**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.ipynb** for activities performed as part of Data understanding and Preparation along with Exploratory Data Analysis.

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

Step2: As part of data pre-processing work, we have converted the categorical value to numerical value i.e. for Column ‘Fire\_vehicles’ , we have converted ‘YES’ to 1 and ‘NO’ to 0. Similarly, for the Column 'Vehicle\_excess', we have replaced 'T' with 1, 't' with 1, 'F' with 0 and 'f' with 0

We have considered every 30 minutes interval starting from 7:00 as interval 1, 7:30 as interval 2, 8:00 as interval 3 and so on and converted the Column ‘Hour’ value accordingly in the code as continuous value, so we *have some 26 continuous interval* from ‘Hour’ perspective for the given 5 days from Monday to Friday between December 14, 2009 to December 18, 2009.

Step3: Missing Values:

We have identified the number of missing values (*along with noise in data*) by comparing the number of records in each column and found missing values for Columns: Accident\_victim (2 missing value), Running\_over (1 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).

Kindly **note**, we have passed **["n/a", "na", "-1", " "]** as missing value in the code by considering these as *list of values* and then replaced such values appropriately.

Step4: We have visualized the distribution for above mentioned columns that have missing data.

Step5: 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.

Step6: Check for Outliers: In order to scale the given data, we have checked for outliers by using Boxplot in the code for each column and then removed those outlier values. Such as, 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 rows.

Broken\_Truck: Contains outlier value as 1000 for 12:30 Hour and removed the entire row corresponding to this outlier value.

Point\_of\_flooding: Contains outlier value as 2000 for 15:00 Hour and removed the entire row corresponding to this outlier value.

Step7: Evaluating the inferences based the Column values before applying / implementing PCA

In-Progress….

Identification of the factors causing the traffic slowness: