**Phase 2**

**Innovation**

**Project 9:** Air Quality Analysis and Prediction in Tamil Nadu

**DESCRIPTION:**

The objective of this project is to employ data science techniques to analyze air quality in various regions of Tamil Nadu, India. 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.

**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 online sources like tn.data.gov.in or https://www.kaggle.com/datasets . This data should identifying pollution hotspots, and building a predictive model for RSPM/PM10 levels and cover various pollutants such NO2, SO2 .

**Content :**

This Data is combined version of air quality level of Tamil Nadu from 2010-2015.It has some of the district wise category of data with air quality parameters. And it is released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP).

**Dataset collection:**

The dataset is being collected from <https://www.kaggle.com/datasets>. Here we go to look up the details of the respected Dataset.

* The dataset is taken and it is descript as Air-Quality-data-Tamil-Nadu.
* The dataset contains the trends of air quality from the year 2010 to 2015 from various districts in Tamil Nadu .

This shows the various columns which is responsible for the air quality.

**Factors taken to check the Dataset:**

The dataset attributes includes like ,

1. Stn code (Station code)
2. Sampling date
3. State
4. City /Town /Village
5. Location of Monitoring Station
6. Type of Location (Industrial / Residential / Rural / Urban)
7. SO2
8. NO2
9. RSPM/PM10
10. SPM Levels

**Elaboration about columns :**

* The first column denotes the station code which is numbered in such a way for easy determination
* Sampling date means the date that a sample was collected from the manufacturer or start of a period for sample collections .
* In Tamil Nadu the various districts which includes varieties of localities that are being monitored by location monitoring system.
* The districts are categorized by the types of areas like rural ,urban and industrial etc.
* The levels of gases like SO2 and NO2 are also noted.
* Respirable Suspended Particulate Matter ( rspm ), these particles cause the worst damage as they can penetrate deep into the lungs.

Particulate Matter (PM10) is also evaluated. These particles can be found in dust and smoke. They can impact our health and used to check air quality.

**Libraries Used :**

A Libraries is a collection of functions that can be performed to form a standard code. Some of them are required to be imported. They are

* Pandas
* Numpy (Numerical python)
* Geopandas
* Math
* Seaborn
* Folium
* HeatMap
* MarkCluster
* TimestampedGeoJson

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data .

Numpy can be used to perform a wide variety of mathematical operations on arrays .

Geopandas extends the datatypes used by pandas to allow spatial operations on geometric types. Geometric operations are performed by shapely. Geopandas further depends on file access and matplotlib for plotting. From this geopandas the geo code can be used .

Math is an built in module that can be used for mathematical tasks. It has set of constants, methods and able to perform useful mathematical calculations within our application.

Folium is a library used for visualizing geospatial data .It can used for plotting interactive maps.

Seaborn is a library in python data visualization based on matplotlib. It provides high-level interface and informative statistical graphics.

Timestamp JSON the part of a log message that marks the time that an event occurred.

**How to import :**

The libraries are installed from the python prompt and it is get imported either in pycharm or the notebook.

**Training and Testing :**

Train/Test is method the accuracy of the model.80% for training and 20% for testing .The model is trained using the training set.

**Metrics and Accuracy check:**

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fractions of predictions our model got right

ACCURACY=NO. OF CORRECT PREDICTIONS / TOTAL NO.OF PREDICTIONS

**Conclusion:**

Thus the innovative idea of air quality analysis and prediction in Tamil Nadu has been explored.