



(RESEARCH ARTICLE)



COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR AIR QUALITY AND POLLUTION ASSESSMENT

Charishma Purna Aki ^{1,*}, Sri Surya Lakshmi Sowjanya Saladi ¹, Prabhu Sagar Telagamsetti ¹, Harish E ¹ and Lalu Prasad Jammana ²

¹ Department of Information Technology (IT), Aditya University, Surampalem - Pin 533437, Andhra Pradesh, India.

² Department of Computer Science and Engineering (CSE), Aditya University, Surampalem - Pin 533437, Andhra Pradesh, India.

* Corresponding Author; Email: charishmaaki@gmail.com

ORCID Details

Charishma Purna Aki: <https://orcid.org/0009-0005-1111-7428>

Sri Surya Lakshmi Sowjanya Saladi: <https://orcid.org/0009-0004-6323-2367>

Prabhu Sagar Telagamsetti: <https://orcid.org/0009-0006-2480-729X>

Harish E: <https://orcid.org/0009-0001-3392-9233>

International Journal of Science and Research Archive, 2026, 21(01), 288–295

Publication history: Received on 31 August 2026; revised on 08 October 2026; accepted on 10 October 2026

Article DOI: <https://doi.org/10.30574/ijrsra.2026.21.1.1874>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Pollution level evaluation needs an organized study on pollutants, weather patterns and environment as they affect measured levels of air quality. Recent research on machine learning for predicting air quality through measurement of pollutants, weather conditions, classification techniques and comparison models [1]-[4]. Studies on Visakhapatnam, India show that there is an application of various machine learning algorithms for predicting air quality based upon pollutants and weather conditions [21]. The paper is an article on comparing machine learning models to classify air quality and predict PM_{2.5} levels based upon the Air Quality and Population Assessment data set [19]. There are 5000 entries of data which includes nine variables such as temperature, humidity, PM_{2.5}, PM₁₀, no₂, so₂, CO, proximity to industrial areas and population density. There are four types of supervised machine learning models to be compared; namely K Nearest Neighbor algorithm, SVM model, linear regression model and gaussian naive bayes classifier. KNN, SVM, and GNB are used to classify air quality as Good, Moderate, Poor, and Hazardous whereas Linear Regression is used for predicting continuous PM_{2.5} concentration. Preprocessing is followed by standardization and then an 80:20 stratified train-test split. Accuracy, precision, recall, F1-score are used for classification whereas MAE, MSE, RMSE and R² are used to evaluate regression models. The results indicate good classification accuracy, whereas Linear Regression has an R² value of 0.9588. Five fold cross validation is used for checking model consistency [20].

Keywords: Air quality assessment; Air pollution; Machine learning; KNN; SVM; Gaussian Naive Bayes; Linear Regression; PM_{2.5}; Model evaluation; Comparative analysis.

1 INTRODUCTION

Pollution of air is a serious issue for environment studies as it varies due to weather, surroundings, industry level activities and people's needs. Machine learning techniques are being used more often to predict and analyse air quality data such as pollutants and weather conditions through comparison of models [1]-[4]. This paper is about a study that focuses on Visakhapatnam, Andhra Pradesh, where several machine learning algorithms were tested to predict air quality index based upon pollutants and weather data [21].

This paper is concerned with an experimental study of comparison between models instead of predicting from just one model. Same data set is used for studying four types of models based on their own assumptions. KNN, SVM, and Gaussian Naive Bayes are used for multiclass air quality classification whereas Linear Regression is used to predict continuous values of PM2.5 levels. Classification vs Regression differentiation remains constant during testing to interpret results according to each specific problem statement.

The project implements a reproducible workflow consisting of data inspection, preprocessing, feature scaling, stratified splitting, model training, testing, metric calculation, and cross-validation. This type of evaluation process is similar to other studies on data preprocessing, validation techniques, feature selection methods, model explanation tools, and generalization capabilities for air quality prediction models [22], [23].

