Prediction of Air Quality Index Using D a t a s c i e n c e

1. Introduction

Humans can only survive because of air. Its quality must be monitored and understood for our wellbeing. Due to air pollution, millions of people around the world suffer from physiological disorders and respiratory death. According to scientific evidence, air pollution poses the single greatest environmental risk. Due to the toxic gas emissions caused by rapid industrialization, population levels have dramatically increased. Our health is suffering greatly as a result of the air

being contaminated by hazardous substances. Due to this unchecked pollution, air quality has significantly declined. AQI is a numerical index used to measure and convey air pollution levels. The 12 parameters (air pollutants) used to calculate the AQI are NO_2 (nitrogen dioxide), SO_2 (sulfur dioxide), CO (carbon monoxide), O_3 (ozone), PM_{10} (particulate matter having a diameter of 10 microns or less), $PM_{2.5}$ (particulate matter having a diameter of 2.5 microns or less), NH_3 (ammonia), and benzene. In other applications, the six pollutants PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, and O_3 are used to calculate the air quality index (AQI). However, the precise selection of contaminants relies on the particular aim and numerous variables, including data accessibility, measurement techniques, and monitoring frequency. A high AQI number indicates severely contaminated air, which can have a serious negative impact on health. Real-time air quality can be monitored using the AQI. Numerous weather stations have also captured daily and hourly AQI data in our own backyard. These data will be mined and harvested with the intention of using them in the suggested work.

As a result, the dataset used contains records of the AQIs in various Indian cities. The three distinct regression analysis techniques will be put into practice, and the best accuracy will be determined through comparison.

The proposed work compares a dataset's effectiveness before and after using the SMOTE algorithm. The major novelty is the usage of SMOTE. Unlike other papers, the impact of an imbalanced dataset has been studied, and hence, SMOTE has been applied to balance it. Furthermore, the whole process has been documented with graphs and metrics which showcase each algorithm, every performance metric, under every dataset—in both its balanced and imbalanced form. The effectiveness of the suggested methods will aid in predicting future AQI levels, which can serve as a warning and emphasize the need of reducing air pollution levels.

2. Literature Survey

They initially looked at the relationship between several air indicators, such as the AQI, $PM_{2.5}$ concentrations, total NO_x (nitrogen oxides) concentrations, and so on, in this study [1]. Second, they built prediction models using random forest regression (RFR) and support vector regression (SVR), and finally, they assessed the regression models' performance using RMSE, coefficient of determination (R-SQUARE), and correlation coefficient *r*. A widely used machine learning method (SVR) is used to quantify pollutant and particle levels and predict the air quality index [2]. According to the findings, hourly concentrations of pollutants such as carbon monoxide, sulfur dioxide, nitrogen dioxide, ground-level ozone, and particulate matter 2.5, as well as the hourly AQI for the state of California, may be consistently predicted using SVR with the RBF kernel. The classification of unseen validation data into six AQI categories provided by the United States Environmental Protection Agency (dataset) was completed with 94.1 percent accuracy.

The prediction of the AQI using ML techniques such as time series analysis and LR. To predict the AQI, MLR and supervised machine learning technique were used. Various quantitative indices were used to assess the performance. Second, to forecast the AQI in the future, the ARIMA time series model was used. Both models were found to be highly accurate and efficient in forecasting the AQI [3]. An integrated model used artificial neural networks and the Kriging method to estimate the quantity of air pollutants at several places in Mumbai and Navi Mumbai. The high *R* values meant that the necessary level of fit between anticipated and observed values had been achieved. In terms of *R* value and forecast, ANN outperformed simple regression models [4]. To predict AQI author concentration based on parameter like $PM_{2.5}$, PM_{10}, SO_2 and NO_2. In conclusion, of the algorithms linear regression, decision tree regression, SVR, and RFR, the random forest regression algorithm yielded the best accuracy of 0.99985 on the test data with the least mean square error of 0.00013 and the mean absolute error of 0.00373 [5].

To forecast the AQI using the previous year's data and projecting over a specified future year as a gradient descent boosted the multivariable regression issue. They outperformed ordinary regression models by improving the model's efficiency by employing cost estimates for the forecasting problem. They also utilized the AHP MCDM technique to assess the order of preference based on how closely the alternatives resembled the ideal solution [6]. Logistic regression [7] was used to determine if the presented data sample of daily weather/environmental conditions in a specific city was polluted or not. Based on previous $PM_{2.5}$ readings, this system attempted to predict $PM_{2.5}$ levels and detect air quality. The results demonstrated that logistic regression and autoregression could be used effectively to detect air quality and predict $PM_{2.5}$ levels in the future. Using 6 years of meteorological and pollutant data, this research [8] offered an ML approach for predicting $PM_{2.5}$ concentrations from wind (speed and direction) and precipitation levels. The findings of the classification model showed good reliability in classifying low (10 g/m^3) against high (>25 g/m^3) $PM_{2.5}$ concentrations, as well as low (10 g/m^3) versus moderate (10–25 g/m^3) $PM_{2.5}$ concentrations. An integrated model used the ANN and the Kriging method to predict the level of air pollutants in Mumbai and Navi Mumbai based on historical data from the meteorological department and the Pollution Control Board [9]. The proposed model was then implemented and tested using the MATLAB application for ANN and the *R* application for the Kriging method. The system helped with analyzing theextensive pollution data and projecting future pollution. The identification of future data points to forecast air pollution was also done using time series analysis. An effective strategy to predict Delhi's AQI using a deep RNN based on LSTM to predict hourly pollutant concentrations was explored. Even in hourly predictions, results were accurate. According to the findings [10], deep learning-based strategies performed better than traditional statistical methods [11].

To predict daily AQI, prediction models included those that used ARIMA as a time series model, PCR as a hybrid regression model, ARIMA and PCR as the first ensemble model, and ARIMA and gene expression programming

(GEP) as the second ensemble model. By utilizing the correlation between urban nature (such as street greenness and street building), urban traffic (such as vehicle volume), and air pollution, a set of periodic-frequent patterns and a $PM_{2.5}$ estimating model were created (e.g., $PM_{2.5}$). They established a link between urban nature, traveling automobiles, and the quality of air pollution. Using this information, people can work toward developing an outstanding strategy to address all of them [12]. Linear regression was used as a machine learning algorithm to predict air quality for the next day using sensor data from three specific locations in the Capital City of India-Delhi and the National Capital Region (NCR). The model's performance was assessed using four performance measures: MAE, MSE, RMSE, and MAPE. This paper looked at AQI prediction using data generated by IoT arrangements [13]. The ANN algorithm predicted hourly criteria pollutants concentration levels and, AQI, AQHI, for Ahvaz, Iran, over a span of 12 months (Aug 2009–Aug 2010). This study demonstrated that the ANN can be used to forecast air quality in cities such as Ahvaz in order to prevent health effects. They came to the conclusion that urban air quality authorities might evaluate the spatial-temporal profile of pollutants and air quality metrics using an artificial neural network [14].

Using air quality and meteorological records, tree-based ensemble learning models were developed to study the urban air quality of the city of Lucknow in India over a five-year period. PCA was used to identify the sources of air pollution. Due to the incorporation of boosting and bagging techniques, the DTF and DTB models performed better in classification and regression than the SVM. The suggested ensemble models for managing urban ambient air quality were successful in predicting it [15]. They focused on air quality index measures and predictions based on past data for the Central Jakarta area. $PM_{2.5}$, one of the most often utilized components in AQI assessment, was used in this investigation. Based on testing data, Brown's weighted exponential moving average accurately predicted future Central Jakarta AQI levels. In terms of precision, it outperformed the WMA, EMA, and BDES approaches [16].

The dataset was collected to predict the AQI [17] in Chennai, Tamil Nadu. After that, it underwent preprocessing to eliminate redundant data and replace missing values. A deep learning model based on SVR and LSTM was used to classify the AQI values. This proposed deep learning method improved prediction accuracy, which would warn the public to reduce air pollution to a justifiable level. They used five regression models for AQI prediction [18]: principal component, partial least square, and principal component with one out, CV, and multiple regression AQI data from numerous Indian cities. They created three classification models to predict the AQI bucket: multinomial logistic regression, KNN, and KNN model with repeat CV classification. In terms of accuracy and AUC, the KNN Model with repeated CV and tune length 10 performed the best. Health problems are predicted by the decision tree and Naive Bayes algorithms. Good, moderate, unhealthy (unhealthy for sensitive groups), and very unhealthy were the AQI categories. Compared to the Naive Bayes method's accuracy of 86.663 percent, the decision tree algorithm achieved 91.9978 percent [19].

A nonmonitoring region's AQI was anticipated. With results that were 92 percent acceptable for one-hour prediction, the temporal dimension model was initially presented based on the improved KNN algorithm to forecast AQI values across monitoring stations. The algorithm was utilized in conjunction with a backpropagation neural network (BPN), where it additionally considered geographic distance, to predict the outcome of air quality in the spatial dimension [20]. They used ML models to forecast Dhaka's air quality levels that include deep learning methods, such as LSTM, and various other techniques. The novel aspect of this approach was that they used a unique parameter(i.e., daily temperature) for predicting air pollution [21]. An ML-based technique was used for correctly predicting the AQI based on data acquired from weather stations and environment monitoring. The prediction method uses a neural network system improved using a new nonlinear autoregressive neural network (ARNN) having an exogenic input model, which is specifically created for time-series prediction. The framework was used in a scholarship involving various weather monitoring sites in the London area [22].

To predict the air quality index of significant pollutants such as $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, and O_3, they employed a variety of classification and regression approaches, including linear regression, SDG regression, and random forest regression. Evaluations were carried out using MSE, MAE, and R-SQUARE, which showed that ANN and SVM worked best for AQI prediction in New Delhi [23]. They read [24] several papers and gained an understanding of how the ANN could be used to predict the AQI. They used Jaccard similarity and deep learning methods in their proposal. The datasets were collected from UC Irvine. They came to the conclusion that deep learning approaches improve prediction accuracy.

To predict the AQI data on smart cities the following algorighms like supervised learning, SVM and neural networks were utilized in this paper. Databases were procured from the CPCB of the Ministry of Environment, Forests, and Climate Change of the GoI. The model performed well in terms of predicting the air quality of Delhi [25]. The K-Means method [26] was proposed to analyze air pollution. Using real-time records for pollutants, the correlation coefficient was calculated. The possibilistic fuzzy c-means (PFCM) algorithm was contrasted with the K-Means algorithm. The findings demonstrated that the enhanced k-means clustering technique delivered AQI values with higher accuracy and lower execution time. We use supervised learning to create prediction models. Experiments have shown that decision trees (classification), SVR, and stacking ensembles work much better than the other methods in their category. Mathematical models, learning, and regression techniques were recommended for developed areas and cities [27].

The advancement of models for anticipating normal air quality levels utilizing computational insight techniques

enables. The models were developed using data from the three checking stations in the Czech Republic, Dukla, Rosice, and Brnenska, in order to predict the normal air quality file and forecast air quality records for each air pollution separately. For examination, they utilized RMSE [28]. For AQI expectations, they used [29] IoT-based gadget information. They performed contamination expectations involving four high-level relapse methods in this paper and introduced a similar report to decide the best model for precisely anticipating air quality as to information size and handling time. For the correlation of these relapse models, the mean MAE and RMSE were used as gauging measures. High-recurrence detail successions WD(D) and low-recurrence surmised groupings WD(A) are produced using wavelet deterioration, as are long transient memory brain organization and autoregressive moving normal model for WD(D) and WD(A) arrangements for the forecast. As for execution measurements, they utilized RMSE, MAE, and R-SQUARE.

In this study [30], a unique machine-learning technique was developed to predict the condensate viscosity in the areas near the wellbore using 5 input variables: pressure, temperature, initial gas to condensate ratio, gas-specific gravity, and condensate gravity. The novel multiple extreme learning machine (MELM), least squares support vector machine (LSSVM), and multilayer perceptron, each of which has been hybridized with a particle swarm optimizer (PSO) and genetic algorithm, were among the nine machine learning and hybrid machine learning algorithms that were evaluated (GA). In this study [31], a unique machine-learning technique was created based on feature selection to anticipate FVDC from a 12-input variable well-log. The fracture density was previously predicted using a hybrid method that incorporates two networks of multiple extreme learning machines (MELMs), multilayer perceptrons (MLPs), genetic algorithms (GAs), and particle swarm optimizers (PSOs). They used an innovative MELM-PSO/GA mixture that has never been used before. The models were MLP-PSO predictions, which the performance accuracy investigation found.

In this work [32], they created a novel deep machine learning model called convolutional neural network (CNN) to predict oil flow rate through an orifice plate using seven input variables, including fluid temperature, upstream pressure, root differential pressure, the ratio of base sediment to water, oil specific gravity, kinematic viscosity, and beta ratio (Qo). Because there were no consistent and accurate methods to determine Qo, deep learning may be a useful replacement for traditional machine learning techniques. The study's findings demonstrated that the CNN model had the highest Qo prediction accuracy of any of the four developed models when used in the dataset of 3303 data records collected from oil fields throughout Iran.

565 data points from different parts of the world were used in this investigation. In this study [33], the multilayer perceptron method (MLP), an artificial intelligence network, and the innovative combination approaches for oil formation volume factor (OFVF)—artificial bee colony (ABC) and firefly (FF) optimization methods—had been used. In terms of RMSE and R-SQUARE, the MLP-ABC models of prediction accuracy were evaluated for this test dataset.

In this study [34], unique methods for pore pressure prediction were created based on the most significant collection of input features. Accuracy, R-SQUARE, and RMSE were utilized as performance metrics in this work.

For pore pressure (PP) prediction utilizing good log data, this paper [35] combined the empirical equations with machine learning methods such as the random forest algorithm, support vector regression algorithm, artificial neural network algorithm, and decision tree algorithm. For this, 2827 data records from three oil field wells (Wells *A*, *B*, and *C*) in the Middle East were employed. The results showed that the DT method outperformed the other three predictive models in terms of performance prediction accuracy.

In this work [36], predicting dispersed fracture densities in reservoir rocks may be possible using hybrid machine-learning-optimizer models applied to a collection of petrophysical logs confirmed using image log data. The diverse characteristics of fractures were addressed by various well logs in various and sophisticated ways.

Three Marun oil field wells (MN#163, MN#225, and MN#179) provided access to the Asmari reservoir section on Iranian soil, and well-log data records were collected for these wells in order to anticipate shear wave velocity (VS) [37]. Two hybrid machine learning prediction models (MELM-PSO and MELM-GA), one deep learning model (CNN), and regularly used empirical methodologies to anticipate VS were evaluated using the compiled dataset. Deep learning successfully predicts VS for the supervised validation subset.

During the overbalance drilling technique, the safe mud weight window (SMWW) was determined in this paper [38] by projecting the permitted upper and lower limits of the bottom hole pressure window. The novel machine learning approach MELM-PSO was developed to anticipate SMWW using ten well-log input variables and feature selection. RMSE, R-SQUARE, and other performance indicators were applied in this study.

In this study [39], a trustworthy machine-learning forecasting model was used to predict the permeability (K) for heterogeneous carbonate gas condensate reservoirs. They used four machine learning models to predict permeability: decision trees (DTs), support vector machines (SVMs), and group way of data management (GMDH). In addition, the GMDH model outperformed the other models.

In this study [40], the rheological performance of three low-solid drilling fluids (based on bentonite, natural polymers, and nanoclay) was developed using the hybrid nanocomposite as an addition. As the polymer/nanoclay-hybrid-nanoparticle concentration increases, the fluids' filtration abilities get better. The rheological behavior of low-solids polymer-based drilling fluid was most positively impacted by the addition made of the clay-based nanocomposite. The ideal nanoclay content in the hybrid-polymer nanocomposite was thought to be around 5 wt%, according to the analysis of the rheological characteristics and filtration loss of the drilling fluids.

In this article [41], the effectiveness of each drilling fluid type was evaluated in terms of its ability to reduce fluid loss

and mud cake thickness, hence avoiding differential pipe sticking. In that instance, drilling fluid filtering qualities were evaluated as a potential predictor of well diameter reduction caused by mud cake, close to permeable formations, and mud cake thickness was modified. The novel results showed that the rheological and filtration properties of drilling fluids were significantly enhanced by nanoparticles.

In this study [42], they created reliable models to forecast the liquid critical-flow rates for operating oil wells. Performance metrics were applied, such as coefficient of determination, root mean square error (RMSE), average relative error (ARE), and average absolute relative error (AARE).

In this work [43], they improved the forecast of the gas flow rate through wellhead chokes for a gas-condensate field by using the Firefly algorithm.

In this study [44], they developed a cutting-edge hybrid machine learning method that successfully predicts the gas flow rate through wellhead choke in gas condensate reservoirs.

This work is innovative since it thoroughly analyzed other papers that have made the same attempt. By balancing and imbalancing the dataset, the regression models that had the highest accuracy were selected and subsequently used. The use of SMOTE is another noteworthy innovation. Unlike other articles, this one has explored the effects of an imbalanced dataset and used SMOTE to balance it. Additionally, graphs and metrics that demonstrate each method, each performance parameter, and each dataset in both balanced and imbalanced forms have been used to document the entire process.

It will be possible to anticipate future AQI levels with the help of the offered techniques, which can serve as a warning and highlight the necessity of lowering air pollution.

The gaps identified from the literature survey are given below.

(i) In India, AQI measurement stations were set up in 2014. The National Air Monitoring Program has been used to measure AQI data in 240 cities across India. No proper system is in place which regularly provides predicted data for the future.

(ii) All the papers usually focus on one city or area, giving a biased outlook.

(iii) The performance of the existing system should be increased.

These gaps are incorporated into the proposed method. The proposed method uses different regression models along with the SMOTE algorithm for multiple cities in order to increase the accuracy of the various models. Moreover, in the papers studied, the following outcomes were found (i.e., accuracy) for the existing algorithms such as Naive Bayes, support vector machine, artificial neural network, gradient boost, decision tree, and k-nearest neighbor.

Table 1 shows the various ML techniques/algorithms used in the existing systems and also states the accuracy achieved by each ML technique such as Naïve Bayes (NB), support vector machine (SVM), artificial neural network (ANN), gradient boost (GB), decision tree (DT), and enhanced k-means.

3. Dataset Description and Sample Data

The link to the dataset used for this work is given below.

https://www.kaggle.com/rohanrao/air-quality-data-in-india.

The dataset includes hourly and daily air quality and AQI (air quality index) data from numerous stations in several Indian cities. The data are for the years 2015 through 2020. The original dataset included 29532 rows and 16 columns, which included all of the cities listed below. The cities are given below:

Ahmedabad, Aizawl, Amaravati, Amritsar, Bangalore, Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, and Visakhapatnam.

The attribute information is given below.

3.1. Date YYYY-MM-DD, City, $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, NH_3, CO, SO_2, O_3, Benzene, Toluene, AQI, and AQI_Bucket. AQI_Bucket has six values such as good, satisfactory, moderate, poor, very poor, and severe. The dataset is cleaned and selected from the 4 cities datasets such as New Delhi, Bangalore, Kolkata, and Hyderabad from the original dataset. The attribute xylene was removed from the dataset due to the fact that the column values were empty for all 4 cities chosen by using Microsoft Excel software. The dataset includes hourly and daily air quality and AQI (air quality index) data from numerous stations in 26 Indian cities. From the original dataset, the data of four cities such as New Delhi, Bangalore, Kolkata, and Hyderabad were extracted. Because these are major cities of India, it is important to analyze the pollution levels in different urban cities of India as they are the major contributors to the pollution. These particular cities have a higher population density and give a good estimate of the pollution.

After cleaning the dataset and dividing it into 4 for each city, the New Delhi dataset had 176 rows and 15 columns, the Bangalore dataset had 1362 rows and 15 columns, the Kolkata dataset had 747 rows and 15 columns, and the Hyderabad dataset had 1615 rows and 15 columns, respectively. The sample dataset for New Delhi, Bangalore, Kolkata, and Hyderabad is shown in Tables 2–5, respectively.

The initial dataset has an imbalanced composition. Using the synthetic minority oversampling technique (SMOTE) algorithm, the imbalanced dataset is transformed into a balanced dataset. Oversampling is employed in this algorithm. Any classes with inadequate rows are supplemented with additional rows to ensure that each class label has an equal number of rows, or more or fewer rows, in the dataset. Asymmetry exists in an imbalanced dataset. An imbalanced dataset produces a skewed class distribution, which affects the model's accuracy in several ways.

Table 1: Some of the existing algorithm accuracy in percentage from the literature survey.

Name of the algorithm	Accuracy in percentage (%)	Comments
Naïve Bayes (NB)	86.663	—
Support vector machine (SVM)	92.40	—
Artificial neural network (ANN)	84–93	After simulating a lot of models, ANN gives within the range.
Gradient boost (GB)	96	—
Decision tree (DT)	91.9978	Predicting the $PM_{2.5}$ with a near 89% accuracy rate.
Enhanced k-means	71.28	The k-means clustering method is 40% more efficient than the PFCM algorithm based on the speed of execution and accuracy.
Support vector regression (SVR)	99.4	—
Random forest regression (RFR)	99.985	Least MSE of 0.00013 and MAE of 0.00373.
CatBoost regression (CR)	99.88	Predicting $PM_{2.5}$ readings with an inaccuracy of just 0.0006 and a 99.88% accuracy.

Table 2: Sample dataset for New Delhi city.

City	Date	$PM_{2.5}$	PM_{10}	NO	NO_2	NO_x	NH_3	CO	SO_2	O_3	Benzene	Toluene	AQI	AQI_bucket
Delhi	02/01/2015	186.18	269.55	62.09	32.87	88.14	31.83	9.54	6.65	29.97	10.55	20.09	454	Severe
Delhi	03/01/2015	87.18	131.9	25.73	30.31	47.95	69.55	10.61	2.65	19.71	3.91	10.23	143	Moderate
Delhi	04/01/2015	151.84	241.84	25.01	36.91	48.62	130.36	11.54	4.63	25.36	4.26	9.71	319	Very poor
Delhi	05/01/2015	146.6	219.13	14.01	34.92	38.25	122.88	9.2	3.33	23.2	2.8	6.21	325	Very poor
Delhi	06/01/2015	149.58	252.1	17.21	37.84	42.46	134.97	9.44	3.66	26.83	3.63	7.35	318	Very poor
Delhi	07/01/2015	217.87	376.51	26.99	40.15	52.41	134.82	9.78	5.82	28.96	4.93	9.42	353	Very poor
Delhi	08/01/2015	229.9	360.95	23.34	43.16	51.21	138.13	11.01	3.31	30.51	5.8	11.4	383	Very poor
Delhi	09/01/2015	201.66	397.43	19.18	38.56	45.6	140.6	11.09	3.48	32.94	5.25	11.12	375	Very poor
Delhi	10/01/2015	221.02	361.74	24.79	46.39	55.19	134.06	9.7	5.91	34.12	4.87	9.44	376	Very poor
Delhi	11/01/2015	205.41	393.2	28.46	47.29	57.88	131.1	10.98	5.54	50.37	5.93	10.59	379	Very poor

As a result, it is necessary to balance the data. It is possible to improve the accuracy of the results by oversampling the positive class label. SMOTE is used in this paper to conduct oversampling. The SMOTE technique, which builds its model on nearest neighbors, increases the frequency of the minority class or minority class group in the given dataset. The given dataset has 6 positive classes and 12 negative classes, and they are shown in Figure 1. This dataset is given as the input of the SMOTE algorithm. After that, it increases the number of occurrences of the minority class (positive) from six to twelve. It aids in dataset balancing, which improves algorithm performance and prevents overfitting problems. SMOTE typically involves finding a feature vector and its closest neighbor, taking the difference between the two, multiplying it by a random number between 0 and 1, finding a new point on the line segment by adding the random number to the feature vector, and repeating the process for all located feature vectors. SMOTE has the advantage of producing synthetic data points as opposed to copies that differ slightly from the original data points.

Table 6 logs the count of the attribute (AQI_Bucket) labels with 6 distinct types of values; they are moderate, satisfactory, good, poor, very poor, and severe. After multiple iterations used in the SMOTE algorithm, the values are much closer to each other. Delhi city did not have any "good" label values in the AQI_BUCKET column in the dataset, and hence, it is marked as 0. Similarly, in Bangalore, there are no "severe" label values in the AQI_BUCKET column and it is marked as 0. The SMOTE algorithm is being utilized in this paper to improve the accuracy of each model being run on the dataset, by balancing the datasets. An imbalanced dataset leads to a skewed class distribution that causes discrepancies inaccuracies of models. Higher accurate models, higher balanced accuracy, and higher balanced detection rate are produced by balanced datasets. Therefore, SMOTE is employed to accomplish this purpose and improve accuracy.

SMOTE has the benefit of not producing duplicate data points but rather artificial data points that are marginally different from the actual data points. By producing examples that are similar to the minority points already in existence, this algorithm aids in overcoming the overfitting issue caused by random oversampling. SMOTE also creates larger and less specific decision boundaries that increase the generalization capabilities of classifiers, thereby improving their performance.

Table 3: Sample dataset for Bangalore city.

City	Date	$PM_{2.5}$	PM_{10}	NO	NO_2	NO_x	NH_3	CO	SO_2	O_3	Benzene	Toluene	AQI	AQI_bucket
Bangalore	14/11/2015	42.42	156.84	7.25	29.94	31.78	21.94	1.56	2.23	31.35	1.82	4.65	130	Moderate
Bangalore	19/11/2015	21.99	39.86	7.08	16.44	19.51	41.96	1.73	2.95	9.98	1.52	2.38	103	Moderate
Bangalore	20/11/2015	13.89	31.44	6.84	12.14	15.35	23.93	1.72	2.5	4.56	0.74	1.48	74	Satisfactory
Bangalore	23/11/2015	19.66	36.84	6.47	16.37	20.87	24.04	1.35	2.83	4.09	1.18	2.17	75	Satisfactory
Bangalore	24/11/2015	20.35	33.97	7.76	20.64	24.75	26.98	1.36	2.59	7.77	1.02	1.9	85	Satisfactory
Bangalore	25/11/2015	34.39	36.29	8.38	28.8	32.28	32.75	2.48	3.76	14.63	1.32	3.17	141	Moderate
Bangalore	26/11/2015	43.91	43.65	11.74	29.33	32.78	55.4	1.52	3.44	14.8	1.53	3.59	90	Satisfactory
Bangalore	27/11/2015	44.14	112.78	7.05	26.64	27.06	32.33	2.18	4.3	25.57	1.69	3.36	126	Moderate
Bangalore	28/11/2015	44.94	114.34	8.47	28.1	29.37	32.75	2.3	4.7	29.1	1.56	2.38	147	Moderate
Bangalore	29/11/2015	29.35	75.79	5.72	21.21	21.4	19.08	1.55	4.55	29.03	1.01	1.15	87	Satisfactory

Table 4: Sample dataset for Kolkata city.

City	Date	$PM_{2.5}$	PM_{10}	NO	NO_2	NO_x	NH_3	CO	SO_2	O_3	Benzene	Toluene	AQI	AQI_bucket
Kolkata	16/06/2018	47.55	128.66	6.01	24.89	24.51	7.4	0.72	7.3	27.24	2.14	0.81	119	Moderate
Kolkata	18/06/2018	50.1	105.68	3.23	33.28	36.5	8.55	1.47	3.02	72.28	1.97	2.62	107	Moderate
Kolkata	19/06/2018	39.25	87.24	2.6	30.86	33.45	12.06	1.35	1.93	81.12	1.59	2.47	148	Moderate
Kolkata	20/06/2018	24.44	53.19	5.77	38.03	43.79	9.14	1.7	6.88	49.58	2.02	3.13	94	Satisfactory
Kolkata	21/06/2018	31.68	60.16	4.46	38.39	43.04	6.52	1.42	1.31	13.47	3.76	5.52	100	Satisfactory
Kolkata	22/06/2018	25.22	48.96	0.99	28.1	29.07	6.53	0.39	2.31	30.32	1.62	2.65	60	Satisfactory
Kolkata	23/06/2018	22.95	44.58	1.14	25.76	26.85	5.38	0.38	1.06	22.84	1.67	2.63	47	Good
Kolkata	24/06/2018	24.61	46.54	0.86	25.49	26.32	3.96	0.4	1.1	23.13	1.51	2.28	48	Good
Kolkata	25/06/2018	28.6	45.36	1.95	43.45	45.37	3.62	0.41	1.11	13.56	2.58	4.17	50	Good
Kolkata	26/06/2018	30.5	46.08	1.27	37.12	38.33	3.19	0.38	2.29	34.84	2.05	4.41	61	Satisfactory

Table 5: Sample dataset for Hyderabad city.

City	Date	$PM_{2.5}$	PM_{10}	NO	NO_2	NO_x	NH_3	CO	SO_2	O_3	Benzene	Toluene	AQI	AQI_bucket
Hyderabad	08/09/2015	91.82	32.94	5.41	28.93	23.37	24.94	0.48	7.98	27.04	1.01	5.74	179	Moderate
Hyderabad	09/09/2015	35.56	40.81	4.02	31.15	24.31	24.81	0.57	4.93	22.48	1.41	7.61	162	Moderate
Hyderabad	10/09/2015	45.64	44.89	7.06	28.96	25.58	24.8	0.73	5.29	24.69	1.25	7.84	76	Satisfactory
Hyderabad	11/09/2015	60.88	51.27	5.15	30.64	24.22	25.86	0.53	5.16	24.11	1.09	5.42	140	Moderate
Hyderabad	12/09/2015	65.61	41.31	3.4	26.03	20.37	24.78	0.57	5.44	25.47	0.83	4.39	128	Moderate
Hyderabad	13/09/2015	60.02	36.67	2.35	19.82	14.51	21.68	0.49	4.02	37.7	0.79	4.07	164	Moderate
Hyderabad	14/09/2015	73.21	35.28	2.82	19.94	15.4	21.4	0.57	5.96	34.11	0.52	2.44	169	Moderate
Hyderabad	01/10/2015	120.75	92.29	1.92	21.65	15.87	27.65	0.64	2.67	15.85	1.21	5.95	340	Very poor
Hyderabad	02/10/2015	29.66	76	2	25.94	16.02	20.45	0.6	3.81	17.4	1.2	5.62	125	Moderate
Hyderabad	03/10/2015	36.56	63.06	3.06	20.11	15.07	18.05	0.64	7.58	19.16	1.2	6.4	75	Satisfactory

The comparison of balanced and imbalanced datasets for the New Delhi, Bangalore, Kolkata, and Hyderabad cities is shown in Figures 2–5, respectively.

4. Methodology

In this paper, the proposed methods use three different algorithms to draw a comparative analysis of the AQI values of New Delhi, Bangalore, Kolkata, and Hyderabad by using parameters such as $PM_{2.5}$, PM_{10}, NO, NO_2, NO_x, NH_3, CO, SO_2, O_3, Benzene, and toluene levels, which will then compare the three algorithms and find the most accurate and efficient algorithm. The aim is to analyze and present it in an efficient way. It would help us discover interesting and insightful information. These particular cities have a higher population density and give a good estimate of the pollution in a major South Asian city. More cities have not been added due to the fact that it makes the research paper way too lengthy. Hence, the major cities of India have been chosen to analyze the pollution levels in different urban cities of India as they are the major contributors to pollution.

Figure 1: New minority class instances added.

Some of the existing algorithms used are Naive Bayes-a Bayes theorem-based classifier, support vector machine-a supervised learning model for classification and regression, artificial neural network-learning methodology inspired by actual

IMBALANCED DATASET
NEW DELHI:

Selected attribute
Name: AQI_Bucket Missing: 0 (0%) Distinct: 5 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	Severe	239	239
2	Moderate	485	485
3	Very Poor	514	514
4	Poor	534	534
5	Satisfactory	108	108

BALANCED DATASET

Selected attribute
Name: AQI_Bucket Missing: 0 (0%) Distinct: 5 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	Severe	478	478
2	Moderate	485	485
3	Very Poor	514	514
4	Poor	534	534
5	Satisfactory	432	432

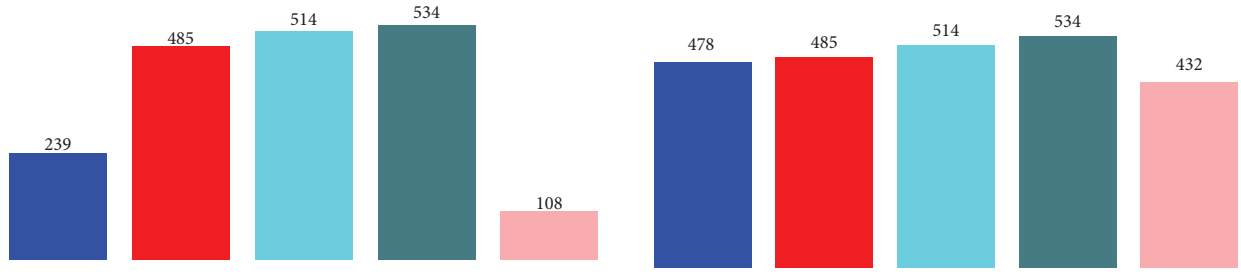

Figure 2: Balanced and imbalanced data values for New Delhi city.

IMBALANCED DATASET
BANGALORE:

Selected attribute
Name: AQI_Bucket Missing: 0 (0%) Distinct: 5 Type: Nominal Unique: 1 (0%)

No.	Label	Count	Weight
1	Moderate	479	479
2	Satisfactory	810	810
3	Poor	12	12
4	Good	59	59
5	Very Poor	1	1

BALANCED DATASET

Selected attribute
Name: AQI_Bucket Missing: 0 (0%) Distinct: 5 Type: Nominal Unique: 1 (0%)

No.	Label	Count	Weight
1	Moderate	958	958
2	Satisfactory	810	810
3	Poor	768	768
4	Good	944	944
5	Very Poor	1	1

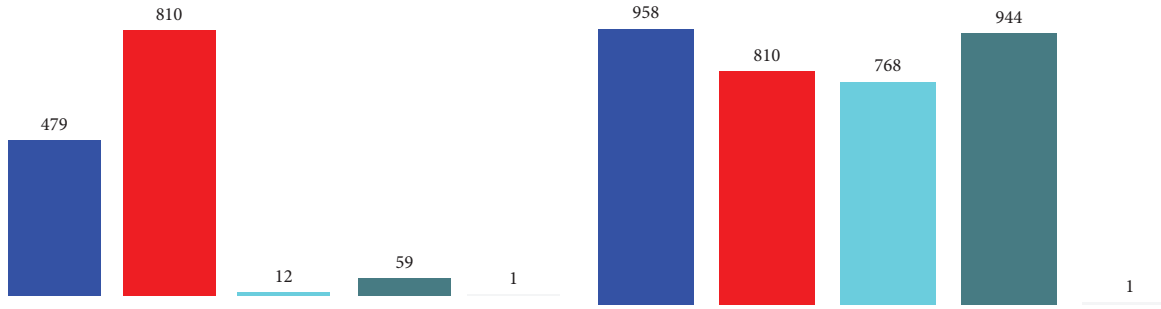

Figure 3: Balanced and imbalanced data values for Bangalore city.

neurons of the brain, gradient boost-techniques utilizing an ensemble of weak prediction models, decision tree-which works by making predictive models using data, and k-nearest neighbor-a lazy learning nonparametric supervised method.

The proposed algorithms used and compared are given below.

4.1. Synthetic Minority Oversampling Technique (SMOTE) Algorithm. Synthetic samples are created for the minority class using this oversampling technique. It aids in making an imbalanced dataset balanced. This approach helps with beating the issue of overfitting brought about by arbitrary oversampling.

4.2. Support Vector Regression. It is a discrete value prediction technique that uses supervised learning. For comparable purposes, SVMs and support vector regression are likewise used. Finding the most appropriate line is the main tenet of SVR. In SVR, the hyperplane with the most points is the line that fits the data the best.

IMBALANCED DATASET
KOLKATA:

Selected attribute
Name: AQI_Bucket Missing: 0 (0%) Distinct: 6 Type: Nominal Unique: 1 (0%)

No.	Label	Count	Weight
1	Moderate	151	151
2	Satisfactory	278	278
3	Good	119	119
4	Poor	119	119
5	Very Poor	66	66
6	Severe	13	13

BALANCED DATASET

Selected attribute
Name: AQI_Bucket Missing: 0 (0%) Distinct: 6 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	Moderate	302	302
2	Satisfactory	278	278
3	Good	238	238
4	Poor	238	238
5	Very Poor	264	264
6	Severe	208	208

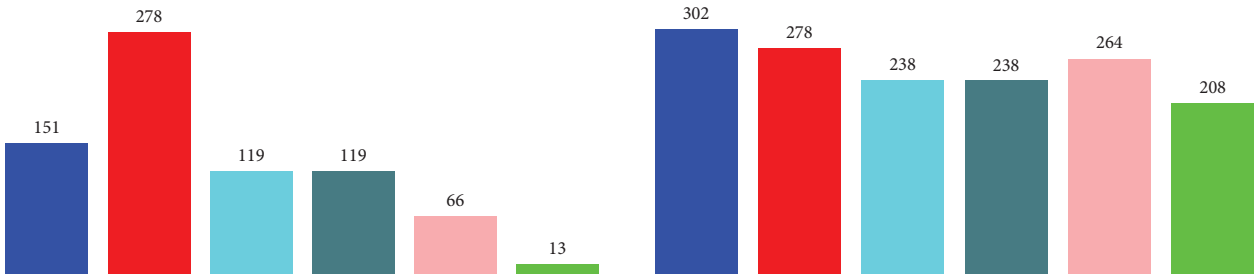

Figure 4: Balanced and imbalanced data values for Kolkata city.

IMBALANCED DATASET
HYDERABAD:

Selected attribute
Name: AQI_Bucket Missing: 0 (0%) Distinct: 6 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	Moderate	806	806
2	Satisfactory	645	645
3	Very Poor	3	3
4	Poor	30	30
5	Severe	4	4
6	Good	126	126

BALANCED DATASET

Selected attribute
Name: AQI_Bucket Missing: 0 (0%) Distinct: 6 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	Moderate	806	806
2	Satisfactory	645	645
3	Very Poor	768	768
4	Poor	960	960
5	Severe	1024	1024
6	Good	1008	1008

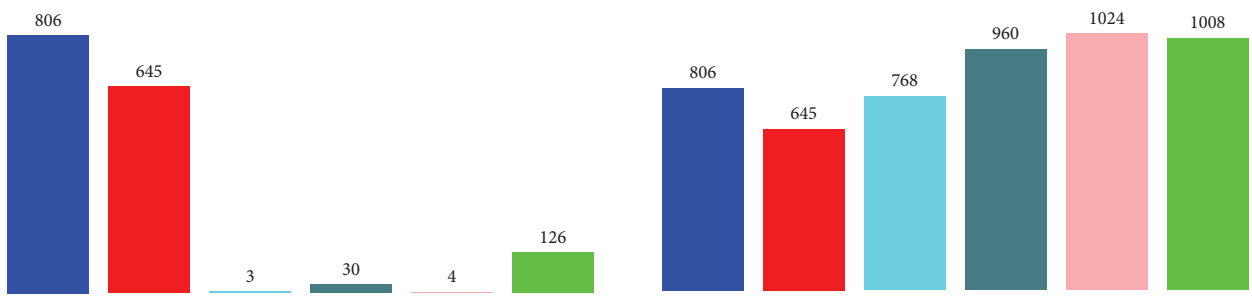

Figure 5: Balanced and imbalanced data values for Hyderabad city.

4.3. Random Forest Regression (RFR) Algorithm. It is a frequently used supervised machine-learning technique for classification and regression problems. It creates decision trees based on a variety of samples, utilizing the average for regression and the classification vote.

4.4. CatBoost Regression (CR) Algorithm. Yandex has developed a library of open-source software. It offers a framework for gradient boosting which, unlike the standard technique, aims at resolving categorical features using an alternative based on permutation.

