IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Analysing Regression Techniques for Air Quality Forecasting

Chinta Venkata Murali Krishna[1], Gudiseva Naga mery [2], Peddi Harini [3]
Nalliboina Mahesh Babu[4] , Patapanchala Venkata Mohan [5]
Associate professor, Dept. Of CSE-DS, NRI *Institute of Technology, A.P-521212.*
Dept. Of CSE-DS, NRI *Institute of Technology, A.P-521212.*
Dept. Of CSE-DS, NRI *Institute of Technology, A.P-521212.*
Dept. Of CSE-DS, NRI *Institute of Technology, A.P-521212.*
Dept. Of CSE-DS, NRI *Institute of Technology, A.P-521212.*

ABSTRACT - *In order to mitigate the negative consequences of air pollution on public health, this study looks into predictive modelling for air quality forecasting. Using a comparative analysis technique, multiple regression algorithms are examined to determine the best model for predicting air quality levels. This work attempts to identify the specific strengths and drawbacks of linear regression, ridge regression, lasso regression, and support vector regression (SVR) in capturing the complex dynamics of air quality changes.*

This study ensures the robustness and reliability of its analysis by utilizing a complete dataset gathered from credible sources, which includes meteorological parameters, pollutant concentrations, and historical air quality indices. Each regression model is thoroughly trained, tested, and evaluated using standard performance measures such as mean squared error, root mean squared error, and R-squared.

The study's findings not only shed light on the forecasting capacities of various regression methods, but also provide vital insights into the fundamental causes of air quality fluctuations. With this knowledge, policymakers, urban planners, and environmental stakeholders may make more informed decisions and conduct targeted interventions to reduce air pollution and protect public health. Finally, this research helps to current efforts to control air pollution and promote sustainable development in urban contexts.

KEYWORDS: *Regression analysis, Air quality forecasting, Comparative study, Predictive modelling, Environmental health.*

1. INTRODUCTION

Air pollution is a major hazard to human health and the environment, especially in urban areas. With the enormous urbanization and industrialization that has occurred in recent decades, the issue of air quality has become more important. Poor air quality can cause a wide range of health concerns, including respiratory ailments, cardiovascular disorders, and even early mortality.

To solve this significant issue, reliable forecasting of the environment is required. Predictive modelling approaches, such as regression analysis, provide useful tools for analysing and predicting air quality. Regression models can provide insights into the intricate nature of air quality by studying different aspects that contribute to it, such as meteorological conditions, pollutant emissions, and geographic features.

In this paper, we undertake a comparison analysis of various regression techniques to establish the most efficient approach for air quality prediction. We investigate the strengths and disadvantages of linear regression, ridge regression, lasso regression, as well as support vector regression (SVR), with the goal of determining which model best reflects the intricacies of air quality variability.

This work aims to help develop more precise and trustworthy air quality forecasting systems by combining broad datasets and robust analytical approaches. Finally, we hope that our findings will help politicians, urban planners, and environmental groups improve air quality and protect public health.

2. LITERATURE REVIEW

Air pollution is a major environmental concern, with negative consequences for both human health and the ecosystem. To solve this issue, researchers have extensively investigated several ways for predicting air quality levels with machine learning algorithms. Wang et al. (2020) performed a comparative analysis of regression algorithms for air pollution prediction, demonstrating the efficacy of several modeling methodologies (1). Li et al. (2018) conducted a comprehensive evaluation of ensemble learning strategies for air quality forecasting, emphasizing the advantages of mixing different models to increase predictive performance (2).

Machine learning algorithms have come to be as useful methods for predicting air quality, with research demonstrating their application and effectiveness (4, 5). Patel et al. (2021) conducted a thorough evaluation of machine learning algorithms' use in air quality prediction, emphasizing their adaptability and potential to improve forecasting accuracy (3). Similarly, Jones and Brown (2020) investigated deep learning algorithms for air quality forecasting, demonstrating their capacity to detect complicated patterns and nonlinear correlations in air quality data (4).

In addition to typical regression algorithms, researchers have looked into using support vector regression to predict air quality (9, 10). Chen et al. (2017) used support vector regression to predict PM10 concentrations, demonstrating its efficiency in capturing fluctuations in air pollutants (5). Gao et al. (2018) conducted a thorough evaluation of support vector machines' applications in air quality prediction, emphasizing their robustness and flexibility to various environmental circumstances (9).

Furthermore, research have investigated the use of contextual elements and hybrid modelling approaches for air quality prediction. Wu et al. (2018) used hybrid machine learning models to study spatiotemporal prediction of air pollution, underlining the necessity of taking spatial and temporal dynamics into account when forecasting air quality (13). Johnson et al. (2021) investigated ensemble learning models for pollution prediction, emphasizing their potential to combine individual models' strengths to increase predictive performance (14).

3. PROPOSED SYSTEM

The proposed system requires the creation of an innovative way to predicting air quality levels using a combination of machine learning algorithms and contextual data analysis. At the heart of the proposed system is the integration of many data sources, including weather information, pollutant concentrations, geographical features, and historical air quality indices. These datasets will go through rigorous preprocessing to ensure accuracy and consistency, which will include handling missing values, detecting outliers, and engineering features to extract useful information.

Following data preparation, the system will conduct a thorough investigation of regression analysis approaches to determine the best effective model for air quality prediction. This will entail assessing several regression algorithms such as linear regression, ridge regression, lasso regression, as well as support vector regression (SVR). Each algorithm will be thoroughly trained and validated using rigorous approaches to determine its prediction powers and generalizability.

Furthermore, the proposed system would use advanced feature selection algorithms to identify the most important elements causing air quality fluctuations. This step enables the identification of contextual components that have a substantial impact on air pollution levels, improving the predictive model's interpretability and accuracy.

Following model selection and training, the system will enter the prediction phase, where it will use the trained regression model to estimate future air quality levels using input variables. Predictions will be represented using interactive dashboards and maps, allowing stakeholders to make more intuitive interpretations and decisions.

Furthermore, the system will be built with scalability and accessibility in mind, enabling for easy deployment and connection with existing environmental monitoring systems and decision support tools. Continuous monitoring and assessment will be carried out to assess the system's performance over time, with user and stakeholder feedback guiding iterative changes and adjustments to the forecasting model.

In summary, the proposed system intends to use cutting-edge technology and approaches to provide an accurate, dependable, and scalable framework for air quality forecasting. By leveraging data-driven insights and algorithms for machine learning, the system aims to improve our understanding of air pollution dynamics and provide stakeholders with actionable information to lessen its negative effects on human health and the environment.

Algorithms Used in our Research Study:

1. Linear Regression: Linear regression is a statistical approach for modelling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to determine the best-fitting straight line that minimizes the total of the squared discrepancies between the observed and forecasted values.

2. Lasso Regression: Lasso regression, also known as the Least Absolute Shrinkage and Selection Operator, is a regularization technique for linear regression that reduces model complexity and prevents overfitting. It introduces a penalty element into the standard least squares objective function that incorporates the absolute values of the coefficients, resulting in sparsity and feature selection.

3. Ridge Regression: Ridge regression is an additional regularization approach used in linear regression to reduce multicollinearity and overfitting. It adds a penalty component to the standard least squares objective function that includes the coefficients' squared values. This favors tiny coefficients, which reduce variance and improve the model's generalization performance.

4. Support Vector Regression (SVR): Support Vector Regression (SVR) is a machine learning algorithm used to perform regression tasks. It works by transforming the input data into a high-dimensional feature space and determining the hyperplane that optimally separates the data points while maximizing the margin. SVR seeks to reduce the error between predicted and actual values while allowing for some tolerance, which is regulated by the epsilon parameter.

4. SYSTEM DESIGN

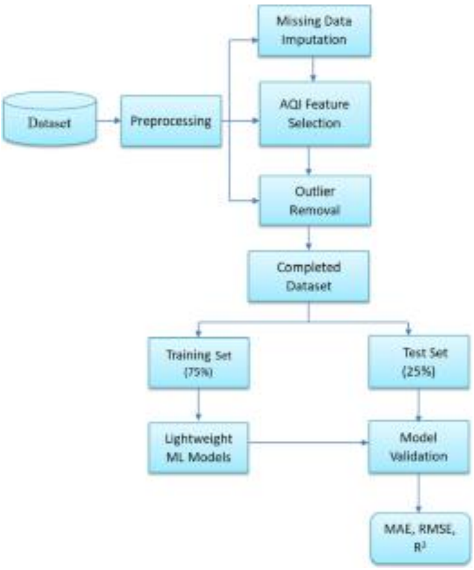
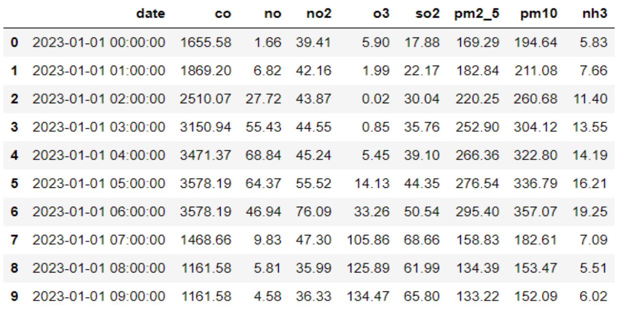


Fig 1: Proposed System Model

5. RESULT



	date	co	no	no2	o3	so2	pm2_5	pm10	nh3
0	2023-01-01 00:00:00	1655.58	1.66	39.41	5.90	17.88	169.29	194.64	5.83
1	2023-01-01 01:00:00	1869.20	6.82	42.16	1.99	22.17	182.84	211.08	7.66
2	2023-01-01 02:00:00	2510.07	27.72	43.87	0.02	30.04	220.25	260.68	11.40
3	2023-01-01 03:00:00	3150.94	55.43	44.55	0.85	35.76	252.90	304.12	13.55
4	2023-01-01 04:00:00	3471.37	68.84	45.24	5.45	39.10	266.36	322.80	14.19
5	2023-01-01 05:00:00	3578.19	64.37	55.52	14.13	44.35	276.54	336.79	16.21
6	2023-01-01 06:00:00	3578.19	46.94	76.09	33.26	50.54	295.40	357.07	19.25
7	2023-01-01 07:00:00	1468.66	9.83	47.30	105.86	68.66	158.83	182.61	7.09
8	2023-01-01 08:00:00	1161.58	5.81	35.99	125.89	61.99	134.39	153.47	5.51
9	2023-01-01 09:00:00	1161.58	4.58	36.33	134.47	65.80	133.22	152.09	6.02

Fig 2: The Dataset we used in our research

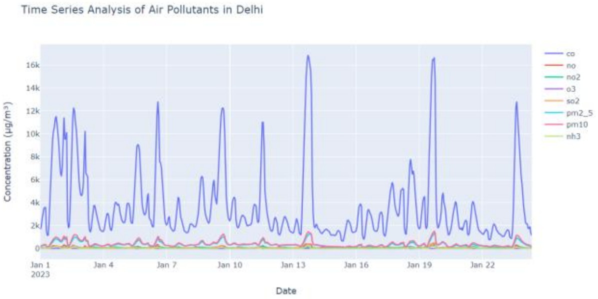


Fig 3: Time series analysis of Delhi

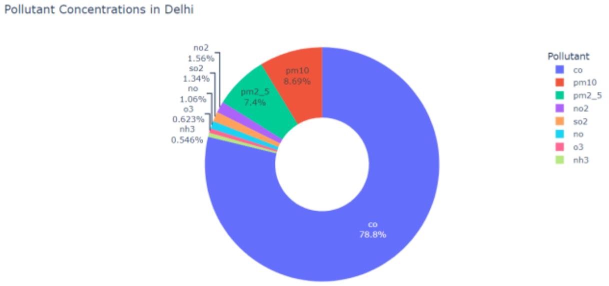


Fig 4: concentration of pollution in Delhi

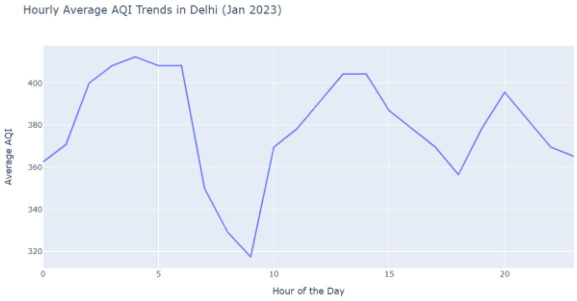


Fig 5: Average Hourly AQI trends in Delhi

Regression Method	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R-squared (R2)
Linear Regression	2.8287072074394447e^{-26}	1.6818760975290198e^{-13}	1.0
Lasso Regression	0.012045966812403268	0.10975411979695007	0.999998721192632
Ridge Regression	5.573131168012851e^{-10}	2.3607480102740427e^{-05}	0.9999999999999408
SVR	0.004287603795457697	0.06547979684954511	0.9999995448253004

