Project Title: A Machine Learning Approach to Air Quality Classification Using Sensor Array Data

- **Team Members:** [Prajwal V – PES2UG23CS424, Sipivistha – PES2UG23CS412]

- **Course:** UE23CS352A Machine Learning
- **Date:** October 5, 2025

---

1. Abstract

This report details the development of a high-accuracy machine learning model for classifying urban air quality. Utilizing a dataset of hourly readings from a multi-sensor array, we implemented a complete data science pipeline. The process included data preprocessing, exploratory data analysis (EDA) to uncover temporal patterns, and strategic feature engineering. A **Random Forest Classifier** was trained to predict pollution levels, achieving an exceptional accuracy of **99.84%** on unseen test data, demonstrating the viability of using sensor arrays for real-time air quality monitoring.

---

2. Data and Methodology

The project followed a systematic approach to transform raw sensor data into actionable predictions.

- **Dataset:** The foundation of this project is the "Air Quality" dataset, containing 9,357 valid hourly observations across 15 chemical and environmental sensors.
- **Data Preprocessing:** The initial step involved a robust cleaning process. Missing values, denoted by -200, were imputed using the **forward-fill (ffill) method**, a suitable strategy for time-series data. The Date and Time columns were merged into a single DateTime index to facilitate temporal analysis. Columns with insufficient data, such as NMHC(GT), were removed to prevent noise.
- **Exploratory Data Analysis (EDA):** To understand the data's underlying structure, we used **Matplotlib** and **Seaborn**. Time-series plots of key pollutants like Benzene (C6H6(GT)) revealed distinct seasonal patterns, with higher concentrations in colder months. A correlation heatmap confirmed the strong, expected relationships between sensors and their target pollutants.
- **Feature Engineering:** This was a critical step to enhance model performance.
  - **Target Variable:** A categorical target, Pollution_Level, was created by segmenting the C6H6(GT) readings into four quartiles: 'Low', 'Medium', 'High', and 'Very High'.

  - **Time-Based Features:** To capture the patterns seen in EDA, we extracted Hour, DayOfWeek, and Month from the DateTime index. These features allow the model to learn daily cycles (e.g., rush hour) and seasonal variations.

---

3. Model Training and Evaluation

- **Algorithm Selection:** We chose the **Random Forest Classifier** from the **Scikit-learn** library due to its high performance, robustness, and ability to handle complex interactions between features.
- **Training Process:** The dataset was partitioned using the train_test_split function, with 80% of the data used for training the model and the remaining 20% reserved for testing. This ensures that the model's evaluation is based on data it has never seen before.

---

4. Results and Conclusion

The model's performance on the unseen test set was outstanding, validating our approach.

- **Accuracy:** The final model achieved an accuracy of **99.84%**.
- **Precision and Recall:** The classification report showed scores of 1.00 for precision, recall, and F1-score across almost all classes, indicating that the model is extremely reliable and makes very few misclassifications.
- **Conclusion:** The results strongly indicate that a Random Forest model, when combined with thoughtful feature engineering, can predict air quality levels from sensor data with near-perfect accuracy. This demonstrates the immense potential for building reliable, automated environmental monitoring systems.

5. Future Work

While the model is highly accurate, future work could explore hyperparameter tuning to further optimize performance or experiment with other advanced models like Gradient Boosting (XGBoost) or simple Neural Networks (ANN).