Predicting Color & Quality of Wine by analyzing its chemical components using Random Forest Analysis

CSE891 Utkarsh Sachaan

Why a Model to predict Quality of Wine?

- Currently Quality of wine is determined by connoisseurs of Wine.
- A master Sommelier could demand upto $150,000 per year.
- Wouldn't it be great if there is an analytical model or scale to determine the quality of wine.

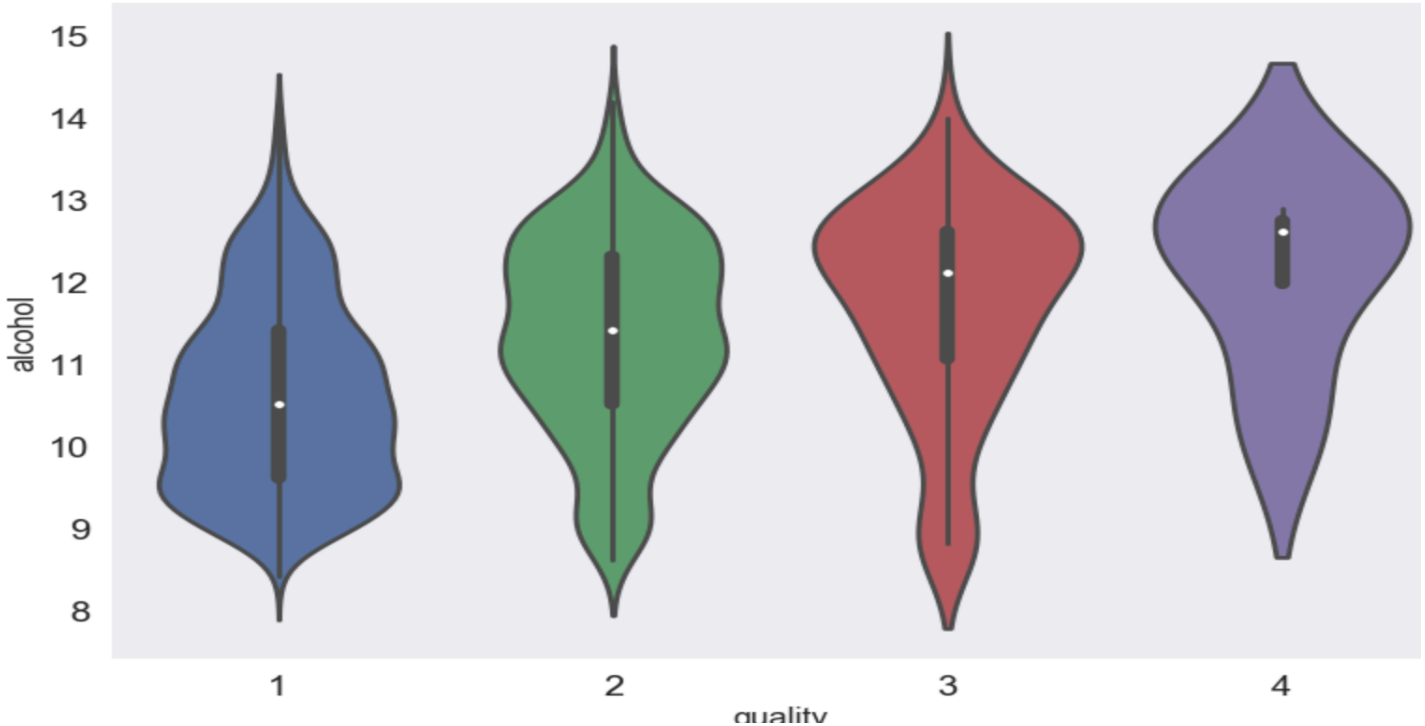
Data Source

- Have used the data from UCI machine learning repository.
- The data has eleven components as fixed acidity, residual sugar, chlorides, density, pH, sulphates etc.

Method

- Collating data for Red & White wine to increase complexity.
- Randomizing dataset for each load.
- Applying Random Forest from sklearn package with exploratory analysis on data.

