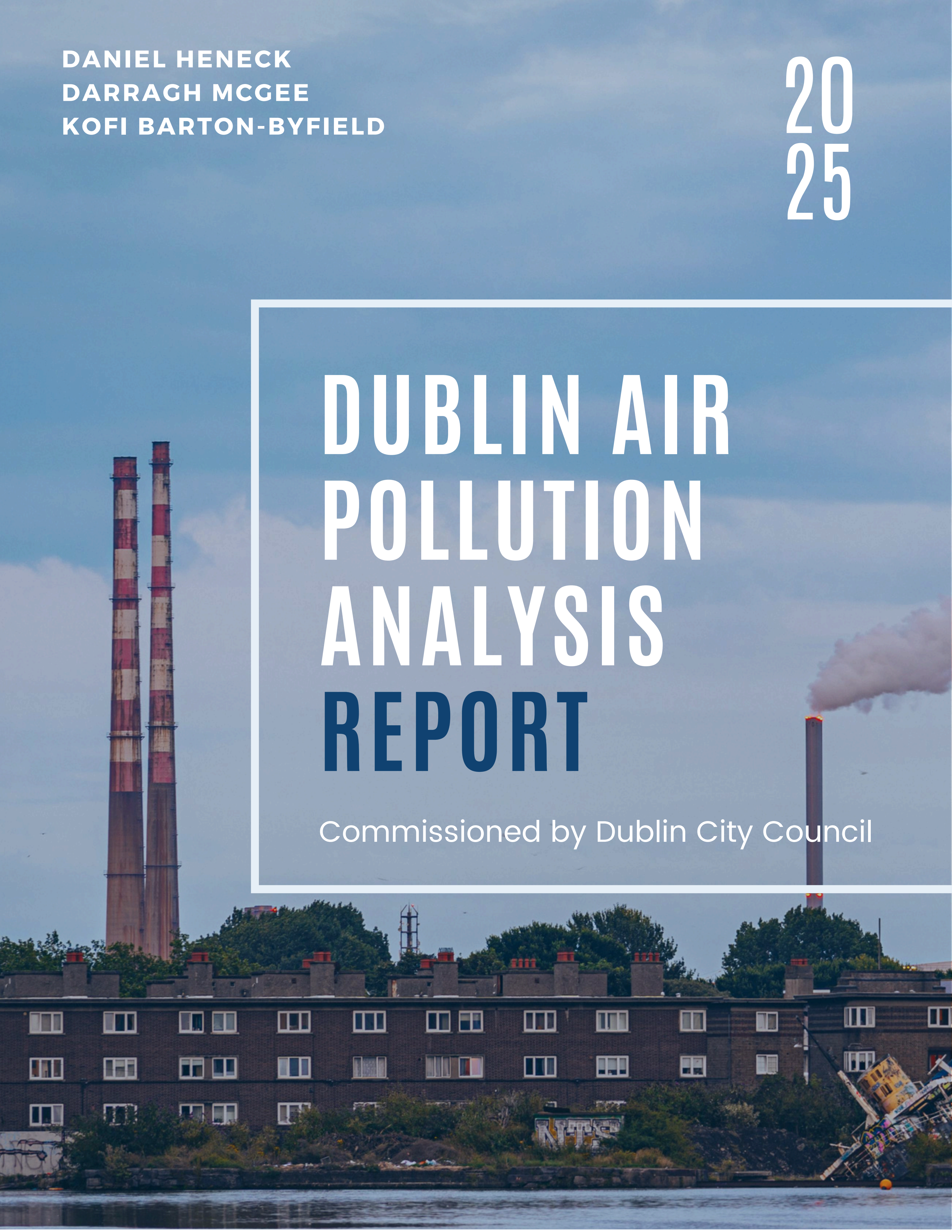


DANIEL HENECK
DARRAGH MCGEE
KOFI BARTON-BYFIELD

2025

DUBLIN AIR POLLUTION ANALYSIS REPORT

Commissioned by Dublin City Council

FUTURE
DATA
DUBLIN

Project Overview and Dataset Assessment

Report Purpose

This report was commissioned by Dublin City Council to investigate key factors contributing to air pollution in Dublin and to support the development of actionable strategies for improving air quality.

Data Exploration

The DCC Google-AirView dataset includes over 5 million data points, capturing second-by-second measurements of Carbon Monoxide (CO), Carbon Dioxide (CO_2), Nitrogen Dioxide (NO_2), Nitric Oxide (NO), Ozone (O_3), and Particulate Matter (PM2.5), including particle counts from 0.3–2.5 µm. Data were collected along typical urban routes in Dublin during weekday working hours (9:00–17:00).

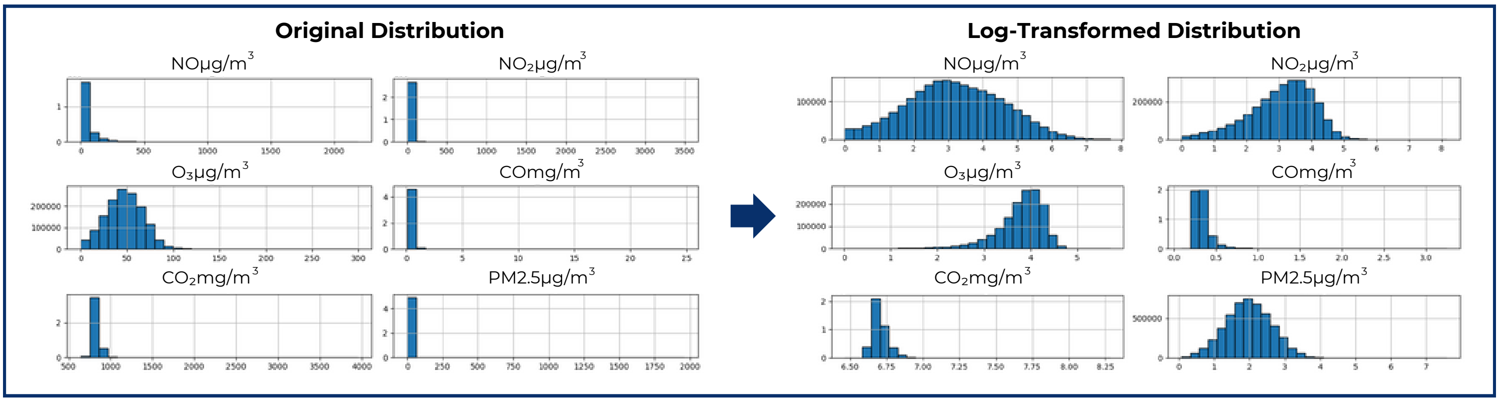
Summary statistics and missingness proportions (Table 1) were used to assess the structure and completeness of the data. PM2.5 showed the lowest proportion of missing data across all pollutants, reducing the risk of bias that can arise from missingness not at random (such as sensor dropouts during pollution spikes) which could bias reliability for downstream analysis.

Variable	Count	Missing %	Mean	SD
CO	4,708,874	6.4%	0.42	0.24
CO_2	4,131,647	17.9%	816.65	52.71
NO_2	3,865,035	23.2%	18.32	34.64
NO	4,611,444	8.3%	20.94	103.19
O_3	1,447,372	71.2%	46.83	20.35
PM2.5	4,940,266	1.8%	7.98	8.55

Table 1: Summary Statistics for Pollutants

Figure 1 shows the original and log-transformed pollutant distributions. Due to strong positive skew and wide value ranges, a log(1 + x) transformation was applied to improve interpretability. Among the transformed variables, PM2.5 stands out with a near-normal distribution, minimal skew, and a concentrated range. In contrast, CO_2 and CO remain right-skewed, while NO, NO_2, and O_3 show partial normalisation but retain some asymmetry.

Figure 1: Original and Log-Transformed Distributions



PM2.5 was selected as the target variable for further analysis based on three key factors: it exhibited the lowest proportion of missing; it showed a near-normal distribution after log transformation; and it has strong public health relevance, as PM2.5 is identified by the World Health Organization as the most harmful pollutant due to its association with serious health effects.

Temporal Structure

Figure 2 shows the mean hourly PM2.5 concentrations, with a clear peak observed during evening "rush hour" traffic.

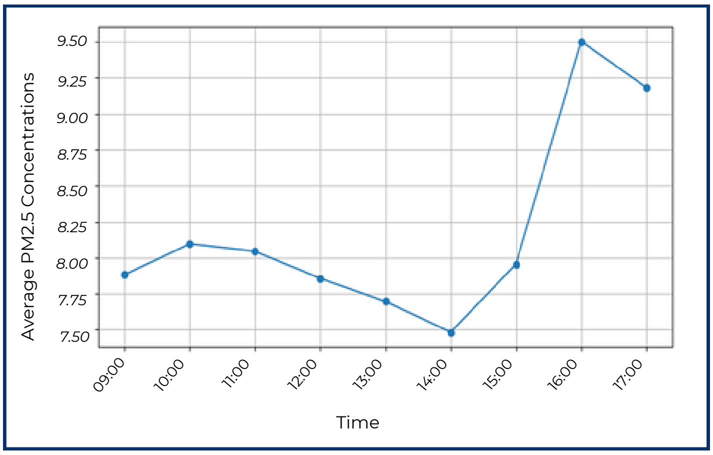


Figure 2: Hourly Variation in PM2.5

Figure 3 plots average monthly PM2.5 concentrations, capturing seasonal variation across the dataset period.

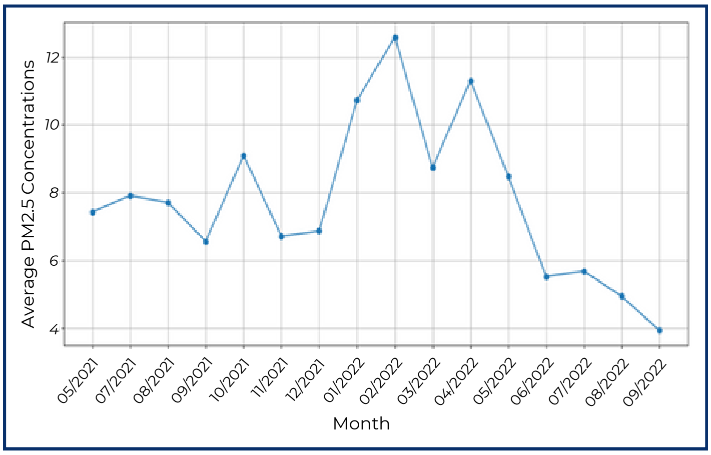


Figure 3: Monthly Variation in PM2.5

Positive and Negative Aspects of the Dataset

Positive Aspects	Negative Aspects
High Temporal Resolution: Over 5 million records collected at one-second intervals.	**Limited Feature Set:** Contains only pollutant readings; lacks explanatory variables.
Broad Pollutant Coverage: PM2.5, NO_2, CO_2, and O_3 are included.	**Lack of Standardisation:** Variables require transformation due to scale differences.
Wide Geographic Scope: Enables spatial comparisons across Dublin.	**Missing Data:** Substantial gaps in O_3 (71%), NO_2 (23%), and CO_2 (17%).

Overall Suitability

The dataset is best suited for high-resolution analysis of PM2.5, given its strong spatial-temporal coverage and low missingness. Incomplete data for other pollutants limit its effectiveness for broader multivariate air quality analysis. The lack of contextual variables further reduces its standalone utility, requiring supplementation from external sources for explanatory modeling.

Factors Influencing Air Pollution

FUTURE DATA DUBLIN

Selection of Explanatory Variables

To investigate air pollution in Dublin, two key factors were selected: weather conditions and road traffic volume. Both are widely recognised in research and policy as major influences on air quality, and their analysis can inform the development of targeted strategies to reduce pollution exposure.

Weather Conditions

Rationale: Meteorological variables have a well-documented impact on air pollution. Temperature, wind speed, humidity, and precipitation can affect the dispersion, dilution, and accumulation of airborne pollutants. For instance, wind promotes dispersion, while stagnant conditions or temperature inversions may intensify pollution.

Dataset: Hourly weather data from Met Éireann Historical Data, covering: Temperature (°C), Wind speed (km/h), Relative humidity (%), and Precipitation (mm).

Traffic Volume

Rationale: Road traffic is a primary source of urban particulate pollution. PM2.5 concentrations often rise during peak commuting hours and in areas with high vehicle density. This relationship is well-established in environmental literature and reflected in preliminary time-series patterns.

Dataset: Vehicle count data from Smart Dublin – Traffic Volume Counts, including: Hourly traffic volume, and Road location and classification.

Correlation Analysis

Weather and traffic data were merged with the Google AirView dataset by aligning timestamps and location references. A correlation matrix (Figure 4) was produced using PM2.5 concentrations and selected predictor variables.

Key findings included:

- Positive correlations with several weather variables, including temperature (0.36), wet bulb temperature (0.41), and vapour pressure (0.38)
- A weak negative correlation with traffic volume (Sum Volume: −0.20; Avg Volume: −0.17), which may reflect dispersion effects during periods of higher traffic.

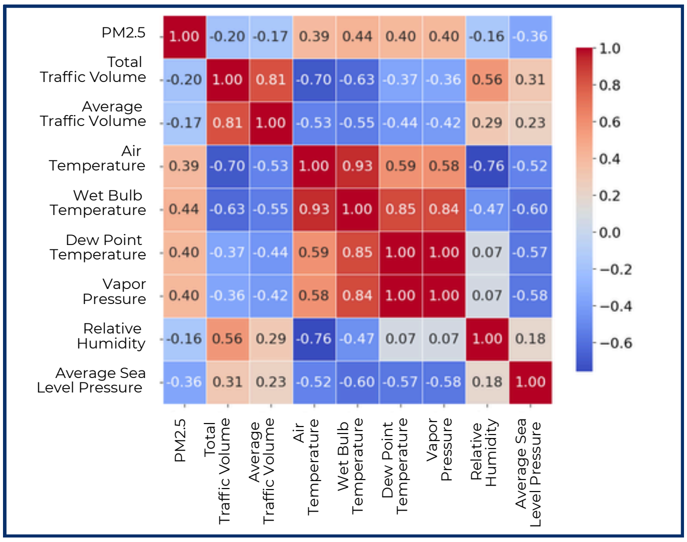


Figure 4: Correlation Matrix of PM2.5 vs. Selected Variables

Summary of Factor Selection

- Weather and traffic were selected for their relevance to observed patterns in PM2.5 concentrations
- The datasets are publicly available, regularly updated, and compatible with the AirView data
- Their integration supports both exploratory analysis and predictive modelling

Model Development and Evaluation

Selection of Influencing Variables

A supervised machine learning framework was employed to examine the relationship between meteorological and traffic-related variables and the likelihood of high pollution days in Dublin. This modelling effort supports data-driven strategies for improving urban air quality, with PM2.5 selected as the primary pollutant of interest.

Data Preparation and Binary Classification

A binary classification label was constructed by categorising days as either Good Pollution (0) or Bad Pollution (1), based on the upper quartile of logged PM2.5 values. This threshold ensured that only days with the most elevated concentrations were assigned to the “bad” category.

Data preprocessing included the removal of outliers beyond three standard deviations and the exclusion of rows containing missing values, resulting in the removal of 6,841 observations. All remaining numeric features were normalised to a [0, 1] range to ensure consistent scale and facilitate model convergence. Class balance was also assessed, with roughly equal proportions of Good and Bad Pollution days, reducing the risk of classifier bias.

Model Development and Evaluation

Four classification models were deployed to predict high pollution days: Logistic Regression, Random Forest, XGBoost, and Decision Tree. Each model was trained on 80% of the data and evaluated on the remaining 20%, using a fixed random seed to ensure reproducibility. The models used a consistent set of features, including *Total Traffic Volume, Average Traffic Volume, Air Temperature, Wet Bulb Temperature, Dew Point Temperature, Vapour Pressure, Relative Humidity*, and *Mean Sea Level Pressure*. Optimised hyperparameters were used for fair model comparison.

