# FIT5202: Data Processing for Big Data - Assignment 2 Marking Rubric

| Part A | Excellent | Very Good | Good | Satisfactory |
|---|---|---|---|---|
| **Step 01: Import pyspark and initialize Spark** | Importing and initializing has been done successfully according to the specification. RDDs have been created with 4 cores and the question has been answered correctly. | Importing and initializing has been done with some minor mistakes. RDDs have been created with single core and the question has been answered correctly. | Importing and initializing has been done with some major mistakes. RDDs have been created and the question has been answered correctly. | Importing and initializing has been done with some major mistakes. RDDs have been created and but the question has not been answered. |
| **Step 02:Load the dataset and print the schema and total number of entries** | The dataset is loaded properly and the total number of entries is calculated and displayed. | The dataset is loaded properly and the total number of entries is calculated and displayed but output is wrong. | The dataset is loaded properly and total number of entries is calculated but not displayed. | Only data is loaded but total entries is not calculated. |
| **Part B** | | | | |
| **Step 03: Delete columns from the dataset** | All the columns specified in the question are deleted or delete some of them with proper justificaition.. | Some of the columns are deleted without proper justification. | Only 2 or 3 columns are deleted. | Only 1 column is deleted. |
| **Step 04: Print the number of missing data in each column.** | Define the function to calculated the missing data in each column and displays the correct number. | Calculate all the columns but display the wrong total. | Calculate only some of the columns and display | Calculate only some of the columns and display wrong total. |
| **Step 05: Fill the missing data with average value and maximum occurrence value.** | Fill all the missing data both for numeric value and non-numeric value with correct information | Partially fill the missing data both for numeric value and non-numeric value | Only the numeric values are filled with proper average but the non-numeric values are not properly calculated. | Only numeric or non-numeric values are properly done. |
| **Step 06: Data transformation** | The type casting is done properly for all the specified columns and StringIndexer method is also applied properly. The columns whose Indexing is done, their original columns are droped. | The type casting is done properly for all the specified columns and StringIndexer method is also applied properly. The columns whose Indexing is done, their original columns are not droped. | The type casting is not done properly for all the specified columns but StringIndexer method is applied properly. The columns whose Indexing is done, their original columns are not droped. | The type casting is not done properly for all the specified columns and StringIndexer method is also applied pertially. The columns whose Indexing is done, their original columns are not droped. |
| **Step 07: Create the feature vector and divide the dataset** | The feature vector is properly created and the dataset is split properly (randomsplit). | The feature vector is properly created and the dataset is split but not randomly. | The feature vector is properly created and the dataset is not splited in described ration. | The feature vector or split is not done properly. |
| **Part C** | | | | |
| **Step 08: Apply machine learning classification algorithms on the dataset and compare their accuracy. Plot the accuracy as bar graph.** | All the machine learning algorithms are properly impletement and their accuracy is calculated. The plot is also properly created. | Only subset of machine learning algorithms are properly impletement and their accuracy is calculated. The plot is properly created. | Only subset of machine learning algorithms are properly impletement but their accuracy is not calculated properly. The plot is not properly created. | Only one of machine learning algorithm is properly impletement and its accuracy is calculated properly. The plot is not properly created. |
| **Step 09: Calculate the confusion matrix and find the precision, recall, and F1 score of each classification algorithm. Explain how the accuracy of the predication can be improved?** | The precision, recall and F1 score are calculated properly for all the algorithms. Properly explain the improvement staregies. | The precision, recall and F1 score are calculated properly for all the algorithms. Partially explain the improvement staregies. | The precision, recall and F1 score are calculated properly for some of the algorithms. Partially explain the improvement staregies. | The precision, recall and F1 score are calculated properly for some of the algorithms. No explanation for the improvement. |