Dataset:

- The dataset used in this analysis was named “Wine Quality Data Set” from the Machine Learning Repository.
- It analysed two different wine variants (red & white) in two seperate files, and presented their various attributes before finally giving a wine quality rating for each row of data.
- The attributes of the dataset were fixed acidity, volatile acidity, citric acidity, residual sugar, chlorides, sulphur dioxide (free and total), density, pH, sulphates, alcohol and quality of wine.

Goal:

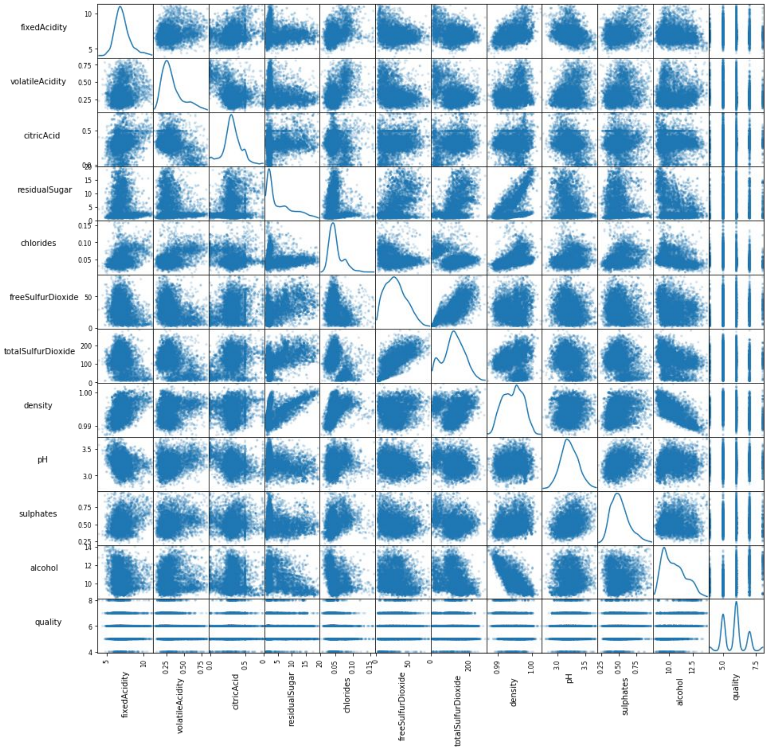
- The aim of this report was to develop a classification model that could enable us to predict the colour of wine, based on the various attributes of wine.

Hypothesis:

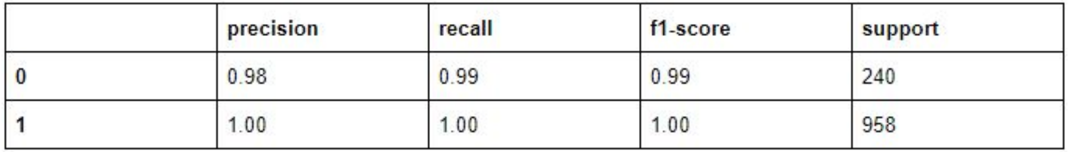
- Red and white wine differ in attributes enough to separate and accurately predict them.

Analysis of Results:

- **K-Nearest Neighbour** (no. of neighbours=61)
  - The highest f1-score split for the K-Nearest Neighbour model was an 80% training sample and a 20% testing sample.
  - The f1-score for the model is 99% for red wine and close to 100% for white wine and the accuracy(test) of the model is 99.5% while the accuracy(train) is 99.1%.
- **Decision-Tree** (min_samples_split=150,max_depth=10,min_samples_leaf=200)
  - Decision-Tree lacked behind with only a highest f1-score of 90% for red wine and 98% for white wine on the same split.

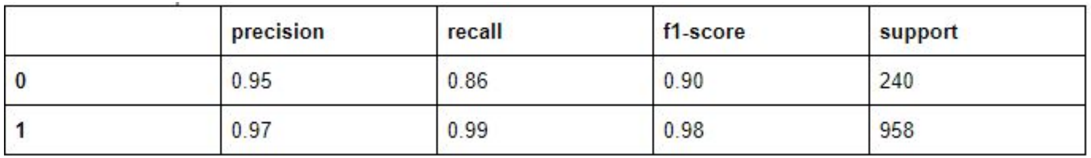


K-Nearest Neighbour: 80% train, 20% test



	precision	recall	f1-score	support
0	0.98	0.99	0.99	240
1	1.00	1.00	1.00	958

Decision Tree: 80% train, 20% test



	precision	recall	f1-score	support
0	0.95	0.86	0.90	240
1	0.97	0.99	0.98	958

Accuracy Train vs Accuracy Test:

- Accuracy Train = 0.9908161135462326
- Accuracy Test = 0.994991652754591
- Not Overfitting (as accuracy train < accuracy test)

Confusion Matrix:

```
[[238   2]
 [  4 954]]
```

Data Preparation Steps:

1. Importing data to Python
2. Checking data was to specifications (e.g. correct length and attributes)
3. Cleaning data entry errors (e.g. extra white spaces, capital letters, etc.)
4. Appending the white wine to red wine dataset to create a joined dataset
5. Adding variable to the end of rows that gives information as to whether or not it is red or white wine.

Conclusion/Recommendation:

- K-Nearest Neighbour with an 80% training sample and 20% testing sample is the most effective classification modelling technique for this dataset and goal.
- Using this model, we should be able to say with a 99.5% accuracy whether a wine is white or red.
- To confirm the hypothesis, red and white wine differ enough to separate and accurately predict them.