Concept Note

Analyzing & Predicting the AQI (Air Quality Index) of Delhi based on 2020 to 2023

Concept of the project

The project focuses on analyzing the impact of air pollution on public health in Delhi, by collecting and analyzing data on pollutants such as CO, NO, NO2, O3, SO2, and particulate matter (PM2.5 and PM10). It aims to establish correlations between pollution levels and public health outcomes, including hospital admissions, respiratory and cardiovascular diseases, and mortality rates. Using data cleaning, exploratory data analysis (EDA), and machine learning models, the project will predict future health risks based on historical pollution data. The goal is to provide actionable insights for policymakers to mitigate health risks and improve public health outcomes in Delhi. By handling real-world data and performing comprehensive analysis, the project seeks to enhance understanding and management of air pollution and its health impacts.

Problem Statement

Air pollution is a significant environmental and public health issue, contributing to various health problems, especially respiratory diseases. This project aims to analyze the correlation between air pollution levels and public health outcomes in Delhi, predicting future health risks based on pollution trends .

Objective of the Project

- To identify patterns and trends in air pollution and public health outcomes.
- To analyze the correlation between air pollution levels and health metrics such as hospital admissions and disease incidence.
- To predict future health risks based on current and historical pollution data.
- To provide actionable insights and recommendations for policymakers to mitigate health risks.

Data Sources used

I have used a kaggle dataset and data based on, from year **2020 to 2023**
Url : https://www.kaggle.com/datasets/deepaksirohiwal/delhi-air-quality

Features

- **Date:** The specific date of data collection.
- **State:** The state where the data was collected (Delhi).

- **Location:** The specific location within the state.
- **Agency:** The agency responsible for collecting the data.
- **CO (Carbon Monoxide):** Concentration levels of CO in the air.
- **NO (Nitric Oxide):** Concentration levels of NO in the air.
- **NO2 (Nitrogen Dioxide):** Concentration levels of NO2 in the air.
- **O3 (Ozone):** Concentration levels of ozone in the air.
- **SO2 (Sulfur Dioxide):** Concentration levels of SO2 in the air.
- **PM2.5 (Particulate Matter 2.5):** Concentration levels of fine particulate matter.
- **PM10 (Particulate Matter 10):** Concentration levels of coarse particulate matter.
- **NH3 (Ammonia):** Concentration levels of NH3 in the air.

Tool for Analysis (Use any tool, even excel)

- **Python**: Pandas, NumPy, SciPy for data manipulation and analysis.
- **Visualization Tools**: Matplotlib, Seaborn for data visualization.
- **Machine Learning Libraries**: Scikit-learn for model training and evaluation.
- **Time Series Analysis**: Statsmodels for ARIMA, Keras/TensorFlow for LSTM.
- **Excel**: For preliminary data exploration and basic analysis.

Hypothesis

Higher levels of air pollution correlate with increased incidence of respiratory diseases and hospital admissions in Delhi.

Methodology

Data Collection:

- Gather air pollution, health, and weather data from respective sources.

Data Cleaning and Preprocessing:

- Handle missing values, remove duplicates, and correct anomalies.
- Normalize and standardize data as needed.

Exploratory Data Analysis (EDA):

- Visualize data to identify patterns and trends.
- Analyze the correlation between pollution levels and health metrics.

Feature Engineering:

- Create pollution indices, average temperature, humidity, and seasonal variation features.
- Aggregate data by time periods for temporal analysis.

Model Selection:

- Choose models like Linear Regression, Random Forest, or Gradient Boosting for health outcome predictions.
- Use ARIMA or LSTM for time series analysis.

Model Training and Evaluation:

- Split data into training and testing sets.
- Train multiple models and evaluate their performance using metrics like MAE,
- RMSE, and R-squared.
- Perform cross-validation for robustness.

Prediction and Visualization:

- Use the best-performing model to predict future health risks.
- Visualize predictions with graphs, charts, and heatmaps.

Conclusion and Recommendations:

- Summarize findings and model performance.
- Provide insights and recommendations for policymakers.

Probable Outcome

- Identification of significant correlations between air pollution levels and health outcomes.
- Predictive models capable of forecasting future health risks based on pollution trends.
- Visualized data insights that highlight critical periods and areas of high health risk.
- Actionable recommendations for reducing air pollution and mitigating its impact on public health.