ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 17, Issue, 11, pp.35191-35196, November, 2025
DOI: https://doi.org/10.24941/ijcr.49772.11.2025

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

A COMPARATIVE STUDY ON AIR QUALITY INDEX BY MACHINE LEARNING ALGORITHMS

[1,*]Damodharan, S., [2]Geetha K. and [3]Madhavi Latha, N.

[1]Department of statistics, Sri Venkateswara University, Tirupati; [2]Department of Mathematics, Sri Sankara Arts and Science College, Kanchipuram; Department of statistics, Sri Venkateswara University, Tirupati

ARTICLE INFO

Article History:
Received 14th August, 2025
Received in revised form
20th September, 2025
Accepted 17th October, 2025
Published online 29th November, 2025

Keywords:

AQI, Air Pollution, Health Impact, Regression,SVM, ANN, KNN, Decision Tree,Random Forest.

***Corresponding author:**
Damodharan, S.,

ABSTRACT

Our atmosphere is predominantly consists of two important gases that are important for livelihood, namely Oxygen and Nitrogen. Clean air is essential to stay healthy as well as to maintain a good environment. Air Quality is measured using a metric called Air Quality Index (AQI). It is a number that tells which pollutant is present in the air, its percentage and how it affects our health and Ecosystem. It keeps a tab on major Air Pollutants namely PM_{10},$PM_{2.5}$,No_2,So_2,CO,O_3,NH_3,Lead. Besides these, Emission of Firecrackers also made an impact of it by releasing harmful chemicals and organic pollutants. In the present study, we had used machine learning models like Linear Regression,Support Vector Machine,Artificial Neural Networks, Decision Trees and Random Forest. The results showed that ANN outperformed other models with the metrics R^2 (1),MAE (1), MAPE (0.7142),RMSE(1) for the city Gummidipoondi. This helps government to take preventive steps that worsens with all reasons related to weather conditions,vehicle emissions and local pollutants that create public health emergencies related to many respiratory and cardiovascular diseases.

Citation: Damodharan, S., Geetha K. and Madhavi Latha, N. 2025. *"*A Comparative Study on Air Quality Index by Machine Learning Algorithms.**".** ***International Journal of Current Research, 17, (11), 35191-35196.***

INTRODUCTION

Air pollution is the presence of substances in the atmosphere that are harmful to all the living organisms and environment. It might be chemical, physical or biological. There are many different types of air pollutants such as gases, particulates and biological molecules. It can happen naturally or by human activities. One such human activities affects due to bursting firecrackers during Diwali due to toxic substances. It consists of four parts:fuel,oxidizers,colorants and binders. By burning these fireworks, pollutes air by producing different toxic gases like sulphur dioxide,carbon dioxide, Carbon monoxide, particulate matter and traces of metals. As a result, it produces short term cardio vascular illness and chronic exposure diseases. When they are busted at an height, the pollutants they release are diluted before they reach people whereas ground-level fireworks have an immediate negative influence on our health.

REVIEW OF LITERATURE

Mohammed Sarmadi et al [1], published an article "An quality Index variation before and after the onset of Covid 19 Pandemic: a comprehensive study on 87 capital, industrial and polluted cities of the word". The authors investigated the changes in the air quality index in industrial, densely populated and capital cities in different countries of the world before and after 2020 of 121 days of 87 cities. AQI results showed that AQI improved except carbon monoxide and ozone in 2020. However, changes in 2021 have been received. Also, their finding says that few cities had decrease of PM2.5, PM_{10}, No_2, So_2,CO and O_3 in 2020 compared to 2019.Shivani Nigam et al[2], in their Research paper"Air Quality Index – A Comparative study of Accessing the status of Air Quality", estimated four pollutants $PM_{2.5}$, PM_{10}, No_2, So_2 of Nagpur from May to October 2014 which stated PM_{10} as the dominant pollutant causing harm to public. Six different methods were used to calculate ambient Air Quality Index based on Break Point Concentration for 24 hourly Averages. Anil Kender et al[3], published an article entitled "Forecasting of Daily air quality Index in Delhi", predicted daily AQI using ARIMA, PCR for Delhi city during 2000 to 2006 in all the four seasons. The percentage of very poor was found in summer and winter seasons which might be due to worst meteorological scenario. Ligia T.Silva et al[4]," City Noise-Air: An Environment quality Index for cities", generated maps of noise and pollutant concentration by applying Simulation models of Portuguese city. The urban quality Index aggregated data for the calculation of air and noise quality. Fabio Murena [5] in the Research article, "Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naples", developed Pollution Index from nine monitoring stations of Italy during 2001-2002. Reports had directed a procedure for evaluation of Pollution Index. It was reevaluated after considering additive effects of air pollutants. George Kyrkilis, et al[6],"Development of an aggregate Air Quality Index for

an urban Mediterranean agglomeration: Relation to potential health effects", developed an aggregate AQI of five pollutants namely PM10, No_2, So_2,CO and O_3 and evaluated for Athens, Greece with data of four monitoring stations during 1983-1999. Also, they revealed Athenian area reached high levels. Huixiang Liu et al[7]," Air Quality Index and Air pollutant concentration Prediction based on Machine learning Algorithms", predicted AQI for Beijing city and No_2 for Italy with two datasets(13-12-2002 to 18-08-2003) and (March 2004 to Feb 2005) by using Support Vector and Random Forest Regression models. They evaluated with the help of Root Mean Squarer Error, Coefficient of Correlation, Coefficient of Determination, of which Support Vector Regression was better in predicting AQI. Huan Li et al[8], "A Visualization approach to Air Pollution Data Exploration- A case study of Air Quality Index($PM_{2.5}$) in Beijing, China", proposed an efficient visualization method of China and found that PM_{10},$PM_{2.5}$ and No_2 may be emitted by strong winds may accelerate the spread of pollutants. The average concentration of $PM_{2.5}$ was higher than AQI value of 50 over six-year study period.

METHODS AND MATERIALS

For the present study, Secondary data of AQI was collected from the portal Central Pollution Control Board of period(25-10-24 to 25-11-24) for 20 cities in Tamil Nadu.

AQI is an indicator to report about how much amount air is clean or polluted. It is acquired by measuring emissions of eight major pollutants which are noted every hour. Each country has their own standards. In our Country, CPCB standard is followed for calculating air quality index. It is computed for 24-hourly average concentration value and health breakpoint concentration range.

AQI is measured in terms of six categories:

AQI	Category
0-50	Good
51-100	Satisfactory
101-200	Moderate
201-300	Poor
301-400	Very Poor
401-500	Severe

Even though measurement of AQIs is similar, the practical implementation of each can differ. Applying AQIs for a dataset can show large differences in the index values and concentration of pollutants. For the present study, we used Linear Regression, Ridge Regression, Lasso Regression, Support Vector Regression.

Regression: Regression is a statistical method used to analyze the relationship between a dependent variable and one or more independent variables. There are several types of Regression.

Linear Regression: Linear Regression is a statistical method that computes the linear relationship between the dependent variableand one or more independent features by fitting a linear equation to dataset. Its interpretation is a notable strength. The model's equation provides clear coefficients that give the impact of each independent variable over dependent one. It is not only apredictive tool, also lays the foundation of various advanced models. The linear relationship between dependent and independent variables is given by

$$Y = \beta_0 + \beta_1 x$$

Support Vector Regression : SVR is an extension of Support Vector Machines (SVM) that can be used to solve regression problems. It minimizes the prediction error that approximates the relationship between dependent and independent variables. It is used for classification, regression and outliers' detection as they are effective in high dimensional spaces.

Artificial Neural Networks: An Artificial Neural Network (ANN) is a machine learning algorithm that works similar to the biological neuron'sbehavior. It is the most powerful machine learning algorithms used to make predictions most of the complicated datasets.It has three layers namely input, hidden and output layers which forms an architecture. The neurons in the hidden layer are activated by a function such as sigmoid or hyperbolic tangent. It learns the procedure by testing and training the dataset for producing output. There are several types of ANN namely Recurrent neural networks,Convolutional neural network, Feedforward Neural Network, Radial Basis,Perception Neural Networks and so on.

k-Nearest Neighbor: The k- nearest neighbor (KNN) is a machine learning algorithm that classifies new data points which resembles similarity of the dataset. It is widely used in Classification and Regression. It is easy to understand but however its performance is affected by the choice of K and distance metric. If the input data has more outliers or noise, high value of k would be considered.

Decision Trees: Decision Trees are machine learningalgorithm which is simple predictive models which classifies the data into sub groups with input variables for getting possible outputs. It is a supervised learning algorithm to draw conclusions. More generally Classification and Regression is used. When outcome is in discrete, Classification tree analysis is used while if it is a real number Regression Tree analysis is used. The different decision tree algorithms are Id3, C4.5, CART, CHAID, and MARS. Pruning and Regularization can be mitigated by this technique.It follows Metrics such as Gini impurity, information gain or Mean Square Error(MSE) for split evaluation.

Random Forest: Random Forest algorithm is a machine learning algorithm that uses multiple decision trees for prediction, which is made up of multiple decision trees. This algorithm is an extension of the bagging method ensuring low correlation. This is a key difference between decision trees and random forests. While decision trees involve all the possible feature splits, while Random Forest selects a subset of those features.

Evaluation Metrics

Mean Absolute Percentage Error: Mean Absolute Percentage Error(MAPE) is used as a Metric for predicting accuracy of a Forecasting Method. It is also known as Mean Absolute Percentage Deviation. It is commonly used as a loss function for Regression problems.

$$MAPE = \sum \frac{\frac{|A-F|}{A}}{n} * 100$$

Where A= actual value, F =Forecasted value and N = number of fitted points

Mean Absolute Error: Mean Absolute Error is an evaluation metric used to calculate the accuracy of a regression model. It measures the average absolute difference between actual and predicted values.

$$MAE = \frac{1}{n}\sum | y - \hat{y} |$$

Where n is the number of observations, y represents the actual values $\hat{y}$ represents the predicted values

Lower MAE value indicates better model performance stating that predictions are closer to the actual values.

Root Mean Square Error: RMSE is measured by applying square root of the average of the squared difference between the prediction and actual value. Simply, it is the standard deviation of the residuals(prediction error).

$$RMSE = \sqrt{\sum(\frac{(y-\hat{y})^2}{n})}$$

Coefficient of Determination(R^2): It is the coefficient of determination is the proportion of the variance in the dependent variable that is predicted from the independent variable which indicates amount of variation in the data

.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where RSS is Residuals of sum of squares, TSS is Total sum of squares A value near 1 says that model is good fit.

RESULTS AND DISCUSSION

In this work, we had used several models like Linear Regression, Support Vector Regression, Artificial Neural Networks, k-Nearest Neighbor, Decision Trees, and Random Forest to analyze the AQI for 20 cities in Tamilnadu.

Performance Metrics like Coefficient of Determination, MAE, MAPE, and RMSE was used to evaluate performance.

