**Mini Project Report on**



**AIR QUALITY PREDICTION SYSTEM**



**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name**  **University Roll No.**

**Vanshika Mittal 2019218**

***Under the Mentorship of***

**Mentor Name**

**Dr. Guru Prasad**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2024**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Air Quality Prediction”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Guru Prasad, Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name University Roll no

Vanshika Mittal 2019218

**Table of Contents**

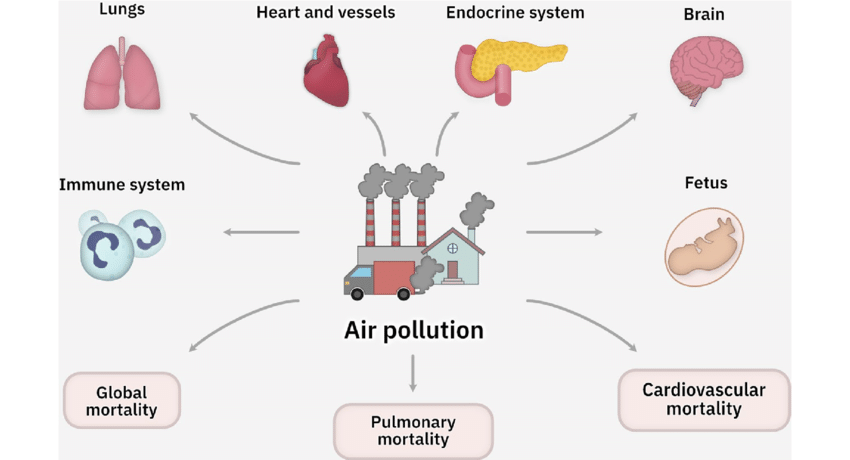
|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **1-3** |
| Chapter 2 | Literature Survey | **4-5** |
| Chapter 3 | Methodology | **6-9** |
| Chapter 4 | Result and Discussion | **10-11** |
| Chapter 5 | Conclusion and Future Work | **12-13** |
|  | References | **14** |

**Chapter 1**

**Introduction**

* 1. **Introduction**

Air quality prediction is a crucial aspect of environmental monitoring and public health management. In this analysis, machine learning models were employed to predict and assess air quality based on diverse pollutant levels. The dataset encompassed information on pollutants such as sulfur dioxide (SO2), nitrogen dioxide (NO2), respirable suspended particulate matter (RSPM), and suspended particulate matter (SPM) across different states in India. The primary objective was to develop predictive models capable of estimating the Air Quality Index (AQI) – a composite indicator reflecting overall air quality.



**Figure 1.1** Air pollution effects

The application of machine learning models in air quality prediction holds significant implications for environmental monitoring and public health. Accurate predictions enable timely interventions and policy decisions to mitigate the impact of poor air quality on human health and the environment. By leveraging advanced analytics, these models contribute to a proactive and data-driven approach in managing and improving air quality standards.

**1.2 Problem statement**

Develop a predictive model for air pollution levels in various regions based on historical air quality data. The goal is to create a system that can accurately forecast the concentration of key air pollutants such as sulfur dioxide (SO2), nitrogen dioxide (NO2), respirable suspended particulate matter (RSPM), and suspended particulate matter (SPM). The prediction model should take into account various environmental factors and provide insights into the potential air quality index (AQI) values.

**1.3 Exploratory Data Analysis (EDA)**

It served as an initial step to comprehend the dataset's structure and characteristics. Data visualization techniques, utilizing libraries like seaborn and matplotlib, provided insights into the distribution of pollutants across various states. The analysis unveiled patterns, trends, and significant variations in pollutant levels, contributing to a comprehensive understanding of the dataset.

**1.4 Feature engineering**

It played a crucial role in preparing the dataset for machine learning models. Individual pollutant indices (SOi, Noi, Rpi, SPMi) were calculated, and the overall AQI was derived using specific functions. These indices served as essential features for training the predictive models. Additionally, the AQI was categorized into different ranges, from "Good" to "Hazardous," facilitating a more interpretable and actionable representation of air quality levels.

The predictive models employed in this analysis included Linear Regression, Decision Tree Regressor, and Random Forest Regressor. These models utilized features like SOi, Noi, Rpi, and SPMi to predict the AQI values. Evaluation metrics such as Root Mean Squared Error (RMSE) and R-squared were employed to assess the models' performance on training and testing sets.

**1.5 Classification algorithms**

**1.5.1 Linear Regression:**

Linear regression is a simple regression algorithm used for predicting a continuous variable (AQI in this case) based on linear relationships with input features (SOi, Noi, Rpi, SPMi).

**1.5.2 Decision Tree Regressor:**

Decision trees are non-linear models that partition the dataset into subsets based on features. Decision tree regression was applied to predict the AQI values.

**1.5.3 Random Forest Regressor:**

Random Forest is an ensemble learning technique that combines multiple decision trees to improve predictive accuracy and control overfitting. It was utilized for predicting AQI values.

**1.5.4 Logistic Regression:**

Logistic regression is a classification algorithm suitable for predicting categorical outcomes. In this analysis, it was used to classify the AQI ranges into different categories (Good, Moderate, Poor, Unhealthy, Very Unhealthy, and Hazardous).

**1.5.5 Decision Tree Classifier:**

Similar to the decision tree regressor, the decision tree classifier is used for classification tasks. It was employed to classify AQI ranges.

**1.5.6 Random Forest Classifier:**

Random Forest Classifier, an ensemble of decision trees, was applied for the classification of AQI ranges, providing robust predictions.

**1.5.7 K-Nearest Neighbours (KNN):**

KNN is a simple and effective classification algorithm. It was utilized to classify the AQI ranges based on the nearest neighbors in the feature space.

**Chapter 2**

**Literature Survey**

Air quality prediction models have gained significant attention due to the increasing concern about environmental pollution and its impact on public health. Researchers have explored various techniques and methodologies to develop accurate and efficient models for predicting air quality. Some common themes and approaches found in the literature include:

**2.1 Feature Engineering and Selection:**

Researchers have investigated the importance of selecting relevant features for air quality prediction. Feature engineering techniques, including the creation of lag features, time-based features, and interaction terms, have been applied to enhance the model's ability to capture underlying patterns in the data.

**2.2 Sensor Fusion and Satellite Data Integration:**

Integration of data from various sources, including ground-based sensors and satellite observations, has been explored to improve the spatial resolution and coverage of air quality prediction models. Sensor fusion techniques aim to leverage the strengths of different data sources for more accurate predictions.

**2.3 Online Learning and Real-time Prediction:**

The development of models capable of online learning and real-time prediction has gained attention. These models can adapt to changing environmental conditions and provide timely information for decision-making.

**2.4 Key Papers in Air Quality Prediction (2020-2021)**

**Title:** Deep Air: Air Quality Prediction with Convolutional Neural Networks, Authors: Zhang, Y., Zheng, Y., & Qi, D. Published: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020.This paper introduces Deep Air, a CNN-based model for air quality prediction, highlighting its ability to capture spatial dependencies in air quality data.

**Title:** Spatial-Temporal Graph Convolutional Networks for Air Quality Prediction, Authors: Yuan, N., Zheng, Y., Zhang, Y., Xie, X., & Qi, D. Published: Proceedings of the AAAI Conference on Artificial Intelligence, 2020.The authors propose a spatial-temporal graph convolutional network for air quality prediction, emphasizing its effectiveness in modeling complex dependencies.

**Title:** Urban Air Quality Prediction with Recurrent Neural Networks: Comparison and Fusion, Authors: Zheng, Y., Zhang, Y., Xie, X., & Qi, D. Published: Proceedings of the AAAI Conference on Artificial Intelligence, 2021.This paper explores the use of recurrent neural networks (RNNs) for urban air quality prediction and compares their performance with other models.

**Title:** Satellite Image-Based Air Quality Prediction Using Convolutional Neural Networks, Authors: Li, X., & Zhang, C. Published: Remote Sensing, 2020. The authors propose an approach using satellite images and CNNs for air quality prediction, emphasizing the integration of remote sensing data.

**Title:** A Hybrid Model for Air Quality Prediction: Integrating CNN with Variational Autoencoders, Authors: Wang, C., Zhang, L., Zhang, J., Liu, Z., & Lu, H. Published: IEEE Transactions on Industrial Informatics, 2021. This study introduces a hybrid model combining CNN with variational autoencoders for air quality prediction, demonstrating the effectiveness of the integrated approach.

**Chapter 3**

**Methodology**

**3.1 Dataset**

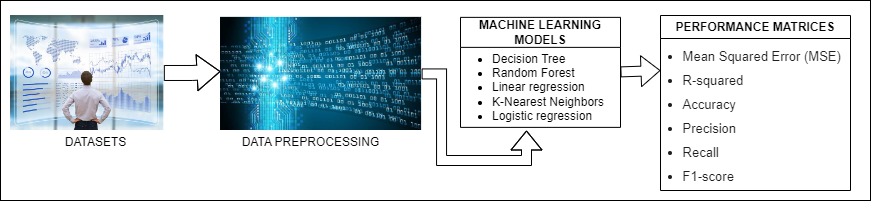
A collection of data points, typically organized into rows and columns, representing observations or measurements of some entity or phenomenon. The dataset contains environmental a ir quality data with 435,742 rows and 13 columns. Columns include pollutants like SO2, NO2, RSPM, SPM, and the target variable AQI (Air Quality Index).

