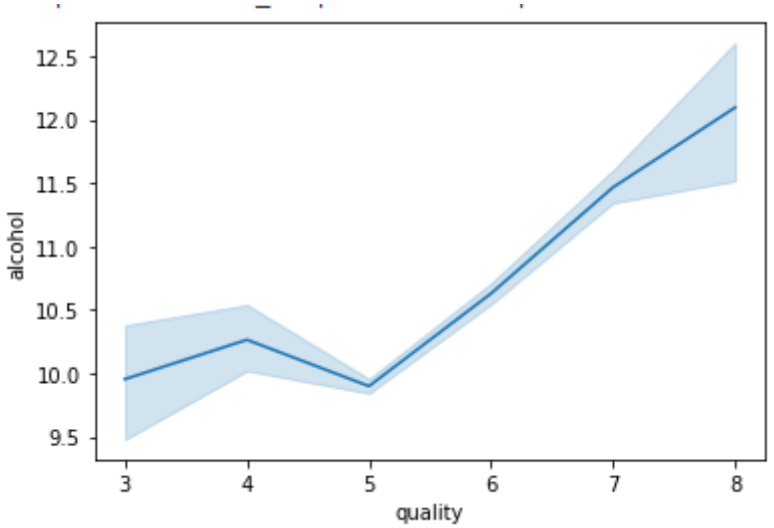




Above there is the histogram for the good quality wine counts, according to the dataset only around 12% of wines are in good quality.



To find the correlation between quality and the components heatmap has been used, according to the heatmap quality mostly depends on alcohol rate. Below there is a graph showing the correlation between alcohol rate and the quality.



The line graph shows that there is a positive correlation between alcohol rate and the quality of wine, higher the alcohol rate, higher the quality.

III Preprocessing

Three steps of preprocessing have been applied to the dataset, first there is no good quality column in the original dataset, so to convert the problem into a binary classification problem wines with quality higher than or equal to 7 is considered good quality and others normal quality. After that data is splitted into features and target label, 11 components have been used as features and the good quality column has been used for target label. Then to measure the accuracy of the models the data has been splitted as training set and test set, 30% of total data has been used for testing the accuracy, the random state seed has been hardcoded to 42.

```
features = ['fixed acidity', 'volatile acidity','citric acid', 'residual sugar', 'chlorides',
            'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol']

label = ['goodquality']
#Choosing necessary features and target label.

#Assigning values
X = data[features].values
Y = data[label].values
```

```
from sklearn.model_selection import train_test_split
#Splitting train and test data.
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state = 42)
```

Codes showing preprocessing steps

IV Experiments

i. Logistic Regression

As the first experiment logistic regression has been used for binary classification of the wines. The algorithm is also called sigmoid function, it takes an input and converts it into a value between 0 and 1 and classifies the input to the closest number. This is the simplest model but it has good accuracy rates in most cases so it was the first experiment.

```
#Applying Logistic Regression
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score
from sklearn.model_selection import cross_val_score
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
prediction_logreg = log_reg.predict(X_test)
cm_logreg = confusion_matrix(y_test, prediction_logreg)
accuracy_logreg = accuracy_score(y_test, prediction_logreg)
lr_cross_val = cross_val_score(log_reg, X_train, y_train, cv=5)
```

Code showing logistic regression steps

Sci-kit learn library has been used for calculations, train and test data are used for training and testing the model. And then to measure the results, confusion matrix, accuracy score and cross validation techniques have been used which will be investigated later in the next section.

ii. K-nearest neighbor

As the second experiment K-nearest neighbor algorithm has been used to classify the wines. It is an algorithm depending on checking the closest K points to related point and classifying it to the class which has (K/2)+1 occurrences. In this project as K, number 5 has been used.

```
#Applying KNN
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
prediction_knn = knn.predict(X_test)
cm_knn = confusion_matrix(y_test, prediction_knn)
accuracy_knn = accuracy_score(y_test, prediction_knn)
knn_cross_val = cross_val_score(knn, X_train, y_train, cv=5)
```

Code showing KNN steps

Same steps have been followed with logistic regression, in this experiment KNeighborsClassifier has been used for calculations.

iii. Decision Tree

As the last experiment decision tree classifier has been used, which is an algorithm similar to a flow-chart structure, following a bunch of steps it classifies the input value into one of the target labels. This algorithm is the most improved amongst others and it is faster than KNN algorithm in execution time.

```
#Applying Decision Tree
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier()
dtc.fit(X_train, y_train)
prediction_dtc = dtc.predict(X_test)
cm_dtc = confusion_matrix(y_test, prediction_dtc)
accuracy_dtc = accuracy_score(y_test, prediction_dtc)
dtc_cross_val = cross_val_score(dtc, X_train, y_train, cv=5)
```

Code showing KNN steps

As seen above it has almost the same structure with previous experiments, this time DecisionTreeClassifier has been used for calculations.

V Results

The general results can be considered good depending on the results of the experiments, it has an average of 86% accuracy. The results have been checked with cross-validation technique to be sure their performances are consistent.

Model	Accuracy	Confusion Matrix
Logistic Regression	86.6%	401 12 51 15
K-nearest Neighbor	85.8%	399 14 54 13
Decision Tree	87.0%	372 41 21 46

As seen in the results table above, the decision tree has the highest accuracy amongst the others but the most important part if we take a look at the confusion matrices of the models logistic regression and K-nearest neighbor algorithms although they have good accuracies they were pretty unsuccessful while classifying high quality wines. While the logistic regression has 22% and KNN has 19% accuracies classifying the good wines, the decision tree has ~69% accuracy which is quite an improvement compared to the other results.

VI Conclusion

As a result, the decision tree is the best algorithm suitable for this classification problem due to its performance and accuracy in classifying the good quality wines. As an improvement in the further studies the good quality wines can be compared with their prices to check if the expensive wines are really in good quality.

References

1. *P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.*