

# **PREDICTIVE ANALYTICS**

## **PROJECT REPORT**

**(AUG – DEC 2025)**

### **Air Quality Prediction and Analysis Using Machine Learning**

**Submitted by**

**Aditya Kumar**

**Reg. No.- 12413688**

**B-Tech**

**Computer Science and Engineering**

**Section – K23BR**

**INT234**

Under the guidance of

**Prof. Vikas Mangotra**

**Assistant Professor**

**UID -31488**

**Discipline of CSE/IT**

**Lovely School Of Engineering**



**L OVELY  
P ROFESSIONAL  
U NIVERSITY**

### **DECLARATION**

I, **Aditya Kumar** student of Bachelor of Technology under CSE/IT Discipline at, **Lovely Professional University, Punjab**, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 15/12/25

Registration No.- **12413688**

Signature

**Aditya Kumar**

## **CERTIFICATE**

This is to certify that **Aditya Kumar** bearing Registration no. **12413688** has completed INT 234 project titled, “**Air Quality Prediction and Analysis Using Machine Learning**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Vikash Mangotra**

**Assistant Professor**

**School of Computer Science and Engineering**

**Lovely Professional University**

Phagwara, Punjab.

Date: 15-12-25

## ACKNOWLEDGEMENT

I would like to express my sincere and heartfelt gratitude to Prof. Vikas Mangotra sir for their invaluable guidance, continuous encouragement, and constructive feedback throughout the successful completion of this project. Their expertise, patience, and insightful suggestions played a vital role in shaping the direction of this work and improving its overall quality.

I am deeply thankful to the department and the institution for providing a supportive academic environment, necessary infrastructure, and access to learning resources that enabled me to explore and apply predictive analytics and machine learning techniques effectively. The knowledge and skills gained through the curriculum served as a strong foundation for carrying out this project.

I would also like to acknowledge the contribution of open-source platforms, datasets, and tools that made it possible to work on real-time air quality data. The availability of such resources allowed me to gain practical exposure to data pre-processing, exploratory data analysis, model building, evaluation, and visualization, thereby enhancing my understanding of real-world data science applications.

I extend my sincere appreciation to all the teachers and mentors who have directly or indirectly contributed to my academic growth and inspired me to develop analytical and problem-solving skills. Their guidance has been instrumental in building my interest in predictive analytics and machine learning.

Finally, I express my heartfelt thanks to my family and friends for their constant support, motivation, and encouragement throughout the duration of this project. Their belief in me and moral support helped me stay focused and complete this work successfully.

[Git hub Link](#)

[LinkedIn Link](#)

**Table of contents**

<b>SI No.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
<b>01</b>	<b>INTRODUCTION</b>	<b>06</b>
<b>02</b>	<b>SOURCE OF DATASET</b>	<b>09</b>
<b>03</b>	<b>DATA PRE-PROCESSING</b>	<b>11</b>
<b>04</b>	<b>ANALYSIS OF DATASET</b>	<b>12</b>
<b>05</b>	<b>MODEL DEVELOPMENT</b>	<b>21</b>
<b>06</b>	<b>MODEL COMPARISON</b>	<b>24</b>
<b>07</b>	<b>BEST MODEL IDENTIFICATION</b>	<b>26</b>
<b>08</b>	<b>CONCLUSION</b>	<b>28</b>
<b>09</b>	<b>FUTURE SCOPE</b>	<b>28</b>
<b>10</b>	<b>REFERENCES</b>	<b>29</b>

# **1.INTRODUCTION**

## **1.1 Overview of Predictive analytics**

- Predictive analytics is a method of studying data to understand what is likely to happen in the future. It works by analyzing past data and current data to find patterns, trends, and relationships. These patterns are then used to make predictions using machine learning and statistical techniques. Instead of only looking at what has already happened, predictive analytics helps in planning ahead.
- Today, large amounts of data are being generated in almost every field. Predictive analytics helps in turning this raw data into useful information. It is widely used in areas such as healthcare, business, transportation, weather forecasting, and environmental monitoring. By using predictive models, organizations and researchers can make better decisions and reduce uncertainty.

## **1.2 Importance of Prediction in Selected domain**

- Air quality is an important factor that affects human health, the environment, and daily life. High levels of air pollution can cause serious health problems such as breathing difficulties, asthma, heart diseases, and long-term lung damage. Children, elderly people, and individuals with existing health issues are especially affected by poor air quality.
- Predicting air pollution levels in advance is useful for both individuals and authorities. It helps people take safety measures, such as avoiding outdoor activities during high pollution periods. Government agencies can use predictions to issue warnings, manage traffic, control industrial emissions, and plan pollution reduction strategies. Therefore, air quality prediction plays an important role in protecting public health and improving environmental conditions.

