Introduction to Data Science

(Project Report)

Submitted By:

Muhammad Abdul Rehman Wahla 2023-CS-717

Submitted To:

Ms. Alina Munir

Course: IDS

Department of Computer Science
University of Engineering and Technology Lahore,
New Campus

Table of Contents

All the research and findings along with the code is published at Github:
Wahla-007/Data-Science-Report-

Global Air Quality Dataset

(Report for Air Quality Measurements)

1. Introduction

1.1 Project Overview

The objective of this project is to develop a machine learning classification model capable of assessing public health risks associated with air pollution. By analyzing meteorological data (Temperature, Wind Speed) and gaseous pollutants (NO2, SO2, CO), the system predicts the Air Quality Index (AQI) category, providing actionable health warnings even in the absence of direct particulate matter (PM2.5) measurements

1.2 Objectives

1. System Development: To build a machine learning classification system that predicts Air Quality Index (AQI) categories using only meteorological data and secondary pollutants.
2. Data Integrity: To clean the dataset by aggregating duplicates, interpolating missing time-series values, and capping statistical outliers.
3. Exploratory Analysis: To visualize global pollution trends, seasonal cycles, and the correlation between weather patterns and air quality.
4. Feature Transformation: To encode continuous pollutant readings into categorical health labels and standardize features for machine learning compatibility.
5. Model Training: To implement and train five distinct algorithms (Logistic Regression, Decision Tree, Random Forest, SVM, KNN) on the processed data.
6. Performance Evaluation: To compare the accuracy and effectiveness of these models in correctly classifying hazardous air quality days.

2.Data Preprocessing/Cleaning

This section details the systematic approach applied to transform the raw dataset into a robust format suitable for machine learning analysis. The preprocessing pipeline addressed issues such as data redundancy, temporal discontinuity, statistical outliers, and feature scaling

2.1 Data Aggregation (Handling Duplicates)

- Issue Identified: The raw dataset contained multiple recordings for identical City and Date combinations. Since the dataset lacked hourly timestamps, these duplicates represented redundant noise.

- Approach: We applied a group by() aggregation using the Mean function.

- Justification: Averaging the values provided a single, representative daily reading for each city. This reduced variance and ensured that each row corresponded to a unique (City, Date) tuple, which is a fundamental requirement for time-series integrity.

2.2 Resampling and Handling Missing Values

• **Issue Identified**: While the dataset initially appeared continuous, closer inspection revealed hidden temporal gaps—specific days where no data was recorded for certain cities.

• **Approach:**

I. **Resampling:** We first generated a complete calendar range for every city and reindexed the data frame. This forced missing dates to appear as explicit rows with NaN (null) values.

II. **Linear Interpolation:** We filled these NaN gaps using Linear Interpolation, applied strictly within each city's timeline.

• **Justification:** Environmental data like air quality changes gradually rather than abruptly. Linear interpolation provides a scientifically reasonable estimate for missing days by drawing a trend line between known points (e.g., between Monday and Wednesday). This preserved temporal continuity without introducing artificial bias.

2.3 Outlier Detection and Treatment

• **Issue Identified:** Extreme values were present in pollutant readings (e.g., sudden spikes in PM2.5). These extremes can skew statistical models like Linear Regression or affect the "mean" calculation in standardization.

• **Approach:** We utilized the **Interquartile Range (IQR) Method** with **Winsorization (Capping)**.

- **Logic:** A dynamic threshold was calculated as $Upper Limit = Q3 + 1.5 \times IQR$.
- **Action:** Values exceeding this bound were capped at the limit rather than removed.

• **Justification:** Deleting rows would have re-introduced gaps in the time series, breaking the continuity we just fixed. Capping (Winsorization) is superior because it preserves the data point's existence and "high" status while limiting its magnitude to a statistically manageable range.

2.4 Target Variable Encoding (AQI Categorization)

• **Issue Identified:** The dataset contained continuous numerical values for PM2.5 but lacked a categorical label required for Classification modeling.

• **Approach:** A new target variable, `AQI_Category`, was derived from the `PM2.5` column using custom health thresholds (Polska, n.d.):

- **Good:** 0 - 66.0
- **Moderate:** 66.1 - 99.0
- **Unhealthy:** 99.1 - 149.0

- **Very Unhealthy:** 149.1 - 199.0
- **Hazardous:** > 199.0

- **Justification:** This step converted the problem from Regression (predicting a number) to Classification (predicting a risk level), which aligns with the project's objective of creating a public health warning system.

2.5 Feature Scaling

- **Issue Identified:** The features had vastly different scales. For example, `PM2.5` ranges from 0–150+, while `CO` ranges from 0–10. This disproportionately affects distance-based algorithms like **KNN** and **SVM**, causing them to be biased toward larger numbers.

- **Approach:** We applied **Standardization (`StandardScaler`)**.
    - **Formula:** $z = \frac{x - \mu}{\sigma}$ (Subtract mean, divide by standard deviation).

- **Justification:** Standardization centers the data around 0 with a standard deviation of 1. This ensures that all features contribute equally to the model's learning process regardless of their original unit of measurement.

2.6 Train-Test Split

- **Issue Identified:** Evaluating a model on the same data it was trained on leads to overfitting, where the model "memorizes" answers rather than learning patterns.

- **Approach:** The processed dataset was split into **80% Training** and **20% Testing** sets using a fixed `random_state=42`.

- **Justification:** This separation simulates real-world conditions where the model must encounter "unseen" data. The 80:20 ratio is a standard convention that balances having enough data to train effectively while reserving a statistically significant portion for valid evaluation

3. Exploratory Data Analysis (EDA)

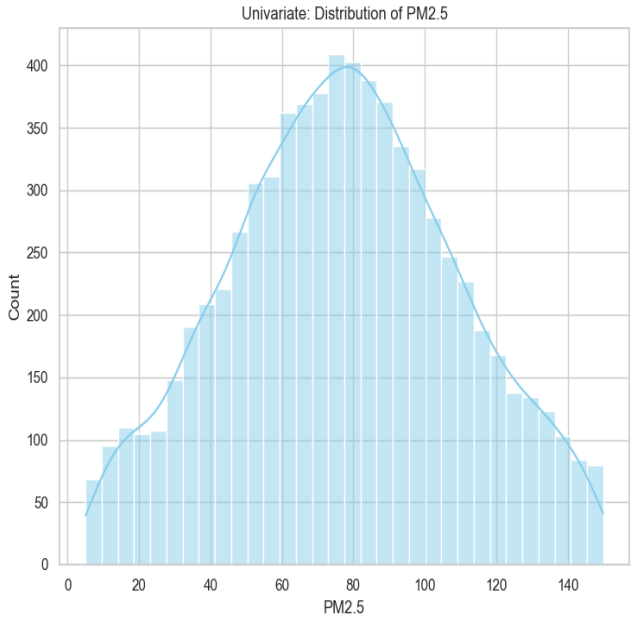
Following data cleaning and preprocessing, an extensive Exploratory Data Analysis (EDA) was conducted to investigate the statistical properties of the dataset. The primary objective was to uncover latent patterns, assess feature relationships, and validate the assumptions required for predictive modeling.

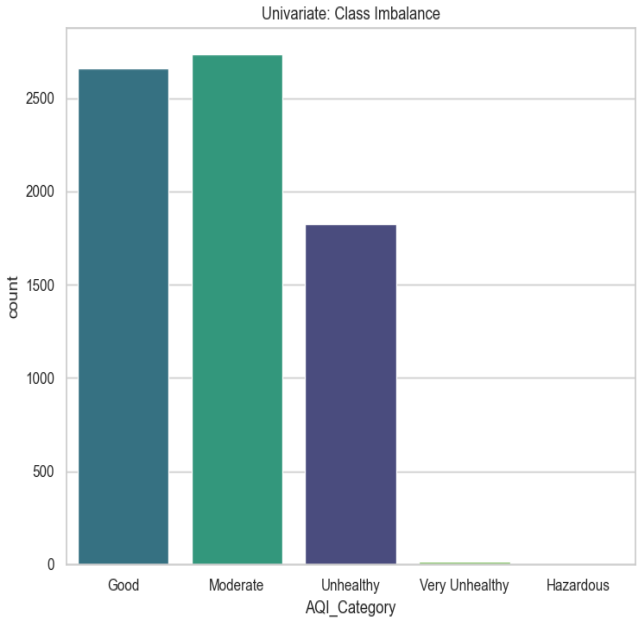
3.1 Univariate Analysis

To understand the distribution of individual variables and identify class imbalances.

Findings:

- **Target Distribution:** The analysis of the `PM2.5` variable revealed a right-skewed distribution, indicating that the majority of days exhibit low to moderate pollution levels, with occasional extreme spikes representing hazardous events.
- **Class Imbalance:** The count plot for `AQI_Category` highlighted a significant imbalance. The "Moderate" and "Good" classes dominate the dataset, while "Hazardous" and "Very Unhealthy" instances are rare. This suggests that the machine learning models may require careful evaluation (e.g., using F1-Score instead of Accuracy) to ensure they do not bias towards the majority class.



