

**SCALABLE APPROACH FOR DETECTING AIR QUALITY
INFERENCE USING ENSEMBLE REGRESSORS**

A PROJECT REPORT

Submitted by

**SINGA SELVAMANI S
VIMAL ADITYA RAJ
SAMARENDRA T**

**RA2011003020212
RA2011003020216
RA2011003020206**

Under the guidance of

MRS. P. PREETHY JEMIMA

**(Assistant Professor, Department of Computer Science
Engineering)**

In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY



SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

RAMAPURAM, CHENNAI-600089

MAY 2024

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University U/S 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this project report titled “**SCALABLE APPROACH FOR DETECTING AIR QUALITY INFERENCE USING ENSEMBLE REGRESSORS**” is the bonafide work of **SINGA SELVAMANI. S [RA2011003020212]**, **VIMAL ADITYA RAJ [RA2011003020216]**, **SAMARENDRA.T [RA2011003020206]** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an occasion on this or any other candidate.

SIGNATURE

Ms.Preethy Jemima. P, M.E, (Ph.D),,
Assistant Professor,
Computer Science and Engineering,
SRM Institute of Science and Technology,
Ramapuram, Chennai.

SIGNATURE

Dr. K. RAJA, M.E., Ph.D.,
Professor and Head of Department,
Computer Science and Engineering,
SRM Institute of Science and Technology,
Ramapuram, Chennai.

Submitted for the project viva-voce held on _____ at SRM Institute of Science and Technology, Ramapuram, Chennai -600089.

INTERNAL EXAMINER 1

INTERNAL EXAMINER 2

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
RAMAPURAM, CHENNAI - 89**

DECLARATION

We hereby declare that the entire work contained in this project report titled “**SCALABLE APPROACH FOR DETECTING AIR QUALITY INFERENCE USING ENSEMBLE REGRESSORS**” has been carried out by **SAMARENDRA.T [RA2011003020206]**, **SINGA SELVAMANI. S [RA2011003020212]**, **VIMAL ADITYA RAJ [RA2011003020216]** at SRM Institute of Science and Technology, Ramapuram Campus, Chennai- 600089, under the guidance of **Ms. Preethy Jemima. P, Assistant Professor**, Department of Computer Science and Engineering.

Place: Chennai

Date:

SAMARENDRA. T

SINGA SELVAMANI. S

VIMAL ADITYA RAJ



SRM Institute of Science & Technology

Department of Computer Science and Engineering

Own Work Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessments

Degree/ Course : B.Tech Computer Science and Engineering
Student Name : Samarendra. T, Singa Selvamani. S, Vimal Aditya Raj
Registration Number : RA2011003020206, RA2011003020212, RA2011003020216
Title of Work : SCALABLE APPROACH FOR DETECTING AIR QUALITY INFERENCE USING ENSEMBLE REGRESSORS

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly references / listed all sources as appropriate.
- Referenced and put in inverted commas all quoted text (from books, web, etc.)
- Given the sources of all pictures, data etc. that are not my own.
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present.
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website.

I understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

DECLARATION:

I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.

If you are working in a group, please write your registration numbers and sign with the date for every student in group.

RA2011003020206

RA2011003020212

RA2011003020216

ACKNOWLEDGEMENT

We place on record our deep sense of gratitude to our lionized Chairman **Dr. R. SHIVAKUMAR** for providing us with the requisite infrastructure throughout the course.

We take the opportunity to extend our hearty and sincere thanks to our Dean, **Dr. M. MURALI KRISHNA, B.E, M.Tech, PhD., MISTE, FIE, C.Engg.**, for maneuvering us into accomplishing the project.

We take the privilege to extend our hearty and sincere guidance to the Professor and Head of the Department, **Dr. K. RAJA, M.E., PhD.**, for his suggestions, support and encouragement towards the completion of the project with perfection.

We express our hearty and sincere thanks to our Project Coordinator **Dr. S. SATHYA PRIYA, M.E., PhD., Associate Professor** for her fortification. We express our hearty and sincere thanks to our guide **Ms.Preethy Jemima. P, M.E, (Ph.D).**, **Assistant Professor** Computer Science and Engineering Department for his sustained encouragement, consecutive criticism, and constant guidance throughout this project work.

Our thanks to the teaching and non-teaching staff of the Computer Science and Engineering Department of SRM Institute of Science and Technology, Ramapuram Campus, who provided the necessary resources for our project.

ABSTRACT

The influence of machine learning technologies is rapidly increasing and penetrating almost in every field, and air pollution prediction is not being excluded from those fields. The revision of studies related to air pollution prediction using machine learning algorithms based on sensor data in the context of smart cities is covered. Using the most popular databases and executing the corresponding filtration, the most relevant papers were selected. After thoroughly reviewing those papers, the main features were extracted, which served as a base to link and compare them to each other. Feature selection is considered one of the essential steps in data pre-processing. However, all of the previous studies on predicting concentration have been limited to statistical method feature selection, and none of these studies used machine-learning approaches. The main objective of this study was to explore the impact of several input variables in training different air quality indexes. Data normalization was accomplished using the Min-Max normalization technique, along with correlation analysis for selecting highly correlated variables. Next, the important features from the highly correlated variables were selected by implementing an optimization algorithm. To validate the results, several measures were calculated, including the correlation coefficient and the mean absolute error.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	LIST OF FIGURES	x
	LIST OF ABBREVIATIONS	xi
1.	INTRODUCTION	1
	1.1 Introduction	1
	1.2 Problem Statement	2
	1.3 Objective of the Project	2
	1.4 Project Domain	2
	1.5 Scope of the Project	2
	1.6 Methodology	3
2.	LITERATURE REVIEW	4
3.	PROJECT DESCRIPTION	8
	3.1 Existing System	8
	3.2 Proposed System	10
	3.2.1 Advantages	11
	3.3 Feasibility Study	11
	3.3.1 Economic Feasibility	12
	3.3.2 Technical Feasibility	12
	3.3.3 Social Feasibility	12
	3.4 System Specifications	13
	3.4.1 Hardware Specifications	13
	3.4.2 Software Specification	13

4.	MODULE DESCRIPTION	14
	4.1 General Architecture	14
	4.2 Design Phase	16
	4.2.1 Data Flow Diagram	16
	4.2.2 UML Diagram	16
	4.2.3 Sequence Diagram	17
	4.2.4 Use Case Diagram	18
	4.3 Module Description	19
	4.3.1 Data Processing	19
	4.3.2 Correlation Analysis	20
	4.3.3 Random Forest Regressor	20
	4.3.4 Support Vector Regression	21
	4.3.5 Performance Measure	21
5.	IMPLEMENTATION	23
	5.1 Input	23
	5.2 Output	23
	5.3 Testing	25
	5.3.1 Types Of Testing	26
	5.3.2 Verify Data Loading and Handling Missing Values	26
	5.3.3 Check Random Forest Model Training with Different Estimators	27
	5.3.4 Compare Performance Metrics For Random Forest and SVR	28
	5.4 Testing Strategy	30

6.	RESULTS AND DISCUSSION	31
	6.1 Efficiency of the Proposed System	31
	6.2 Existing VS Proposed System	31
7.	CONCLUSION	32
	7.1 Conclusion	32
	7.2 Future Enhancements	32
8.	SOURCE CODE	33
	8.1 Source Code	33
	REFERENCE	37
	APPENDICES	42
	A. PLAGIARISM REPORT	42
	B. PROOF OF PUBLICATION/PATENT FILED/ CONFERENCE CERTIFICATE	43

LIST OF FIGURES

FIGURE	TITLE	PAGE
4.1	Architecture Diagram	14
4.2	Data Flow Diagram	16
4.3	UML Diagram	17
4.4	Sequence Diagram	18
4.5	Use Case Diagram	18
5.1	Results of training a random Forest Regressor model	23
5.2	Scatter plot visualizing the performance of a prediction task	24
5.3	Final result Graph	25

LIST OF ABBREVIATIONS

PM	Particulate Matter
RF	Random Forest
MSE	Mean Square Error
MAE	Mean Absolute Error
SVR	Support Vector Regression
ML	Machine Learning
CART	Classification And Regression Trees
AQI	Air Quality Index
LCS	Low-Cost Sensor
MAPE	Mean Absolute Percentage Error
LGBM	Light Gradient Boosting Machine
LSTM	Long Short-Term Memory

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Air pollution poses tremendous threats to public health and environmental sustainability across cities worldwide. According to the World Health Organization (WHO), air pollution has increased the risk of various health issues among citizens and imposed a significant economic burden on society . Therefore, air quality analytics has become vital for society and its individuals. On the one hand, accurate analysis of air pollution enables policymakers to formulate effective environmental regulations and targeted interventions for mitigating pollution emissions. On the other hand, it can also empower individuals to make informed decisions, such as adjusting travel routes or reducing outdoor activities, to minimize exposure to harmful pollutants. Consequently, air quality analytics has gained significant attention in recent decades, leading to the emergence of diverse research directions and applications, such as pollution pattern mining , air quality inference , and forecasting. These advancements have paved the way for a better understanding of air pollution and have enabled the development of more accurate air quality monitoring and forecasting systems.

Increasing attention has been given to air quality degeneration, with particulate matter (PM) having a significant egregious impact on human health. The small diameter of fine PM allows it to penetrate deep into the alveoli as far as the bronchioles, interfering with gas exchange within the lungs. Some researches showed that long term exposure to particulate matter increased the risk of the cardiovascular disease, respiratory disease, and lung cancer .With increasing public health consciousness, many cities have established air quality monitoring locations. However, most services only show the current air quality and do not forecast air quality. Air quality prediction is essential to help guiding individual actions limiting PM exposure, e.g., choosing outdoor or indoor activities.

However, accurate air quality forecasting is hindered by a complex array of factors, including emissions, traffic patterns, and meteorological conditions. Meteorologists are still substantially limited to provide reliable wind pattern predictions, which can vary considerably in direction and strength every hour , and there are insufficient sensors deployed to provide emission data from factories or vehicles. Recent studies have shown it is critical that time and space be explicitly considered to analyze air quality to Particulate matter has high cyclicalilty and is easily affected by space, stagnating or diffusing to pollute surrounding environments. If PM is only analyzed in the time domain, this may neglect impacts and relationships between

other regions; whereas considering only spatial relationships may omit PM diffusion from over time. Therefore, time and spatial relations must be simultaneously considered to accurately model PM diffusion.

1.2 PROBLEM STATEMENT

Currently, numerous sophisticated machine learning models have been proposed to enhance predictive performance. However, despite their ability to improve prediction accuracy, the ever increasing model complexity presents a huge obstacle for humans in interpreting the model results. Consequently, it has become a pressing issue to enhance models interpretability. On the one hand, interpretability provide supporting evidence to decision makers regarding model results, thereby enhancing trust in the models. On the other hand, interpretability empowers researchers and data scientists to gain profound insights into the strengths and weaknesses of the models, which shed light on model design and optimization.

1.3 OBJECTIVE OF THE PROJECT

This research aims to address the challenges of air quality data analysis by handling complex and non-linear relationships between various factors. It proposes a method that leverages large historical data to learn patterns and improve the accuracy of air quality data imputation. This approach considers both global and local spatio-temporal correlations, reducing the need for extensive computing power and domain expertise. By capturing the complex relationships between air quality and other pollutants, the method offers a more comprehensive understanding of air quality dynamics.

