

Red and White Wine Analysis

DAB - Data Analytics Boys

Atharva S Gadad	PES1UG20CS088
Aaditya Vikram	PES1UG20CS528
Dhruv Jyoti Garodia	PES1UG20CS527
Alan S Paul	PES1UG20CS624

INTRODUCTION

Wine is an alcoholic drink that is made up of fermented grapes. Depending on the variety of grape used wine is available in two types: Red wine and White wine.

White wine is primarily made with white grapes, and the skins are separated from the juice before the fermentation process. Red wine is made with darker red or black grapes, and the skins remain on the grapes during the fermentation process.

Studies have shown that moderate wine consumption may be good for your heart and circulatory system. There is also evidence that drinking wine in moderation may lower your risk of type 2 diabetes.

Nutritional value of wine

Red wine

Calories:
125 per glass

Carbohydrates:
4 grams per glass

Alcohol content:
3.1 grams per glass

White wine

Calories:
115 per glass

Carbohydrates:
5 grams per glass

Alcohol content:
2.9 grams per glass



DATASET USED

Link to download dataset:

<https://www.kaggle.com/code/danielpanizzo/red-and-white-wine-quality/data>

Description:

The dataset used contains data about various chemical properties of wine, quality of wine and the type of wine(Red or white). The quality of wine was evaluated by three experts who provided score between 0(Bad) and 10(Excellent) for each wine.



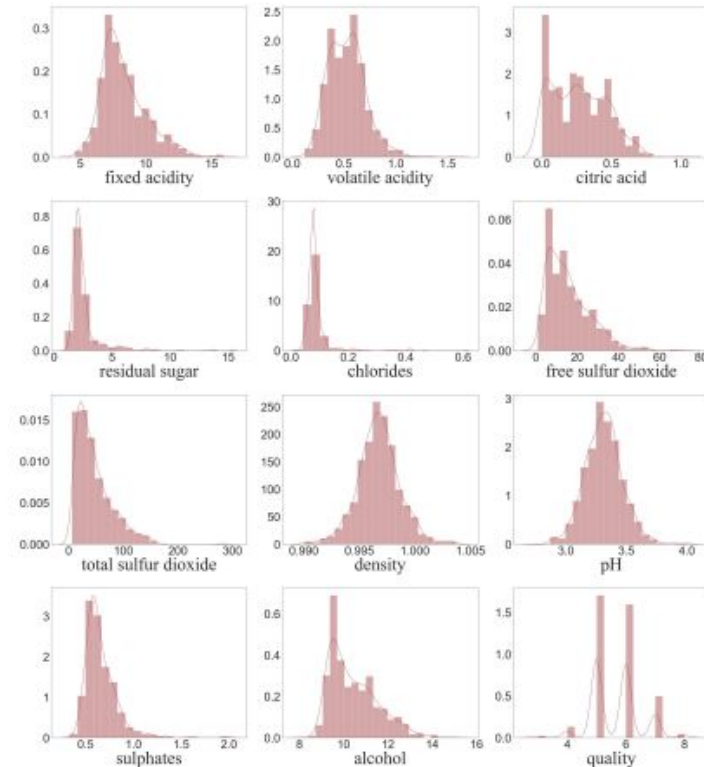
FEATURES PRESENT IN THE DATASET AND ANALYSIS

->During the process of fermentation some amount of sugar gets converted to alcohol. The remaining sugar content is residual sugar. Most of the residual sugar content is less than 4 g/l indicating that most of the alcohol in the dataset is dry in nature.

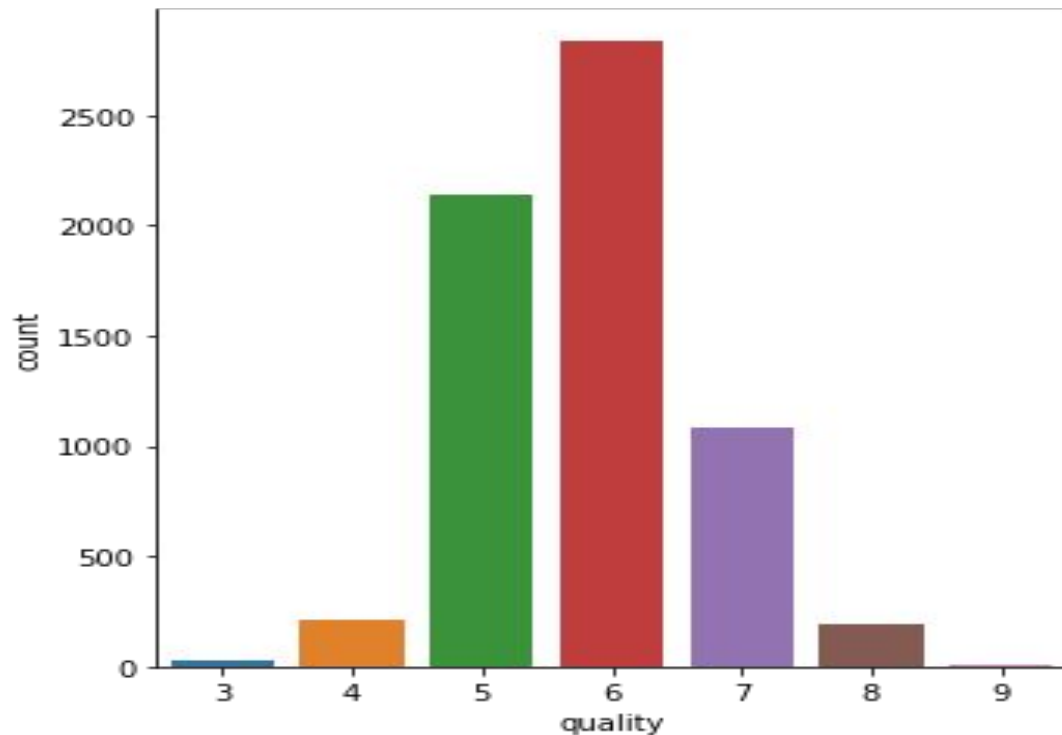
->Acids in wine are off two types - volatile acidity and fixed acidity. Volatile acidity influences the smell of wine and fixed acidity influences the taste.

->Citric acid is mainly used in cheap quality wines.

->PH is used for indicating acidity which in turn is a good indicator of taste as well