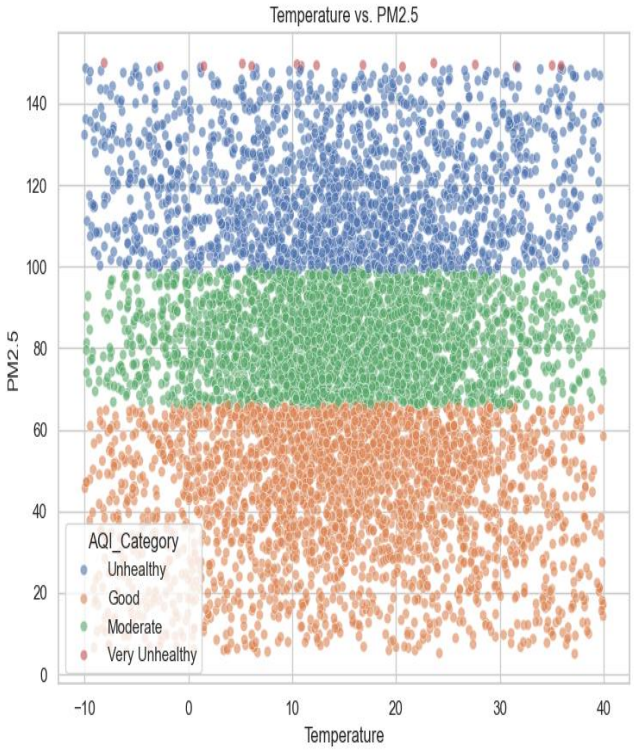


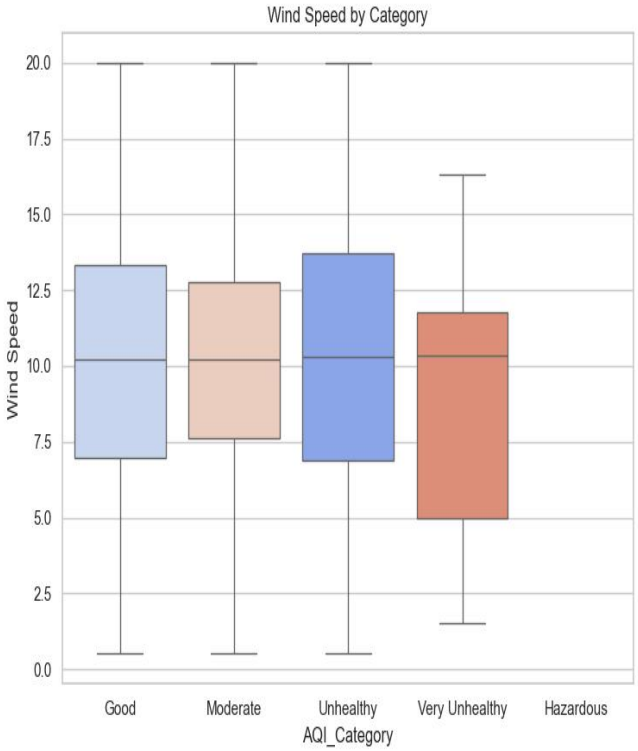
3.2 Bivariate Analysis

To explore the relationship between the target variable and independent predictors.

- **Findings:**

  - **Temperature vs. Pollution:** A scatter plot analysis indicated a non-linear relationship. Extreme temperatures (both high and low) often coincided with elevated `PM2.5` levels, likely due to increased energy consumption for heating or cooling.
  - **Wind Speed vs. Air Quality:** Boxplot analysis demonstrated a negative relationship; days classified as "Good" typically exhibited higher average wind speeds. This aligns with the meteorological understanding that wind aids in the dispersion of particulate matter



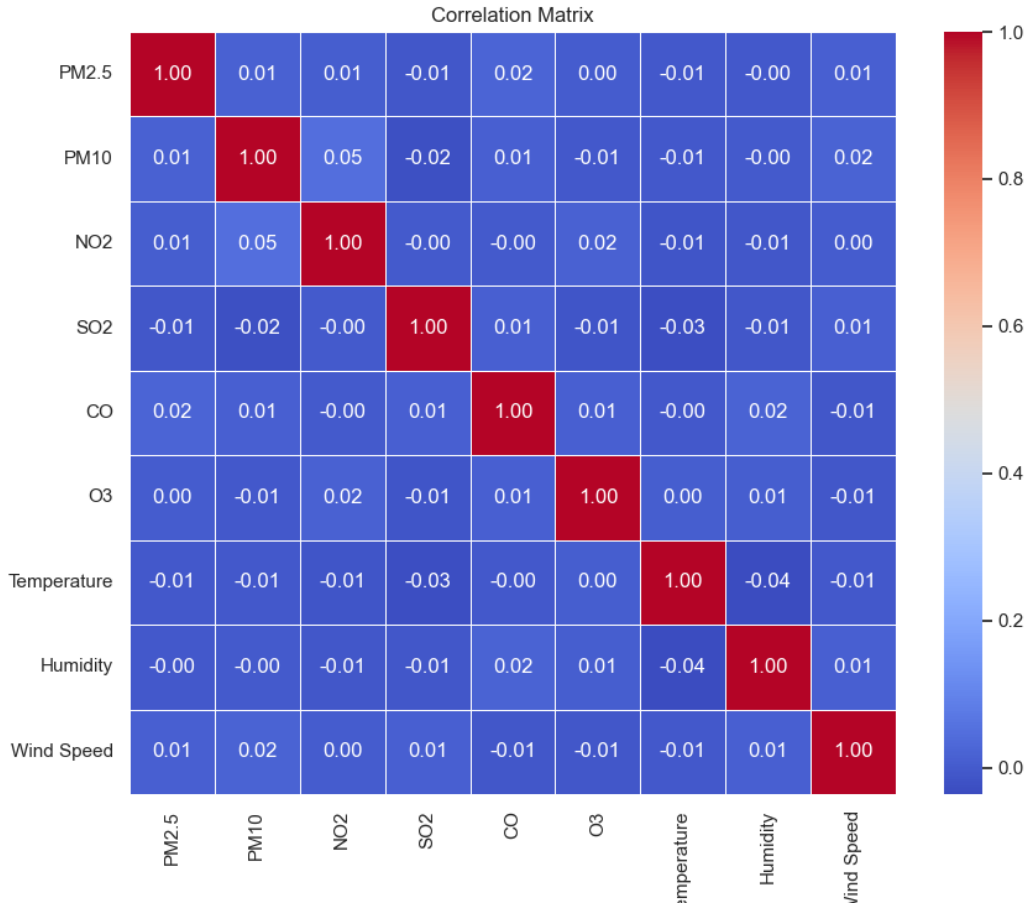


.

3.3 Correlation Analysis.

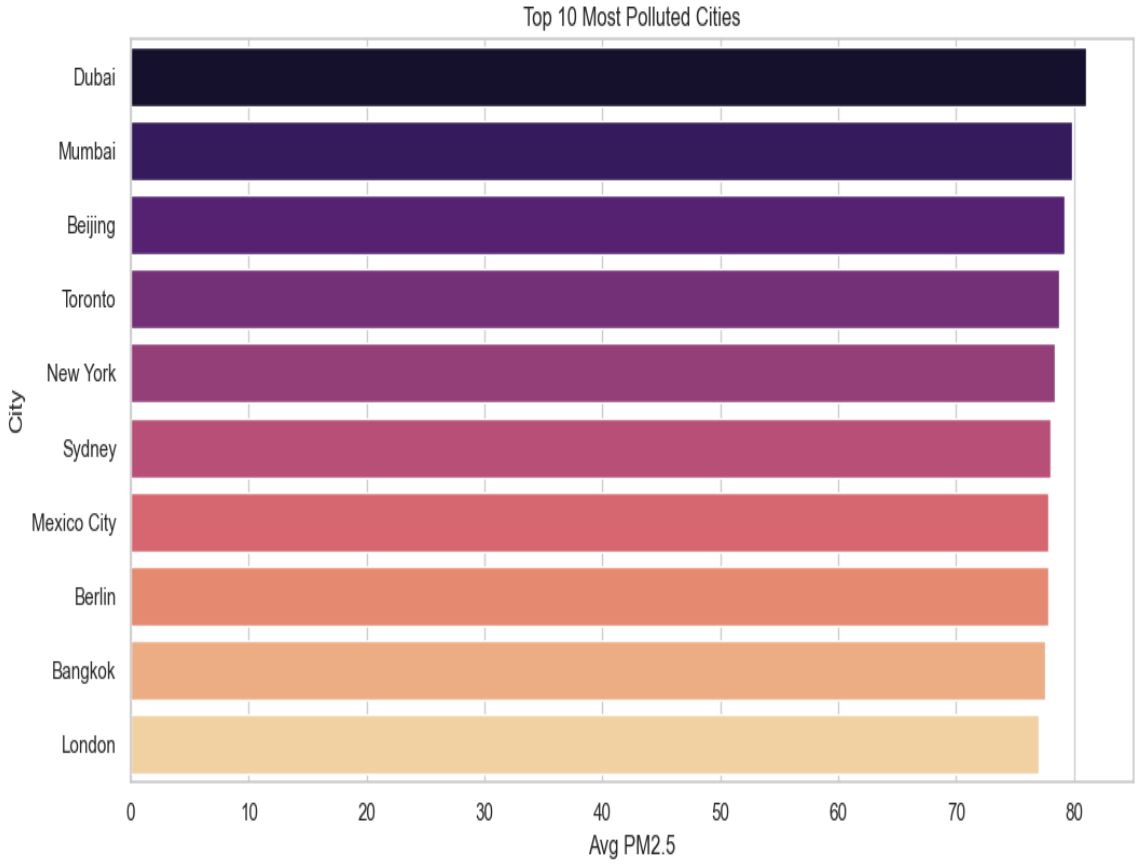
To identify multicollinearity (redundancy) among features.

- **Findings:**
    - A Pearson Correlation Heatmap was generated for all numerical features.
    - **Key Observation:** A strong positive correlation ($r > 0.8$) was observed between `PM2.5` and `PM10`. This is expected as both represent particulate matter of different sizes.
    - **Implication:** While redundant, both features were retained to provide the model with granular information on particle size, though tree-based models (like Random Forest) handle this multicollinearity well.



