**CSE891 Python Project Write-up**

**Project Overview:**

**T**he goal of the project is to develop a model which predicts the Color & Quality of wine by analyzing its chemical components with an acceptable accuracy.

**Scope of Project:**

**T**here has always been lot of debate on effectiveness & accuracy on a Sommelier’s work, evidently a sommelier could earn anywhere from $50K-$100K/year, this huge range in earnings is because a sommelier with more years of experience is considered better. However, a sommelier’s accuracy is not consistent due to human errors as traditional methods involve observing the color, swirl & smell of wine. Thus, this project develops a model/scale that could predict the quality of wine by analyzing its chemical components.

**Source of Data:**

**T**he data for the project is collected from UCI Machine Learning Repository provided below:

<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>

The data contains about 5000 instances divided in Red & Wine Datasets, with 12 attributes providing a decent veracity & volume to generate a predictive model. Few of the Input variables are listed below as an example:

1 - fixed acidity   
2 - volatile acidity   
3 - citric acid   
4 - residual sugar, etc.

**Method:**

**D**eveloping a single script that effectively predicts Color & Quality of Wine; The data from Red, White wine dataset is collated creating a data-dump of ‘N’ rows (greater than 2000, less than 3000) randomized on each load, effectively creating a unique dataset on each load for analyses.

Though the Quality of wine, in raw dataset ranges from 1-9, the model focuses on predicting only on high-quality wine 6-9, also due to very few rows with wine quality 1-5 these are ignored for building an optimized model.

Graphing variables to understand co-relation, then test & Train Datasets are created & Random Forest modelling technique is applied to fit the test set using sklearn package library to predict Color & Quality of wine.

**Data Exploration & Visualizations:**

**P**erforming various Data Exploratory analysis provides below insights from the data:

* Performing Parallel Co-Ordinated Plot technique to uncover hidden trends in the variables, does not give any reasonable direction (Figure 1).

