Wine Quality Prediction

Submitted to : Prof Si Thu Aung

- Ali Hussain Ladiwala
- Karan Pinakinbhai Jariwala
- Varun kumar Ejanthkar
- Kandi Sandeep Reddy

Goal

The goal of this project is to predict the quality of wine given metrics such as acidity, citric acid, residual sugar, chlorides etc

Dataset Description

Attributes

There are 6497 samples of white wine in the data sets. Each wine sample (row) has the following characteristics (columns):

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol
- Quality (score between 0 and 10)

Data Visualization

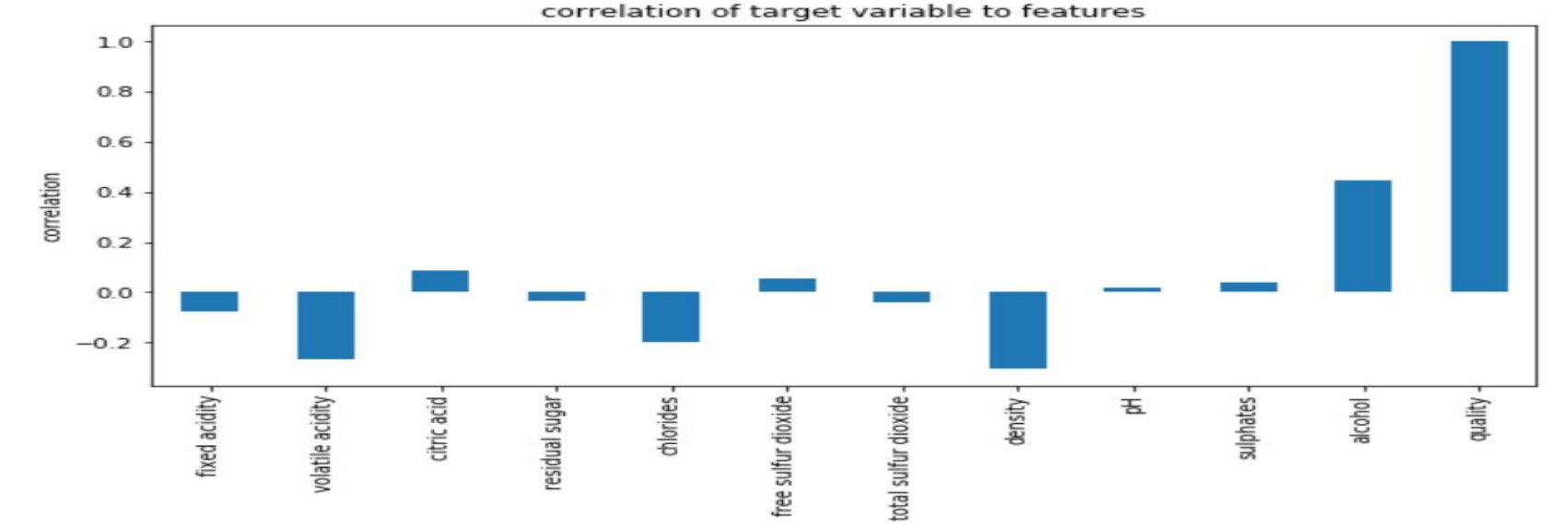
Data Description

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

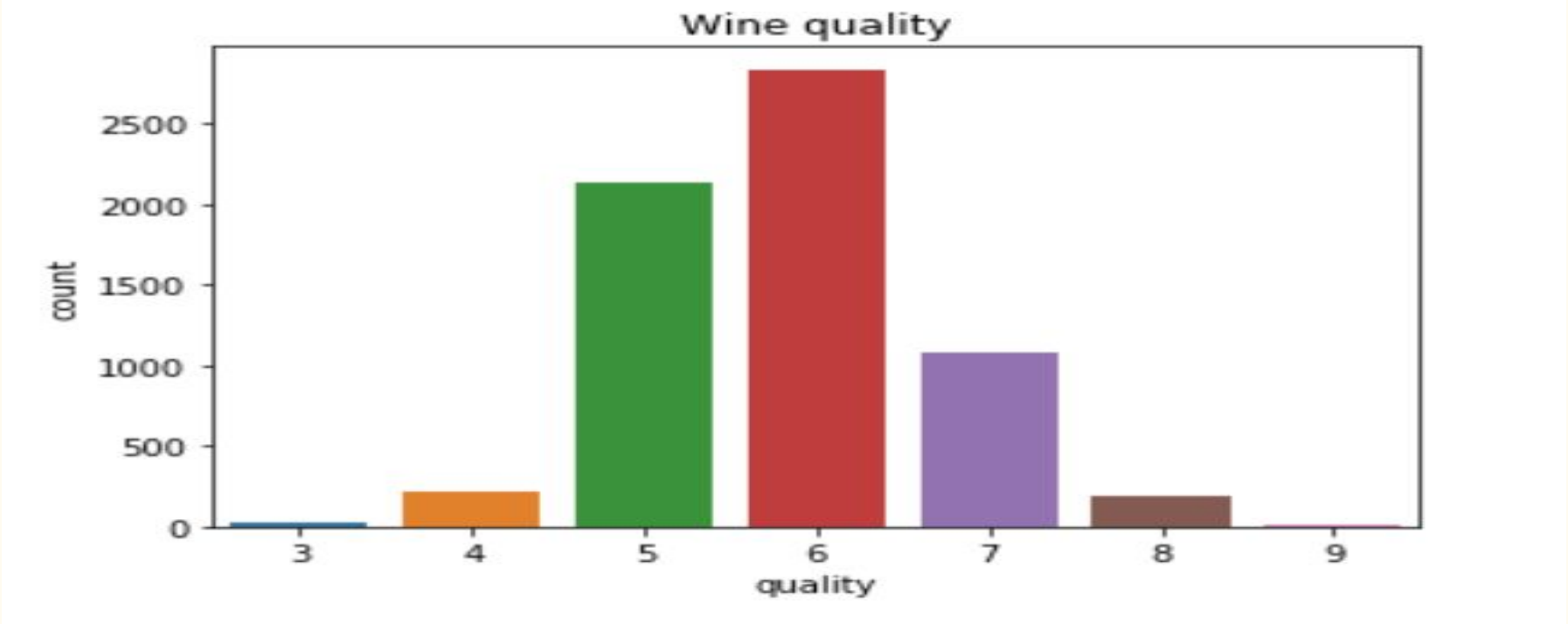
Data Description

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	6487.000000	6489.000000	6494.000000	6495.000000	6495.000000	6497.000000	6497.000000	6497.000000	6488.000000	6493.000000	6497.000000
mean	7.216579	0.339691	0.318722	5.444326	0.056042	30.525319	115.744574	0.994697	3.218395	0.531215	10.491801
std	1.296750	0.164649	0.145265	4.758125	0.035036	17.749400	56.521855	0.002999	0.160748	0.148814	1.192712
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.720000	0.220000	8.000000
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340	3.110000	0.430000	9.500000
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890	3.210000	0.510000	10.300000
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990	3.320000	0.600000	11.300000
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	1.038980	4.010000	2.000000	14.900000

