**Red Wine Quality Prediction Project**

**1.Project Description**

The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).  
  
This dataset can be viewed as classification task. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Also, we are not sure if all input variables are relevant. So, it could be interesting to test feature selection methods

2.Data Analysis

My analysis will use Red Wine Quality Data Set, available on the <https://github.com/FlipRoboTechnologies/ML-Datasets/blob/main/Red%20Wine/winequality-red.csv>

My first step was to clean and prepare the data for analysis. I went through different steps of data cleaning. First, I checked the data types focusing on numerical and categorical to simplify the correlation’s computation and visualization. Second, I tried to identify any missing values existing in our data set. Last, I researched each column/feature’s statistical summary to detect any problem like outliers and abnormal distributions.

3.EDA Concluding Remarks

In EDA part we plot Heatmap to see which variables are likely to affect the quality of red wine the most, I ran a correlation analysis of our independent variables against our dependent variable, quality. This analysis ended up with a list of variables of interest that had the highest correlations with quality.

In order of highest correlation, these variables are:

1. Alcohol: the amount of alcohol in wine

2. Volatile acidity: are high acetic acid in wine which leads to an unpleasant vinegar taste

3. Sulphates: a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant

4. Citric Acid: acts as a preservative to increase acidity (small quantities add freshness and flavour to wines)

5. Total Sulphur Dioxide: is the amount of free + bound forms of SO2

6. Density: sweeter wines have a higher density

7. Chlorides: the amount of salt in the wine

8. Fixed acidity: are non-volatile acids that do not evaporate readily

9. pH: the level of acidity

10. Free Sulphur Dioxide: it prevents microbial growth and the oxidation of wine

11. Residual sugar: is the amount of sugar remaining after fermentation stops. The key is to have a perfect balance between — sweetness and sourness (wines > 45g/litres are sweet

In this dataset contains1599 rows and 12 columns which 1 our dependent variable, quality, and remaining 11 is dependent variables I found the popularity of the medium/average values of quality: 5 and 6. Considering the dependent variable’s transformation, I found out that our data is normally distributed.

4.Pre -Processing Pipeline

In this phase before model building as per the tip which provided on given dataset so they were say like we are finding the wine quality which is greater than equal to 7 or higher than we can consider it as a good quality means 1 otherwise all remainder as bad quality means0

Perform features selection on this I replaces quality column to good quality and apply for loop and check it is greater than or equal to to 7 or not check its good quality or bad. Then we separating features and label so drop quality column in x axis and store good quality in y axis so we our x is independent variable and y is dependent variable.

5.Building Machine Learning Models

Based on the EDA and correlation analysis, three potential models were used in the modelling part. First, we finding random state model then apply classification algorithm.

we check train and test data and also confusion matrix and precision, recall ,f1 score.

1. Random Forest model it gives 94% accuracy
2. Logistic Regression model it gives 81%accuracy
3. Support vector machine model it gives84% accuracy
4. Extra tree classifier model it gives 96% accuracy and its gives us best accuracy

We check cross validation and we analyse extra tree classifier model is best model we apply hyper parameter tunning on this for best accuracy.

6.Concluding Remark

By analysing the physicochemical tests samples data of red wines from the north of Portugal, I was able to create a model that can help predict the quality of red wine products and have a better understanding of each critical and up-to-date features. I have found that the Model Extra tree classifier -based feature sets performed better than others. In general, using Model 4 as our best model for prediction, I determined four of the features as the most influential: volatile acidity, citric acid, sulphates, and alcohol. To be more specific, high-quality wines seem to have lower volatile acidity, higher alcohol, and medium-high sulphate values. Meanwhile, lower-quality wines tend to have low values for citric acid First, the main problem came from the fact that our data set was unbalanced. A majority of the quality values were “regular” (5 and 6), which made no significant contribution to finding an optimal model. These values made it harder to identify each factor’s different influence on a “high” or “low” quality of the wine, which was the main focus of this analysis. In order to improve our predictive model, we need more balanced data the data set was it only had 12 attributes, which can narrow down the accuracy of our predicting quality of red wine.