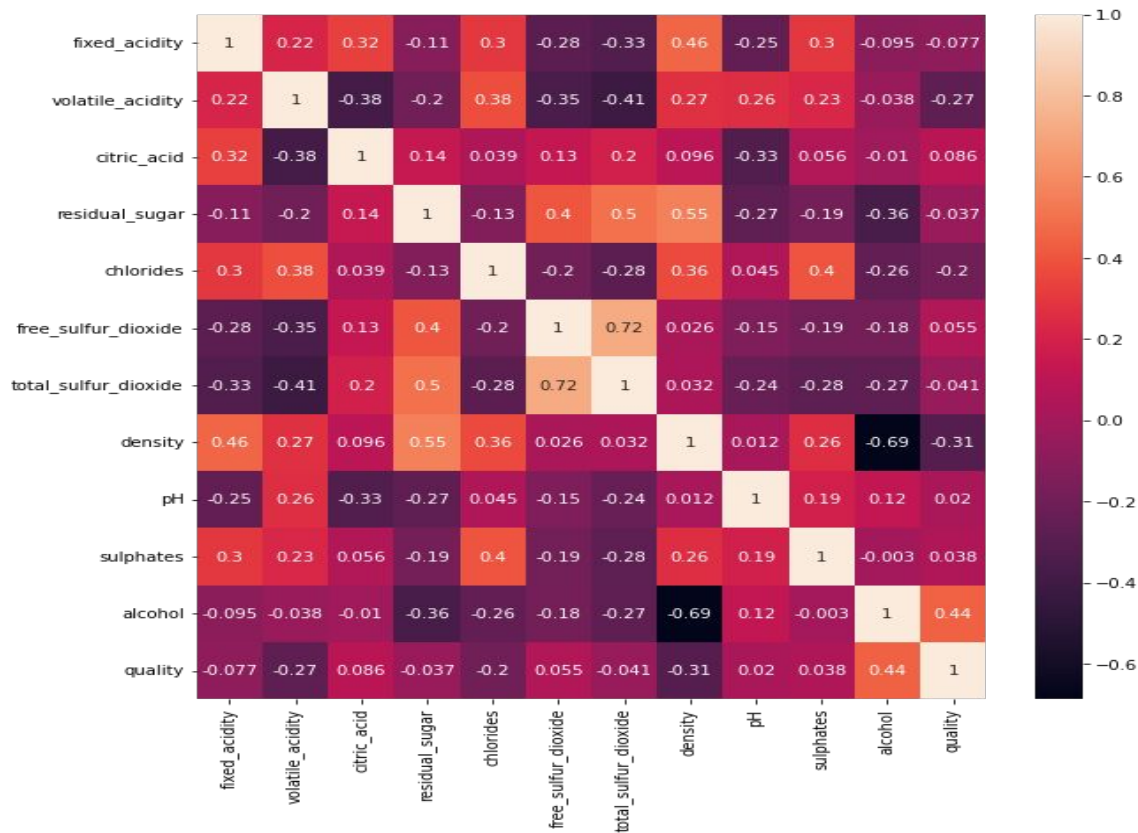
EDA ON WINE DATASET



Bar plot for the **Quality** column

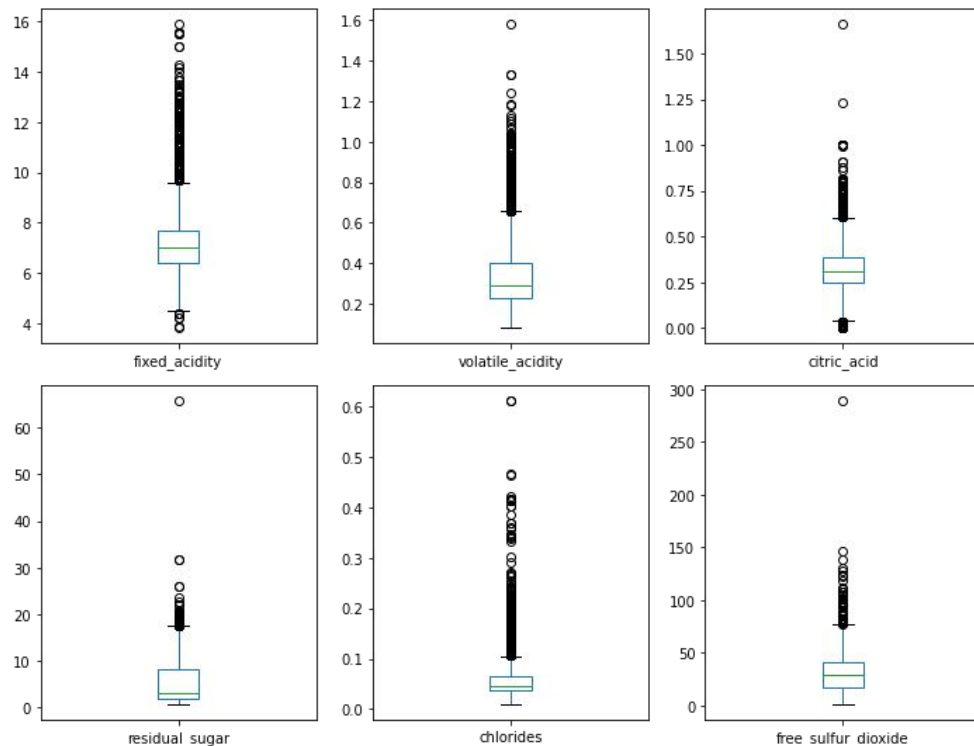
We find that an average wine has a quality metric of **6**

EDA ON WINE DATASET



Correlogram of wine dataset representing correlation between different features in the dataset

