**2. Data Acquisition**

For the problem identified , find an appropriate data set (Your data set must be unique with minimum **20 features and 10k rows**) from any public data source.

**2.1 Download the data directly**

**2.2 Code for converting the above downloaded data into a dataframe**

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

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

**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?

**3. Data Preparation**

If input data is numerical or categorical, do 3.1, 3.2 and 3.4

If input data is text, do 3.3 and 3.4

**3.1 Check for**

* duplicate data
* missing data
* data inconsistencies

**3.2 Apply techniques.**

* to remove duplicate data
* to impute or remove missing data
* to remove data inconsistencies

**3.3 Encode categorical data.**

In [ ]:

**3.4 Text data**

1. Remove special characters
2. Change the case (up-casing and down-casing).
3. Tokenization — process of discretizing words within a document.
4. Filter Stop Words.

**3.4 Report**

Mention and justify the method adopted

* to remove duplicate data, if present
* to impute or remove missing data, if present
* to remove data inconsistencies, if present

OR for textdata

* How many tokens after step 3?
* how may tokens after stop words filtering?

If the any of the above are not present, then also add in the report below.

Score: 2 Marks (based on the dataset you have, the data prepreation you had to do and report typed, marks will be distributed between 3.1, 3.2, 3.3 and 3.4)

**3.5 Identify the target variables.**

* Separate the data from the target such that the dataset is in the form of (X,y) or (Features, Label)
* Discretize / Encode the target variable or perform one-hot encoding on the target or any other as and if required.
* Report the observations

Score: 1 Mark

**4. Data Exploration using various plots**

**4.1 Scatter plot of each quantitative attribute with the target.**

Score: 1 Mark

**4.2 EDA using visuals**

* Use (minimum) 2 plots (pair plot, heat map, correlation plot, regression plot...) to identify the optimal set of attributes that can be used for classification.
* Name them, explain why you think they can be helpful in the task and perform the plot as well. Unless proper justification for the choice of plots given, no credit will be awarded.

Score: 2 Marks

**5. Data Wrangling**

**5.1 Univariate Filters**

**Numerical and Categorical Data**

* 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 **two**)

**For Text data**

1. Stemming / Lemmatization.
2. Forming n-grams and storing them in the document vector.
3. TF-IDF (From the above 2 you are required to use only any **two**)

Score: 3 Marks

**5.2 Report observations**

Write your observations from the results of each method. Clearly justify your choice of the method.

Score 1 mark

**6. Implement Machine Learning Techniques**

Use any 2 ML algorithms

1. Classification -- Decision Tree classifier
2. Clustering -- kmeans
3. Association Analysis
4. Anomaly detection
5. Textual data -- Naive Bayes classifier (not taught in this course)

A clear justification have to be given for why a certain algorithm was chosen to address your problem.

Score: 4 Marks (2 marks each for each algorithm)

**6.1 ML technique 1 + Justification**

​

**6.2 ML technique 2 + Justification**

**7. Conclusion**

Compare the performance of the ML techniques used.

Derive values for preformance study metrics like accuracy, precision, recall, F1 Score, AUC-ROC etc to compare the ML algos and plot them. A proper comparision based on different metrics should be done and not just accuracy alone, only then the comparision becomes authentic. You may use Confusion matrix, classification report, Word cloud etc as per the requirement of your application/problem.

Score 1 Mark

**8. Solution**

What is the solution that is proposed to solve the business problem discussed in Section 1. Also share your learnings while working through solving the problem in terms of challenges, observations, decisions made etc.

Score 2 Marks

##NOTE All Late Submissions will incur a penalty of -2 marks. Do ensure on time submission to avoid penalty.

Good Luck!!!