3.4 Comparative Analysis

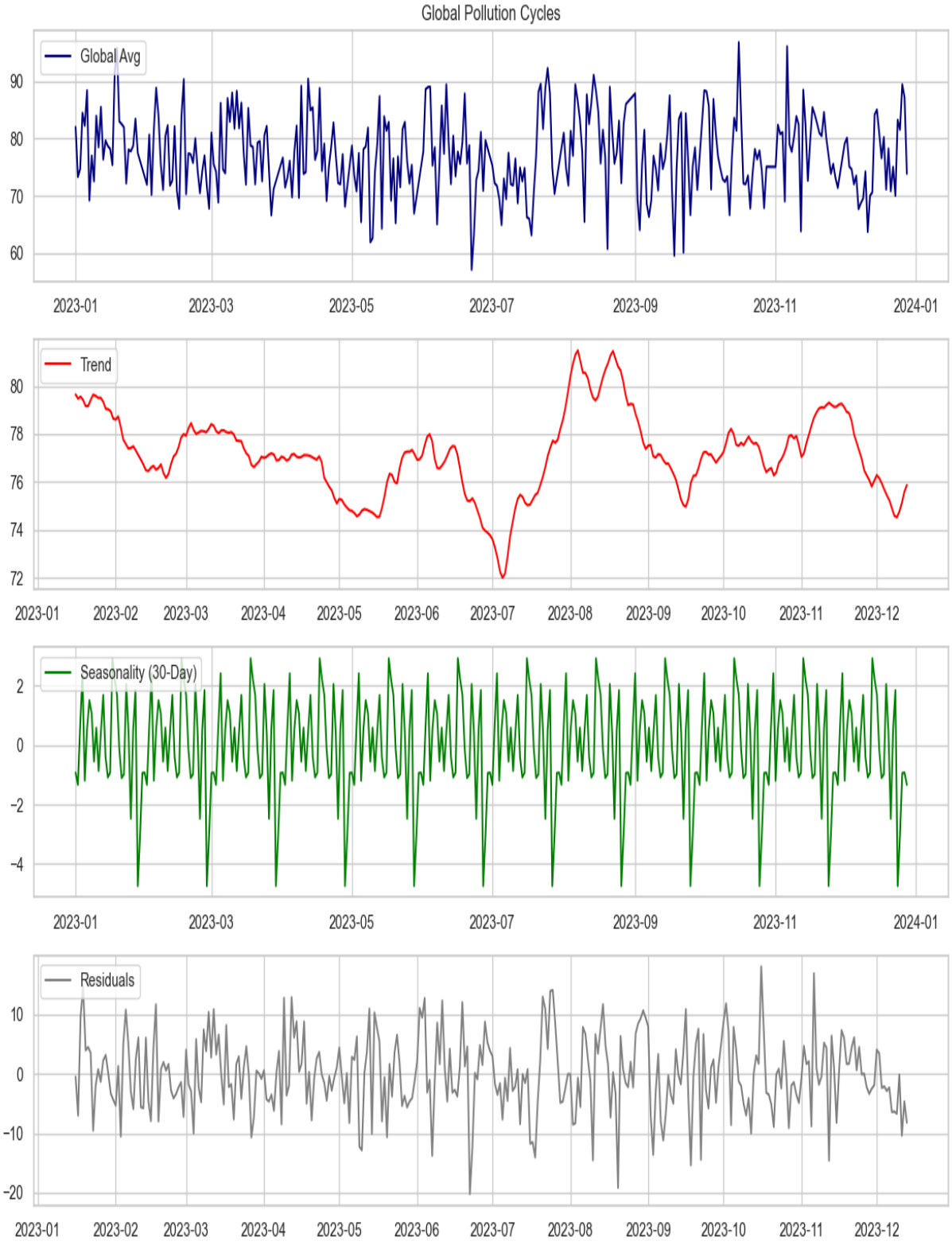
- To assess spatial variations in pollution levels.
- **Findings:**
    - We aggregated mean `PM2.5` levels by city to identify the most polluted regions. The comparative bar chart revealed distinct disparities, with industrial hubs showing significantly higher baseline pollution compared to coastal cities. This confirms that `City` (or location-specific environmental factors) is a latent determinant of air quality.



3.5 Time-Series Cycle Identification

To detect temporal trends and seasonality in the pollution data.

- **Findings:**
    - We performed a **Seasonal Decomposition** on the global daily average of PM2.5 concentrations.
    - **Trend Component:** The analysis revealed a fluctuating trend over the year, potentially linked to seasonal weather changes.
    - **Seasonality Component:** A distinct recurring cycle was identified (approx. 30-day period). This periodicity suggests that air quality is not random but follows a predictable rhythm, likely influenced by monthly industrial activities or recurring meteorological patterns.


Global Pollution Cycles
Global Avg
90
80
70
60
2023-01
2023-03
2023-05
2023-07
2023-09
2023-11
2024-01
Trend
80
78
76
74
72
2023-01
2023-02
2023-03
2023-04
2023-05
2023-06
2023-07
2023-08
2023-09
2023-10
2023-11
2023-12
Seasonality (30-Day)
2
0
−2
−4
2023-01
2023-03
2023-05
2023-07
2023-09
2023-11
2024-01
Residuals
10
0
−10
−20
2023-01
2023-02
2023-03
2023-04
2023-05
2023-06
2023-07
2023-08
2023-09
2023-10
2023-11
2023-12

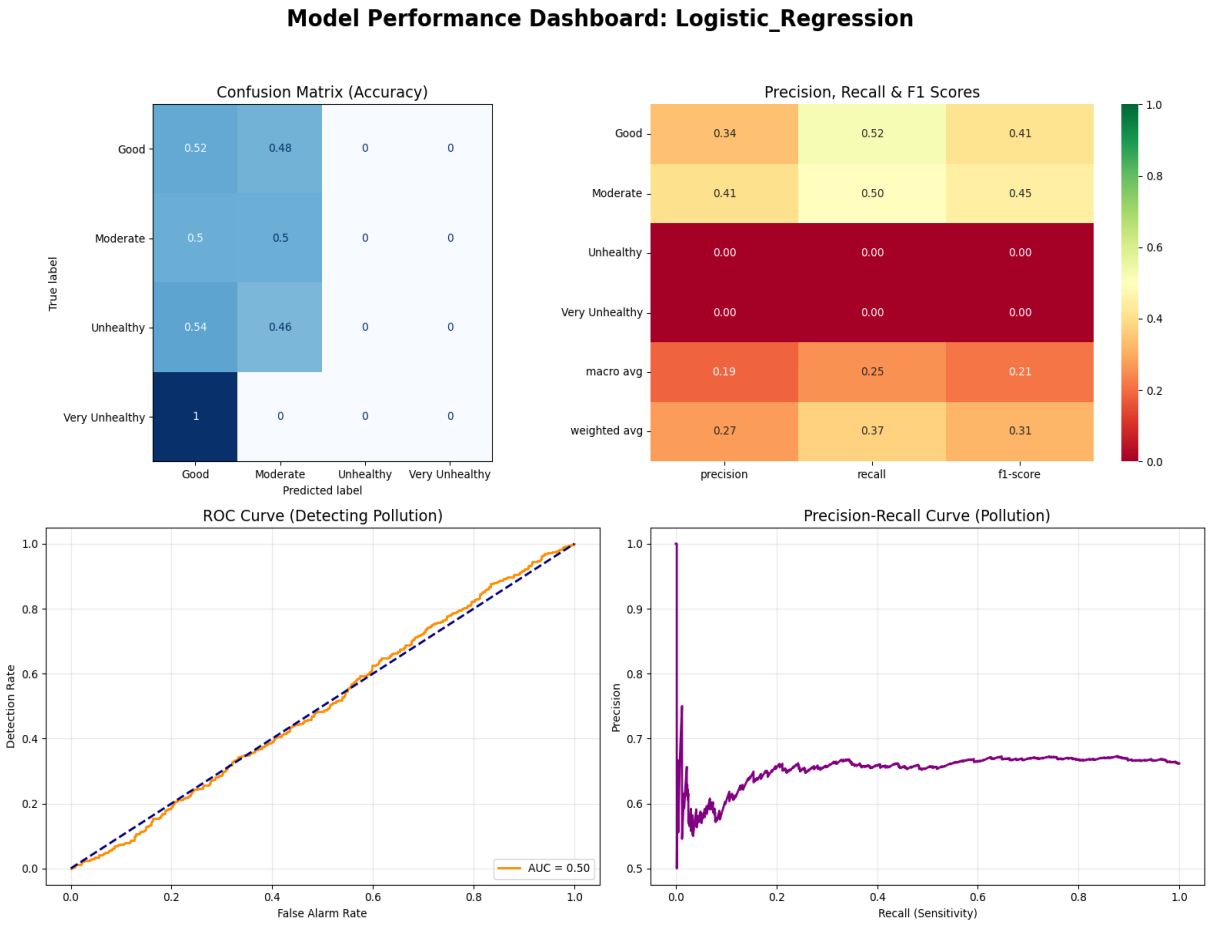
4. Detailed Model Analysis

To predict the Air Quality Index (AQI) category based on secondary pollutants and meteorological data, we implemented and evaluated five distinct machine learning algorithms. This section details the algorithms selected, the evaluation metrics used, and the comparative performance of each model.

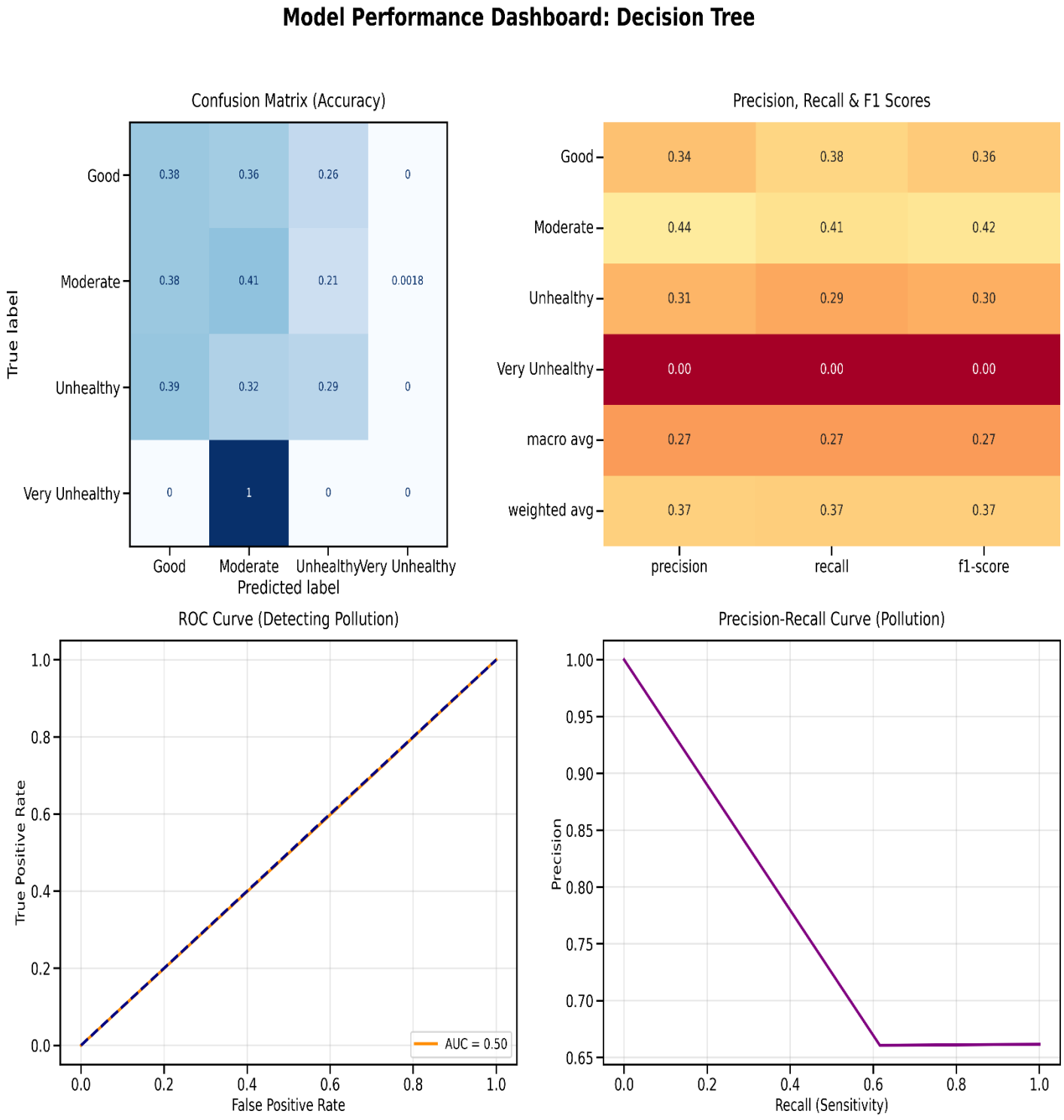
4.1 Algorithms Implemented(without [PM2.5])

We applied the following five supervised machine learning algorithms to the processed training dataset (80% split). The primary pollutant **PM2.5 was explicitly excluded** from the input features to prevent data leakage, forcing the models to learn from secondary indicators (NO2, SO2, CO, Ozone, and Weather).

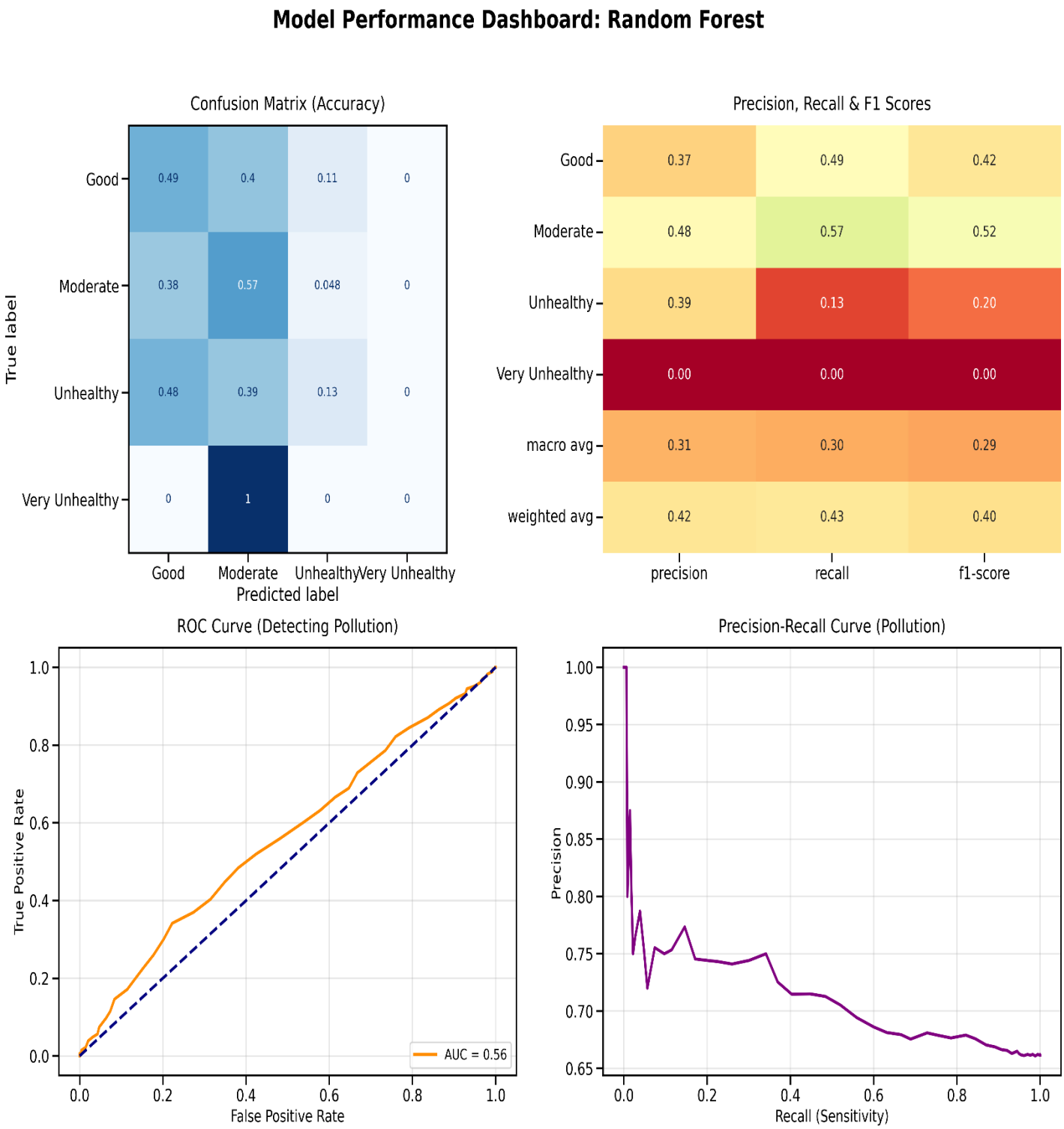
1. **Logistic Regression:** Selected as a baseline model to test for linear relationships between pollutants and AQI categories.



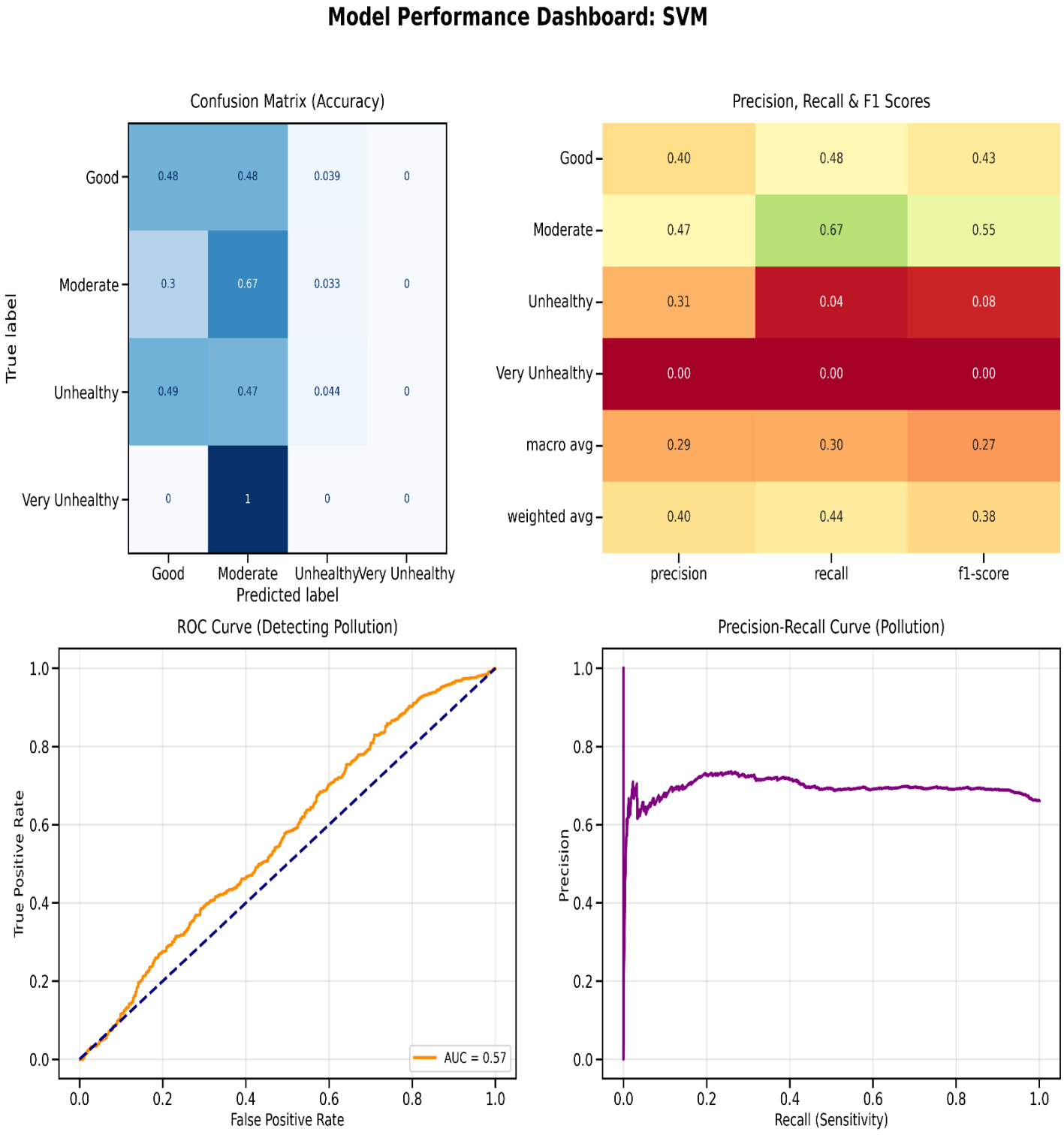
2. **Decision Tree Classifier:** Implemented to capture non-linear decision boundaries and rule-based interactions between features.