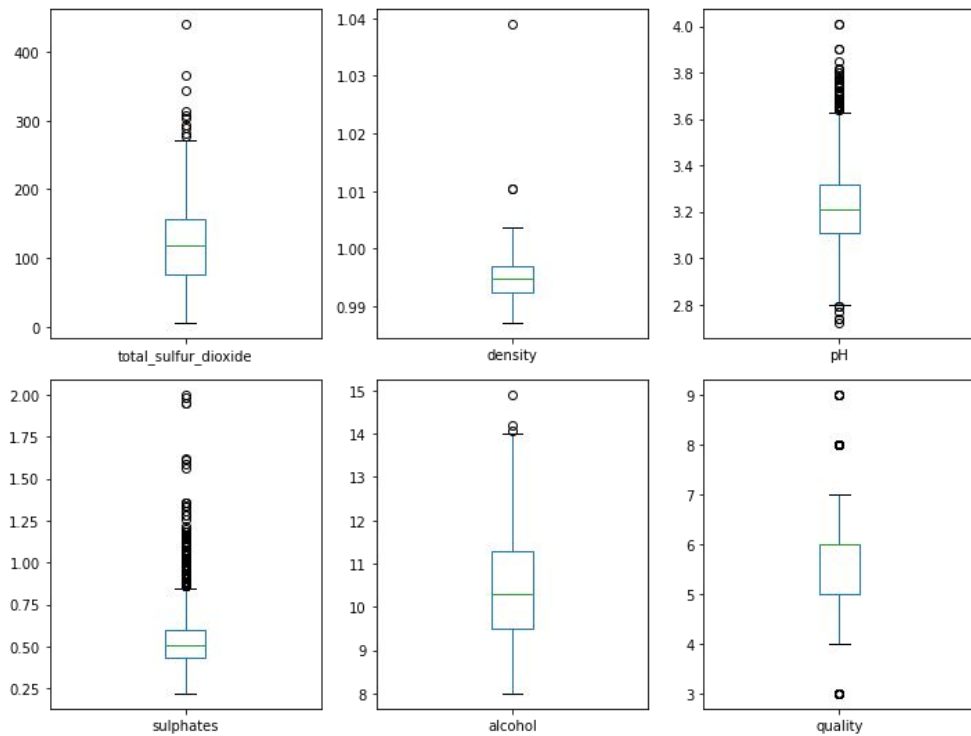
EDA ON WINE DATASET



Boxplot for all attributes of wine dataset.

We can observe that other than density and alcohol attributes “remaining attributes have multiple outliers.

