**Feature Engineering – Mini Project - Submissions from Group ()**

**Activity 1: Data Understanding and Preparation along with EDA**

Approach: File Processing and Data Pre-Processing

Please refer file named # **FE\_Project\_Groupname\_.ipynb** for activities performed as part of Data understanding and Preparation along with EDA

Step1: We have converted the given .arff file into .csv file format before doing the pre-processing work.

Step2: Missing Values:

We have identified the number of missing values (if any) by comparing the number of records in each column and found missing values for Columns: Accident\_victim (2 missing value), Incident\_involving\_dangerous\_freight (2 missing value), Manifestations (1 missing value), Defect\_in\_the\_network\_of\_trolleybuses (2 missing value), Semaphore\_off (1 missing value)

Step3: We have visualized the distribution for above mentioned columns that have missing data.

Step4: Impute missing values: After visualizing, we have imputed these missing values with appropriate values as under:

Accident\_victim (2 missing value): Has missing values for 12:00 and 19:30 Hours respectively. This has been replaced with 0 and 1 respectively, as for 12:00 Hour, other values are also 0 while for 19:30 hour with same Broken\_Truck combination as 4, it has 1.

Incident\_involving\_dangerous\_freight (2 missing value): Has missing value for 13:30 and 15:30 Hours. Both these blanks are replaced with 0 as during these hours, it has all other value 0 only.

Manifestations (1 missing value): Has one negative value (-1) at 13:00 Hour. This has been replaced with 0 as all other values for this column at this timestamp is 0.

Defect\_in\_the\_network\_of\_trolleybuses (2 missing value): It has 2 blank values at 9:00 and 18:30 Hour respectively along with one negative value (-8) at timestamp 7:00 Hour. These values are replaced with average value for the respective timestamps.

Semaphore\_off (1 missing value): Has one missing value at 19:00 Hour. This is also replaced with average value as 1.

Step5: Check for Outliers and Noise in data: In order to scale the given data, we have identified the outlier value at timestamp 12:30 for Broken\_Truck and at timestamp 15:00 for Point\_of\_flooding and removed these with the average value of respective hour/timestamp

Broken\_Truck: Contains outlier value as 1000 for 12:30 Hour and replaced with the average.

Point\_of\_flooding: Contains outlier value as 2000 for 15:00 Hour and replaced with average.

Identification of the factors causing the traffic slowness: