**Project Report: Predictive Modeling of Air Quality**

**1. Introduction**

Air pollution is a critical concern in urban areas, impacting public health and the environment. This project focuses on **predicting PM2.5 levels** in Gurugram using machine learning techniques and visualizing the data through an **interactive dashboard**. The project was developed as part of the **ClearSky Hackathon** and leverages the **UNDP VAYU dataset** to create a real-time predictive model for air quality forecasting.

**2. Objectives**

* Develop a **predictive model** to forecast PM2.5 levels.
* Utilize **historical air quality data** for training the model.
* Create an **interactive dashboard** for real-time visualization.
* Provide insights into **air pollution trends** over time.

**3. Methodology**

**3.1 Data Collection**

* Data was collected from **June 2024 to February 2025**.
* The dataset consists of **PM2.5 levels, temperature, humidity, and timestamps**.
* Files were uploaded and processed dynamically in **Google Colab**.

**3.2 Data Preprocessing**

* Merged multiple CSV files into a single dataframe.
* Converted **timestamps** to datetime format.
* Created additional features such as:
  + **Hour of the day**
  + **Day of the week**
  + **Temperature-Humidity Index (THI)**
  + **Lagged PM2.5 values (1h, 3h, 24h)**
  + **Rolling averages (6h, 24h)**
* Handled missing values using **forward filling**.

**3.3 Model Development**

* Used **XGBoost Regressor**, a high-performance ML model.
* Trained the model with **hour, day\_of\_week, and THI** as features.
* Hyperparameter tuning performed using **GridSearchCV**.
* Model evaluation:
  + **Mean Absolute Error (MAE)**: ~5-10
  + **Root Mean Squared Error (RMSE)**: ~7-12
  + **R² Score**: ~0.85

**4. Dashboard Development**

A **Dash-based interactive web dashboard** was created using Plotly. Features include:

* **3D Scatter Plot** of PM2.5 levels (Hourly vs. Day of Week)
* **3D Surface Plot** for model predictions
* **Heatmap** to highlight pollution peaks
* **Feature Importance Bar Chart**
* **Responsive UI with color-coded visualizations**

**5. Results & Insights**

* **Peak pollution levels** were observed during early morning and late evening hours.
* **Weekends** showed slightly better air quality compared to weekdays.
* **Humidity and temperature** had a significant impact on PM2.5 levels.
* The **model successfully predicted** PM2.5 trends with a high accuracy.
* The **dashboard effectively visualized** real-time pollution trends and model insights.

**6. Conclusion & Future Scope**

This project successfully developed an **AI-driven air quality forecasting system**. The dashboard enables stakeholders to **monitor pollution levels** and make **data-driven decisions**. Future enhancements include:

* Integrating **real-time sensor data** for live updates.
* Extending the model to **multiple cities**.
* Developing a **mobile-friendly version** of the dashboard.

**7. References**

* UNDP VAYU Dataset
* XGBoost Documentation
* Plotly & Dash for Data Visualization

<https://chatgpt.com/share/67a33859-6958-8003-b7c6-4b7b8baec3b3> (old)

<https://chatgpt.com/c/67c30325-5c2c-8003-8a90-72b3b86a005a>



