**Feature Engineering – Project Submission from Group-ID 25: 2019AIML083, 2019AIML066, 2019AIML016**

**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: 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 ["-1", " ", ""] as missing value in the code by considering these as list of values and then replaced such values appropriately.

Post identification of missing values, we have removed spaces (wherever applicable) from the file.

Step3: Evaluating the non-contributing features/attributes based the Column values before applying PCA

Before implementing Principal Component Analysis, we have removed/dropped the following 5 Columns, considering they are non-contributing / non-influencing factors. Please see below details against each column reasoning for removal:

1. Fire\_vehicles: This is dropped as it has only 1 occurrence with Yes (i.e. 1), so not a contributing attribute/feature.
2. Occurrence\_involving\_freight: This also has only 1 occurrence having 1, so not a contributing attribute/feature.
3. Incident\_involving\_dangerous\_freight: This also has only 1 occurrence having 1, so not a contributing attribute/feature.
4. Fire: This also has only 1 occurrence having 1, so not a contributing attribute/feature.
5. Intermittent\_Semaphore: This also has only 1 occurrence having 1, so not a contributing attribute/feature.

Now, we are left with 12 Columns on which we will apply PCA to further reduce dimensions/features towards model building activity.

Step4: As part of data pre-processing work, we have converted the categorical value to numerical/continuous value i.e. for the Column 'Vehicle\_excess', we have replaced 'T' with 1, 't' with 1, 'F' with 0 and 'f' with 0

Similarly, we have considered every 30 minutes slots 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 27 slots/continuous interval* from ‘Hour’ perspective for the given 5 days from Monday to Friday between December 14, 2009 to December 18, 2009.

Note: The column ‘Fire\_vehicles’ is not required to be converted as we have already dropped this column.

Step5: Impute missing values: We have identified the missing value counts (wherever applicable) and imputed these missing values with appropriate values as under:

1. Accident\_victim (2 missing value): Has missing values for 12:00 and 19:30 Hours respectively. These values have been replaced with floor of mean.
2. Running\_over (1 missing value): Has been replaced with 0.
3. Vehicle\_excess: 1 value is replaced with 0.
4. Manifestations (1 missing value): Has one negative value (-1) at 13:00 Hour. This has been replaced with 0.
5. 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 0.
6. Semaphore\_off (1 missing value): Has one missing value at 19:00 Hour. This is also replaced with 0.
7. Note: For columns ‘Incident\_involving\_dangerous\_freight’ & ‘Fire\_vehicles’, missing value imputation are not required as these columns were already dropped.

Step6: We have removed duplicates as part of next step, before applying PCA.

Step7: Check for Outliers: In order to scale the given data, we have checked for outliers by using Boxplot visualization in the code for relevant columns 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.

Now, we are left with 132 record sets along with 13 Columns.

Step8: Feature Selection: In order to perform feature selection, we have considered ‘Slowness\_in\_traffic\_percent’ as response variable y, so we have 12

And, split the data into train and test model as 70:30 ratio.

We have Identified features that have high co-relation with respect to target variable (Slowness\_in\_traffic\_percent) using heatmap and correlation matrix. Check co-linearity among the features and choose the feature with high information gain.

Step 9: Now, we have normalized the features, applied PCA to convert it into 5 new components and printed PCA transformed data.

Then, we have created the regression model as:

Coefficient = Theta1 = [ 1.59570782 -1.25875204 0.30672479 0.88592319 0.21111515]

Intercept = Theta0 = 10.030434782608696

Thus, we have developed Model 1 as under:

Y = 10.030 + 1.596X1 -1.26X2 + 0.31X3 + 0.886X4 + 0.21X5

Identification of the factors causing the traffic slowness: