IBM NAN MUDHALVAN PHASE:4

|  |  |
| --- | --- |
| DATE | 26-10-2023 |
| TEAM ID | PROJ\_224711\_TEAM\_2 |
| PROJECT TITLE | AIR QUALITY ANALYSIS AND PREDICTION IN TAMILNADU |

Air Quality Analysis and Prediction in Tamil Nadu

PREDICTION:

Technological advancements lead to the emissions of air pollutants over the decades.

Major concerns in industrial cities which experience air pollution, can be harmful not only

for the environment but also for human health. Due to this urban resident are more likely to

live in less polluted neighborhoods to avoid the health impact of air pollution. Atmospheric

pollution can be classified into three types based on the sources mobile, stationery and area

sources. Mobile sources are due to the motor vehicles, airplanes, locomotives and other

engines and equipment that are able to move to different locations. Stationary sources

include foundries, fossil fuel burning, food processing plants, power plants, refineries and

other industrial sources. Area sources is caused by certain local actions. Air pollution can be

caused due to the pollutants which are emitted directly from a source or which are not

directly emitted as such. It can result in the degradation of ambient air quality in the

industrial cities. Also daily exposure of people to air pollution results in diseases like

asthma, wheezing, and bronchitis.

FEATURE ENGINEEERING:

**Temporal Features:** Extract features like hour of the day, day of the week, month, etc., from timestamps to capture temporal patterns.

**Lagged Features:** Create lag features (past values of the target variable) to capture trends and seasonality.

**Weather Data Integration:** Integrate weather data (temperature, humidity, wind speed, etc.) as features, as weather conditions significantly affect air quality.

**Statistical Features:** Compute statistics like mean, median, standard deviation, etc., for numerical features to provide more information to the model.

**Feature Scaling:** Normalize or standardize features to bring them to a similar scale, especially for models sensitive to feature scales (e.g., SVMs, k-NN).

DATASET:

The data is obtained from https://tn.data.gov.in/resource/location-wise-daily-ambient-air-

quality-tamil-nadu-year-2014

DATA PREPARATION:

**Data Collection:** Gather relevant air quality data, which may include parameters like PM2.5, PM10, CO2 levels, temperature, humidity, wind speed, etc.

**Data Cleaning:** Handle missing values, outliers, and inconsistencies in the dataset.

**Feature Selection:** Identify which features are relevant for your prediction task. You might need domain knowledge or use feature selection techniques to make informed choices.

COLUMNS USED:

From Tamil Nadu \_Air quality analytics.csv data the following columns are used

* stn code
* Sampling Date
* State
* City/Town/Village
* Location of agency
* Type of location
* SO2
* NO2
* RSPM/PM10
* PM2.5

LIBRARIES USED:

The Python 3 environment comes with many helpful analytics libraries installed and several

helpful packages to load.

The essential libraries used in this project are :

* Importing OS (for kaggle inputs)
* Numpy and Pandas libraries
* Matplotlib
* Seaborn

TRAIN AND TEST :

Training the dataset by describe(), isnull ().sum(), drop(), show(), and by using k-means

algorithm we train the data

Testing the data by importing sklearn.cluster from k-means with ensuring the plot range and axis

labels producing the k value, scattering the data by kmeans. clusters\_centers and producing 3D plot

REST OF THE EXPLANATIONS:

Data Collection

The samples are collected from NAMP stations are analysed for the Respirable Suspended

Particulate matter (RSPM) and gaseous pollutants such as Sulphur dioxide(SO2) and Nitrogen

dioxides(NO2)

Data analysis

ANOVA (one way), Tukey HSD, and Pearson correlation coefficient (r) were computed

using self-coded software on Microsoft Excel 2019 to statistically analyze the collected

data.

ALGORITHM USED:

Apply clustering algorithms like K-Means, DBSCAN, or hierarchical clustering to

segment customers.

Visualization: Visualize the customer segments using techniques like scatter plots, bar

charts, and heatmaps. Interpretation: Analyze and interpret the characteristics of each customer

segment to derive actionable insights for marketing strategies.

MODEL SELECTION AND TRAINING:

**Choose Algorithms:** Select appropriate machine learning algorithms for regression tasks, such as Linear Regression, Decision Trees, Random Forest, Gradient Boosting, or even more advanced techniques like Neural Networks.

**Model Training:** Split the data into training and testing sets. Train your selected models on the training data.

MODEL EVALUATION:

* **Metrics:** Use appropriate evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), or R-squared to assess the model's performance.
* **Cross-Validation:** Implement techniques like k-fold cross-validation to get a better estimate of the model's performance

**Cross-Validation:** Implement techniques like k-fold cross-validation to get a better estimate of the model's performance and to ensure that the model generalizes well to unseen data

**Hyperparameter Tuning:** Fine-tune hyperparameters of your models using techniques like grid search or random search to optimize their performance.

PREDICTION AND ANALYSIS:

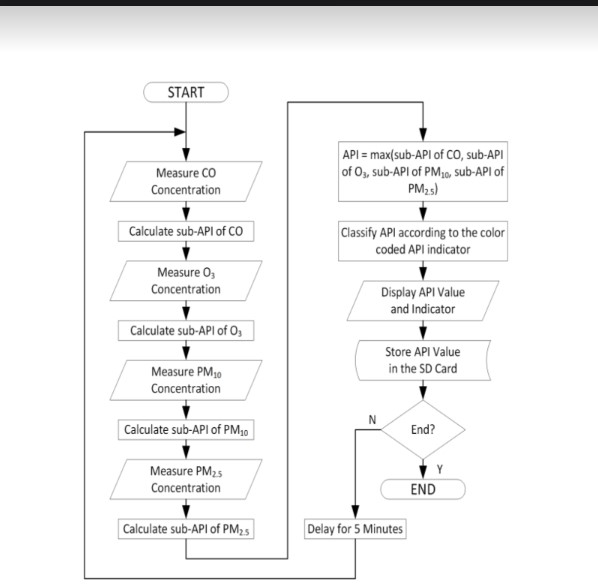
**Make Predictions:** Use the trained model to make predictions on new or unseen data.