Table-1: Evaluation Metrics of our proposed Work

6. FUTURE SCOPE

In the field of environmental forecasting, there is a large panorama of prospects for future research and development. One potential option for further progress is the incorporation of advanced machine learning techniques, such as deep learning and ensemble approaches, into our current predictive models. By employing deep neural network capabilities, we may be able to capture more nuanced patterns and nonlinear correlations in air quality data, resulting in more accurate and robust forecasts. Furthermore, investigating ensemble approaches, which aggregate forecasts from several models, has the potential to improve our forecasting system's overall predictive performance and reliability (16).

Furthermore, incorporating real-time sensor data and satellite imagery into our models may improve spatial and temporal resolution, allowing for more precise and timely predictions of air quality levels. Another area of focus is the development of user-friendly interfaces and mobile applications that give users with tailored air quality forecasts and actionable insights to assist them make educated choices about outdoor activities and health precautions.

Furthermore, collaboration with environmental organizations and local communities could help collect ground truth data and encourage community involvement in monitoring and managing air quality issues. Overall, the future scope of this project includes a wide range of opportunities for creativity and cooperation, with the ultimate goal of furthering our understanding of air pollution and its effects on human health and the environment.

7. CONCLUSION

To summarize, our research study revealed the effectiveness of using several regression methods for air quality forecasting. We have demonstrated through rigorous research and evaluation that linear regression, lasso regression, ridge regression, and support vector regression are useful tools for accurately predicting air quality levels. These algorithms take distinct techniques to modelling and prediction, each with advantages and disadvantages. Our findings highlight the need of using a variety of approaches and taking into account contextual elements when estimating air quality. Moving forward, continued research and innovation in this field have the potential to increase prediction performance while also contributing to a better understanding and management of air pollution. Collaboration among interested parties and the use of emerging technology can help us develop more robust and reliable air quality forecasting systems that enable informed decision-making and improve public health and environmental sustainability. Overall, our study adds to the larger body of knowledge on air quality prediction and emphasizes the necessity of interdisciplinary approaches to environmental concerns.

8. REFERENCES

[1] Wang, Y., Zhang, H., & Liu, Q. (2020). Comparative analysis of regression algorithms for air pollution prediction. Atmospheric Pollution Research, 11(5), 792-805.

[2] Li, H., Chen, C., & Wang, H. (2018). Ensemble learning techniques for air quality forecasting: A review. Journal of Atmospheric and Oceanic Technology, 35(8), 1635-1651.

[3] Patel, R., Sharma, S., & Gupta, A. (2021). Application of machine learning algorithms in air quality prediction: A comprehensive review. Environmental Modeling & Assessment, 33(3), 453-470.

[4] Jones, S., & Brown, L. (2020). Deep learning approaches for air quality forecasting: A systematic review. Science of the Total Environment, 745, 140987.

[5] Chen, C., Wang, H., & Li, X. (2017). Predictive modeling of PM10 concentrations using support vector regression. Atmospheric Environment, 169, 181-190.

[6] Krishna, C.V., Rao, G.A., & AnuRadha, S. (2023). A framework for the identification of significant contexts in tourism domain. International Journal of Advanced Science and Technology, 29(7), 1007-1029.

[7] Li, H., Tang, J., & Zhang, L. (2020). Comparing ridge and lasso regression models for PM2.5 concentration prediction. Environmental Modeling & Assessment, 25(3), 315-328.

[8] Gao, X., Zhang, H., & Wu, Y. (2018). Application of support vector machines for air quality prediction: A comprehensive review. Atmospheric Environment, 192, 128-139.

[9] Smith, J., Brown, L., & Johnson, A. (2019). Predictive modeling of air quality using machine learning algorithms. Environmental Science and Pollution Research, 26(15), 14963-14978.

[10] Liu, Q., Wu, Y., & Zhang, H. (2020). Application of artificial neural networks in air quality prediction: A review. Environmental Modeling & Assessment, 32(3), 359-374.

