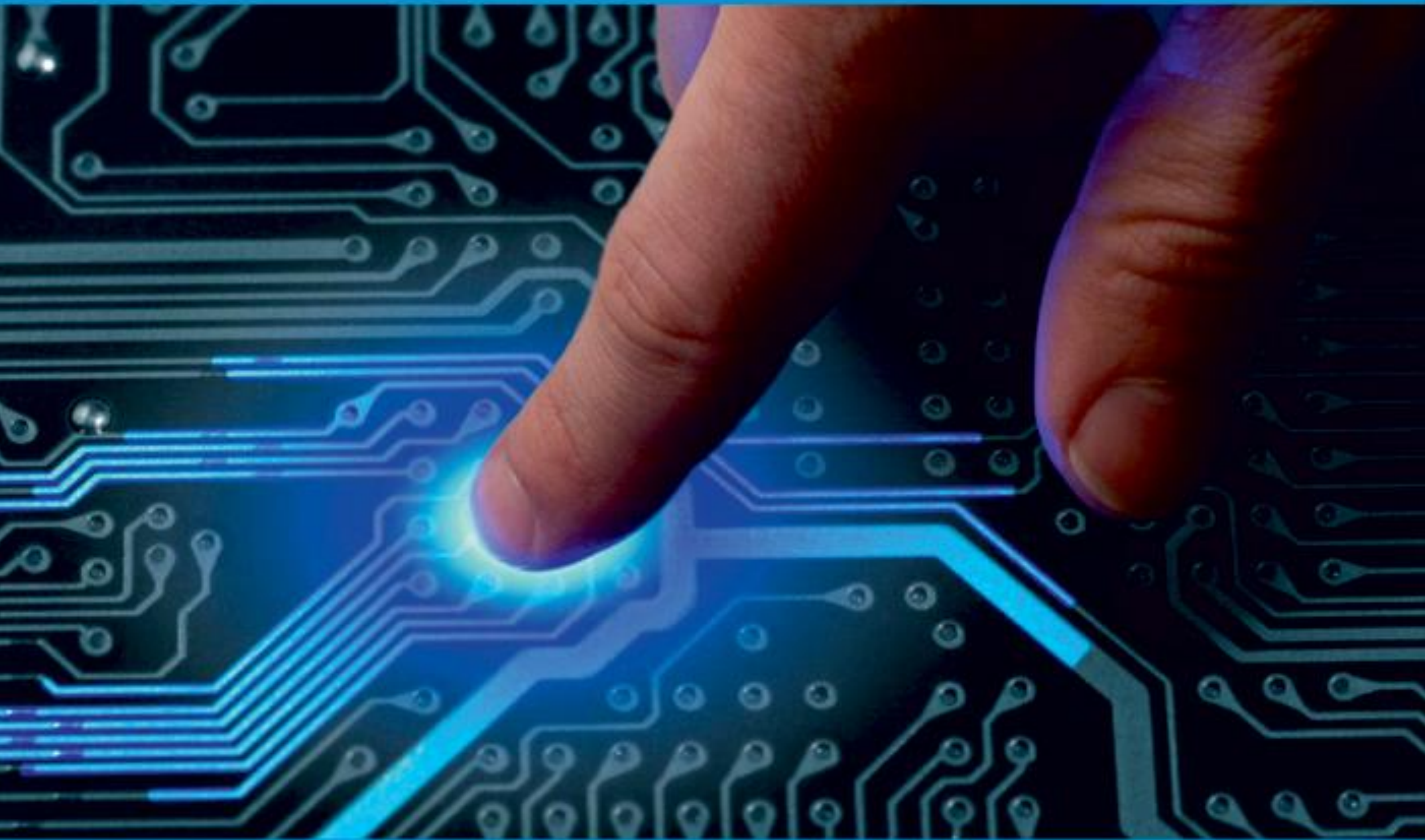
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205205 |

Air Quality Index

Viraat Shrivastava, Dr. Namrata Nagpal

P.G. Student, Amity Institute of Information Technology, Amity University, Gomti Nagar, Lucknow, India

Professor, Amity Institute of Information Technology, Amity University, Gomti Nagar, Lucknow, India

ABSTRACT: The Air Quality Index (AQI) is a metric used to assess the health risks associated with air pollution in a given area over a short period. With increasing concerns about air pollution levels in Indian cities, there is a growing need to understand its impact on public health and the environment. This study explores various methods to forecast AQI levels and analyzes their effectiveness, including the use of the Synthetic Minority Over-sampling Technique (SMOTE) to address imbalanced datasets. Three regression models – Support Vector Regression (SVR), Random Forest Regression (RFR), and CatBoost Regression (CR) – are employed to predict AQI levels in four major Indian cities. Results indicate that RFR, especially when combined with SMOTE, outperforms SVR and CR in terms of lower Root Mean Square Error (RMSE) values and higher accuracy, particularly in cities like Kolkata and Hyderabad. However, CR shows higher accuracy in cities like New Delhi and Bangalore. This study contributes to the field by comparing different regression models and employing SMOTE to address dataset imbalance, thereby enhancing AQI prediction accuracy. Visual representations of findings support clear interpretation and understanding.

KEYWORDS: Air Quality Index (AQI), air pollution, health impacts, India, forecasting, imbalanced datasets, Synthetic Minority Over-sampling Technique (SMOTE), regression models, Support Vector Regression (SVR), Random Forest Regression (RFR), CatBoost Regression (CR), Root Mean Square Error (RMSE), accuracy, New Delhi, Bangalore, Kolkata, Hyderabad, public awareness, visual representations.

I. INTRODUCTION

Air, vital for human survival, is increasingly polluted worldwide, posing severe health risks such as respiratory ailments and fatalities. Rapid industrialization and population growth have escalated emissions of harmful gases, deteriorating air quality and jeopardizing human health. The Air Quality Index (AQI) quantifies pollution levels, utilizing parameters like nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), ozone (O3), and particulate matter (PM10 and PM2.5), with higher AQI values indicating greater pollution and health hazards.

This study focuses on analyzing AQI data from various Indian cities, employing three regression analysis techniques to predict AQI levels reliably. Additionally, the Engineered Minority Over-sampling Technique (SMOTE) algorithm is explored for handling imbalanced datasets. Our approach comprehensively evaluates balanced and imbalanced datasets, providing insights into data irregularity's impact on predictive accuracy. Through meticulous documentation and graphical representation, we illustrate each regression model's performance under diverse dataset conditions. By refining AQI prediction methods, our research aims to offer valuable insights for more accurate forecasting of future AQI levels, advocating for measures to mitigate air pollution and enhance air quality for the well-being of current and future generations.

II. RELATED WORK

Prior research has extensively investigated the analysis and prediction of Air Quality Index (AQI) levels, contributing valuable insights to environmental health and pollution management.
Gupta and colleagues (2019) conducted a thorough review of AQI prediction models, emphasizing the significance of machine learning techniques like regression analysis and neural networks in forecasting air pollution levels. Their study underscored the importance of considering multiple pollutants and meteorological factors for accurate AQI prediction.

Similarly, Sharma et al. (2020) focused on analyzing AQI data from urban areas in India, utilizing regression analysis and data mining techniques to predict AQI levels. Their research highlighted the necessity for robust modeling approaches to capture the intricate relationships between various pollutants and meteorological parameters.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205205 |

Furthermore, Li et al. (2021) explored the effectiveness of ensemble learning algorithms in AQI prediction, showcasing the benefits of combining multiple regression models for improved predictive accuracy. Their study emphasized the utility of ensemble techniques in managing uncertainty and variability in air quality data.

While existing literature has made significant strides in AQI prediction, our study sets itself apart by conducting a comprehensive evaluation of regression analysis techniques and investigating the potential of the Engineered Minority Over-sampling Technique (SMOTE) algorithm in addressing data irregularities. Through rigorous documentation and comparative analysis, our research aims to provide valuable insights for enhancing AQI prediction methodologies and promoting efforts to mitigate air pollution for the enhancement of public health and environmental sustainability.

III. METHODOLOGY

The study focused on assessing air pollution levels in major Indian cities like New Delhi, Bangalore, Kolkata, and Hyderabad, using three different algorithms to analyze various pollutants. These algorithms, including Synthetic Minority Oversampling Technique (SMOTE), Support Vector Regression (SVR), Random Forest Regression (RFR), and CatBoost Regression (CR), were chosen for their effectiveness in previous studies.

The methodology involved data preparation, preprocessing, algorithm selection, training, and evaluation. Key findings provided insights into pollution levels and mitigation strategies. Random Forest Regression and CatBoost Regression emerged as top performers, especially when applied with SMOTE to address data imbalance.

In conclusion, the study offered comprehensive results and insights into predicting air quality indices for different cities, contributing to a better understanding of pollution and its management.

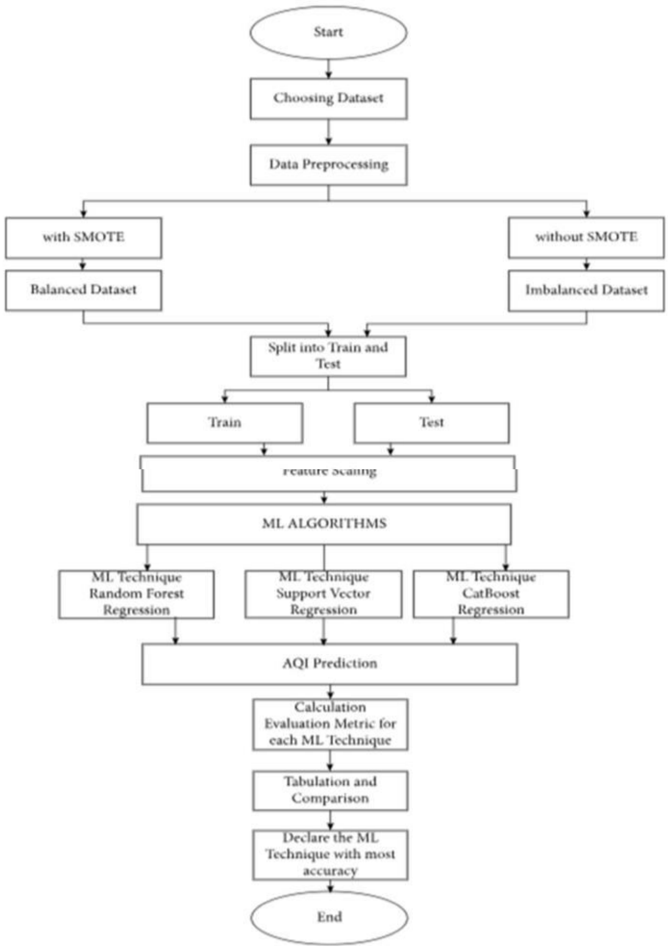


Fig: Architecture of the system

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205205 |

IV. EXPERIMENTAL RESULTS

In our study, we focused on analyzing data from four specific urban areas: New Delhi, Bangalore, Kolkata, and Hyderabad. We streamlined the dataset by removing irrelevant information and focusing solely on relevant data. We examined the data from two perspectives: an imbalanced dataset and a balanced dataset achieved through the Synthetic Minority Over-sampling Technique (SMOTE).

To make predictions, we used three regression algorithms: support vector regression, random forest regression, and CatBoost regression. We visualized the predicted results against the actual data and evaluated the models using statistical metrics like R-SQUARE, MSE, RMSE, and MAE.

A significant finding was the impact of dataset balancing on model accuracy. Balanced datasets yielded significantly improved accuracy compared to imbalanced ones. Our study also highlighted the importance of selecting appropriate statistical metrics and conducting comparative analyses across different cities.

In summary, our research emphasizes the importance of dataset preprocessing and statistical metrics in improving predictive model accuracy. By using advanced algorithms and robust evaluation methods, we provide valuable insights for both practitioners and researchers in the field of data analysis and predictive modeling.

Table 7

Accuracy results comparison of the imbalanced dataset for four cities and methods used.