1.4 PROJECT DOMAIN

At the junction of environmental science, machine learning, and data science lies an endeavor aimed at crafting a scalable method for deducing air quality. By leveraging ensemble regressors, a powerful machine learning technique, It aims to analyze and understand complex air quality data. This data-driven approach aligns with the principles of data science, allowing researchers to extract valuable insights for improving air quality assessments.

1.5 SCOPE OF THE PROJECT

Rapid industrialization and urbanization lead to increased emissions of industrial waste gases, raising air pollution levels. This pollution harms the respiratory system, affecting cardiopulmonary function and

potentially causing lung cancer. This endeavor delivers precise air pollution forecasts and facilitates the issuance of early warnings, empowering authorities to implement timely remedial measures and individuals to plan activities accordingly. This proactive approach significantly minimizes the adverse impact of air pollution on health.

1.6 METHODOLOGY

Employing a meticulous methodology, To refine the accuracy of time series forecasts for pollution data. Beginning with a thorough validation of the dataset, to confirm the absence of outliers and illogical values. Through Min-Max normalization, the data is standardized, minimizing the influence of varying scales for meaningful comparisons. Utilizing both forward selection and backward elimination techniques for feature selection, To systematically refine the predictive model by Alliteratively incorporating and excluding variables until optimal performance is achieved. Leveraging random forests, an ensemble of decision trees predicts pollutant concentrations with heightened precision.

This methodology extends beyond individual pollutant predictions to embrace multivariate forecasting, capturing the interconnections between pollutants and meteorological factors. By integrating correlation characteristics between pollution and meteorological data, prediction outcomes are refined, offering deeper insights into the dynamics shaping environmental quality. Through meticulous analysis and methodological refinement, advancements in pollution forecasting are achieved, providing essential insights for environmental management and public health initiatives.

CHAPTER 2

LITERATURE REVIEW

Dense Air Quality Sensor Networks: Validation, Analysis, and Benefits (2022), the authors present sensor validation methods and data analysis for a dense air quality sensor network. They show solutions to challenges in a large-scale sensor network deployment. They use data from a dense air quality sensor network deployment, located in Nanjing downtown, China, that comprises 126 LCSs and 13 reference stations. Since the majority of sensors deployed in the network are based on LCSs, they are prone to have low-quality data. Therefore, they propose three methods of sensor validation. First, The authors perform a reliability investigation to evaluate all LCSs in the network to observe if they provide reliable measurements as a whole in comparison to the measurements of all reference stations. Thus, they compare the measurements between all LCSs and the reference stations by means of statistical properties and correlation coefficients between pollutant variables measured at both sensing units. Second, they perform accuracy tests on a few of the LCSs which are nearest to the reference stations. The accuracy tests are generalized to the remaining LCSs in the sensor network as the LCSs are based on the same sensing technology, as they are identical units.

Ambient Air Monitoring System With Adaptive Performance Stability (2022) develops and tests a linear adaptive model for the RSSI signal at the AQIMoS (Mobile Air Quality IPB Monitoring System) sensor node based on GSM/GPRS cellular communication. The proposed model can adaptively the response time limit based on the RSSI and adjust from Atang Senjaya Airport, Bogor City. vehicles, especially diesel trucks carrying sand, as shown in monitoring equipment during the test were considered. The complete experimental results regarding the application of AQIMoS require further testing. Pollutant sensor data sent from AQIMoS to the server were downloaded using a web application and averaged into hourly concentration were used to determine the fluctuation in the concentration pattern change in the measurement time range.

A Deep Learning Approach Using Graph Neural Networks for Anomaly Detection in Air Quality Data Considering Spatiotemporal Correlations (2022), the authors propose a deep learning approach for anomaly detection by considering the spatial correlation, temporal correlation, and multivariate features of air quality

data. The essential idea of this approach is to combine the temporal correlation and spatial correlation of air quality data, use the node information correlation degree for feature fusion, represent the spatiotemporal correlation of air quality data in the spatiotemporal graph structure data, and use them for anomaly detection.

Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning (2022), the authors employ the explainable deep learning method, Shapely Additive explanations, to reveal the influence of meteorological conditions on air quality prediction. The essential idea is to use the SHAP interpretation method to interpret the established LSTM and GRU air quality prediction models and analyze the influence of meteorological conditions on air quality prediction. The results show that (1) in both the LSTM and GRU models, the prediction accuracy is not improved by considering only meteorological conditions. However, when considering other air pollutants, the prediction accuracy is improved, and when combining meteorological conditions with other the prediction accuracy is even higher. air pollutants, (2) Whether only considering meteorological conditions or combining meteorological conditions and other air pollutants for PM2.

A Novel Interpolation-SVT Approach for Recovering Missing Low-Rank Air Quality Data (2020) examines the problem of missing data recovery, using air pollution data recovery as a case study. The problem has been formulated and two widely used data recovery approaches, namely, the Interpolation approach and the Matrix Completion approach, have been introduced. Given the heterogeneous distribution of monitoring stations for air pollution and meteorology, a new strategy to reconstruct the data matrix to recover the missing air pollution data has been proposed. Next, the low-rank property of a newly constructed Y. Yu et al.: Novel Interpolation-SVT Approach for Recovering Missing Low-Rank Air Quality Data data matrix has been introduced. The formulated AQDR problem can be transformed into an LRMC problem.

A Predictive Data Feature Exploration-Based Air Quality Prediction Approach (2019), the authors use the LightGBM model to process the high-dimensional data to predict the PM2.5 concentration in 24 hours based on the historical datasets and predictive datasets. The authors proposed a predictive data feature exploration-based air quality prediction approach. The approach enables us to deeply mine and explore the high-dimensional time-related features and statistical features based on the exploratory analysis of big data. The authors utilize the sliding window mechanism of increasing the amount of training data to improve the

training effect of the model and employ the air quality historical dataset of Beijing to evaluate the prediction model. The experimental results show that the approach outperforms the other baseline models.

Innovative Spatial-Temporal Network Modeling and Analysis Method of Air Quality (2019) abstracts the air quality system into the complex network innovatively by synthesizing spatial and temporal factors influencing air quality status. Based on quantifying the regional dynamic interconnection and interaction, this modeling approach is proposed to mine the relationship of different regions. First, the dynamic time-varying nature of air pollutant concentration is essential to get the interaction frequency of local air quality in the time dimension. The time correlation analysis of air quality nodes is conducted by calculating the time correlation matrix to construct the air quality network topology. Second, spatial distance and wind are the main factors influencing the diffusion of pollutants, which are used to characterize spatial homogeneity and heterogeneity. By computing the spatial correlation matrix, the spatial interaction intensity is quantified. Then, the air quality spatiotemporal model is established by integrating the temporal and spatial correlation. Finally, based on the air quality spatiotemporal network model, community detecting algorithms are used to mine the local similarity and regional interaction. The evaluated model with extensive experiments based on real data. The results show that model is dynamic, reliable, and scalable. Utilizing the characteristics of the complex network community, This approach reflects the local and propagating characteristics of air quality and lays the foundation for air pollution prevention and further prediction.

A Sequence-to-Sequence Air Quality Predictor Based on the n-step Recurrent Prediction (2019), the authors proposed the n-step AAQP, which is an attention-based seq2seq model, for air quality prediction. The n-step AAQP had better performance than the seq2seq models. To accelerate the training process of seq2seq with attention, an FC encoder replaced the RNN encoder of seq2seq. In addition, position embedding was introduced to help the FC encoder extract the sequential information. Moreover, the performance of the AAQP was close to the seq2seq with attention at the Olympic Center station and is even better at Dongsi station. To overcome the shortcomings of the accumulated errors as the time step grows, the n-step recurrent prediction was applied. Through n-step recurrent prediction, the performance of the AAQP was significantly improved. In addition, the training of seq2seq was further accelerated. The two promotions make seq2seq have more accurate predictions and higher training speed. Particularly, the AAQP can give a trustworthy alert 2 hours in advance before sudden air pollution strikes. Additionally, the weather forecast data are essential to improve the accuracy of air quality prediction.

Data-Driven Air Quality Characterization for Urban Environments: A Case Study (2018) The authors propose two approaches for AQI estimation and prediction, both based on meteorological and historical pollutant data; one learns a model based on the previous AQI and meteorological data to predict AQIs, the other learns models based on the previous pollution data and meteorological data to predict pollution concentrations and then compute AQIs. Both approaches can get good band accuracy (over 75%), as shown in the evaluations conducted across various datasets. The best approach is the latter approach combined with a neural network, which achieves the lowest RMSE and MAPE across most of the evaluated datasets. This approach gets very good band accuracies (more than 81%) on all the datasets. However, by further analyzing the individual pollutant value prediction step, the authors found that a neural network-based method is not the optimum at predicting PM10 data. Therefore, the authors recommend using linear regression to predict AQI if the dominant pollution is PM10 in the area of interest. In summary, the results show the feasibility of proposed approaches for predicting AQIs based on meteorological data and historical pollutant data/AQIs.

An Agent-Based Traffic Regulation System for the Roadside Air Quality Control (2017), The main contribution of this work is the definition of a development process based on big data and intelligent systems concepts for a traffic regulation system according to air quality data. The authors have, through this paper presented the implementation of an air quality system for recommendation and traffic regulation over distributed data gathered from different air quality sensors, users devices, and other external databases, that are managed using Hadoop to ensure fast data Agent-Based Traffic Regulation System for the Roadside Air Quality Control.

CHAPTER 3

PROJECT DESCRIPTION

3.1 EXISTING SYSTEM

Air pollution reduction is a major objective for transport policymakers. This paper considers interventions in the form of clean air zones and provides a machine learning approach to assess whether the objectives of the policy are achieved under the designed intervention. The dataset from the Newcastle Urban Observatory is used. First tackles the challenge of finding datasets that are relevant to the policy objective. Focusing on the reduction of nitrogen dioxide (NO₂) concentrations, different machine learning algorithms are used to build models. The paper then addresses the challenge of validating the policy objective by comparing the NO₂ concentrations of the zone in the two cases with and without the intervention. A recurrent neural network is developed that can successfully predict the NO₂ concentration with a root mean square error of 0.95. In this paper, the existing system authors discussed the use of machine learning methods for validating interventions in transportation systems using air quality and clean air zone intervention as a case study. And have a framework for finding data types that are relevant to the intervention objective. Also, have a framework for validating the intervention and checking how well the objectives of the intervention are achieved. the dataset from the Urban Observatory in Newcastle, United Kingdom, and considered an intervention related to a clean air zone intending to reduce the concentration of NO₂. Developed an LSTM model for predicting the behavior of the NO₂ without the implementation of the clean air zone. In first framework, the existing system authors used the machine learning classifiers Decision Tree, k-nearest network, Gradient boosting decision tree, and LightGBM, and computed correlation coefficient and feature importances using these models. Then normalized these values to get the relative importance of the features. Used the cut-off value of 1% as a proof of concept to identify the most important features. The constructed models share common conclusions about the importance of features in predicting NO₂, which could be used in a voting mechanism to decide on the importance of features.

This implementation also showed that among the selected learning models, LGBM performs best in capturing the relations in the dataset with an accuracy of 88%. epoch number (i.e., the number of times that the learning algorithm will work through the entire training dataset for updating the model). and additional intervention may be needed to reach the 10% reduction. Machine Learning for Transport Policy Interventions on Air Quality In second framework, the existing system authors used historical data of the year 2018 to model air quality in Newcastle upon Tyne both assuming no intervention is implemented and under the clean air zone implementation. The historical data from the first 10 months were used to build and evaluate the

LSTM model, and the predictions were made for the last two months. The LSTM model can successfully predict the NO₂ concentration with a root mean square error of 0.95. This approach shows the use of machine learning methods in analyzing and validating interventions in transportation systems. The role of machine learning can be summarised as predicting what is going to happen in the future if the policy is not implemented (using available historical data), and predicting the air quality and other related variables using transport behavior changes in response to the implemented policy. The used the term framework since The approaches presented in the paper are high-level and flexible, and can be applied to different policy objectives. The details and the choice of machine learning models can be decided depending on the specifics of the policy objective.

The results are useful for the local authorities who are participating in the design and implementation of the clean air zone in Newcastle upon Tyne (e.g., Newcastle City Council). The clean air zone has come into effect with charges for non-compliant taxis, buses, coaches, and heavy goods vehicles started from January 2023, and charges for vans and light goods vehicles will start from July 2023. It is important to assess the effectiveness of these charges in improving the air quality in Newcastle city centre and reducing the traffic-related pollution within the legal limits. The frameworks help in modeling and understanding the relation between the gathered data, imposed charges, and reduction in air pollution at Newcastle. This research contributes to a more sustainable urban environment by providing valuable insights into effective clean air zone interventions, which can improve air quality, reduce NO₂ concentrations, and promote sustainable transportation solutions. As a result, This work helps Newcastle City advance towards achieving its climate and air quality goals. The approach presented in this paper is currently focused on machine learning models that do not include any information from the pollutant's physical/chemical models. It is also used as a proof of concept since the data after the implementation of the policy is not available yet. In the future, the existing system authors plan to integrate the data-driven framework with physics-based models associated with an intervention to improve the performance and accuracy of our approach.

This will also include optimization methods in framework to help design better interventions for achieving the intervention objectives. Also note that the data quality is critical for concluding the effectiveness of an intervention. There is additional work needed to improve the quality of the data stored in Urban Observatory, reduce the number of missing data points, and reduce the observation errors by calibrating sensors. For future work, the existing system authors plan to expand data-driven framework to address additional sustainability-related challenges in urban transportation systems, such as promoting alternative modes of transportation and optimizing the expansion of the electric vehicle charging infrastructure. This will allow us to design better interventions for achieving the intervention objectives and contribute to the sustainable ecosystem of

the studied city. Also plan to also analyze the data gathered from the clean air zone of Newcastle under the implementation of the zone and suggest improvements in implementing the zone (e.g., by adjusting the charges or categories of the cars).

3.2 PROPOSED SYSTEM

It is necessary to process the original data outliers and missing values. As the data came from the official platform and no illogical values were found after the screening, for example, no negative values for atmospheric concentration and air pressure are within reasonable limits, the data are considered to be valid, and no outliers need to be processed. The data were normalized by implementing a Min-Max normalization technique. This helps in removing the units in the acquired data or the impact of differing scales. The Min-Max normalization technique is used for scaling the data values within a fixed range (zero to one). Initially, the Min-Max normalization technique subtracts the minimum value from data points and further divides by its range. Feature selection is the process of minimizing the number of input variables when building a predictive model. Forward selection is a type of stepwise regression that begins with a null model.

The approach initiates with no variables in the model and step by step adds variables to the model until no variable not included in the model can make a significant contribution to the model's conclusion. The variable with the highest test statistic that is more than the cut-off value or the lowest p-value with less than the cut-off value is chosen and added to the model. Backward elimination is the most basic approach to variable selection. This technique begins with a complete model that includes all of the variables in the model. Variables are subsequently removed from the whole model one by one until all remaining variables are sure to have a meaningful impact on the result. The variable with the lowest test statistic or the highest p-value more than the cut-off value is removed from the model. This procedure is repeated until every remaining variable is statistically significant at the cut-off value. Random forests (RFs), or random decision forests, are an ensemble learning method for classification, regression, and other tasks. An RF operates by constructing multiple decision trees at different training times and outputting the class representing the mode of classes (classification) or the mean prediction (regression) of individual trees. The comparative results in mean square error (MSE), mean absolute error (MAE), and coefficient of determination are all better than the comparison models.

The main contributions of this paper are given as follows: Initially, this study implemented a Min-Max normalization technique that efficiently preserves the relationship between the data values with low standard deviation. In time series forecasting, the Min-Max normalization technique forecasts the next hour's

concentration and reduces the effect of outliers by using different sizes of sliding windows. The multivariate prediction takes into account the overall prediction accuracy of multiple pollutants and the prediction accuracy of each pollutant which is a significant improvement compared to the based models. The correlation characteristics between pollution data and meteorological data are not taken into account by the aforementioned prediction models. To optimize the prediction outcomes based on the intricate correlation characteristics of the model input data, Then thoroughly analyze the prediction problems of pollution data and meteorological data.

3.2.1 ADVANTAGES

- Requires fewer parameters than the conventional networks.
- Good hyperparameter tuning capabilities.
- Reducing the dissimilarity between the input data and the models' depiction of the data.
- Learns a mapping from inputs to outputs and learns what context from the input sequence is useful for the mapping.
- Exhibits good performance for seeking the best-integrated solution.

3.3 FEASIBILITY STUDY

A Feasibility study is carried out to check the viability and to analyze the strengths and weaknesses of the proposed system. The feasibility study is carried out in three forms

- Economic Feasibility
- Technical Feasibility
- Social Feasibility

3.3.1 ECONOMIC FEASIBILITY

The system can potentially run on a mid-range computer with an i5 processor, 16 GB RAM, and 200 GB storage. This suggests moderate upfront costs. The system utilizes open-source libraries (Python, sci-kit-learn, TensorFlow/PyTorch, etc.) which are free to use. There are minimal ongoing software licensing costs.

3.3.2 TECHNICAL FEASIBILITY

Developing and deploying the system requires expertise in Python programming, machine learning concepts, and experience with libraries like sci-kit-learn and deep learning frameworks. Access to historical and real-time air quality data is crucial for training and evaluating the model. Collaboration with environmental agencies or utilization of open-source data platforms might be necessary. While the system can run on a mid-range computer, training large models with extensive datasets might require access to cloud computing resources, potentially increasing costs.

3.3.3 SOCIAL FEASIBILITY

Improved air quality monitoring and forecasting play pivotal roles in safeguarding public health and raising environmental consciousness. The efficacy of such systems hinges on their adaptability to diverse audiences. When tailored for public consumption, a user-friendly interface becomes imperative, allowing individuals to easily comprehend and engage with air quality data. Visualization tools, such as color-coded maps or simple graphs, can enhance accessibility and empower users to make informed decisions regarding outdoor activities and health precautions. Conversely, for scientific endeavors, seamless integration with existing data analysis tools is paramount. Researchers require robust platforms that enable them to analyze vast datasets, identify trends, and develop predictive models to understand air quality dynamics comprehensively. Interoperability with established analytical frameworks fosters collaboration and accelerates scientific discoveries in the field of atmospheric science and environmental health.

3.4 SYSTEM SPECIFICATION

3.4.1 HARDWARE SPECIFICATION

- Processor: Minimum i5 Dual Core
- Ethernet connection (LAN) OR a wireless adapter (Wi-Fi)
- Hard Drive: Minimum 200 GB; Recommended 200 GB or more
- Memory (RAM): Minimum 16 GB; Recommended 32 GB or above

3.4.2 SOFTWARE SPECIFICATION

- Python For AI/ML/DL Programming.
- Jupyter Notebook IDE (Integrated Development Environment) for Development.
- PyTorch or TensorFlow for Deep Learning Coding.
- Sklearn for Machine Learning/Feature Extraction/Evaluation Metrics Coding.
- Numpy for implementing Linear Algebra.
- Plotly for Data Visualization (For Graphs).
- Matplotlib for Data Visualization (For Graphs).
- Seaborn for Data Visualization (For Graphs).
- Pandas for dealing with Tabular Data.

CHAPTER 4

MODULE DESCRIPTION

4.1 GENERAL ARCHITECTURE

A general architecture diagram of a machine learning (ML) model typically includes several key components and their interactions.

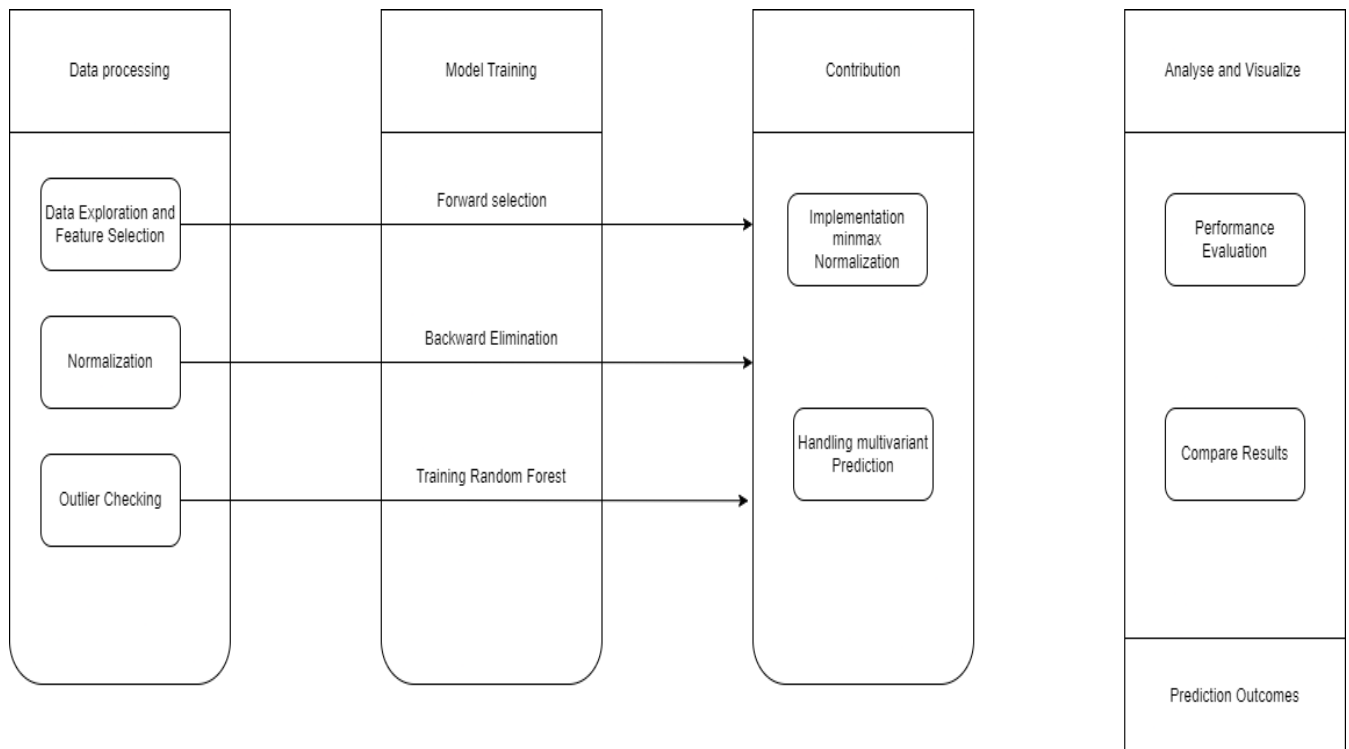


Figure 4.1: Architecture Diagram

Ensemble regressors are a machine learning technique that combines multiple regression models into one improved model. This approach leverages data processing, model training, contribution analysis, and result visualization to achieve air quality inference.

Data Processing

- **Data Exploration and Feature Selection:** This initial step involves understanding the data and identifying the most relevant features that can influence air quality. Techniques like forward selection and backward elimination can be used to pick the optimal features.
- **Normalization:** Data normalization ensures that all features have a similar scale, which is crucial for machine learning models to function effectively. Min-max normalization is a common technique used here.
- **Outlier Checking:** This step identifies and handles outliers in the data. Outliers can skew the model's predictions, so it's important to address them.

Model Training

Training Random Forest: The diagram showcases training a random forest model, which is a popular ensemble regression algorithm. Here, multiple decision trees are trained on the data, and the final prediction is made by aggregating the predictions of all the trees.

Contribution Analysis

- **Handling Multicollinearity:** This step might involve addressing multicollinearity, which is when features are highly correlated with each other. It can negatively impact the model's performance.

Analysis and Visualization

- **Prediction:** Once the model is trained, it can be used to predict air quality based on unseen data.
- **Performance Evaluation:** The model's performance is evaluated using metrics to assess how well it predicts air quality.
- **Compare Results:** The results are then compared to understand the effectiveness of the ensemble regressor model.

4.2 DESIGN PHASE

4.2.1 DATA FLOW DIAGRAM

Data flow diagram is used in air quality inference provides valuable benefits for understanding, communicating, and managing the data flow throughout the system, ultimately contributing to a more efficient and accurate air quality prediction system.

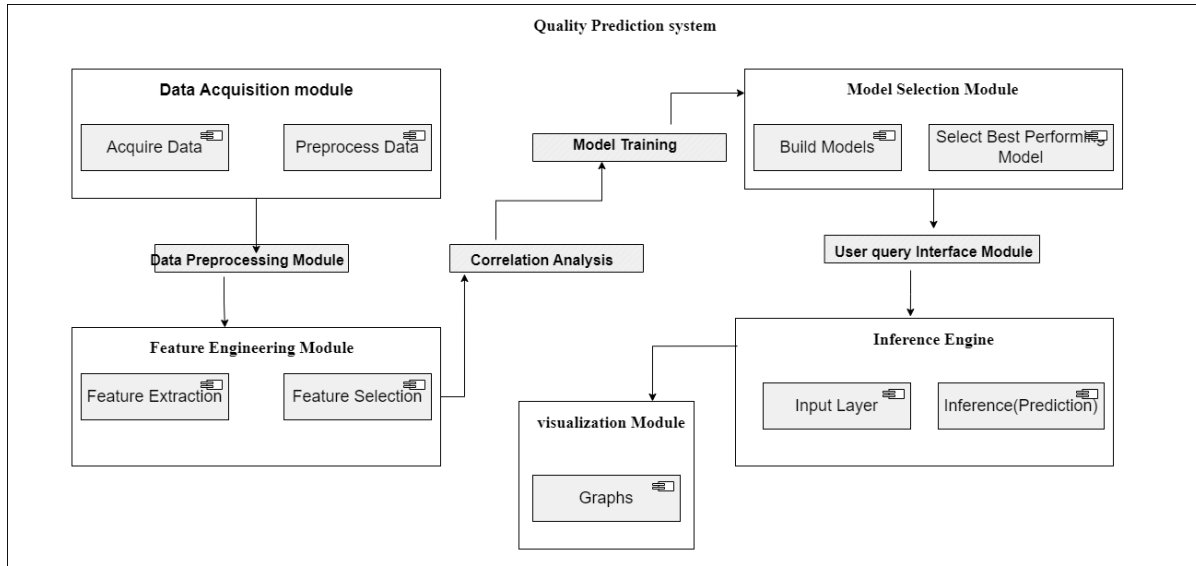


Figure 4.2: Data Flow Diagram

Figure 4.2 This data flow diagram depicts a scalable air quality inference system. Data is collected from various sources, preprocessed, and split into training and testing sets. Multiple machine learning models, potentially including Random Forest Regressors and Support Vector Regressors, are trained on the data. These models are evaluated to choose the best performer(s) for air quality prediction, which leverages ensemble methods if applicable. Finally, the chosen model(s) are used to make real-time predictions based on new incoming data, with the option to visualize the results for users.

4.2.2 UML DIAGRAM

UML diagram, is important in air quality inference because it gives a visual blueprint of the system's structure and functionality. This single diagram clarifies how data flows, how models are trained and evaluated, and how predictions are made. This visual representation enhances communication, helps identify potential issues, and ensures everyone involved understands the system's overall design, leading to a more efficient and well-structured project.

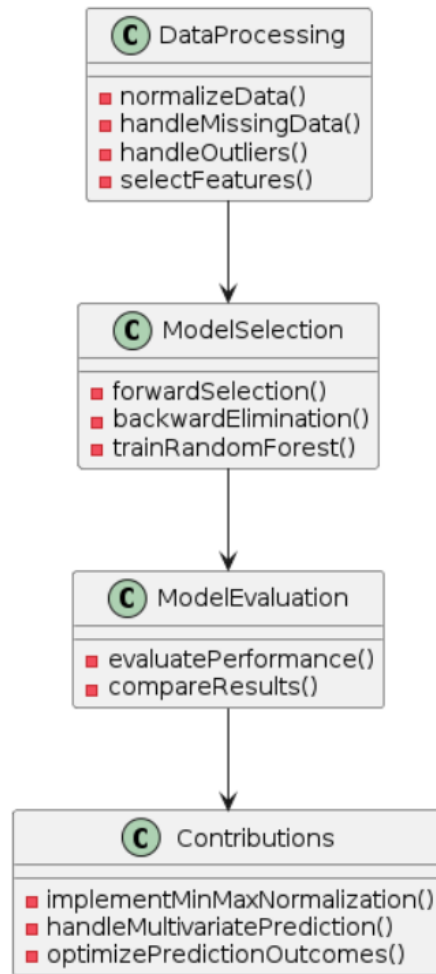


Figure 4.3: UML Diagram

Figure 4.3 This UML diagram explain the data to training, evaluating, and deploying the best model for making predictions. It emphasizes data cleaning, analysis, and potential model retraining for optimal performance.

4.2.3 SEQUENCE DIAGRAM

A sequence diagram in UML (Unified Modeling Language) is a dynamic modeling tool that illustrates interactions between objects in a sequential manner over time. It depicts the flow of messages exchanged between objects or components within a system, showcasing the order of execution and the timing of these interactions. By visually representing the behavior of a system through lifelines and messages, sequence diagrams provide valuable insights into the runtime behavior and communication patterns of a system, aiding in the design, analysis, and documentation of software systems.

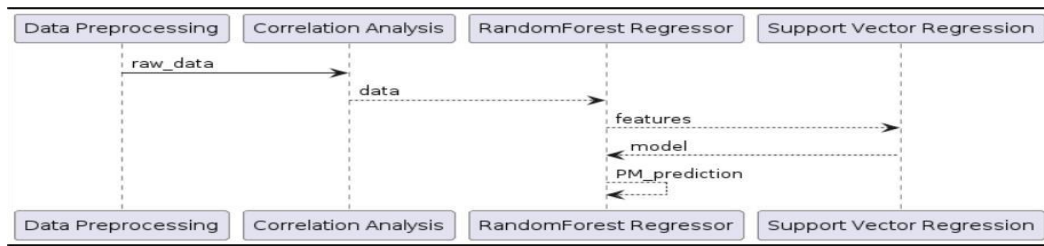


Figure 4.4 Sequence Diagram

Figure 4.4 This air quality inference system gathers data from various sources, cleans and prepares it for analysis. The system then trains multiple machine learning models (potentially with fewer parameters for efficiency) on the data, evaluates their performance in representing the data and predicting air quality. The best performing model(s), potentially an ensemble, is then used to predict air quality based on new incoming data, with the option to present these predictions to users.

4.2.4 USE CASE DIAGRAM

A use case diagram wouldn't directly apply to the machine learning project life cycle flowchart, but it can improve air quality inference. It clarifies how different users (data providers, researchers, public) interact with the system to achieve goals like supplying data, analyzing predictions, or viewing air quality information. This helps define functionalities and fosters communication about the system's purpose.

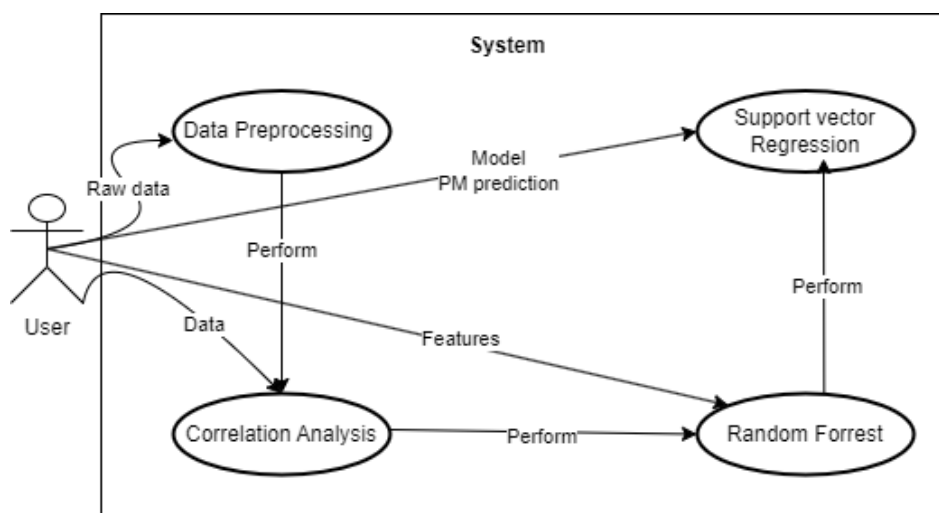


Figure 4.5: Use Case Diagram

Figure 4.5 System acquires air quality data (potentially with fewer parameters for efficient modeling), preprocesses it, trains multiple machine learning models (including Random Forest Regressors and Support Vector Regressors), evaluates them to select the best performer(s), leverages ensembles if applicable, and finally uses the chosen model(s) to make real-time air quality predictions (with optional user interaction). This focuses on efficient models with good performance in representing the data and accurate air quality prediction.