2 LITERATURE

Machine learning is used for predicting current levels of air quality (AQI) and concentrations of pollutants from recent studies. Research studies compare traditional machine learning models to ensembles and deep learning techniques; results show differences due to data set selection, feature choice, preprocessing methods, and validation strategies [1]-[4], [22], [23]. There are many papers on classification techniques like SVM, Random Forest, and Naive Bayes to classify AQI and air pollution [5], [6], [9]-[12]. There is also a lot of literature that deals with predicting PM2.5 levels through linear regression models and other types of regression analysis [13]-[18]. According to this paper, we compare KNN, SVM, Linear Regression, and Gaussian Naive Bayes models for air quality classification and PM2.5 prediction based upon those results. Evaluation metrics for standard models are used to evaluate their effectiveness and reliability [20].

2.1 Existing system

Current methods for assessing air quality are based upon measured pollutants levels, predetermined limits of pollution concentration index (AQI), statistical modeling techniques such as regression analysis; machine learning algorithms like decision trees etc. Web based predictive models are available through published systems [2] and [5]. A lot of studies are done on certain algorithms or some types of sophisticated models from specific areas/time periods. These models have benefits due to certain goals; however they may vary on datasets, feature sets, targets definitions, data processing techniques as well as performance measurement criteria among others.

2.2 Disadvantages of Existing System

- No comparison among various ML models for a given set of data.
- Low precision because of dependence upon an individual model instead of comprehensive analysis.
- Lack of proper data cleaning results in poor performance models.
- Expensive price tag and poor range for conventional sensors' monitoring system.
- Lack of standardization for evaluating performance through precision, recall, f1 score and r2 score.
- There is no clear method to choose an appropriate model among various options available when making predictions or classifying data points.

2.3 Proposed System

This paper proposes an integrated machine learning model to assess air quality based on the Air Quality and Population Assessment dataset [19]. Preprocessing pipelines are used here to enhance data quality and prepare it for use by models during training. This paper divides its data set among four groups for testing KNN algorithm, SVM model, linear regression analysis technique and gaussian naive bayes classifier using same procedure or method of evaluation. KNN, SVM, and Gaussian Naive Bayes are used for classifying air quality into Good, Moderate, Poor, and Hazardous categories whereas Linear Regression is used to predict PM2.5 concentration as a continuous value. The 80/20 train/test ratio is used to evaluate models by applying feature normalization if necessary. The five fold cross validation method was used for checking model consistency. Classification model performance is measured by accuracy, precision, recall, F1 score and confusion matrix whereas linear regression is measured by mean absolute error, mean squared error, root mean square error and coefficient of determination (R2). Results are shown as tabular data and graphs that provide an overview to compare various models used by us for classifying air quality problems and predicting pollutants.

2.4 Advantages of Proposed System

- Enhanced predictive power by comparing several models to them.
- Classification model evaluation by means of regression analysis on same data set.
- Effective data cleaning techniques such as imputation of missing values and normalization are needed to obtain accurate outcomes.
- Standardized, multi-metric evaluation (Accuracy, Precision, Recall, F1-Score, R^2 Score) for reliable comparison.
- A clear method to choose an appropriate model among many options available for predicting/classifying data.
- A reliable and reusable method for wider use of environmental monitoring systems.

2.4.1 Block Diagram

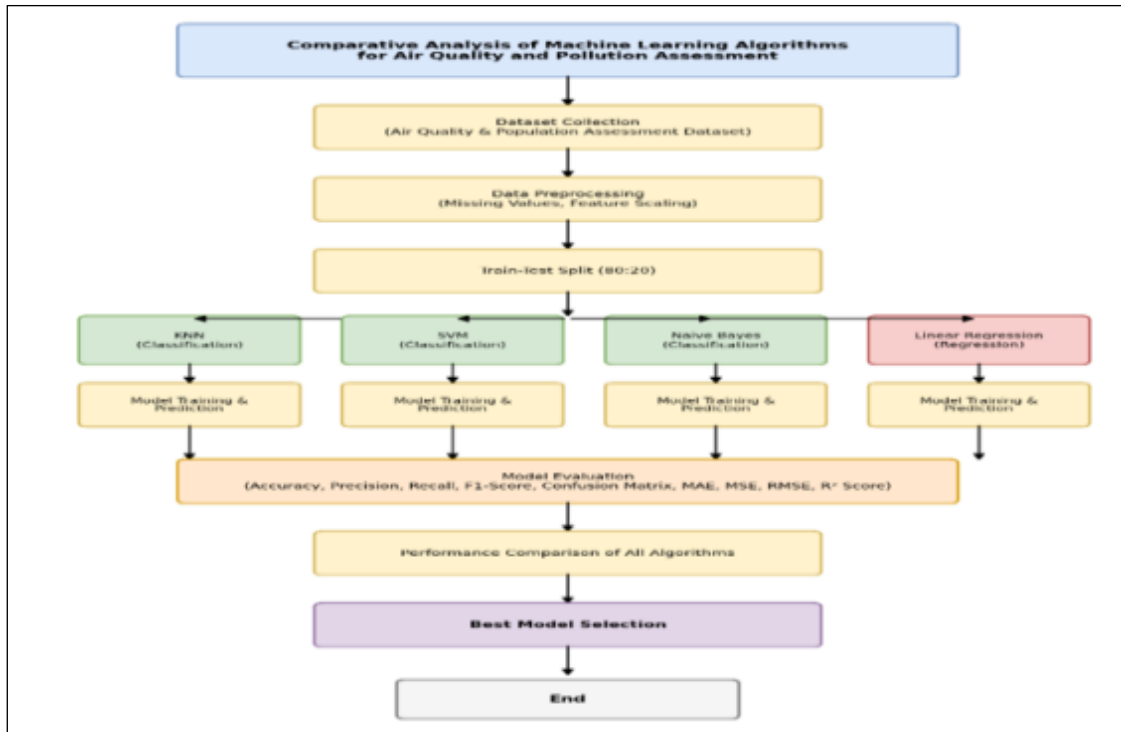


Figure.1: Block Diagram showing the ML pipeline and comparison of four algorithms within the system.

3 ABOUT DATASET

This study is based on an air quality and population assessment dataset from Kaggle [19]. There are 5000 entries to this data set which has ten columns. There are nine numeric inputs for this model while there is a single air quality target column. Temperature Humidity Particulate Matter Gases Proximity Population Density are given by this data set.

Air Quality has four levels such as good moderate poor hazardous. The targets for classifying these models have been labeled according to what is needed from them. Regression models treat PM2.5 as an univariate continuous variable. Before modeling data needs to be checked on missingness and duplication while also normalizing numeric attributes as needed from a machine-learning perspective.

Table 1: Dataset Attributes and Their Roles

Attribute	Type	Role
Temperature	Numerical	Input feature
Humidity	Numerical	Input feature
PM2.5	Numerical	Input feature / regression target
PM10	Numerical	Input feature
NO2	Numerical	Input feature

SO2	Numerical	Input feature
CO	Numerical	Input feature
Proximity to Industrial Areas	Numerical	Input feature
Population Density	Numerical	Input feature
Air Quality	Categorical	Classification target

4 METHODOLOGY

The proposed methodology is based on a typical machine-learning pipeline that includes data preprocessing, separation of data from the targets, scaling, splitting of training and testing sets, model training, model prediction and model evaluation. The same preprocessing and evaluation procedure is applied to four machine-learning models.

4.1 Machine Learning Models and Experimental Design

4.1.1 K-Nearest Neighbour (KNN)

The KNN algorithm assigns an item to a category by comparing it with other items that have similar characteristics as theirs. Standardization of features is used since k-nearest neighbours (KNN) relies upon distances between points. For classification purposes of air quality as good, moderate, poor or hazardous category.

4.1.2 Support Vector Machine (SVM)

The SVM classifies data points through maximizing margin separation between them. Four class of air quality classification system it is applied to. Performance Measurement Standardization is used to measure performance by means of accuracy, precision, recall, F1 score, and a confusion matrix.

4.1.3 Linear Regression

Linear regression model is applied for predicting continuous values of PM2.5. Performance evaluation of this model was done by calculating MAE, MSE, RMSE and R2 value.

4.1.4 Gaussian Naive Bayes (GNB)

The Gaussian Naive Bayes model is an example of probability-based classifier that follows Bayes' rule for prediction purposes. This model classifies air quality as a categorical variable where each feature is assumed to be independent of others.

4.1.5 Data Preprocessing and Evaluation Protocol

- Check for type of data structures, distributions to targets and range values.
- Look out for null entries or duplicates among them.
- Feature separation vs Target variable separation.
- Normalize numeric values by means of standard scaling if necessary.
- Divide it to an 80/20 split between training and test set by means of stratified sampling for classification purposes.
- Train and Test Model Separately.
- Classification evaluation by means of accuracy, precision, recall, F1 score and confusion matrix.
- Assess regression by mean absolute error, mean squared error, root mean square.
- Use a 5 fold validation for checking models' consistency.

4.2 Project Objectives

- Create an ML model to classify air quality and predict PM2.5 levels.
- Consistent pre-processing without any data leakage is required here.
- Try out k-Nearest Neighbors (KNN), Support Vector Machines, Linear Regression and Gaussian Naive Bay.
- Assess classifiers and regressors by means of suitable evaluation measures.
- Compare models by means of table graphs, chart plots, and cross validation.
- Study of various assumption about algorithms' impact on models' effectiveness.
- Create a scalable model for supporting more data sets and machine learning models.

5 OUTPUT

Models developed for air quality classification and PM_{2.5} prediction are tested here. Accuracy, precision, recall, and F1 score were used to evaluate KNN, SVM, and Gaussian Naive Bayes models. For continuous PM_{2.5} prediction, Linear Regression was evaluated using MAE, MSE, RMSE, and R² score. All three models have an accuracy rate greater than 90% according to these classifications. Linear regression model had an R squared value of 0.9588 which indicates a good correlation between actual and predicted PM_{2.5} levels. The five fold cross validation is done for checking accuracy and precision of this model.

Table 2: Classification Model Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
K-Nearest Neighbors (KNN)	92.80	92.79	92.80	92.68
Support Vector Machine (SVM)	94.20	94.19	94.20	94.18
Gaussian Naive Bayes (GNB)	91.00	91.11	91.00	91.03