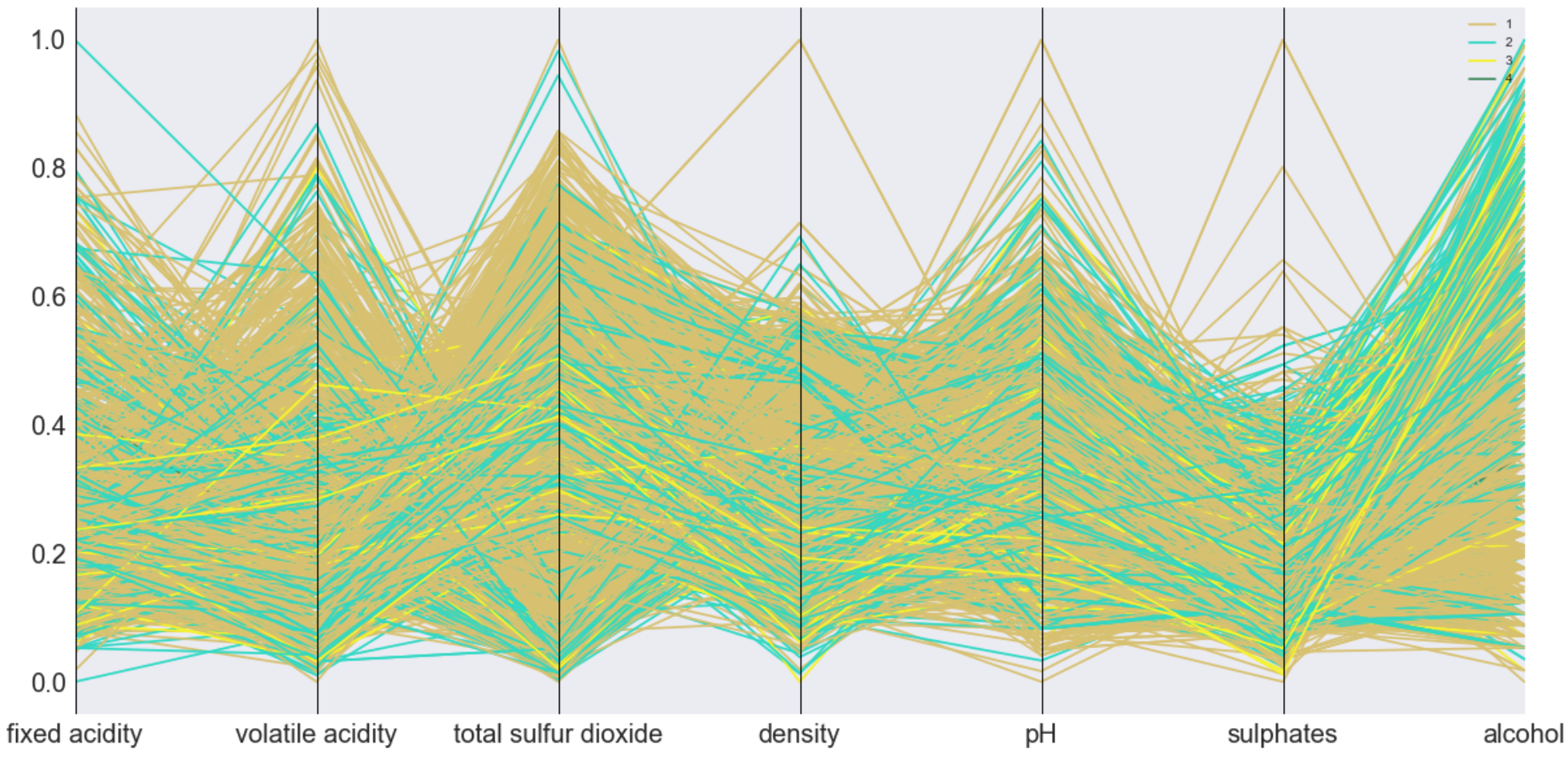


Figure 1

* Visualizing pH vs Quality by Violin plot proves that high Quality Wine has much stable pH value (Figure 2).

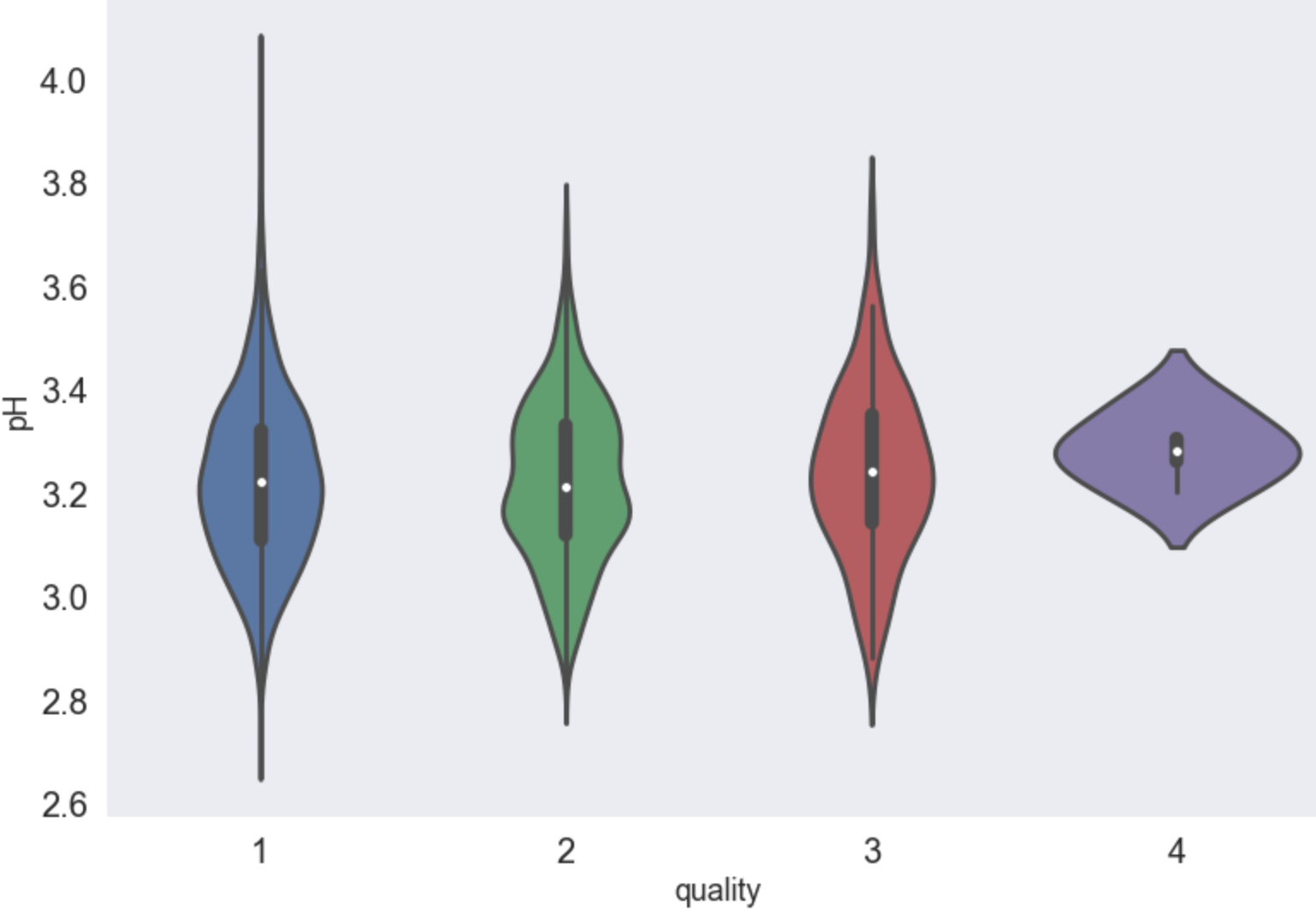


Figure 2

* Visualizing Alcohol vs Quality suggests that better Quality wine on an average has higher alcohol content (Figure 3).