All the three algorithms showed promising results in other works which had been studied through the literature survey. These three algorithms were chosen due to their high accuracy in previous different works (Table 1), and with the proposed work, the aim is to draw a comparative analysis and find the one with the best accuracy with balanced and imbalanced datasets. The aim is to use them and apply them

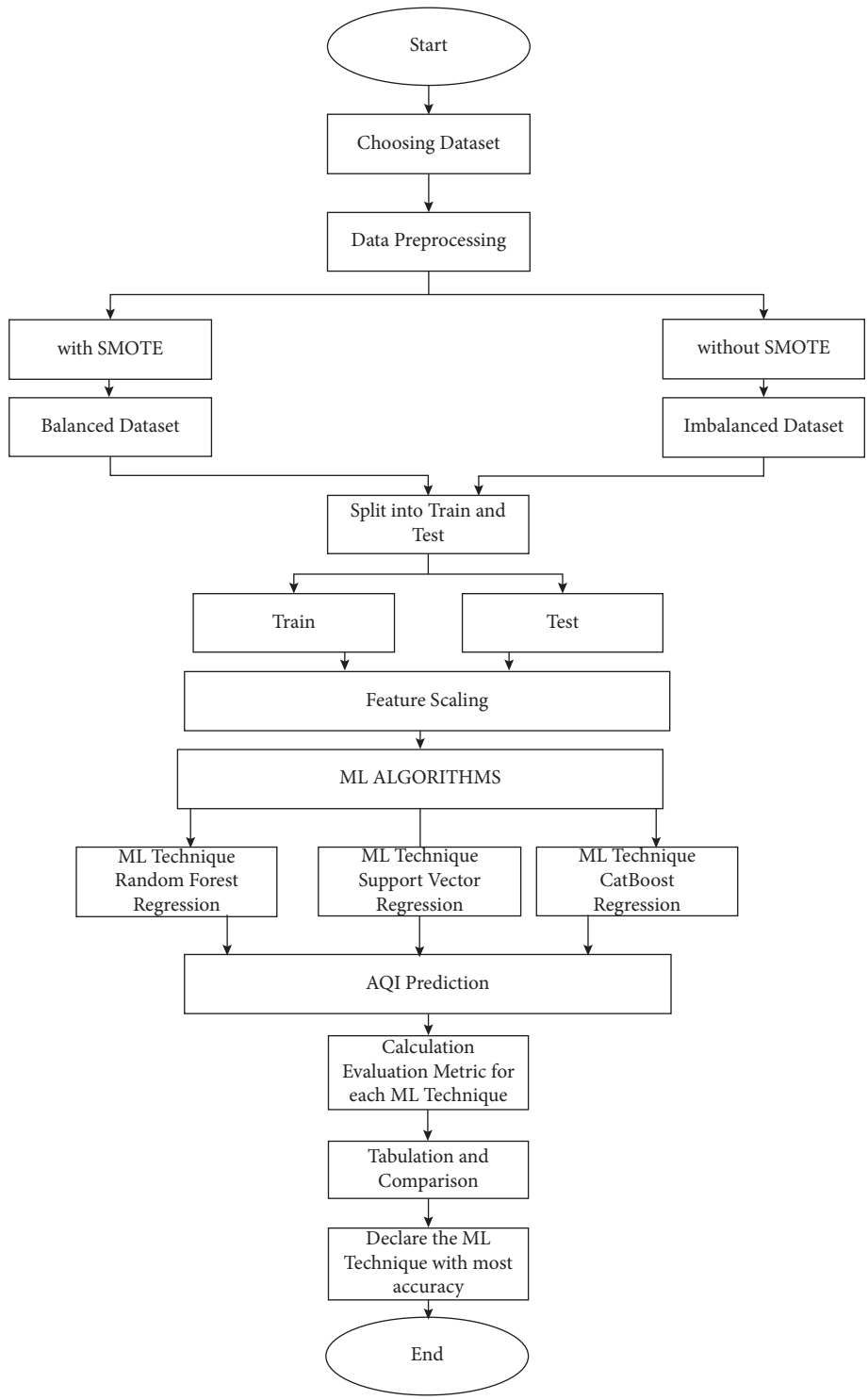

Figure 6: Flowchart for the proposed methodology.

to the Bangalore, Kolkata, Hyderabad, and New Delhi datasets and compare their accuracies to figure out what best fits our use case.

The picked algorithms have the highest accuracy based on our extensive literature survey as logged in Table 1, used for the AQI prediction. The algorithms being used for prediction are support vector regression (SVR), random forest regression (RFR), and CatBoost regression (CR). These algorithms will be provided with a suitably large dataset of cities, such as New Delhi, Bangalore, Kolkata, and Hyderabad, and will provide a practical environment.

The dataset used will be cleaned, reduced, and prepared according to our requirements and the data will be split into training and testing data. The plan is to use the simplest, most

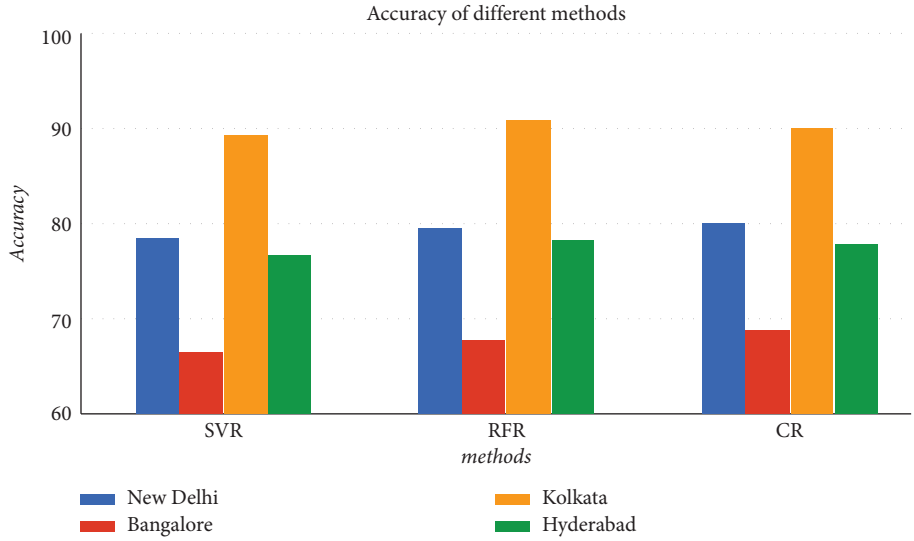

Figure 7: Accuracy comparison of algorithms for four cities.

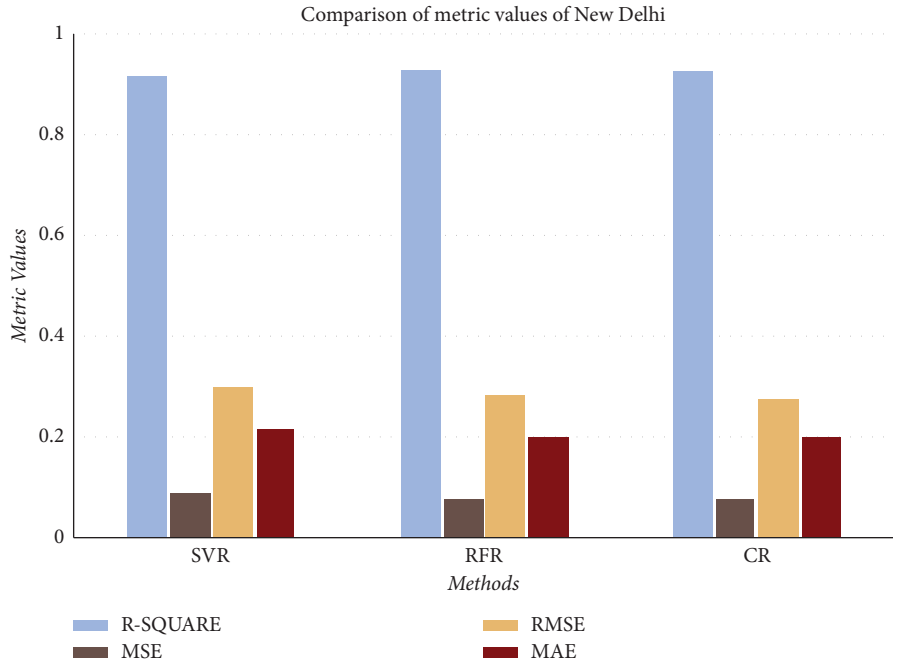

Figure 8: The comparison between R-square, MSE, RMSE, and MAE of support vector regression, random forest regression, and CatBoost regression of the New Delhi city imbalanced dataset.

straightforward implementation in order for the algorithms to be applied easily in a real-life use case. Then, different parameters will be taken to finalize and draw up a comparison between these 3 algorithms and then come to the conclusion to show which is the most accurate. The comparison can bring out important information about AQI prediction methods and even help us choose the most suitable one. A comparison of the accuracy levels obtained with an imbalanced dataset and a balanced dataset with the help of the SMOTE algorithm will also be done.

Hence, the methodology is a step-by-step process in which the first step is to find a suitable dataset and clean it. After this, further data preprocessing is applied which makes use of SMOTE in order to balance the dataset. Both balanced and imbalanced datasets will be preserved and used in order to bring to light any differences in performance that may arise due to balancing. Following this, in a standard machine learning procedure, the dataset is split into train and test to train the models and test their accuracies against real data. Feature scaling and normalization are carried out.

Now, each regression model which has been picked, namely, random forest, support vector regression, and CatBoost, are used for prediction and its accuracy is gauged, for each balanced and imbalanced dataset as mentioned previously. They are compared using metrics such as RMSE and R-SQUARE. Finally, all the data and results have been

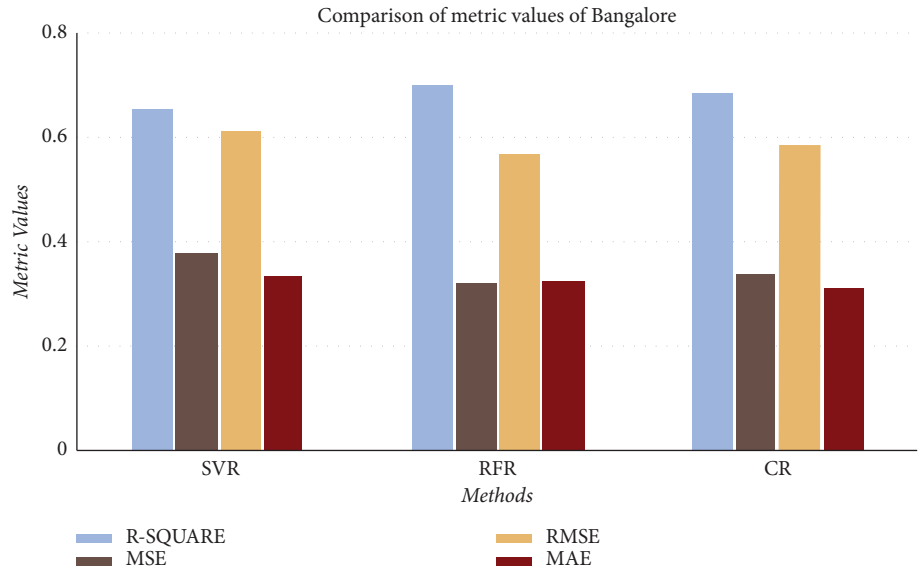

Figure 9: The comparison between R-square, MSE, RMSE, and MAE of support vector regression, random forest regression, and CatBoost regression for the Bangalore city imbalanced dataset.

displayed using clear figures, graphs, and charts which easily make one understand what exactly has led to the increase in accuracy and hence help future research.

Figure 6 shows the various steps which will be performed during the implementation of this work to achieve the determined result. The flowchart is a process-based flowchart that shows the steps of the process in a detailed manner. It has been derived from the actual working out into running these models and extracting results. The process flowchart is drawn in Western ANSI standards.

Step 1. Choosing a dataset

Choosing an extensive dataset from Kaggle according to our requirements and downloaded its CSV file.

Step 2. Data preprocessing

In data preprocessing, they cleaned the original dataset and extracted the New Delhi, Bangalore, Kolkata, and Hyderabad city data. Because these are major cities in India, it is important to analyze the pollution levels in different urban cities in India as they are the major contributors to the pollution. These particular cities have a higher population density and give a good estimate of the pollution. Each of these datasets was cleaned by removing all null value rows, and the attribute xylene was removed from the dataset due to the fact that the column values were empty for all 4 cities chosen, hence making it a redundant attribute. Microsoft Excel software is used to remove unnecessary, irrelevant, and erroneous data.

Step 3. Applying the SMOTE algorithm

After the cleaning of the dataset, the synthetic minority oversampling technique (SMOTE) is used to correct the class imbalances in the AQI_Bucket values. Delhi, Bangalore, Kolkata, and Hyderabad required 3, 11, 9, and 24 manual iterations to achieve a suitable level of balance. This is carried out to create a balanced version of the dataset.

Step 4. Not applying the SMOTE algorithm

Here, the synthetic minority oversampling technique (SMOTE) is not applied to the dataset it is being used directly just after removing unnecessary, irrelevant, and erroneous data in it and hence is in its imbalanced form.

Step 5. Splitting of the dataset

The datasets are split into training and test data at an 80 : 20 ratio. These are used to train the model and then test it against the original data. The values predicted by the machine learning algorithms are corroborated with the original data to predict accuracy.

Step 6. Training the dataset

Empirical studies show that the best results are obtained if 80% of the data is used for training. Random sampling is used as a way to divide the data into train and test sections. It is widely accepted and is very popular.

Step 7. Testing the dataset

Empirical studies show that the best results are obtained if the remaining 20% of the data is used for testing. Random sampling is used as a way to divide the data into train and test sections. It is widely accepted and is very popular.