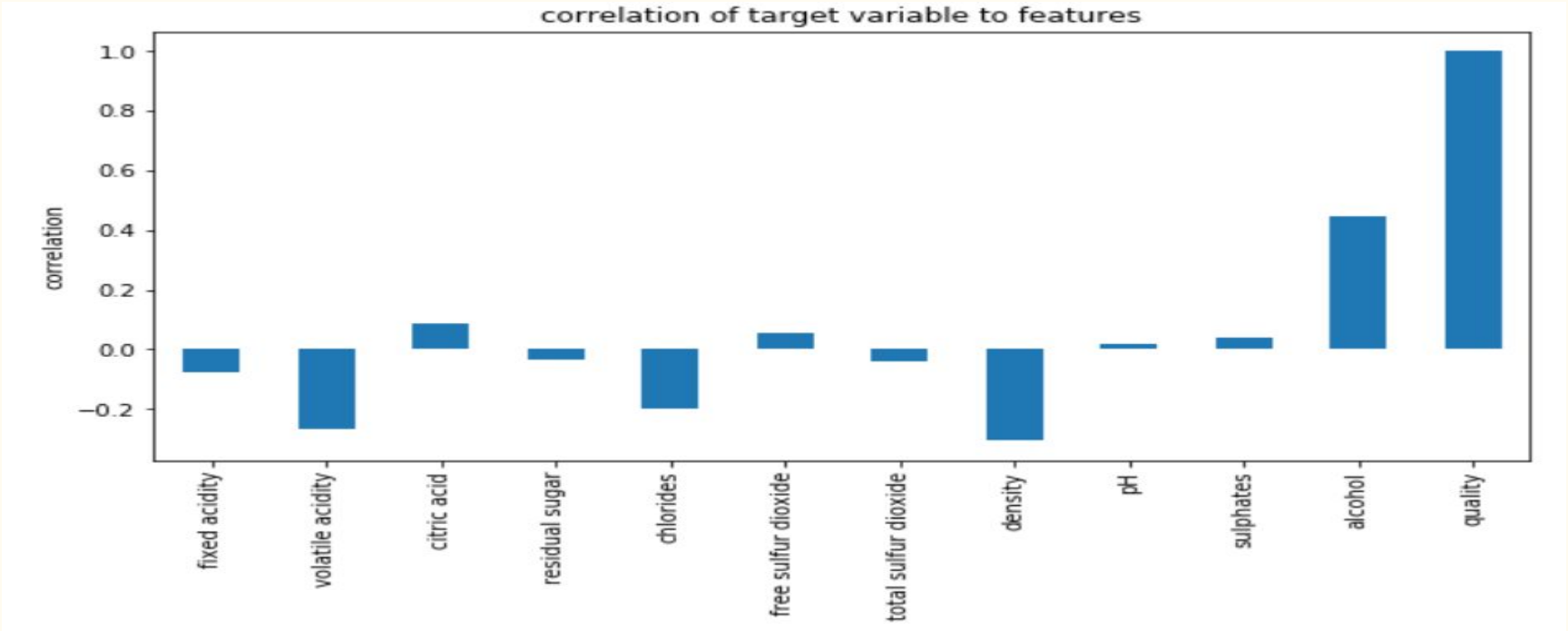
Data Correlation



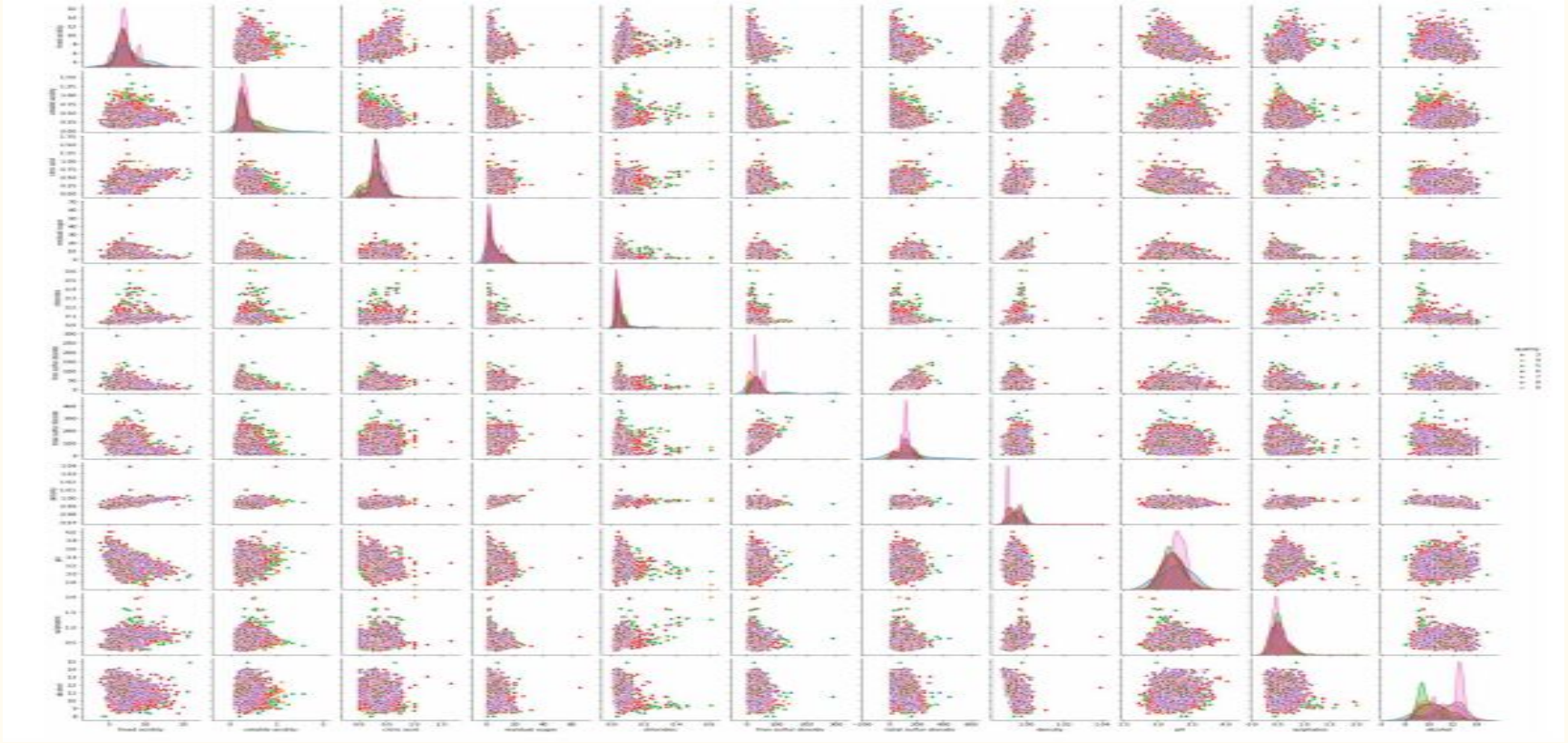
Distribution of quality