EDA ON WINE DATASET



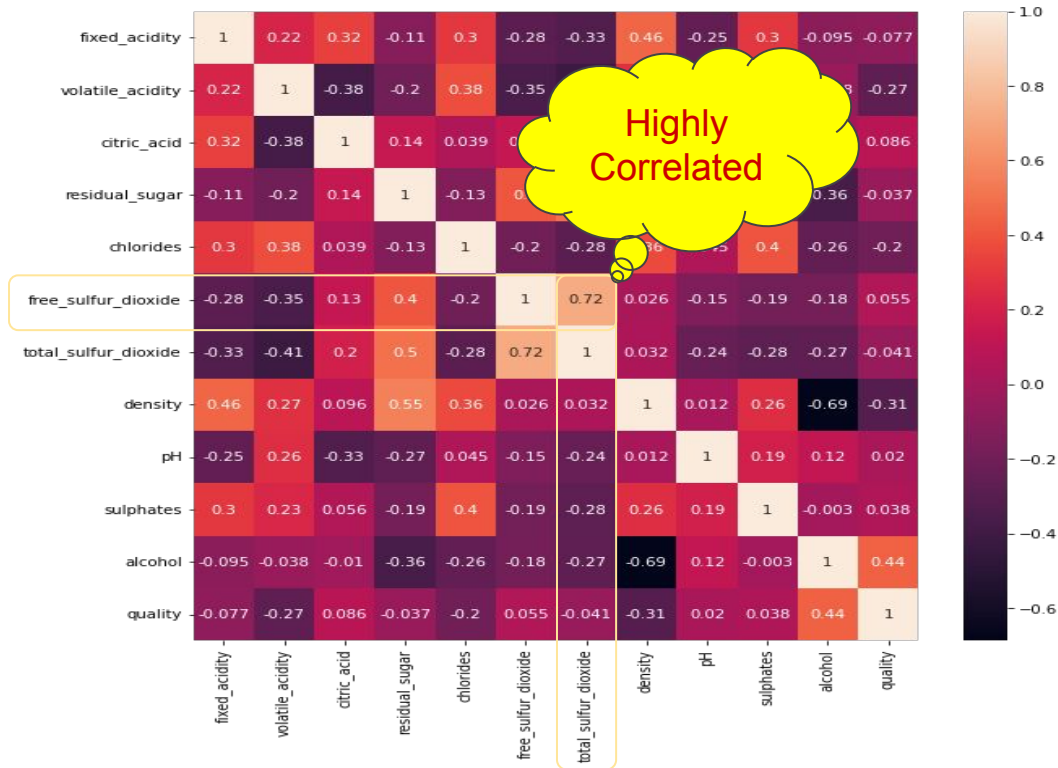
Feature Selection

To analyze data we will reduce features to be used for analysis or else analyzing data will become computationally expensive and also the model might not perform well due to unnecessary variables which are not relevant to the target variable.

- 1) Correlation
- 2) Variance Threshold Based Selection
- 3) Mean Absolute Difference (MAD)
- 4) Information Gain
- 5) Random Forest based Selection
- 6) Backward Feature Selection
- 7) Forward Feature Elimination



(1) Inference from Correlogram Plot



From the plot we see that total and free sulphur dioxide are highly correlated, so in our model prediction we are just using **Total Sulphur Dioxide**.

Similarly, for Alcohol and Density, we choose Density.

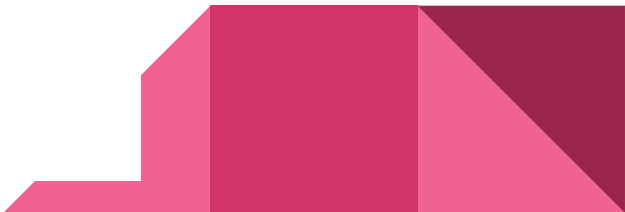
(2) Variance based feature Selection

Variance Threshold:

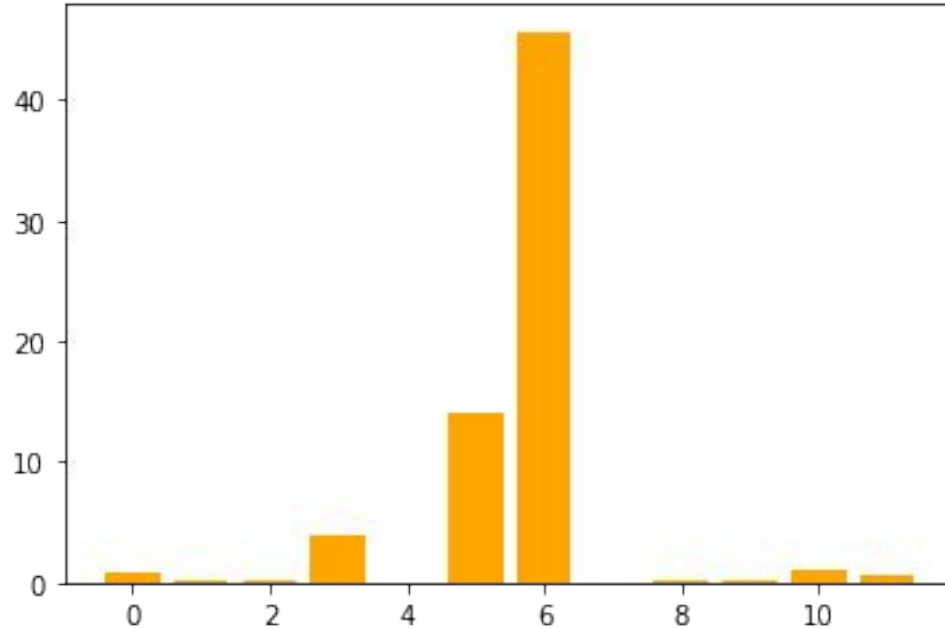
Variance Threshold is a feature selector that removes all the low variance features from the dataset that are of no great use in modeling.

