Predicting Wine Quality

Using ML algorithms

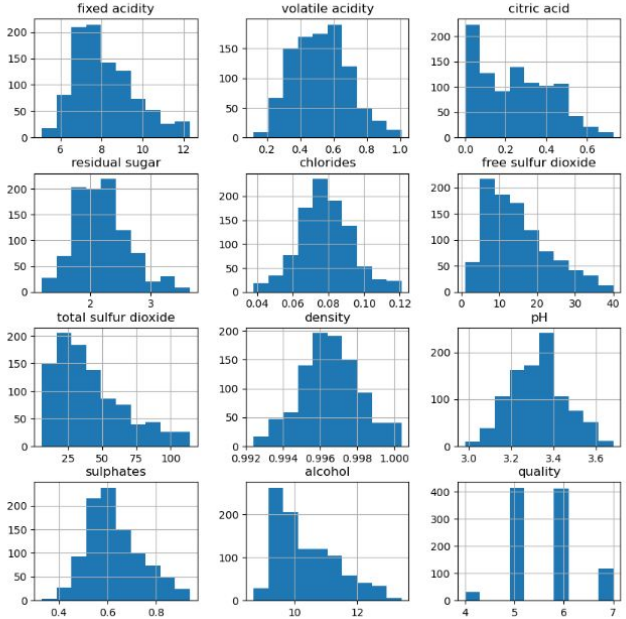
Problem Statement

- To analyze the quality of wine based on factors like fixed acidity, volatile acidity, citric acid, residual sugar, chlorides etc.
- To find quality of wine which can be rated between 1-10 or bad - good.
- To predict quality of wine.

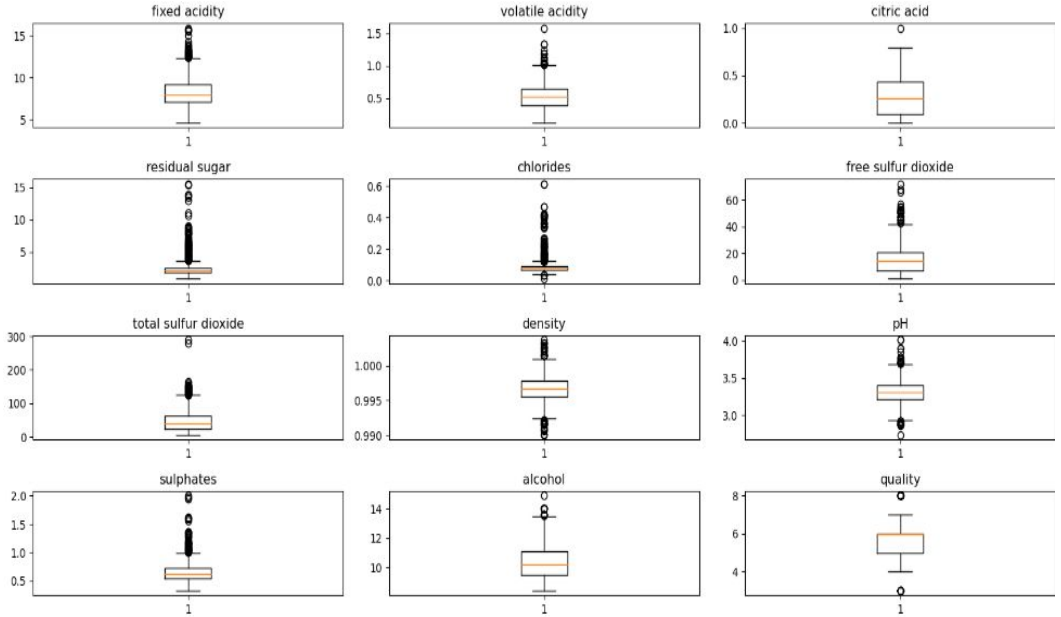
EDA (Exploratory data analysis)

- **Data Inspection** - Basic Statistics, Information about Data such as data types of each columns, no. of rows and columns etc.
- **Missing values**, **Duplicate rows** - Identified using simple functions provided by pandas library.
- **Outliers** - Outlier detection of all variables using box plots.
- **Correlation Analysis** - Correlation between all variables using heatmap.
- **Feature Analysis -** Understanding how data of each column is spread using histogram.

Feature analysis



Outlier Detection



Data Pre - processing

- **Handling Duplicates** - Redundant data needs to be removed.
- **Handling Outliers** - Removing outliers since it did not have significant contribution to the data.
- **Feature Selection** - The correlation matrix shows week correlation between volatile acidity, chlorids, total sulphur dioxide, density
- **Splitting Data** - Split into train and test data for training and testing our model.
- **Feature Scaling** - Helps the data to be on the same scale, since different features seem to have variety of ranges.

Model Selection

- Logistic Regression, Decision Tree, Random Forest and KNN

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.63	[0.0, 0.69, 0.62, 0.43]	[0.0, 0.75, 0.6, 0.42]	[0.0, 0.72, 0.61, 0.43]
1	Decision Tree	0.63	[0.0, 0.69, 0.62, 0.43]	[0.0, 0.75, 0.6, 0.42]	[0.0, 0.72, 0.61, 0.43]
2	Random Forest	0.62	[1.0, 0.67, 0.6, 0.5]	[0.0, 0.74, 0.62, 0.32]	[0.0, 0.7, 0.61, 0.39]
3	KNN	0.62	[1.0, 0.67, 0.6, 0.5]	[0.0, 0.74, 0.62, 0.32]	[0.0, 0.7, 0.61, 0.39]

- The target variable was extremely imbalanced which resulting in overfitting in one of the categories.
- Changing the target variable from multiclass to binary class seems to resolve this issue.
- The quality of wine can also be classified as good or bad

MODEL SELECTION FOR BINARY CLASSIFICATION

- Random Forest is the best classification algorithm for prediction of wine quality.
- Prediction of wine quality prediction is correct 90% of the time. (Accuracy)
- Wine is classified as Good among the total prediction 53% of the time. (Precision)
- Wine is classified as Good when its actually good 29% of the time. (Recall)
- Reduced values of Precision and Recall can be due to imbalanced data.

Challenges

- **Data Quality** - Analyzing the target variable, attempting to balance the data using oversampling and undersampling. Conversion of target variable to binary to increase accuracy.
- **Feature Selection** - Removal of features with less collinearity between the features and target variable did not contribute any significant difference to the model.

Conclusion

- Understanding of poor wine quality is important as its a threat to human safety. The ML model provides solution to this issue by predicting quality of wine.
- Random forest ML model provides a solution that is economical with good accuracy, high efficiency and least human interaction.