**Machine Learning Project**

**Submitted by-**

1. Rishabh Gautam
2. Sayan Das

3. Tanisha Banik

4. Aisika Roy

Description

In this project, we were asked to experiment with a real world dataset, and to explore how machine learning algorithms can be used to find the patterns in data. We were expected to gain experience using a machine learning library, and were expected to submit a report about the dataset and the algorithms used. After performing the required tasks on a dataset of our choice, herein lies our final report.

The dataset used is a sample of air qualities in different states of India. The dataset that was used for this project is collected from UCI Repository and has the following feature vectors:

1. State
2. City
3. Location
4. Type: Type of area
5. Number of monitoring days: Number of days observed
6. Respirable Suspended Particulate Matter(RSPM)- Annual average (µg/m3): RSPM is that **concentration of particulate matter** which is readily inhaled by humans through their respiratory system and in general, considered as particulate matter with their diameter (aerodynamic) less than 2.5 micrometers. Larger particles would be filtered in the nasal duct.
7. Standard Deviation:
8. Percentage- exceedence(24 hourly): the percentage by which the concentration of air pollutants exceeds the corresponding values every 24 hours
9. Air Quality: corresponding air qualities (low, moderate, high and critical)

Since, our dataset is multiclass we are using random forest here as it can handle very large number of features and it is also capable of using continuous and categorical data, robust against overfitting and good at dealing with outliers in training data.

METHODOLOGIES

1. Collect and prepare the data
2. Visualise the data
3. Import required libraries
4. Import datasets
5. Preprocessing the data
6. Encoding the missing values
7. Dummy variable trap
8. Feature scaling
9. Splitting the dataset into training and test set
10. Creating the appropriate model
11. Fitting the model to the training set
12. Predicting test set results

**Conclusion:**

We conclude that the dataset is not a complete set, and there are still other feature vectors missing from it. What we were attempting to generalize is a subset of the actual input set, where the other dimensions are not known, and hence none of the classifiers were able to do better (without overfitting) than 66.17% (Random Forest Classifier). In the future, if similar studies are conducted to generate the dataset used in this report, more feature vectors need to be calculated so that the classifiers can form a better idea of the problem at hand.