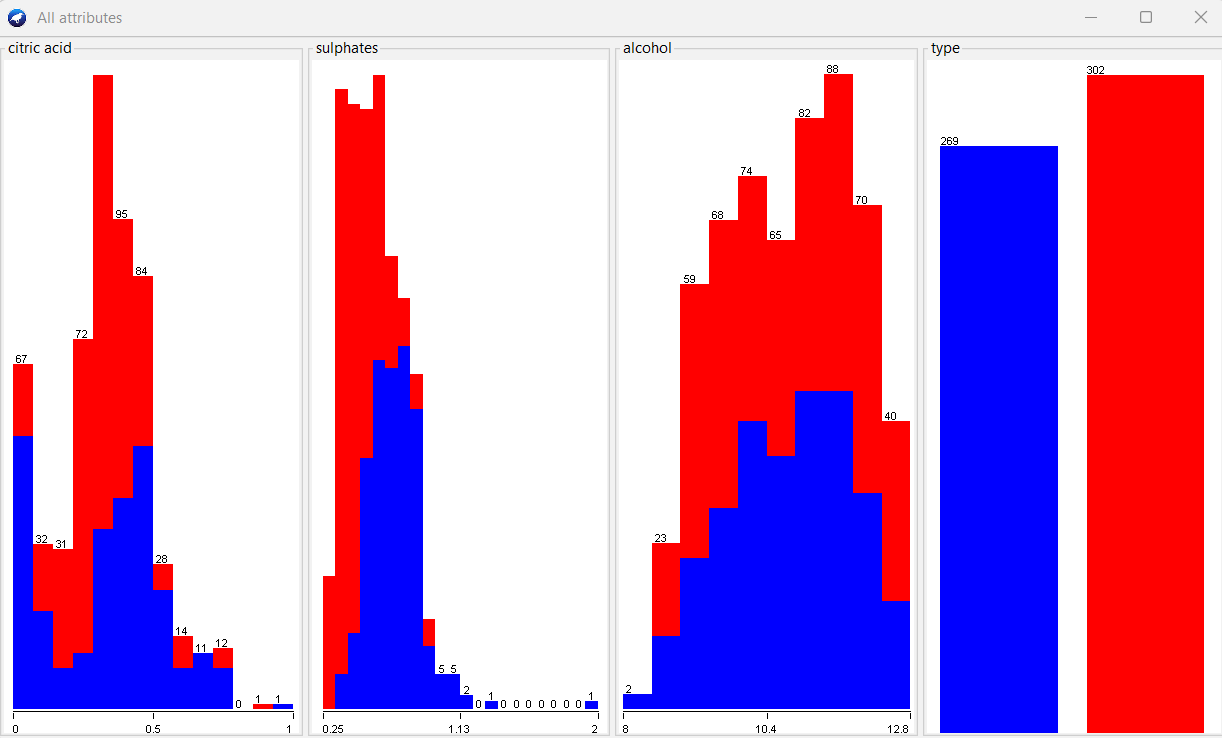
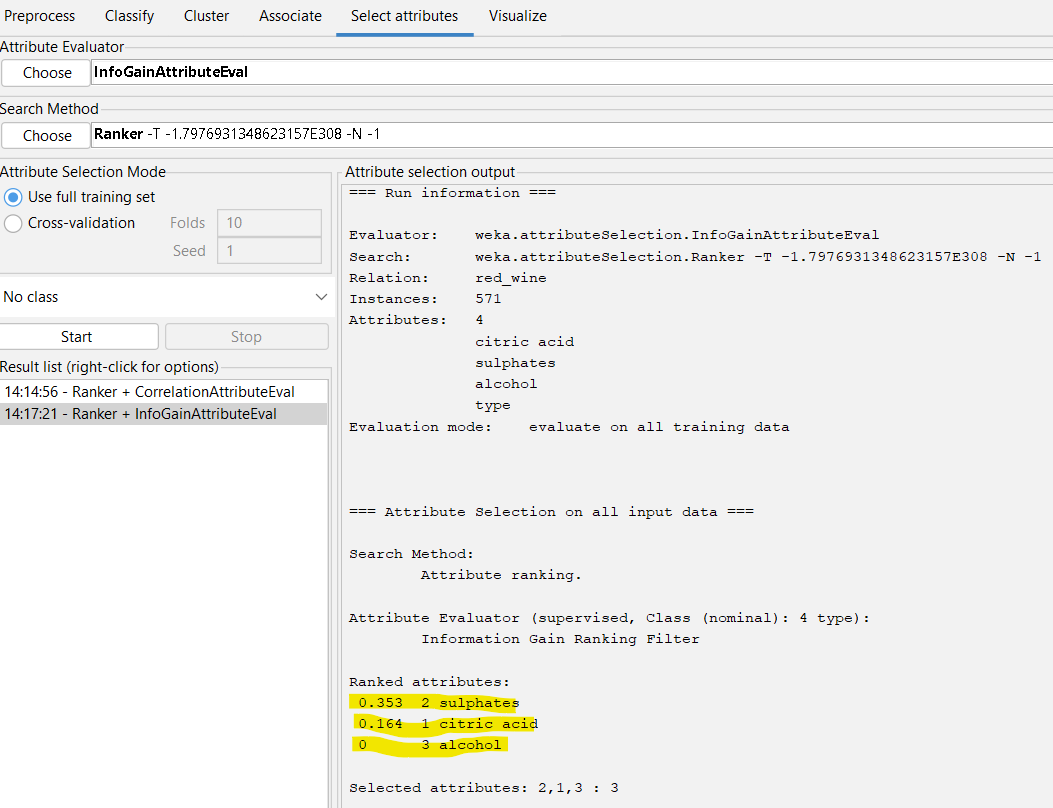
**WEKA TASKS:**