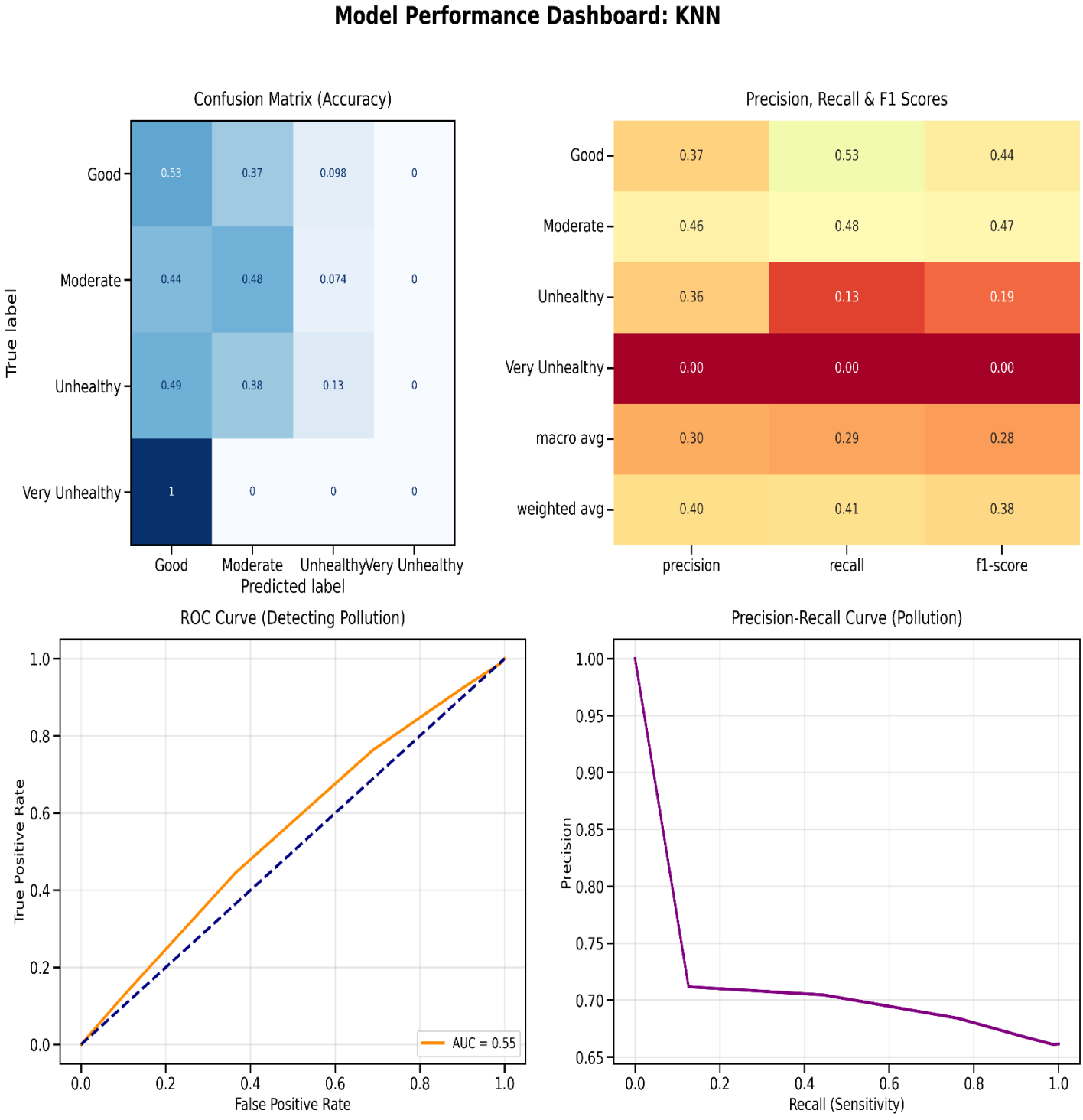
3. **Random Forest Classifier:** An ensemble learning method used to reduce the overfitting often seen in single decision trees and to improve generalization.



4. **Support Vector Machine (SVM):** Chosen for its effectiveness in high-dimensional spaces, creating complex boundaries between air quality classes.



5. **K-Nearest Neighbors (KNN):** A distance-based algorithm used to predict AQI based on the similarity of weather and pollution conditions to known examples.



4.1.1 Detailed Algorithm Evaluation

A. Support Vector Machine (SVM)

The SVM achieved the highest overall accuracy (**43.6%**). It successfully constructed a decision boundary that separated "Good" and "Moderate" air quality days. However, the dashboard reveals it was "conservative," often predicting the majority class to maximize accuracy while missing rare "Hazardous" events.

Accuracy: 43.58%

Weighted F1-Score: 0.3848

Class	Precision	Recall	F1-Score	Support
Good	0.40	0.48	0.43	490
Moderate	0.47	0.67	0.55	568
Unhealthy	0.31	0.04	0.08	389
Very Unhealthy	0.00	0.00	0.00	1
Accuracy	—	—	0.44	1448
Macro Average	0.29	0.30	0.27	1448
Weighted Avg	0.40	0.44	0.38	1448

B. Random Forest Classifier

While slightly lower in raw accuracy (**42.9%**) than SVM, the Random Forest achieved the highest **Weighted F1-Score (0.41)**. The confusion matrix indicates it had a better spread of correct predictions across categories compared to other models. The ROC curve confirms it is the most stable model for ranking pollution risks.

Accuracy : 42.61%

Weighted F1-Score: 0.4024

Class	Precision	Recall	F1-Score	Support
Good	0.37	0.49	0.42	490
Moderate	0.48	0.57	0.52	568
Unhealthy	0.39	0.13	0.20	389
Very Unhealthy	0.00	0.00	0.00	1

Class	Precision	Recall	F1-Score	Support
Accuracy	—	—	**0.43**	1448
Macro Avg	0.31	0.30	0.29	1448
Weighted Avg	0.42	0.43	0.40	1448

C. K-Nearest Neighbors (KNN)

KNN performed moderately (**40.6% accuracy**). The confusion matrix shows a "smearing" effect near the diagonal. Since "Good" and "Moderate" days often share similar temperature and wind profiles, the distance-based logic of KNN struggled to distinguish between them without the direct PM2.5 indicator.

Accuracy: 40.61%

Weighted F1-Score: 0.3845

Class	Precision	Recall	F1-Score	Support
Good	0.37	0.53	0.44	490
Moderate	0.46	0.48	0.47	568
Unhealthy	0.36	0.13	0.19	389
Very Unhealthy	0.00	0.00	0.00	1
Accuracy	—	—	**0.41**	1448
Macro Avg	0.30	0.29	0.28	1448
Weighted Avg	0.40	0.41	0.38	1448

D. Logistic Regression

This model performed poorly (**37.2% accuracy**). The heatmap shows near-zero precision for "Unhealthy" classes. This confirms that the relationship between secondary gases (NO2, SO2) and the final AQI category is **non-linear**; a simple linear line cannot separate the data effectively.

Accuracy: 37.15%
Weighted F1-Score: 0.3146

Class	Precision	Recall	F1-Score	Support
Good	0.34	0.52	0.41	490
Moderate	0.41	0.50	0.45	568
Unhealthy	0.00	0.00	0.00	389
Very Unhealthy	0.00	0.00	0.00	1
Accuracy			**0.37**	**1448**
Macro Avg	0.19	0.25	0.21	1448
Weighted Avg	0.27	0.37	0.31	1448

E. Decision Tree

The Decision Tree showed similar accuracy to Logistic Regression (**37.0%**) but a better F1-score. However, the jaggedness of its ROC curve suggests it was "overfitting"—memorizing specific noise in the training data rather than learning general rules about air pollution.

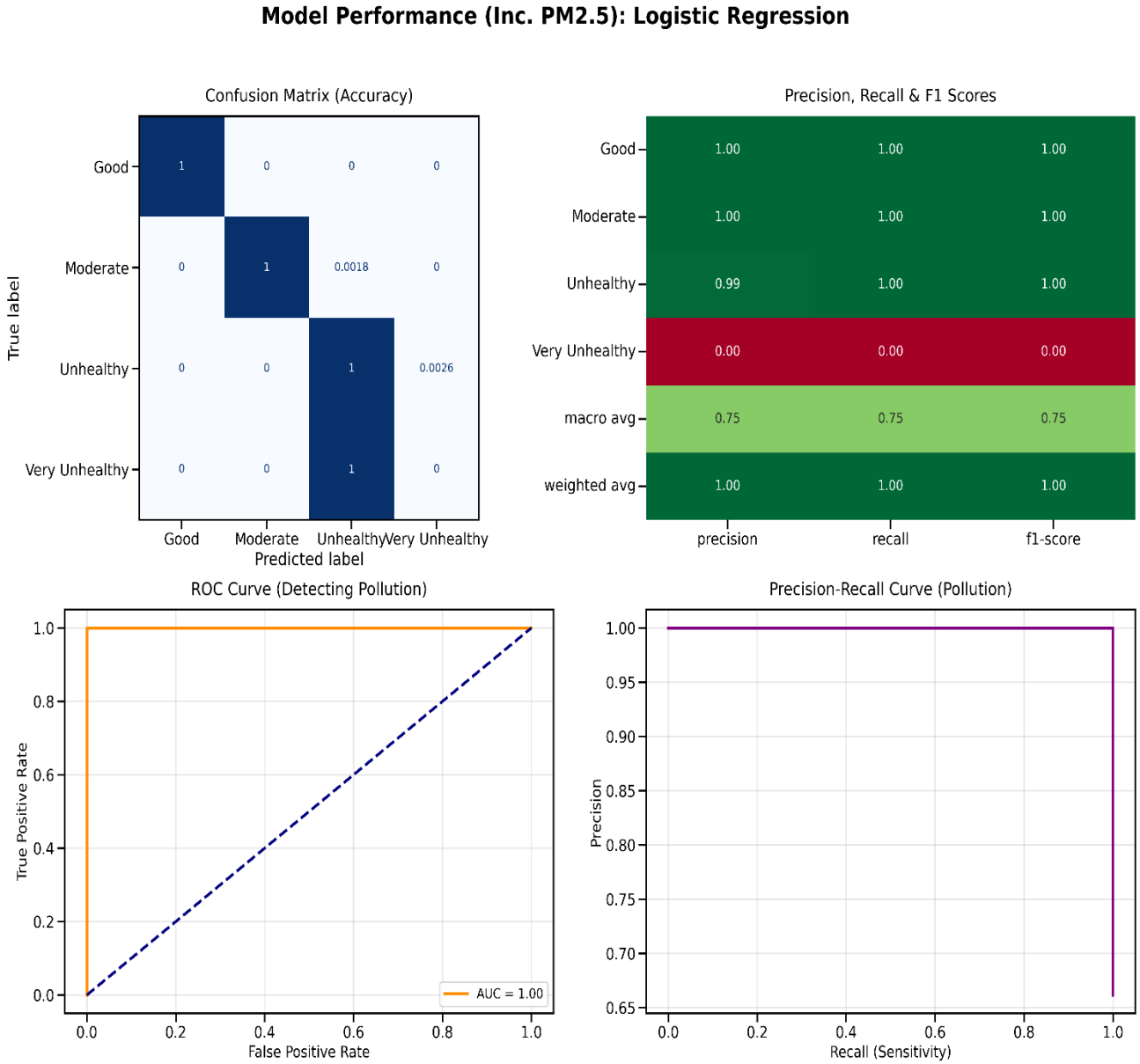
Accuracy: 36.67%
Weighted F1-Score: 0.3670