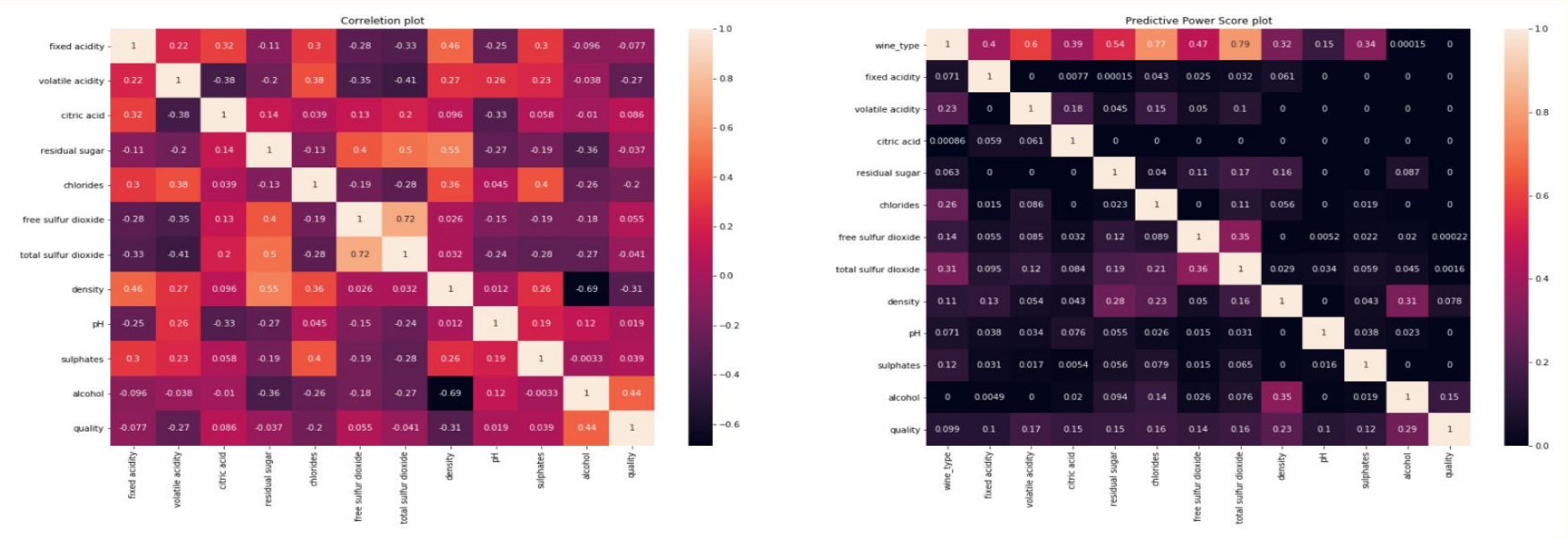
Data Correlation



Data Correlation

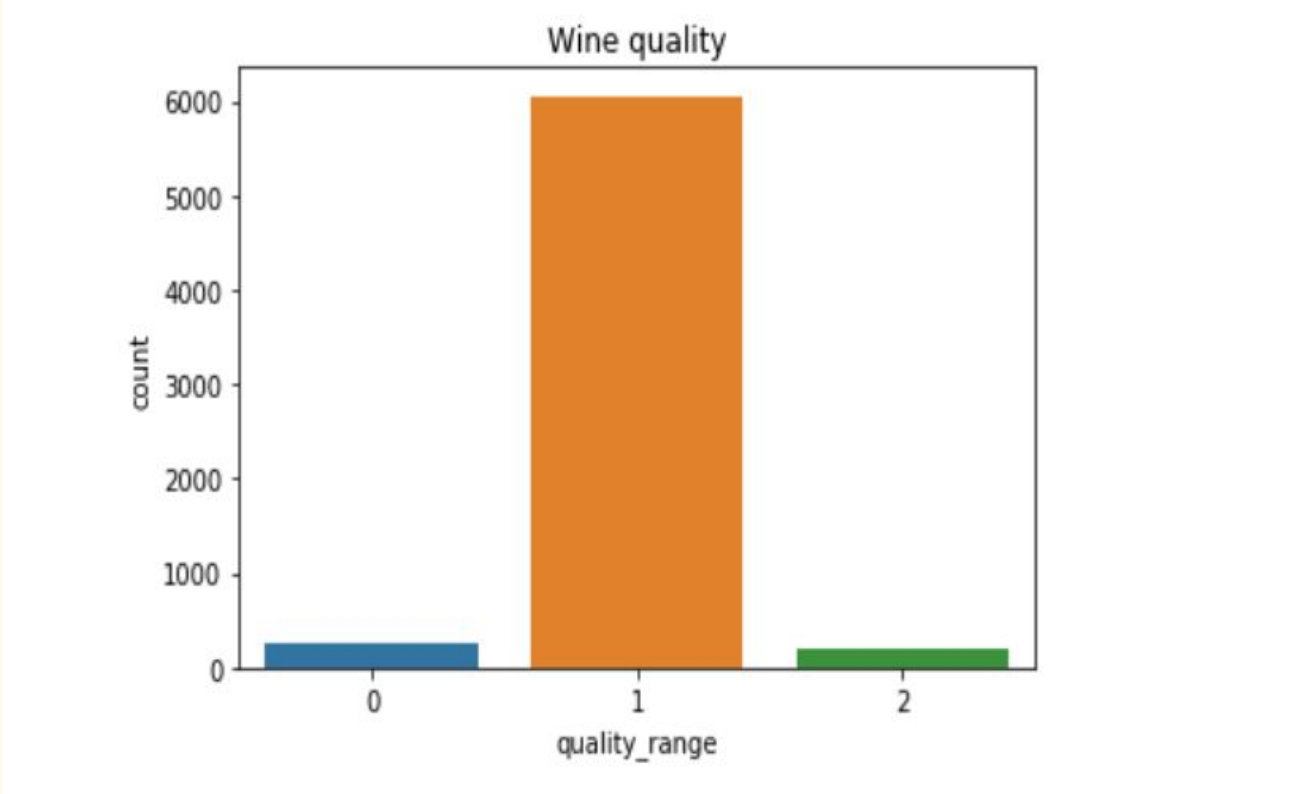


Data Correlation



Data Pre-Processing

Dividing quality into 3 bins of (low, medium Average)



Data Preprocessing/Cleaning

- Handling missing values and replacing it by the mean.