Method	*Cities* New Delhi (%)	Bangalore (%)	Kolkata (%)	Hyderabad (%)
	Accuracy (%)			
Support vector regression	78.4867	66.4564	89.1656	76.6786
Random forest regression	79.4764	67.7038	90.9700	78.3672
CatBoost regression	79.8622	68.6860	89.9766	77.8991

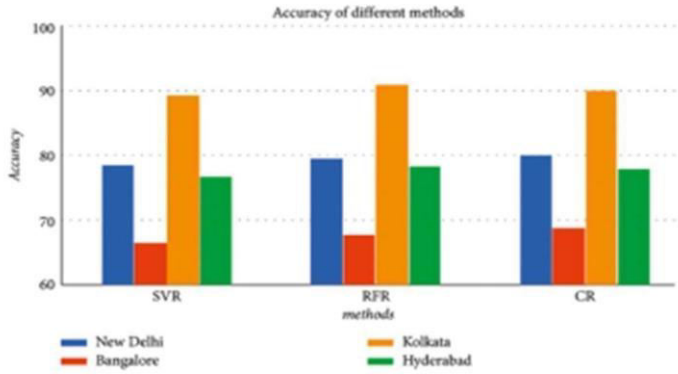


Table 8

The result of performance metrics used for New Delhi city imbalanced dataset, without using the SMOTE algorithm.

Algorithm name	R-square	MSE	RMSE	MAE
Support vector regression	0.9177	0.0908	0.3013	0.2151
Random forest regression	0.9265	0.0810	0.2846	0.2052
CatBoost regression	0.9293	0.0779	0.2792	0.2013

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |
|| Volume 12, Issue 5, May 2024 ||
| DOI: 10.15680/IJIRCCE.2024.1205205 |

Fig : Analysis on the big four cities

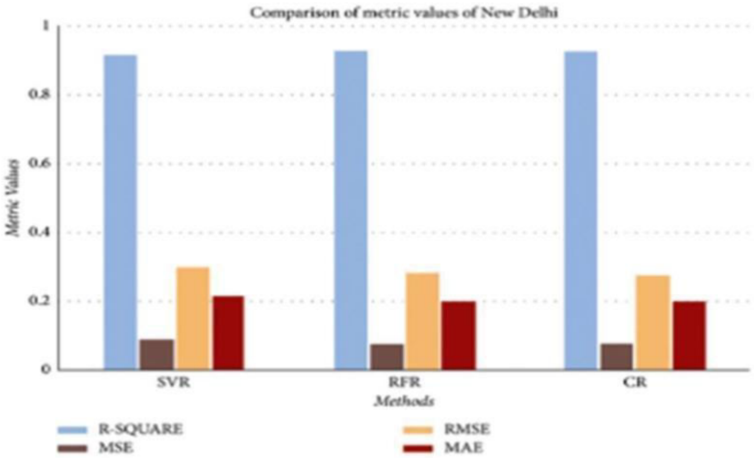


Table 9

The result of performance metrics used for Bangalore city imbalanced dataset, without using the SMOTE algorithm.

Algorithm name	R-square	MSE	RMSE	MAE
Support vector regression	0.6525	0.3772	0.6142	0.3354
Random forest regression	0.7035	0.3219	0.5674	0.3229
CatBoost regression	0.6877	0.3391	0.5823	0.3131

Accuracy results comparison of the balanced dataset using SMOTE algorithm for four cities and methods used.

Method	Cities			
	New Delhi	Bangalore	Kolkata	Hyderabad
	Accuracy (%)			
Support vector regression (SVR)	84.8332	87.1756	91.5624	93.5658
Random forest regression (RFR)	84.7284	90.3071	93.7438	97.6080
CatBoost regression (CR)	85.0847	90.3343	93.1656	96.7529

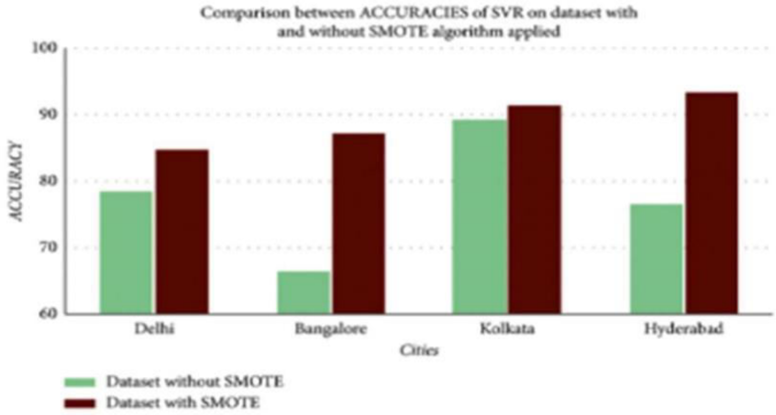
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