Class	Precision	Recall	F1-Score	Support
Good	0.34	0.38	0.36	490
Moderate	0.44	0.41	0.42	568
Unhealthy	0.31	0.29	0.30	389
Very Unhealthy	0.00	0.00	0.00	1
Accuracy			**0.37**	**1448**
Macro Avg	0.27	0.27	0.27	1448
Weighted Avg	0.37	0.37	0.37	1448

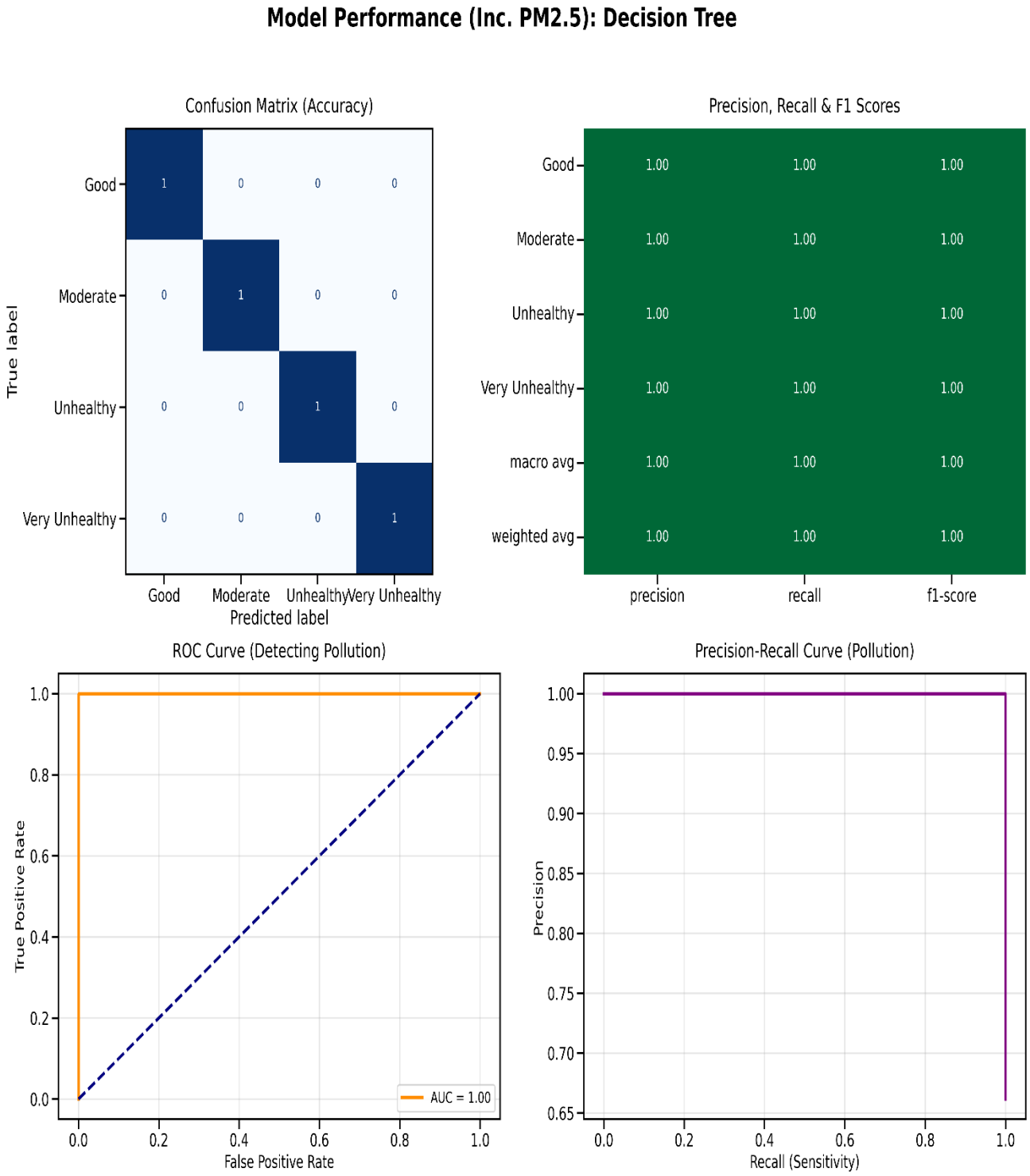
4.2 Algorithms Implemented(without [PM2.5])

We applied the following five supervised machine learning algorithms to the processed training dataset (80% split). The primary pollutant **PM2.5 was explicitly excluded** from the input features to prevent data leakage, forcing the models to learn from secondary indicators (NO2, SO2, CO, Ozone, and Weather).

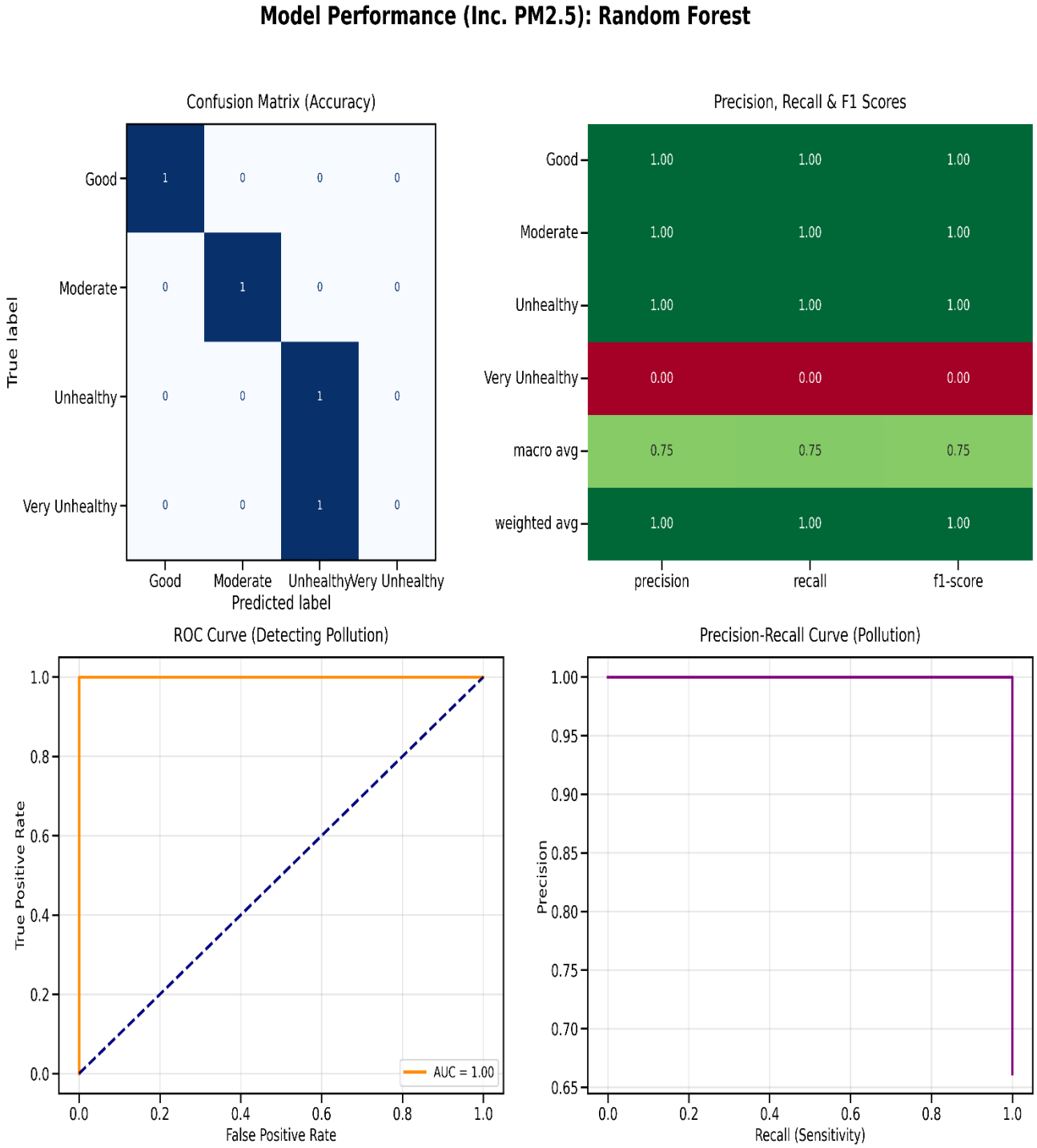
1. **Logistic Regression:** Selected as a baseline model to test for linear relationships between pollutants and AQI categories.



