Assignment 2

**Python Tasks (50 points)**

1. Read  **red-wine.csv** into Python as a data frame, use a pandas profiling tool (<https://github.com/pandas-profiling/pandas-profiling>) to create an HTML file, and paste a screenshot of the HTML file here (10 points)

Graphical user interface, application

Description automatically generated Chart

Description automatically generated Chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart

Description automatically generated Graphical user interface, table

Description automatically generated

1. Repeat the same experiments in WEKA Question 2,  and report the same metrics as in Question 2. To receive full credit, you will need to write a script to assemble the result as above in the form of Pandas data frame. Paste a screenshot of your result from your Python notebook here. Please make sure that there is a reasonable number of significant digits in reporting your output. (20 points)

Calendar

Description automatically generated with low confidence

1. Plot the ROC curve of the Random Forest classifier from the Python package, and paste a screenshot of your ROC curve here (10 points)

Chart, line chart

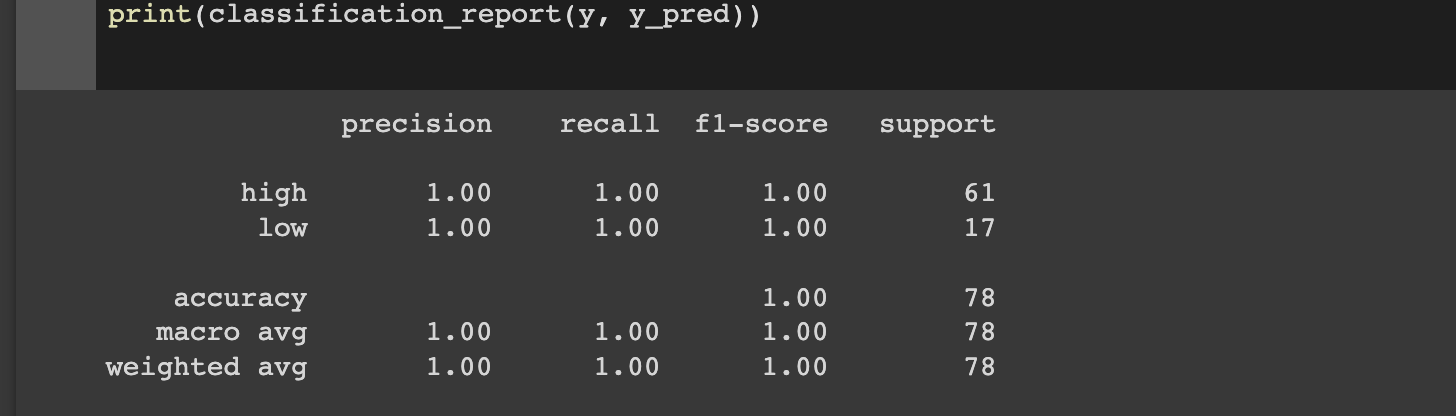
Description automatically generated

1. Using the best model obtained above in Q2 (python)  and running the model on **white-wine.csv** and reporting the AUC score, comment on the performance. (5 points)

The Random Forest classifier performed well on the white wine dataset, as seen by an **AUC value of 0.889**. A score of 0.5 indicates that the classifier is guessing at random, while a score of 1.0 indicates that it is accurately predicting the future.

The AUC value of 0.889 in this instance indicates that the classifier can discriminate between positive and negative samples with a respectably high degree of accuracy. It is crucial to keep in mind, nevertheless, that the model's performance might change based on the particular use case and the analysis's objectives. To gain a deeper understanding of the model's performance, it's also crucial to take additional measures into account, such as accuracy, recall, and F1-score.

1. Suppose all the models have comparable performance, which model would you prefer if the wine-tasting experts would like to gain some insights into the model? Note: there could be multiple model types fitting this criterion. (5 points)

 Diagram

Description automatically generated