# Sangam 2019 - ML Hackathon by IITMAA

## Approach:

We are given information about a subset of the traffic volume dataset and asked to build a predictive model that predicts the traffic volume for given time duration and climate conditions.We are given 14 basic explanatory variables, including date_time, is_holiday, weather_type, among the others. This is a regression problem, and we will be implementing a random forest regressor.

1.Preprocessing the training dataset.

2.Feature engineering on training dataset

3.Splitting the training data into testing and training.

4.Normalizing the values(Standard scaler)

5.Training the model using random forest regressor.

6.Preprocessing the testing dataset.

7.Feature engineering on testing dataset

8.Normalizing testing dataset values.

9.Prediction on test dataset.

## Preprocessing:

1.)**Duplicate data** in the given data set were discarded to make sure the predictive model was training correctly on given data.

2.)The data set was checked for any kind of **null,Nan or infinity** values and corrective measures were taken accordingly.

1.) Various **outliers** in the dataset like the ones in the rain_p_h,snow_p_h and temperature columns which were causing the predictive model to waver were discarded .

2.)The **date_time** feature which was a object feature was firstly converted to date time feature so that extracting parameters related to date like the **year,month,day** and parameters related to time like **hour** can be easily done and also parameters like **week of year, day of week and day of year** were extracted

4.)The data columns like **air_pollution,wind_direction,dew_point** were **dropped** as their significance in predicting the traffic volume according to the correlation matrix was noted to be least.

5.)The string data in the **weather_type** and **weather_description** were then **label encode** to fit them into the training model.

6.) In the dataset the **holidays** were mentioned only for a particular row and not for whole day, hence by using a **mapping function** we gave the holiday value to the complete day

7.)Since all the data related to **date_time** feature was extracted into multiple columns it was **dropped** then.

8.)The string datatype of i**s_holiday** was encoded simply by traversing through the dataset and assigning values accordingly.

## Feature Engineering:

1.Two new features were created like the **is_weekend** feature and **peak feature(peak_time)**

**Is_weekend:**this feature gives information whether the given day is weekend or no. The weights given in this feature depends on which **day of week** it is, the weights were decide by plotting various graphs using the matplotlib.pyplot libraries.

**peak**:this feature gives information whether the given hour is peak time of traffic or no. The weights given in this feature depends on which **time of the day** it is, the weights were decided by plotting various graphs using the matplotlib.pyplot libraries.

## Tools Used:

Numpy : We used Numpy library for computing scientific/mathematical data like Numerical Analysis, Linear algebra, Matrix computations.

Pandas: We used Pandas because it provides fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive.

Matplotlib:Matplotlibis a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

Seaborn:Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing informative statistical graphics.