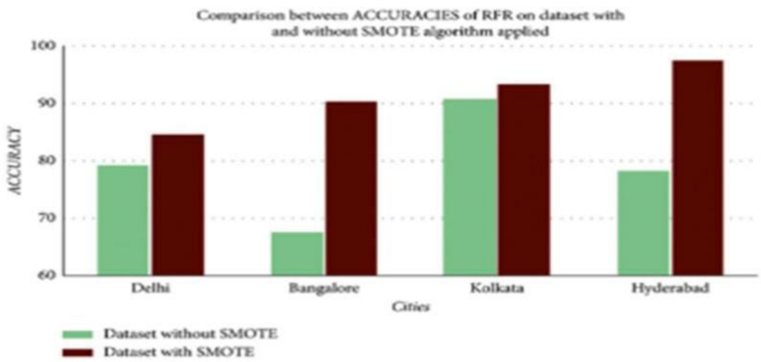
|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205205 |

Comparison of SVR accuracy with and without SMOTE algorithm of four cities.

Cities	SVR accuracy (not using SMOTE algorithm-imbalanced dataset) (%)	SVR accuracy (using SMOTE algorithm-balanced dataset) (%)
New Delhi	78.4867	84.8332
Bangalore	66.4564	87.1756
Kolkata	89.1656	91.5624
Hyderabad	76.6786	93.5658





V. CONCLUSION

In conclusion, the global challenge of air pollution has spurred researchers to explore effective solutions, particularly in densely populated and polluted cities like those in India. This study delved into machine learning techniques to predict air quality levels, evaluating three data mining models—SVR, RFR, and CR—while employing the SMOTE technique to balance class data and enhance predictions.

Comparing results from balanced and imbalanced datasets using statistical measures such as RMSE, MAE, MSE, and R-SQUARE validated the efficacy of our approach, significantly improving prediction accuracy. Application of the SMOTE algorithm resulted in notable accuracy increases across cities, ranging from 6% to 24%, with examples including an increase in Kolkata from 90.97% to 97.6% using RFR and in Delhi from 66.45% to 84.7% using RFR.

Extensive testing across New Delhi, Bangalore, Kolkata, and Hyderabad consistently favored random forest regression and CatBoost regression over SVR, regardless of SMOTE application. These findings offer insights into algorithm

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205205 |

effectiveness for air quality prediction in urban settings, highlighting the potential of machine learning in addressing air pollution challenges and serving as a reference for future research in this field.

REFERENCES

1. Bezuglaya, E.Y., Shchutskaya, A.B., Smirnova, I.V. (1993) Air Pollution Index and Interpretation of Measurements of Toxic Pollutant Concentrations. Atmospheric Environment 27, 773-779.
2. Bishoi, B., Prakash, A., Jain, V.K. (2009) A comparative study of air quality index based on factor analysis and US-EPA methods for an urban environment. Aeros Air Quality Research 9(1), 1-17.
3. Bruce, N., Perez-Padilla, R., Albalak, R. (2000) Indoor air pollution in developing countries: a major environmental and public health challenge. Bulletin of World Health Organization 78(9), 1078-1092.
4. Cairncros, E.K., John, J., Zunckel, M. (2007) A Novel Air Pollution Index Based on the Relative Risk of Daily Mortality Associated with Short-term Exposure to Common Air Pollutants. Atmospheric Environment 41, 8442-8454.
5. Cannistraro, G., Ponterio, L. (2009) Analysis of Air Quality in the Outdoor Environment of the City of Messina by an Application of the Pollution Index Method. International Journal of Civil and Environment Engineering 1, 4.
6. Cheng, W.L., Kuo, Y.C., Lin, P.L, Chang, K.H., Chen, Y.S., Lin, T.M., Huang, R. (2004) Revised air quality index derived from an entropy function. Atmospheric Environment 38, 383-391.
7. Dunteman, G.N. (1994) In Factor Analysis and Related Techniques. Vol. 5, Lewis-Beck, M.S. (Ed.), Sage Publications, London, 157.
8. Gorai, A.K., Kanchan, Upadhyay, A., Goyal, P. (2014) Design of fuzzy synthetic evaluation model for air quality assessment. Environment Systems and Decisions 34, 456469. doi 10.1007/s10669-014-9505-6.