**Title Page :**

# PREDICTION OF AIR QUALITY INDEX USING MACHINE LEARNING TECHNIQUES WITH RANDOM FOREST REGRESSION OVER SUPPORT VECTOR REGRESSION

K.Pavani[1] , Dr.C.Anitha[2]

K.Pavani[1]

Research Scholar,

Department of Computer Science and Engineering,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105

pavanik1946.sse@saveetha.com

Dr.C.Anitha[2]

Project Guide, Corresponding Author,

Department of cloud computing,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

SaveethaUniversity,Chennai,Tamilnadu,India, Pincode: 602105.

anitha.sse@saveetha.com

**Keywords:**Air Quality Index (AQI),Random Forest Regression,Support Vector Regression,Prediction,Machine Learning,Air Pollution,

## ABSTRACT

**Aim:** The main objective of the research work is to predict the AQI value by machine learning algorithm and get the high rate of accuracy value using Random Forest Regression algorithm compared with Support Vector Regression algorithm. **Materials and methods:** By using algorithms of Random Forest Regression and the Support Vector Regression algorithm with the sample size of 10 each and the Google colab is used. Two sample groups tested with 20 samples, which is tested at G power as 85% with t-test analysis with the significance value of 0.000. **Results:** By using the dataset the sample size is N=10 and test results prove that the Random Forest Regression algorithm has an average accuracy of 94.00%, which seems to be better than the Support Vector Regression algorithm algorithm accuracy of 86.50%. The significance value was 0.000 ($p<0.05$) showing that there is a statistically significant difference between the two algorithms. **Conclusion:** As a result, it is discovered that there is a significant difference for the accuracy using Random Forest Regression algorithm over Support Vector Regression algorithm.



**Keywords:**Air Quality Index (AQI),Random Forest Regression,Support Vector Regression,Prediction,Machine Learning,Air Pollution.


## INTRODUCTION

One of the biggest environmental problems today is air pollution, which endangers ecosystems, human health, and the sustainability of the planet(Sankar 2023). Due to the quickening rates of urbanization, industrialization, and economic growth, pollutants are being released into the atmosphere at alarming rates, which is having a negative influence on air quality all around the world. Air pollution affects every location equally, from crowded urban centers to isolated rural areas, which is why it's a critical worldwide issue(Morapedi and Obagbuwa 2023).The effects of bad air quality are many and varied, impacting not just human health but also the integrity of ecosystems, the productivity of agriculture, and economic success**.**Particulate matter (PM), sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), and carbon monoxide (CO) are examples of air pollutants that can cause a variety of health issues(Zukaib et al. 2023). These include respiratory conditions, heart issues, and even early mortality. The detrimental impacts of air pollution can exacerbate pre-existing health disparities and inequalities in vulnerable populations, including children, the elderly, and people with pre-existing medical illnesses.Moreover, air pollution has a significant impact on ecosystem health and environmental sustainability(Ravindiran et al. 2023). Smog, acid rain, and ground-level ozone are created when pollutants are released into the atmosphere. These phenomena can deteriorate soil quality, taint water sources, and endanger plant and animal species. The effects of air pollution pose a threat to the vital ecosystem services that ecosystems like forests, wetlands, and aquatic habitats provide, such as carbon sequestration, water purification, and habitat provision(Chatterjee 2022).

Apart from its adverse effects on the environment and human health, air pollution has a substantial financial cost to society. Every year, the financial burden that air pollution places on governments, corporations, and individuals alike amounts to billions of dollars due to medical expenses, lost productivity, agricultural loss, and property damage. Furthermore, marginalized and underprivileged populations frequently bear a disproportionate share of the economic costs associated with air pollution, which exacerbates social injustices and feeds the cycle of poverty and suffering(Jitkajornwanich et al. 2024).Due to the complexity of the air pollution issue, creative solutions that can efficiently monitor, reduce, and control air quality on a local, regional, and worldwide basis are desperately needed. While they offer useful data, traditional methods of monitoring air quality such as satellite-based remote sensing and fixed monitoring stations have limitations in terms of their predictive power, temporal resolution, and spatial coverage(Wang, Ren, and Xia 2023). To get around these, scientists and decision-makers have been using more sophisticated computational methods, especially machine learning, to create prediction models that can predict air quality metrics more precisely and accurately.As a branch of artificial intelligence, machine learning includes a wide range of computational methods and algorithms that let computers recognize patterns in data, learn from them, and make judgment calls or predictions without the need for explicit programming.Machine learning algorithms can extract complicated associations and correlations that traditional statistical methods would miss by using massive datasets of historical air quality measurements, meteorological observations, land use data, and other pertinent variables. This makes it possible to create prediction models that more precisely and consistently forecast air quality characteristics like pollution concentrations or air quality indices(Kajewska-Szkudlarek 2023).

Machine learning has become an effective tool for predicting air quality in recent years, and it has the potential to completely change how we monitor and control air quality in industrial and urban settings. Machine learning algorithms can detect subtle patterns and trends in the dynamics of air quality by evaluating large amounts of heterogeneous data from many sources. This allows for the early identification of pollution hotspots, the identification of emission sources, and the forecasting of future air quality conditions.In this work, we concentrate on the machine learning prediction of the Air Quality Index (AQI), a composite measure that encompasses several air contaminants and the health concerns associated with them. We specifically investigate the predictive power of two popular regression techniques, Support Vector Regression (SVR) and Random Forest Regression (RFR), in AQI level forecasting, given a wide range of input factors(Saminathan and Malathy 2023). Pollutant concentrations, weather patterns, topographical characteristics, and other pertinent elements that affect the dynamics of air quality may be among these variables.The decision to look into RFR and SVR was made because of their dependability, adaptability, and suitability for working with complicated, high-dimensional datasets that are frequently encountered in air quality prediction assignments. Modeling the intricate dynamics of air pollution is a good fit for RFR, an ensemble learning technique based on decision tree algorithms, since it is excellent at capturing nonlinear correlations and interactions among input variables(Saminathan and Malathy 2023). Conversely, SVR is a potent regression method that

excels at managing high-dimensional data, preventing overfitting, and making good generalizations to previously unseen data

By comparing and contrasting RFR and SVR, we hope to determine which method is best for air quality level forecasting and assess each method's advantages and disadvantages when it comes to AQI prediction. Using real-world air quality data and performance metrics like coefficient of determination, mean absolute error, and root mean square error, we aim to evaluate each algorithm's predictive accuracy, robustness, and generalization abilities across a range of environmental factors and spatiotemporal scales.The knowledge gathered from this research could help with evidence-based policy development and decision-making that aims to reduce air pollution, safeguard public health, and advance sustainable development.Through the advancement of our comprehension of the intricate relationships governing the dynamics of air quality and the improvement of our forecasting skills, we can enable stakeholders and policymakers to proactively tackle air pollution and its consequences on human well-being, ecological systems, and the financial sector.In conclusion, this work is a vital step toward using machine learning to address the worldwide problem of air pollution(Zukaib et al. 2023). We can facilitate the development of more efficient methods for monitoring and reducing air pollution on a local, regional, and global scale by creating predictive models for air quality forecasts that are accurate and transparent. We can work toward a future where everyone has access to clean, healthy air through interdisciplinary collaboration and creative research methodologies, ensuring the prosperity and well-being of current and future generations.

## MATERIALS AND METHODS

The study of the research was conducted in the Programming lab, Saveetha School of Engineering, SIMATS. There are two groups identified. Group 1 Random Forest Regression Algorithm and Group 2 isSupport Vector Regression Algorithm. It has the good layout structure of SSE. The sample size of (N=10) and calculated from the SPSS analysis is carried out with the level of significance 0.000 ($p<0.05$), the dataset is collected from Kaggle.com.

**Random Forest Regression Algorithm**

By building an ensemble of decision trees, Random Forest Regression is a potent machine learning technique that may be used to predict continuous variables. To increase the precision and resilience of the model in the context of regression analysis, Random Forest Regression aggregates the predictions of several decision trees. The steps for putting Random Forest Regression into practice are as follows:

Step 1: Assemble the target variable and input characteristics for the training dataset.

Step 2: For every tree, choose a subset of data and features at random.

Step 3: Use feature randomness and bootstrapping to create several decision trees.

Step 4: To create an ensemble prediction, combine the predictions from each tree.

Step 5: To assess the performance of the model, compute the mean squared error or another suitable measure.
Step 6: Use cross-validation to fine-tune hyperparameters like the number of trees and maximum depth.
Step 7: Use a test dataset to validate the model's performance.
Step 8: Implement a model to forecast fresh data.

**Support Vector Regression Algorithm**

A machine learning approach called Support Vector Regression (SVR) uses the concepts of Support Vector Machines (SVM) to do regression tasks. To reduce the discrepancy between expected and actual values, SVR fits a hyperplane in a higher-dimensional space. The use of support vectors, which determines the ideal location for the hyperplane, is crucial to SVR. During training, SVR uses a kernel function to alter input information in an effort to reduce error and model complexity. It works well for identifying non-linear relationships and producing precise forecasts.Because SVR is flexible and resilient in regression modeling, it finds use in environmental science, engineering, and finance.All things considered, SVR provides a stable and adaptable method for regression modeling, able to identify intricate patterns in the data and produce precise predictions even when noise and anomalies are present. Regression tasks in both academic and industry contexts have found SVR to be a popular choice due to its effectiveness, interpretability, and ease of use.
Step 1: Assemble the target variable and input characteristics for the training dataset.
Step 2: To guarantee uniform scaling, standardize or normalize characteristics.
Step 3: Choose the proper kernel parameters and kernel function (such as a linear, polynomial, or radial basis function).
Step 4: Use training data to train the SVR model.
Step 5: Use cross-validation to optimize hyperparameters like epsilon and C (regularization parameter).
Step 6: Use a test dataset to validate the model's performance.
Step 7: Conduct analysis and interpretation of the model.
Step 8: Implement a model to forecast fresh data.

Google colab is a free and open-source distribution of the Python and R programming languages for scientific computing, data science, and machine learning. It includes a wide range of tools and libraries for data manipulation, analysis, and visualization, as well as tools for building and deploying machine learning models.

**Statistical Analysis**

Using the SPSS statistical package, the analysis of mean accuracy forAQI Value using Random Forest Regression algorithm and Support Vector Regression algorithm was carried out by

applying an independent sample t-test to obtain the accuracy of 94.0%. The speed and file size are independent variables and type of the file is dependent variables.

## RESULTS

One of the most important metrics for determining the present or expected levels of air pollution is the Air Quality Index (AQI). It allocates a number that represents the degree of pollution, ranging from safe to dangerous. A higher AQI value indicates increased pollution levels, which can be harmful to health, particularly for older adults, children, and people with respiratory conditions. It is imperative to monitor AQI levels with vigilance, particularly during periods of elevated pollution. Protecting the public's health necessitates taking proactive steps to reduce exposure, such as reducing outdoor activities and pollution exposure.

Table 1 shows both algorithms and it represents the values of the Random Forest Regression algorithm and the Support Vector Regression algorithm that produces the outcome as the exact rate Accuracy.

In Table 2, the values for Random Forest Regression include N (number of observations) of 10, a mean of 89.70, a standard deviation of 2.04, and a standard error mean of 0.6. The values for Support Vector Regression in the same table include N of 10, a mean of 85.4, a standard deviation of 0.96, and a standard error mean of 0.3.

Table 3 displays the results of a T-Test conducted on two statistically independent samples. The mean difference between the Random Forest Regression and Support Vector Regression groups is 4.3, with a standard error difference of 0.7. The significance value is smaller than 0.05, indicating that there is statistically significant difference between the two groups and thus improving the performance with high energy efficiency.

Figure 1 represents the difference between the proposed algorithm and the comparison algorithm; they are the Random Forest Regression algorithm and the Support Vector Regression algorithm . The difference is shown in the bar graphs that shows the higher value of the proposed algorithm and it also shows the exact value of mean accuracy.

## DISCUSSION

In this study of enhancing the AQI, the Random Forest Regression algorithm has accuracy (94.00) and the accuracy of the Support Vector Regression algorithm is 86.50. Thus there is no statistical difference between the accuracy of the Random Forest Regression algorithm and that with the Support Vector Regression algorithm with the significance value of 0.000 ($p<0.05$) and statistical analysis tool SPSS has been utilized to run the independent sample t-tests on the data set.

For the purpose of environmental management and public health, the air quality index (AQI) must be predicted. Using machine learning techniques such as Support Vector Regression (SVR) and Random Forest Regression (RFR) has the potential to improve prediction accuracy(Morapedi and Obagbuwa 2023). RFR's ensemble of decision trees shows promise in managing a wide range of datasets that include different environmental elements that have an impact on air quality(Boudreault, Campagna, and Chebana 2024). It is a desirable option for AQI prediction tasks due to its ease of use and capacity for handling huge datasets. Furthermore, RFR models provide interpretability by emphasizing the significance of various variables in AQI level determination, assisting stakeholders in comprehending the underlying dynamics of air quality(Wang, Ren, and Xia 2023).

SVR, on the other hand, offers a strong substitute, especially in situations where there are intricate and non-linear correlations between input features and AQI. Through the use of kernel functions to translate data into higher-dimensional spaces, SVR is able to catch complex patterns that may be difficult for linear models to identify.In actuality, the decision between RFR and SVR for AQI prediction depends on a number of variables(Saminathan and Malathy 2023), such as the properties of the dataset, the need for interpretability, the limitations of computing, and the intended prediction performance(Chiu et al. 2023). Furthermore, continual research and experimentation are necessary to continuously improve the efficacy of machine learning techniques in predicting AQI and supporting well-informed decision-making for environmental management, given the dynamic nature of environmental data and the evolving understanding of air quality dynamics.

## CONCLUSION

The Random Forest Regression algorithm has an accuracy of 94.00 and the Support Vector Regression algorithm’s accuracy is 86.50 which shows that the accuracy of the Random Forest Regression has no significant difference with the Support Vector Regression algorithm.

## DECLARATIONS

### Conflict of Interests

No conflict of interest in this manuscript.

### Author Contribution

Author VBR was involved in data collection, data analysis, manuscript writing. Author TPA was involved in conceptualization, data validation and critical review of manuscript.

### Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (formerly known as Saveetha University) for providing necessary infrastructure to carry out this work successfully.

**Funding**

We thank the following organizations for providing financial support that enabled us to complete this study:

1. VRT Techno Solutions Pvt. Ltd.
2. Saveetha School of Engineering.
3. Saveetha University
4. Saveetha Institute of Medical Technical Sciences.

## REFERENCES

Sankar, Ganesh S. 2023. *Air Quality Index Prediction Using Machine Learning Techniques*. Independent Author.

Morapedi, Tshepang Duncan, and Ibidun Christiana Obagbuwa. 2023. "Air Pollution Particulate Matter (PM2.5) Prediction in South African Cities Using Machine Learning Techniques." *Frontiers in Artificial Intelligence* 6 (October): 1230087.

Zukaib, Umer, Mohammed Maray, Saad Mustafa, Nuhman Ul Haq, Atta Ur Rehman Khan, and Faisal Rehman. 2023. "Impact of COVID-19 Lockdown on Air Quality Analyzed through Machine Learning Techniques." *PeerJ. Computer Science* 9 (March): e1270.

Ravindiran, Gokulan, Gasim Hayder, Karthick Kanagarathinam, Avinash Alagumalai, and Christian Sonne. 2023. "Air Quality Prediction by Machine Learning Models: A Predictive Study on the Indian Coastal City of Visakhapatnam." *Chemosphere* 338 (October): 139518.

Chatterjee, Aditya. 2022. *Predict Air Pollution Level Using Machine Learning: Get Practical Hands-on Experience*.