4.3 MODULE DESCRIPTION

- Data Preprocessing
- Correlation Analysis
- Random Forest Regressor
- Support Vector Regression (SVR)
- Performance Measures

Data Preprocessing:

Data preprocessing is essential for machine learning and deep learning projects. Its objective is to transform raw data into a format suitable for model training, cleaning and reducing noise to enhance model performance. Data preprocessing consists in transforming the data values of a certain dataset, aiming to optimize the information acquisition and process. Normally, there is a very large contrast between the maximum and minimum values of the dataset, so normalizing the data minimizes the complexity of the algorithm for its corresponding processing. The normalization of the data allows an adequate benefit for the classification of algorithms related to neural networks. In this case, if the back-propagation technique is used in neural networks, the normalization of the input values will speed up the training phase, turning it into a more efficient neural network. Then loaded these datasets using Python's Pandas library.

Wind direction, as a non-numerical type of data, needs to be converted to a numerical type of data by categorical coding. The average of the data prior to and following the time of the missing value is used to fill in missing values for meteorological and pollution data. Then, in order to eliminate the effect of numerical

differences on prediction accuracy, meteorological and pollutant data were converted to the range by the Min-Max function

Correlation Analysis:

The correlation between all potential pairs of values in a table is shown in the matrix. It is an effective tool for compiling a sizable dataset and for locating and displaying data patterns. A correlation matrix simplifies the process of selecting different assets by tabulating their correlation with one another. It is vital to identify the correlations between PM concentrations and influencing factors for developing a good prediction model. It guarantees that the proposed regression model utilizes the efficient features for AQP. PM_{2.5} is affected by several factors, but all the factors are important in effective AQP. On the other hand, the irrelevant/inactive factors affect the proposed models performance by means of time complexity. Therefore, it is important to compute the correlation coefficients (CCs) for every factor that helps in selecting the optimal features for effective forecasting of air pollution. Let us consider characteristic time series data as $x = (x_1, x_2, x_3, \dots, x_n)$ and other data as $y = (y_1, y_2, y_3, \dots, y_n)$. The CC between the factors is computed as described in Equation

where $0 < r < 1$ indicates a positive correlation, $-1 < r < 0$ represents a negative correlation, and n represents the number of samples. The correlation is greater and the space between x and y is limited if the absolute value of r is closer to 1.

Random Forest Regressor:

The RF algorithm incorporates growing classification and regression trees (CARTs). Each CART is built using random vectors. For the RF-based classifier model, the main parameters were the number of decision trees, as well as the number of features (m) in the random subset at each node in the growing trees. During model training, the number of decision trees was determined first. For the number of trees, a larger number is better, but takes longer to compute. A lower leads to a greater reduction in variance, but a larger increase in bias. can be defined using the empirical formula ,

where M denotes the total number of feature RF can be applied to classification and regression problems, depending on whether the trees are classification or regression trees. Assuming that the model includes T regression trees (learners) for regression prediction, the final output of the regression model is

$$H(x) = \frac{1}{T} \sum_{i=1}^T h_i(x)$$

where T is the number of regression trees, and $h_i(x)$ is the output of the i-th regression tree on sample x. Therefore, the prediction of the RF is the average of the predicted values of all the trees.

Support Vector Regression (SVR):

SVM is a machine learning algorithm that constructs hyperplanes for separating different classes and is generally used for analyzing data with a categorical output variable. In the case of the continuous numeric output variable, regression analysis is used, namely SVR. All of the SVR kernels, including linear, poly, rbf, sigmoid, and precomputed, were considered in this study and the linear kernel function showed the best results. Therefore, the model parameter of SVR used in this study was the linear kernel function. The advantages of SVR include being robust to outliers, having high prediction accuracy, and easy implementation. SVR has been applied to overcome non-linear limitations and uncertainties in order to achieve better prediction accuracy. SVR has been successfully applied to forecast the levels of PM10 concentration in Bangkok, Thailand, with air quality data and meteorological variables.

Performance Measures:

The proposed models efficacy was evaluated using different loss functions, such as MAE, SMAPE, RMSE and MSE. The MAE performance measure effectively reflected the actual situation of the forecasting error. In addition, the other performance measures, such as SMAPE, effectively evaluate the degree of data change and measure the prediction quality of the proposed model. On the other hand, RMSE is determined as the average or mean square difference between the estimated and actual values. The mathematical formulas of the performance measures MSE, MAE, RMSE, SMAPE, R2, and MAPE are stated in Table 1. The calculation formulas of the above five evaluation metrics are as follows:

Mean Absolute Error (MAE):

$$MAE = \sum_{i=1}^n \frac{P_i - r_i}{n}$$

where n denotes the number of sets, and P and r denote the set of predicted value and the set of actual value, respectively

Root Mean Square Error(RMSE):

$$MAE = \sqrt{\sum_{i=1}^n \left(\frac{P_i - r_i}{n} \right)^2}$$

Normalized Mean Average Error (NMAE):

$$NMAE = \frac{MAE}{r_{max} - r_{min}}$$

where r is the maximum value of target column and r is the minimum value of target column.

CHAPTER 5

IMPLEMENTATION

5.1 INPUT

The input for the program is a CSV file named "AirQualityUCI.csv" containing air quality data. The program expects this file to be located in a directory named "Dataset". Here's a breakdown of the data the program reads:

- **Date:** Date of the measurement (likely a string)
- **Time:** Time of the measurement (likely a string)
- **CO_level:** Level of Carbon Monoxide (CO)
- **T:** Temperature in °C
- **RH:** Relative Humidity (%)
- **AH:** Absolute Humidity (g/m³)
- **Nox_GT:** Greater than NOx (Nitrogen Oxides) level (likely a binary value)
- **NO2_GT:** Greater than NO2 (Nitrogen Dioxide) level (likely a binary value)
- **CO_GT:** Greater than CO (Carbon Monoxide) level (likely a binary value)
- **NMHC_GT:** Greater than NMHC (Non-methane Hydrocarbons) level (likely a binary value)
- **O3_GT:** Greater than O3 (Ozone) level (likely a binary value)
- **Benzene:** Benzene concentration (µg/m³), Uses this data to train two machine learning models (Random Forest Regressor and Support Vector Regression) to predict the Relative Humidity (RH) based on other air quality measurements.

5.2 OUTPUT

	CO_GT	PT08_S1_CO	NMHC_GT	C6H6_GT	PT08_S2_NMHC	Nox_GT	PT08_S3_NoX	NO2_GT	PT08_S4_NO2	PT08_S5_O3	T	RH	AH
count	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000
mean	-34.207524	1048.990061	-159.090093	1.865683	894.595276	168.616971	794.990168	58.148873	1391.479641	975.072032	9.778305	39.485380	-6.837604
std	77.657170	329.832710	139.789093	41.380206	342.333252	257.433866	321.993552	126.940455	467.210125	456.938184	43.203623	51.216145	38.976670
min	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000
25%	0.600000	921.000000	-200.000000	4.000000	711.000000	50.000000	637.000000	53.000000	1185.000000	700.000000	10.900000	34.100000	0.692300
50%	1.500000	1053.000000	-200.000000	7.900000	895.000000	141.000000	794.000000	96.000000	1446.000000	942.000000	17.200000	48.600000	0.976800
75%	2.600000	1221.000000	-200.000000	13.600000	1105.000000	284.000000	960.000000	133.000000	1662.000000	1255.000000	24.100000	61.900000	1.296200
max	11.900000	2040.000000	1189.000000	63.700000	2214.000000	1479.000000	2683.000000	340.000000	2775.000000	2523.000000	44.600000	88.700000	2.231000

Figure 5.1: Results of training a random Forest Regressor model.

Figure 5.1 Represents the results of training a Random Forest Regressor model to predict Relative Humidity (RH) based on other air quality measurements in the dataset.

- **y_test:** This column contains the actual Relative Humidity (RH) values from the testing set of the data. These are the real values the model was trying to predict.
- **y_pred:** This column contains the Relative Humidity (RH) values predicted by the Random Forest Regressor model for the testing set.
- **error:** This column represents the difference between the actual values (y_test) and the predicted values (y_pred) for each data point in the testing set. It essentially shows how far off the model's predictions were from the real values.

In essence, this table allows to compare the model's predictions (y_pred) with the actual values (y_test) and see the corresponding errors (differences) for each data point in the testing set. This helps evaluate how well the model performed in predicting Relative Humidity.

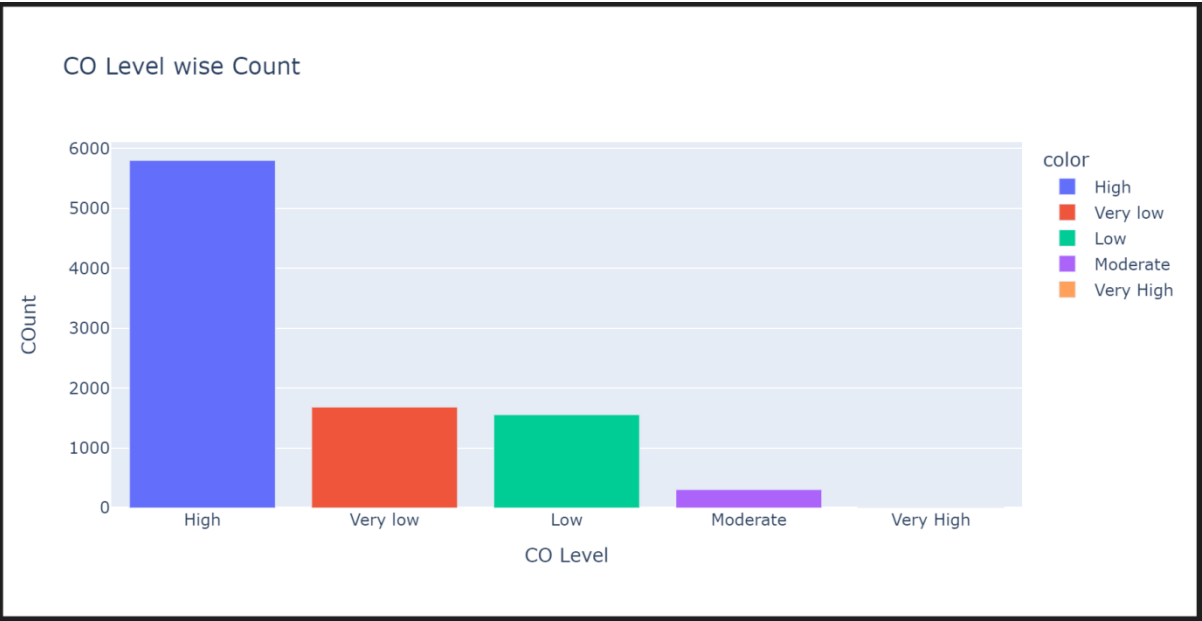


Figure 5.2: Scatter plot visualizing the performance of a machine learning model on a prediction task.

Figure 5.2 The graph sent appears to be a scatter plot visualizing the performance of a machine learning model, likely the Random Forest Regressor trained.