1. Comparing the plots, in my understanding, since the plot for alcohol is more distributed and with less number of outliers, alcohol is more predictive. Since sulphates attribute has more outliers, the prediction of wine quality based on sulphates may not be accurate. Also, according to general knowledge, the 4 factors to determine wine quality are Acidity, Tannin, Alcohol, and Sweetness. Alcohol also adds viscosity and helps balance sweetness and acidity. Hence, alcohol is the most predictive attribute in my opinion.
2. My speculation was not consistent with the results produced after running the logistic regression model. After running the logistic regression model, with the help of information gain evaluation and correlation attribute evaluation, it is found that sulphates is more predictive than alcohol and citric acid. The results are as follows:

AUC score: 0.828 and Accuracy: 79.33%



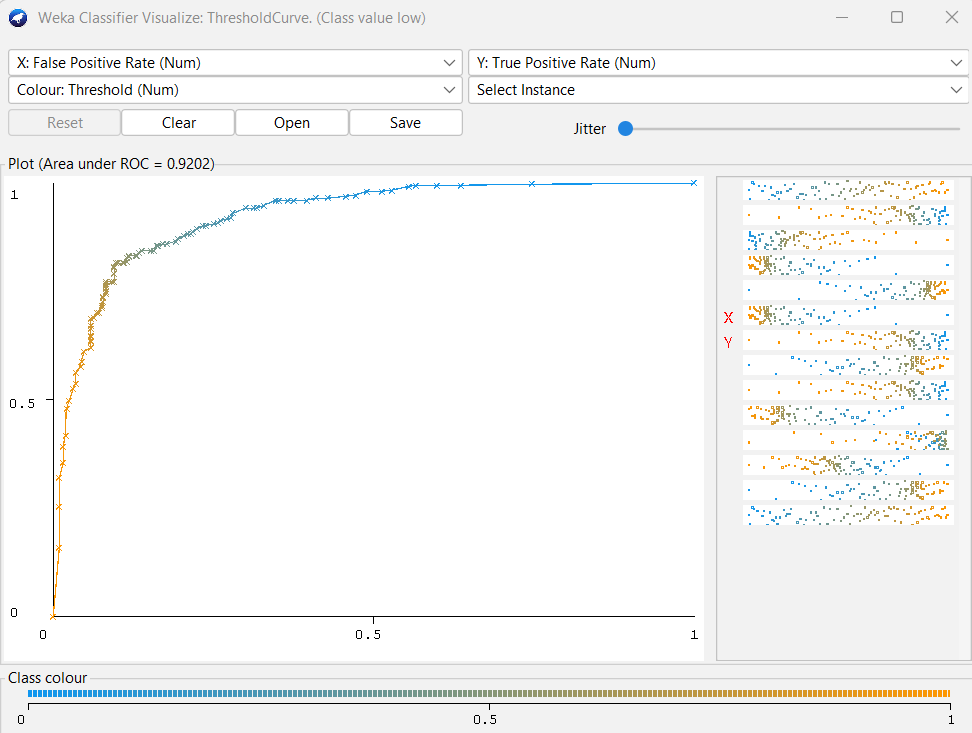
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **ZeroR** | **OneR** | **LR** | **NB** | **DT** | **SVM** | **RF** |
| **AUC** | N/A | N/A | 0.8735 | 0.890 | 0.877 | 0.781 | 0.920 |
| **Accuracy** | 52.8897% | 78.4588% | 79.3345% | 82.486% | 84.232% | 78.2837% | 85.289% |

1. Observing the above results, the best performance for the red-wine dataset was given by the Random Forest model with an accuracy of 85.289% and an Area Under ROC of 0.9202 which is higher when compared to the other models.

**Random Forest:** In a Random Forest model, a set of decision trees is trained on randomly selected subsets of the training data. During the prediction phase, the Random Forest combines the outputs of all the individual trees to produce a final prediction.

**ROC:** ROC is a graphical representation of the performance of a binary classification model by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different classification thresholds.

**AUC:** AUC is a metric that measures the overall performance of a binary classification model based on the area under the ROC curve. The AUC ranges between 0 and 1, with a higher value indicating a better performance of the model.

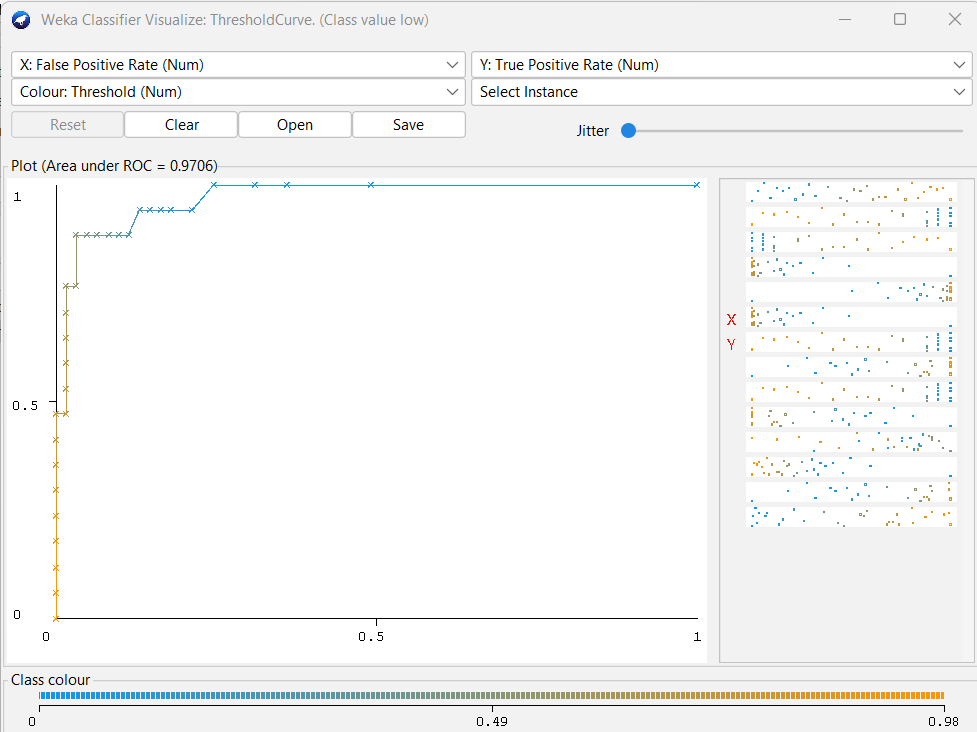


**Comment:** The accuracy of 85.289% suggests that the model correctly predicts the wine quality for 85.289% of the samples in the dataset. The AUC score of 0.9202 suggests that the model has a good ability to distinguish between the positive and negative classes. Overall, it is a good model to predict the quality of the wine with the given attributes.