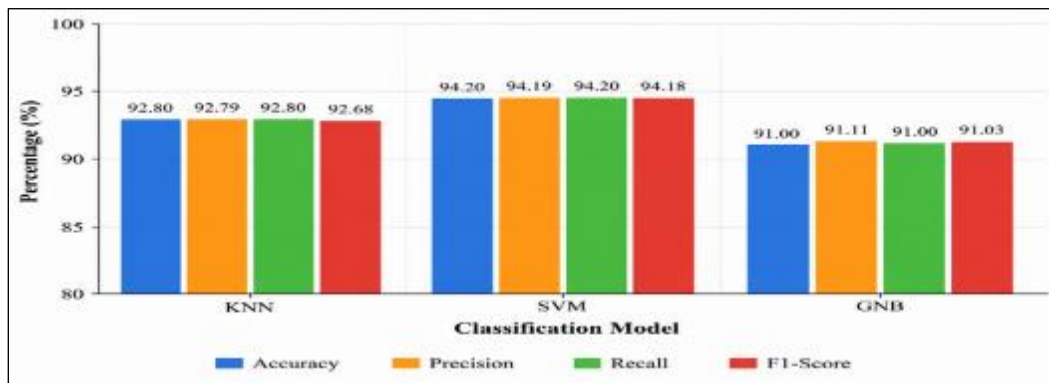


Figure. 2: Classification Model Performance Comparison

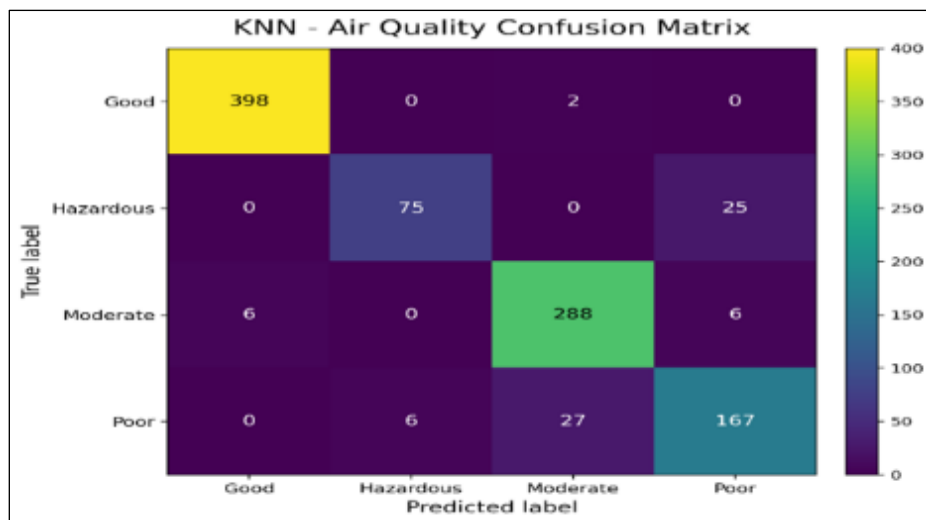


Figure. 3: Confusion Matrix for KNN

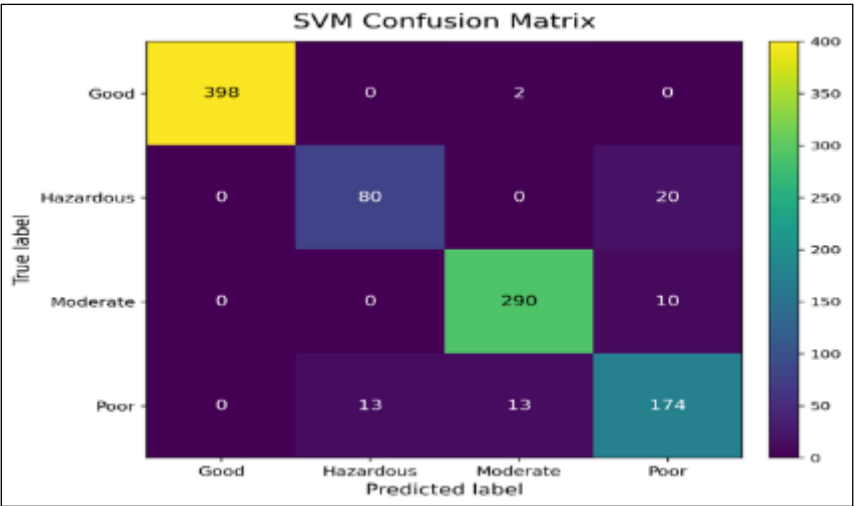


Figure. 4: Confusion Matrix for SVM

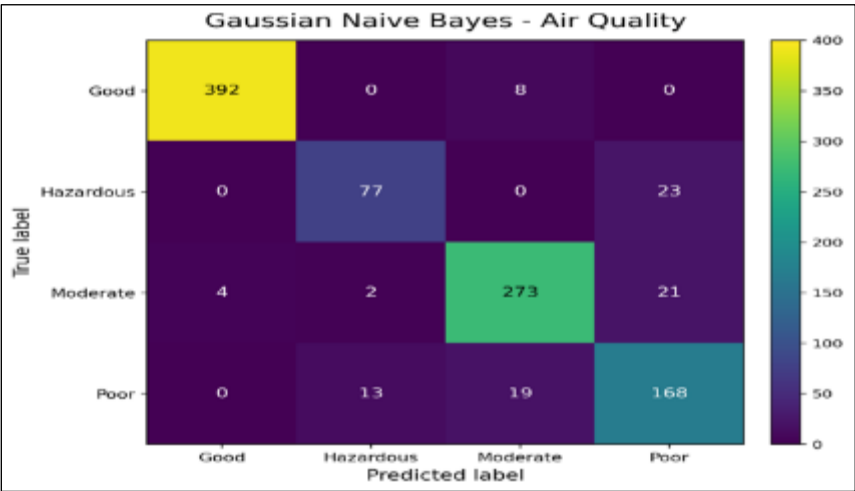


Figure. 5: Confusion Matrix for Gaussian Naive Bayes (GNB)

Table 3: Performance Metrics of Linear Regression for PM2.5 Prediction

Metric	Value	Description
Mean Absolute Error (MAE)	3.3086	Average absolute difference between actual and predicted PM2.5 values.
Mean Squared Error (MSE)	20.8521	Average squared difference between actual and predicted PM2.5 values.
Root Mean Squared Error (RMSE)	4.5664	Square root of MSE, indicating the average prediction error in the same units as PM2.5.
R ² Score (Coefficient of Determination)	0.9588	Proportion of variance in PM2.5 values explained by the model.