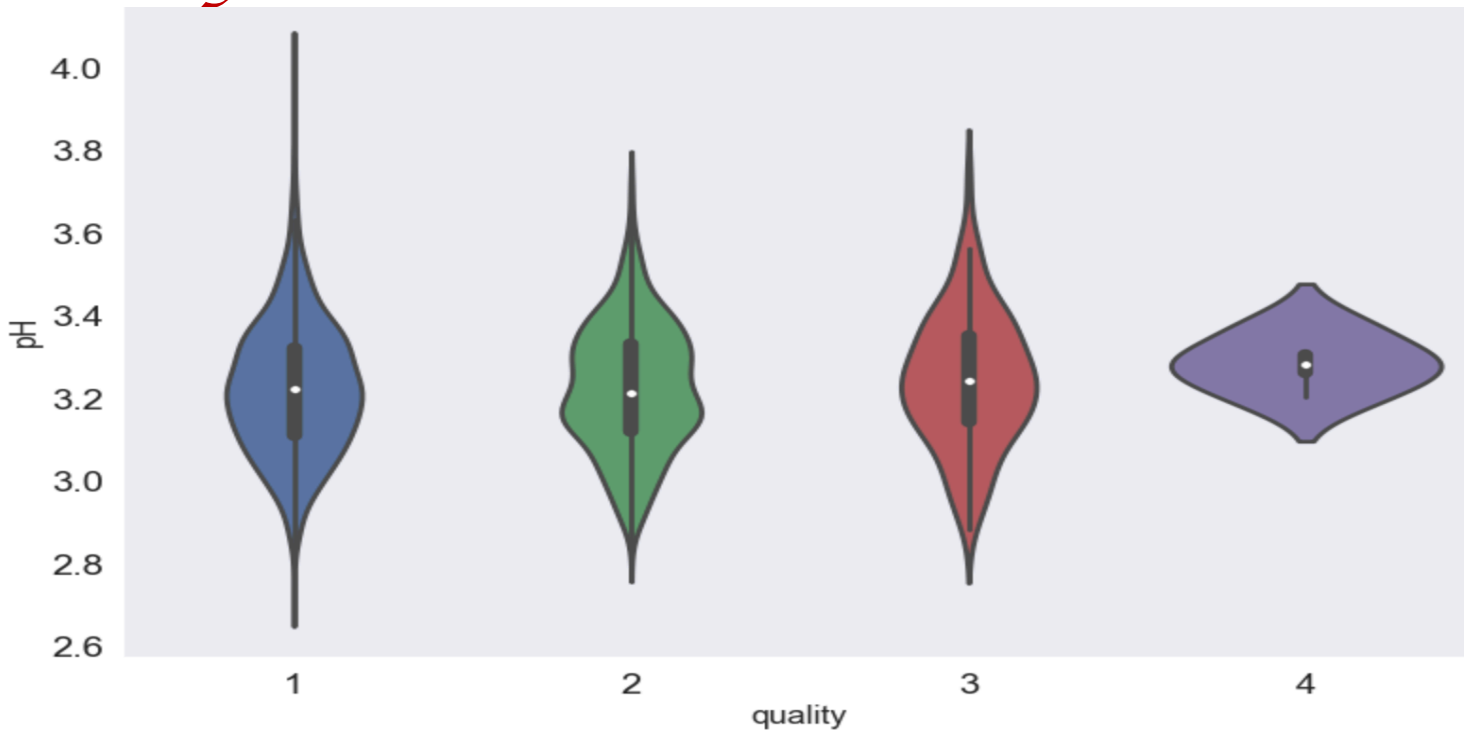
Insights from Data exploratory Analysis



Sample graph depicting better quality wine has higher alcohol content.

Continued..

Insights from Data exploratory Analysis



Sample graph proving that better quality wine has much balanced pH level.

