Phase 1: Problem Definition and Design Thinking

Project 9: Air Quality Analysis and Prediction in Tamil Nadu

PROBLEM DEFINITON:

The objective of this project is to employ data science techniques to analyze air quality in various regions of Tamil Nadu, India. This analysis aims to provide insights into air pollution levels, identify key contributors, and develop predictive models for air quality. The project will use historical and real-time data sources to create actionable recommendations for improving air quality in the region.

PROJECT OBJECTIVES :

Aim :

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.

Project Team :

The project is being developed by the team to make an analysis on air quality.

Justification :

Air quality data helps us determine if an area is meeting the air quality standards devised by CPCB, WHO. The data collected from air quality monitoring would primarily help us identify polluted areas, the level of pollution and air quality level.

Purpose :

The main purpose to build this project is to make an analysis on air quality and identify the over and less polluted areas in order to maintain an polluted free environment.

Outcome :

By executing this project, we aim to provide valuable insights into air quality in Tamil Nadu, promote data-driven decision-making, and contribute to efforts to combat air pollution and improve the overall well-being of the region's residents

DESIGN THINKING:

Gather relevant data sources. In Tamil Nadu, this might include air quality monitoring stations, weather data, satellite imagery, and pollution sources data.

Data Collection :

Gather historical and real-time air quality data from government agencies, environmental sensors, and satellite sources. This data should identifying pollution hotspots, and building a predictive model for RSPM/PM10 levels and cover various pollutants such NO2, SO2  .

Pollutant Analysis:

Analyze the concentrations of major air pollutants including SO2,NO2,O3.

Geographic Variation :

Identify regions with highest levels of air pollution and understand the factors contributing to this variation.

Building Predictive Model :

The dataset will be using from <https://tn.data.gov.in/resource/location-wise-daily-ambient-air-quality-tamil-nadu-year-2014>

* First and foremost, import the necessary python libraries like matplotlib, scikit, seaborn etc.
* Explore the data set. Python has several [functions](https://365datascience.com/tutorials/python-tutorials/python-functions/) that will help you with your explorations.
* Feature the selections that contribute more to dataset.
* Select only those that have the strongest relationship with the predicted variable.
* To build model, train and split the dataset and the testing is done using regression techniques.

Data Cleaning:

Clean and preprocess the data to remove any outliers or inconsistencies. Ensure that the data is in a format suitable for visualization.

Data Preprocessing

Clean and preprocess the collected data to handle missing values,outliers, and ensure data consistency..

Choose Visualization Tools :

Select appropriate data visualization tools and libraries. Popular options include Python libraries like Matplotlib, Seaborne etc.

Select Visualization Types:

Consider time series plots or line charts.

To compare air quality across different locations, use bar charts or heat maps.

For geographical distribution, create maps with color-coded markers.

Create Visualizations:

Use bar charts to compare air quality indices across cities or regions in Tamil Nadu.

Generate heatmaps to visualize variations in air quality across different times and locations.

Include labels, legends, and annotations to make visualizations easy to understand

Color Coding:

Use appropriate color schemes, such as green for good air quality and red for poor air quality, to make the visualizations intuitive.

Data Splitting:

Split the dataset into training, validation, and test sets to evaluate model performance effectively.

Model Training:

Train your selected models on the training data, tuning hyperparameters as needed.

Model Evaluation :

We can construct the testing set from given dataset and use the testing errors and metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), or R-squared

Iterate :

Keep refining the system based on ongoing feedback and emerging technologies.

Interpret Results:

Interpret the model outputs and identify key factors affecting air quality in Tamil Nadu. Understand the significance of variables and their impact.

Documentation:

Maintain clear documentation of the data sources, code, and methodology for reproducibility.

Summarizing historical trends and future prediction based on analysis done on quality of air in Tamil Nadu.