مجلة ساحل المعرفة للعلوم الإنسانية والتطبيقية
Sahel Almarifah Journal of Humanities and Applied Sciences
تصدر عن الاكاديمية الليبية فرع الساحل الغربي
المجلد الثاني–عدد خاص-S1
الصفحات (E-828 - E-798)

Classification of Air Pollution Risk Levels Using a Soft Voting Ensemble Model Based on Real-World Air Quality Monitoring Data

Maher Shoikh [1,*], Ghadeer Suleiman [2]
[1] Department of Forestry Sciences, Bryansk State University of Engineering and Technology, Bryansk -Russia
[2] Department of Computer Science, Prince Hussein Bin Abdullah II Faculty of Information Technology, Al al-Bayt University, Mafraq- Jordan
.
[*]Corresponding author: mahershoikh@gmail.com

تصنيف مستويات مخاطر تلوث الهواء باستخدام نموذج التصويت المرن الجماعي بناءً على بيانات رصد جودة الهواء الواقعية

ماهر شويخ [1،*] غدير سليمان [2]
[1] قسم علوم الغابات، جامعة بريانسك الحكومية للهندسة والتكنولوجيا، بريانسك -روسيا
[2] قسم علوم الحاسوب، كلية الأمير حسين بن عبد الله الثاني لتكنولوجيا المعلومات، جامعة آل البيت، المفرق - الأردن.

Abstract

Air pollution is one of the major environmental and public health challenges in rapidly urbanizing regions around the world. Elevated concentrations of airborne pollutants such as particulate matter and gaseous emissions can threaten both human health and environmental sustainability. Thus, accurate recognition of pollution risk degrees is vital for developing environmental monitoring systems and supporting decision-making in urban environmental management. We present an ensemble machine learning-based framework to classify the levels of air pollution risk using environmental, meteorological and temporal indicators extracted from real-world air quality monitoring data collected through different urban locations. Our dataset consists of major atmospheric pollutants as well as meteorological variables capturing the air pollution processes in urban areas over varying seasons. Initially evaluated several machine learning algorithms like Random Forest, Extra Trees, Support Vector Machine, Logistic Regression and Extreme Gradient Boosting. A Soft Voting ensemble model was then designed to combine the prediction strengths of all best-performing classifiers. The proposed model attained an accuracy of about 82.7% with the weighted F1-score being 0.828, thus performing better than any single models. Cross-validation validated the framework's robustness and stability, allowing analysis of feature importance to highlight PM2. 5 = the most important determinant of pollution risk levels. The findings highlight the utility of ensemble machine learning methodologies for environmental monitoring, providing greater insight into pollution exposure and informing data-driven decision making to promote sustainable air quality management.

Keywords: Air pollution; Air quality index; Ensemble learning; Environmental risk assessment; Urban environmental monitoring.

الملخص

يُعد تلوث الهواء أحد أبرز التحديات البيئية والصحية العامة في المناطق الحضرية سريعة التوسع حول العالم. إذ يُمكن أن تُهدد التركيزات المرتفعة للملوثات المحمولة جوًا، مثل الجسيمات الدقيقة والانبعاثات الغازية، صحة الإنسان واستدامة البيئة على حدٍ سواء. لذا، يُعدّ التحديد الدقيق لدرجات مخاطر التلوث أمرًا بالغ الأهمية لتطوير أنظمة الرصد البيئي ودعم عملية صنع القرار في الإدارة البيئية الحضرية. نقدم إطار عمل قائم على التعلم الآلي لتصنيف مستويات مخاطر تلوث الهواء باستخدام مؤشرات بيئية وأرصادية وزمنية مُستخرجة من بيانات رصد جودة الهواء الواقعية التي جُمعت من مواقع حضرية مختلفة. تتكون مجموعة بياناتنا من ملوثات جوية رئيسية، بالإضافة إلى متغيرات الأرصاد الجوية التي تُغطي عمليات تلوث الهواء في المناطق الحضرية على مدار فصول السنة المختلفة. في البداية، قمنا بتقييم العديد من خوارزميات التعلم الآلي، مثل الغابة العشوائية، والأشجار الإضافية، وآلة المتجهات الداعمة، والانحدار اللوجستي، وتعزيز التدرج الشديد. ثم صُمم نموذج تجميعي للتصويت الناعم لدمج نقاط قوة التنبؤ لأفضل المصنفات أداءً. حقق النموذج المقترح دقة بلغت حوالي 82.7%، مع قيمة F1 المرجحة 0.828، متفوقًا بذلك على جميع النماذج الفردية. وقد أكدت عملية التحقق المتبادل متانة واستقرار الإطار، مما سمح بتحليل أهمية الميزات لتسليط الضوء على أن الجسيمات الدقيقة PM2.5 هي العامل الأكثر أهمية في تحديد مستويات مخاطر التلوث. تُبرز هذه النتائج فائدة منهجيات التعلم الآلي الجماعية في الرصد البيئي، إذ توفر فهمًا أعمق للتعرض للتلوث، وتُسهم في اتخاذ قرارات قائمة على البيانات لتعزيز الإدارة المستدامة لجودة الهواء.

الكلمات الدالة: تلوث الهواء؛ مؤشر جودة الهواء؛ التعلم الجماعي؛ تقييم المخاطر البيئية؛ الرصد البيئي الحضري.

1. Introduction

One of the biggest threats to the environment and public health in the globe today is air pollution. The decline of air quality in many urban areas has been mostly caused by rapid urbanization, industrial growth, and an increase in transportation [1]. Exposure to high levels of airborne pollutants, including both particulate matter (PM_{10}, $PM_{2.5}$) and harmful gases such as nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3), has been linked to a number of detrimental health outcomes, such as respiratory conditions, heart problems, and early death [2], [3].

Therefore, these gases are also a core focus of this research, not merely the solid particles.

In response to the growing concern over air pollution, both Arab and foreign governments, along with environmental organizations, have implemented monitoring systems intended to quantify pollutant concentrations and evaluate environmental risks. The Air Quality Index (AQI). The Air Quality Index (AQI), which compiles data from various contaminants to categorize air quality situations and alert the public of possible health hazards, is one often used indicator [4,5,6].

However, because of the intricate relationships between pollutant concentrations, weather patterns, and temporal fluctuations, it is still difficult to effectively estimate and forecast air pollution danger levels. Numerous interrelated factors, such as weather, seasonal fluctuation, atmospheric processes, and geographic features, affect the dynamics of air pollution. These intricacies necessitate sophisticated analytical techniques that can recognize nonlinear interactions in environmental datasets [7, 8].

1.2 Literature Review

The application of machine learning algorithms for environmental risk assessment and air quality prediction has been widely studied in recent years.

A voting-based ensemble classifier was created in a study by [9] to forecast air quality levels using ambient air quality data gathered in Bangalore over a six-year period (2014–2019). Several machine learning techniques, such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Logistic Regression, were integrated into the suggested framework. The effectiveness of ensemble learning techniques for air quality prediction was highlighted by

the results, which showed that the voting ensemble model outperformed individual models in terms of predictive accuracy.

The creation of inexpensive monitoring systems for air pollution assessment is the focus of another line of inquiry. A self-made air quality monitoring system based on ESP32 microcontrollers and the ESP-NOW communication protocol was presented in [10].
was suggested as a self-made air quality monitoring system built on ESP32 microcontrollers and the ESP-NOW communication protocol. Pollutant concentrations, including NO_2, SO_2, O_3, CO, CO_2, and H_2S, were measured by the system using Alphasense sensors. The system demonstrated the potential of low-cost monitoring technologies by producing results similar with professional monitoring devices, according to experimental validation.
The classification of the Air Quality Index (AQI) has also made extensive use of machine learning algorithms. In [11], a number of machine learning algorithms were assessed for AQI classification using data gathered from several monitoring stations, including Logistic Regression, Decision Trees, KNN, Random Forest, Support Vector Machine, and Naïve Bayes. The study demonstrated the efficacy of machine learning techniques in air pollution analysis by using random oversampling strategies to overcome class imbalance issues and achieving a classification accuracy surpassing 99%.
For the purpose of monitoring air quality, new sensing technologies have also been investigated. In [12], gaseous pollutants and particle matter were measured along transportation routes and atmosphere vertical profiles using drones fitted with laser sensors. The findings demonstrated that drone-based monitoring systems can supplement conventional monitoring stations and offer useful information for the analysis of regional pollution.

Typical Sniffer4D ecosystem. Source: Soarability.com

Additionally, crowdsourced and ensemble machine learning algorithms have been suggested to increase the accuracy of air pollution forecasts. Several models, including as Decision Trees, Random Forests, Neural Networks, and AdaBoost, were assessed in [13]. When compared to individual algorithms, ensemble techniques showed better prediction accuracy.

Furthermore, the success of combining inexpensive sensors with Internet of Things (IoT) technology for real-time air quality monitoring was shown by the Enviro-IoT system described in [14]. During long-term monitoring sessions, the system obtained good measurement accuracy for $PM_{2.5}$ and PM_{10}, and NO_2.

1.3 Problem Statement

Accurately categorizing environmental risk levels related to air pollution is still a difficult undertaking, despite the increasing availability of air quality monitoring devices. The intricate nonlinear interactions between environmental factors may be missed by traditional statistical methods, which frequently rely on threshold-based assessments of pollutant concentrations. Numerous interrelated factors, such as pollutant concentrations, weather, seasonal fluctuations, and temporal dynamics, affect the dynamics of air pollution. It is challenging to create trustworthy models that can precisely determine pollution danger levels using traditional analytical techniques because of these interactions.

Many current research rely on single algorithms or small feature sets, which may lower predictive accuracy and model resilience, despite the fact that machine learning approaches have shown great promise for environmental data analysis.

1.4 Research Gap

There are still a number of issues with the current literature, despite the growing use of machine learning approaches in air quality analysis. Rather than categorizing environmental risk levels, many previous research have concentrated on forecasting pollutant concentrations. Furthermore, many research do not fully integrate environmental, meteorological, temporal, and engineered information inside a single analytical framework; instead, they rely on individual machine learning models. Additionally, neither the optimization of ensemble-based classification algorithms for environmental risk assessment nor the evaluation of the contribution of various feature groups by ablation analysis have received much attention.

Consequently, a reliable and comprehensible ensemble-based machine learning system that can precisely categorize air pollution risk levels using actual monitoring data is required.

1.5 Research Motivation

One of the most important environmental issues affecting human health, urban sustainability, and environmental quality globally is still air pollution. Large amounts of environmental data have been produced by the expanding availability of environmental monitoring systems, opening up new possibilities for data-driven environmental analysis.

Strong tools for examining intricate environmental information and spotting trends that conventional analytical methods might miss are offered by machine learning techniques. Specifically, by integrating several machine learning algorithms, ensemble learning techniques have shown enhanced prediction performance.

In addition to improving environmental monitoring systems and decision-making procedures, the development of strong ensemble-based machine learning models that can integrate environmental, meteorological, and temporal variables can greatly improve air pollution risk assessment.

1.6 Research Contributions

The following is a summary of this study's primary contributions:

- The creation of an ensemble-based machine learning system that uses temporal, meteorological, and environmental data to categorize air pollution risk levels.
- Using a Soft Voting ensemble model to integrate several machine learning methods to increase model robustness and classification accuracy.

• Using feature engineering methods to capture intricate connections between weather patterns and environmental contaminants.
• A thorough assessment of feature contributions via ablation analysis.
• Using feature importance analysis to interpret model results, important environmental factors impacting pollution risk levels are identified.
• The suggested model is validated using cross-validation techniques to guarantee its stability and capacity for generalization.

Aim of the Study

The objective of this research is to create a reliable ensemble-based machine learning model that uses environmental, meteorological, and temporal indicators obtained from actual monitoring data to reliably categorize air pollution risk levels.

2.1 Research Objectives

The following goals are pursued by the study in order to accomplish this goal:
O1: Examine and prepare data from actual air quality monitoring.
O2: Create engineered features that illustrate how pollution and meteorological factors interact.
O3: Create and assess various machine learning models for classifying the risk of air pollution.
O4: Construct a Soft Voting ensemble model that integrates several classifiers.
O5: Optimize ensemble weights to improve classification performance.
O6: Conduct ablation studies to evaluate the influence of different feature groups.
O7: Assess model robustness using cross-validation techniques.
O8: Identify the most influential environmental variables affecting pollution risk levels.

2.2 Research Questions

This study addresses the following research questions:
RQ1: Can ensemble-based machine learning models effectively classify air pollution risk levels using environmental, meteorological, and temporal indicators?
RQ2: Does integrating multiple machine learning algorithms through a Soft Voting ensemble improve classification performance compared with individual models?
RQ3: How do different feature groups (environmental pollutants, meteorological variables, temporal indicators, and engineered features) influence classification accuracy?
RQ4: Which environmental variables contribute most significantly to predicting air pollution risk levels?
RQ5: How robust and stable is the proposed ensemble model when evaluated using cross-validation techniques?

1. Methodology

3.1. Study design and methodological framework

We employed an environmental data-driven approach to classify air pollution risk levels in major urban areas based on real-world monitoring observations. To give an interpretable assessment of environmental risk patterns, we designed a methodological framework linking atmospheric pollution indicators with meteorological and temporal conditions. Instead of focusing on the air pollution issue only as a numeric prediction task, we proposed to approach air pollution by exploring it from an environmental risk classification perspective, in order to classify urban

pollution conditions into low, medium and high concern situations for urban environmental quality and the management of urban pollutants.

The complete workflow included five basic steps: dataset acquisition and exploration, data cleaning and pre-processing, building environmental features, risk classification using machine learning methods, and spatial interpretation of predicted risk patterns. That design was chosen to ensure that the classification process captures not just pollutant concentrations, but also the wider environmental context in which air quality conditions evolve (e.g., seasonal variability, urban setting, meteorological influence).

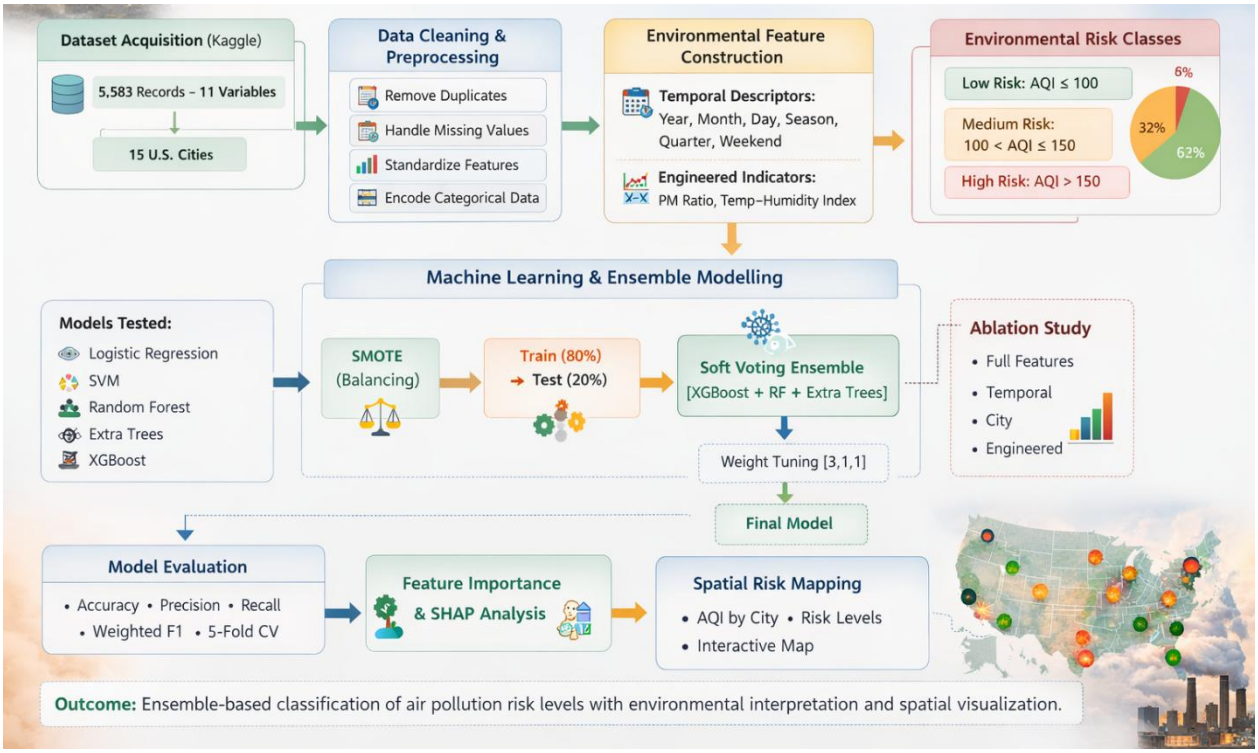

Figure 1: Methodological framework for air pollution risk classification using real-world monitoring data and a soft voting ensemble model.

Figure 1 provides an overview of the methodological workflow employed in this study for classifying air pollution risk levels by major urban areas. The framework starts with the collection of real-time air quality monitoring data and moving through steps of raw-data cleaning, feature-engineering processes, construction of environmental features using machine learning classification models. A soft voting ensemble model combining XGBoost, Random Forest, and Extra Trees was selected as a final model after algorithm evaluation and weight tuning. The model performance was evaluated by accuracy, precision, recall, weighted F1-score and five-fold cross-validation. Moreover, ablation experiments were performed to evaluate the contribution of individual feature groups and interpretability techniques (e.g., feature importance and SHAP analysis) were applied. The predicted pollution risk levels were further spatialized to facilitate environmental interpretation and urban air quality evaluation.

3.2. Dataset description

It was based on real-world air quality data collected from Kaggle, a repository for collection of monitoring data in different cities in USA. The original dataset had 5,583 rows and 11 Features

corresponding to daily air quality observations and meteorological conditions. The final working dataset consisted of 5,387 observations after eliminating duplicate records, invalid dates and rows with missing AQI values [15].

The dataset includes 15 major cities — Austin, Charlotte, Chicago, Columbus, Dallas, Fort Worth, Houston, Jacksonville Los Angeles New York Philadelphia Phoenix San Antonio San Diego san Jose Such urban diversity is an adequate framework for studying spatial differences in air pollution characteristics and behavior in other metropolitan environments.

The original variables included:

- **Date**
- **City**
- **Air Quality Index (AQI)**
- **PM2.5**
- **PM10**
- **Ozone**
- **NO2**
- **CO**
- **SO2**
- **Temperature**
- **Humidity**

These Features jointly characterize the atmospheric environmental status by indicating both direct pollutant loads and related meteorological circumstances.

The selected features are very relevant for the assessment of air pollution from an environmental perspective. Fine particulate matter (PM2.5) and coarse particulate matter (PM10) are some of the most ultimate atmospheric quality indicators due to their direct correlation with visibility degradation, respiratory stress, and urban pollution accumulation. Ozone is a product of photochemical activity in the lower atmosphere and can be exacerbated under specific thermal and solar conditions. Gaseous pollutants such as nitrogen dioxide (NO2), carbon monoxide (CO) and sulfur dioxide (SO2) are directly linked to combustion sources, traffic activity, and industrial emissions. Temperature and humidity were included as they have a strong impact on pollutant dispersion, chemical reactivity, atmospheric stability, and human thermal perception of the environmental conditions.

3.3. Data cleaning and preprocessing

Before developing the models, there was a preprocessing step series applied to make sure that data reliability. First, we filtered duplicates out from the data. Second, the date field was standardized in a datetime format to maintain temporal consistency. Exclusion of records with invalid dates or missing (Air Quality Index) AQI values was due to the fact that AQI was the basis of environmental risk class derivation.

Several environmental features/variables, including PM2. PM2.5, PM10, Ozone, NO2, CO, SO2, Temperature and Humidity. Missing values were instead imputed during preprocessing with median for numerical features and most frequent for categorical features rather than removing large quantities of observations. This method maintained the environmental structure of the dataset and avoided information loss.

Numerical variables were standardized using z-score normalization for training the models. This step was vital as the dataset includes environmental variables of varying scale (ranging from gas concentrations to meteorological quantities), encompassing a wide range. In model fitting, standardization promoted more balanced contributions from all predictors.

Both categorical variables — City and Season — were one-hot encoded, enabling the machine learning models to consider spatial and seasonal context without introducing arbitrary ordinal relationships.

3.4. Environmental feature construction

From the original variables-Features, additional environmental features aimed at better representing atmospheric processes and pollution behavior were derived. This was meant to compensate for the fact that not all interactions can be captured using only raw measurements as variables.

3.4.1. Temporal descriptors

Due to the significant impact of seasonality and short-term temporal cycles on air pollution dynamics, various time-related Features (variables) were derived from the date field:

- **Year**
- **Month**
- **Day**
- **Day of week**
- **Quarter**
- **Weekend indicator**
- **Season** (Winter, Spring, Summer, Autumn)

These descriptors were included to account for variations in anthropogenic activity, dynamic meteorological shifts, seasonal emission characteristics and ambient exposure conditions across urban environments. For instance, seasonal variation may influence atmospheric stagnation, temperature inversions, photochemical ozone formation and pollutant washout processes. Effects attributable to weekends may also be influenced by differences in traffic intensity and urban activity.

3.4.2. Engineered environmental indicators

Two more variables were created to aid the environmental interpretation of the dataset:

1. **PM ratio (PM2.5 / PM10)**
   This ratio indicates the potentially dominant role of fine particles on total suspended particulate matter. Higher values may indicate greater influence from combustion related fine aerosols that have been shown to have higher environmental and health considerations than the larger coarse particles.
2. **Temperature–humidity index**
   This was calculated as a product of temperature and humidity to also reflect combined atmospheric conditions that affect pollutant persistence, discomfort, and environmental stress. It was added as a simple interaction term to capture the combined effect of thermal and moisture conditions on atmospheric behavior.

The need to introduce engineered features stems from the realization that the severity of air pollution is rarely managed by a single variable; rather, it results from interactions between pollutant loads, climatic conditions and urban setting.

3.5. Environmental risk classification scheme

Rather than forecasting AQI as a continuous measure, this study converted AQI into a three-tier environmental risk classification system. This method was chosen to offer clearer output for environmental monitoring and decision-making support.