It looks only at the features (x), not the desired outputs (y), and can thus be used for unsupervised learning.

Default Value of Threshold is 0

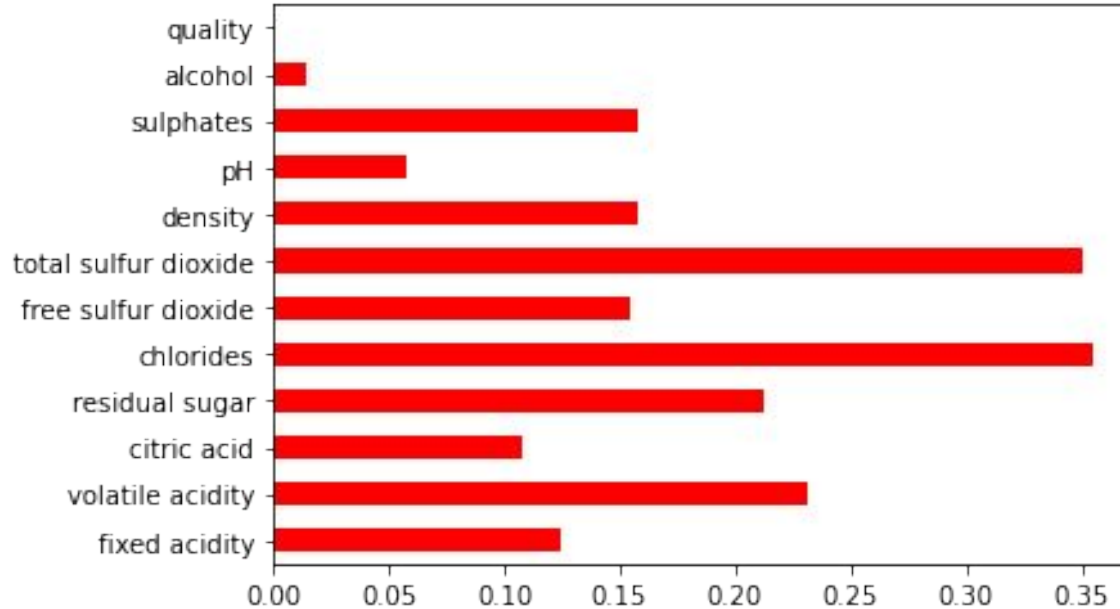
- If Variance Threshold = 0 (Remove Constant Features)
 - If Variance Threshold > 0 (Remove Quasi-Constant Features)
- 

(3) Mean Absolute Difference (MAD)



'fixed acidity',
'volatile acidity',
'citric acid',
'residual sugar',
'Chlorides',
'free sulfur dioxide',
'total sulfur dioxide',
'Density',
'pH',
'Sulphates',
'Alcohol',
'quality'

(4) Information Gain



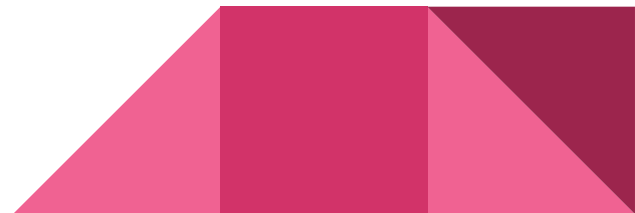
'Chlorides',
'total sulfur dioxide',
'volatile acidity',
'residual sugar',
'Sulphates',
'Density',
'free sulfur dioxide',
'fixed acidity',
'citric acid',
'pH',
'Alcohol'

(5) Random Forest based Selection

Importances	Features
0.047959	fixed acidity
0.127354	volatile acidity
0.015404	citric acid
0.043000	residual sugar
0.282498	chlorides
0.057292	free sulfur dioxide
0.285175	total sulfur dioxide
0.052422	density
0.022876	pH
0.053445	sulphates
0.010291	alcohol
0.002283	quality

Random forests consist of 4 –12 hundred decision trees, each of them built over a random extraction of the observations from the dataset and a random extraction of the features.

