Group No

Group Member Names:

1.
2.
3.

1. Business Understanding

Students are expected to identify a classification problem of your choice. You have to detail the Business Understanding part of your problem under this heading which basically addresses the following questions.

1. What is the problem that you are trying to solve?
2. What data do you need to answer the above problem?
3. What are the different sources of data?
4. What kind of analytics task are you performing?

Unsupported Cell Type. Double-Click to inspect/edit the content.

2. Data Acquisition

For the problem identified by you, students have to find the data source themselves which should be a website which has the required data in it. You have to write Python crawler code to scrape data from the respective website rather than downloading ready-made dataset as such from sources like Kaggle etc.

(Data downloaded from website like Kaggle will be awarded negative marks.)

2.1 Code for scraping data from website

```
[ ] ##---------Type the code below this line-----------------##
```

2.2 Code for converting the above scraped data into a dataframe

```
[ ] ##---------Type the code below this line-----------------##
```

2.3 Confirm the data has been correctly by displaying the first 5 and last 5 records.

```
[ ] ##---------Type the code below this line-----------------##
```

2.4 Display the column headings, statistical information, description and statistical summary of the data.

```
[ ] ##---------Type the code below this line-----------------##
```

2.5 Write your observations from the above.

1. Size of the dataset
2. What type of data attributes are there?
3. Is there any null data that has to be cleaned?

Unsupported Cell Type. Double-Click to inspect/edit the content.

3. Data Preparation

3.1 Display how many unique values are present in each attribute

```
[ ] ##---------Type the code below this line-----------------##
```

3.2 Check for the presence of duplicate data, identify the attributes with duplicate data, report the attributes. Mention the method adopted to remove duplicate data if present. Report the results again.

```
[ ] ##---------Type the code below this line-----------------##
```

3.3 Show whether there are any missing values in each attribute. Report the same.

```
[ ] ##---------Type the code below this line-----------------##
```

3.4 Clean the missing data using any imputation technique, mention the method used and again report the change after cleaning the data.

```
[ ] ##---------Type the code below this line-----------------##
```

3.5 Check if all the attributes are following the same format and are consistent. If not, report all such attributes and what inconsistencies are present.

```
[ ] ##---------Type the code below this line-----------------##
```

3.6 Correct the data if there are inconsistencies from 3.5. Report or print the data after correction.

```
[ ] ##---------Type the code below this line-----------------##
```

3.7 Identify the target variables.

```
[ ] ##---------Type the code below this line-----------------##
```

3.8 Separate the data front the target such that the dataset is in the form of (X,y) or (Features, Label)

```
##---------Type the code below this line-----------------##
```

+ Code + Text

3.9 Discretize the target variable or perform one-hot encoding on the target or any other as and if required.

```
[ ] ##---------Type the code below this line-----------------##
```

4. Data Exploration using various plots

4.1 Scatter plot of each attribute with the target.

```
[ ] ##---------Type the code below this line-----------------##
```

4.2 Pair plot of each attribute to identify the linear relationships among the attributes.

```
##---------Type the code below this line-----------------##
```

4.3 Regression plots to identify the linear relationship between each attribute with the target variable.

```
[ ] ##---------Type the code below this line-----------------##
```

4.4 Can any other plot help to identify the optimal set of attributes that can be used for classification. The plot will be based on linear or nonlinear separations. If there is/are such plots, name them, explain why you think they can be helpful in the task and perform the plot as well.

```
[ ] ##---------Type the code below this line-----------------##
```

5. Data Wrangling

+ Code + Text

5.1 Display correlation heatmap of each attribute against the target and report which features are significant.

```
[ ] ##---------Type the code below this line-----------------##
```

5.2 Univariate Filters – Identify top 5 significant features by evaluating each feature independently with respect to the target variable by exploring

1. Mutual Information (Information Gain)
2. Gini index
3. Gain Ratio
4. Chi-Squared test
5. Fisher Score (From the above 5 you are required to use only any **three**)

Write your observations from the results of each method and report the top 5 significant features for each of the above methods. Also plot a graph of significant features for each of them for better visualization.

```
[ ] ##---------Type the code below this line-----------------##
```

5.3 Train a "DecisionTreeClassifier" on the entire data and use the classifier to extract the top 5 significant features. Plot graph of significant features for better visualization.

```
[ ] ##---------Type the code below this line-----------------##
```

5.4 Using "mlxtend" library perform SequentialFeatureSelector to identify top 5 features.

```
[ ] ##---------Type the code below this line-----------------##
```

5.5 Conclude the top 3 significant features with necessary justifications.

----------Type the answers below this line----------