Step 8. Feature scaling

The data have been normalized in order to make the dataset flexible and consistent. StandardScaler from Scikit-Learn Library has been used to do so. It normalizes the features by deleting the mean and scaling the unit variance.

Step 9. Applying machine learning (ML) techniques

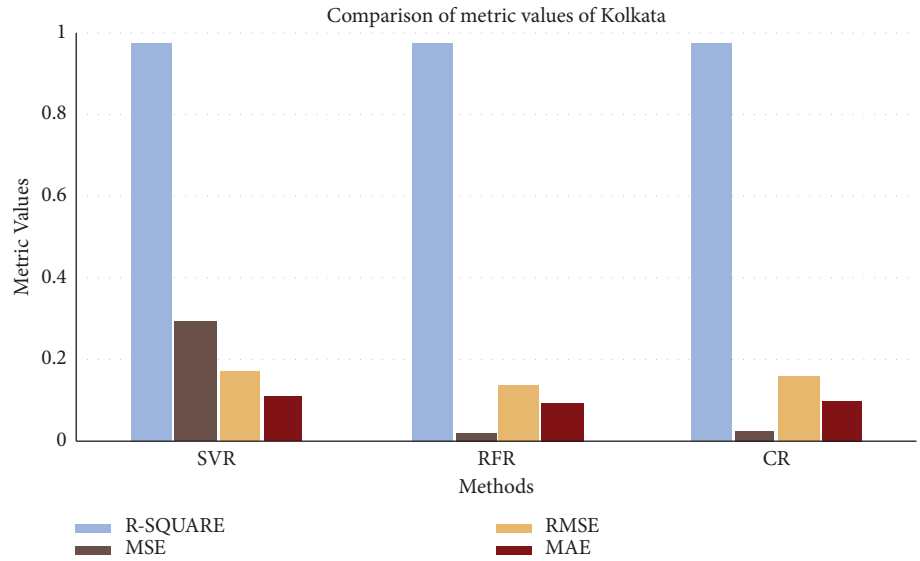

Figure 10: The comparison between R-square, MSE, RMSE, and MAE of support vector regression, random forest regression, and CatBoost regression for the Kolkata city imbalanced dataset.

After normalizing the range of features in the datasets, various algorithms, namely, CatBoost regression, random forest regression, and support vector regression are used to forecast air quality index, and then, they are compared to show which algorithm gives the best accuracy level for each city, respectively.

Step 10. Applying ML technique-random forest regression

Random forest is a supervised machine learning algorithm that is used for classification and regression problems. It creates decision trees from several samples, using the majority vote for classification and the average in the case of regression. A random forest produces precise predictions that are easy to understand. Effective handling of large datasets is possible.

Step 11. Applying ML technique-support vector regression

Support vector regression is a supervised machine learning algorithm that is used for regression problems. Discrete values can be predicted using it. The core idea of SVR is locating the best fit line. The SVR best-fitting line is the hyperplane with the most points. The flexibility of SVR allows us to decide how much error in our model is acceptable.

Step 12. Applying ML technique-CatBoost regression

A supervised machine learning approach called CatBoost regression is based on gradient-boosted decision trees. During training, a number of decision trees are constructed progressively. To generate a powerful, competitive predictive model through greedy search, the main objective of boosting is to successively integrate a large number of weak models or models that only marginally outperform chance. It has a quick inference process since it uses symmetric trees and its boosting techniques aid in lowering overfitting and enhancing model quality.

Step 13. AQI prediction

Machine learning techniques are used to aid in this process, and the accuracy level of AQI for each city is estimated. The values are tabulated and graphs depicting the accuracy levels of all 4 cities are plotted.

Step 14. Calculation of evaluation metric for each ML technique

The metrics used for the proposed work are R-SQUARE, MSE, RMSE, MAE, and the accuracy (1-MAE) of CatBoost regression, random forest regression, and support vector regression.

Step 15. Tabulation and comparison

Taking all the metric values obtained after running the machine learning techniques (i.e.,) R-SQUARE, MSE, RMSE, MAE, and the accuracy of the algorithms. For comparison tabulating, the predicted values and actual values for each city and model and plot multiple graphs such as line graphs, density plots, and scatter plots are analyzed. All metric values and accuracy values of each city and model are further tabulated, plotting bar graphs to compare the accuracy of each model city-wise and also plot bar graphs to compare R-SQUARE, MSE, RMSE, and MAE values of each model city-wise. Here, the accuracy is calculated using various cities datasets with SMOTE applied to them, repeating the same steps from Step 10 to Step 15 after using the dataset with the SMOTE algorithm applied.

Step 16. Final comparative results (declare the ML technique with the highest accuracy)

Once tabulated all the values, the next step is to compare the metric values of all the used algorithms and see what best fits the scenario. In the proposed work, random forest and CatBoost regression are the best performances overall. RFR

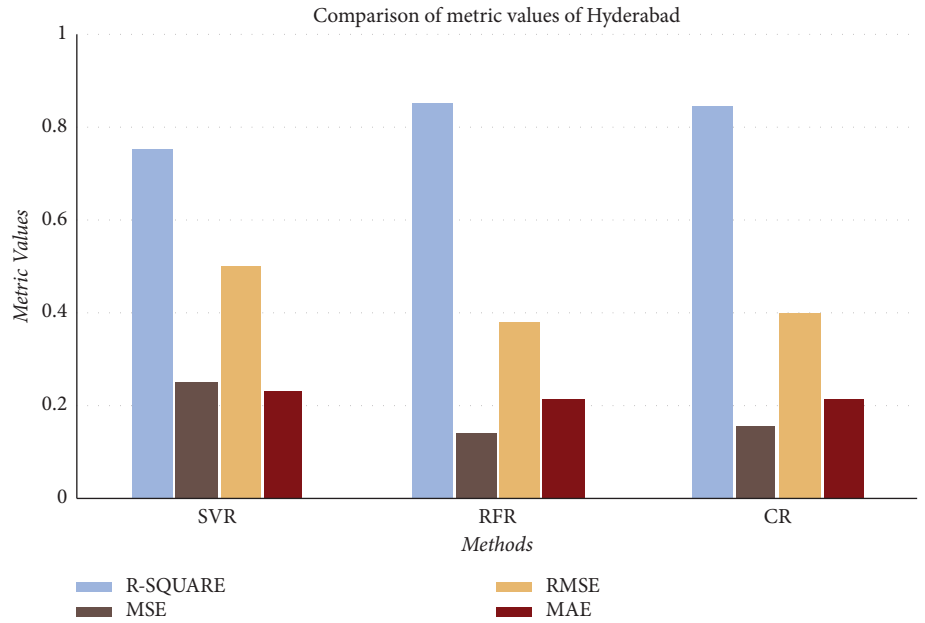

Figure 11: The comparison between R-square, MSE, RMSE, and MAE of support vector regression, random forest regression, and CatBoost regression for the Hyderabad imbalanced dataset.

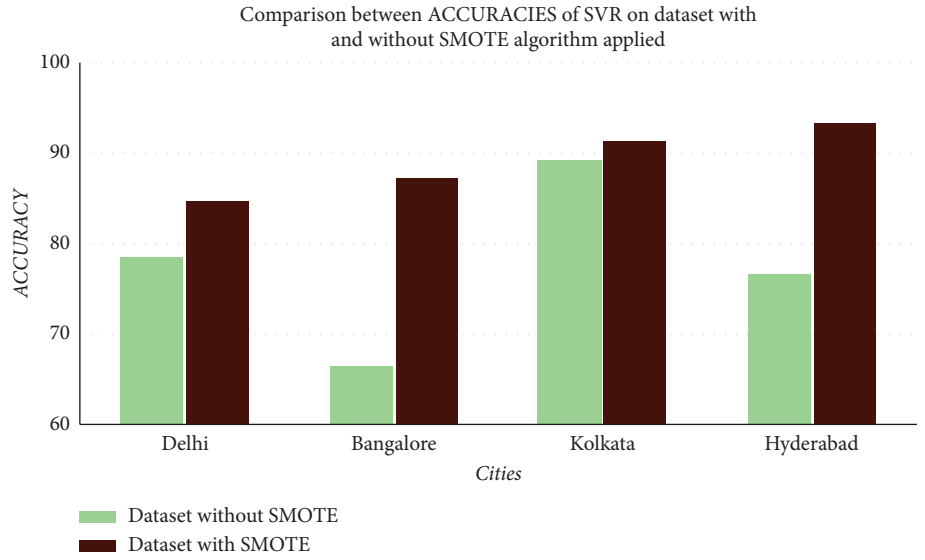

Figure 12: Comparison between the accuracy of SVR on the balanced and imbalanced dataset (with and without using the SMOTE algorithm).

got the best RMSE values in Bangalore, Kolkata, and Hyderabad, whereas CatBoost regression performed best in Delhi. The highest accuracy was obtained by random forest regression for the cities of Kolkata and Hyderabad and New Delhi and Bangalore. CatBoost regression gave the highest accuracy. The tabulated values are compared with metric values before and after applying SMOTE on the dataset to find what gives better accuracy. In the proposed work, random forest and CatBoost were the best performances overall.

5. Discussion on Metrics Used

The metrics used in the proposed work are R-SQUARE, mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and accuracy.

(i) R-SQUARE indicates to what extent the regression model is in line with the observed data. A higher *R* square value denotes a better model fit, the *R* Square equation is shown by equation

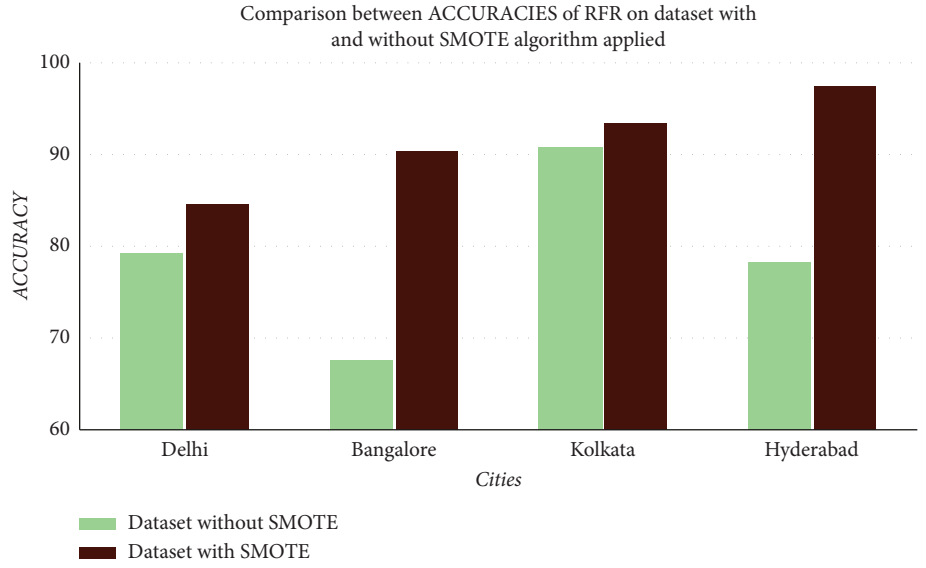

Figure 13: Comparison between the accuracy of RFR on a dataset with and without the SMOTE algorithm.

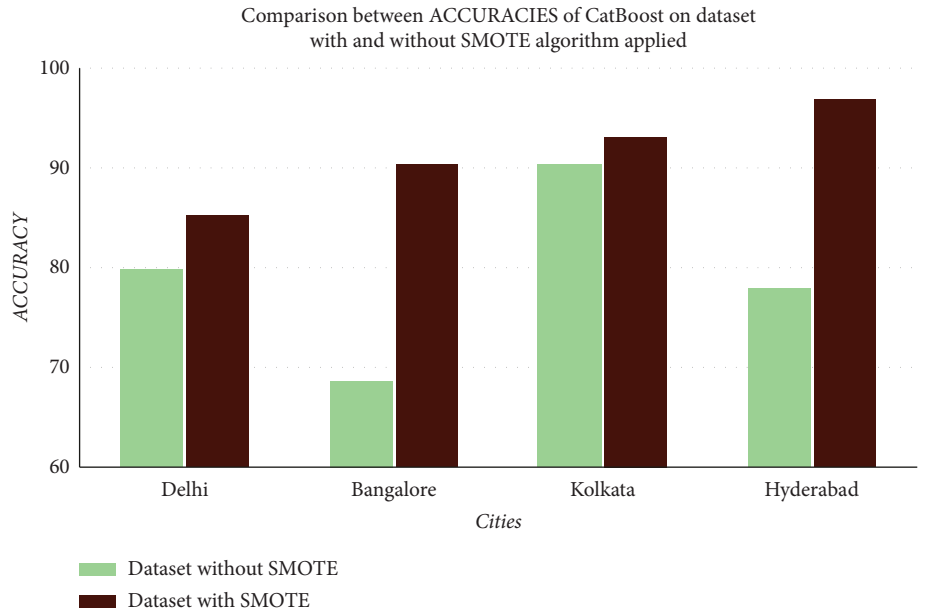

Figure 14: Comparison between the accuracy of CR on a dataset with and without the SMOTE algorithm.

$$R - \text{SQUARE} = \frac{\text{SSregr}}{SStt}. \tag{1}$$

The sum of squares due to regression is denoted by *SS*regr (explained sum of squares), while the sum of squares overall is denoted by *SStt*. The degree to which the regression model fits the data well is shown by the sum of squares due to regression. The total sum of squares is used to determine how much the observed data has changed (data utilized in regression modeling).

(ii) MSE is a parameter that measures how closely a fitted line resembles a set of data points. The lower the value, the closer it is to the line, and hence the better. If the MSE value = 0, the model is perfect. It is shown in equation

$$MSE = \sum_{i-1}^{n} \frac{\left(X_i - X_i^{\wedge}\right)^2}{n}, \tag{2}$$

where $A = \pi r^2$,

(a) x_i = The i^{th} observed value
(b) $x_i^{\wedge}$ = The corresponding predicted value
(c) n = The number of observations

(iii) RMSE indicates how densely the data are distributed along the line of best fit. RMSE values in the

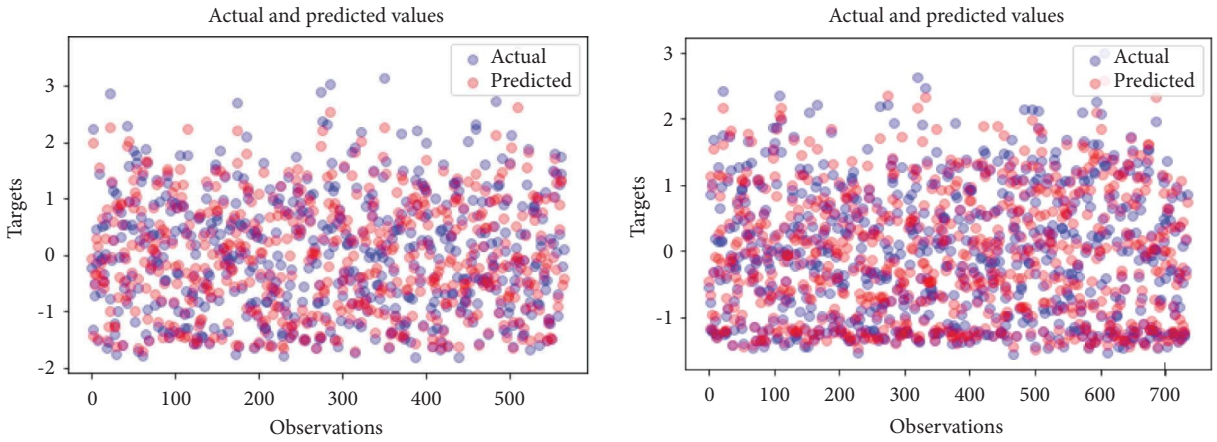

Figure 15: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for New Delhi-SVR.

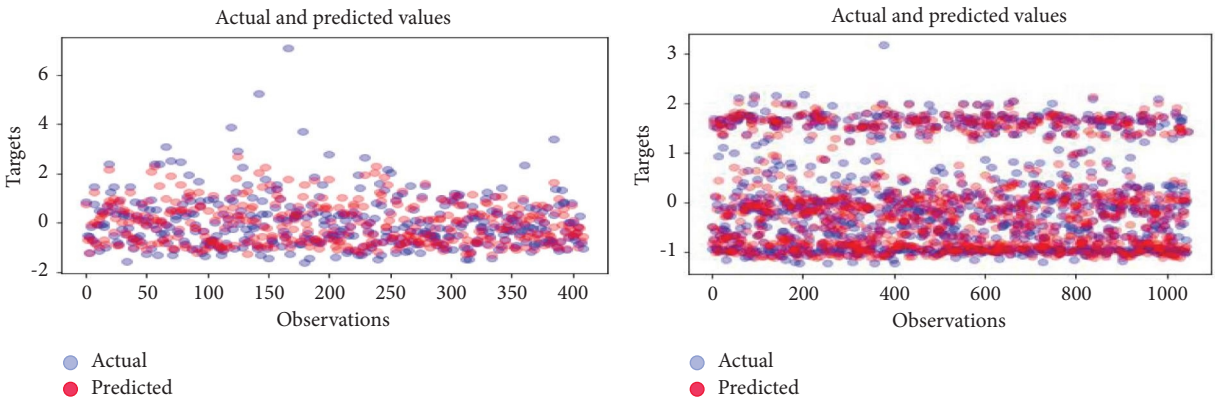

Figure 16: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Bangalore-SVR.

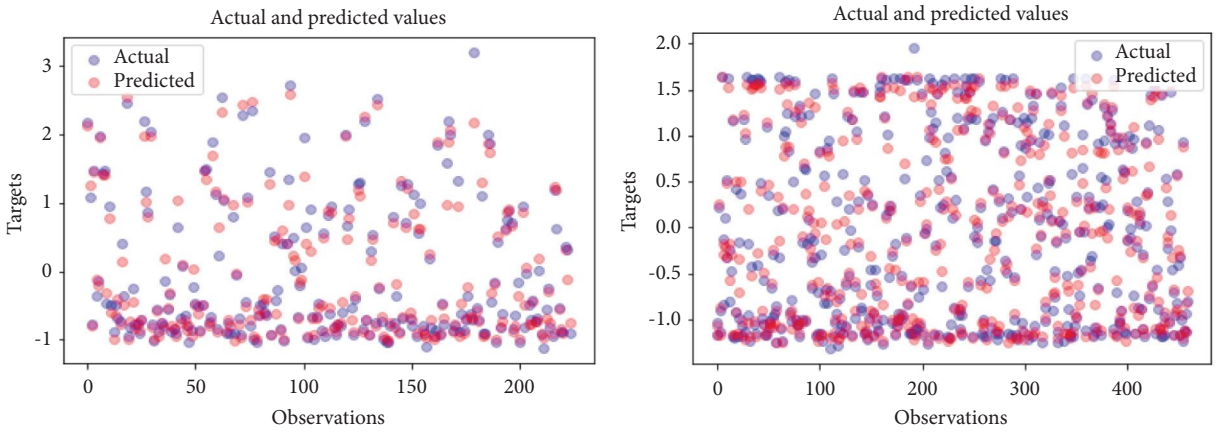

Figure 17: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Kolkata-SVR.

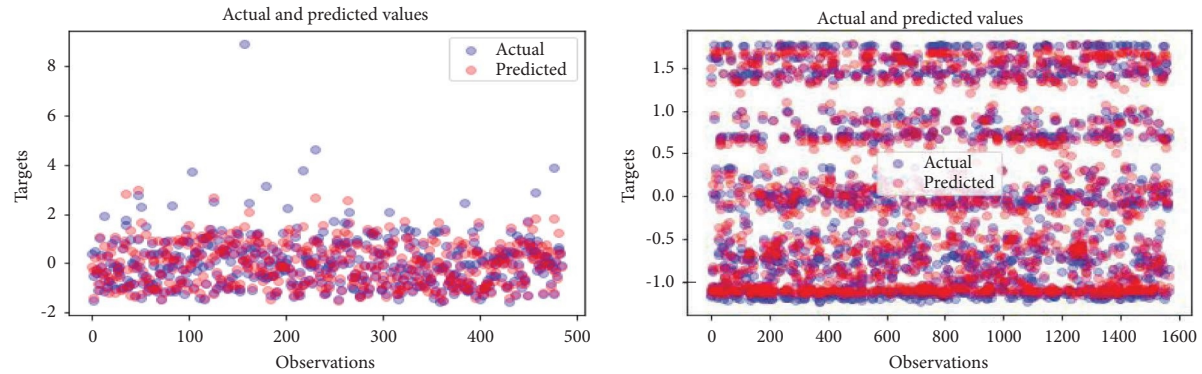

Figure 18: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Hyderabad-SVR.

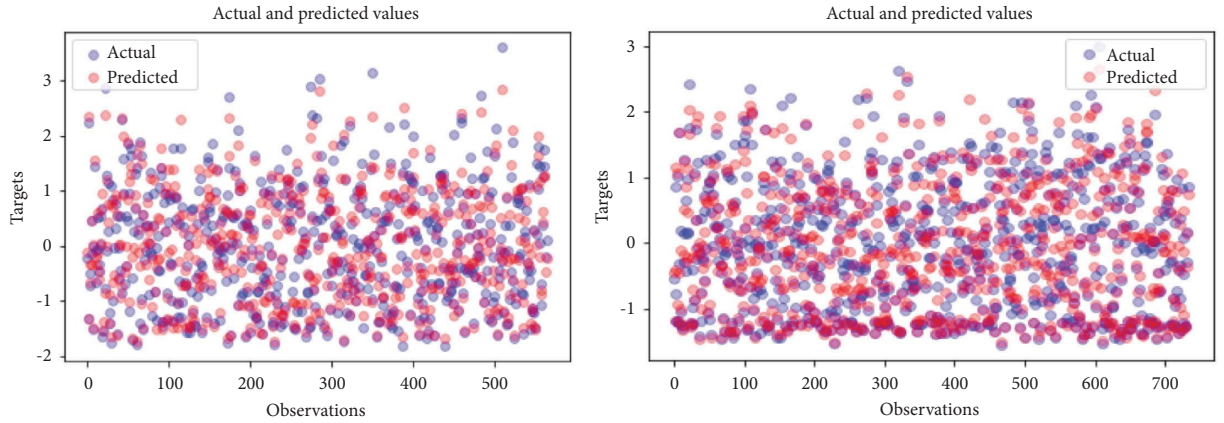

Figure 19: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for New Delhi-RFR.

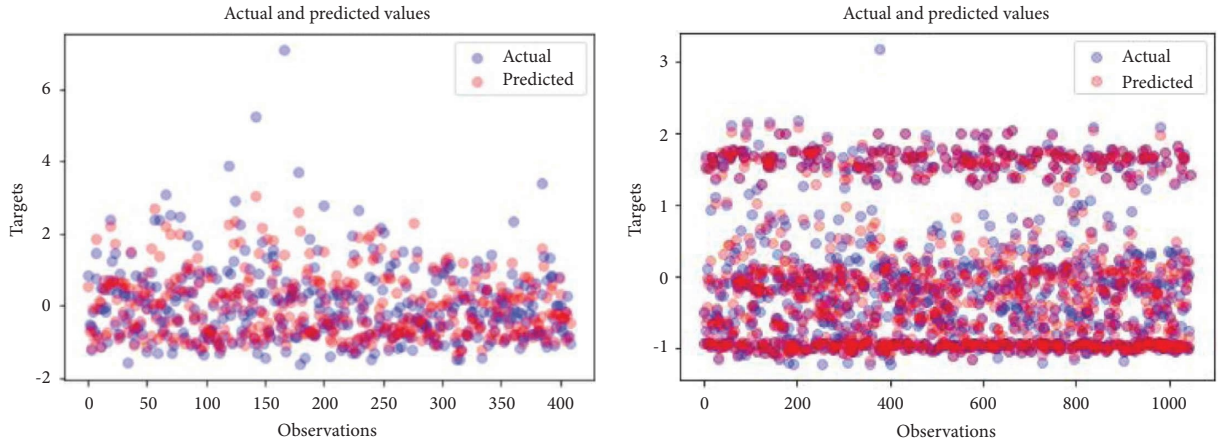

Figure 20: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Bangalore-RFR.

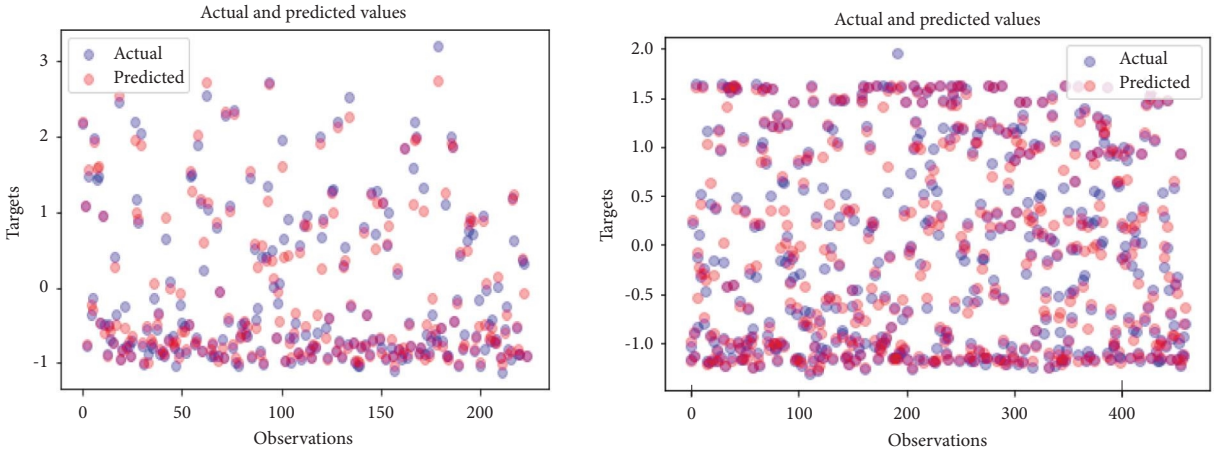

Figure 21: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Kolkata–RFR.

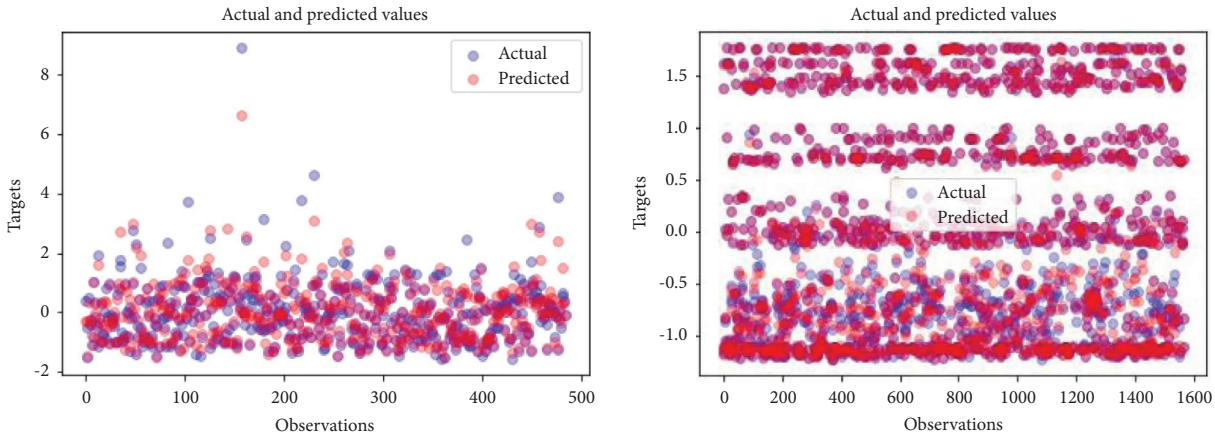

Figure 22: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Hyderabad–RFR.

range of 0.2–0.5 demonstrate that the model can reasonably predict the data. It is shown in the equation

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \hat{X_i})^2}{m}}, \tag{3}$$

where

(a) x_i = The i^{th} observed value
(b) $\hat{x_i}$ = The corresponding predicted value
(c) n = The number of observations

(iv) MAE evaluates the absolute distance of the observations to the predictions on the regression line. It is shown in the equation

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^{n} |X_i - X|, \tag{4}$$

where

(a) n is the number of errors
(b) Σ is the summation symbol (which means "add them all up")
(c) $|x_i - x|$ is the absolute errors

(v) Accuracy is used as a measurement to calculate how well a model is finding patterns and identifying relations in the dataset and it is shown in the equation

$$\text{Accuracy} = (1 - MAE) * 100. \tag{5}$$

This gives the accuracy in percentage.

6. Results and Discussion

In the proposed work, the dataset mentioned above has been cleaned such that it only has the values for the cities of New Delhi, Bangalore, Kolkata, and Hyderabad. The dataset was used in two ways, once in an imbalanced version and then in a balanced version using SMOTE. Graphs were plotted and it was seen that there was an increase in the accuracies of the models which had the balanced dataset. For prediction purposes, three algorithms were run on it, namely, support

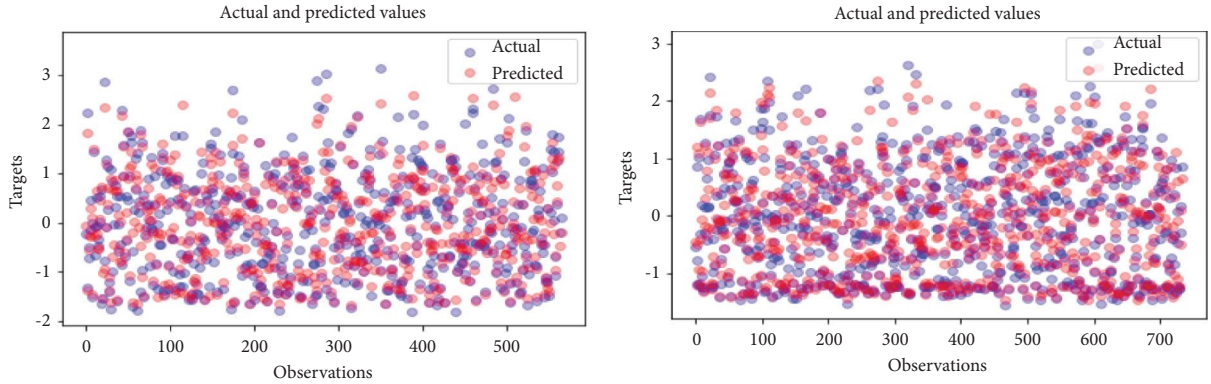

FIGURE 23: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for New Delhi–CR.

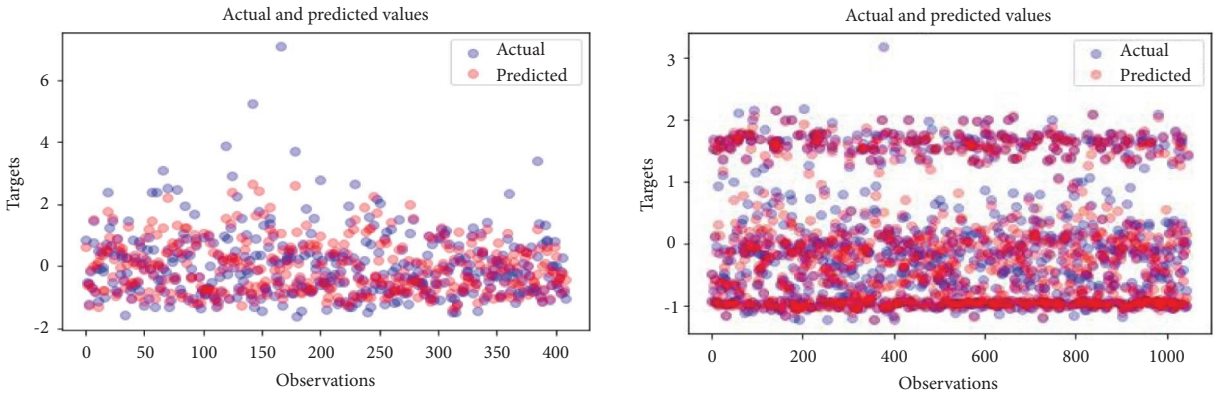

FIGURE 24: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Bangalore–CR.

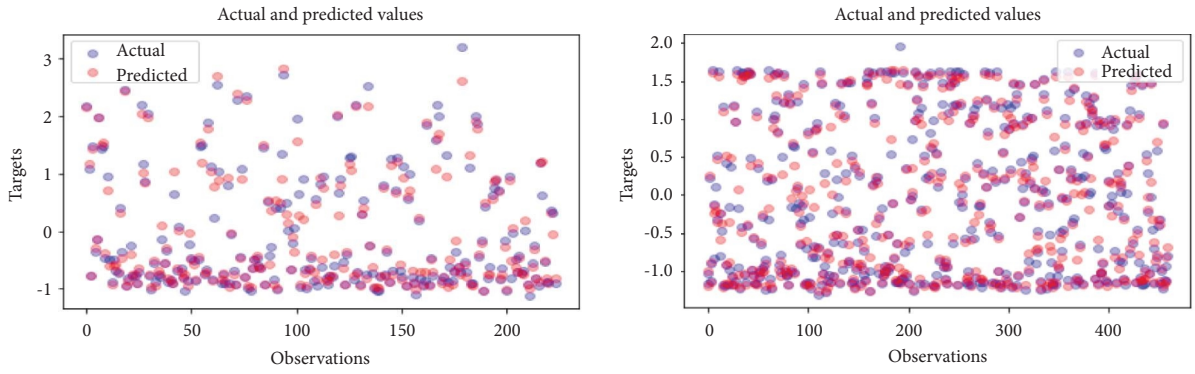

FIGURE 25: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Kolkata-CR.

vector regression, random forest regression, and CatBoost regression. Plotted graphs between the test data and the predicted data were shown as well. The metrics calculated in each algorithm are R-SQUARE, mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). Comparative tables, graphs and scatter plots were drawn for balanced and imbalanced dataset results to show how using a balanced dataset when used provides higher accuracies in each algorithm.

According to the research in this paper, the choice to use statistical metrics, such as RMSE, R-SQUARE and so on, has been understood and referred to in papers [30–33], as well as how to effectively implement them. Metrics are used to track and gauge a model's performance (during training and testing). These metrics provide information on the precision of the forecasts, and the amount of departure from the actual values since all of the algorithms utilized are based on regression models.

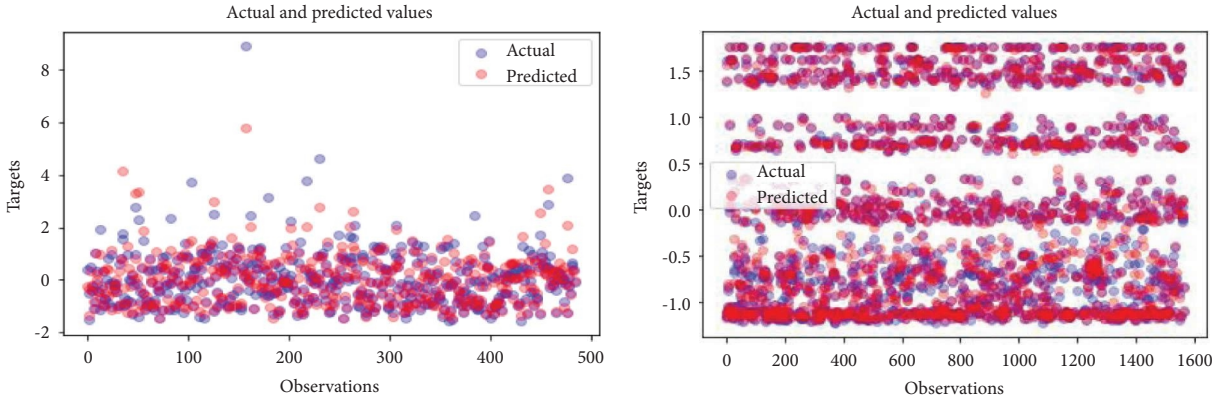

Figure 26: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Hyderabad–CR.

Accuracy results' comparison of the imbalanced dataset without using SMOTE algorithm for all the 4 cities such as Delhi, Bangalore, Kolkata, and Hyderabad obtained by the machine learning techniques such as support vector regression, random forest regression, and CatBoost regression is shown in Table 7. Among the four cities, the Kolkata city dataset gives the maximum accuracy for these three algorithms, whereas the Bangalore city dataset gives the minimum accuracy. The dataset used was imbalanced.

Figure 7 depicts the accuracy achieved by various ML techniques such as SVR, RFR, and CR to estimate AQI in four different cities using a bar graph.

Table 8 logs the result of performance metrics used for the New Delhi city imbalanced dataset (i.e.,) without using SMOTE algorithm are R-SQUARE, MSE, RMSE, and MAE values for all 3 algorithms such as support vector regression, random forest regression, and CatBoost regression. The CatBoost regression algorithm gives the best result in comparison to support vector regression and random forest regression.

In Figure 8, the comparison between R-SQUARE, MSE, RMSE, and MAE of support vector regression, random forest regression and CatBoost regression of New Delhi city imbalanced dataset (i.e.,) without using SMOTE algorithm through graphical representation is shown. It depicts that CatBoost regression has the highest R-SQUARE, and the lowest RMSE, MSE, and MAE values.

Table 9 logs the result of performance metrics used for the Bangalore imbalanced dataset (i.e.,) without using the SMOTE algorithm are R-SQUARE, MSE, RMSE, and MAE values for all 3 algorithms such as support vector regression, random forest regression, and CatBoost regression. The random forest regression gives the best result when compared to support vector regression and CatBoost regression except for the fact that CatBoost regression gives a lesser MAE than random forest regression.

In Figure 9, the comparison between R-SQUARE, MSE, RMSE, and MAE of support vector regression, random forest regression, and CatBoost regression is shown. It depicts that random forest regression has the highest R-SQUARE, and the lowest RMSE, MSE value and CatBoost regression has the lowest MAE value.

Table 10 logs the result of performance metrics used for the Kolkata city imbalanced dataset (i.e.,) without using SMOTE algorithm are R-SQUARE, MSE, RMSE, and MAE values for all 3 algorithms such as support vector regression, random forest regression, and CatBoost regression. The random forest regression gives the best result in comparison to the support vector regression and CatBoost regression algorithm.

In Figure 10, the comparison between R-SQUARE, MSE, RMSE, and MAE of support vector regression, random forest regression and CatBoost regression for Kolkata city imbalanced dataset (i.e.,) without using SMOTE algorithm is shown. It depicts that CATBOOST has the highest R-SQUARE, and the lowest RMSE, MSE, and MAE values.

Table 11 logs the result of performance metrics used for the Hyderabad city imbalanced dataset (i.e.,) without using the SMOTE algorithm are R-SQUARE, MSE, RMSE, and MAE values for all 3 algorithms such as support vector regression, random forest regression, and CatBoost regression. The random forest regression gives the best result in comparison to the support vector regression and CatBoost regression.

In Figure 11, the comparison between R-SQUARE, MSE, RMSE, and MAE of support vector regression, random forest regression, and CatBoost regression imbalanced dataset (i.e.,) without using the SMOTE algorithm is shown. It depicts that random forest regression has the highest R-SQUARE, and the lowest RMSE, MSE, and MAE values.

Accuracy results comparison of the balanced dataset using SMOTE algorithm for all the 4 cities such as Delhi, Bangalore, Kolkata, and Hyderabad obtained by the machine learning techniques such as support vector regression, random forest regression, and CatBoost regression are shown in Table 12. Among the four cities, the Hyderabad city dataset gives the maximum accuracy for these three algorithms, whereas the New Delhi city dataset gives the minimum accuracy. In the proposed work, the original

Table 6: Comparison of dataset size with and without the SMOTE algorithm.