### **1.3 Problem Statement**

- Although air quality data is collected in real time from various monitoring stations, making accurate predictions from this data is not easy. The dataset often contains missing values, different pollutant types, and large variations across locations and time. The relationship between pollution levels and environmental factors is complex and not always linear.
- Traditional data analysis methods are limited in handling such complex datasets. As a result, there is a need for a predictive analytics approach that can clean the data, handle inconsistencies, and apply machine learning techniques to predict pollution levels accurately. This project aims to address these challenges by using supervised and unsupervised learning methods.

### **1.4 Objectives of the Study**

- The main objectives of this project are:
- To collect and pre-process real-time air quality data from multiple locations.
- To analyze the dataset and understand relationships between different air quality features.
- To build regression models for predicting pollution intensity.
- To classify air quality levels into Low, Moderate, and High categories using classification algorithms.
- To identify pollution patterns and group similar locations using clustering techniques.

## **1.5 Scope of Project**

- The scope of this project is limited to the analysis of the given real-time air quality dataset. The project focuses on applying predictive analytics techniques such as regression, classification, and clustering. It does not include live data streaming or real-time deployment.
- The project is mainly intended for academic and learning purposes. However, the methods used can be extended further by adding weather parameters, real-time data updates, and a web or mobile-based interface for public use.

## **1.6 Expected Outcomes**

The expected outcomes of this project include:

- A clean and well-prepared dataset suitable for machine learning analysis.
- Reliable prediction of pollution values using regression models.
- Clear classification of air quality into Low, Moderate, and High levels.
- Identification of pollution patterns across different geographical areas.
- Improved understanding of how predictive analytics can be applied to environmental data.

## **2. Source of Dataset**

### **2.1 Name of Dataset**

The dataset used in this project is titled “**Real-Time Air Quality Index from Various Locations**”. It contains air quality information collected from multiple monitoring stations across different regions.

### **2.2 Source of Dataset**

The dataset has been obtained from the **Government of India Open Data Portal (data.gov.in)**, which provides publicly accessible datasets for research and academic purposes.

**Source Link:**

<https://www.data.gov.in/resource/real-time-air-quality-index-various-locations>

**Source Type:** Government Portal (Open Data Source)

The data is provided in a structured format and includes real-time air quality measurements collected by authorized monitoring agencies.

## 2.3 Description of Attributes

The dataset includes the following attributes:

Attribute Name	Description
latitude	Latitude of the air quality monitoring station
longitude	Longitude of the air quality monitoring station
country	Country where the monitoring station is located
state	State of the monitoring location
city	City of the monitoring location
station	Name of the air quality monitoring station
pollutant_id	Type of pollutant being measured
pollutant_avg	Average pollution value
pollutant_min	Minimum pollution value recorded
pollutant_max	Maximum pollution value recorded
last_update	Date and time of the observation

## 2.4 Number of Rows and Columns

- **Number of Rows:** Approximately 3200 or more records
- **Number of Columns:** 11 attributes

*(The number of rows may slightly change after data cleaning.)*

## 2.5 Type of Prediction

This project involves the following types of predictions:

- **Regression:** To predict the average pollution level.
- **Classification:** To classify air quality into Low, Moderate, and High categories.
- **Clustering:** To identify patterns and group locations based on pollution levels

### **3. DATASET PREPROCESSING**

Dataset preprocessing is an important step in any predictive analytics project. Raw data often contains missing values, duplicate records, and inconsistent formats, which can affect model performance. In this project, several pre-processing steps were applied to ensure that the air quality dataset was clean, consistent, and suitable for machine learning models.

#### **3.1 Data Cleaning**

The dataset was first checked for duplicate records. Duplicate entries can lead to biased results, so they were removed from the dataset. Missing values were mainly present in numerical attributes related to pollution measurements. These missing values were handled by replacing them with the mean of their respective columns. This approach helps maintain the overall data distribution without removing useful records.

#### **3.2 Handling Date and Time**

The `last_update` column contains date and time information. This column was converted into a proper date-time format to ensure consistency and avoid errors during analysis. Invalid or incorrect date values were safely handled during conversion.

#### **3.3 Encoding Categorical Variables**

Machine learning models require numerical input. Therefore, categorical attributes such as country, state, city, station, and pollutant ID were converted into numerical form using label encoding. This step allowed the models to process location and pollutant-related information effectively.

#### **3.4 Feature Selection**

Relevant features were selected based on their importance and relationship with pollution levels. Attributes such as latitude, longitude, pollutant type, state, city, and station were chosen for model training. Irrelevant or redundant features were excluded to improve model accuracy and reduce complexity.

### **3.5 Feature Scaling**

Feature scaling was applied to ensure that all numerical features were on a similar scale. Standardization was performed using a standard scaling technique, which transforms data to have a mean of zero and a standard deviation of one. This step is especially important for distance-based and margin-based algorithms such as KNN and SVM.

### **3.6 Final Prepared Dataset**

After preprocessing, the dataset became clean, structured, and ready for analysis. The processed dataset was then used for exploratory data analysis, regression modeling, classification, and clustering tasks.

## **4. ANALYSIS ON DATASET**

The analysis of the dataset was carried out based on the defined objectives of the project. Different analytical techniques were applied to understand the data, evaluate model performance, and extract meaningful insights.

### **4.1 General Description**

The dataset consists of real-time air quality information collected from multiple monitoring stations across different locations. It includes geographical details such as latitude and longitude, location-based attributes like state and city, and pollution-related attributes such as pollutant type and pollution values.

After preprocessing, the dataset became structured and suitable for analysis. The data was then analyzed using exploratory data analysis, supervised learning models, and unsupervised learning techniques to meet the objectives of the project.

## 4.2 Analysis for Objective 1: Data Understanding and Exploration

### Specific Requirements

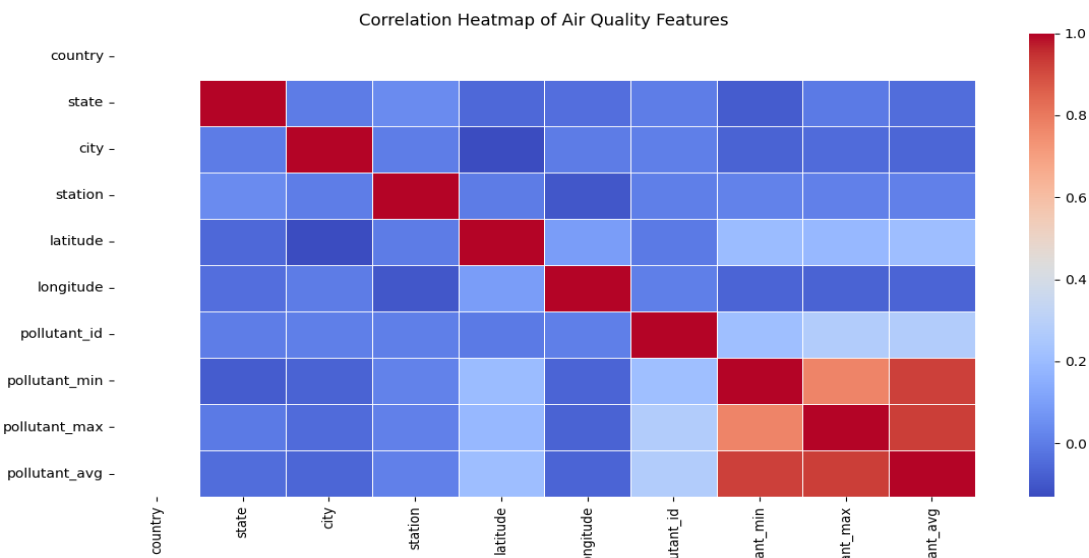
- Understand the structure and distribution of the dataset
- Identify relationships between numerical features
- Detect patterns and correlations in air quality data

### Analysis Results

Exploratory data analysis showed that pollution levels vary significantly across different locations and pollutants. Some numerical features showed strong relationships with pollution values, indicating their importance in prediction.

### Visualization

A **correlation heatmap** was used to visualize relationships between numerical attributes. The heatmap helped in identifying features that have a higher influence on pollution levels and supported feature selection for model building.



## **4.3 Analysis for Objective 2: Pollution Prediction Using Regression Models**

### **Specific Requirements**

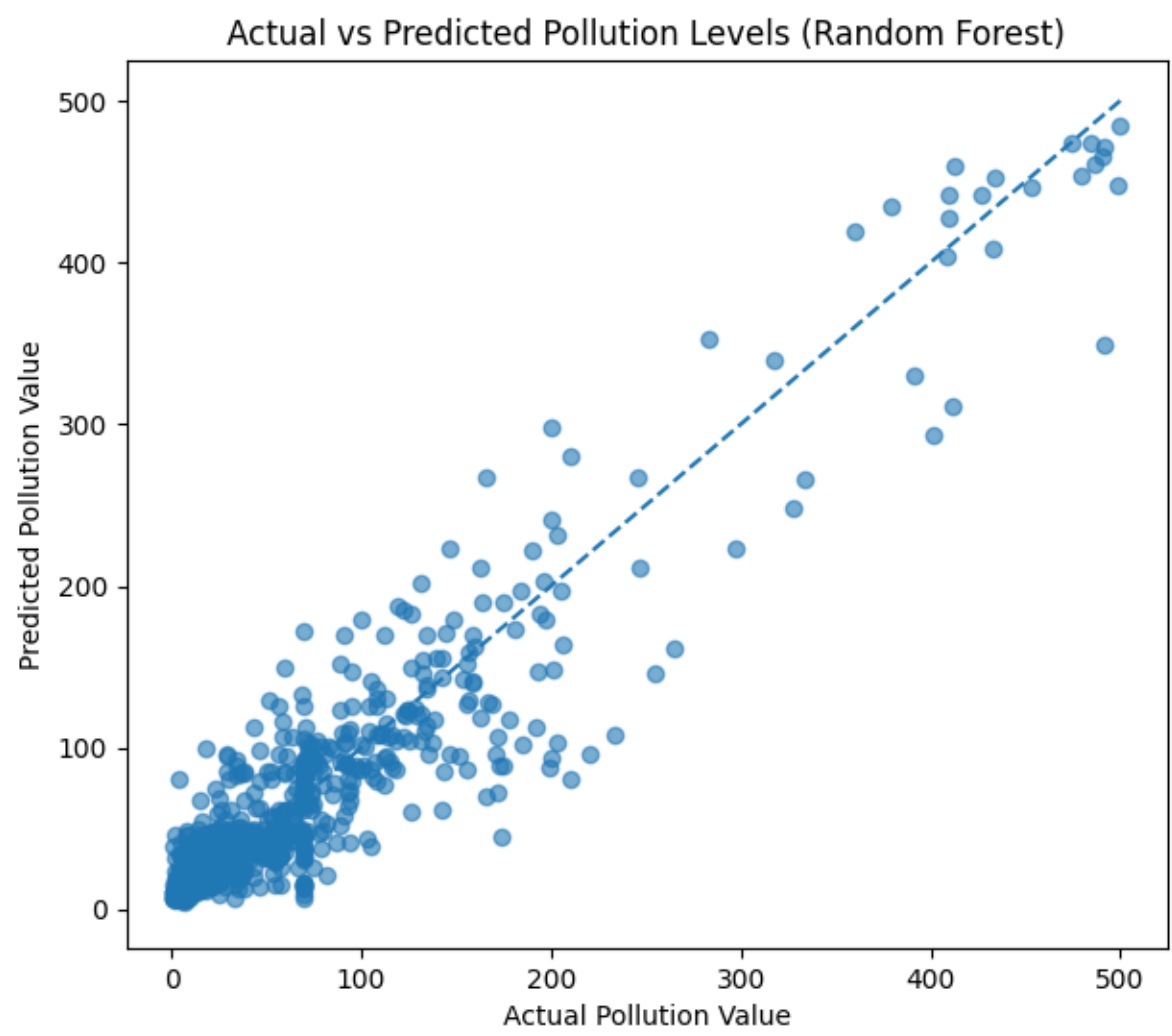
- Predict average pollution levels
- Compare the performance of multiple regression models

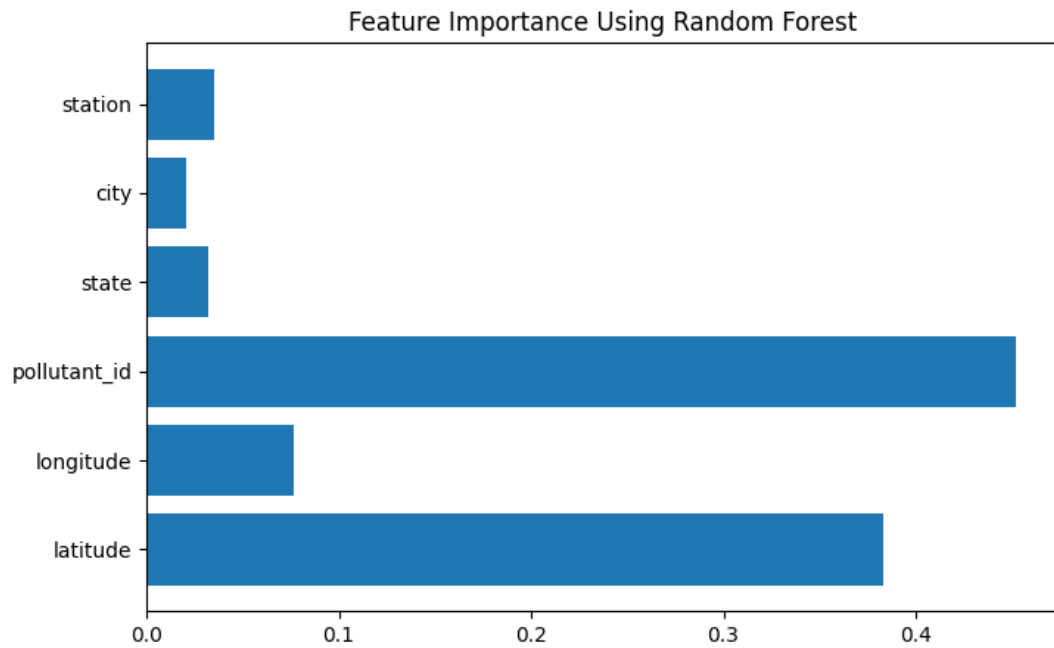
### **Analysis Results**

Three regression models were applied: Linear Regression, Polynomial Regression, and Random Forest Regression. Linear Regression provided a basic prediction, while Polynomial Regression captured non-linear patterns. Random Forest Regression performed the best, achieving the highest  $R^2$  score and lowest prediction error.

The comparison clearly showed that ensemble-based models handle complex air quality data more effectively.

**Visualization**





```
Regression Model Comparison
      Model  R2 Score      RMSE
0   Linear Regression  0.107121  85.264287
1 Polynomial Regression  0.180016  81.709693
2   Random Forest    0.867909  32.794945
```

```
Classification Model Comparison
      Model  Accuracy  F1 Score
0       KNN    0.652106  0.640724
1  Naive Bayes  0.594384  0.499516
2 Decision Tree  0.678627  0.680346
3       SVM    0.652106  0.613999
```

```
Clustering Summary
Cluster
1    1091
2    1088
0    1022
Name: count, dtype: int64
```

```
Project Executed Successfully
```

## 4.4 Analysis for Objective 3: Air Quality Classification

### Specific Requirements

- Categorize air quality into Low, Moderate, and High levels
- Evaluate classification models

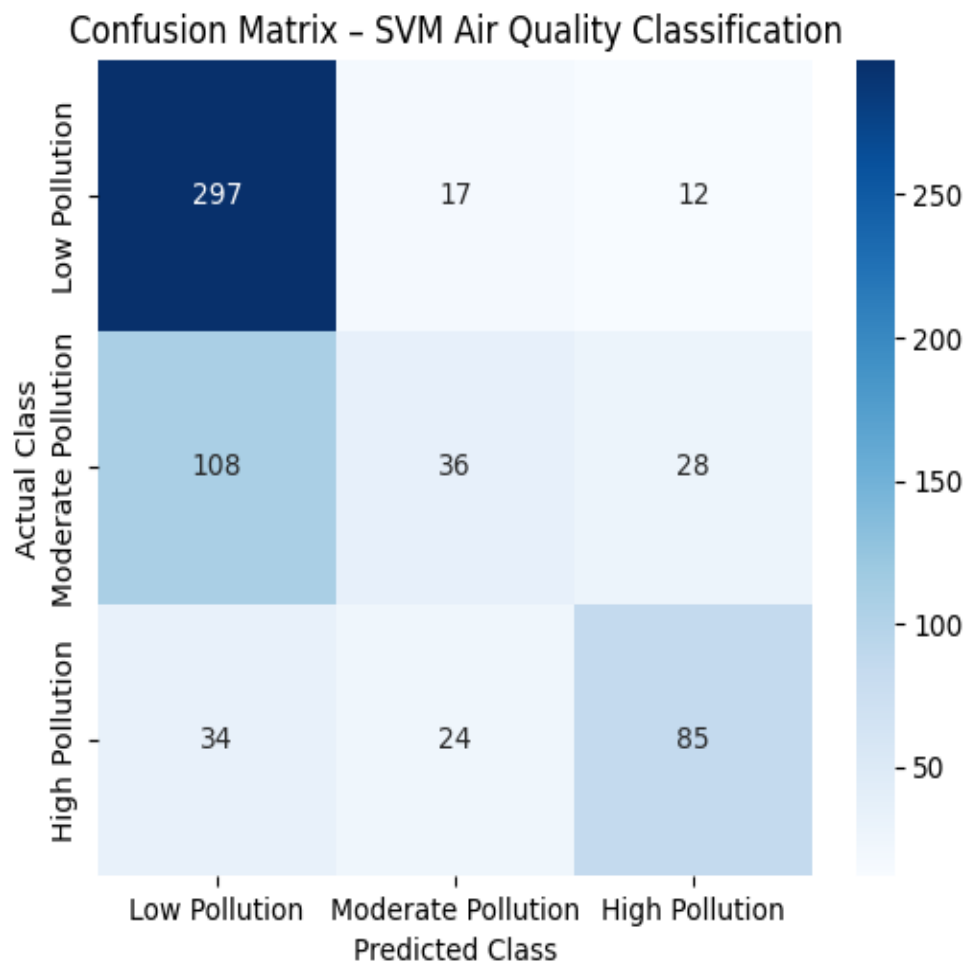
### Analysis Results

The pollution values were converted into three classes based on predefined thresholds. Four classification models were applied: KNN, Naive Bayes, Decision Tree, and SVM. Among these, SVM achieved the highest accuracy and F1 score, indicating better classification performance.

Some misclassification was observed between Moderate and High pollution levels, which is expected due to overlapping pollution ranges.

### Visualization

- **Classification comparison table** showing accuracy and F1 score
- **Confusion matrix**, which clearly displayed correct and incorrect classifications for each pollution category



## **4.5 Analysis for Objective 4: Pollution Pattern Identification Using Clustering**

### **Specific Requirements**

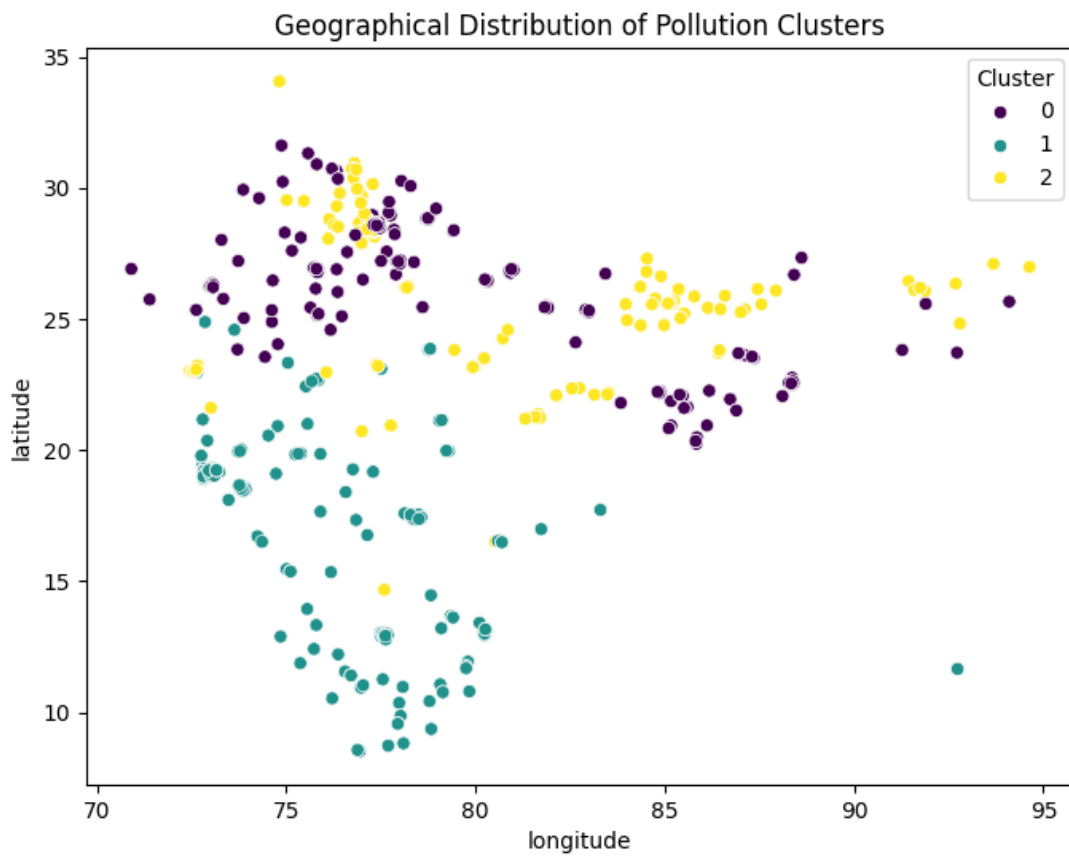
- Group locations based on pollution characteristics
- Identify hidden patterns without labeled data

