AIR QUALITY ANALYSIS AND PREDICTION

INTRODUCTION:

An "Air Quality Analysis" project is an essential undertaking that focuses on the assessment, monitoring, and improvement of the quality of the air we breathe. It plays a pivotal role in safeguarding public health, preserving the environment, and promoting sustainable urban development. This introductory section will provide an overview of the project's purpose, significance, and objectives.

DATA SOURCES:

Data has been collected from the field instruments, primary data includes air quality parameters like particulate matters (PM2.5 and PM10), gaseous pollutants and meteorological data, air quality indexes and quality control data.

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

External data sources such as Meteorological Data, Traffic and Transportation Data, Industrial Emission Data, Geospatial Data, Population and Demographic Data, Health Data, Air Quality Regulations, Satellite and Remote Sensing Data, Historical Air Quality Data and Environmental Events Data.

Rigorous data cleaning and preprocessing are essential to ensure data accuracy. Steps include handling missing data, standardizing formats, and addressing outliers.

COLUMNS USED:

The columns used in this air quality analysis project from the Location wise daily Ambient Air Quality of Tamil Nadu for the year 2014 dataset are SO2, NO2 and RSPM/PM10.

SO2: SO2 in air quality analysis datasets is crucial for assessing air quality, regulatory compliance, understanding environmental and health impacts, and taking steps to mitigate air pollution. It is one of the key pollutants routinely monitored in air quality studies and is an important component in efforts to maintain and improve air quality.

NO2: NO2 data in air quality analysis is used to identify the health impacts, Air quality regulation, Temporal and spatial variability, Data correlation and Public Awareness.

RSPM/PM10: RSPM (Respirable Suspended Particulate Matter) and PM10 (Particulate Matter with a diameter of 10 micrometers or less) are used for analysis of environmental concerns, regulatory compliance, Air Quality Index, Source identification and Research and Epidemiology.

LIBRARIES USED:

To work with location-wise daily ambient air quality data for Tamil Nadu in the year 2014, you'll likely need to use several libraries and tools for data manipulation, analysis, and visualization.

1. **Pandas:** Pandas is a powerful library for data manipulation and analysis. It's particularly useful for handling and exploring tabular data.

   Install Pandas: You can install Pandas using pip with the following command: pip install pandas.

2. **Numpy:** NumPy is a library for numerical and array operations in Python. It's essential for working with multi-dimensional data arrays.

   Install NumPy: You can install NumPy using pip with the following command: pip install numpy.

3. **Matplotlib and Seaborn:** Matplotlib and Seaborn are libraries for creating static, animated, or interactive visualizations to help you understand the data.

   Install Matplotlib and Seaborn: You can install both libraries using pip: pip install matplotlib seaborn.

4. **Scikit-Learn:** Scikit-Learn is a machine learning library that includes tools for data preprocessing, modeling, and evaluation. You may use it for predictive modeling in your project.

   Install Scikit-Learn: You can install Scikit-Learn using pip: pip install scikit-learn.

TRAIN AND TEST:

Data Cleaning:

Dealing with missing values, which may involve imputation or removal of incomplete records. Identifying and addressing outliers that can skew the analysis. Ensuring data consistency by addressing discrepancies or errors in data entries. Use visualizations such as line charts, heat maps, and geographic maps to uncover trends, correlations, and disparities in vaccine distribution and effectiveness.

Data Collection and Preprocessing:

Historical air quality data, including various pollutant concentrations (e.g., PM2.5, PM10, NO2, CO, etc.), meteorological data, and other relevant features. Preprocess the data by handling missing values, outliers, and normalizing or standardizing the features as necessary.

Data Splitting:

The dataset is split into two subsets: a training set and a testing set. A common split is 80% for training and 20% for testing, but the split ratio is based on the dataset's size and characteristics. It can be done using train_test_split model.

Model Selection:

Common machine learning and deep learning models include linear regression, decision trees and long short term memory can be used. The choice of the model depends on the complexity of the problem and the dataset.

Model Training:

The training dataset to train the selected model. The model learns the relationships between the input features and the target variable, which, in this case, is SO2, NO2 and RSPM/PM10.

CODE LINK:

https://colab.research.google.com/drive/1en3zjIpjibbbepN3TsVehtSWz9BskKO6

ACCURACY CHECK:

Model Evaluation:

After training the model, its performance is evaluated on the testing dataset. Common evaluation metrics for air quality prediction include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) for regression models.

Hyper-parameter Tuning:

There is a need of fine-tune hyper parameters to optimize its performance. Techniques like grid search or random search can help find the best combination of hyper parameters.

CONCLUSION:

In this air quality analysis project conducted in Tamil Nadu, we aimed to assess the ambient air quality in the region during the year 2014 to understand the distribution of air pollutants, their sources, and their impact on public health and the environment.