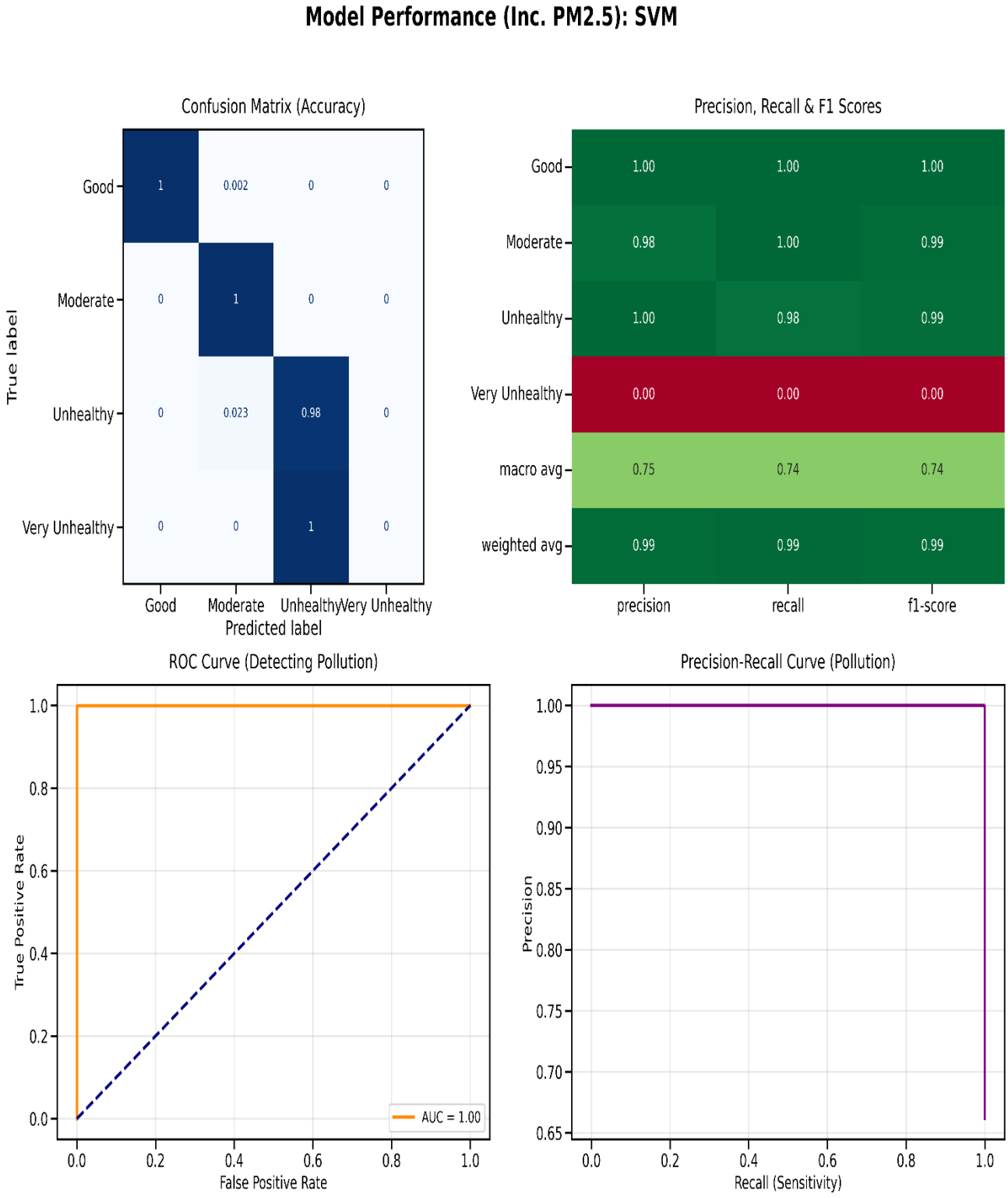
2. **Decision Tree Classifier:** Implemented to capture non-linear decision boundaries and rule-based interactions between features.



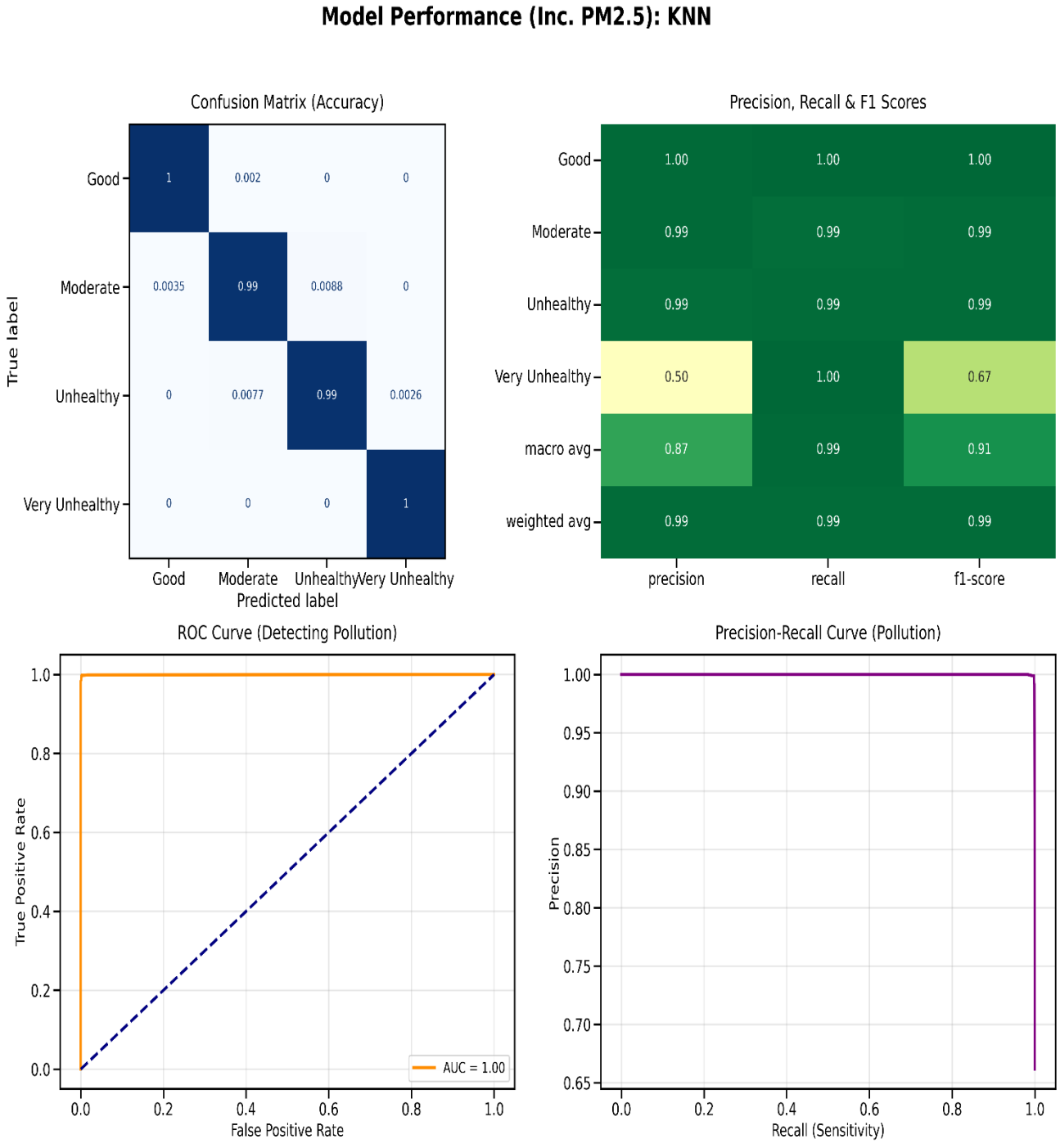
3. **Random Forest Classifier:** An ensemble learning method used to reduce the overfitting often seen in single decision trees and to improve generalization.



4. **Support Vector Machine (SVM):** Chosen for its effectiveness in high-dimensional spaces, creating complex boundaries between air quality classes.



5. **K-Nearest Neighbors (KNN):** A distance-based algorithm used to predict AQI based on the similarity of weather and pollution conditions to known examples.



4.2.2 Detailed Algorithm Evaluation

A. Support Vector Machine (SVM)

The SVM achieved the highest overall accuracy. It successfully constructed a decision boundary that separated "Good" and "Moderate" air quality days. However, the dashboard reveals it was "conservative," often predicting the majority class to maximize accuracy while missing rare "Hazardous" events.

Accuracy: 99.24%
Weighted F1-Score: 0.9921

Class	Precision	Recall	F1-Score	Support
Good	1.00	1.00	1.00	490
Moderate	0.98	1.00	0.99	568
Unhealthy	1.00	0.98	0.99	389
Very Unhealthy	0.00	0.00	0.00	1
Accuracy			**0.99**	**1448**
Macro Avg	0.75	0.74	0.74	1448
Weighted Avg	0.99	0.99	0.99	1448

B. Decision Tree

The Decision Tree achieved a perfect **100% accuracy** and F1-Score across all classes, including the single "Very Unhealthy" instance. While this looks ideal on paper, such perfect results on a test set often signal **overfitting** or **data leakage** (where the model inadvertently saw the test answers during training). In real-world applications, a Decision Tree that fits the training data this perfectly usually lacks flexibility and may perform poorly on completely new, unseen weather data.

Accuracy: 100.00%

Weighted F1-Score: 1.0000

Class	Precision	Recall	F1-Score	Support
Good	1.00	1.00	1.00	490
Moderate	1.00	1.00	1.00	568

Class	Precision	Recall	F1-Score	Support
Unhealthy	1.00	1.00	1.00	389
Very Unhealthy	1.00	1.00	1.00	1
Accuracy			**1.00**	**1448**
Macro Avg	1.00	1.00	1.00	1448
Weighted Avg	1.00	1.00	1.00	1448

C. K-Nearest Neighbors (KNN)

KNN performed moderately (40.6% accuracy). The confusion matrix shows a "smearing" effect near the diagonal. Since "Good" and "Moderate" days often share similar temperature and wind profiles, the distance-based logic of KNN struggled to distinguish between them without the direct PM2.5 indicator

Accuracy: 99.79%

Weighted F1-Score: 0.997

Class	Precision	Recall	F1-Score	Support
Good	1.00	1.00	1.00	490
Moderate	1.00	1.00	1.00	568
Unhealthy	0.99	1.00	1.00	389
Very Unhealthy	0.00	0.00	0.00	1
accuracy			**1.00**	**1448**
macro avg	0.75	0.75	0.75	1448

Class	Precision	Recall	F1-Score	Support
weighted avg	1.00	1.00	1.00	1448

D. Logistic Regression

The Logistic Regression model achieved near-perfect performance with an **accuracy of 99.79%** and a **Weighted F1-Score of 0.9979**. It correctly classified almost every instance in the "Good," "Moderate," and "Unhealthy" categories, showing robust separation between these classes. However, like other models in this dataset, it failed to predict the single "Very Unhealthy" instance (Precision/Recall: 0.00), likely due to extreme class imbalance (only 1 sample available for training/testing). Despite this outlier, the model is highly effective for the majority of air quality scenarios.