1. From question 2, the highest accuracy and AUC score are for Random Forest Model on red wine dataset. Fitting the model on white wine data gives the following results.

Accuracy: 92.3077%

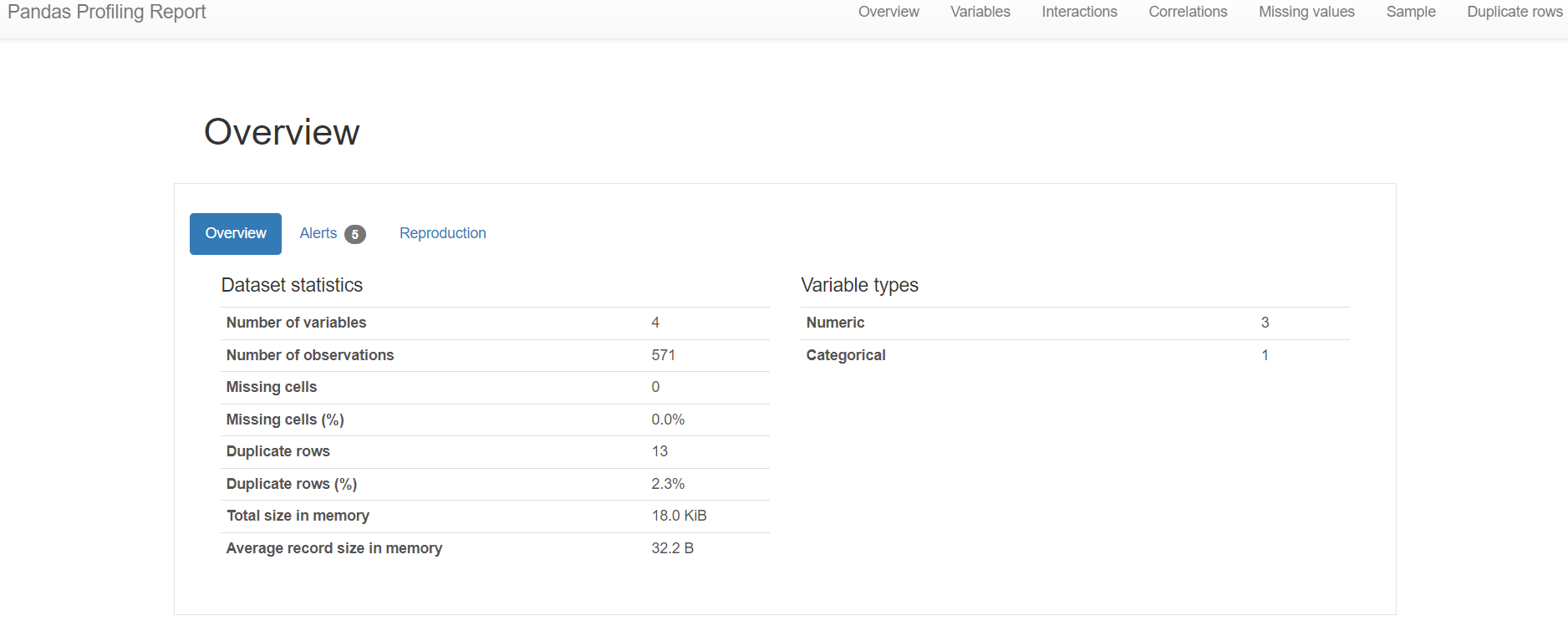
AUC Score: 0.971

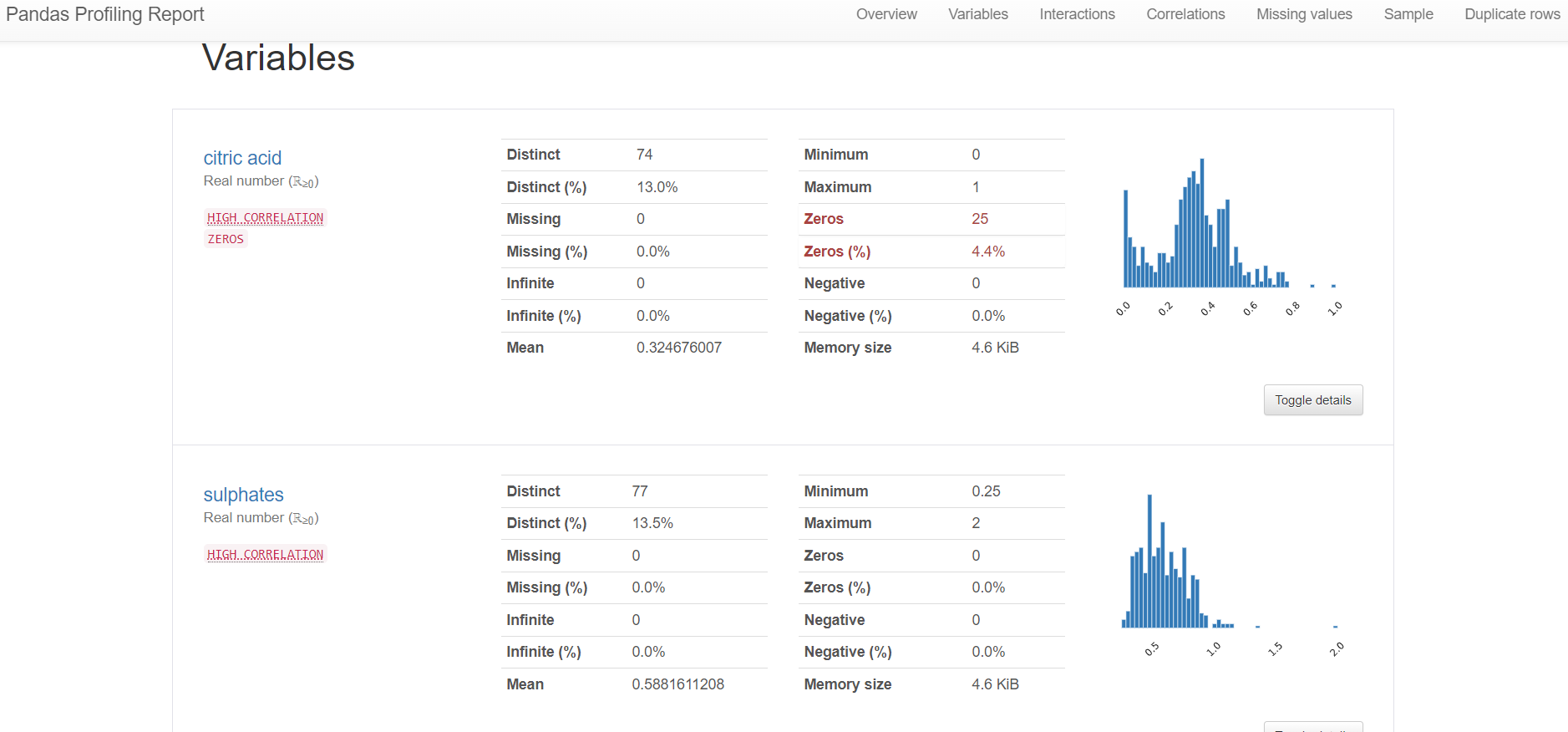


**Comment:** The accuracy of 92.3077% suggests that the model correctly predicts the wine quality for 92.3077% of the samples in the dataset. The AUC score of 0.971 suggests that the model has a good ability to distinguish between the positive and negative classes.

**PYTHON:**

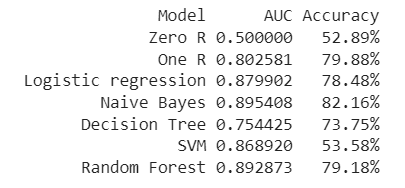
1. The screenshot of the overview of the red wine dataset and the html file:



