Task 1 (15 marks)

Part 1

In the first dataset(breast cancer dataset, i applied several data preprocessing techniques to make the dataset into the right format. After data loading into my colab environment, i found that the dataset had no column names. I started off by naming all the 10 columns of the dataset for ease of use later in the code.

I also found that the dataset had special characters in the "nodes.caps". I did the cleaning of the column by removing the rows containing the special characters. I opted to remove since they were few rows which could not have impact on model building and training.

After checking the glimpse of the whole dataset, i realized that all the columns which were supposed to be factor type were all in character format. I therefore converted all the columns from character type to factor(categorical type) before building the model.

Up to this point since the dataset was now fine with every part, i was now able to proceed on model building.

In the second dataset(German credit data, i applied several data preprocessing techniques to make the dataset into the right format also. After data loading into my colab environment, i found that the dataset had no column names. I started off by naming all the 10 columns of the dataset for ease of use later in the code.

I also found that the dataset had two columns into one. For example, there was a column containing the status of personality for an individual as well as their sex type. I did the cleaning of the column by separating the column into two one containing the status and another one containing the sex of an individual.

After checking the glimpse of the whole dataset, i realized that some of the columns which were supposed to be factor type were all in character format. I therefore converted all of them from character type to factor(categorical type) before building the model.

Up to this point since this dataset was also now fine with every part, i was now able to proceed on model building.

Task 2 (20 marks)

Part 1.

In matters concerning parameter choice to improve the model accuracy, i decided to use:

- Max_depth() - this parameter was to help me in deploying the models on 100 different test segments.
- Minsplit – to set the minimum number of observations in the node before the algorithm perform a split.
- Minbucket - to set the minimum number of observations in the final note i.e. The leaf

Task 5 (15 marks)

I choose decision tree to PART classifier. This is because after the whole process of parameter tuning, the two classifiers had an improved accuracy however with decision tree leading with a higher accuracy when compared to PART classifier. Decision tree classifier thus has shown higher stability when compared to the PART classifier in both datasets used during this data mining project