Wine Quality Prediction

09.12.2022

MACHINE LEARNING PROJECT

—

GROUP NAME:

SIDDHARTH AGRAWAL 20C21001
BHARGAV CHAUDHARY 20C21010
JAINAM SHAH 21C21510

BRIEF DESCRIPTION OF THE TOPIC

Nowadays, industries are using product quality certifications to promote their products. This is a time taking process and requires the assessment given by human experts which makes this process very expensive. This paper explores the usage of machine learning techniques such as Random forest algorithm for product quality.

Nowadays people try to lead a luxurious life. They tend to use the things either for show off or for their daily basis. These days the consumption of red wine is very common to all. So it became important to analyze the quality of red wine before its consumption to preserve human health.

Hence this research is a step towards the quality prediction of the wine using its various attributes. Dataset is taken from the sources and the techniques such as Random Forest, Support Vector Machine ETC are applied. Various measures are calculated and the results are compared among the training set and testing set and accordingly the best out of the three techniques depending on the training set results is predicted.

Better results can be observed if the best features from other techniques are extracted and merged with one another to improve the accuracy and efficiency.

NEED FOR THE TITLE:

Wine is the most commonly used beverage globally, and its values are considered important in society. Quality of the wine is always important for its consumers, and mainly for producers in the present competitive market to raise the revenue.

Historically, wine quality used to be determined by testing at the end of the production; to reach that level, one already spends lots of time and money. If the quality is not good, then the various procedures need to be implemented from the beginning, which is very costly. Every person has their own opinion about the taste, so identifying a quality based on a person's taste is challenging.

With the development of technology, the manufacturers started to rely on various devices for testing in development phases. So, they can have a better idea about wine quality, which, of course, saves lots of money and time. In addition, this helped in accumulating lots of data with various parameters such as quantity of different

chemicals and temperature used during the production, and the quality of the wine produced.

These data are available in various databases (UCL Machine Learning Repository, and Kaggle). With the rise of ML techniques and their success in the past decade, there have been various efforts in determining wine quality by using the available data [. During this process, one can tune the parameters that directly control the wine quality. This gives the manufacturer a better idea to tune the wine quality by tuning different parameters in the development process. Besides, this may result in wines with multiple tastes, and at last, may result in a new brand. Hence, the analysis of the basic parameters that determine the wine quality is essential.

In addition to humanitarian efforts, ML can be an alternative to identify the most important parameters that control the wine quality. In this work, we have shown how ML can be used to identify the best parameter on which the wine quality depends and in turn predict wine quality.

According to experts, the wine is differentiated according to its smell, flavor, and color, but we are not a wine expert to say that wine is good or bad. What will we do then? Here's the use of Machine Learning comes,

ALGORITHM USED:

In wine quality prediction we have used a supervised learning algorithm.

Supervised learning

Supervised learning is the type of machine learning in which machines are trained using well "labeled" training data, and on the basis of that data, machines predict the output. The labeled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y)

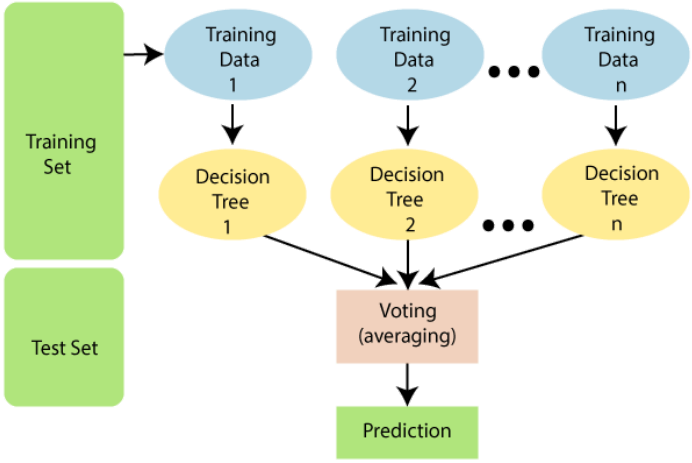
In supervised learning there is a wide range of machine learning algorithms such as linear regression, logistic regression, support vector machine, and kernel methods, neural networks, and many others are available for the learning process. Each technique has its strengths and weaknesses. In this work, we use the following supervised learning algorithms to predict wine quality.