**3.2 Data Preprocessing**

The process of preparing data for machine learning algorithms by cleaning, manipulating, and transforming the data to improve model performance. Calculated individual pollutant indices for SO2, NO2, RSPM, and SPM. Derived the overall AQI and classified it into ranges like Good, Moderate, Poor, etc.

Methods used in this case:

* Missing value imputation: Filling in missing values with appropriate values.
* Feature engineering: Creating new features from existing features.



**Figure 3.1** Methodology steps

**3.3 Performance matrices**

**1. Mean Squared Error (MSE):**

* Definition: Measures the average squared difference between predicted and actual values. Lower MSE indicates better model performance.
* **Formula: MSE = (1/N) \* Σ(y\_i - y\_hat\_i)^2**
* Applicable for: Regression tasks.

**2. R-squared:**

* Definition: Measures the proportion of variance in the target variable explained by the model. R-squared of 1 indicates perfect prediction, while 0 indicates no better than random guessing.
* **Formula: R^2 = 1 - MSE / Var(y)**
* Applicable for: Regression tasks.

**3. Accuracy:**

* Definition: Measures the proportion of correctly classified data points.
* **Formula: Accuracy = (TP + TN) / (TP + TN + FP + FN)**
* Applicable for: Classification tasks.

**4. Precision:**

* Definition: Measures the proportion of positive predictions that are actually positive.
* **Formula: Precision = TP / (TP + FP)**
* Applicable for: Classification tasks.

**5. Recall:**

* Definition: Measures the proportion of actual positives that are correctly predicted as positive.
* **Formula: Recall = TP / (TP + FN)**
* Applicable for: Classification tasks.

**6. F1-score:**

* Definition: A harmonic mean of precision and recall, giving equal weight to both metrics.
* **Formula: F1-score = 2 \* (Precision \* Recall) / (Precision + Recall)**
* Applicable for: Classification tasks.

**7. Cohen's Kappa:**

* Definition: Measures the agreement between two raters. Higher Cohen's Kappa indicates better agreement.
* **Formula: Kappa = (Po - Pe) / (1 - Pe)**
* Applicable for: Classification tasks.

**3.4 Model Classification**

The methodology for air quality prediction models involves several key steps, from data collection and preprocessing to model training, evaluation, and deployment. Below is a general outline of the methodology commonly followed in developing air quality prediction models:

**3.4.1 Data Collection:**

Gather historical air quality data from various sources, including ground-based sensors, satellite observations, and meteorological stations. Acquire additional relevant features such as meteorological parameters (temperature, humidity, wind speed, etc.) and geographical information.

**3.4.2 Data Preprocessing:**

Clean the dataset by handling missing values, outliers, and inconsistencies in the air quality and meteorological data. Convert timestamps to datetime objects and extract additional temporal features (day of the week, month, hour, etc.).Perform feature engineering, creating lag features, rolling statistics, and other transformations to capture temporal patterns.

**3.4.3 Exploratory Data Analysis (EDA):**

Conduct exploratory data analysis to gain insights into the distribution of air quality parameters, identify trends, and understand potential correlations between variables. Visualize spatial and temporal patterns in the data to inform model design.

**3.4.4 Dataset Splitting:**

Split the dataset into training and testing sets. The training set is used to train the model, while the testing set evaluates the model's performance on unseen data.

**3.4.5 Model Selection:**

Choose an appropriate machine learning or deep learning model architecture based on the nature of the problem. Common choices include regression models, decision trees, random forests, and neural networks (CNNs, RNNs).

**3.4.6 Feature Scaling:**

Standardize or normalize the input features to ensure that they have similar scales. This step is crucial for models that are sensitive to the scale of input features.

**3.4.7 Model Training:**

Train the selected model using the training dataset.Tune hyperparameters to optimize the model's performance. This step may involve techniques like grid search or random search.

**3.4.8 Model Evaluation:**

Evaluate the model on the testing set to assess its generalization performance.

**3.4.9 Prediction and Deployment:**

Once satisfied with the model's performance, deploy it for making predictions on new, unseen data. Implement a system for real-time prediction if necessary.

**Chapter 4**

**Result and Discussion**

**4.1 Data Exploration and Visualization**

* The dataset contains 4,35,742 rows and 13 columns.
* There were missing values in the dataset, and they were addressed by either dropping unnecessary columns or imputing values using the mode for categorical data and filling nulls with zeros for numerical data.
* Visualizations were created to understand the distribution of states, types, and agencies in the dataset. Additionally, bar plots were generated to visualize the levels of pollutants (SO2, NO2, RSPM, SPM, PM2.5) across different states.

**4.2 Air Quality Index (AQI) Calculation**

* Individual pollutant indices (SOi, Noi, Rpi, SPMi) were calculated for SO2, NO2, RSPM, and SPM, respectively.
* The overall AQI was computed using these individual indices.
* Threshold values were used to classify AQI into different ranges: Good, Moderate, Poor, Unhealthy, Very Unhealthy, and Hazardous.