	Sum	Percentage
wine_type	0	0.000000
fixed acidity	10	0.001539
volatile acidity	8	0.001231
citric acid	3	0.000462
residual sugar	2	0.000308
chlorides	2	0.000308
free sulfur dioxide	0	0.000000
total sulfur dioxide	0	0.000000
density	0	0.000000
pH	9	0.001385
sulphates	4	0.000616
alcohol	0	0.000000
quality	0	0.000000

	Sum	Percentage
wine_type	0	0.0
fixed acidity	0	0.0
volatile acidity	0	0.0
citric acid	0	0.0
residual sugar	0	0.0
chlorides	0	0.0
free sulfur dioxide	0	0.0
total sulfur dioxide	0	0.0
density	0	0.0
pH	0	0.0
sulphates	0	0.0
alcohol	0	0.0
quality	0	0.0

Model

Models

Logistic Regression

Accuracy Score: 0.911

Logistic Regression is a classical linear method for binary classification.

Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function

It fits a line to best separate the two classes.

Logistic Regression ability to provide probabilities and classify new samples using continuous and discrete measurements makes it a popular ML method.

K-Nearest Neighbour (Score: 0.938)

Simplest Classification Technique.
K refers to number of nearest neighbours that will be used to predict.
The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

- Initialize K to your chosen number of neighbors

- Calculate the distance between the query example and the current example from the data. Add the distance and the index of the example to an ordered collection

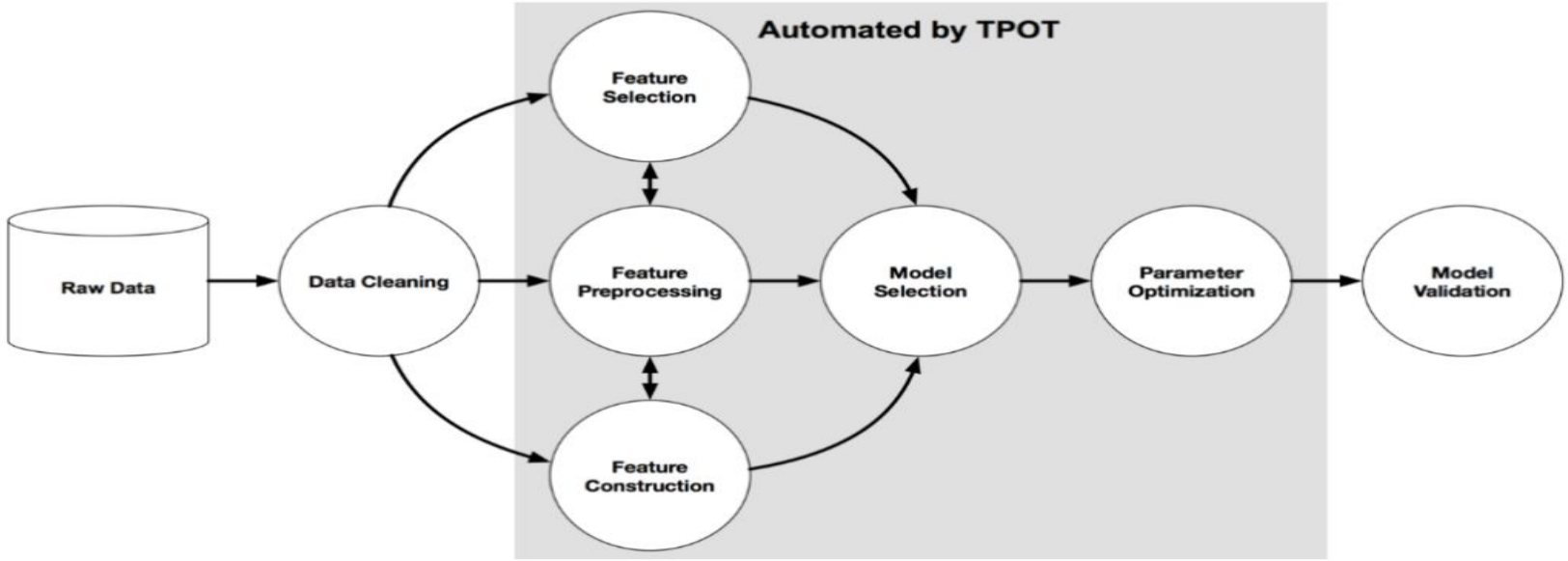
- Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
- Pick the first K entries from the sorted collection
- Get the labels of the selected K entries. Return the mode of the K labels

Exploring AutoML

What is AutoML ?

