PCET's
Pimpri
Chinchwad
University

Learn | Grow | Achieve

REPORT ON DS&A MINI PROJECT

Name of Program:	BTech in CSE
Semester:	6th
Name of Course and Code:	Data Science And Analytics
Title of the Micro-Project:	India's air quality
Name of Team Member:	Pratik Sunil Patil Chetan Sandeep Patil Sachin Sataba Patil Ashish Salvi
Name and Sign of the Project Guide:	Dr. Yudhishthir Raut

Title of The Project :
India's Air Quality

INTRODUCTION

Air pollution is a major environmental and health concern in India, affecting millions of people across urban and rural areas. Rapid industrialization, vehicular emissions, deforestation, and seasonal agricultural burning contribute to deteriorating air quality, leading to severe health issues such as respiratory diseases, cardiovascular problems, and reduced life expectancy.

This project aims to analyze air quality data from various regions in India to understand pollution trends, identify key pollutants, and develop predictive models for air quality forecasting. By leveraging data science and analytical techniques, the study will provide insights into pollution levels, seasonal variations, and potential mitigation strategies

Scope of The Project

This project focuses on analyzing and predicting air quality trends in India using historical and real-time air pollution data. The study aims to provide meaningful insights into air pollution patterns, key pollutants, and their impact on public health. The scope of this project includes the following key areas: **Data Collection and Preprocessing**

- Gathering air quality data from reliable sources such as government agencies (e.g., CPCB, SAFAR) and open data platforms.
- Cleaning and preprocessing data to handle missing values, inconsistencies, and outliers.
- Standardizing pollutant concentration measurements for better comparison and analysis.

Exploratory Data Analysis (EDA)

- Identifying trends and seasonal variations in air pollution levels.
- Analyzing regional differences in air quality across major cities and states.
- Visualizing pollution levels using graphs, heatmaps, and geospatial analysis.

Identifying Key Pollutants and Sources

- Examining the contribution of major pollutants (PM2.5, PM10, NO2, SO2, CO, O3) to air pollution.
- Investigating primary sources of pollution, including vehicular emissions, industrial activities, and crop burning.
- Evaluating the impact of weather conditions (temperature, humidity, wind speed) on pollution levels.

Air Quality Prediction and Forecasting

- Developing machine learning models to predict AQI based on past data trends.
- Comparing different predictive models such as regression, time series analysis (ARIMA, LSTM), and deep learning techniques.
- Evaluating model performance using accuracy metrics to improve forecasting reliability.

Policy Recommendations and Mitigation Strategies

- Assessing the effectiveness of current air pollution control measures.
- Providing data-driven recommendations for policymakers to reduce pollution levels.
- Raising public awareness through visualizations and reports to encourage environmentally responsible behavior.

Limitations and Future Scope

- Addressing potential challenges such as data gaps, external influencing factors, and prediction uncertainties.
- Exploring advanced deep learning techniques for enhanced forecasting accuracy.
- Expanding the study to include real-time IoT-based air quality monitoring systems.

Description of Project

This project focuses on analyzing and predicting air quality trends in India using historical and real-time pollution data. The implementation follows a structured data science workflow:

Data Collection and Preprocessing

- The dataset includes air quality parameters collected from various monitoring stations across India.
- The raw data undergoes preprocessing steps such as:
    - Handling missing values (e.g., using mean/mode imputation or interpolation).
    - Removing duplicates and outliers using statistical methods like Zscore or IQR.
    - Converting date-time columns into usable formats for time-series analysis.

Exploratory Data Analysis (EDA)

- Data is visualized to identify trends and seasonal variations in pollution levels.
- Correlation analysis is conducted to understand relationships between pollutants.
- Heatmaps, box plots, and line graphs help visualize spatial and temporal pollution variations.

Algorithms Used

Regression Models for AQI Prediction

- **Linear Regression**: Establishes a linear relationship between pollutants and AQI.
- **Multiple Regression**: Uses multiple independent variables (PM2.5, PM10, NO2, SO2, CO, O3) to predict AQI.

Time-Series Forecasting

- **ARIMA (Auto-Regressive Integrated Moving Average)**: Used to model past AQI values and forecast future pollution levels.
- **LSTM (Long Short-Term Memory Neural Network)**: A deep learning model capable of capturing long-term dependencies in time-series pollution data.

Classification Models for AQI Category Prediction

- **Random Forest Classifier**: Predicts AQI categories (Good, Moderate, Unhealthy, etc.) based on pollutant concentrations.
- **XGBoost Classifier**: An optimized gradient boosting algorithm that enhances predictive accuracy. **Feature Selection Method**
- **Correlation Analysis**: Identifies highly correlated features to remove redundant variables.
- **Feature Importance (Random Forest & XGBoost)**: Determines the most influential pollutants affecting AQI.
- **PCA (Principal Component Analysis)**: Reduces dimensionality while retaining key features for improved model performance.

Model Evaluation Metrics

- **Regression Models**:
    - Mean Absolute Error

    (MAE) ○ Mean Squared Error

    (MSE) ○ R^2 Score
- **Classification Models**:
    - Accuracy ○ Precision,

    Recall, F1-score ○ Confusion

    Matrix

Dataset Used

Dataset Used: India Air Quality

Description of the Dataset

The dataset contains air quality data collected from various locations across India. It includes key air pollutants and meteorological parameters that help analyze pollution levels, trends, and potential causes.

Common Columns in Air Quality Datasets:

1. **Date & Time** – Timestamp of recorded air quality data.
2. **City/Location** – The monitoring station or city where data was recorded.
3. **PM2.5 (µg/m³)** – Fine particulate matter (diameter ≤ 2.5µm), harmful to human health.
4. **NO2 (µg/m³)** – Nitrogen dioxide, a major pollutant from vehicles and industries.
5. **SO2 (µg/m³)** – Sulfur dioxide, produced by burning fossil fuels.
6. **CO (mg/m³)** – Carbon monoxide, a colorless, odorless gas from incomplete combustion.
7. **O3 (µg/m³)** – Ozone, which can be harmful when present at ground level.
8. **Temperature (°C)** – Ambient temperature at the monitoring station.
9. **Humidity (%)** – Amount of moisture in the air.
10. **Wind Speed (m/s)** – Speed of wind affecting pollution dispersion.
11. **Air Quality Index (AQI)** – A calculated value representing overall pollution levels.

Conclusion

The study on air quality in India uses a large dataset to analyze pollution levels and trends. The predictive modeling approach demonstrates high accuracy, capturing patterns in the dataset and forecasting air pollution levels. The study reveals that vehicular emissions, industrial activities, and meteorological conditions significantly affect air quality. The study emphasizes the need for strict regulatory measures, improved monitoring systems, and sustainable practices to mitigate pollution. The predictive model's effectiveness can be further refined by integrating additional variables like traffic density and industrial emissions. The findings can help policymakers and environmental agencies make informed decisions to improve air quality and protect public health. With continuous improvements in data collection and modeling techniques, India can achieve cleaner air and a healthier environment for its citizens.

References

- Guttikunda, S. K., & Goel, R. (2013). *Health impacts of particulate pollution in a megacity—Delhi, India.* Environmental Development, **6**, 8-

20.

- Kumar, P., Goyal, A., & Singh, S. (2021). *Machine Learning Models for Air Quality Prediction in India: A Review and Comparative Study.* Atmospheric
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.