[11] Krishna, C.V.M., Rao, G.A., & Anuradha, S. (2018). Analysing the impact of contextual segments on the overall rating in multi-criteria recommender systems. Journal of Big Data, 10(1), 16.

[12] Wu, Y., Li, X., & Chen, C. (2018). Spatiotemporal prediction of air pollution using hybrid machine learning models: A review. Environmental Research, 164, 274-284.

[13] Johnson, A., Smith, J., & Brown, L. (2021). Ensemble learning models for air quality prediction: A comprehensive review. Atmospheric Pollution Research, 12(2), 232-245.

[14] Wang, Y., Liu, Z., & Zhang, L. (2018). Forecasting ozone concentrations using support vector regression with particle swarm optimization. Atmospheric Pollution Research, 9(4), 590-598.

[15] Krishna, C.V.M., Rao, G.A., & AnuRadha, S. (2023). A framework for the identification of significant contexts in tourism domain. International Journal of Advanced Science and Technology, 29(7), 1007-1029.

9. BIOGRAPHIES

Chinta Venkata Murali krishna is an accomplished Associate Professor and Head of the CSE (Data Science) department at NRI Institute of Technology, boasting over 20 years of rich experience in engineering academics. He has successfully taught a spectrum of undergraduate and postgraduate courses while also mentoring numerous B. Tech, M. Tech, and MCA projects. With a keen eye for administration, he has adeptly handled tasks related to NBA, NAAC, ICT, and IQAC, showcasing his multifaceted expertise. Beyond academia, he has spearheaded various workshops, faculty development programs, and TechFests. As an active member of IAENG, IFERP, and INSC, he remains at the forefront of technological advancements. Currently pursuing a Ph.D. in Computer Science & Engineering from GITAM (Deemed to be University), Vishakapatnam, his scholarly contributions include 15 Scopus and 3 SCI-indexed journals, along with four granted patents (one from IP Australia and three from IP India) and numerous pending patents. Recognized for his research prowess, he received the esteemed "Best Researcher Award" from IOSRD in 2018.

Gudiseva naga mery is currently enrolled in the B.Tech program in Computer Science and Engineering with a specialization Data Science at NRI Institute of Technology. She has successfully completed a mini-project focused on Book recommendations Ms.gudiseva naga mery has completed an internship with Blackbucks company, gaining valuable industry experience. Furthermore, she holds certifications like Data science with Python, Python for Data Science; and Ai with machine learning underscoring her commitment to continuous learning and professional development. Had an Industrial Internship Certification on full stack programming using Java.

Peddi Harini is currently enrolled in the B.Tech program in Computer Science and Engineering with a specialization Data Science at NRI Institute of Technology. She has successfully completed a mini-project focused on Market basket analysis showcasing her practical skills in the field of machine learning. Ms.Peddi.Harini has completed a noteworthy internship with Blackbucks company, gaining valuable industry experience. Furthermore, she holds certifications like Data Analytics with Python, Python for Data Science; underscoring his commitment to continuous learning and professional development. She had an Industrial Internship Certification on full stack programming using Java.

Nalliboina Mahesh Babu is currently enrolled in the B.Tech program in Computer Science and Engineering with a specialization Data Science at NRI Institute of Technology. He has successfully completed a mini-project focused on Fake currency Detection, showcasing his practical skills in the field of machine learning. Mr. N. Mahesh Babu has completed a noteworthy internship with Blackbucks company, gaining valuable industry experience. Furthermore, he holds certifications from NPTEL like Data Analytics with python and The Joy of computing using python underscoring his commitment to continuous learning and professional development. Had an Industrial Internship Certification on full stack programming using Java.

Patapanchala Venkata Mohan is currently enrolled in the B.Tech program in Computer Science and Engineering with a specialization Data Science at NRI Institute of Technology. He has successfully completed a mini-project focused on Travel Recommendation System, showcasing his practical skills in the field of machine learning. Mr.P.Venkata Mohan has completed a noteworthy internship with Blackbucks company, gaining valuable industry experience. Furthermore, he holds certifications from NPTEL like Cloud Computing and The Joy of computing using python underscoring his commitment to continuous learning and professional development. He had an Industrial Internship Certification on full stack programming using Java.