Section 1: Lab Assessment based on BreastCancer dataset. The dataset will be provided to the students. You are required to write Python scripts in order to answer each of the questions. Your scripts should be written in Jupyter notebook/Colab. Submit your ipynb notebook.

The test duration is 1 hour. **(20 Marks)**

1) Create a dataframe that contains all the data from BreastCancer.csv

(1 Mark)

2) Check if there is any missing value. If yes, remove the instance.

(1 Mark)

3) Perform a correlation analysis. (Note: Your label is *BreastCancer*).

(1 Mark)

4) Sort the features based on their correlation strength with *BreastCancer* in descending order. Note: You need to consider only the absolute correlation value to disregard the correlation direction.

(2 Marks)

5) Split your data into 70% training and 30% testing. Create a *for* loop to iteratively train a Random Forest Classifier model to predict *BreastCancer* and report the accuracy.

- In the first iteration, use only the top one feature with the highest correlation strength to train the model and report the accuracy.
- In the second iteration, use only the top two features with the highest correlation strength to train the model and report the accuracy.
- In the third iteration, use only the top three features with the highest correlation strength to train the model and report the accuracy.
- Repeat the iteration until you use all the features.

Plot a graph of accuracy vs iteration. Label your axes. Comment on what you have observed.

(15 Marks)