Random forest algorithm:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "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." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



DATASET USED:

NAME: **winequalityN.csv**

SOURCE: KAGGLE (https://www.kaggle.com/datasets/rajyellow46/wine-quality)

If you download the dataset, you can see that several features will be used to classify the quality of wine, many of them are chemical, so we need to have a basic understanding of such chemicals.

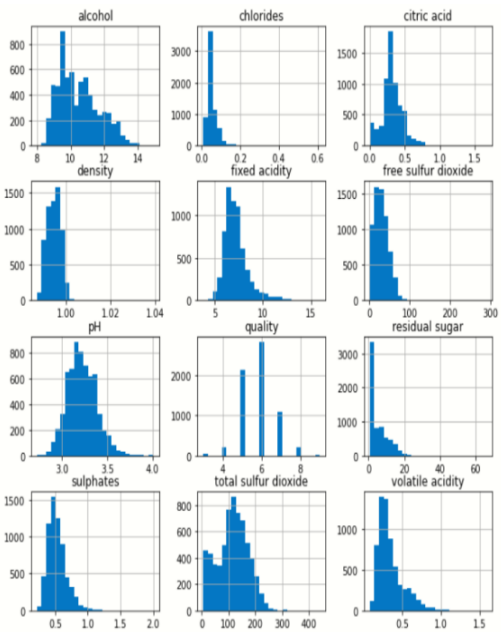
- volatile acidity : Volatile acidity *is the* gaseous acids present in wine.
- fixed acidity : Primary fixed acids found in wine are tartaric, succinic, citric, and malic
- residual sugar : Amount of sugar left after fermentation.
- citric acid : It is weak organic acid, found in citrus fruits naturally.
- chlorides : Amount of salt present in wine.
- free sulfur dioxide : So2 is used for prevention of wine by oxidation and microbial spoilage.
- total sulfur dioxide
- pH : In wine pH is used for checking acidity
- density
- sulfates : Added sulfites preserve freshness and protect wine from oxidation, and bacteria.
- alcohol : Percent of alcohol present in wine.

DATA VISUALIZATION:

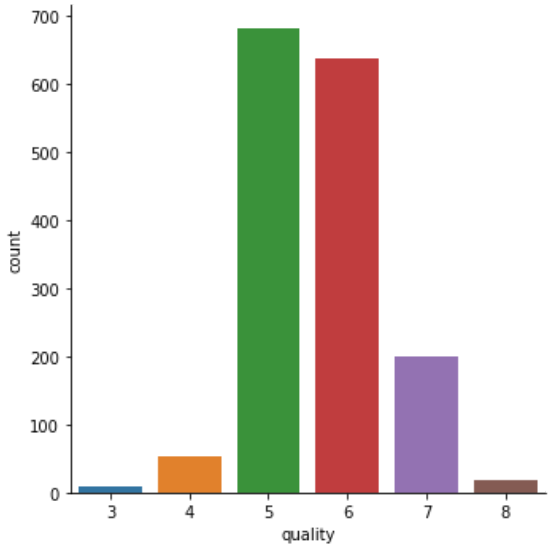
We know that the “image speaks everything” here the visualization came into the work, we use visualization for explaining the data. In other words, we can say that it is a graphic representation of data that is used to find useful information.

```
df.hist(bins=25,figsize=(10,10))
# display histogram
plt.show()
```

