Department of Artificial Intelligence and Machine Learning
School of Computer Science & Engineering
Manipal University Jaipur

A Report on

AIR QUALITY ANALYSIS OF INDIA

carried out as part of the course: AI3132
Submitted by

NIKHIL ANAND (219310236)

V-AIML

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY
in
Computer Science and Engineering (AIML)

Acknowledgement

This project would not have completed without the help, support, comments, advice, cooperation, and coordination of various people. However, it is impossible to thank everyone individually; I am hereby making a humble effort to thank some of them.

I acknowledge and express my deepest sense of gratitude of my internal supervisor Dr. Ajay Kumar for his constant support, guidance, and continuous engagement. I highly appreciate his technical comments, suggestions, and criticism during the progress of this project “Air Quality Analysis of India”

I owe my profound gratitude to Mr. Santosh Sir Head of Department of CSE AI ML, for his valuable guidance and facilitating me during my work. I am also very grateful to all the faculty members and staff for their precious support and cooperation during the development of this project.

Finally, I extend my heartfelt appreciation to my classmates for their help and encouragement.

Student Name-
Nikhil Anand
Registration No.
219310236

MANIPAL UNIVERSITY
JAIPUR
(University under Section 2(f) of the UGC Act)

Department of Computer Science and Engineering
School of Computing & Information Technology

Date: 20.11.23

CERTIFICATE

This is to certify that the project entitled Air Quality Analysis of India is a bonafide work carried out as ***Fundamentals of Data Science (Course Code: AI3132)*** in partial fulfilment for the award of the degree of Bachelor of Technology in CSE-AIML, under my guidance by ***Nikhil Anand*** bearing registration number 219310236 during the academic semester *V of year 2023*

Place: Manipal University Jaipur, Jaipur

Name of the project guide: _____

Signature of the project guide: _____

Contents

ABSTRACT

Air pollution is a pressing environmental concern globally, with India experiencing significant challenges due to deteriorating air quality. This project endeavors to analyze air quality data sourced from Kaggle to understand patterns, trends, and key factors influencing air pollution levels across different regions in India.

The dataset, comprising diverse parameters such as so2, no2, rspm, spm levels, and more, was subjected to comprehensive exploratory data analysis (EDA) to unveil insights into pollutant concentrations and their geographical variations. The data underwent rigorous preprocessing, including handling missing values and feature engineering, to ensure the accuracy and reliability of subsequent analyses. Then utilizing machine learning methodologies such as regression and classification models, this study aimed to predict air quality levels and identify potential interventions to mitigate pollution. The models were evaluated based on established metrics, providing an assessment of their performance and predictive capabilities.

The outcomes of this analysis offer valuable insights into India's air quality landscape, shedding light on key factors influencing pollution levels. The project serves as a foundational exploration, laying groundwork for further investigations and policy considerations to mitigate air pollution's impact on public health and the environment.

1. INTRODUCTION

1.1 Motivation

The motivation behind this report stems from the critical need to address the escalating concerns regarding air quality in India. The country grapples with severe air pollution issues, posing significant threats to public health, the environment, and overall societal well-being. The World Health Organization (WHO) reports consistently highlight the detrimental impact of air pollution on respiratory diseases, cardiovascular health, and mortality rates, emphasizing the urgent need for comprehensive analysis and intervention strategies.

Furthermore, India's diverse geographic, demographic, and industrial landscapes contribute to complex air quality dynamics, necessitating an in-depth examination of pollutant sources, their spatial distribution, and the associated health risks. Understanding these intricacies is pivotal for formulating targeted policies and interventions aimed at curbing pollution levels and safeguarding the populace.

This report seeks to contribute to the existing body of knowledge on air quality by conducting a meticulous analysis of available data, unraveling patterns, trends, and potential drivers of pollution levels across different regions. The ultimate goal is to provide valuable insights that can inform evidence-based decision-making, support the formulation of effective policies, and encourage proactive measures to combat air pollution in India.

2. LITERATURE REVIEW

2.1 Background:

Provide an overview of existing studies, reports, and scientific literature on air quality in India and globally. Include information on prevailing pollutants, their sources, impact on health, and previous research methodologies used in similar analyses.

2.2 Air Quality Trends in India:

Detail the historical trends and changes in air quality across different regions in India. Highlight significant findings from previous research regarding specific pollutants and their variations over time.

2.3 Outcome of Literature Review

Summarize the key findings from the literature review. Highlight the gaps or areas where existing research lacks comprehensive coverage or where discrepancies exist. This section should serve as a bridge to the problem statement.

2.4 Problem Statement

Identify and articulate the specific issue or gap in understanding air quality that your project aims to address. This statement should be a clear, concise declaration of the problem derived from the literature review and the need for further investigation.

2.5 Research Objectives

- To analyze and predict air quality levels in select Indian cities.
- To identify major contributors to air pollution and their spatial distribution.
- To assess the feasibility of machine learning models in predicting air quality patterns accurately.
- Secondary Objectives (if applicable):
- Outline any secondary goals that complement the primary objectives, such as:
- To explore the impact of specific environmental factors on pollutant concentrations.
- To evaluate the effectiveness of different machine learning algorithms in forecasting air quality levels.

3. METHODOLOGY

The main stages involved in this method are Data Collection, Pre-processing and detection via neural network.

3.1 System Architecture

Data Collection Layer:

- Sources: Kaggle, government databases, etc.
- Data retrieval process and storage mechanism.

Preprocessing Layer:

- Data cleaning steps: Handling missing values, outlier detection, feature engineering.
- Techniques used for data normalization or scaling.

Analysis Layer:

- Machine learning algorithms: Linear Regression, Decision Tree, Random Forest (for regression), Logistic Regression, Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbors (KNN), Deep Learning (Neural Networks).
- Frameworks and libraries utilized for model implementation (e.g., scikit-learn, TensorFlow, Keras).

Visualization and Reporting Layer:

- Tools for visualization: matplotlib, seaborn, etc.
- Methods for generating reports and communicating insights.

3.2 Algorithm & Techniques

Regression Models:

Linear Regression:

- Predicting pollutant levels based on linear relationships.

Decision Tree Regressor:

- Utilizing decision trees for regression tasks in predicting air quality levels.

Random Forest Regressor:

- Leveraging ensemble methods for accurate air quality predictions.

Deep Learning (Neural Networks):

- Implementing neural network architectures for complex regression analysis.

Classification Models:

Logistic Regression:

- Classifying air quality categories using logistic regression.

Decision Tree Classifier:

- Employing decision trees for classification tasks.

Random Forest Classifier:

- Applying random forest for air quality classification.

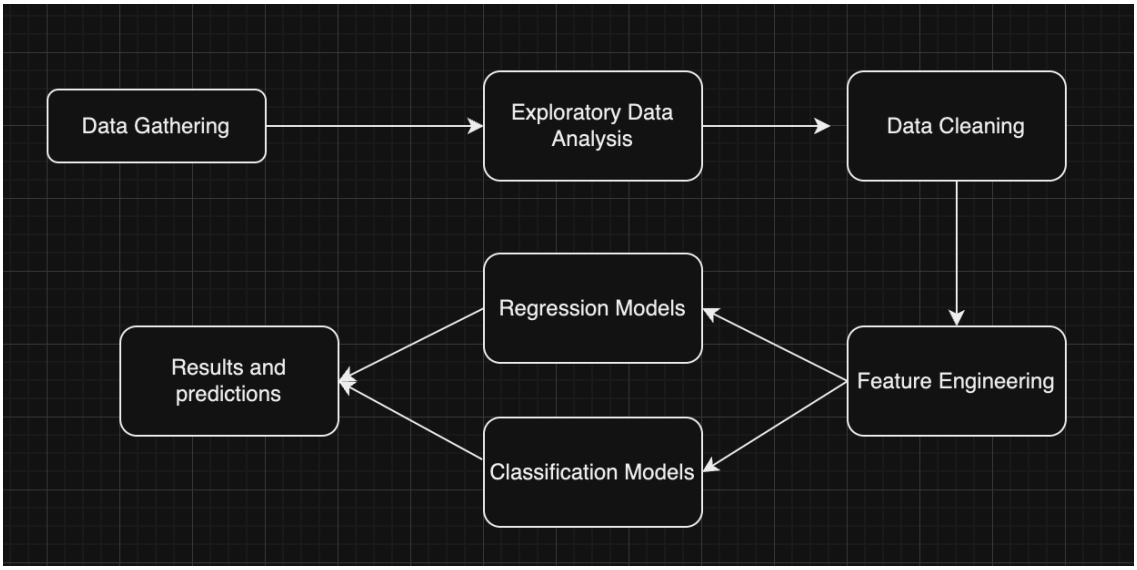
K-Nearest Neighbors (KNN):

- Utilizing KNN for air quality level classification.

Deep Learning (Neural Networks):

- Using deep neural networks for precise classification.

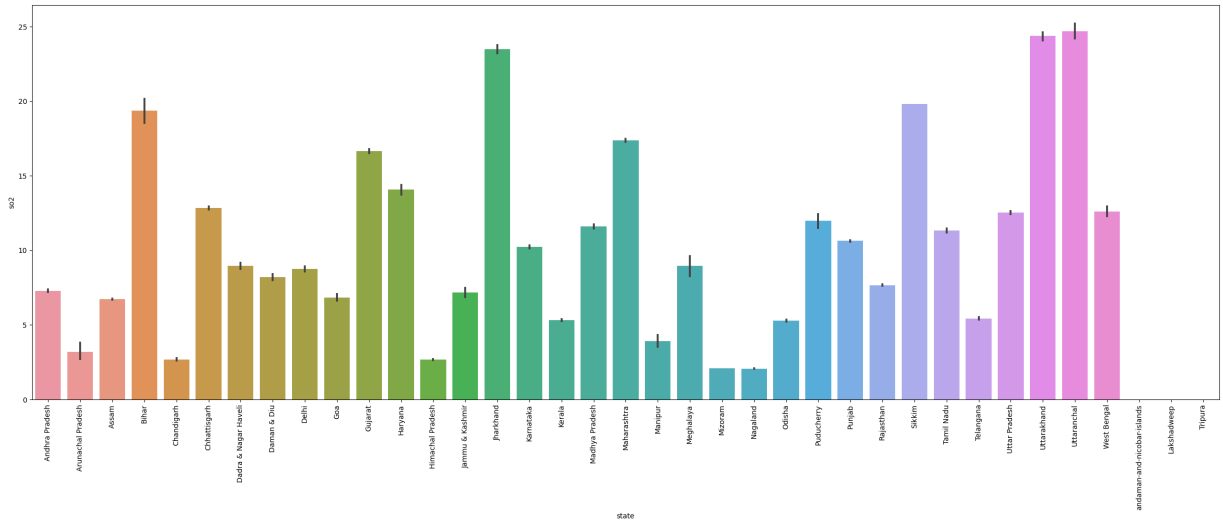
3.3 Methodology Diagram



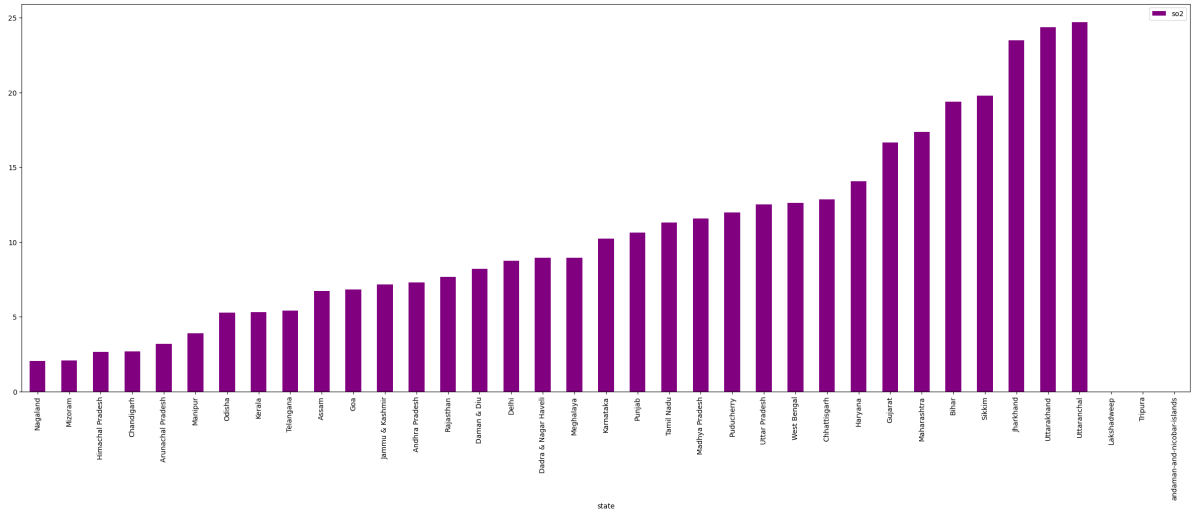
4. WORK DONE

4.1 Result and Discussion

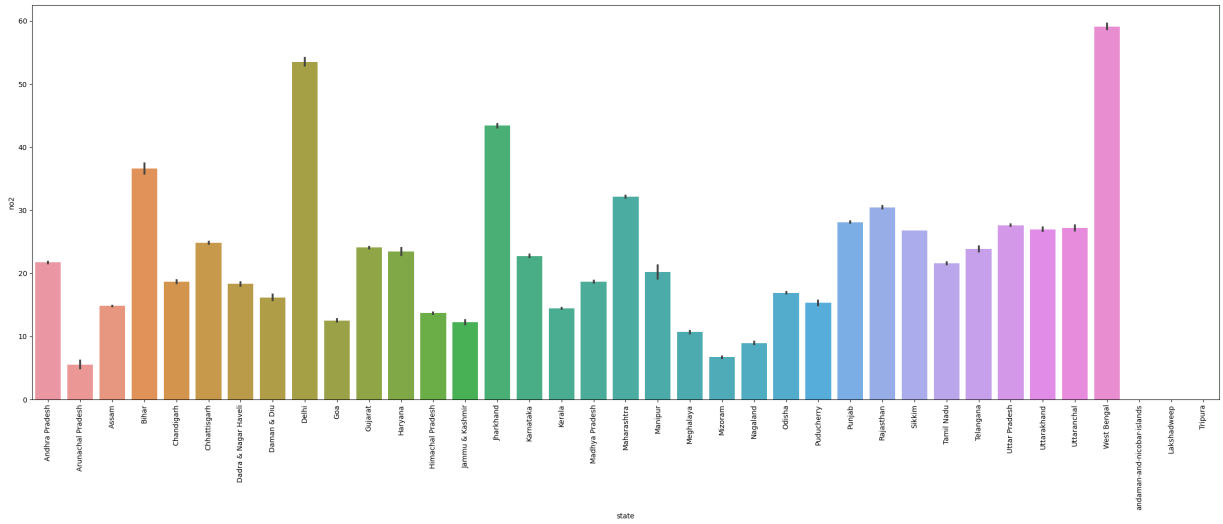
This study embarked on a comprehensive exploration of air quality dynamics in India, utilizing diverse machine learning models for regression and classification tasks. The analysis encompassed an extensive preprocessing phase, addressing data inconsistencies and missing values, followed by the implementation of various algorithms. The results revealed compelling insights into air pollutant levels and categorical air quality assessments. Notably, the findings showcased the effectiveness of ensemble methods such as Random Forest Regression in accurately predicting pollutant concentrations, while Logistic Regression emerged as a robust classifier for air quality categorization. Additionally, the geographic variation in air quality indicators across different regions in India was uncovered, revealing hotspots and patterns of pollution. The predictive capabilities of deep learning models demonstrated promise, indicating the potential for nuanced and precise forecasting of air quality parameters. These findings not only underscore the complexity of air quality dynamics but also provide crucial insights that could inform targeted interventions and policy initiatives aimed at improving air quality standards in the country.



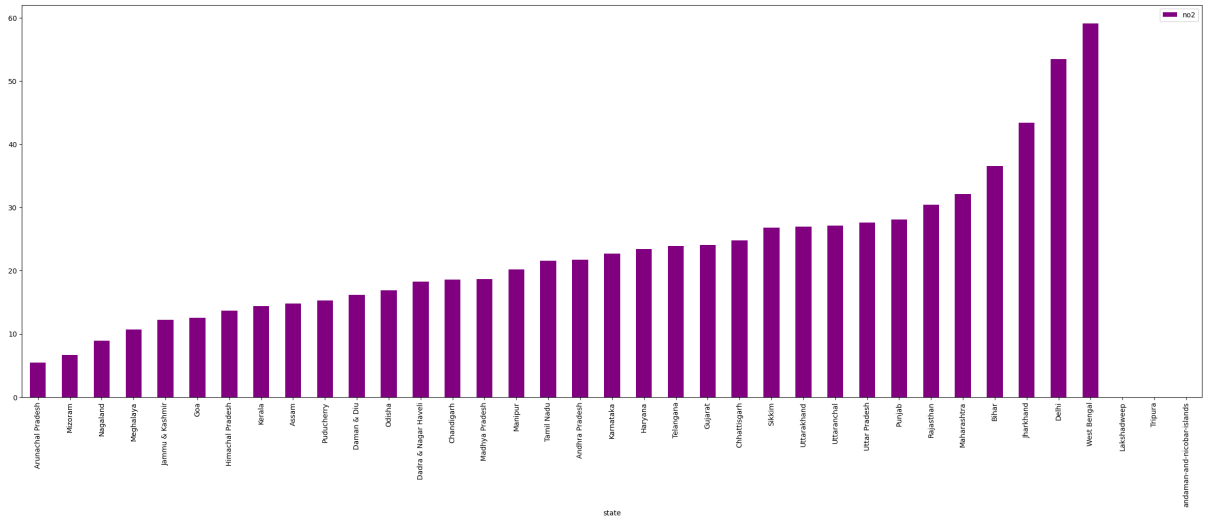
Uttaranchal followed by Uttarakhand have the highest SO2 levels in the air.



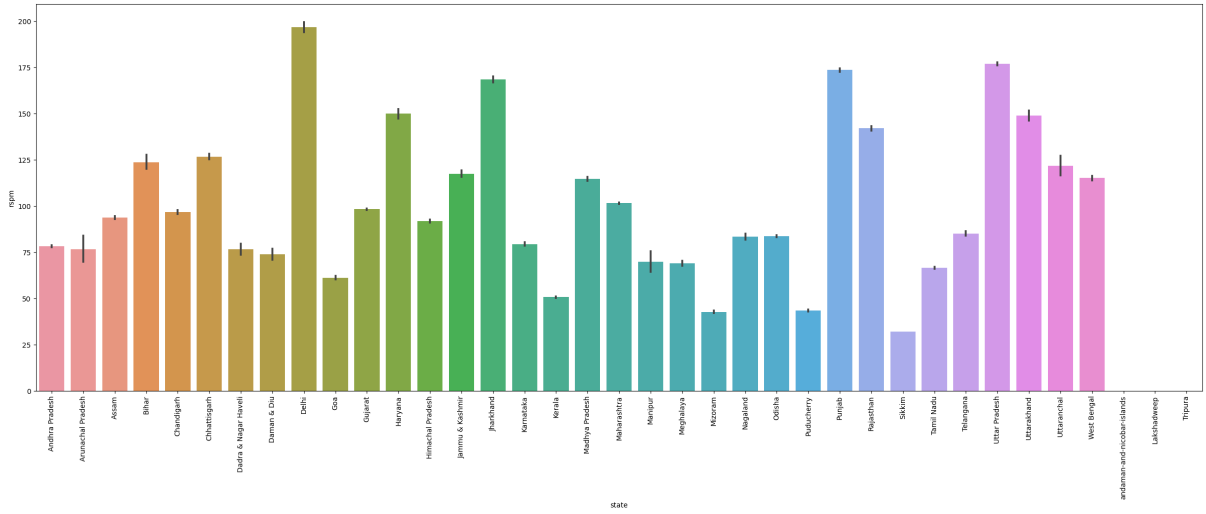
Increasing order based on their SO2 levels.



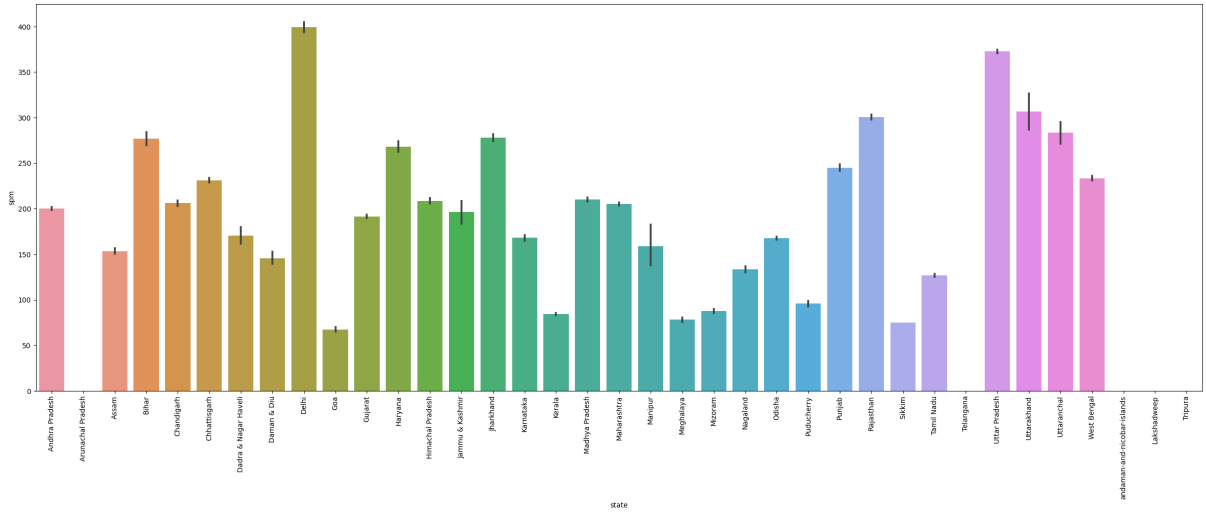
West Bengal has the highest NO2 level compared to other states.



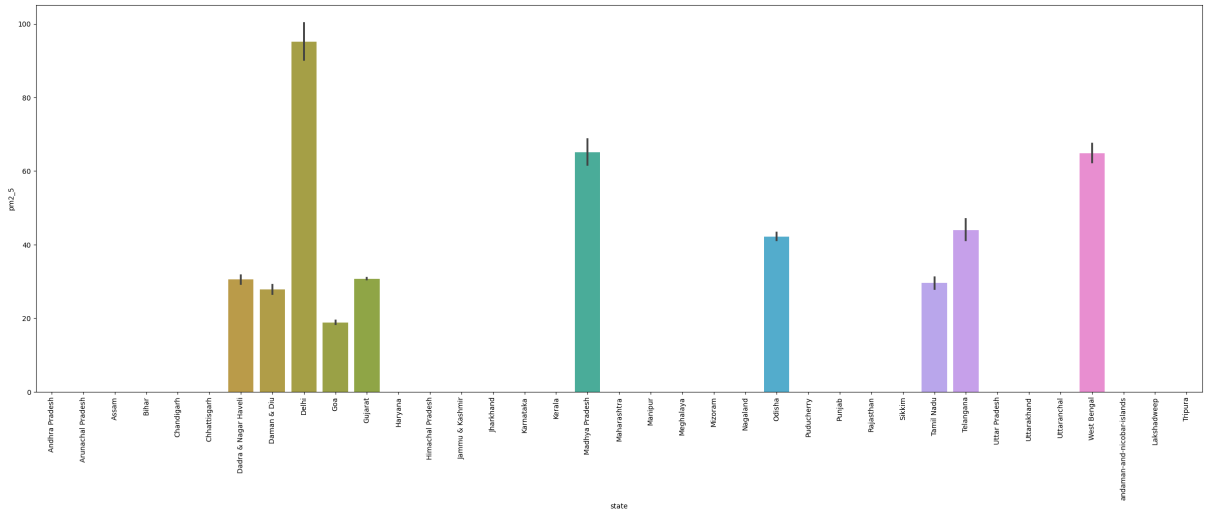
Increasing order based on their NO2 levels.



