Air Quality Analysis and Prediction in Tamil Nadu

Phase 2

10-10-2023

Problem 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.

**Innovation:**

**1. Data Cleaning and Preprocessing**: (Using Pandas)

a. **Drop Unwanted Columns**: Remove columns like stn Code, State, and any other columns that are not relevant to the analysis.

b. **Convert Sampling Date**: Change the data type of the Sampling Date column to a datetime object to enable time-based analysis.

c. **Rename Columns**: Clean up column names by removing spaces and converting them to lowercase to ensure consistency.

d. **Group Data by Area**: Group the data by Area to aggregate and analyze air quality statistics for specific regions.

**2. Exploratory Data Analysis (EDA):**

Conduct various visualizations using matplotlib and seaborn to explore the data, such as:

* **Violin Charts for Overall Data Distributions**: Visualize the distribution of SO2, NO2, and RSPM/PM10 using violin charts to understand the range, median, and spread of these pollutants.
* **Histograms**: Create histograms to visualize the frequency distribution of pollutant levels, helping identify patterns and outliers.
* **Violin Charts by Month**: Analyze data distribution by month to identify seasonal trends and variations.
* **Violin Charts by Weekdays**: Examine data distribution by weekdays to identify any day-of-week patterns.
* **Line Charts with Important Holidays**: Plot line charts showing SO2, NO2, and PM10 levels in each area and location type, highlighting important holidays to identify potential spikes or changes in air quality.
* **Heat Map of Correlations**: Generate a heatmap of the correlations matrix to explore the relationships between SO2, NO2, and PM10.
* **Regression Plots**: Create regression plots for SO2, NO2, and PM10 for each area and location type to visualize trends and relationships.
* **GeoSpatial Air Quality Distribution Heatmaps**: Visualize the spatial distribution of air quality in Tamil Nadu using heatmaps to identify areas with higher or lower pollution levels.

**3. Extract Insights:**

Analyze the visualizations and summary statistics to extract insights, such as:

* Identifying areas with consistently high or low pollutant levels.
* Determining the impact of seasonality and weekdays on air quality.
* Assessing the correlation between different pollutants.
* Recognizing any significant trends or changes in air quality over time.

**4. Build a Predictive Model:**

1. **Model Selection**: Based on the insights from EDA, choose appropriate machine learning models for predicting air quality levels (e.g., Random Forest for SO2, NO2, and PM10).
2. **Data Splitting**: Split the dataset into training and testing sets to train and evaluate the model's performance.
3. **Feature Engineering**: Select relevant features and engineer new ones if necessary.
4. **Model Training**: Train the selected machine learning models on the training data.
5. **Model Testing**: Evaluate the model's performance on the testing data, using appropriate metrics.
6. **Fine-Tuning**: Iteratively fine-tune the model parameters for better performance.

**5. Present Results and Findings**

Communicate the results and insights obtained form the analysis and modeling. This can be don through:

* Creating a report or presentation summarizing the key findings.
* Visualizing and explaining the model's predictions and performance.
* Documenting insights for decision-makers or stakeholders to take action on air quality issues.