Not every tree sees all the features or all the observations, and this guarantees that the trees are de-correlated and therefore less prone to over-fitting.



Chosen Attributes

After applying feature selection technique these are the final list of features we will be using to train models we build.

- 1) Total Sulfur Dioxide
- 2) Chlorides
- 3) Volatile Acidity
- 4) Residual Sugar
- 5) Sulfates
- 6) Fixed Acidity



(6) Backward Feature Selection

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.  
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 3.6s remaining: 0.0s  
[Parallel(n_jobs=1)]: Done 12 out of 12 | elapsed: 52.1s finished  
  
[2022-10-23 03:38:29] Features: 11/7 -- score: 0.9597706824880855 [Parallel(n_jobs=1)]: Using  
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 4.6s remaining: 0.0s  
[Parallel(n_jobs=1)]: Done 11 out of 11 | elapsed: 37.6s finished  
  
[2022-10-23 03:39:07] Features: 10/7 -- score: 0.9593957839041553 [Parallel(n_jobs=1)]: Using  
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 3.4s remaining: 0.0s  
[Parallel(n_jobs=1)]: Done 10 out of 10 | elapsed: 30.1s finished  
  
[2022-10-23 03:39:37] Features: 9/7 -- score: 0.9604432571057409 [Parallel(n_jobs=1)]: Using  
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 7.0s remaining: 0.0s  
[Parallel(n_jobs=1)]: Done 9 out of 9 | elapsed: 27.2s finished  
  
[2022-10-23 03:40:04] Features: 8/7 -- score: 0.9604571579560454 [Parallel(n_jobs=1)]: Using  
[Parallel(n_jobs=1)]: Done 1 out of 1 | elapsed: 2.4s remaining: 0.0s  
[Parallel(n_jobs=1)]: Done 8 out of 8 | elapsed: 19.3s finished  
  
[2022-10-23 03:40:23] Features: 7/7 -- score: 0.9602836858094093
```

'volatile acidity',

'residual sugar',

'chlorides',

'total sulfur dioxide',

'Density',

'pH',

'alcohol'



ANN TO CLASSIFY TYPE OF WINE

We constructed multiple artificial neural network models to classify the type of wine having different layers, optimizers and activation function. We recorded the accuracy for each model.

- 1) Type 1: 1 hidden layer having 128 neurons with relu function as activation function and output layer having sigmoid function as activation function. Adam optimizer used for the model. Test accuracy achieved is 97.61538505554199%.
- 2) Type 2: 4 hidden layers used having 64,128,128,256 hidden neurons respectively. The output layer has sigmoid activation function. Adam optimizer used. Test accuracy achieved is 97.76923060417175%.
- 3) Type 3: Same as Type 2 but it uses SGD as optimizer. Test accuracy achieved is :97.76923060417175%.

We can see that as we increased the number of layers the prediction accuracy also increases.



CLASSIFICATION OF WINE TYPE

Logistic Regression

It is statistical analysis method to predict binary outcome , such as Yes or No , based on the prior observations of the data set