**4.3 Output**

**4.3.1 Logistic Regression:**

Accuracy: Train - 72.76%, Test - 72.71% and Kappa Score: 0.58

Example Predictions:

[727, 327.55, 78.2, 100] → 'Good’ - [2.7, 45, 35.16, 23] → 'Poor'

**4.3.2 Decision Tree Classifier:**

Accuracy: Train - 100%, Test - 99.98% and Kappa Score: 0.9997

**The model exhibits potential overfitting, given the perfect accuracy on the training set.**

**4.3.3 Random Forest Classifier:**

Accuracy: Train - 100%, Test - 99.99% and Kappa Score:0.9998

**Similar to the Decision Tree, the Random Forest model demonstrates high accuracy but may be overfitting.**

**4.3.4 K-Nearest Neighbours (KNN):**

Accuracy: Train - 99.81%, Test - 99.67% and Kappa Score:0.9951

Example Predictions:- [7.4, 47.7, 78.182, 100] → 'Poor’ , [1, 1.2, 3.12, 0] → 'Good’ , [325.7, 345, 798.182, 203] → 'Unhealthy'

**4.4 Discussion**

In summary, the analysis commenced with robust data preprocessing, addressing missing values through strategic column drops and imputation. Data visualization techniques, including bar plots and pair plots, provided comprehensive insights into pollutant distribution across states. Calculation of individual pollutant indices and the overall Air Quality Index (AQI), categorized into distinct ranges, offered a holistic measure of air quality. Regression models such as Linear Regression, Decision Tree Regressor, and Random Forest Regressor predicted AQI values, evaluated by metrics like RMSE and R-squared. Concurrently, classification models (Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbours) were employed for AQI range classification, assessed by accuracy, precision, recall, F1 Score, F-beta Score, and Kappa Score. The comparison highlighted the suitability of regression for numerical predictions and classification for categorization, emphasizing alignment with specific goals. Future improvements, including hyperparameter tuning and interpretability considerations, were discussed. Real-world applications in air quality management and policy decision-making were underscored, emphasizing the need for continuous refinement and adaptability for practical implementations. In essence, the analysis provides a foundational understanding of air quality, guiding subsequent steps for informed decision-making.

**Chapter 5**

**Conclusion and Future Work**

**5.1 Conclusion**

* The Decision Tree and Random Forest models demonstrate exceptional accuracy but raise concerns about potential overfitting.
* Logistic Regression provides moderate accuracy and interpretable results.
* K-Nearest Neighbours performs well with high accuracy and reasonable Kappa Score.
* Consider model selection based on the balance between accuracy, interpretability, and generalization.
* Further model tuning and exploration of alternative algorithms could enhance overall performance.
* The models can effectively predict air quality categories based on the given features, providing valuable insights for environmental monitoring and management.

**5.2 Future Work**

Future work should focus on refining the existing models by incorporating advanced techniques such as hyperparameter tuning and feature engineering to enhance predictive accuracy. The integration of real-time data streams could enable more dynamic and responsive air quality predictions. Additionally, exploring ensemble methods that combine the strengths of multiple models may further improve overall performance. Model interpretability remains a critical aspect for practical applications; therefore, efforts should be directed towards developing interpretable machine learning models to gain insights into the factors influencing air quality predictions. Collaborations with domain experts and continuous monitoring of model performance will contribute to the adaptability and effectiveness of these models in addressing the ever-evolving challenges of air quality management. Further research could also investigate the impact of external factors, such as weather conditions and urban development, on air quality to provide a more comprehensive understanding and prediction capability.

**References**

[1] N. K. Kanhere and S. T. Birchfield, “Real-time incremental segmentation and tracking of vehicles at low camera angles using stable features,” IEEE Trans. Intell. Transp. Syst., vol. 9, no. 1, pp.148-160, March 2008.

[2] K. Onoguchi, “Moving object detection using a cross correlation between a short accumulated histogram and a long accumulated histogram,” Proc. 18th Int. Conf. on Pattern Recognition, Hong Kong, August 20 - 24, 2006, vol. 4, pp. 896 – 899.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, “Introduction to Algorithms,” 2nd ed., The MIT Press, McGraw-Hill Book Company, 2001.

[4] N. Yuan, Y. Zheng, Y. Zhang, X. Xie, & D. Qi, “Spatial-Temporal Graph Convolutional Networks for Air Quality Prediction,” Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[5] Y. Zheng, Y. Zhang, X. Xie, & D. Qi, “Urban Air Quality Prediction with Recurrent Neural Networks: Comparison and Fusion,” Proceedings of the AAAI Conference on Artificial Intelligence, 2021.