Model performance was assessed using four key metrics: accuracy, precision, F1-score, and ROC-AUC, as summarised in Table 2. The Decision Tree model achieved the highest overall accuracy (0.80) and F1-score (0.70), indicating balanced classification across both pollution categories. Random Forest achieved the highest ROC-AUC (0.82), reflecting strong discriminative performance across thresholds. XGBoost recorded the highest precision (0.81), highlighting its effectiveness in reducing false positives.

	Accuracy	Precision	F1-Score	AUC-ROC
Logistic Regression	0.76	0.80	0.59	0.76
Random Forest	0.79	0.78	0.69	0.82
XGBoost	0.79	0.81	0.67	0.81
Decision Tree	0.80	0.77	0.70	0.78

Table 2: Model Performance in Predicting High Pollution Days

Confusion Matrix Analysis

To further assess classification performance, confusion matrices were examined for each model (Table 3). Tree-based classifiers achieved more accurate detection of high pollution days, with higher true positive rates and fewer false negatives than Logistic Regression. The Decision Tree model recorded the lowest number of false negatives (4,768), while Random Forest provided a strong overall balance between precision and recall. These results reinforce the earlier performance metrics and support the reliability of non-linear models in this classification task.

	TP	FP	TN	FN
Logistic Regression	6,271	1,544	20,727	6,993
Random Forest	8,155	2,250	20,021	5,109
XGBoost	7,481	1,706	20,565	5,783
Decision Tree	8,496	2,514	19,757	4,768

Table 3: Confusion Matrix Components for Each Model

ROC Curves

Figure 5 presents the ROC curves for all four models. Tree-based models consistently outperformed Logistic Regression by maintaining higher true positive rates across a range of false positive thresholds. This visual confirms that models capable of capturing non-linear relationships are better suited to this classification problem. The close proximity of the curves for the tree-based models also suggests that all three offered similarly strong discriminative performance.

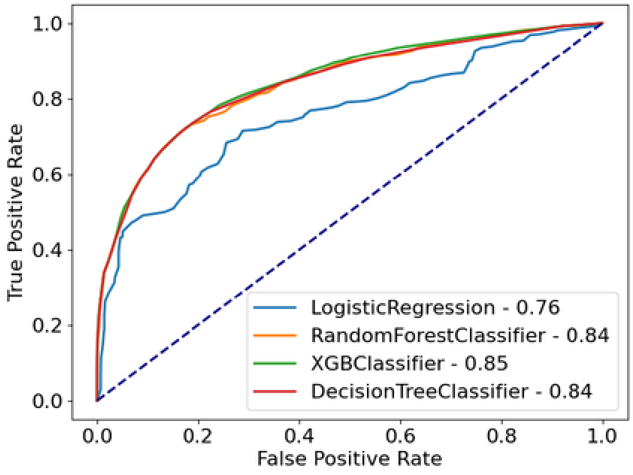


Figure 5: ROC Curves for All Classification Models

Feature Importance and Interpretation

Feature importance scores from all four models are shown in Figure 6. *Air Temperature* emerged as the most influential variable, particularly in the tree-based models. Other consistently important predictors included *Wet Bulb Temperature, Mean Sea Level Pressure*, and *Relative Humidity*. In contrast, *Total Traffic Volume* and *Average Traffic Volume* contributed minimally across all models. These results suggest that meteorological factors are more critical than traffic-related features in predicting daily PM2.5 concentrations.

