**Kaggle Project 10**

**Wine Quality Prediction**

**Prepared By: Dwaipayan Mukherjee (2211569, AIA – AIML)**

**About the Project:**

Here the aim is to predict the quality of wine based on given features. We use the wine quality dataset available on Internet for free. This dataset has the fundamental features which are responsible for affecting the quality of the wine. Using several Machine learning models, I will predict the quality of the wine.

**Steps that I follow throughout the project:**

1. **Imports**: I have done the necessary imports.
2. **Loading the data and performing EDA**:

**[2.1] EDA on train data:**

First I have loaded the train data as a Pandas Dataframe and checked out the Info of the dataset as well as the descriptive statistics.

Next, I have explored if there are any duplicate values and missing values in this dataset. I found there are none.

Next, I have explored some data visualizations where I have investigated the relation between each of the columns in the data and the distributions of each column separately.

I checked for the skewness of the columns in the dataset. The column **residual sugar** had outliers, so I have worked on them. I have done some simple monotonic transformations on the columns like square root and logarithm. I have taken log of values in **chloride** column, square root of the values in **total sulfur dioxide** column, log of values in the **sulphates** column, log of values in the **fixed acidity** column, log of values in the **alcohol** column, log of values in the **volatile acidity** column, and log of values in the **free sulfur dioxide** column.

Next, I have checked for the correlation of the columns in the train dataset.

I have label encoded the target column **quality**.

Lastly, I have prepared the training data by splitting the target and the features columns. I have dropped the irrelevant **Id** column from the features data, since it does not provide any signal to the quality of wine.

**[2.2] EDA on the test dataset:**

**I have repeated the same steps as on the training set, except label encoding the target column since the test data does not have the target column.**

1. **Models:**

I have used four models to train on the training dataset, namely Random Forest Classifier, XGBoost Classifier, LightGBM Classifier and the CatBoost Classifier. I have used a 5-fold cross validation while training and evaluating each model. The metric used for model evaluation is **Quadratic Weighted Kappa (QWP, henceforth)**.

**[3.1] Random Forest Classifier:**

I have used the Random Forest Classifier with random state as 2 and maximum depth as 4. The resultant 5-fold average QWP came out to be **0.4329**.

**[3.2] XGBoost Classifier:**

I have used the XGBoost classifier with objective as multi:softmax, number of classes as 6, learning rate as 0.01, number of estimators as 500, maximum depth as 4, subsample as 0.7, column sample by tree as 0.7 and random state as 2. The resultant 5-fold QWP came out to be **0.5178**.

**[3.3] LightGBM Classifier:**

I have used the LGBM Classifier with objective as multiclass, number of classes as 6, learning rate as 0.01, number of estimators as 500, maximum depth as 4, subsample as 0.7, column sample by tree as 0.7 and random state as 2. The resultant 5-fold QWP came out to be **0.5139**.

**[3.4] CatBoost Classifier:**

I have used the CatBoost Classifier with loss\_function as MultiClass, maximum depth as 4, learning rate as 0.01 and number of estimators as 500. The resultant 5-fold QWP came out to be **0.4949**.

1. **Final Prediction:**

Since the XGBoost Classifier performed the best (highest QWP), I have used this model to train on the entire training data and then predict on the test data.

1. **Submission:**

I have stored the predicted quality in the submission dataset, used inverse transformation on these labels to get back the original labelling and saved it as a CSV file.