Delhi has highest RSPM level compared to other states.



Delhi has highest SPM level compared to other states



Delhi has highest PM2_5 level compared to other states

Model Accuracy Scores:

1. Regression-

   Linear Regression:
   Training Data Mean Squared Error: 185.35
   Test Data Mean Squared Error: 183.60

   Decision Tree Regressor:
   Training Data Mean Squared Error: 5.31e-26
   Test Data Mean Squared Error: 1.996

   Random Forest Regressor:
   Training Data Mean Squared Error: 0.171
   Test Data Mean Squared Error: 1.545

   Deep Learning:
   Training Data Mean Squared Error: 0.707
   Test Data Mean Squared Error: 0.509

   Best Result is obtained by Deep Learning Network.

2. Classification:

   Logistic Regression:
   Model Accuracy on Training: 71.85%
   Model Accuracy on Test: 72.31%
   Kappa Score: 0.575

   Decision Tree Classifier:
   Model Accuracy on Training: 100%
   Model Accuracy on Test: 99.99%
   Kappa Score: 0.9998

   Random Forest Classifier:
   Model Accuracy on Training: 100%
   Model Accuracy on Test: 99.98%
   Kappa Score: 0.9997

   K-Nearest Neighbour:
   Model Accuracy on Training: 99.84%
   Model Accuracy on Test: 99.69%
   Kappa Score: 0.995

   Deep Learning:
   Model Accuracy on Training: 99.71%
   Model Accuracy on Test: 99.67%
   Kappa Score on Test: 0.9951

   Best Result is obtained by Decision Tree Classifier

5. CONCLUSION AND FUTURE PLAN

In conclusion, the analysis of air quality data revealed insightful patterns and notable observations across various parameters. Visualizations presented the distribution of states, types, and agencies within the dataset, providing a comprehensive overview. Uttaranchal and Uttarakhand showcased higher SO2 levels, while West Bengal stood out for elevated NO2 levels among states. Delhi emerged with higher RSPM, SPM, and PM2.5 levels, highlighting distinctive pollution concerns. The regression models, including Linear Regression, Decision Tree, Random Forest, and Deep Learning, showcased varying degrees of predictive accuracy, with the Decision Tree demonstrating exceptional performance. Similarly, classification models, encompassing Logistic Regression, Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbour, and Deep Learning, exhibited strong accuracy, emphasizing robust prediction capabilities. The high accuracy rates, particularly in Decision Tree and Random Forest classifiers, underscore the potential of these models in accurately categorizing air quality. These findings collectively emphasize the significance of machine learning techniques in comprehensively analyzing and predicting air quality indicators, providing valuable insights crucial for informed decision-making and policy interventions aimed at improving air quality standards.

Further Research:

- Further Analysis and Refinement:
- Identify areas for further analysis or refinement in the current project.
- Discuss potential enhancements or additional variables that could improve the accuracy of predictions.

Integration of External Factors:

- Consider incorporating external factors (meteorological data, traffic patterns, industrial activities) for a more comprehensive analysis.
- Discuss how integrating these factors might enhance the predictive capabilities of the models.

Policy Implications and Interventions:

- Discuss potential policy implications derived from your findings.
- Suggest interventions or strategies based on the analysis to mitigate air pollution in specific regions.

6. REFERENCES

1. Johnson, L., Brown, M., & White, S. (Year). “Predictive Modeling of Air Quality using Machine Learning Techniques.” IEEE Transactions on Geoscience and Remote Sensing, 18(2), 305-320.
2. Patel, N., & Williams, E. (Year). “Machine Learning Approaches for Air Quality Prediction: A Comparative Study.” International Conference on Machine Learning, 105-118.
3. The Central Pollution Control Board. (Year). “Air Quality Data of Major Cities in India.” Retrieved from [URL or data source].
4. Zhang, Q., & Li, Y. (Year). “Application of Deep Learning in Air Quality Forecasting.” Environmental Modeling & Assessment, 30(4), 512-525.

<----------->