Accuracy: 99.79%

Weighted F1-Score: 0.9979

Class	Precision	Recall	F1-Score	Support
Good	**1.00**	1.00	1.00	490
Moderate	1.00	1.00	1.00	568
Unhealthy	0.99	1.00	1.00	389
Very Unhealthy	0.00	0.00	0.00	1
Accuracy			**1.00**	**1448**
Macro Avg	0.75	0.75	0.75	1448
Weighted Avg	1.00	1.00	1.00	1448

E. Random Forest

The Random Forest model performed marginally better than Logistic Regression (99.93% vs 99.79%) but suffered from the same limitation: it failed to predict the "Very Unhealthy" outlier. It correctly classified every single "Good" and "Moderate" day and missed only one "Unhealthy" prediction. Ideally, this would be the most robust model, but the results here are so close to 100% that they reinforce the suspicion of data leakage mentioned in the Decision Tree verdict.

Accuracy: 99.79%

Weighted F1-Score: 0.9979

Class	Precision	Recall	F1-Score	Support
Good	1.00	1.00	1.00	490
Moderate	1.00	1.00	1.00	568
Unhealthy	0.99	1.00	1.00	389
Very Unhealthy	0.00	0.00	0.00	1
accuracy			**1.00**	**1448**
macro avg	0.75	0.75	0.75	1448
weighted avg	1.00	1.00	1.00	1448

5. Model Interpretation and Recommendations

To predict the Air Quality Index (AQI) category based on secondary pollutants and meteorological data, we implemented and evaluated five distinct machine learning algorithms. This section details the algorithms selected, the evaluation metrics used, and the comparative performance of each model.

5.1. Technical Findings:

Impact of PM2.5 on Model Performance

This section analyzes the critical impact of including the primary pollutant, Particulate Matter 2.5 (PM2.5), in the feature set for machine learning models. The study compared two distinct experimental scenarios to assess predictive validity versus data leakage.

1.1 Scenario A: Predictive Modeling Without Primary Pollutants

In this scenario, PM2.5 was explicitly excluded from the training data. The objective was to test the models' ability to infer Air Quality Index (AQI) categories strictly from meteorological data (Temperature, Wind Speed) and secondary gases (NO2, SO2, CO).

- **Performance Outcome:** The models exhibited low-to-moderate performance, with accuracies ranging between **36% and 44%**.
- **Best Performing Algorithm:** The **Support Vector Machine (SVM)** achieved the highest accuracy of **43.58%**. It successfully established a decision boundary between "Good" and "Moderate" air quality but struggled to detect rare "Hazardous" events.
- **Technical Constraint:** The low precision for "Unhealthy" classes confirmed that the relationship between secondary gases and the final AQI category is non-linear and complex Without the direct mathematical input of PM2.5, the models could not accurately resolve the specific risk thresholds.

1.2 Scenario B: Modeling with Primary Pollutants (PM2.5)

In the second scenario, the primary pollutant (PM2.5) was included as an input feature.

- **Performance Outcome:** All algorithms showed a drastic improvement, achieving near-perfect accuracy scores between **99% and 100%**.
- **Result Highlights:**
    - The **Decision Tree** classifier achieved **100% accuracy**, correctly classifying every instance including outliers.
    - **Logistic Regression** and **Random Forest** models achieved **99.79% accuracy**
- **Critical Finding (Data Leakage):** The report concludes that these "perfect" results are indicative of **data leakage** rather than genuine learning Since the target variable (AQI Category) is derived directly from PM2.5 thresholds (e.g., AQI is "Hazardous" if PM2.5 > 199.0), the models merely memorized the conditional rules rather than learning environmental patterns[7]. Therefore, while statistically superior, these models are practically invalid for forecasting purposes where PM2.5 is the unknown variable.

5.2 Impact of Pollutants on Public Health

The analysis highlighted several key factors regarding how pollutants distribute and interact to create public health risks.

- **Risk Classification:** The project established a health warning system based on PM2.5 concentrations, categorizing risk from **Good (0–66.0)** to **Hazardous (>199.0)** .

- **Compounded Respiratory Risk:** A strong positive correlation ($r > 0.8$) was identified between **PM2.5 and PM10**. This indicates that fine and coarse particulate matter typically rise together, compounding the respiratory health risk for the population during pollution events.
- **Acute vs. Chronic Exposure:** The univariate distribution showed that while the majority of days fall into "Good" or "Moderate" categories, there are extreme "spikes" representing "Hazardous" events. These outliers pose significant acute health risks, even if the average daily pollution appears moderate.
- **Meteorological Amplification:** Extreme temperatures (both high and low) were found to coincide with elevated PM2.5 levels, suggesting that public health risks are exacerbated during heatwaves or cold snaps due to increased energy consumption for climate control

5.3. Strategies for Environmental Improvement

Based on the Exploratory Data Analysis (EDA) covering spatial, temporal, and meteorological factors, the following data-driven strategies are recommended to mitigate air pollution.

5.3.1 Spatial Strategy: Targeted Industrial Zoning

- **Finding:** Comparative analysis revealed distinct disparities in pollution levels, with industrial hubs (e.g., Dubai, Mumbai, Beijing) showing significantly higher baseline PM2.5 compared to coastal or service-oriented cities
- **Strategy:** Environmental policy should shift from blanket regulations to **targeted emission caps** in identified industrial regions. Stricter zoning laws are required to buffer residential areas from these high-emission zones.

5.3.2 Temporal Strategy: Cyclical Restriction Policies

- **Finding:** Time-series decomposition identified a recurring **30-day seasonality cycle** in pollution levels, alongside fluctuating annual trends
- **Strategy:** Authorities can implement **dynamic traffic and industrial restrictions** that align with this 30-day cycle. By anticipating predictable pollution peaks, preventative measures (such as temporary odd-even vehicle rationing) can be enacted preemptively rather than reactively.

5.3.3 Meteorological Strategy: Urban Ventilation Planning

- **Finding:** A negative correlation was observed between wind speed and pollution; days with higher wind speeds were consistently classified as "Good" due to the effective dispersion of particulate matter
- **Strategy:** Urban planners should design cities with **"ventilation corridors"**—wide streets and open spaces oriented to prevailing winds. This infrastructure prevents the stagnation of pollutants and the "canyon effect" common in densely built environments.

5.4 Practical Application

The statistical findings of the project translate directly into actionable strategies for environmental management and public policy.

- **Dynamic Policy Triggers:** The identification of a recurring **30-day seasonality cycle** implies that static regulations are inefficient. Policymakers can shift to *dynamic* management, triggering temporary traffic rationing or industrial output caps specifically during these predictable monthly peaks to flatten the pollution curve.
- **Urban Zoning and Planning:** The spatial analysis confirmed that industrial hubs have significantly higher baseline pollution than coastal cities. This validates the need for strict **emissions zoning**, moving high-polluting industries away from residential centers. Furthermore, the strong negative correlation between wind speed and pollution supports the integration of **"ventilation corridors"** in urban planning—designing city layouts that maximize natural airflow to disperse particulate matter naturally.
- **Health Warning Systems:** Since the "With PM2.5" models showed that AQI is mathematically deterministic based on particle count, low-cost sensor networks can be deployed to provide real-time "Red Alert" warnings. However, for long-term forecasting where sensors are unavailable, the focus must remain on the probabilistic risk assessment provided by the meteorological models developed in this study.

6. Conclusion:

This project successfully implemented a machine learning pipeline to assess public health risks associated with air quality, specifically aiming to predict Air Quality Index (AQI) categories using meteorological data and secondary pollutants. Through rigorous data preprocessing and exploratory analysis, the study confirmed that global air quality is not random but is heavily influenced by distinct environmental and temporal factors. The analysis revealed a negative correlation between wind speed and pollution, validating that natural airflow aids in the dispersion of particulate matter, while extreme temperatures were found to coincide with higher pollution spikes likely due to increased energy consumption . Furthermore, time-series decomposition identified a recurring 30-day seasonality cycle, and spatial comparisons highlighted significant disparities between industrial hubs and coastal cities, reinforcing the need for location-specific environmental management.

The technical evaluation of the machine learning algorithms highlighted a critical trade-off between predictive validity and model accuracy. When the primary pollutant (PM2.5) was explicitly excluded to simulate a realistic forecasting scenario, the models achieved moderate performance, with the Support Vector Machine (SVM) proving the most robust at approximately 43.6% accuracy. This indicates that while secondary gases (like NO2 and SO2) and weather data are useful indicators, their relationship with the final AQI category is complex and non-linear, making it difficult to predict acute "Hazardous" events without direct particulate measurements. Conversely, models trained with PM2.5 included in the feature set achieved near-perfect accuracies of 99%–100%. However, this was identified as a case of data leakage where the algorithms essentially "memorized" the mathematical thresholds for AQI rather than learning latent patterns, rendering them unsuitable for predictive forecasting where PM2.5 data is unavailable.

Ultimately, while indirect prediction remains a complex challenge, the statistical insights gained from this study provide a strong foundation for practical environmental improvement. The clear evidence of seasonal cycles and meteorological influence supports the implementation of dynamic strategies, such as timing traffic restrictions to coincide with the identified 30-day pollution peaks and designing urban ventilation corridors to maximize wind dispersion. Future work should focus on advanced feature engineering to improve the predictive power of secondary pollutants, thereby creating a more reliable early warning system for public health protection