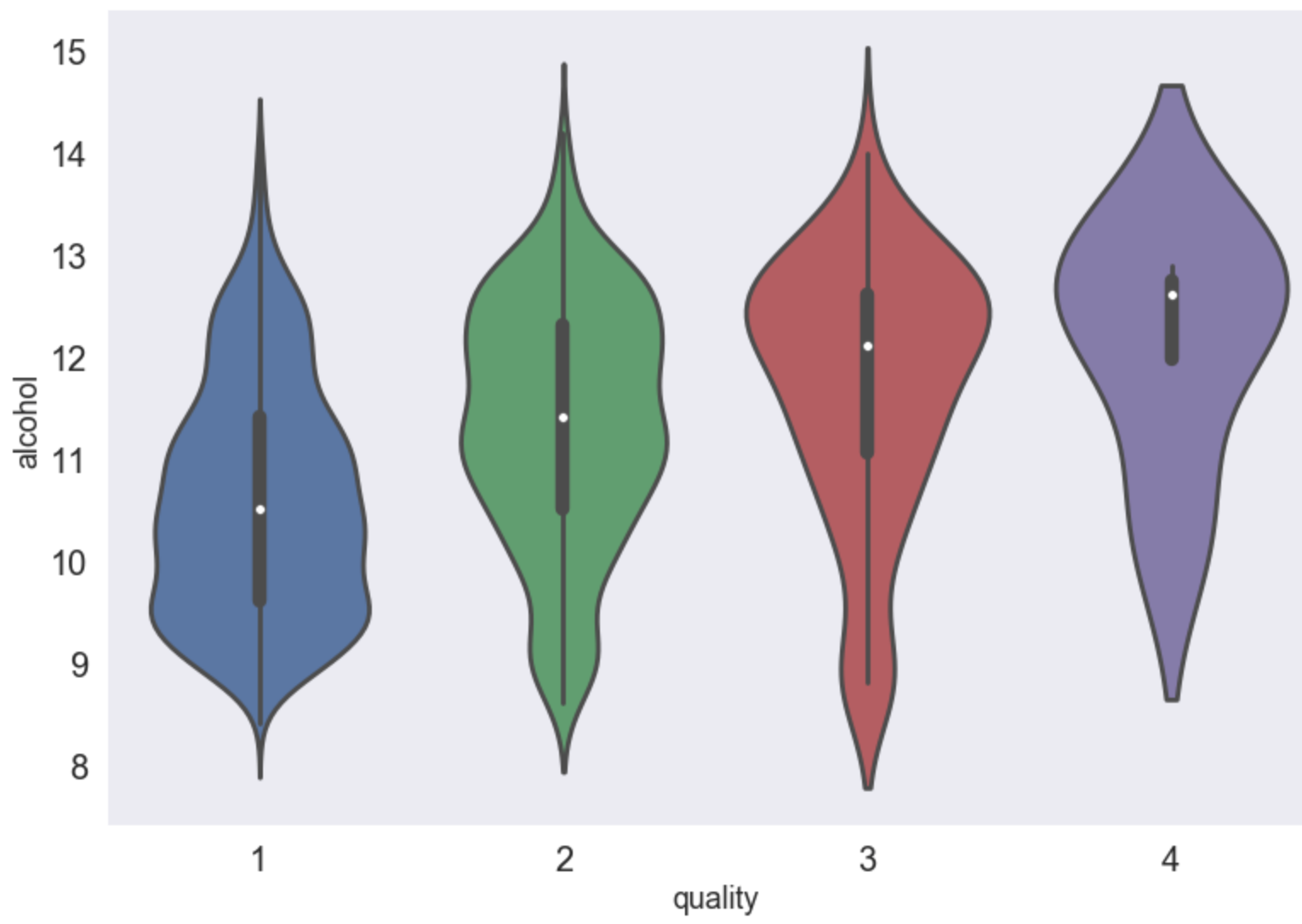


Figure 3

**Insights & Conclusion:**

**U**sing Random forest predictive modelling technique, the model gives an accuracy of 97% in classifying if the Wine is Red or White, the model utilizes all the variables from the dataset.

After removing multi-collinearity (removing variables) and re-creating test/train datasets with the normalized data the random forest model predicted the quality of wine by an accuracy of 70% at around19th Tree, cross-validation was carried out by LDA (Linear Discriminant Analysis).

Thus, the model currently gives at-best an accuracy of 72% to predict the Quality of Wine.