| S.NO. | Name | Sap id | Roll no. |
|---|---|---|---|
| 1 | Vanshika parashar | 500076703 | R177219209 |
| 2 | Shrey Chauhan | 500076344 | R177219174 |

# Wine Quality Prediction Using Machine Learning

## Basic overview of the process:

1. Basic understanding of Wine.
2. Data description
3. Importing modules
4. Study dataset
5. Visualization
6. Handle null values
7. Split dataset
8. Normalization
9. Applying model
10. Save model

## 1.Basics understanding of Wine:

Wine is differentiated according to its smell, flavor, and color.

## 2.Dataset description:

There are several features which will be used to classify the quality of wine, many of them are chemicals. Some classifications are given below:

- 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
- sulphates : Added sulfites preserve freshness and protect wine from oxidation, and bacteria.
- alcohol : Percent of alcohol present in wine.

## 3.Importing modules:

Pandas are used for data analysis, NumPy is for n-dimensional array ,seaborn and matplotlib both have similar functionalities which are used for visualization.

## 4.Study dataset:

Here we will check the technical information contained in the data and with this information, we will process our next work.

## 5.Visualization:

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.
The output image will reveal how that data is easily distributed on features.
We plot the bar graph in which we check what value of alcohol is able to make changes in quality.

## 6.Correlation:

Now, we will perform a correlation on the data to see how many features there are correlated to each other.
For checking correlation we use a statistical method that finds the bonding and relationship between two features.
Now, we have to find those features that are fully correlated to each other by reducing the number of features from the data.
Here we write a python program with that we find those features whose correlation number is high, as you see in the program we set the correlation number greater than 0.7 it means if any feature has a correlation value above 0.7 then it was considered as a fully correlated feature, at last, we find the feature total sulfur dioxide which satisfy the condition.

## 7.Handle null values:

In the database, there is so much notice data present, which will affect the accuracy of our ML Model. In machine learning, there are many ways to handle null or missing values. Now, we will use them to handle our unorganized data. We will simply fill null values with the help of fillna()function.

## 8.Splitting dataset:

Now we perform a split operation on our dataset.

## 9.Normalization:

We do normalization on numerical data because our data is unbalanced; it means the difference between the variable values is high so we convert them into 1 and 0.

## 10.Applying Model:

Here we will use *Random Forest Classifier* because it was the only ML model that gives the 88% accuracy which was considered as the best accuracy.

## ***11.saving Model:***

Then we save our machine learning model.

**Our machine learning prediction is over**.