DAC_PHASE5

Project:

Air Quality Analysis

Problem Statement:

The project aims to analyze and visualize air quality data from monitoring stations in Tamil Nadu. The objective is to gain insights into air pollution trends, identify areas with high pollution levels, and develop a predictive model to estimate RSPM/PM10 levels based on SO2 and NO2 levels. This project involves defining objectives, designing the analysis approach, selecting visualization techniques, and creating a predictive model using Python and relevant libraries.

Approach:

Design Thinking:

Project Objectives: Define objectives such as analyzing air quality trends, identifying pollution hotspots, and building a predictive model for RSPM/PM10 levels.

Analysis Approach: Plan the steps to load, preprocess, analyze, and visualize the air quality data.

Visualization Selection: Determine visualization techniques (e.g., line charts, heatmaps) to effectively represent air quality trends and pollution levels.

Checking the requirements:

First we install the required software and install the modules required for the project and then we set up the test environment by changing the path variables and we launch the application.

Data collection and warehousing:

We collect various data in the form of an excel spreadsheet and convert it into a CSV file and save it in the same folder where we are going to implement the algorithm.

We have several data on:

Stn Code : Contains pincode of the city

Sampling Date : contains date of sampling

State : contains the name of the state

City/Town/Village/Area : contains the name of the City/Town/Village/Area

Location of Monitoring Station : contains the place where the location of monitoring station is located.

Agency : contains the state/central pollution control board details

Type of Location : states whether the area is industrial/rural.

SO2 : Sulphur di oxide content

NO2 : Nitrogen di oxide content

RSPM/PM : Respirable Suspended Particulate Matter measured.
PM2.5 : It represents the value of particulate matter measured.

Approach for making design:

Data Mining:

Data mining or Knowledge Discovery (KD) is used to read and analyze large datasets and then finding/extracting patterns from the data. It is used for predicting the future trends or forecast patterns over a period. Data mining algorithms are usually based on wellknown mathematical algorithms and techniques. There are two types of data mining learning algorithms: 1) Supervised algorithms and 2) Unsupervised algorithms.

We are going to make optimal use of these to train our machine learning model for better prediction. The dataset is provided in the Government website.

Dataset link : https://tn.data.gov.in/resource/location-wise-daily-ambient-air-quality-tamil-nadu-year-2014

Unsupervised learning algorithm:

The Unsupervised algorithm is the process in which the training dataset contains only the input set and not the corresponding target vectors. The main criterion is to find groups or patterns of similar examples within the dataset, called as clustering.

Supervised learning algorithm:

The Supervised algorithm is the process in which the training data comprises of both the training and the corresponding output target vectors. In this project, a supervised learning algorithm called Artificial Neural Network (ANN) has been used for training, validation and testing the dataset. In addition, to the ANN, a Multiple Linear Regression (MLR) model has been used for comparing the performance against the ANN. The below section introduces the processes of Artificial Neural Network (ANN) and Multiple Linear Regression (MLR).

Test execution:

We use the pandas,scikit,matplotlib modules in python in order to implement the supervised machine learning algorithm and to visualize it in a realistic manner.

Conclusion:

These steps are considered optimal for getting the desired output.