**Part 1- Project Report**

Find a data set from the following online portals:

1. https://data.sfgov.org/Public-Safety/Traffic-Crashes-Resulting-in-Injury/ubvf-ztfx

You should follow CRISP-DM while working on the data. Data Understanding and Preparation is very important.

**Data Understanding**

As part of your data understanding, you should describe and explore data and verify its quality.

**Data Preparation**

Once you understand the data, you can select, clean, construct, integrate, and format data. After data preparation, you should have the original prepared data, normalized data, standardized data, dataset reduced by **PCA, and dataset reduced by LDA (**5 datasets).

**Modeling**

Outlier detection: Detect outliers using **LOF, ISF and OCSVM** methods.

Classification: Perform classification using **kNN, DT, RF, Naïve Bayes, SVM and MLP**.

Clustering: Perform clustering **using kMeans, DBSCAB, EM, and Agglomerative clustering**.

Resampling: Apply undersampling, oversampling and SMOTE techniques to create new datasets. Then perform classification again to see how your results change.

You need to include nice visualizations using Tableau/PowerBI. (You can leave this part out).

You have at least 8 sets of data – original data, normalized data, standardized data, dataset reduced by PCA, and dataset reduced by LDA, undersampled data, oversampled data, SMOTE data, etc.

So, it is better to write a function that has all models defined in it.

**Discussion of Results**

Wherever possible, combine your models and results. Also, in detail, discuss your findings and results. When you report classification results, include accuracy, precision, recall and F1-score.

You should have 5 main sections – **Data Understanding, Data Preparation, Modeling, Discussion of results, and conclusion**. The report should have a cover page, table of contents, tables, pictures, etc., introduction, conclusion, and references.