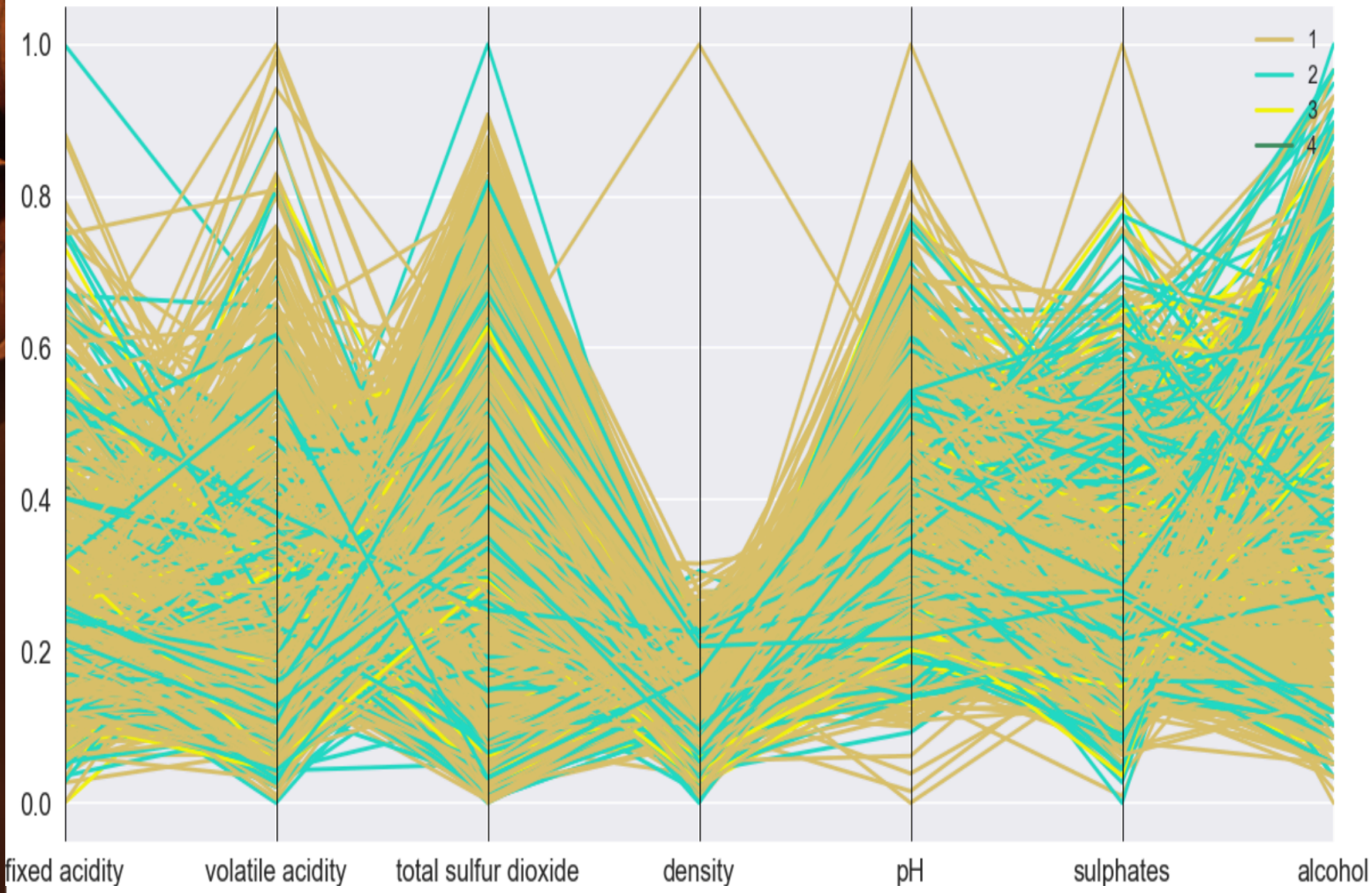
Insights from First run with all Variables

The random Forest Model differentiates Red from White wine with a efficiency of 96%.

```
-1.84734973477
0.969152485784
```