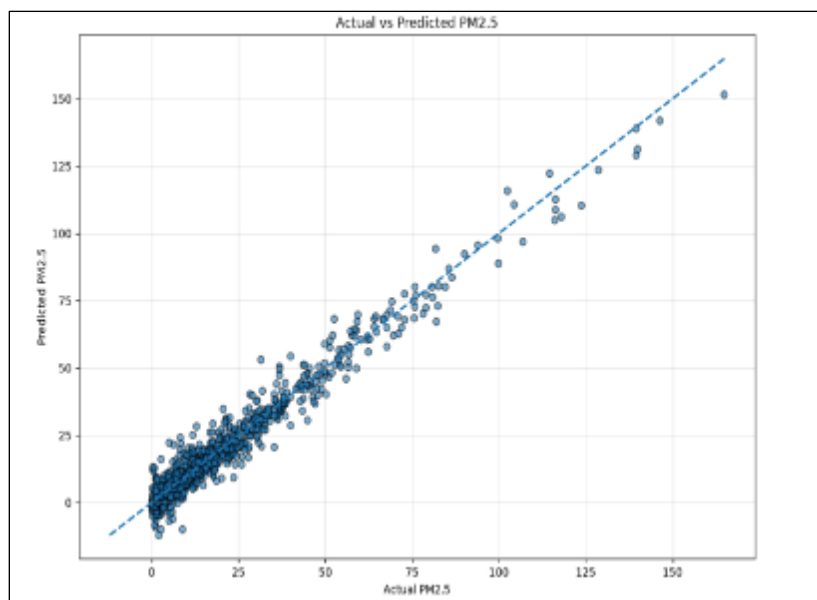


Figure. 6: Actual versus Predicted PM2.5 Concentrations Using Linear Regression

6 COMPARATIVE DISCUSSION

Model evaluation was done through suitable measures of accuracy and precision to classify data points as well as predict continuous values from them. KNN, SVM, and Gaussian Naive Bayes had an accuracy rate of 92.80%, 94.20% and 91.00%. The linear regression model had an R^2 value of 0.9588 for predicting PM2.5, while its mean cross-validation R^2 was about 0.967. The findings show that these models are consistently good at predicting what they were designed to do.

7 CONCLUSION

The paper is an analysis of this model that compares models to classify air quality and predict PM2.5 levels from the Air Quality and Population Assessment data set. Various types of supervised machine-learning models are used to evaluate them based on accuracy rate or precision score for both classification problems as well as regression problem. This paper shows how to use machine learning models on environmental information such as air quality and pollution statistics. This proposal may also be improved through larger real-time data sets, extra environmental variables, and sophisticated machine learning / deep learning methods for better air quality monitoring.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The authors thank Aditya University for the facilities and support provided during this research.

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding this manuscript

REFERENCES

- [1] Air Quality Prediction System Using Machine Learning Models. Water, Air, & Soil Pollution, 2024. <https://link.springer.com/article/10.1007/s11270-024-07390-0>
- [2] AirNet: Predictive Machine Learning Model for Air Quality Forecasting Using Web Interface. Environmental Systems Research, 2024. <https://link.springer.com/article/10.1186/s40068-024-00378-z>
- [3] A Comparative and Systematic Study of Machine Learning (ML) Approaches for Particulate Matter (PM) Prediction. Archives of Computational Methods in Engineering, 2023. <https://link.springer.com/article/10.1007/s11831-023-09994-x>