Jitkajornwanich, Kulsawasd, Nattadet Vijaranakul, Saichon Jaiyen, Panu Srestasathiern, and Siam Lawawirojwong. 2024. "Enhancing Risk Communication and Environmental Crisis Management through Satellite Imagery and AI for Air Quality Index Estimation."

*MethodsX* 12 (June): 102611.

Wang, Siyuan, Ying Ren, and Bisheng Xia. 2023. “Estimation of Urban AQI Based on Interpretable Machine Learning.” *Environmental Science and Pollution Research International* 30 (42): 96562–74.

Kajewska-Szkudlarek, Joanna. 2023. “Predictive Modelling of Heating and Cooling Degree Hour Indexes for Residential Buildings Based on Outdoor Air Temperature Variability.” *Scientific Reports* 13 (1): 17411.

Saminathan, S., and C. Malathy. 2023. “Ensemble-Based Classification Approach for PM2.5 Concentration Forecasting Using Meteorological Data.” *Frontiers in Big Data* 6 (June): 1175259.

Zukaib, Umer, Mohammed Maray, Saad Mustafa, Nuhman Ul Haq, Atta Ur Rehman Khan, and Faisal Rehman. 2023. “Impact of COVID-19 Lockdown on Air Quality Analyzed through Machine Learning Techniques.” *PeerJ. Computer Science* 9 (March): e1270.

Morapedi, Tshepang Duncan, and Ibidun Christiana Obagbuwa. 2023. “Air Pollution Particulate Matter (PM2.5) Prediction in South African Cities Using Machine Learning Techniques.” *Frontiers in Artificial Intelligence* 6 (October): 1230087.

Boudreault, Jérémie, Céline Campagna, and Fateh Chebana. 2024. “Revisiting the Importance of Temperature, Weather and Air Pollution Variables in Heat-Mortality Relationships with Machine Learning.” *Environmental Science and Pollution Research International* 31 (9): 14059–70.

Wang, Siyuan, Ying Ren, and Bisheng Xia. 2023. “Estimation of Urban AQI Based on Interpretable Machine Learning.” *Environmental Science and Pollution Research International* 30 (42): 96562–74.

Saminathan, S., and C. Malathy. 2023. “Ensemble-Based Classification Approach for PM2.5 Concentration Forecasting Using Meteorological Data.” *Frontiers in Big Data* 6 (June): 1175259.

Chiu, Yueh-Hsiu Mathilda, Ander Wilson, Hsiao-Hsien Leon Hsu, Harris Jamal, Nicole Mathews, Itai Kloog, Joel Schwartz, et al. 2023. “Prenatal Ambient Air Pollutant Mixture Exposure and Neurodevelopment in Urban Children in the Northeastern United States.” *Environmental Research* 233 (September): 116394.

## TABLES & FIGURES

**Table 1.** Shows the comparison of accuracy rate by using Random Forest Regression Algorithm and Support Vector Regression Algorithm it shows the both values and the rate of Accuracy (94.00).

| S.NO | RANDOM FOREST REGRESSION ACCURACY (%) | SUPPORT VECTOR REGRESSION ACCURACY (%) |
|---|---|---|
| 1 | 90.45 | 82.60 |
| 2 | 93.97 | 85.24 |
| 3 | 95.35 | 86.01 |
| 4 | 90.21 | 80.00 |
| 5 | 97.00 | 87.01 |
| 6 | 92.81 | 80.54 |
| 7 | 92.00 | 85.00 |
| 8 | 91.00 | 82.10 |
| 9 | 94.80 | 87.65 |
| 10 | 94.00 | 86.00 |

**Table 2.** Shows Statistical Analysis values of Mean accuracy (89.70), Standard Deviation(2.045), and Standard error deviation(0.645). The Random Forest Regression Algorithm and the Support Vector Machine algorithm have the values of the Mean accuracy, Standard Deviation, and Standard Error .

| Algorithm | N | Mean | std.deviation | Std.error mean |
|---|---|---|---|---|
| Accuracy | | | | |
| Random Forest Regression Algorithm | 10 | 89.7 | 2.04 | 0.64 |
| Support Vector Regression Algorithm | 10 | 85.4 | 0.9 | 0.30 |