### **Analysis Results**

K-Means clustering was applied to group locations into three clusters. Each cluster represented locations with similar pollution characteristics. The clustering results showed a fairly balanced distribution of data points among the clusters, indicating meaningful grouping.

### **Visualization**

A **geographical scatter plot** was used to visualize clusters based on latitude and longitude. This visualization helped in understanding how pollution patterns vary across different regions.



## 4.6 Summary of Analysis

The analysis successfully met all the objectives of the project. Exploratory analysis provided useful insights into the dataset, regression models accurately predicted pollution values, classification models effectively categorized air quality levels, and clustering revealed hidden pollution patterns. The use of visualizations made the analysis clear and easy to interpret.

## **5. MODEL DEVELOPMENT**

Model development is a key stage of this project, where machine learning algorithms were trained using the preprocessed air quality dataset. The goal of this stage was to build predictive models that can estimate pollution intensity, classify air quality levels, and analyze patterns in the data. Both supervised and unsupervised learning techniques were used to meet the project objectives.

### **5.1 Regression Model Development**

Regression models were developed to predict the average pollution value based on location and pollutant-related features. Three different regression models were trained and evaluated to compare their performance.

#### **Linear Regression**

Linear Regression was used as a baseline model to understand the relationship between input features and pollution values. It assumes a linear relationship between features and the target variable. This model helped in setting a reference point for comparing more complex models.

#### **Polynomial Regression**

Polynomial Regression was applied to capture non-linear relationships present in the air quality data. By transforming input features into polynomial terms, this model was able to improve prediction accuracy compared to simple linear regression.

#### **Random Forest Regression**

Random Forest Regression is an ensemble learning technique that combines multiple decision trees to make predictions. This model performed the best among the regression models as it can handle non-linear relationships and complex data patterns effectively. It also provided feature importance scores, which helped in understanding the contribution of each feature.

## **5.2 Classification Model Development**

Classification models were developed to categorize air quality into three levels: Low, Moderate, and High. These models were trained using labeled pollution categories derived from pollution values.

### **K-Nearest Neighbors (KNN)**

KNN is a distance-based classification algorithm that assigns a class based on the majority class of its nearest neighbors. It was used to understand how proximity-based learning performs on air quality data.

### **Naive Bayes**

Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes independence between features and provides fast classification. This model was included to compare probabilistic approaches with other classifiers.

### **Decision Tree**

Decision Tree classification works by splitting the dataset into branches based on feature values. It provides clear decision rules and is easy to interpret. This model performed well due to its ability to capture non-linear decision boundaries.

### **Support Vector Machine (SVM)**

Support Vector Machine was used to create an optimal decision boundary between different air quality classes. It achieved the highest classification accuracy and F1 score among the tested classifiers, making it the best-performing classification model in this project.

### **5.3 Clustering Model Development**

In addition to supervised learning models, an unsupervised learning approach was also applied.

#### **K-Means Clustering**

K-Means clustering was used to group locations based on pollution characteristics without using predefined labels. The data was divided into three clusters, representing different pollution patterns. This helped in identifying regions with similar air quality behavior.

### **5.4 Summary of Model Development**

In total, multiple predictive models were trained and evaluated:

- **Three regression models**
- **Four classification models**
- **One clustering model**

This multi-model approach ensured robust analysis and allowed meaningful comparison between different machine learning techniques for air quality prediction and analysis.

## 6. MODEL COMPARISON

Model comparison was carried out to evaluate the performance of different machine learning models used in this project. The models were compared using standard evaluation metrics to identify the most accurate and reliable models for air quality prediction and classification.

### 6.1 Regression Model Comparison

Three regression models were compared to predict average pollution values:

- Linear Regression
- Polynomial Regression
- Random Forest Regression

The comparison was based on **R<sup>2</sup> Score** and **Root Mean Squared Error (RMSE)**. The results showed that Linear Regression provided basic predictions, while Polynomial Regression improved accuracy by capturing non-linear patterns. **Random Forest Regression performed the best**, achieving the highest R<sup>2</sup> score and the lowest RMSE. This indicates that Random Forest Regression was more accurate and reliable for predicting pollution intensity.