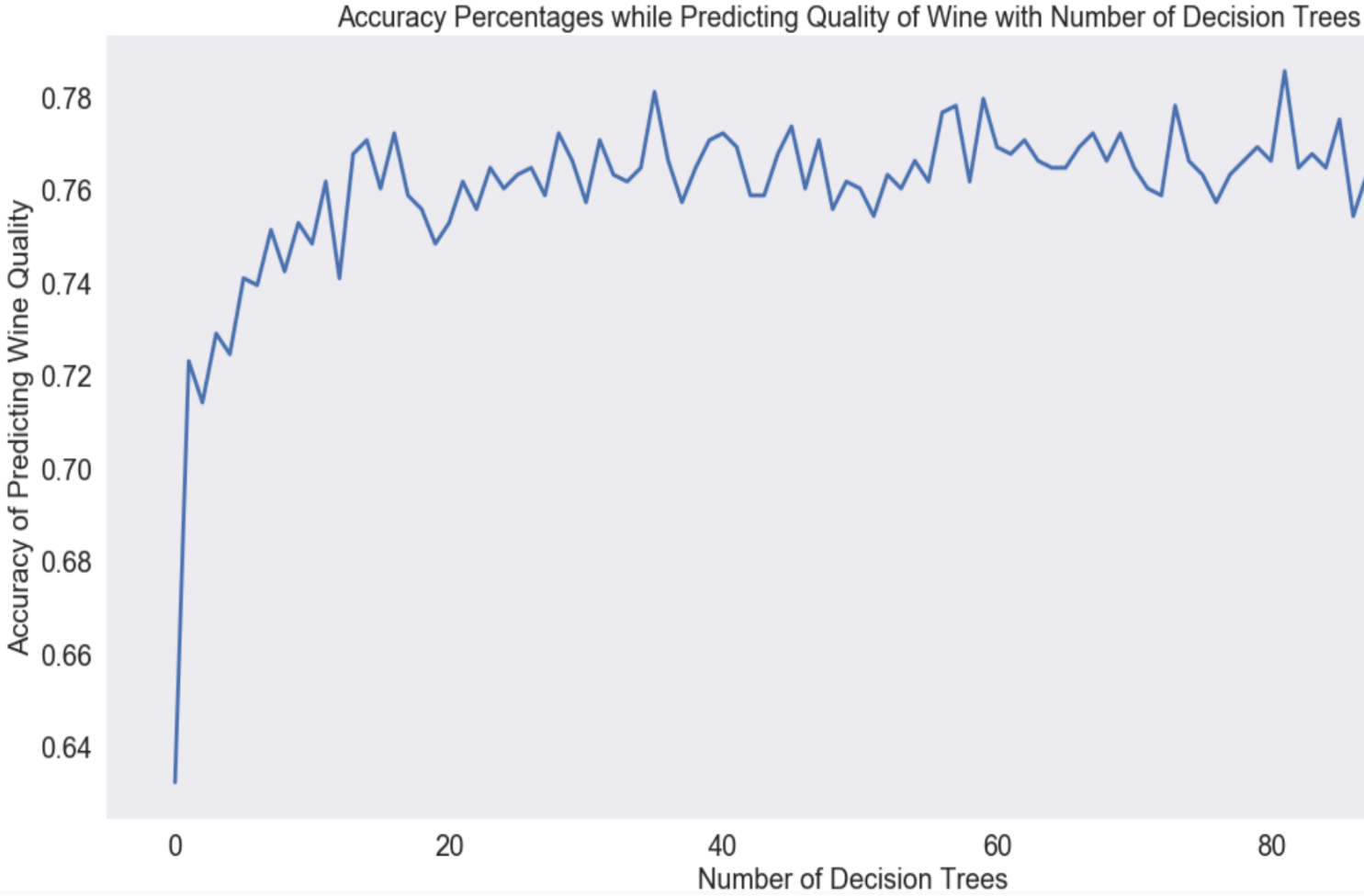
However, model fails miserably to predict the correct Quality of Wine while working with all variables.

Applying parallel coordinated plot to visualize the interaction of parameters



This does not provide any significant insights.

Running Random Forest after removing correlated variables



Accuracy of 76% while predicting Quality f Wine

Work In-Progress..

Currently, I am looking at algorithms as LDA to further understand the Interaction between variables, so that a better accuracy can be achieved.