The following classes were defined:

- **Low Risk:** AQI ≤ 100
- **Medium Risk:** $100 < AQI \leq 150$
- **High Risk:** AQI > 150

There were two primary reasons why this categorization was chosen. First, it is indicative of real differences in air quality conditions on the ground, between relatively reasonable urban air and levels that are likely to warrant increased environmental and health concern. Second, the original AQI distribution was presented with a high degree of imbalance when categorized into finer levels; thus, a three-class structure offered a more stable and analytically reliable classification framework.

After categorization, the final class distribution consisted of:

- **Medium Risk:** 3,404 records
- **Low Risk:** 1,731 records
- **High Risk:** 252 records

This distribution shows that the conditions of medium risk were most common in the dataset, while those of high risk were less frequent but ecologically significant.

3.6. Exploratory environmental analysis

Before to model building, an exploratory analysis was performed to comprehend the environmental framework of the dataset. Several visual and statistical analyses were used, including:

- distribution of the risk classes,
- AQI frequency distribution,
- correlation analysis among numerical variables,
- boxplot comparison of major pollutants across risk levels,
- average AQI by city,
- average AQI by season.

This stage also yielded significant environmental insights. Response variable AQI distribution indicated a majority of the observations were grouped in moderate pollution conditions with decreasing number of extreme high-risk events. The correlation analysis further gave a positive association of AQI with PM10, Ozone, Temperature and to some extent with PM2. 5, whereas negative correlations were found with Humidity. These patterns are global realizations of an environmentally meaningful anomaly in which a warmer, drier atmosphere favors pollutant accumulation and ozone formation under some conditions present in urban environments.

The boxplot representation also confirmed the increasing trend of PM10 and Ozone with risk levels, which indicates their significance in distinguishing pollution severity classes. The city-level

and seasonal analyses further reflected the variability in average AQI values due to both urban setting and seasonal context that should be considered in the classification framework.

3.7. Selection of predictor variables

AQI was removed from the predictor set to prevent information leakage as our predicted variable came directly from AQI. The final predictor set thus included the 18 variables detailed in the following list:

- City
- PM2.5
- PM10
- Ozone
- NO2
- CO
- SO2
- Temperature
- Humidity
- Year
- Month
- Day
- DayOfWeek
- Quarter
- IsWeekend
- Season
- PM_Ratio
- Temp_Humidity_Index

The objective of this feature structure was to capture pollution intensity, meteorological controls, temporal variation, and urban context within a unified framework for environmental classification.

3.8. Machine learning models for environmental risk classification

Five machine learning models were initially tested to classify air pollution risk levels:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest
- Extra Trees
- Extreme Gradient Boosting (XGBoost)

We chose these algorithms to capture both traditional statistical-based classification approaches as well as more flexible nonlinear learning methods. This comparison between several models was carried out in order to recognize the best model for the multidimensional and partial imbalance of environmental dataset.

The training pipeline implemented the Synthetic Minority Over-sampling Technique (SMOTE) since the High_Risk class had fewer observations than other classes. SMOTE was only applied to the training partition in order to mitigate class imbalance without contaminating the test set. It enhanced the models' capability to use its learning of environmentally critical high-risk cases. The

dataset was split into 80% training data and 20% testing data, employing stratified sampling to maintain class proportions in both subsets.

3.9. Ensemble modelling and model refinement

XGBoost performed the best out of all tested classifiers after baseline models were established. Two ensemble strategies were explored to further boost predictive robustness:

1. **Soft Voting Ensemble**
2. **Stacking Ensemble**

The Soft Voting model integrated the three strongest base learners:

- XGBoost
- Random Forest
- Extra Trees

In this approach, the final prediction of class assignment was performed from aggregate probabilities of class produced by component models. We also performed stacking using the same base learners, of which logistic regression was the meta-learner. The comparison indicated that Soft Voting is the best option among both individual models and Stacking, thus deeming it as the final ensemble strategy.

To improve the ensemble further a lightweight weight-tuning experiment was performed. Testing of different weight combinations was done to find the most effective contribution of each base learner. The optimal configuration was discovered to be the weight setting [3, 1, 1], hinting that XGBoost has heavily contributed to making the final decision whereas Random Forest and Extra Trees provided valuable complementary information.

3.10. Ablation analysis

An ablation study was conducted to assess the impact of several environmental feature groups across four scenarios:

1. **Full feature set**
2. **Without temporal descriptors**
3. **Without city information**
4. **Without engineered environmental features**

In this study, we used the combined feature structure to determine whether the observed predictive performance was due to a broad set of variables, rather than a narrow subset.

Results demonstrated that the highest classification accuracy occurred when using the full feature configuration. The quality of prediction dropped when temporal features were removed, suggesting that both seasonal and short-term temporal patterns meaningfully contribute to air pollution behavior. When city information was excluded, performance decreased, thus confirming the influence of spatial heterogeneity on urban air quality. The importance of engineered variables also weakened the model, demonstrating that environmental interaction features (e.g., PM ratio and the temperature–humidity index) introduce valuable discriminatory information from the data.

Environmentally, this ablation analysis shows that air pollution severity is best seen as the result of interacting pollutant loads, meteorological controls and urban-spatial context.

3.11. Model evaluation criteria

Model performance was evaluated using the following classification metrics:

- **Accuracy**

- **Precision**
- **Recall**
- **Weighted F1-score**

In particular, the weighted F1-score was emphasized due to the imbalanced nature of our test dataset across risk levels; using the weighted F1-score provides a more balanced view of predictive quality overall across classes. Moreover, classification reports and confusion matrices were utilized to evaluate the discriminatory power of each model for low-, medium-, and high-risk air pollution conditions.

Hold-out validation: Stratified 5-fold cross-validation was performed on the final selected model to evaluate model robustness. This step gave an estimation of how stable generalization is over partitions of the dataset.

3.12. Environmental interpretation of model outputs

Interpretability constituted a crucial component of the technique, as the study aimed not only to identify pollution risk but also to elucidate the environmental factors that most significantly impact that categorization.

For this purpose, two complementary interpretation techniques were used:

1. **Feature importance analysis**
2. **SHAP (Shapley Additive Explanations)**

Feature importance adopted methods to identify which variables contributed the most strongly to the model output, while SHAP analysis offered a deeper view of how higher or lower values of individual variables predicted the risk classes.

According to interpretability analysis, PM2. PM2.5, PM10 and Ozone (O3) variables were the most impactful ones for our predictions, along with a handful of city-related variables. This is environmentally consistent, because fine particulate pollution is widely understood to be one of the most damaging aspects of degraded urban air quality. Additionally, the influence of city variables suggests that environmental setting at the local scale and structure of emissions from urban areas remain pertinent factors determining risk differentials for pollution.

3.13. Spatial representation of predicted environmental risk

To expand the analysis beyond numerical classification, we generated a spatial map of predicted environmental risk per city. The final ensemble model was applied on the full data and then the predicted class for each observation was aggregated by city to produce:

- average AQI,
- average pollutant concentrations,
- dominant predicted risk class,
- ratio of predicted High_Risk cases.

Approximate geographic coordinates were assigned to each city, and the results were visualized using two complementary map formats:

1. **Interactive city risk map**
2. **Static spatial distribution figure**

The interactive map layered city markers, color-coded risk levels and popup overviews of environmental conditions. Simultaneously, a static scatter-based spatial figure was produced for

direct inclusion in the manuscript. A heatmap layer was applied as well to highlight regions with high predicted ratios of risk.

This spatial component enhances the study's ecological significance by converting model outputs into data that can aid geographic interpretation, urban comparisons and environmental planning.

3.14. Methodological rationale

Following the adoption of a methodology that would enable treating air pollution as an integrated environmental phenomenon rather than as an isolated numerical signal. The inclusion of pollutant concentrations, meteorological conditions, temporal structure, derived interaction variables and spatial urban context illustrates the multidimensional nature of atmospheric pollution. In this context, they considered machine learning as an analytical tool for environmental classification and decision support rather than a purpose in itself.

Thus, the final methodological framework offered is a composite approach of environmental interpretation, data led analysis and urban risk mapping suitability for application in areas relating to environmental monitoring, pollution screening and sustainable management realms.

4. Results and Discussion

4.1 Distribution of environmental risk levels

From the generated environmental risk classification, it can be seen that Medium_Risk factors are reporting more than Low_Risk while High_Risk cases are very few in the dataset. In the end, the distribution resulted in 3,404 Medium_Risk records, 1,731 Low_Risk records and 252 High_Risk records, suggesting that most urban observations correspond to moderate pollution stress - little extreme clean or severely polluted conditions.

The fact that Medium_Risk conditions dominate from an environmental standpoint means most urban environments suffer extended periods of moderate atmospheric stress. And while these did not always approach hazardous limits, long-term exposure to moderate pollution can compromise environmental quality and ecosystem functioning, as well as human health. By contrast, High_Risk events are sparser, more omnivorous in terms of extreme weather episodes but especially relevant for monitoring environmental concerns and the use of early-warning systems.

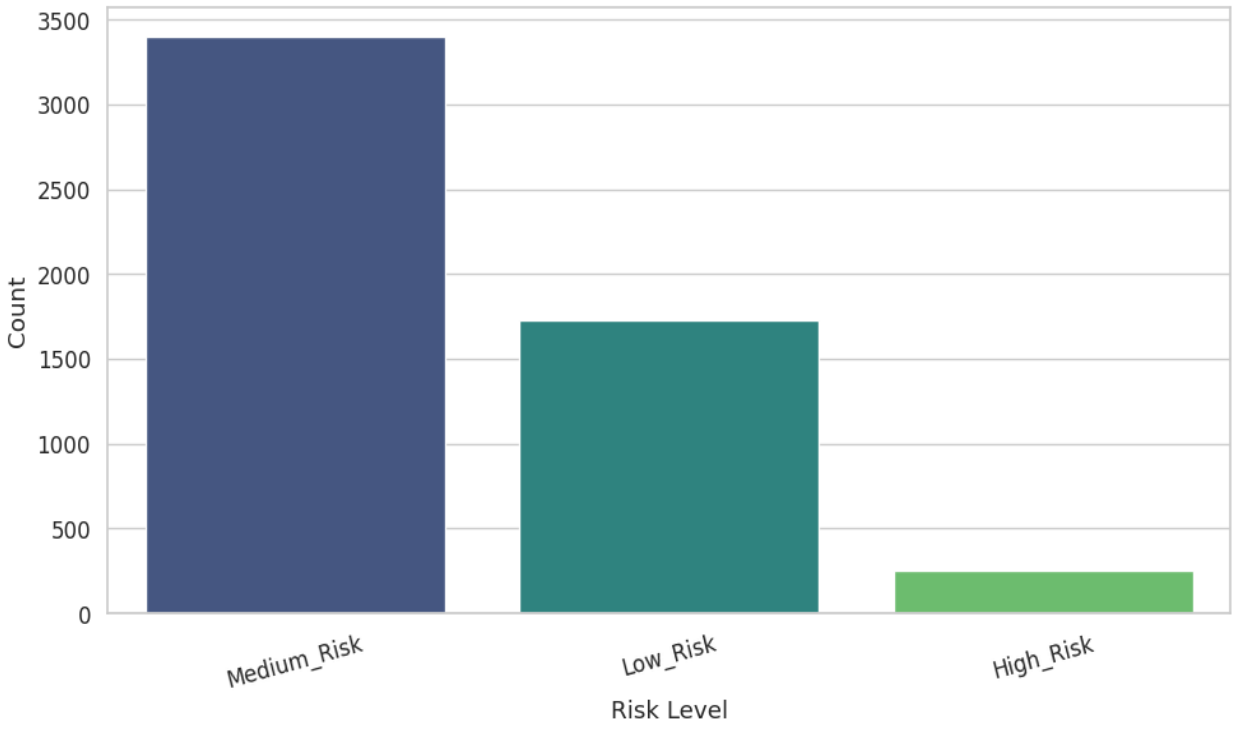

Figure 2: Distribution of Environmental Risk Levels (3-Class Target).

The class distribution used for environmental risk classification is presented in (Figure 1) after transforming AQI into three categories. Note that the preponderance of the Medium_Risk class indicates a commonality of moderate atmospheric stress across monitored cities and a need for imbalance-aware modelling approaches.

4.2 AQI distribution and environmental variability

The distribution of AQI is heavily weighted toward moderate pollution range (e.g. 80-140) and has a distinct right-skew tail indicating rarer high-pollution events. This trend parallels the episodic characteristics of acute air pollution incidents in urban environments.

Moderate pollution levels may well be attributed to normal emissions and typical atmospheric conditions, while high AQI values can result from complexes of negative meteorological stability with increased anthropogenic emissions or photochemical processes. This approach yields a database that reflects both background urban pollution levels in an area and episodic high-pollution episodes, information essential for modeling environmental risk.

The histogram with density curve shown in Figure 3 portrays the statistical distribution of Air Quality Index (AQI) values captured by the dataset. This distribution indicates a large majority of observations fall within the mid-pollution band, in this case possibly from about 80-140 AQI units, showing that most states end up with intermediate air quality levels. The tail of this distribution is right-skewed, representing extreme pollution occurrences that are rare but ecologically important. These extremes may occur under conditions of poor atmospheric stability, increased anthropogenic emissions or enhanced photochemical activity in urban areas. The distribution pattern proves that the dataset well represents both general background pollution conditions and episodic high-risk pollution episodes, which is important for developing reliable machine learning models for environmental risk classification.

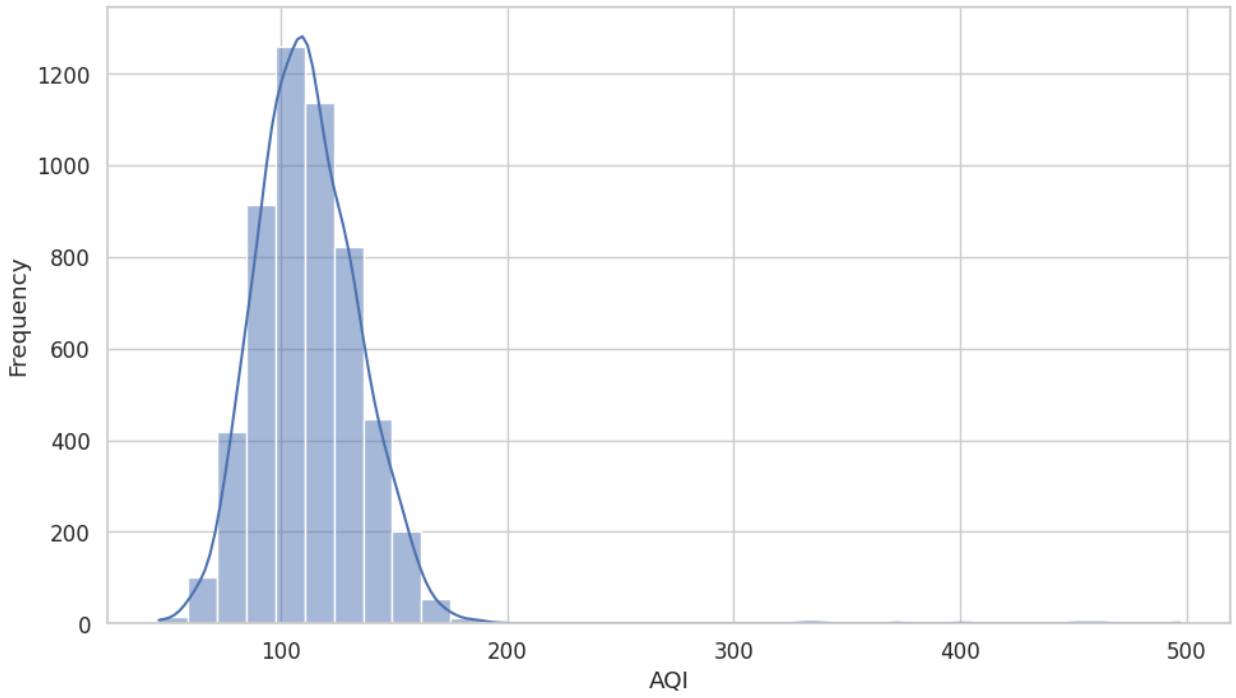

Figure 3: Distribution of Air Quality Index (AQI) values across the dataset

4.3 Relationships among pollutants and meteorological variables

Relationship among pollutants, meteorological conditions, and AQI were found to be significant using correlation analysis. Correlations of AQI were highest with PM10 (r = 0.58) followed by Ozone (r = 0.47), Temperature (r = 0.41') and PM2.5 (r = 0.32). Conversely, AQI was negatively associated with Humidity (r = −0.37).

These relationships are consistent with known atmospheric behavior. Increased particulate levels lead to greater pollution loading, and ozone formation is also usually enhanced under high temperature via photochemistry. This negativity with humidity might indicate improved pollutant dispersion and atmospheric chemistry under humid conditions. Further correlational analysis between meteorological variables revealed a significant relationship between Temperature with Ozone (r = 0.71) and Humidity (r = −0.84), suggesting that thermal and humidity conditions have pivotal links to atmospheric pollution dynamics processes.

The correlation heatmap shown in Figure 4 captures relationships between the Air Quality Index (AQI), major atmospheric pollutants, meteorological variables, and engineered environmental indicators applied in this research paper. The heatmap shows some meaningful patterns that reflect the environmental dynamics of urban air pollution. AQI has the highest positive correlations with PM10 (r = 0.58), Ozone (r = 0.47), Temperature (r = 0.41), PM2.5 (r = 0.32), suggesting that high concentrations of particulates and photochemistry are determinant to higher pollution levels. AQI is inversely correlated with Humidity (r = −0.37), indicating that higher moisture conditions may improve atmospheric dispersion and change the chemical reaction to reduce pollutant accumulation. The heatmap clearly shows high correlations between different meteorological

variables, in particular the positive correlation of Temperature to Ozone (r = 0.71) and the highly negative temperature to Humidity correlation (r = − 0.84). These results highlight the simultaneous impact of both pollutant emissions and meteorological conditions in determining urban air quality behavior and reinforce a need to incorporate both environmental and climatic features in the machine learning classification framework.

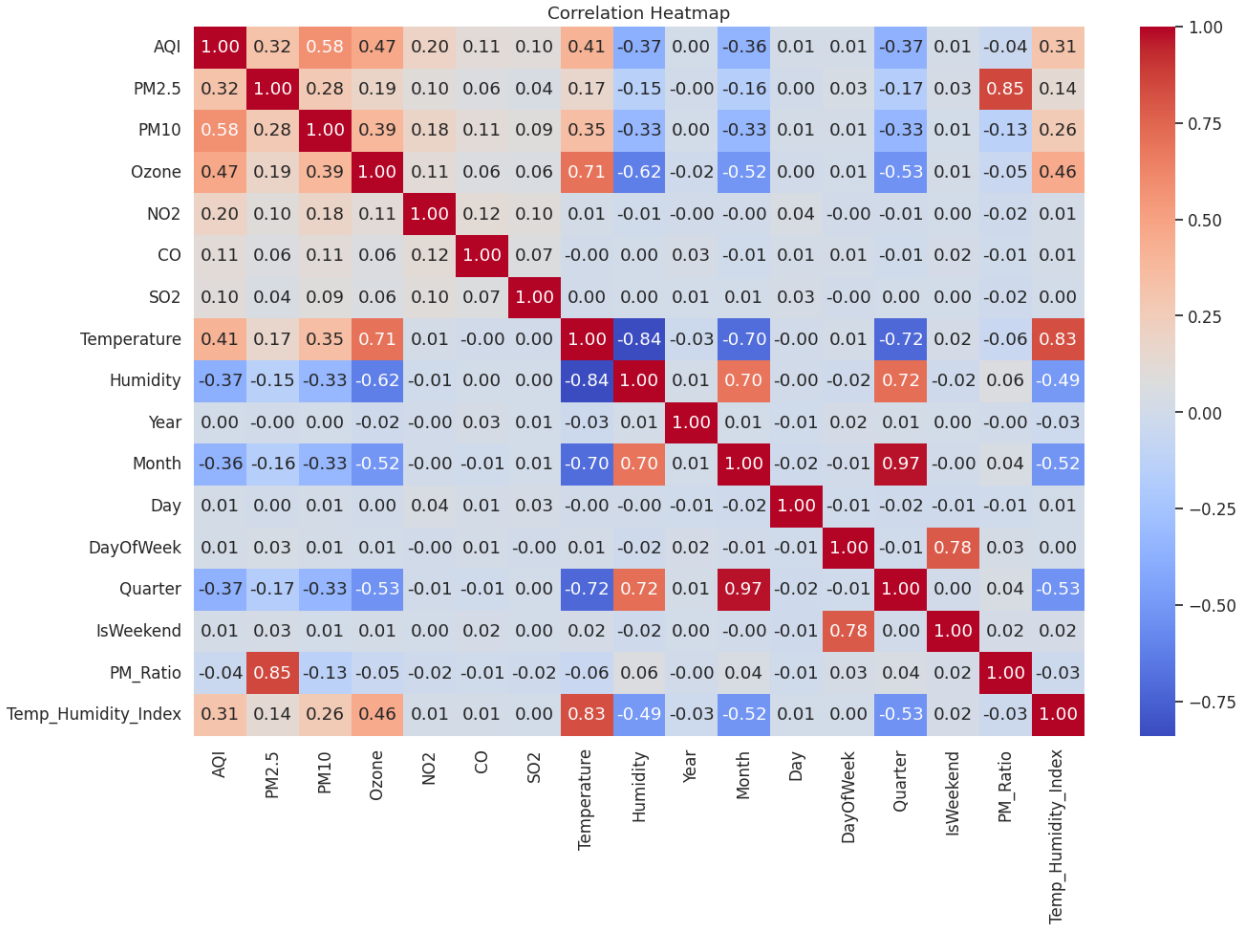

Figure 4: Correlation Heatmap of AQI, Pollutants, Meteorological Variables, and Derived Features

4.4 Pollutant behavior across environmental risk levels

Boxplot showed that the concentration of the pollutants generally increased with increasing level of environmental risk although the clearest gradients were for PM10 and Ozone. The central values, as well as the upper distributions of these two pollutants were particularly pronounced in the High_Risk class compared to the Low_Risk and Medium_Risk classes. In contrast, other air pollutants like PM2. 5, NO2, CO and SO2 were less class separable (although several extreme values were detected).

From an environmental point of view, PM10 and Ozone seem to reflect two principal and complementary pathways of urban atmospheric degradation. PM10 represents direct particle emission linked to traffic activity, construction activities, resuspension of road dust and other

sources from urban surfaces. Ozone, on the other hand represents secondary photochemical pollution produced by gaseous reactions in the troposphere under permissive thermal conditions. The co-elevation of these pollutants under High_Risk conditions implies that extreme air pollution event days in the studied cities is not dominated by a single governing mechanism, but rather progressive through interplay between particulate build-up and photochemical augmentation.

These results are in agreement with the general environmental interpretation of the data set and further reinforce the application of both direct indicator (i.e., pollutants) and sensitive meteorological variables for classification within this framework. This stronger class separation for PM10 and Ozone also accounts for why these variables significantly contributed to separating environmental risk levels in the following machine learning analysis.

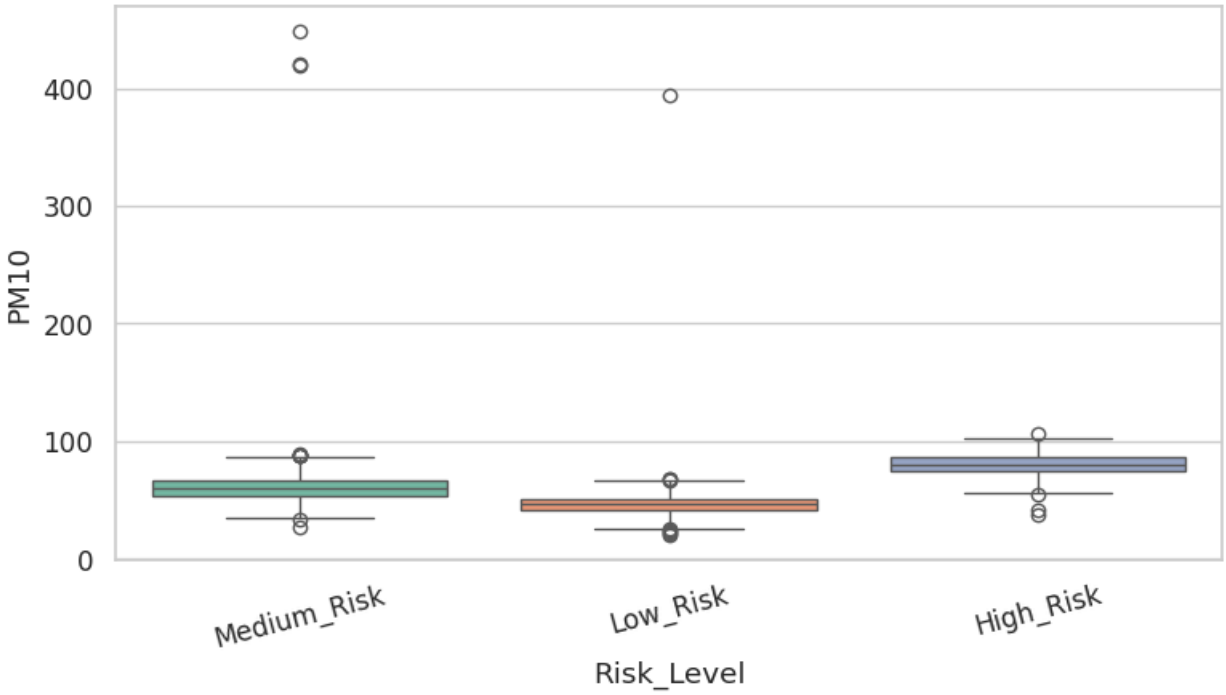

Figure 5: PM10 concentration by environmental risk level.

The distributions of PM10 concentrations by environmental risk class are illustrated in Figure 5. The High_Risk category reveals a higher central values and a wider range of concentration than the Low_Risk class, showing that with pollution severity increases coarse particulate loading This pattern indicates that non-fulminating afferent particles, including those from urban traffic, dust resuspension and surface emissions are significant in separating high risk air pollution states to those with lower risk atmospheric depositions.

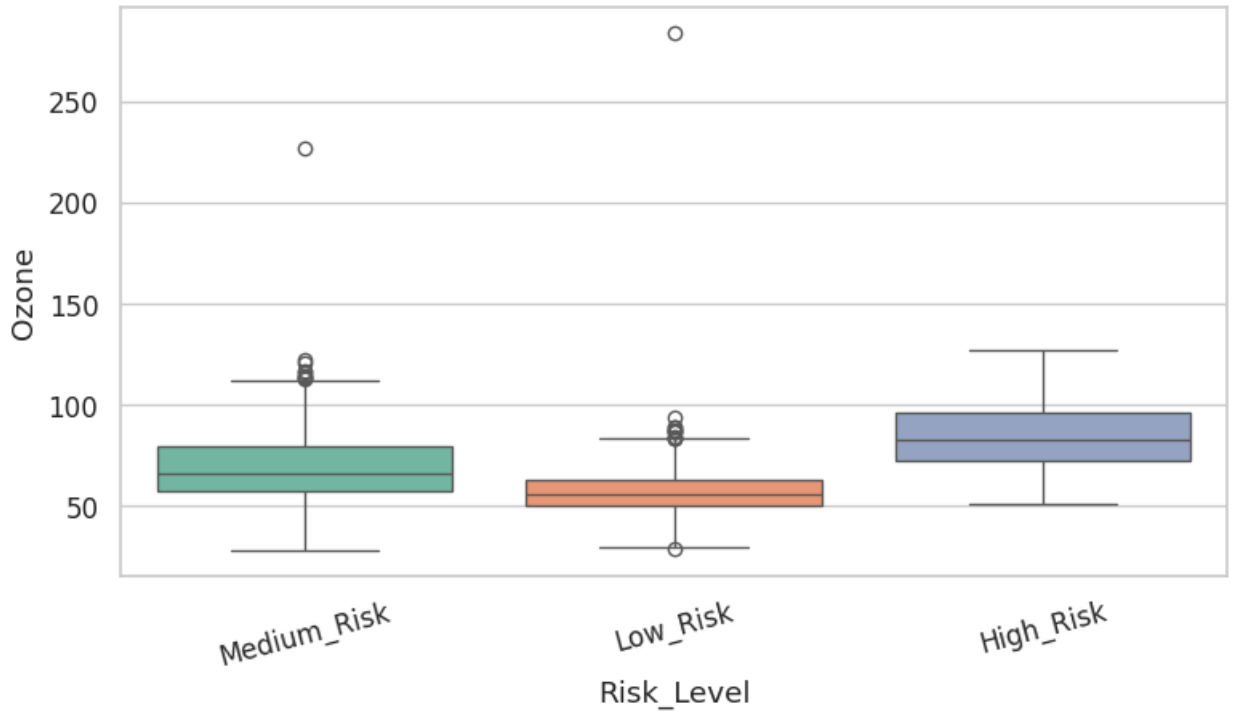

Figure 6: Ozone concentration by environmental risk level.

Figure 6 that distribution of ozone concentrations in the three environmental risk classes This means that ozone concentration is generally lowest for Low_Risk and highest for the High_Risk class on the environmental risk scale, with a clear shift of the central distribution towards higher concentrations as we move up in environmental terms. This pattern underscores the importance of photochemical pollution in extreme air quality events and demonstrates a contribution from secondary atmospheric reactions, particularly under warmer, more conducive meteorological conditions, to heightened urban pollution risk.

4.5 Spatial and seasonal patterns of air pollution

Considerable spatial heterogeneity of average AQI was found among the 15 cities. Los Angeles had the highest mean AQI, followed by Houston, Phoenix, Dallas and Chicago to round out the top five cities with the dirtiest air on average; San Diego and San Jose ranked as having the cleanest overall pollution levels. This spatial variability is a result of variations in urban emissions, local climatic conditions, and atmospheric circulation patterns.

The seasonal analysis showed that Spring had the highest average AQI, followed by Winter and Summer, while Autumn was found to be the least polluted season. Seasonal variation could result from differences in atmospheric stability, mixing conditions, photochemical activity and emissions. These spatial and temporal patterns reinforce the need to incorporate city-level and seasonal variables into the classification framework.

Average AQI of the cities studied differs, as demonstrated in Figure 7. The figure shows that several large metropolitan areas, especially Los Angeles, Houston and Phoenix, have inflation rates of pollution notably excessive than other cities in the dataset. This difference reflects the fact that urban environmental conditions vary widely between cities due to different sources of emissions, urban structure and regional atmospheric conditions.

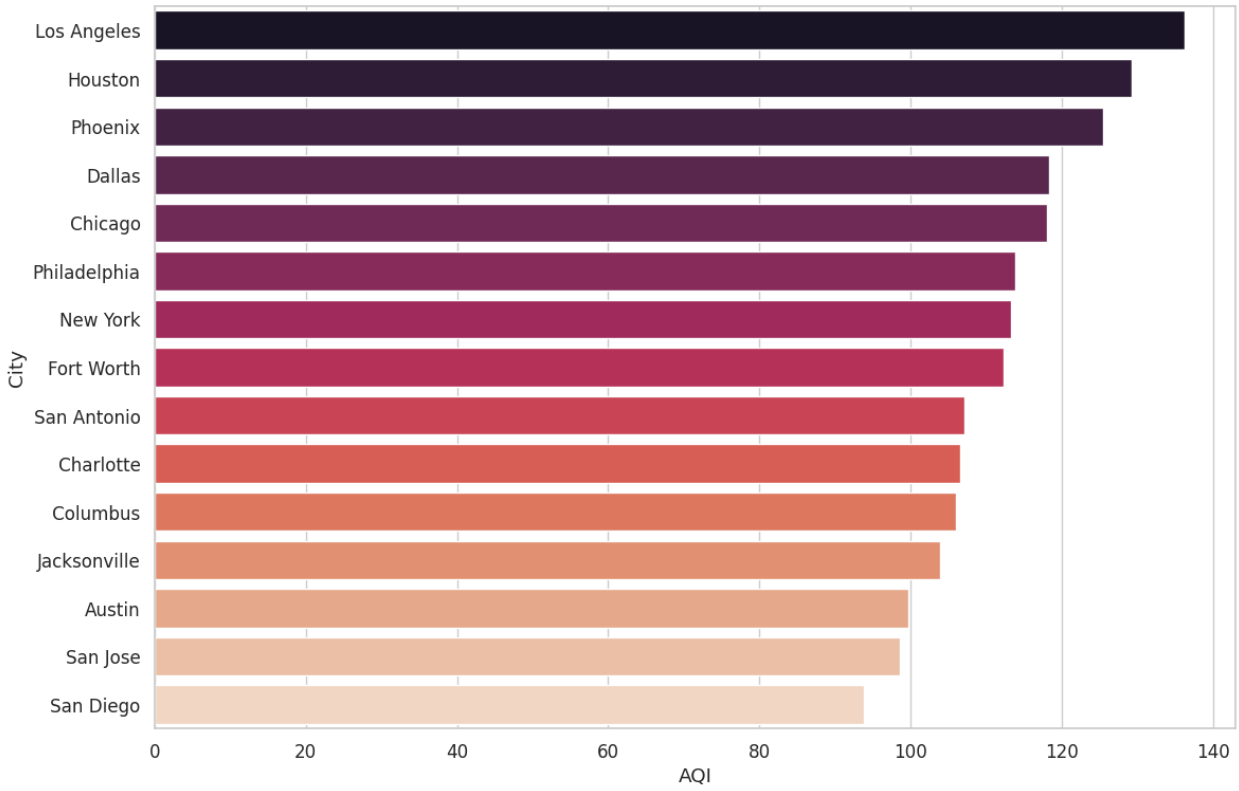

Figure 7: Average AQI by City.

The seasonal distributions of average AQI values are depicted in Figure 8. It can be observed from the above figure that AQI levels tend to be higher during Spring and Winter, and lowest in Autumn. The seasonal differences indicate how the atmospheric environment and seasonal emissions influence pollutant accumulation and dispersion in cities.

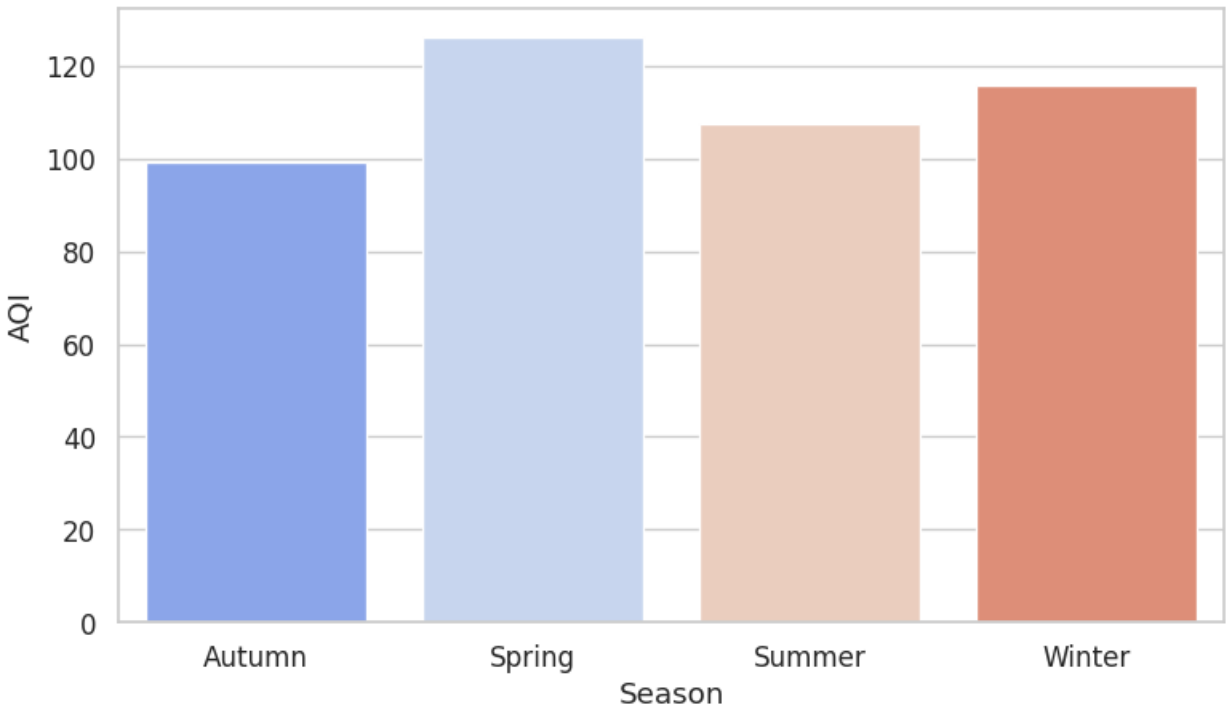

Figure 8: Average AQI by Season.

4.6 Baseline model performance

For baseline models, tree-based algorithms outperformed linear and kernel-based approaches. Of the five classifiers, XGBoost had the best performance overall with a weighted F1-score of 0.8243, followed by Random Forest (0.8159) and Extra Trees (0.8053). SVM and Logistic Regression performed relatively poorly. These results suggest that air pollution risk classification involves complex nonlinear relationships among particulate matter, gaseous pollutants, meteorological conditions, and temporal variables, which are more effectively captured by tree-based ensemble models

Table 1: Performance of baseline machine learning models for environmental risk classification.

Model	**Accuracy**	**Precision**	**Recall**	**Weighted F1-score**
XGBoost	0.8237	0.8251	0.8237	0.8243
Random Forest	0.8126	0.8225	0.8126	0.8159
Extra Trees	0.8033	0.8087	0.8033	0.8053
SVM	0.7922	0.8123	0.7922	0.7980
Logistic Regression	0.7291	0.7823	0.7291	0.7426

As shown in Table 1, tree-based learning methods, especially XGBoost outperform other model types for environmental risk classification problems. This outcome illustrates the nonlinear nature,

interplays that exist between particulate matter, gaseous pollutants, meteorological variables and temporal influences.

4.7 Ensemble modelling and final model selection

To enhance the predictive robustness, Soft Voting and Stacking ensemble methods were employed by aggregating three best classifiers (XGBoost, Random Forest, Extra Trees) together. The Soft Voting model yielded the best test performance overall, with a weighted F1-score of 0.8270, and outperformed all individuals as well as the Stacking models by slight margin. While the gain over XGBoost alone was modest, the ensemble provided a more stable performance in different pollution conditions, which suggests that complementary models contribute to improving predictive reliability when combined.

Table 2: Comparison between baseline and ensemble models.

Model	**Accuracy**	**Precision**	**Recall**	**Weighted F1-score**
Soft Voting	0.8256	0.8294	0.8256	0.8270
XGBoost	0.8237	0.8251	0.8237	0.8243
Stacking	0.8247	0.8216	0.8247	0.8224
Random Forest	0.8126	0.8225	0.8126	0.8159
Extra Trees	0.8033	0.8087	0.8033	0.8053
SVM	0.7922	0.8123	0.7922	0.7980
Logistic Regression	0.7291	0.7823	0.7291	0.7426

The Soft Voting ensemble performed the best overall of all tested models, as confirmed in Table 2. The increase over XGBoost was small but consistent, suggesting that the improvement in performance is one of robustness, rather than accidental gain.

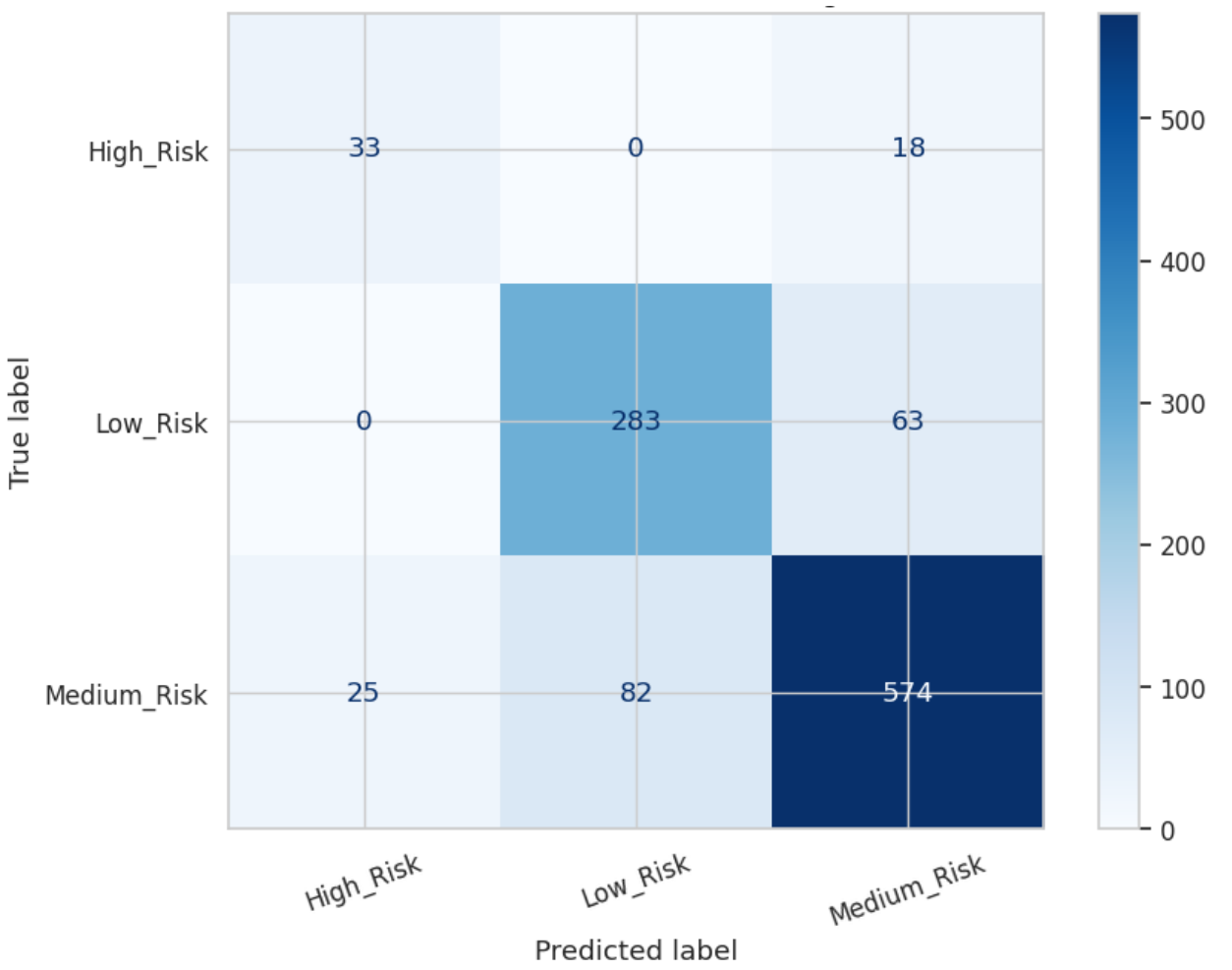

Figure 9: Confusion Matrix for the Soft Voting Ensemble Model.

Figure 9 shows the confusion matrix for the classification performance of the final Soft Voting ensemble model over the three levels of environmental risk. The matrix indicates that the model has correctly predicted 574 of the observations to be Medium_Risk classes. In a related case, 283 Low_Risk observations were accurately classified in the same manner, indicating that strong discrimination can be achieved between conditions of reduced pollution and all other categories.

The model gives 33 cases correctly predicted as High_Risk and relatively few false positive, misclassifying them as Medium_Risk. Note however, this pattern is consistent with the fact that there are fewer high-risk observations so naturally this is a challenging class to classify perfectly. Hence, the confusion matrix demonstrates that through the implementation of this ensemble model over others performed, there is a good representation across all pollution risk types, while remaining predictive performance, for most common environmental conditions.

4.8 Tuned ensemble performance and ablation analysis

Tuning the Soft Voting ensemble's weights, it was found that [3, 1, 1] resulted in that best-performing model with the highest contribution to XGBoost but support from Random Forest and Extra Trees. Finally, the trained model under a tuned set of hyperparameters obtained a weighted F1-score of 0.8283, which indeed reflects its best performance among all configurations that were applied.

An ablation study showed that removing feature groups consistently worsened the model performance. Excluding temporal, spatial or engineered features resulted in lower classification accuracy, and the full feature set provided the highest performance. This aligns with the conclusion that risk from urban air pollution is most appropriately viewed as a joint product of pollutant levels, meteorological conditions, seasonal patterns, and city-specific context.

Table 3: Ablation study results for the tuned Soft Voting ensemble.

Scenario	Best Weight Setting	Accuracy	Precision	Recall	Weighted F1-score
Full_Features	[3, 1, 1]	0.8275	0.8298	0.8275	0.8283
Without_Temporal	[3, 1, 1]	0.8200	0.8318	0.8200	0.8236
Without_City	[3, 1, 1]	0.8182	0.8250	0.8182	0.8205
Without_Engineered	[5, 2, 1]	0.8173	0.8208	0.8173	0.8184

In Table 3 shows that no reduced variant outperformed the full feature configuration. Omitting temporal descriptors, city information or engineered variables reduced the weighted F1-score indicating that all three classes of features contained important environmental information for the classification framework.

4.9 Model robustness and generalization

Evaluation of model robustness employed 5-fold stratified cross-validation yielding a mean weighted F1-score of 0.8237 ± 0.0132. The comparatively small standard deviation suggests consistent predictive performance across various data partitions.

Stabilization indicates that the proposed classification framework is capturing robust relationships between pollutants, meteorological conditions and urban context instead of over fitting to a particular data split.

Table 4 confirms that the selected model generalizes well across multiple data partitions. The consistency of the weighted F1-score across folds indicates that the ensemble framework is stable enough for practical environmental screening and comparative urban assessment.

Table 4: Cross-validation results for the final Soft Voting model.

Fold	**Weighted F1-score**
Fold 1	0.8169
Fold 2	0.8323
Fold 3	0.8043
Fold 4	0.8429
Fold 5	0.8222
Mean ± SD	0.8237 ± 0.0132

4.10 Environmental interpretation of influential predictors

Interpretability analysis identified PM2. 5 by the feature importance, which is much higher than other variables. Another important features ranking included PM10, Ozone and a few city indicators. This result is congruent with the environment, since fine particulate matter is among the most dangerous of atmospheric pollutants because it has easy access to the respiratory system, where among other chronic diseases it can cause cancer and can remain relatively fixed in the atmosphere. The relevance of city indicators further validates that pollution severity is highly dependent on local environmental factors and emission distributions.

Table 5: Top influential predictors identified by XGBoost.

Rank	**Feature**	**Importance**
1	PM2.5	0.3127
2	PM10	0.0518
3	City_Los Angeles	0.0371
4	City_Philadelphia	0.0345
5	City_Phoenix	0.0315
6	City_Houston	0.0290
7	City_Fort Worth	0.0281
8	Ozone	0.0275
9	City_Dallas	0.0260
10	City_Chicago	0.0256

Table 5 shows that the top predictor of environmental risk classification was fine particulate matter, followed by larger particulate loading, ozone and indicators at the city level. This is ecologically sensible and suggests that the final model was responsive to physically meaningful atmospheric variables.

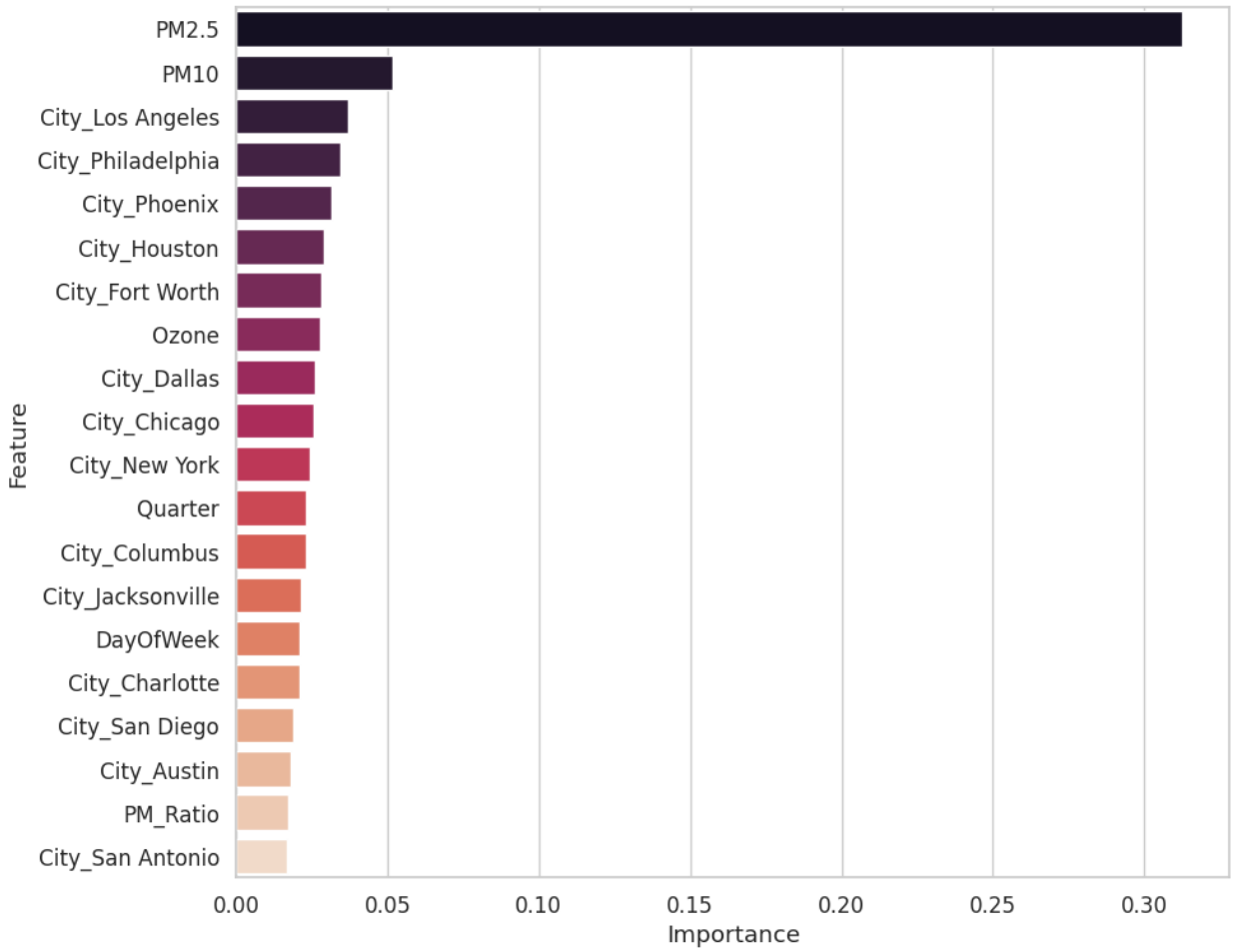

Figure 10: Ranking of the most influential predictors in the XGBoost model.

4.11 Spatial interpretation of predicted environmental risk

There were distinct urban differences in predicted pollution risk with the spatial analysis. This resulted in the largest number of predicted High_Risk observations in Los Angeles (0.292), Houston (0.178), and Phoenix (0.111) respectively. Conversely, a number of cities like Austin and San Diego had virtually no high-risk proportions. The results show that the classification model learned meaningful spatial structures of pollution, which emphasized urbanized hotspots of high environmental risk.

Table 6: City-level summary of predicted environmental risk.

City	Avg_AQI	Dominant Predicted Risk	High_Risk_Ratio
Los Angeles	136.4	Medium_Risk	0.292
Houston	129.5	Medium_Risk	0.178
Phoenix	126.0	Medium_Risk	0.111
Dallas	118.3	Medium_Risk	0.046

Chicago	118.0	Medium_Risk	0.032
New York	113.4	Medium_Risk	0.020
Fort Worth	112.4	Medium_Risk	0.019
Philadelphia	113.9	Medium_Risk	0.008
San Antonio	107.2	Medium_Risk	0.008
Jacksonville	101.5	Medium_Risk	0.008
Columbus	106.0	Medium_Risk	0.005
Charlotte	106.5	Medium_Risk	0.003
San Jose	98.8	Low_Risk	0.003
Austin	99.7	Low_Risk	0.000
San Diego	93.8	Low_Risk	0.000

Table 6 indicates that the highest predicted high-risk burden was clustered in Los Angeles, Houston, and Phoenix. This pattern of spatial differentiation confirms that the classification model captured and preserved meaningful differences among urban units and is, therefore, applicable to environmental screening comparative risk assessment across cities.

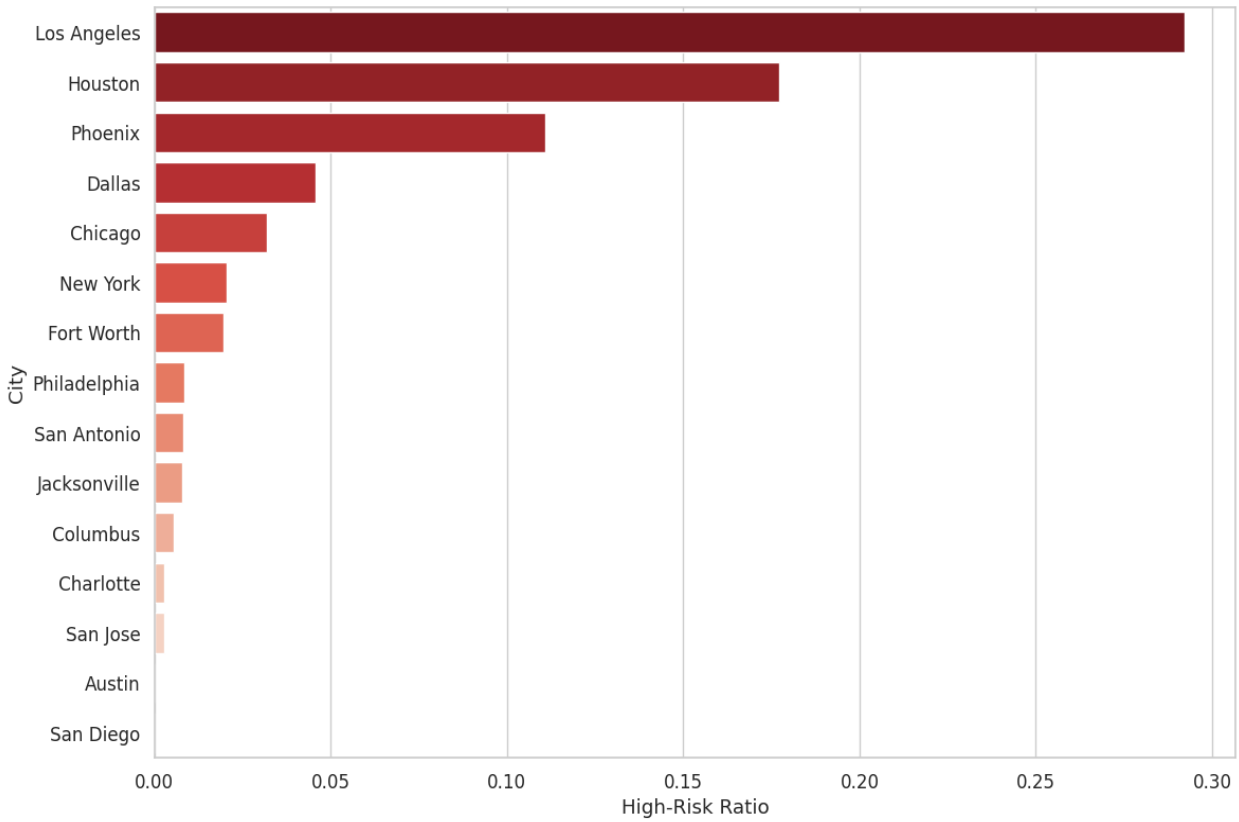

Figure 11: Predicted High-Risk pollution ratio by city.

