**Capstone Project:**

**Title: Air Quality Prediction and Classification**

**For dataset column descriptions refer to this link:** [**https://archive.ics.uci.edu/dataset/360/air+quality**](https://archive.ics.uci.edu/dataset/360/air+quality)

**Dataset**

The dataset contains **9,357 hourly air quality measurements** collected in an Italian city from March 2004 to February 2005. It includes pollutant concentrations (CO, NOx, Benzene, etc.), sensor readings, and weather conditions (temperature, humidity).

**Problem Context**

Air pollution is a critical public health issue. Monitoring and predicting pollutant levels can help in decision-making and preventive actions. This project involves applying **machine learning techniques** to analyze air quality data, predict pollutant concentrations, and classify air quality levels.

**Project Objectives**

1. **Exploratory Data Analysis (EDA):**
   * Study pollutant trends over time (daily, weekly, seasonal).
   * Analyze relationships between pollutants, sensor responses, and weather conditions.
2. **Regression Task:**
   * Build regression models to predict pollutant concentration.
   * Example target variables: **CO(GT)** (Carbon Monoxide), **NOx(GT)**, or **C6H6(GT)** (Benzene).
   * Evaluate using metrics such as **RMSE** and **MAE**.
3. **Classification Task:**
   * Define categories for air quality (e.g., Good, Moderate, Unhealthy) based on pollutant levels.
   * Build classification models to predict these categories.
   * Evaluate using metrics such as **Accuracy, Precision, Recall, and F1-score**.
4. **Model Comparison:**
   * Compare performance of different algorithms (e.g., Linear Regression vs. Random Forest for regression, Logistic Regression vs. Decision Tree for classification).
   * Select the best-performing models.
5. **Reporting:**
   * Summarize findings with plots, tables, and insights.
   * Highlight which features (sensors/weather) are most important in predicting air quality.

✅ **Deliverables:**

* A Jupyter Notebook with full analysis, model building, and results.
* A short project report (2–3 pages) summarizing methodology, results, and conclusions.

**📊 Air Quality UCI Dataset:**

**Dataset Overview**

* **Records:** 9,357 hourly measurements (about one year).
* **Domain:** Environmental monitoring (air quality in an Italian city, 2004–2005).
* **Goal:** Analyze and model pollutant concentrations using sensor and weather data.

**Column Descriptions**

1. **Date** – Monitoring date
2. **Time** – Monitoring hour
3. **CO(GT)** – True hourly averaged CO concentration (mg/m³)
4. **PT08.S1(CO)** – CO-sensitive sensor response
5. **NMHC(GT)** – True hourly averaged Non-Methane Hydrocarbons (µg/m³)
6. **C6H6(GT)** – True hourly averaged Benzene concentration (µg/m³)
7. **PT08.S2(NMHC)** – NMHC-sensitive sensor response
8. **NOx(GT)** – True hourly averaged Nitrogen Oxides concentration (ppb)
9. **PT08.S3(NOx)** – NOx-sensitive sensor response
10. **NO2(GT)** – True hourly averaged Nitrogen Dioxide concentration (µg/m³)
11. **PT08.S4(NO2)** – NO2-sensitive sensor response
12. **PT08.S5(O3)** – Ozone-sensitive sensor response
13. **T** – Ambient temperature (°C)
14. **RH** – Relative Humidity (%)
15. **AH** – Absolute Humidity