### 6.2 Classification Model Comparison

Four classification models were compared to classify air quality levels:

- K-Nearest Neighbors (KNN)
- Naive Bayes
- Decision Tree
- Support Vector Machine (SVM)

The models were evaluated using **Accuracy** and **F1 Score**. Among all models, the **Decision Tree classifier achieved the highest accuracy and F1 score**, followed by SVM. KNN and Naive Bayes showed comparatively lower performance. The results indicate that the Decision Tree model was the most effective for classifying air quality levels in this project.

### **6.3 Clustering Model Evaluation**

K-Means clustering was used to group locations based on pollution characteristics. The dataset was divided into three clusters. The clusters were found to be balanced and meaningful, indicating that the algorithm successfully grouped locations with similar pollution patterns. Since clustering is an unsupervised task, evaluation was done based on cluster distribution and visual inspection rather than accuracy metrics.

## 7. BEST MODEL IDENTIFICATION

After training and comparing multiple machine learning models, the best-performing models were identified based on their evaluation results. The selection was done separately for regression and classification tasks to ensure fair and accurate assessment.

### 7.1 Best Regression Model

Among the regression models used in this project ,Linear Regression, Polynomial Regression, and Random Forest Regression. **Random Forest Regression** was identified as the best model.

This model achieved the **highest  $R^2$  score** and **lowest RMSE**, indicating that its predictions were closest to the actual pollution values compared to the other models. Random Forest Regression performed better because it can capture non-linear relationships and complex interactions between features. It also reduces overfitting by combining predictions from multiple decision trees.

Therefore, Random Forest Regression was selected as the best model for predicting pollution intensity.

### 7.2 Best Classification Model

For air quality classification, four models were evaluated: KNN, Naive Bayes, Decision Tree, and Support Vector Machine (SVM). Based on **accuracy and F1 score**, the **Decision Tree classifier** was identified as the best classification model.

The Decision Tree model achieved the highest accuracy and F1 score among all tested classifiers. It was effective in classifying air quality into Low, Moderate, and High categories by learning clear decision rules from the data. Its ability to handle non-linear patterns helped it perform better than the other classification models.

Hence, the Decision Tree classifier was chosen as the best model for classifying air quality levels.

### 7.3 Reason for Model Selection

The best models were selected based on the following reasons:

- Higher prediction accuracy
- Lower error values
- Better handling of non-linear and complex data patterns
- Consistent performance across evaluation metrics

### 7.4 Summary

In conclusion:

- **Random Forest Regression** is the best model for predicting pollution values.
- **Decision Tree classifier** is the best model for classifying air quality levels.

These models provide reliable results and are suitable for air quality prediction and analysis based on the given dataset.

## 8. CONCLUSION

This project successfully applied predictive analytics and machine learning techniques to analyze real-time air quality data. The dataset was cleaned and pre-processed to handle missing values, duplicates, and categorical features. Exploratory data analysis helped in understanding the relationships between different air quality attributes.

Multiple regression, classification, and clustering models were developed and evaluated. Among the regression models, **Random Forest Regression** performed the best by providing more accurate pollution predictions with lower error values. For classification, the **Decision Tree model** achieved the highest accuracy and F1 score, making it the most effective model for classifying air quality into Low, Moderate, and High levels. K-Means clustering also successfully identified meaningful pollution patterns across different locations.

Overall, the project demonstrated that machine learning models can be effectively used to analyze and predict air quality data. The results highlight the importance of data-driven approaches in understanding environmental conditions and supporting better decision-making.

## 9. FUTURE SCOPE

Although the project achieved its objectives, there is scope for further improvement and extension. In the future, additional weather-related parameters such as temperature, humidity, wind speed, and rainfall can be included to improve prediction accuracy. Real-time data streaming can be integrated to provide live pollution predictions.

Advanced machine learning and deep learning models can also be explored for better performance. The system can be extended by developing a web or mobile application to make the predictions easily accessible to users. Integration with alert systems can help notify people about high pollution levels in advance.

## 10. REFERENCES

- [1] Government of India, “Real-Time Air Quality Index from Various Locations,” *Open Government Data (OGD) Platform India*. [Online]. Available: <https://www.data.gov.in/resource/real-time-air-quality-index-various-locations>. [Accessed: Aug. 2025].
- [2] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2019.
- [3] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2017.