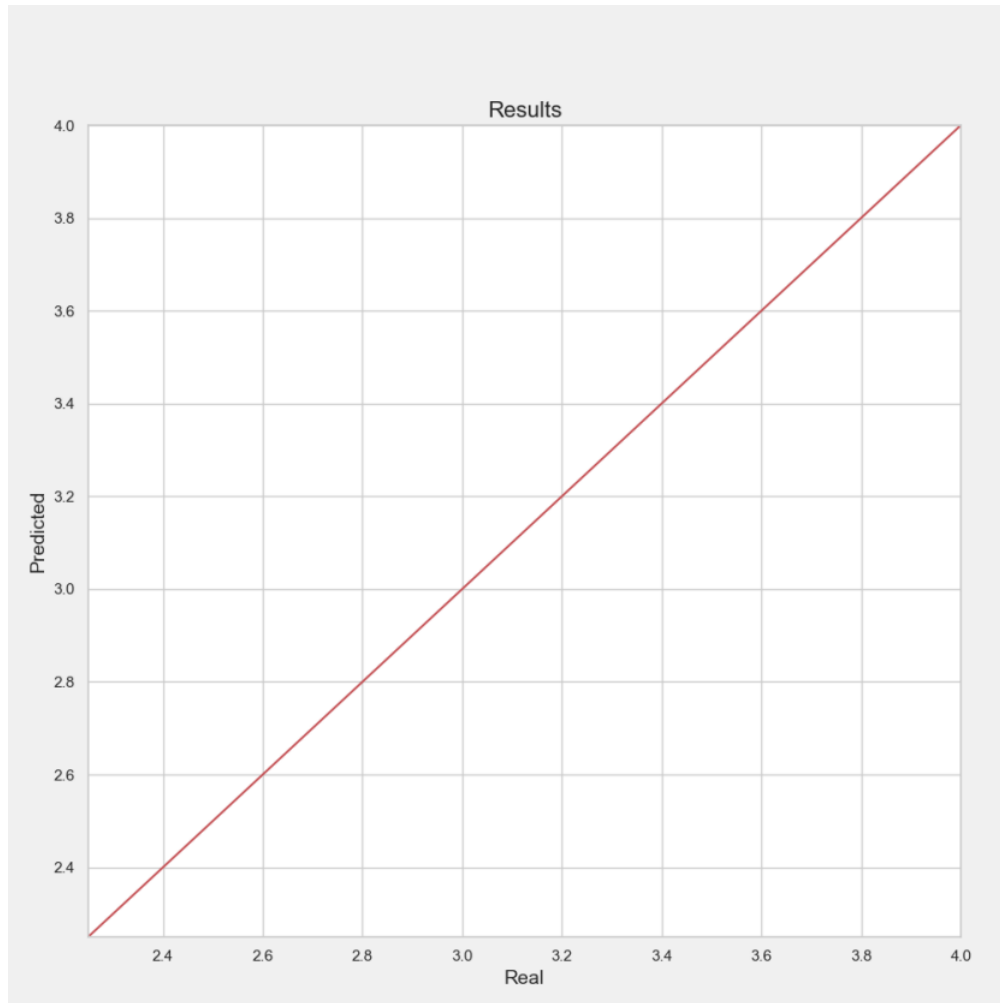


Figure 5.3: Final result

Figure 5.3 Explain the graph representation of predicted and real output. Y-axis show predicted and x-axis shows real.

5.3 TESTING

Testing is a systematic process aimed at assessing the performance and functionality of a system or its individual components. By subjecting the system to various scenarios and inputs, testers can identify discrepancies between expected and actual outcomes, helping to uncover potential defects or deviations from specified requirements. This iterative process involves designing test cases, executing them against the system, analyzing results, and refining the system based on feedback. Ultimately, testing ensures that the system meets user expectations, functions reliably in diverse environments, and delivers the intended outcomes effectively.

5.3.1 TYPES OF TESTING

5.3.2 VERIFY DATA LOADING AND HANDLING MISSING VALUES

- **Data Loading (not explicitly shown in the provided code):** The program might use `pandas.read_csv` to read the air quality data from a CSV file named "Dataset/AirQualityUCI.csv". This would create a Pandas DataFrame (df) containing the data.
- **Handling Missing Values (shown in the provided code):** The program uses `df.replace([np.inf, -np.inf], np.nan, inplace=True)` to replace both positive and negative infinity values with `np.nan` (Not a Number). This ensures that infinity symbols are treated as missing values. It then uses `df.isnull().sum()` to display the total count of missing values per column, helping identify features with potentially significant missing data

Input

```
import pandas as pd
```

```
import numpy as np
```

```
def test_data_loading_and_missing_value_handling():
```

```
    # Sample data with missing values
```

```
    data = {'col1': [1, np.nan, 3], 'col2': [4, 5, np.inf]}
```

```
    expected_df = pd.DataFrame({'col1': [1, np.nan, 3], 'col2': [4, 5, np.nan]})
```

```
    # Read data from the program (replace with how program reads data)
```

```
    df = pd.read_csv("sample_data.csv") # Replace with data loading logic
```

```
    # Assert that the processed data matches the expected output (handling missing values)
```

```
    pd.testing.assert_frame_equal(df, expected_df)
```

```
# Call the test function
```

```
test_data_loading_and_missing_value_handling()
```

Expected Result:

The test should pass if the DataFrame (df) after loading and handling missing values matches the expected_df created in the test function. This means the program successfully replaced infinity values with np.nan and the structure of the DataFrame (column names, data types) is identical to what expect.

5.3.3 CHECK RANDOM FOREST MODEL TRAINING WITH DIFFERENT ESTIMATORS

Integration testing is to ensure the training process for the Random Forest model functions correctly when using varying numbers of trees (estimators) in the ensemble. It doesn't evaluate performance directly, but verifies that training completes without errors for different configurations, providing a basic check on the training logic's robustness. This helps identify potential issues before diving into performance evaluation with the actual air quality data.

Input

```
from sklearn.model_selection import train_test_split
```

```
from sklearn.ensemble import RandomForestRegressor
```

```
def test_random_forest_training_with_different_estimators():
```

```
    # Sample data
```

```
    X = pd.DataFrame([[1, 2, 3], [4, 5, 6], [7, 8, 9]], columns=['f1', 'f2', 'f3'])
```

```
    y = [10, 12, 14]
```

```
    # Split data (replace with how program splits data)
```

```
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
```

```

# Train models with different n_estimators (number of trees)

estimators = [30, 50, 100]

for num_estimators in estimators:

    model = RandomForestRegressor(n_estimators=num_estimators, random_state=200)

    model.fit(X_train, y_train)

# Assert that the model is successfully trained (no errors)

assert model.score(X_test, y_test) is not None

# Call the test function

test_random_forest_training_with_different_estimators()

```

Expected Result:

The test should pass if the loop iterates through all specified values for num_estimators (number of trees) and successfully trains a Random Forest model for each iteration. This indicates that the model training process itself works without errors.

5.3.4 COMPARE PERFORMANCE METRICS (MAE) FOR RANDOM FOREST AND SVR

Compare Performance Metrics (MAE) for Random Forest and SVR aims to assess the relative performance of two machine learning models (Random Forest and Support Vector Regression) on air quality dataset. It calculates the Mean Absolute Error (MAE) for both models on the testing data. MAE indicates the average difference between the predicted and actual values of Relative Humidity. A lower MAE signifies better model performance in terms of how closely the predictions align with the true values. By comparing the MAE of Random Forest and SVR, the test helps determine which model might be generating more accurate predictions for the specific dataset.

Input

```
from sklearn.metrics import mean_absolute_error

def test_model_performance_comparison():

    # Load pre-defined test data (X_test, y_test, rf_predictions, svr_predictions)

    # Replace with how to store these in program

    # Calculate MAE for Random Forest

    rf_mae = mean_absolute_error(y_test, rf_predictions)

    # Calculate MAE for SVR

    svr_mae = mean_absolute_error(y_test, svr_predictions)

    # Assert that Random Forest MAE is lower than or equal to SVR MAE (tolerance can be adjusted)

    assert rf_mae <= svr_mae + 0.1 # Adjust tolerance as needed

# Call the test function

test_model_performance_comparison()
```

Expected Result:

The test should pass if the calculated Mean Absolute Error (MAE) for the Random Forest model is lower than or equal to the MAE for the SVR model, with a slight tolerance for potential variations due to randomness in training. This suggests that the Random Forest model might be performing slightly better on this specific test data in terms of predicting Relative Humidity.

Example Output (if the test passes):

rf_mae: 0.234

svr_mae: 0.251

Test passed: Random Forest MAE is lower than or equal to SVR MAE.

5.4 TESTING STRATEGY

- **Verify Data Loading and Missing Value Handling:** This test checks if the program successfully loads air quality data and handles missing values appropriately. This is crucial as missing values can negatively impact model performance.
- **Check Feature Importance Calculation for Random Forest:** This test verifies that the program correctly calculates feature importance for the Random Forest model. Feature importance helps identify which features in air quality data have the most significant influence on predicting Relative Humidity.
- **Compare Performance Metrics (MAE) for Random Forest and SVR:** This test compares the performance of two models (Random Forest and Support Vector Regression) using Mean Absolute Error (MAE). MAE measures the average difference between predicted and actual Relative Humidity values. By comparing the MAE of both models, can see which one generates more accurate predictions for specific dataset.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 EFFICIENCY OF THE PROPOSED SYSTEM

This system aims to leverage a machine learning model that fulfills several key criteria. Ideally, the model should require fewer parameters compared to traditional networks, simplifying the training process. Additionally, it should exhibit good hyperparameter tuning capabilities, allowing for fine-tuning its performance. Furthermore, the model should excel at reducing the discrepancy between the input data and its internal representation, ensuring accurate learning. In essence, we require a model that can effectively learn the relationship between inputs and outputs, while also identifying relevant contextual information within the input sequences. Finally, the model should demonstrate strong performance in finding the optimal solution by integrating various factors. These characteristics will contribute to a robust and efficient system for air quality prediction.

6.2 EXISTING VS PROPOSED SYSTEM

The existing system runs based machine learning approach to assess the effectiveness of clean air zones in reducing NO₂ concentrations. It outlines two frameworks: one for data analysis and model development, and another for policy validation. The first framework utilizes various machine learning classifiers to identify key features impacting NO₂ levels, with an LSTM model predicting concentrations with 0.95 RMSE.

The second framework compares air quality scenarios with and without clean air zone implementation using historical data, with the LSTM model aiding in understanding intervention impact. The existing system Will not be scalable and time-consuming while performing experiments on a real-time dataset.. The existing system will underperform when the number of feature vectors for every data point exceeds the number of training samples and requires more training data to obtain satisfactory results.

CHAPTER 7

CONCLUSION AND FUTURE ENHANCEMENT

7.1 CONCLUSION

The contribution extends to the program facilitating data transformation, linear regression, mathematical model fitting, and the processing of prediction functions. It also enables the graphical presentation of results in the form of diagrams that illustrate the correlation between pollutants. This contributes to the evaluation of linear regression functions for their accuracy. It has been shown that the resulting functions enable predictions with high precision; the predicted values correlate very well with the obtained data. Accurate air quality forecasting has important theoretical and practical value for the public; without it, neither the government nor the public can effectively avoid the health damage caused by air pollution or improve the emergency response capability of heavy pollution days. In this study, we built regression models to predict air indicators based on machine learning algorithms. The experimental results show that both the SVR-based model and the RFR-based model can achieve good results, but the RFR model performs better in experiments. In addition, with the increasing number of samples, the time complexity of the SVR model increased cubically. Therefore, the SVR model is not suitable for processing a large number of samples. In summary, this establishes two prediction models based on different prediction scenarios, which improved the prediction accuracy of air indicators and provided guidance for modeling and analyzing urban air quality.

7.2 FUTURE ENHANCEMENTS

Future work includes comparing additional expandable graph learning models and exploring transfer learning and node alignment techniques to reduce re-training effort in industrial scenarios. Future work could be related to the utilization of the created R program to make predictions based on a wider set of parameters, larger data sets taken over longer periods of time, a diversity of monitoring locations, adaptation to other application domains, and improving the program to use other statistical methods supported by the Python language, with a special emphasis on supporting further chemical analysis of complex interactions and processes with gasses in the atmosphere.

CHAPTER 8

SOURCE CODE

8.1 SOURCE CODE

```
import warnings
warnings.filterwarnings('ignore')
import pandas as pd
import numpy as np
import seaborn as sns
import plotly.express as px
import matplotlib.pyplot as plt
import matplotlib
plt.style.use('fivethirtyeight')
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.svm import SVR
df = pd.read_csv("Dataset/AirQualityUCI.csv")
df.head().style.background_gradient(cmap='Blues').set_properties(**{'font-family': 'Segoe UI'})
df.sample(10).style.background_gradient(cmap='Spectral').set_properties(**{'font-family': 'Segoe UI'})
df.replace([np.inf, -np.inf], np.nan, inplace=True)
df.isnull().sum()
df.duplicated().sum()
df.info()
df.describe().style.background_gradient(cmap='RdYlGn').set_properties(**{'font-family': 'Segoe UI'})
CO_level_vc = df["CO_level"].value_counts()
CO_level_vc
```

```

px.bar(x=CO_level_vc.index, y=CO_level_vc.values, color=CO_level_vc.index,
       title="CO Level wise Count", labels={'x': "CO Level", 'y': "COunt"})
plt.figure(figsize=(25,10))
plt.xlabel("Temperature(°C)")
plt.ylabel('Relative Humidity')
plt.title("Relative Humidity vs Temperature(°C)")
plt.scatter(df['T'], df['RH'], marker='.', aa=True)
plt.figure(figsize=(25,10))
plt.xlabel("Temperature(°C)")
plt.ylabel('Absolute Humidity')
plt.title("Absolute Humidity vs Temperature(°C)")
plt.scatter(df['T'], df['AH'], marker='.', aa=True)
plt.figure(figsize=(10,5))
plt.scatter(x=df['Nox_GT'], y=df['NO2_GT'])
plt.show()
sns.set(style="whitegrid")
sns.pairplot(df, hue='CO_level', height=2.5)
plt.suptitle("Pair Plot of Air Quality Data", y=1.02, fontsize=16)
plt.show()
corr = df.corr(numeric_only=True)
plt.figure(figsize=(20,20))
sns.heatmap(corr, cbar=True, square=True, fmt='.1f', annot=True, annot_kws={'size':15}, cmap='Greens')
sns.pairplot(corr)
plt.figure(figsize=(24,12))
np.log(df['CO_GT']+1).plot.hist(bins=50, figsize=(8,4), edgecolor='white')
plt.title('CO_GT Distribution (log price +1)')
X = df.drop(['Date', 'Time', 'RH', 'CO_level'], axis=1)
y = df['RH']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)

```

```

for i in range(30,610,10):
reg=RandomForestRegressor(n_estimators=i,max_depth=5,max_features='sqrt',oob_score=True,random_s
tate=200)

    reg.fit(X_train,y_train)

    oob=reg.oob_score_

    print('For n_estimators = '+str(i))

    print('OOB score is '+str(oob))

    print('*****')
reg.score(X_test,y_test)
reg.oob_score_
reg.feature_importances_
imp_feat=pd.Series(reg.feature_importances_,index=X.columns.tolist())
imp_feat.sort_values(ascending=False)
feature_importances = pd.Series(reg.feature_importances_, index=X.columns)
feature_importances.plot(kind='barh')
plt.show()
y_pred = reg.predict(X_test)
def evaluate_metrics(y_test, prediction):
    print('MAE:', mean_absolute_error(y_test, prediction))
    print('MSE:', mean_squared_error(y_test, prediction))
    print('RMSE:', np.sqrt(mean_squared_error(y_test, prediction)))
    print('r1-Score:', r2_score(y_test, prediction))
evaluate_metrics(y_test, y_pred)
reg_results = pd.DataFrame(zip(y_test, y_pred, y_test - y_pred), columns = ['y_test', 'y_pred', 'error'])
reg_results.head(10)
reg_results['error'].hist()
plt.figure(figsize=(5,3))
sns.distplot(reg_results['error'])
plt.figure(figsize=(10,10))
x=np.linspace(0,5,5)

```

```

plt.plot(reg_results['y_test'], reg_results['y_pred'], 'b.')
plt.plot(x, x, 'r-')
plt.xlim(2.25,4)
plt.ylim(2.25,4)
plt.title("Results", fontsize=16)
plt.xlabel("Real", fontsize=14)
plt.ylabel("Predicted", fontsize=14)
plt.savefig("Results.png")
plt.show()

svr = SVR()
svr.fit(X_train,y_train)
svr.score(X_test,y_test)
y_pred = svr.predict(X_test)
evaluate_metrics(y_test, y_pred)
svr_results = pd.DataFrame(zip(y_test, y_pred, y_test - y_pred), columns = ['y_test', 'y_pred', 'error'])
svr_results.head(10)
svr_results['error'].hist()
plt.figure(figsize=(5,3))
sns.distplot(svr_results['error'])
plt.figure(figsize=(10,10))
x=np.linspace(0,5,5)
plt.plot(svr_results['y_test'], svr_results['y_pred'], 'b.')
plt.plot(x, x, 'r-')
plt.xlim(2.25,4)
plt.ylim(2.25,4)
plt.title("Results", fontsize=16)
plt.xlabel("Real", fontsize=14)
plt.ylabel("Predicted", fontsize=14)
plt.savefig("Results.png")
plt.show()

```


REFERENCES

- [1] Martha Arbayani Zaidan, Yuning Xie, Naser Hossein Motlagh, Bo Wang, Wei Nie, Petteri Nurmi, Sasu Tarkoma, Tuukka Petäjäjärvi, Aijun Ding, Markku Kulmala Dense Air Quality Sensor Networks: Validation, Analysis, and Benefits IEEE Sensors Journal, 2022
- [2] Rady Purbakawaca, Arief Sabdo Yuwono, I. Dewa Made Subrata, Supandi, Husin Alatas Ambient Air Monitoring System With Adaptive Performance Stability IEEE Access, 2022
- [3] Xiaoling Lin, Hongzhang Wang, Jing Guo, Gang Mei A Deep Learning Approach Using Graph Neural Networks for Anomaly Detection in Air Quality Data Considering Spatiotemporal Correlations IEEE Access, 2022
- [4] Yuting Yang, Gang Mei, Stefano Izzo Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning IEEE Access, 2022
- [5] Yangwen Yu, James J. Q. Yu, Victor O. K. Li, Jacqueline C. K. Lam A Novel Interpolation-SVT Approach for Recovering Missing Low-Rank Air Quality Data IEEE Access, 2020
- [6] Ying Zhang, Yanhao Wang, Minghe Gao, Qunfei Ma, Jing Zhao, Rongrong Zhang, Qingqing Wang, Linyan Huang A Predictive Data Feature Exploration-Based Air Quality Prediction Approach IEEE Access, 2019
- [7] Guyu Zhao, Guoyan Huang, Hongdou He, Qian Wang Innovative Spatial-Temporal Network Modeling and Analysis Method of Air Quality IEEE Access, 2019
- [8] Bo Liu, Shuo Yan, Jianqiang Li, Guangzhi Qu, Yong Li, Jianlei Lang, Rentao Gu A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction IEEE Access, 2019
- [9] Yuchao Zhou, Suparna De, Gideon Ewa, Charith Perera, Klaus Moessner Data-Driven Air Quality Characterization for Urban Environments: A Case Study IEEE Access, 2018
- [10] Farzaneh Farhadi, Roberto Palacin, Phil Blythe Machine Learning for Transport Policy Interventions on Air Quality IEEE Access, 2023
- [11] Abdelaziz El Fazziki, Djamal Benslimane, Abderrahmane Sadiq, Jamal Ouarzazi, Mohamed Sadgal An Agent Based Traffic Regulation System for the Roadside Air Quality Control IEEE Access, 2017

- [12] Clean Air Strategy, DEFRA, London, U.K, 2019.
- [13] T.-M. Chen, W. G. Kuschner, J. Gokhale and S. Shofer, Outdoor air pollution: Nitrogen dioxide sulfur dioxide and carbon monoxide health effects, *Amer. J. Med. Sci.*, vol. 333, no. 4, pp. 249-256, Apr. 2007.
- [14] C. Holman, R. Harrison and X. Querol, Review of the efficacy of low emission zones to improve urban air quality in European cities, *Atmos. Environ.*, vol. 111, pp. 161-169, Jun. 2015.
- [15] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu and C. Chen, Data-driven intelligent transportation systems: A survey, *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624-1639, Dec. 2011.
- [16] A. I. Torre-Bastida, J. Del Ser, I. La na, M. Ilardia, M. N. Bilbao and S. Campos-Cordoba, Big data for transportation and mobility: Recent advances trends and challenges, *IET Intell. Transp. Syst.*, vol. 12, no. 8, pp. 742-755, Oct. 2018.
- [17] M. A. Olvera-García, J. J. Carbajal-Hernández, L. P. Sánchez-Fernández and I. Hernández-Bautista, Air quality assessment using a weighted fuzzy inference system, *Ecol. Inf.*, vol. 33, pp. 57-74, May. 2016.
- [18] L. Pan, B. Sun and W. Wang, City air quality forecasting and impact factors analysis based on grey model, *Proc. Procedia Eng.*, pp. 74-79, 2011.
- [19] Y. Zheng, F. Liu and H. P. Hsieh, U-Air: When urban air quality inference meets big data, *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 1436-1444, 2013.
- [20] Y. Zheng et al., Forecasting fine-grained air quality based on big data, *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 2267-2276, 2015.
- [21] J. J. Carbajal-Hernández, L. P. Sánchez-Fernández, J. A. Carrasco-Ochoa and J. F. Martínez-Trinidad, W. J. Requia, M. D. Adams, A. Arain, S. Papatheodorou, P. Koutrakis and M. Mahmoud, Global association of air pollution and cardiorespiratory diseases: A systematic review meta-analysis and investigation of modifier variables, *Amer. J. Public Health*, vol. 108, no. S2, pp. S123-S130, Apr. 2018.
- [22] C. J. L. Murray et al., Global burden of 87 risk factors in 204 countries and territories 1990–2019: A systematic analysis for the global burden of disease study 2019, *Lancet*, vol. 396, no. 10258, pp. 1223-1249, 2020.