**Table 3.** Shows Comparison of Significance Level with value $p<0.05$. Both Random Forest Regression Algorithm and the Support Vector Regression Algorithm have a confidence interval of 95% with the significance value 0.000( $p<0.05$).

| | | F | Sig | T | Dif | Sig( 2-Tailed | Mean Difference | Std.Error Difference | Lower | Upper |
|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy | Equal Variance Assumed | 2.99 | 0.101 | 6.087 | 18 | 0.000 | 4.34800 | 0.71433 | 2.84724 | 5.84876 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Equal Variance Not Assumed | | | 6.087 | 12.858 | 0.000 | 4.34800 | 0.71433 | 2.84724 | 5.84876 |

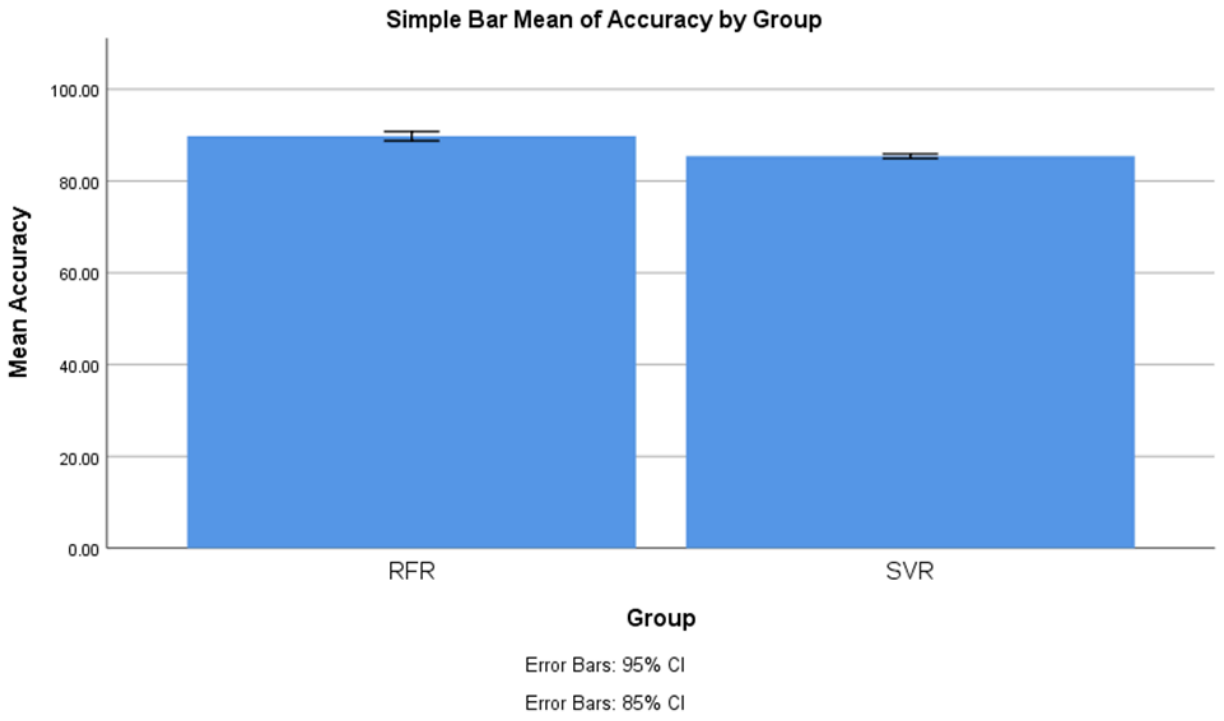


**Fig. 1** Comparison of the Random Forest Regression Algorithm accuracy of (94.00) and it has the mean accuracy of the Support Vector Regression Algorithm (86.50) The mean accuracy of the Random Forest Regression Algorithm has significant difference with the Support Vector Regression Algorithm with the significance value is 0.000 ($p<0.05$) . X Axis: Random Forest Regression Algorithm vs Support Vector Regression Algorithm Y Axis: Mean accuracy ± 2 SD.