	Imbalanced dataset size (not using the SMOTE algorithm)				Balanced dataset size (using the SMOTE algorithm)			
	Cities							
AQI_bucket values	Delhi	Bangalore	Kolkata	Hyderabad	Delhi	Bangalore	Kolkata	Hyderabad
	Size							
Moderate	485	479	151	806	485	958	302	806
Satisfactory	108	810	278	645	432	810	278	645
Good	0	59	119	126	0	944	238	1008
Poor	534	12	119	30	534	768	238	960
Very poor	514	1	66	3	514	1	264	768
Severe	239	0	13	4	478	0	208	1024

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

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

dataset is used and SMOTE is applied to it as mentioned above and cleaned it to only have the values for cities New Delhi, Bangalore, Kolkata, and Hyderabad. 3 algorithms have been implemented on it such as support vector regression, random forest regression, and CatBoost regression for prediction purposes, and plotted graphs between the test data and the predicted data as well.

Table 13 shows a comparison of SVR accuracy with and without SMOTE algorithm of four cities. Bangalore city has the lowest accuracy of 66.46% and Kolkata city has the highest accuracy of 89.17% from the dataset without SMOTE algorithm. Hyderabad city has the highest accuracy of 93.57% and New Delhi city has the lowest accuracy of 84.83% from the dataset with SMOTE algorithm. It is clearly observed that the dataset with SMOTE algorithm applied has higher accuracies. It is shown in Figure 12.

The accuracy comparison of SVR, RFR, and CR on balanced and imbalanced datasets (i.e., with and without using SMOTE algorithm) is shown in Figures 12–14. The accuracy for the balanced datasets for the four cities are increased when compared to the accuracy for the imbalanced datasets. The scatter plots show the actual and predicted values for an imbalanced dataset (without using SMOTE) and a balanced dataset (with using SMOTE) using SVR. The scatter plots for the four cities such as New Delhi, Bangalore, Kolkata, and Hyderabad for SVR are shown in Figures 15–18.

Table 10: The result of performance metrics used for Kolkata city imbalanced dataset,vwithout using the SMOTE algorithm.

Algorithm name	R-square	MSE	RMSE	MAE
Support vector regression	0.9714	0.2942	0.1715	0.1083
Random forest regression	0.9808	0.0197	0.1403	0.0902
CatBoost regression	0.9752	0.0255	0.1597	0.1002

Table 11: The result of performance metrics used for the Hyderabad city imbalanced dataset, without using the SMOTE algorithm.

Algorithm name	R-square	MSE	RMSE	MAE
Support vector regression	0.7599	0.2512	0.5012	0.2332
Random forest regression	0.8600	0.1464	0.3826	0.2163
CatBoost regression	0.8474	0.1596	0.3995	0.2210

Table 14 shows a comparison of RFR accuracy with and without SMOTE algorithm of four cities. Bangalore city has the lowest accuracy of 67.70% and Kolkata city has the highest accuracy of 90.97% from the dataset without SMOTE algorithm. Hyderabad city has the highest accuracy of 97.61% and New Delhi city has the lowest accuracy of 84.73% from the dataset with SMOTE algorithm. It is clearly observed that the dataset with SMOTE algorithm applied has higher accuracies. It is shown in Figure 13.

The accuracy comparison of RFR on balanced and imbalanced datasets (i.e.,) with and without using SMOTE

Table 12: 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

Table 13: 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

Table 14: Comparison of RFR accuracy with and without the SMOTE algorithm of four cities.

Cities	RFR accuracy (not using SMOTE algorithm, imbalanced dataset) (%)	RFR accuracy (using SMOTE algorithm, balanced dataset) (%)
New Delhi	79.4764	84.7284
Bangalore	67.7038	90.3071
Kolkata	90.9700	93.7438
Hyderabad	78.3672	97.6080

Table 15: Comparison of CR accuracy with and without the SMOTE algorithm of four cities.

Cities	CR accuracy (not using SMOTE algorithm, imbalanced dataset) (%)	CR accuracy (using SMOTE algorithm, balanced dataset) (%)
New Delhi	79.8622	85.0847
Bangalore	68.6860	90.3343
Kolkata	89.9766	93.1656
Hyderabad	77.8991	96.7529

Table 16: Overall comparison between accuracy values of the dataset with and without SMOTE algorithm of four cities.

	Cities							
Method	Delhi	Bangalore	Kolkata	Hyderabad (%)	Delhi	Bangalore	Kolkata	Hyderabad
	Accuracy of the imbalanced dataset (without SMOTE algorithm) (%)				Accuracy of the balanced dataset (with SMOTE algorithm) (%)			
SVR	78.4867	66.4564	89.1656	76.6786	84.8332	87.1756	91.5624	93.5658
RFR	79.4764	67.7038	90.9700	78.3672	84.7284	90.3071	93.7438	97.6080
CatBoost	79.8622	68.6860	89.9766	77.8991	85.0847	90.3343	93.1656	96.7529

algorithm is shown in Figure 13. The accuracy for the balanced datasets for the four cities are increased when compared to the accuracy for the imbalanced datasets. The scatter plots show the actual and predicted values for an imbalanced dataset (without using SMOTE) and a balanced dataset (with using SMOTE) using RFR. The scatter plots for the four cities such as New Delhi, Bangalore, Kolkata, and Hyderabad are shown in Figures 19–22.

Table 15 shows a comparison of CR accuracy with and without SMOTE algorithm of four cities. Bangalore city has the lowest accuracy of 68.69% and Kolkata city has the highest accuracy of 89.98% from the dataset without the SMOTE algorithm. Hyderabad city has the highest accuracy of 96.75% and New Delhi city has the lowest accuracy of 85.08% from the dataset with the SMOTE algorithm. It is clearly observed that the dataset with SMOTE algorithm applied has higher accuracies. It is shown in Figure 14.

The accuracy comparison of CR on balanced and imbalanced datasets (i.e.,) with and without using SMOTE algorithm is shown in Figure 14. The accuracy for the balanced datasets for the four cities is increased when compared to the accuracy for the imbalanced datasets. The scatter plots show the actual and predicted values for an

imbalanced dataset (without using SMOTE) and a balanced dataset (with using SMOTE) using CR. The scatter plots for the four cities such as New Delhi, Bangalore, Kolkata, and Hyderabad are shown in Figures 23–26, respectively.

Table 16 shows the overall comparison between the accuracy values of the dataset with and without the SMOTE algorithm of four cities. It can be seen that in the dataset without SMOTE algorithm the Kolkata city dataset gives the maximum accuracy for these three algorithms, whereas the Bangalore city dataset gives the minimum accuracy. The dataset with SMOTE algorithm is that the Hyderabad city dataset gives the maximum accuracy for these three algorithms, whereas the New Delhi city dataset gives the minimum accuracy. The dataset with the SMOTE algorithm clearly shows an increase in accuracy levels. It can also be seen clearly, how each city's accuracy has changed drastically.

The results from the imbalanced dataset show that random forest regression produces the lowest RMSE values in Bangalore (0.5674), Kolkata (0.1403), and Hyderabad (0.3826), as well as higher accuracy, compared to SVR and CatBoost regression for Kolkata (90.9700%) and Hyderabad (78.3672%), whereas CatBoost regression produces the lowest RMSE value in New Delhi (0.2792) and the highest accuracy for New Delhi (79.8622%) and Bangalore (68.6860%). In contrast to SVR and CatBoost regression, random forest regression yields the least RMSE values in Kolkata (0.0988) and Hyderabad (0.0628) and higher accuracies for Kolkata (93.7438%) and Hyderabad (97.6080%) for the balanced dataset, which is the dataset with the synthetic minority oversampling technique (SMOTE) algorithm applied to it. CatBoost regression yields higher accuracies for New Delhi (85.0847%) and Bangalore (90.3071%) and the least accurate results for Kolkata (0.0988%) and Hyderabad (0.0628%). RMSE values for Bangalore and New Delhi are 0.2148 and 0.1895. Therefore, it was evident from this that datasets that had the SMOTE algorithm applied to them produced higher accuracy.

It is observed that when SMOTE is applied, the accuracy for New Delhi with SVR goes from 78.4867% to 84.8332%, with RFR it goes from 79.4764% to 84.7284%, and with CatBoost regression, it goes from 79.8622% to 85.0847%. In the Bangalore dataset again, it is noticed that once the SMOTE algorithm is applied to the dataset, those datasets help achieve that accuracies are considerably higher when models are applied to them than those with imbalanced datasets (without SMOTE). When SMOTE is applied, the accuracy for Bangalore with SVR goes from 66.4564% to 87.1756%, with RFR goes from 67.7038% to 90.3071%, and with CatBoost regression goes from 68.6860% to 90.3343%. It is noticed that when SMOTE is applied, accuracy for Kolkata with SVR jumps from 89.1656% to 91.5624%, with RFR from 90.9700% to 93.7438%, and with CatBoost Regression from 89.9766% to 93.1656%. To establish the trend more, even Hyderabad shows increased accuracies from models when SMOTE is applied, like when it is used with SVR, the accuracy goes from 76.6786% to 93.5658%, with RFR, 93.5658% to 97.6080%, and with CatBoost Regression, 77.8991% to 96.7529%.

So, this gives quite a clear picture of the importance of balanced datasets. Having a dataset properly balanced can give more equal importance to each class. If there is too much of a gap between the number of values present for each class, it does not give an accurate portrayal of the actual scenario, and hence, the model fails. SMOTE creates multiple synthetic examples for the minority class and brings about a balance to the dataset. This makes the models work to the best of their ability, hence bringing better accuracy. This paper, hence makes clear about the importance of using SMOTE-applied datasets. Furthermore, these metrics also help show the best regression models for the particular use case and help in further research.

7. Conclusion and Future Work

Air pollution is a global problem; researchers from all around the world are working to discover a solution. To accurately forecast the AQI, machine learning techniques were investigated. The present study assessed the performance of the three best data mining models (SVR, RFR, and CR) for predicting the accurate AQI data in some of India's most populous and polluted cities. The synthetic minority oversampling technique (SMOTE) was used to equalize the class data to get better and consistent results. This unique approach of balancing the datasets, then using them, and then carefully comparing the results of both imbalanced and balanced ones for being highly accurate and then using statistical methods such as RMSE, MAE, MSE, and R-SQUARE to confirm the better results were very clearly successful in getting higher accuracy. The fresh research on balanced versus imbalance datasets used in such an application is well-tabulated and can be used as a reference for further research.

The algorithms were run using both datasets (with and without the SMOTE algorithm), and an increase of 6 to 24% was found. Our maximum accuracy in any city also went from 90.97% for Kolkata using RFR to 97.6% in the same city and algorithm. Our lowest accuracy went from 66.45% in Bangalore using SVR to 84.7% in Delhi for RFR. Overall, there was a major increase in accuracy. In the proposed work, using extensive testing of all three algorithms in New Delhi, Bangalore, Kolkata, and Hyderabad, it came to our notice that consistently, random forest regression and CatBoost regression provided promising results. In both cases, before using the SMOTE algorithm and after applying SMOTE, they outperformed SVR. The other metric comparison with and without the SMOTE algorithm is given below.

(i) Regarding R-SQUARE for unbalanced data,

  (a) In New Delhi—CatBoost gave the highest R-SQUARE.
  (b) In Bangalore, Kolkata, and Hyderabad: random forest got the highest R-SQUARE.

(a) In New Delhi: CatBoost gave the lowest MSE value.

dom forest got the lowest MSE value.

(iii) Regarding MAE for unbalanced data: in terms of accuracy which was calculated using MAE, it concluded as follows:

(a) In New Delhi and Bangalore: CatBoost gave the highest accuracy.
(b) In Kolkata and Hyderabad: random forest gave the highest accuracy.

(iv) Regarding RMSE for unbalanced data,

(a) In New Delhi: CatBoost got the least RMSE value albeit by a slight margin.
(b) In Bangalore, Kolkata, and Hyderabad: random forest got the least RMSE value.

(v) Regarding R-SQUARE for balanced data,

(a) In New Delhi and Bangalore, CatBoost gave the highest R-SQUARE.
(b) in Kolkata and Hyderabad, random forest gave the highest R-SQUARE.

(vi) Regarding MSE for balanced data,

(a) In New Delhi and Bangalore: CatBoost got the lowest RMSE value.
(b) in Kolkata and Hyderabad: random forest got the least RMSE value.

(vii) Regarding MAE for balanced data: in terms of accuracy which was calculated using MAE, it concluded as follows:

(a) In New Delhi and Bangalore: CatBoost gave the highest accuracy.
(b) In Kolkata and Hyderabad: random forest regression gave the highest accuracy.

(viii) Regarding RMSE for balanced data,

(a) In New Delhi and Bangalore: CatBoost got the lowest RMSE value.
(b) in Kolkata and Hyderabad: random forest got the least RMSE value.

So, it seems that in the use case of AQI in India, the CatBoost and random forest algorithms, coupled with SMOTE applied datasets, can provide great results to estimate air quality, which can prompt local and national governments, as well as other civic bodies to act and regulate the air quality. As very evident from the abovementioned metrics, the application of these regression models on the 2015 to 2020 AQI data has been successful in demonstrating that our innovation of using the SMOTE algorithm has paid off well and increased the accuracy values of these regression models. This innovative approach can be applied to future research and its benefits reaped.

For future work, there are plans to use satellite imagery and more extensive data to provide estimations for individual areas of a city as well. Another avenue to explore would be artificial intelligence (AI) to make the models more effective and innovative. This would help in figuring out which industrial areas contribute the most to pollution. Extending the study and trying new algorithms would also make our work more detailed. The aim is to find patterns and provide solutions on how to improve the air quality index of a city. The factors that contribute the most and ways to minimize them in an efficient way are an area worth exploring. In addition, further analyzing our dataset more to see if there are any intriguing patterns, such as the AQI's increase or reduction level during the holidays, or particular months and seasons, will be fruitful for our cause. [45].