- [4] Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction. *Water, Air, & Soil Pollution*, 2025. <https://link.springer.com/article/10.1007/s11270-025-08122-8>
- [5] Intelligent Real-Time Air Quality Index Classification. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 16, No. 3. https://thesai.org/Downloads/Volume16No3/Paper_31-Intelligent_Real_Time_Air_Quality_Index_Classification.pdf
- [6] Evaluating the Performance of SVM and Random Forest in Air Quality Monitoring and Prediction. *Springer Nature Link*. https://link.springer.com/chapter/10.1007/978-3-032-06253-6_21
- [7] Air Quality Prediction Model Based on mRMR-RF Feature Selection and ISSA-LSTM. *PMC*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10406845/>
- [8] Air Quality Prediction Models Based on Meteorological Factors and Real-Time Data of Industrial Waste Gas. *Scientific Reports*, DOI: 10.1038/s41598-022-13579-2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9166716/>
- [9] Apriani, N.F., et al. Classification and Forecasting Air Pollution Using Naive Bayes and Prophet: A Use Case of Air Quality Index in Jakarta. *IEEE Xplore*, 2024. <https://ieeexplore.ieee.org/document/10732634/>
- [10] Kumari, R., et al. Predicting the Air Quality Using Machine Learning Algorithms: A Comparative Study. *Springer Nature Link*. https://link.springer.com/chapter/10.1007/978-981-97-1320-2_12
- [11] Efficient Weighted Naive Bayes Classifiers to Predict Air Quality Index. *Earth Science Informatics*, 15, 541–552, 2022. <https://link.springer.com/article/10.1007/s12145-021-00755-7>
- [12] Probabilistic Classification of the Severity Classes of Unhealthy Air Pollution Events. *PubMed*, 2024. <https://pubmed.ncbi.nlm.nih.gov/38717514>
- [13] Rybarczyk, Y., Zalakeviciute, R. Regression Models to Predict Air Pollution from Affordable Data Collections. *IntechOpen*, DOI: 10.5772/intechopen.71848, 2017. <https://www.intechopen.com/chapters/57822>
- [14] Bai, L., Li, H., Zeng, B., Huang, X. Design of a Combined System Based on Multi-Objective Optimization for Fine Particulate Matter (PM_{2.5}) Prediction. *Frontiers in Environmental Science*, 2022. <https://www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2022.833374/pdf>
- [15] Trenchevski, A., Kalendar, M., Gjoreski, H., Efnusheva, D. Prediction of Air Pollution Concentration Using Weather Data and Regression Models. https://repo.bibliothek.uni-halle.de/bitstream/1981185920/32933/1/2_2_Trenchevski.pdf
- [16] Zhou, Z. Predicting PM_{2.5} in Urban and Suburban Beijing: A Comparative Study of Random Forest and Linear Regression Models. *IAMPA 2025 Proceedings*, DOI: 10.5220/0013825300004708. <https://scitepress.org/publishedPapers/2025/138253/pdf/index.html>
- [17] PM_{2.5} Concentration Prediction Using Linear Regression, Decision Tree, RF, KNN, and Lasso Regression. *Springer Nature Link*, 2024. <https://link.springer.com/10.1007/s44288-024-00043-z>
- [18] Hu, X., Shi, J., Liang, R. Study on PM_{2.5} Concentration Prediction Model Based on Bagging-Gradient Boosting Decision Tree. *IOS Press*. <https://ebooks.iospress.nl/pdf/doi/10.3233/FAIA231342>
- [19] Air Quality and Population Assessment Dataset. *Kaggle*. <https://www.kaggle.com/datasets/mujtabamatin/air-quality-and-pollution-assessment>
- [20] Pedregosa, F., et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830, 2011.
- [21] Air Quality Prediction by Machine Learning Models: A Predictive Study on the Indian Coastal City of Visakhapatnam. *Chemosphere*, 2023, 338, 139518. <https://doi.org/10.1016/j.chemosphere.2023.139518>
- [22] A Review of Machine Learning for Modeling Air Quality: Overlooked but Important Issues. *Atmospheric Environment*, 2024. <https://www.sciencedirect.com/science/article/pii/S0169809524000437>
- [23] Machine Learning for Air Quality Prediction and Data Analysis: Review on Recent Advancements, Challenges, and Outlooks. *Science of the Total Environment*, 2025, 1002, 180593. <https://doi.org/10.1016/j.scitotenv.2025.180593>