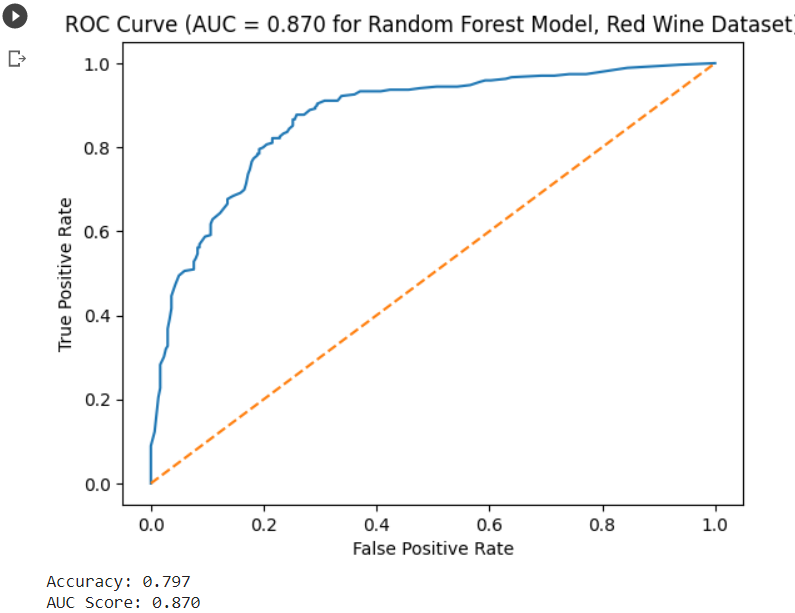




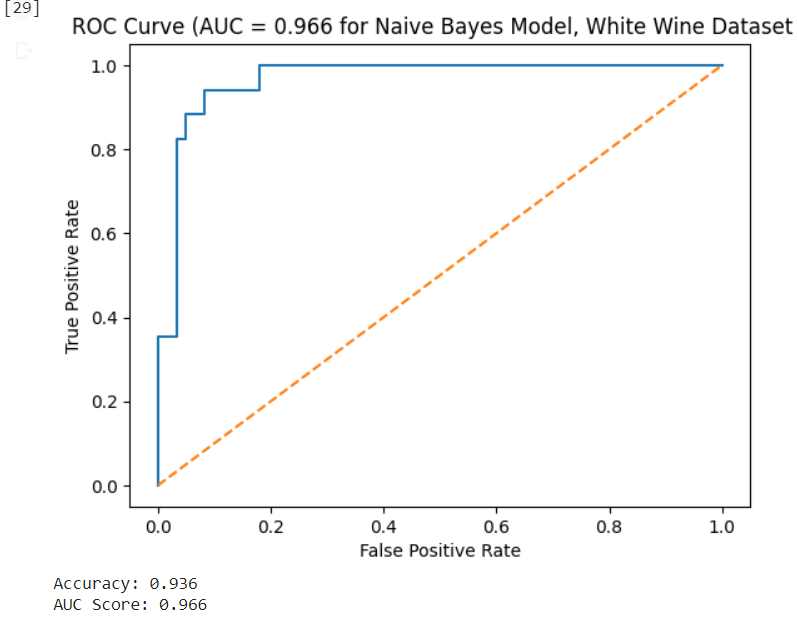
1. Repeating the models in WEKA question 2 in Python provides the following results. The code is present in the Python Notebook link.



1. ROC curve of the Random Forest classifier on red wine dataset



1. The best model obtained from Python Question 2 on red wine dataset is Naïve Bayes Model with an Accuracy of 93.6% and AUC score of 0.96



**Comment:** Since, AUC is a good metric to determine the performance of a model, we can conclude that, Naïve Bayes is the best model with AUC score of 0.96.

1. Supposing that all the models have comparable performance, the following models can be considered by wine-tasting experts.
2. Decision Tree: A decision tree is a model that uses a tree-like graph which recursively splits the dataset into smaller subsets by selecting the feature that gives the most information gain until a stopping criterion is met, resulting in a tree-like model that can be used for prediction. The tree can be easily understood by the wine-tasting experts to understand what attributes are taken into consideration in a hierarchy.
3. Random Forest Model: Random Forest is also similar to a decision tree. It is a supervised algorithm that is made up of multiple decision trees. This can also be easily understood by the wine-tasting experts, and help them understand how the decision is made using decision trees. However, this process is long and requires rigorous training.

**Github link:**

**Python Notebook link:** <https://colab.research.google.com/drive/1kLA5GDIJIsV_1LGxNft1zB-ELA-0n01-#scrollTo=Caf8yq6sa0-F>