PROJECT SYNOPSIS

ANTHROPOGENIC AIR PREDICTION USING MACHINE LEARNING

(Tentative project Name)

Submitted towards the partial fulfillment of the criteria for award of Post Graduate In Data Analytics by Imarticus

Submitted By:

HARIHARASUDHAN P (IL031248)

Course and Batch: Post Graduate in Data Analytics & Feb-2022(PGA-22)

Scope & Objective:

Predicting air quality is necessary step to be taken by government as it is becoming the major concern among the health of human beings. Air quality Index measure the quality of air. Various air pollutants causing air pollution are Carbon dioxide, Nitrogen dioxide, carbon monoxide etc.. that are released from burning of natural gas, coal and wood, industries, vehicles etc. Air Pollution can cause severe disease like lungs cancer, brain disease and even lead to death. Machine learning algorithms helps in determining the air quality index. Various research is being done in this field but still results are still not accurate. Dataset are available from Kaggle, air quality monitoring sites and divided into two Training and Testing. With help of Machine Learning algorithms.

Business Problem Statement:

An air Quality Index (AQI) value of 210 to 300 translates to very unhealthy air quality conditions for survival, with high levels of health concern. This World Meteorological Day, the government wishes to monitor the air pollution index in your city so that they can take action accordingly. Due to an increase in the number of vehicles in your city, the air pollution level is increasing and is consequently affecting nature. My task is to predict the air pollution index based on the historical data provided to help the government to administer the same.

AQI Values	Levels of Health Concern	Colors
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

Here I will be using the both Regression and Classification. As it is best approaches for analysis.

Data Sources:

1. The repository that we used for dataset is Kaggle.
2. There are csv files which contains 4,35742 instances and 13 attributes.
3. The data set from Kaggle repository contains attributes of each state details including their State_code, Sampling_date, State, Location, Agency,Type, So2, No2,Rspm,Spm,Location_monitoring_station,Pm2_5, and Date. This dataset is taken as the training data.

Analytics Tools:

- **Jupyter Notebook:** collaborative work capabilities.
- **Pandas:** A python data analysis library enhancing analytics and modelling.
- **Matplotlib, Seaborn:** A python machine learning library for quality visualizations.
- **Statsmodels:** A python that enables us to estimate and analyze various statistical Models.

Analytics Approach:

- To create algorithms to comparative study among the proposed technique and check the performance of the model with statistical.
- Air quality evaluation has been conducted using conventional approaches in all these years. These approaches involve manual collection and assessment of raw data. The traditional approaches for air quality prediction use mathematical and statistical techniques. In these techniques, initially a physical model is designed and data is coded with mathematical equations.

KPIs,Timelines,Milestones(proposed):

KPIs:

For Regression:

- R-Square
- Adjusted R-Square
- Mean Square Error(MSE)
- Root Mean Square Error(RMSE)
- Mean Absolute Error(MAE)
- Score

For Classification:

- Accuracy
- Precision
- Recall
- F1-Score
- Cohen's Kappa

Timelines:

Expected to complete the project by 15-08-2022

Milestone:

Expected processing and making prediction based on the dataset achieve above 90%.

File Format:

The datasets are provided in CSV format, with the following features:

Retrieved from kaggle:

1) State_code
2) Sampling_date
3) State
4) Location
5) Agency
6) Type
7) So2
8) No2
9) Rspm
10) Spm
11) Location_monitoring_station
12) Pm2_5
13) Date