**Result Analysis:** Analyze the model predictions, identify any patterns or discrepancies, and iteratively improve the model if necessary.

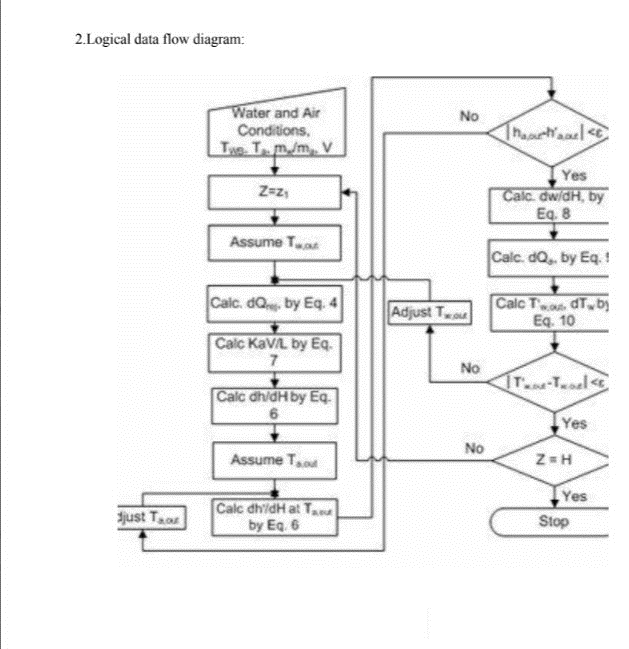
DESIGN AND DATAFLOW

1.Physical data flow diagram:

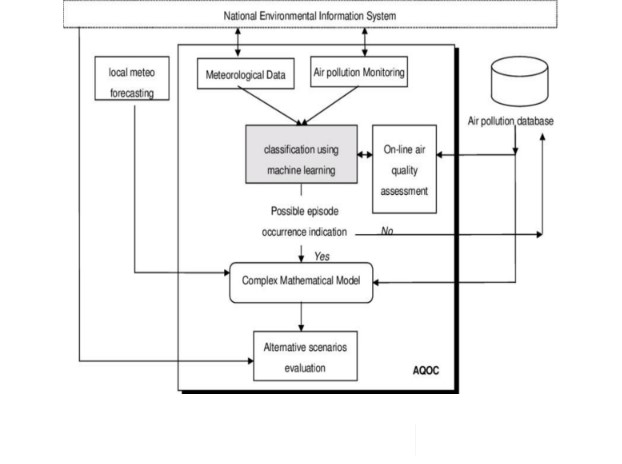
AIR QUALITY ANALYSIS AND PREDICTION



2.Logical data flow diagram:



3. Data flow diagram



Code:

AQI: The air quality index is an index for reporting air quality on a daily basis. In other

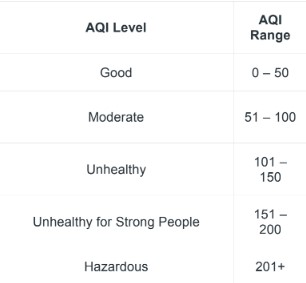
words, it is a measure of how air pollution affects one’s health within a short time period.

The AQI is calculated based on the average concentration of a particular pollutant

measured over a standard time interval. Generally, the time interval is 24 hours for most

pollutants, and 8 hours for carbon monoxide and ozone.

We can see how air pollution is by looking at the AQI



# importing pandas module for data frame

import pandas as pd

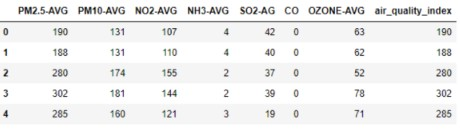
# loading dataset and storing in train variable

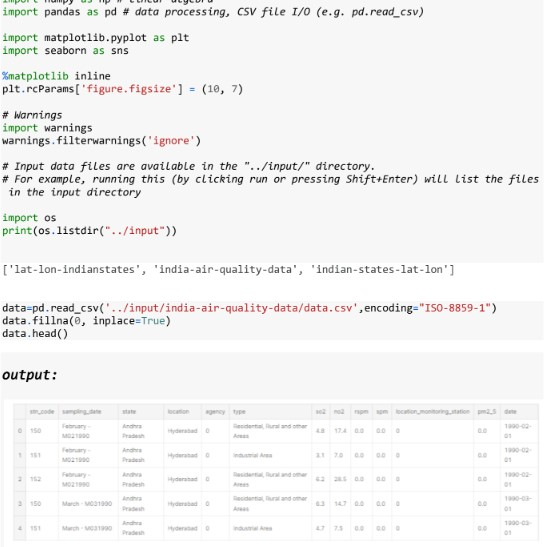
train=pd.read\_csv('AQI.csv')

# display top 5 data

train.head()

Output:





DOCUMENTATION:

**Document the Process:** Maintain a detailed record of the steps you've taken, the decisions you've made, and the results obtained.

CONCLUSION:

The dataset has been visualized and apply model evaluation and building successfully completed

|  |  |
| --- | --- |
| TEAM ID | PROJ\_224711\_TEAM\_2 |
| NAME | PAVITHRA M |
| TEAM MEMBERS | DIVYADHARSHINI S, SHARMILA S, ARTHI M |
| REGISTER NUMBER | 420421205014 |