```
Training accuracy : 0.9788587848932676
Testing accuracy : 0.9747692307692307
```

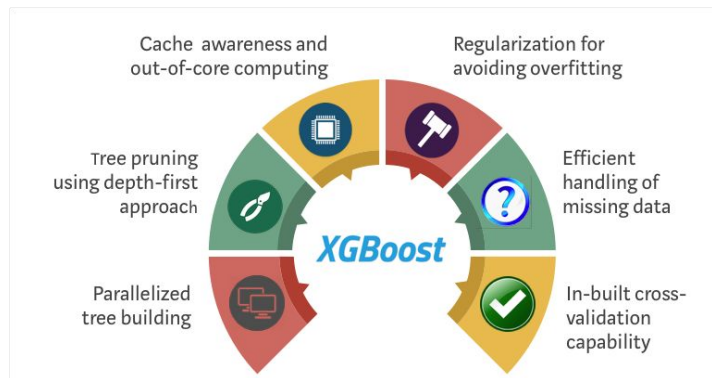
	precision	recall	f1-score	support
0	0.98	0.99	0.98	1233
1	0.97	0.92	0.95	392
accuracy			0.97	1625
macro avg	0.97	0.96	0.96	1625
weighted avg	0.97	0.97	0.97	1625

SVM

We created a SVM with rbf(Radial Basis Function) Kernel. We then fed the training data as well as expected type of wine to the model and trained it. After training the SVM model, the accuracy achieved was 0.9323076923076923%

Accuracy for SVM model:
0.9323076923076923

XGBoost



XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost.

Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree.

These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

XGBoost CLASSIFIER TO CLASSIFY TYPE OF WINE

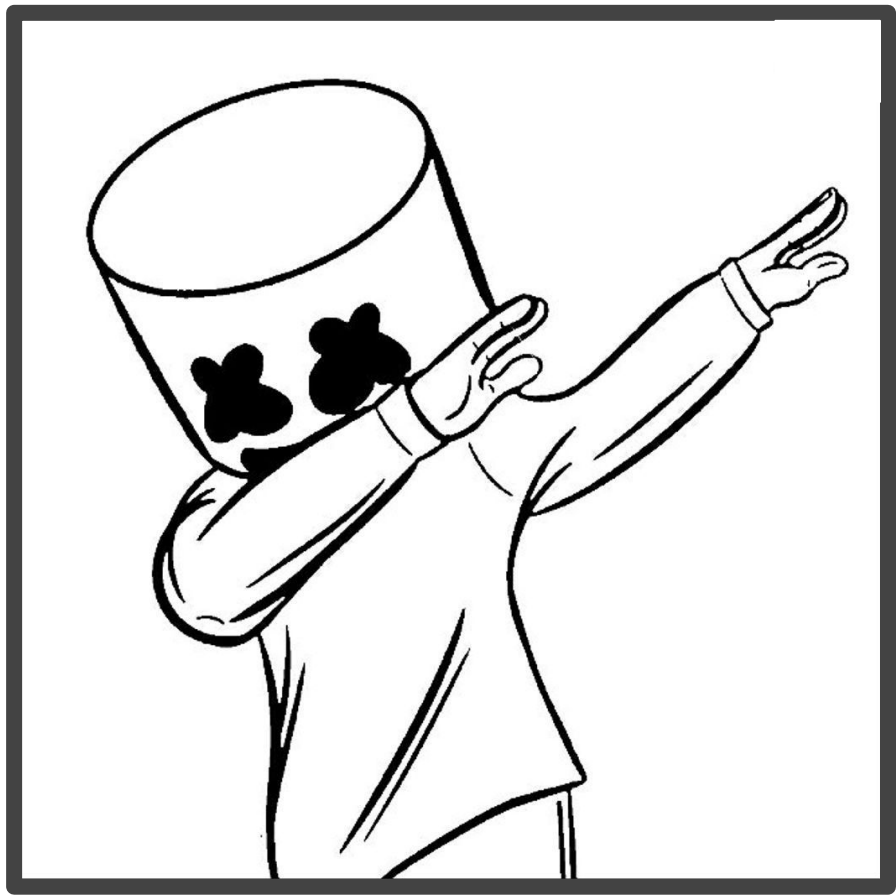
We fed both input data and expected output data to the Classifier to train the model and computed the accuracy over testing data. The accuracy that was achieved is 99%

```
▶ y_pred = model.predict(x_test)

[ ] accuracy = accuracy_score(y_test, y_pred)
    accuracy

0.99
```





Thanks from team

DAB