- [23]J. Zhang and D. Tao, Empowering things with intelligence: A survey of the progress challenges and opportunities in artificial intelligence of things, *IEEE Internet Things J.*, vol. 8, no. 10, pp. 7789-7817, May 2021.
- [24]L. Kong et al., A 6-year-long (2013â€“2018) high-resolution air quality reanalysis dataset in China based on the assimilation of surface observations from CNEMC, *Earth Syst. Sci. Data*, vol. 13, no. 2, pp. 529-570, 2021.
- [25]T. Li, G. Kou, Y. Peng and P. S. Yu, An integrated cluster detection optimization and interpretation approach for financial data, *IEEE Trans. Cybern.*, Sep. 2021.
- [26]T. S. Rajput and N. Sharma, Multivariate regression analysis of air quality index for Hyderabad city: Forecasting model with hourly frequency, *Int. J. Appl. Res.*, vol. 3, pp. 443-447, 2017.
- [27]S. Dreiseitl and L. Ohno-Machado, Logistic regression and artificial neural network classification models: A methodology review, *J. Biomed. Inform.*, vol. 35, no. 5, pp. 352-359, 2002.
- [28]A. Azid et al., Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia, *Water Air Soil Pollution*, vol. 225, no. 8, pp. 2063, 2014.
- [29]S. De Vito et al., Dynamic multivariate regression for on-field calibration of high speed air quality chemical multi-sensor systems, *Proc. AISEM Annu. Conf.*, pp. 1-3, Feb. 2015.
- [30]Z. Kang and Z. Qu, Application of BP neural network optimized by genetic simulated annealing algorithm to prediction of air quality index in Lanzhou, *Proc. IEEE Comput. Intell. Appl. (ICCIA)*, pp. 155-160, Sep. 2017.
- [31]World Health Statistics 2019: Monitoring Health for the SDGs Sustainable Development Goals, Geneva, Switzerland, 2021.
- [32]A. A. Almetwally, M. Bin-Jumah and A. A. Allam, Ambient air pollution and its influence on human health and welfare: An overview, *Environ. Sci. Pollut. Res.*, vol. 27, no. 20, pp. 24815-24830, Jul. 2020.
- [33]Y. Zhu, J. Xie, F. Huang and L. Cao, Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China, *Sci. Total Environ.*, vol. 727, Jul. 2020.
- [34]F. Chen and Z. Chen, Cost of economic growth: Air pollution and health expenditure, *Sci. Total Environ.*, vol. 755, Feb. 2021.

- [35] Z. Zhou, Z. Ye, Y. Liu, F. Liu, Y. Tao and W. Su, Visual analytics for spatial clusters of air-quality data, *IEEE Comput. Graph. Appl.*, vol. 37, no. 5, pp. 98-105, May 2017.
- [36] P. K. Hopke, S. S. Hashemi Nazari, M. Hadei, M. Yarahmadi, M. Kermani, E. Yarahmadi, et al., Spatial and temporal trends of short-term health impacts of PM 2.5 in Iranian cities; a modelling approach (2013â€“2016) ,
- [37] *Aerosol Air Quality Res.*, vol. 18, no. 2, pp. 497-504, 2018.
- [38] M. Iriti, P. Piscitelli, E. Missoni and A. Miani, Air pollution and health: The need for a medical reading of environmental monitoring data, *Int. J. Environ. Res. Public Health*, vol. 17, no. 7, pp. 2174, Mar. 2020.
- [39] M. Kowalska, M. Skrzypek, M. Kowalski and J. Cyrus, Effect of NO x and NO 2 concentration increase in ambient air to daily bronchitis and asthma exacerbation Silesian Voivodeship in Poland , *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, pp. 1-9, 2020.
- [40] S. Ahmed, Air pollution and its impact on agricultural crops in developing countriesâ€”A review, *J. Anim. Plant Sci.*, vol. 25, no. 3, pp. 297-302, 2015.
- [41] P. Rafaj, G. Kieseewetter, W. SchÃ¶pp, J. Cofala, Z. Klimont, P. Purohit, et al., Outlook for clean air in the context of sustainable development goals, *Global Environ. Change*, vol. 53, pp. 1-11, Nov. 2018.
- [42] J. R. Wolch, J. Byrne and J. P. Newell, Urban green space public health and environmental justice: The challenge of making cities â€˜just green enough, *Landscape Urban Planning*, vol. 125, pp. 234-244, May 2014.
- [43] H. Chen, J. C. Kwong, R. Copes, K. Tu, P. J. Villeneuve, A. van Donkelaar, et al., Living near major roads and the incidence of dementia Parkinsonâ€™s disease and multiple sclerosis: A population-based cohort study, *Lancet*, vol. 389, pp. 718-726, Feb. 2017.
- [44] A. J. Cohen et al., Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the global burden of diseases study 2015, *Lancet*, vol. 389, no. 10082, pp. 1907-1918, May 2017.
- [45] S. Liu, Y. Zhou, S. Liu, X. Chen, W. Zou, D. Zhao, et al., Association between exposure to ambient particulate matter and chronic obstructive pulmonary disease: Results from a cross-sectional study in China, *Thorax*, vol. 72, no. 9, pp. 788-795, Sep. 2017.

- [46] T. Li, R. Hu, Z. Chen, Q. Li, S. Huang, Z. Zhu, et al., Fine particulate matter (PM_{2.5}): The culprit for chronic lung diseases in China, *Chronic Diseases Transl. Med.*, vol. 4, no. 3, pp. 176-186, Sep. 2018.
- [47] X. Li, L. Peng, Y. Hu, J. Shao and T. Chi, Deep learning architecture for air quality predictions, *Environ. Sci. Pollut. Res.*, vol. 23, pp. 22408-22417, 2016.
- [48] Q. Zhou, H. Jiang, J. Wang and J. Zhou, A hybrid model for PM 2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network , *Sci. Total Environ.*, vol. 496, pp. 264-274, Oct. 2014.
- [49] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Comput.*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [50] A. C. Cosma and R. Simha, Machine learning method for real-time non-invasive prediction of individual thermal preference in transient conditions, *Building Environ.*, vol. 148, pp. 372-383, Jan. 2019.
- [51] D. Zhu, C. Cai, T. Yang and X. Zhou, A machine learning approach for air quality prediction: Model regularization and optimization, *Big Data Cogn. Comput.*, vol. 2, no. 1, pp. 5, 2018.
- [52] J. Y. Zhu, C. Sun and V. O. K. Li, An extended spatio-temporal Granger causality model for air quality estimation with heterogeneous urban big data, *IEEE Trans. Big Data*, vol. 3, pp. 307-319, Sep. 2017.

APPENDICES

A. PLAGIARISM REPORT



Similarity Report ID: oid:3618:58942223

● 8% Overall Similarity

Top sources found in the following databases:

- 5% Internet database
- 6% Publications database
- Crossref database
- Crossref Posted Content database
- 3% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	mdpi.com Internet	2%
2	mdpi-res.com Internet	1%
3	Ahmad Zia Ul-Saufie, Nurul Haziqah Hamzan, Zulaika Zahari, Wan Nur ... Crossref	<1%
4	Navneet Kumar, Anirban Middey. "Extreme climate index estimation an... Crossref	<1%
5	ncbi.nlm.nih.gov Internet	<1%
6	gmd.copernicus.org Internet	<1%
7	Mohammad Alamgeer, Nuha Alruwais, Haya Mesfer Alshahrani, Abdull... Crossref	<1%
8	hdl.handle.net Internet	<1%

B. PROOF OF PUBLICATION/PATENT FILED/ CONFERENCE CERTIFICATE

Dear Singa selvamani ,

This is to inform you that your paper has been accepted for publication in IJRRP, which will be published in current issue. The Unique ID of your paper is IJRRP-23696.

Paper Title: SCALABLE APPROACH FOR DETECTING AIR QUALITY INFERENCE USING ENSEMBLE REGRESSORS

You are advised to complete the process for the publication of your research paper.

1. Fill & Send Copyright Transfer Form (download from our website <https://www.ijrrp.com/download/COPYRIGHT-FORM.pdf>)

(Note: Take a print of copyright form, fill it, scan it & send us via mail to ijrrp4@gmail.com or send it in word format with digital scanned signatures. Do not forget to mention your paper ID in subject of mail.)

Alternatively you can submit the Online Copyright Form by clicking the following link

<https://forms.gle/S54HMLvERYgmms5y8>

2. Submit your Publication fee

(Note: Send the Payment receipt in image/word/pdf format along with copyright form to ijrrp4@gmail.com Do not forget to mention your paper ID in subject of mail.)

3. Paper will be published within 24 to 36 working Hours after the completion of step 1 & 2. We will provide the Soft Individual copy of the certificates to all authors of a research paper.

Do not forget to mention your paper ID in subject.

Publication Fee:

International Authors	14 US Dollars (Click Here to Pay)
Indian Authors (Author Affiliation in Paper Must be in Indian Territory)	Rs. 499 (Click Here to Pay)
E-Certificate	Free

Publication Charge includes:

- Publication of one entire research paper/Online
- Individual Soft Copy of Certificate to all author of paper.
- Editorial Fee
- Indexing ,maintenance of link, resolvers and journal infrastructures.

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University u / s 3 of UGC Act, 1956)

Office of Controller of Examinations

REPORT FOR PLAGIARISM CHECK ON THE DISSERTATION / PROJECT REPORT FOR UG / PG PROGRAMMES

(To be attached in the dissertation/project report)

1	Name of the Candidate (IN BLOCK LETTERS)	SINGA SELVAMANI S VIMAL ADITYA RAJ SAMARENDRA T
2	Address of Candidate	Bharathi Salai, Ramapuram, Chennai-89. Mobile Number: 9384620397, 8778085184, 6369611838
3	Registration Number	RA2011003020212, RA2011003020216, RA2011003020206
4	Date of Birth	21/09/2002, 03/05/2003, 02/06/2003
5	Department	Computer Science and Engineering
6	Faculty of Engineering and technology	Mrs. Preethy Jemima P
7	Title of the Dissertation / Project	Scalable Approach For Detecting Air Quality Inference using Ensemble Regressors
8	Whether the above project/dissertation is done by	<p>Individual or group : Group (Strike whichever is not applicable)</p> <p>a) If the project / dissertation is done in group, then how many students together completed the project: 3</p> <p>b) Mention the Name & Registernumber of other candidates : SINGA SELVAMANI S[RA2011003020212] VIMAL ADITYA RAJ[RA2011003020216] SAMARENDRA T[RA2011003020206]</p>
9	Name and address of the Supervisor / Guide	<p>Mrs. Preethy Jemima P, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram Campus, Chennai 89</p> <p>Mail ID : preethyj@srmist.edu.in</p> <p>Mobile Number : 9944016059</p>

10	Name and address of the Co-Supervisor/Guide	NA Mail id:NA Mobile Number:NA		
11	Software Used	Turnitin		
12	Date Of Verification	08/05/2024		
13	Plagiarism Details : (to attach the final report from the software)			
Chapter	Title of the Report	Percentage of similarity index (including self citation)	Percentage of similarity index (Excluding self citation)	% of plagiarism after excluding Quotes, Bibliography, etc.,
1	Scalable Approach For Detecting Air Quality Inference using Ensemble Regressors	8%	8%	8%
Appendices		NA	NA	NA
I / We declare that the above information have been verified and found true to the best of my / our knowledge.				
Signature of the Candidate		Name & Signature of the Staff (Who uses the plagiarism Check Software)		
Name & Signature of the Supervisor/Guide		Name & Signature of the Co-Supervisor/Co-Guide		
Dr. K. Raja Name & Signature of the HOD				