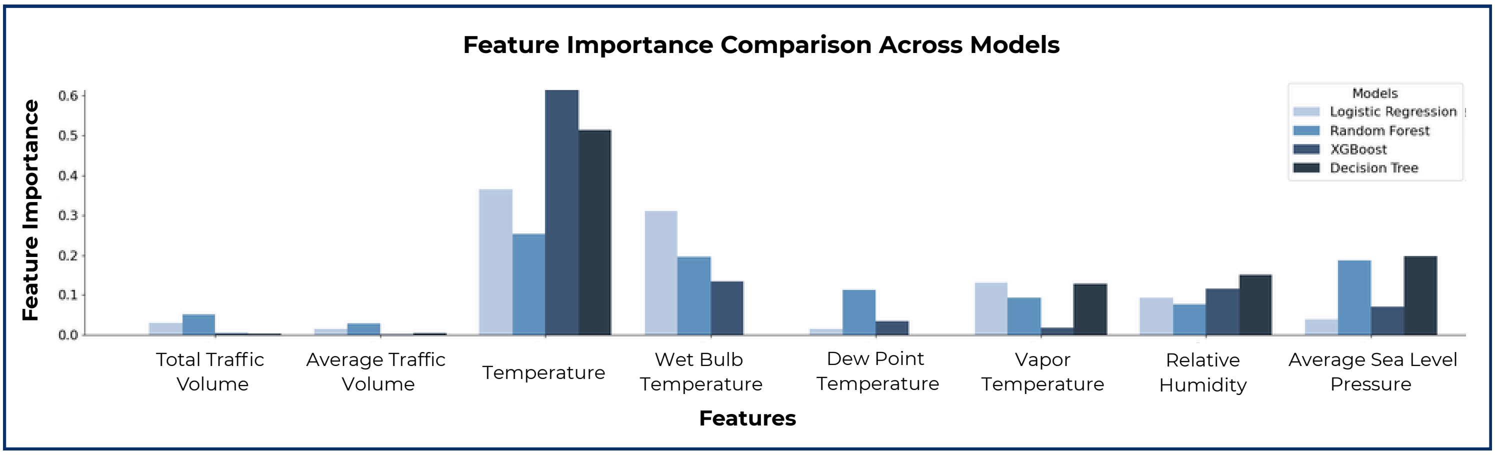


Figure 6: Feature Importance Comparison across Classification Models

Model Selection

- Based on overall performance across key evaluation metrics, the Random Forest classifier is recommended as the most effective model for predicting high pollution days. It achieved a strong balance of accuracy (0.79), F1-score (0.69), and ROC-AUC (0.82), with a high true positive rate and relatively few false positives. Its ability to model non-linear relationships and provide interpretable feature importance scores further enhances its suitability for this task.
- These characteristics make it a strong candidate for integration into real-time air quality forecasting systems and public alert platforms, particularly in an urban policy context such as that of Dublin City Council.

Recommendations/ Solutions

FUTURE DATA DUBLIN

1. Establish Fixed Monitoring Stations in High-Pollution Zones

Limited data availability for pollutants such as NO_2 and O_3 constrained the scope of this analysis. The modelling process underscored the importance of reliable, continuous pollutant data. While PM2.5 was the focus of this analysis, the absence of consistent NO_2 measurements limited the ability to assess traffic-related impacts, despite their known significance.

Recommendation:
Install fixed-site monitoring stations in high-pollution areas to provide continuous, multi-pollutant data (PM2.5, NO_2, O_3, CO_2). This would support more accurate forecasting, enable multi-pollutant modelling, and inform targeted interventions. *Example: London's ULEZ pairs traffic regulation with continuous monitoring and has successfully reduced NO_2 levels.*

2. Implement a Forecast-Based Air Quality Alert System

PM2.5 levels in Dublin were more strongly influenced by meteorological conditions than by traffic volume. These fluctuations create short-term health risks, particularly for vulnerable populations. This suggests that PM2.5 pollution management should prioritise weather-informed forecasting to enable timely public alerts and targeted interventions.

Recommendation:
Develop a Random Forest-based, forecast-driven alert system to identify high-pollution periods and locations, and inform the public—particularly vulnerable groups. Alerts should be delivered via mobile apps or text messages and include clear behavioural guidance (e.g. limiting outdoor activity, mask use) to help reduce exposure. *Example: South Korea's national system uses real-time forecasts to inform the public and guide daily exposure decisions.*