Project Definition and Design Thinking for Air Q Assessment TN

Project Definition:

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.

Design Thinking:

Objective:

The project aims to analyze and visualize air quality data collected from monitoring stations across Tamil Nadu. The primary objectives are:

Gain Insights into Air Pollution Trends: Understand the historical patterns and trends of air pollution in Tamil Nadu.
Identify High Pollution Areas: Identify specific regions within Tamil Nadu that consistently experience high levels of air pollution.
Develop Predictive Model for RSPM/PM10: Create a machine learning model that can estimate RSPM/PM10 levels based on the concentrations of SO2 and NO2.

Data Sources:

Location wise Ambient air quality of TamilNadu for the year 2014.

The dataset has columns such as station code,sampling date,city/town/village,agency,type of location and levels of SO2, NO2, RSPM, PM 2.5 which enables us to measure the air quality in different regions in TamilNadu.

Analysis Approach:

Data Collection and Loading: Obtain air quality data from monitoring stations in Tamil Nadu.

Load the data into a structured format (e.g., CSV, Excel, or database) using Python's Pandas library.
Data Exploration:

Conduct an initial exploration of the data to understand its structure and content.
Check for missing values, outliers, and data types.

Data Preprocessing:

Handle Missing Values:
Imputing or removing missing values using appropriate techniques (e.g., mean imputation, interpolation).

Handle Outliers:
Identifying and addressing outliers through methods like Winsorization or removing extreme values.

Data Normalization (if needed):
Scaling numerical features to a similar range.

Feature Engineering (if applicable):
Creating new features that may enhance predictive performance (e.g., deriving additional air quality indices).

Time Series Analysis (Optional):
If the data is time-stamped, performing time series analysis to identify trends, seasonality, and cyclical patterns.

Exploratory Data Analysis (EDA):
Generating summary statistics and visualizations to gain insights:
Histograms, box plots, and descriptive statistics for each air quality parameter.
Correlation matrix to identify relationships between variables.
Identifying High Pollution Areas:
Using geographical data visualization techniques (e.g., maps) to pinpoint regions with consistently high pollution levels.

Feature Selection (if needed):

Selecting the most relevant features (SO2 and NO2 levels) for building the predictive model.

Model Development (Predictive Modeling):
Choosing a regression model to predict RSPM/PM10 levels based on SO2 and NO2 levels.
Splitting the data into training and testing sets for model evaluation.
Training the model on the training set.

Model Evaluation:
Assessing the model's performance using appropriate metrics (e.g., Mean Absolute Error, Mean Squared Error, R-squared).
Adjusting hyperparameters if necessary.

Data Visualization:

Time Series Plots:
Using line charts for displaying changes in air quality parameters (e.g., RSPM/PM10, SO2, NO2) over time, allowing for the identification of trends, seasonality, and anomalies.

Heatmaps:
Displaying pollution levels across different monitoring stations on a geographical map. This can effectively identify areas with consistently high pollution levels.

Box Plots:
Utilizing box plots to visualize the distribution of pollution levels for each pollutant. This aids in understanding the central tendency, spread, and potential outliers.

Scatter Plots:
Displaying the relationships between different pollutants. For example, plotting SO2 levels against NO2 levels to observe any correlations.

Bar Charts:
Utilizing bar charts to compare pollution levels between different monitoring stations or different time periods.