Figure 11 depicts the estimated rate of High_Risk pollution cases at each city level and indicates significant spatial heterogeneity in urban environmental stress. Los Angeles had the highest high-risk ratio, followed by Houston and Phoenix. Austin and San Diego, on the other hand, demonstrated minimal high-risk proportions as a function of predicted environmental stress. The resulting figure thus offers a straightforward and interpretable comparison of the severity of city-level pollution as well as validating that the proposed framework was indeed capable of capturing meaningful spatial differences in predicted air pollution risk.

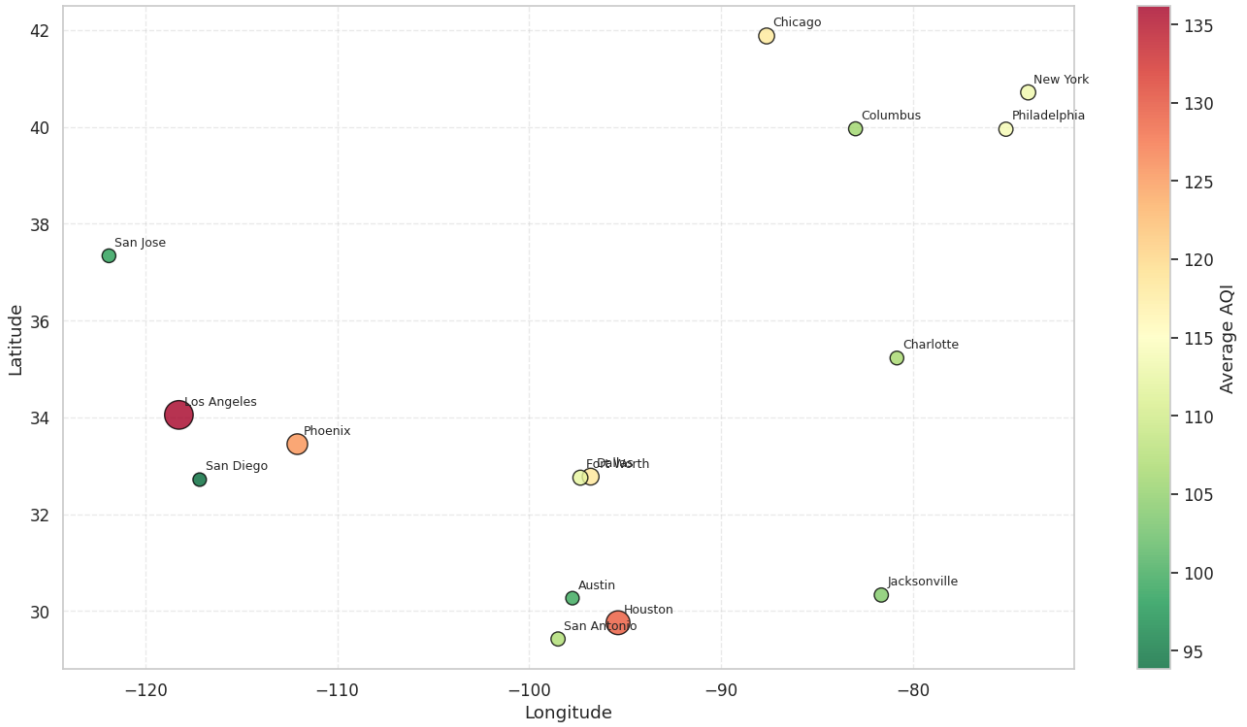

Figure 12: Spatial distribution of predicted environmental risk by city based on average AQI levels in the studied U.S. urban areas.

The spatial distribution of predicted environmental risk for U.S. cities based on their average AQI levels is illustrated in Figure 12. Geographically, it shows clean cuts and Los Angeles, Houston and Phoenix have higher pollution burden as compared to other few cities. On the opposite side of the spectrum, cities find themselves like San Diego and San Jose in lower pollution range. This spatial pattern indicates that the predicted environmental risk is not homogeneously diluted, but meaningful urban heterogeneity from local emission intensity correlated with atmospheric conditions and the meteorological state of each city. Hence, this figure would reinforce the geographical interpretation of the model outputs and support the potentiality of using this framework for spatial environmental screening and urban pollution evaluation.

4.12 Environmental implications

The findings highlight the potential for machine learning to be a useful analytical tool for environmental risk assessment if conducted on physically relevant atmospheric variables. The dominance of particulate matter and ozone, among the most influential predictors, confirms that the model outputs are environmentally valid.

By integrating pollution indicators, meteorological variables, and spatial context, the proposed framework provides a useful approach for urban environmental monitoring, pollution hotspot identification, and comparative risk assessment across cities. Although the present dataset covers selected U.S. cities, the methodological framework is transferable to other urban regions where pollution risk is shaped by similar environmental processes.

4.13 Response to the research questions

The present findings directly answer the research questions proposed in the Introduction. Ensemble-based machine learning models proved effective in classifying air pollution risk levels

using environmental, meteorological, and temporal indicators. The Soft Voting ensemble outperformed both standalone classifiers and the Stacking framework, confirming the added value of ensemble consensus. The ablation analysis showed that temporal, spatial, and engineered features all improved predictive performance, while the interpretability analysis identified PM2.5, PM10, ozone, and city-level variables as the most influential predictors. Finally, the cross-validation results demonstrated that the proposed model is stable and generalizable.

4.14 Limitations and Future Research Directions

Although the proposed ensemble-based framework has shown promising results, there are several limitations discussed below. First, the dataset is limited to the urban space of United States area, which reduces the possibility of geographic generalization of this model. Second, we still have class imbalance with few High_Risk observations. Third, a number of relevant atmospheric and anthropogenic drivers including wind speed, atmospheric pressure, solar radiation, traffic density and industrial emissions are not included among the available predictors.

Future research may seek to combine environment data drawn from multiple sources such as satellite observations and atmospheric transport models. Moreover, advanced spatio-temporal modelling and deep learning approaches have the potential to better represent complex environmental patterns. Integration of other GIS-based tools with the outputs of machine learning may also increase the practical utility of the framework for environmental planning and sustainable air-quality management.

Conclusion

This study an environmental framework was created for classifying the risk of air pollution using real-world urban air quality data in addition to meteorological, temporal and engineered indicators. The results showed that the Soft Voting ensemble model outperformed baseline classifiers and demonstrated robust performance in classifying pollution risk levels. The results validated that the severity of air pollution is driven not just by pollutant concentrations, but also seasonal variation, atmospheric conditions and urban context. While ablation and interpretability tests showed that temporal, geographic, and engineering aspects all made significant contributions to model performance, particulate matter and ozone were the most important factors in all models. Additionally, the study demonstrates how environmental evaluation might benefit from artificial intelligence. AI was used to transform complicated environmental data into interpretable risk buckets that may guide urban monitoring, pollution "hotspot" detection, and sustainable air-quality management, rather than only serving as a forecasting tool.

Overall, the framework provides a useful foundation for conducting data-driven environmental risk screening in urban settings.

References

[1] Li, G., Tang, Y., & Yang, H. (2022). A new hybrid prediction model of air quality index based on secondary decomposition and improved kernel extreme machine learning. Chemosphere, 305, 135348.

[2] Harrison, R. M., & Yin, J. (2000). Particulate matter in the atmosphere: Which particle properties are important for its effects on health? Science of the Total Environment, 249(1–3), 85–101.

[3] Jo, E. J., Lee, W. S., Jo, H. Y., Kim, C. H., Eom, J. S., Mok, J. H., Kim, M. H., Lee, K., Kim, K. U., & Lee, Mects of particulate matter on respiratory disease and the impact of meteorological factors in Busan, Korea. Respiratory Medicine, 124, 79–87.. K. (2017). Eff

[4] Perrino, C., Tiwari, S., Catrambone, M., Dalla Torre, S., Rantica, E., & Canepari, S. (2011). Chemical characterization of atmospheric PM in Delhi, India, during different periods of the year including Diwali festival. Atmospheric Pollution Research, 2(4), 418–427.

[5] Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., & Baklanov, A. (2012). Real-time air quality forecasting, part I: History, techniques, and current status. Atmospheric Environment, 60, 632–655.

[6] Dye, T. S. (2013). Guidelines for developing an air quality (ozone and PM2.5) forecasting program. United States Environmental Protection Agency, Washington, DC, USA.

[7] Zheng, H., Li, H., Lu, X., &Ruan. T. (2018). A multiple kernel learning approach for air quality prediction. Advances in Meteorology, 2018.

[8] International Journal of Artificial Intelligence in Medical Issues. (2025). https://doi.org/10.56705/ijaimi.v312.322

[9] Kumar, S., Vishwakarma, A., Srivastava, M. K., Perwej, Y., & Akhtar, N. (2026). Ensemble machine learning for reliable air pollution prediction and sustainable environmental management. International Journal of Scientific Research in Science and Technology. Available at: www.ijsrst.com

[10] Moskal, A., Jagodowicz, W., Penconek, A., Zaraska, K. Low-Cost Sensor System for Air Purification Process Evaluation. Sensors 2024, 24, 1769.

[11] Nuwairy El Furqany. (2025). Optimizing air quality index classification using multiple machine learning models and oversampling techniques. International Journal of Artificial Intelligence in Medical Issues, 312. ISSN 3025-4167.

[12] Jaron, A., Berucka, A., Delis, P., & Sekrecka, A. (2024). An assessment of the possibility of using unmanned aerial vehicles to identify and map air pollution from infrastructure emissions. Energies, 17, 577.

[13] Bemacki, J., & Schence, R. (2025). A comprehensive review of data-driven techniques for air pollution concentration forecasting. Sensors, 25, 6044. https://doi.org/10.3390/25196044

[14] Johnson, T.; Woodward, K. Enviro-IoT: Calibrating Low-Cost Environmental Sensors in Urban Settings. arXia 2025, arXiv:2502.07596.

[15] EnviroDataScience. “Air Quality Dataset.” *Kaggle*, 1 Sept. 2025, www.kaggle.com/datasets/price438/air-quality-dataset.