output:-



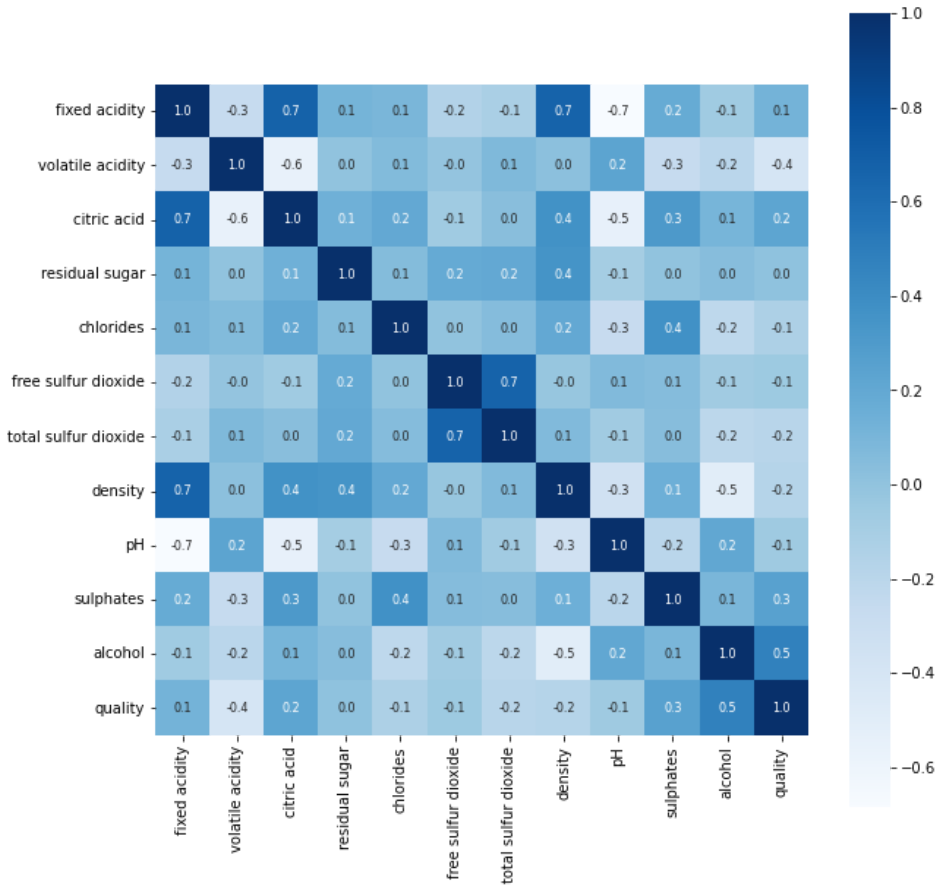
QUALITY BAR GRAPH



```
# constructing a heatmap to understand the correlation between the columns

plt.figure(figsize=(10,10))

sns.heatmap(correlation, cbar=True, square=True, fmt = '.1f', annot =
True, annot_kws={'size':8}, cmap = 'Blues')
```



DATA SPLITTING :

Main aim of the data splitting is to split the data set in test and train data.On train data the our machine learning model is trained and on the test data our model is tested for the result.

This data splitting can be done by sklearn library in python

```
from sklearn.model_selection import train_test_split
```

```
# separate the data and Label

X = wine_dataset.drop('quality',axis=1)

Y = wine_dataset['quality'].apply(lambda y_value: 2 if y_value>=7 else(1 if y_value == 6 else 0))
```

Train & Test Split

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=3)
```

PREDICTIVE SYSTEM :

```
input_data = (7.9,0.32,0.51,1.8,0.341,17.0,56.0,0.9969,3.04,1.08,9.2)

# changing the input data to a numpy array

input_data_as_numpy_array = np.asarray(input_data)

# reshape the data as we are predicting the label for only one instance

input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)

print(prediction)

if (prediction[0]==2):

  print('Good Quality Wine')

elif(prediction[0]==1):

  print('average quality wine')

else:

  print('Bad Quality Wine')
```

OUTPUT:

```
[1]
average quality wine
/usr/local/lib/python3.8
  warnings.warn(
```