It is tool for automating the process of finding the best ML algorithm for your dataset and use case.

Library used in our project



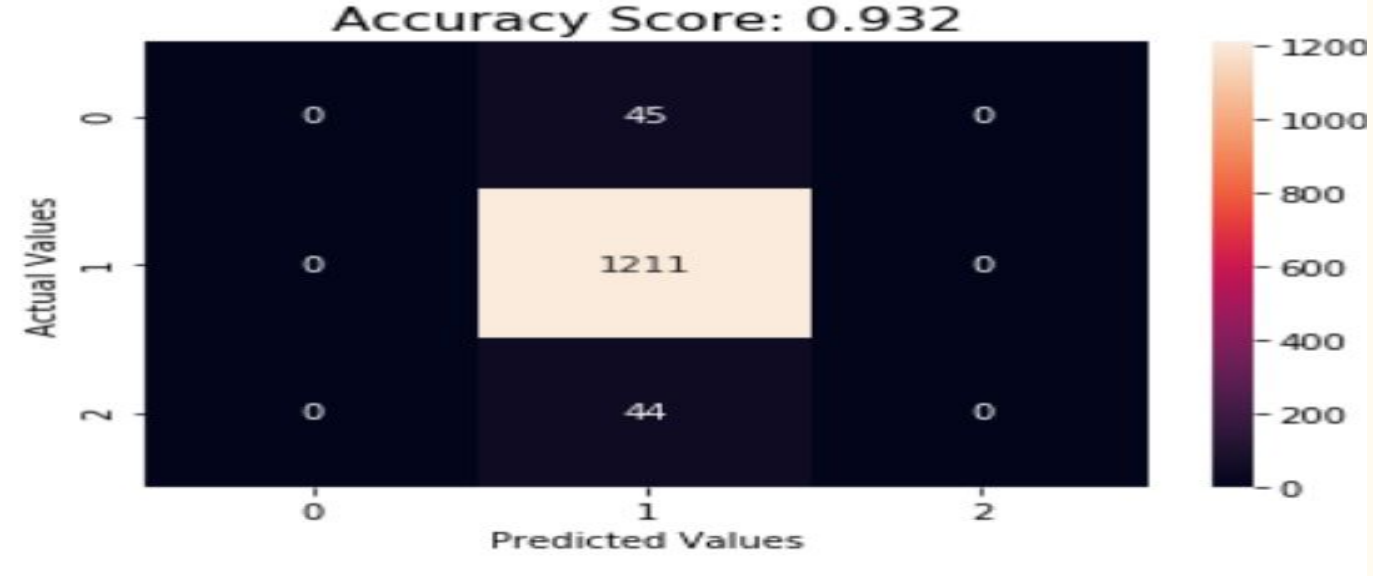
source: TPOT Documentation

Output from AutoML

```
Warning: xgboost.XGBClassifier is not available and will not be used by TPOT.

HBox(children=(FloatProgress(value=0.0, description='Optimization Progress', max=1100.0, style=ProgressStyle(d…

Generation 1 - Current best internal CV score: 0.9428516695047012
Generation 2 - Current best internal CV score: 0.9428516695047012
Generation 3 - Current best internal CV score: 0.9428516695047012
Generation 4 - Current best internal CV score: 0.9428516695047012
Generation 5 - Current best internal CV score: 0.9430439771970089
Generation 6 - Current best internal CV score: 0.943044162286222
Generation 7 - Current best internal CV score: 0.943044162286222
Generation 8 - Current best internal CV score: 0.943044162286222
Generation 9 - Current best internal CV score: 0.9432366550677427
Generation 10 - Current best internal CV score: 0.9432366550677427

Best pipeline: KNeighborsClassifier(RFE(input_matrix, criterion=entropy, max_features=0.55, n_estimators=100, step=0.6500000000000001),
n_neighbors=49, p=2, weights=distance)
```

Using Automl to Predict

- Automl is used for prediction which in turn uses K Nearest Neighbours to predict values.
- Output:



WebApp/Deployment

Web Application

Frontend: Streamlit
Backend: Python
Deployment: Heroku

Deployed Web-app

Demo

Deployed Web-app

Thank you