School of Engineering and Applied Science (SEAS), Ahmedabad University

B.Tech (CSE Semester VI):
Machine Learning (CSE 523)

Project Abstract Submission #1

Submission Deadline: January 31, 2020 (11:59 PM)

- **Group No.: 21**
- **Project Area: Environment And Climate Change**
- **Project Title: Air Quality Index Prediction**
- **Name of the group members :**

1. Vishal Saha (AU1741004)
2. Rushil Shah (AU1741009)
3. Muskan Matwani (AU1741027)
4. Mohit Vaswani (AU1741039)

Abstract

Background and Motivation
In today's time, urbanization, rapid growth and improved lifestyle has led to an increase in air pollution considerably. It is one of the major concerns for the developing countries like India which is considered as one of the most polluted countries in the world. It is a leading risk factor for many health problems such as asthma, throat diseases, lung cancer, and other respiratory problems. Air pollution is due to particulate matter, sulphur dioxide (SO2), ground-level ozone (O3), nitrogen dioxide(NO2) and carbon monoxide(CO) suspended in the air. As a result of air pollution and ozone damage, there has been a huge increase in the level of global warming and irregular changes in climate. Hence, Precise air quality prediction and appropriate action plans to reduce air pollution, is our prime goal. Numerous machine learning techniques have been adopted that help us predict the air quality accurately.

Brief of Base Article
Air pollution is one of the major environmental concerns in a smart city environment. There has been research that came up with linear statistical techniques for the prediction of the air pollution. However, they were not suitable enough to gain good accuracy due to the complexity of the time series data. Hence, over the years, machine learning approaches have been developed for handling such complex methods. In this article [1], the authors have performed air quality prediction with four different regression techniques namely Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and ANN Multi-layer perceptron Regression. These techniques are used to predict the value of the pollutant PM2.5 and helps in calculating the AQI (known as Air Quality Index) based on the concentration of this pollutant. The dataset consists of various meteorological variables and PM2.5 recorded from 5 cities of China which include Guangzhou, Chengdu, Beijing, Shanghai and Shenyang. The four regression techniques are evaluated on 2 metrics- MAE (Mean Absolute Error) and RMSE (Root Mean Square Error). It has also analyzed the processing time by fitting the hyper parameter tuning on Apache Spark to evaluate the best suited model having least error rate and less processing time.

Project Motives

The objective of this project is to apply four main techniques which are considered as the pillars of machine learning, namely - Regression, Classification, Dimensionality Reduction and Density Estimation. These methods will be applied to a dataset that includes data from different states of India [2]. It consists of parameters like SO2, NO2, RSPM, SPM, PM2.5, and other meteorological parameters useful for air quality prediction. We will first preprocess and clean the dataset so as to train our model well. This model aims to use existing regression techniques to predict the value of the pollutant's concentration and then use it to calculate the AQI (Air Quality Index) using Environmental Protection Agency (EPA) method. Moreover, Classification techniques helps in classifying the air quality into different categories which include good, moderate, unhealthy, etc. It is categorized based on the predicted AQI which acts as a good measure to analyze the air pollution in a given area. Furthermore, We can apply dimensionality reduction techniques like PCA that is used to account new set of orthogonal parameters which are computed from linear combination of original set of parameters. It basically classifies the variables to obtain parsimonious prediction model which only depends on a few necessary variables and helps to achieve the same accuracy with the reduced set of parameters. Moreover, Density Estimation is used to find an estimate probability density function. This can be incorporated with machine learning to produce better estimates. At last, we aim to train our model with dataset consisting of different states data using these different algorithms and compare these techniques by analyzing estimation results on several evaluation metrics.

References

[1] Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, Saif ul Islam, Muhammad Nabeed Asghar. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access.

[2] Bhargava, S. (2017, July 22). India Air Quality Data. Retrieved from https://www.kaggle.com/shrutibhargava94/india-air-quality-data