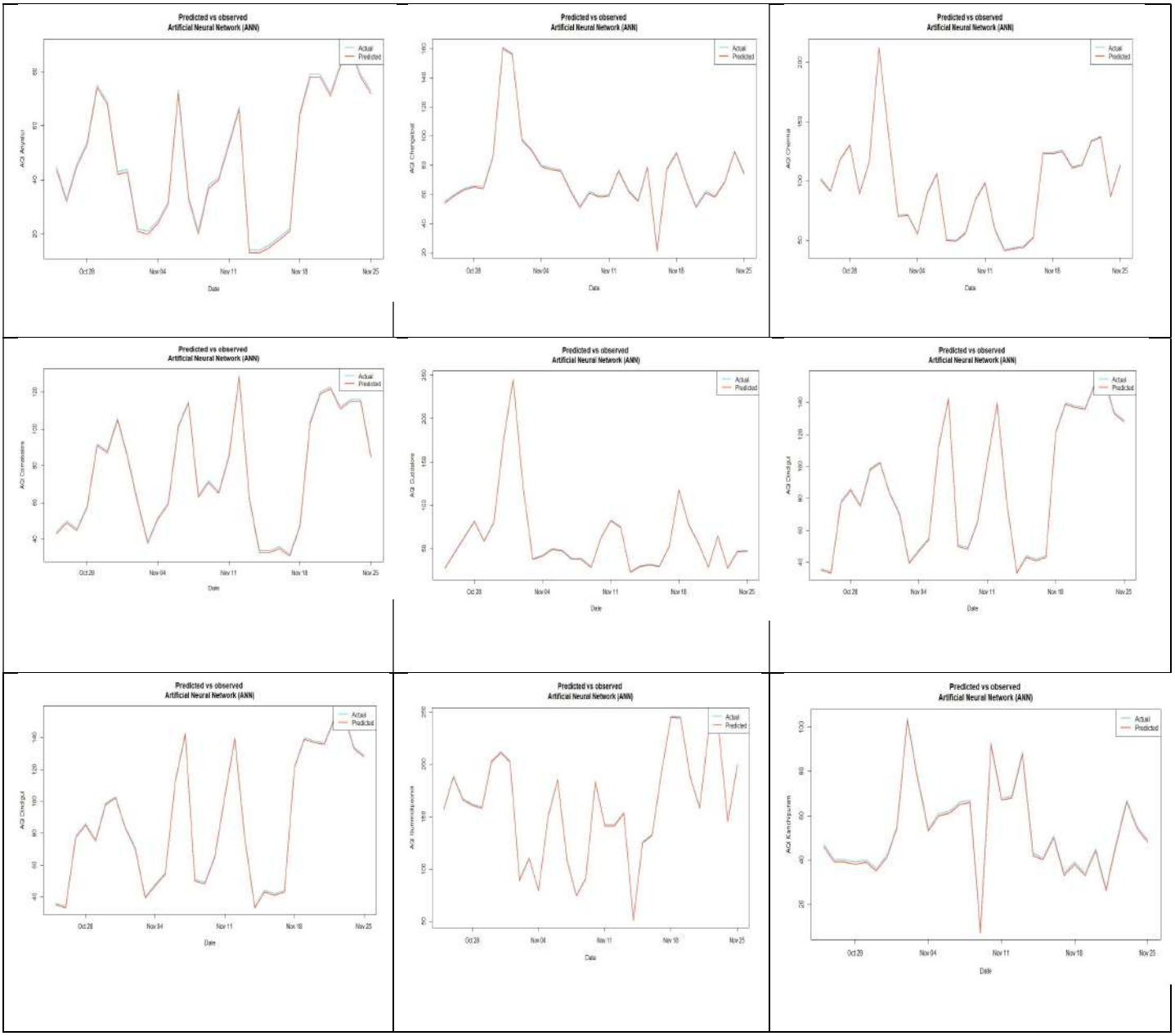
Place	Linear Regression				SupportVector Machine				ArtificialNeural Networks			
	R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE
Ariyalurr	0.1198	19.3056	64.3062	22.191	0.4013	13.862	50.4047	18.4189	1	1	2.8546	1
Chengalpattu	0.0559	17.6018	26.8279	25.5395	0.4493	17.6018	26.8279	25.5395	1	1	1.5127	1
Chennai	0…16	30.4858	39.7898	37.405	0.4179	22.0385	26.3487	28.8046	1	1	1.2322	1
Coimbatore	0.1153	24.6197	41.5543	28.634	0.2711	20.5962	34.8398	26.0383	1	1	1.5747	1
Cuddalore	0.0719	30.1744	55.2605	44.3077	0.2929	24.5717	33.0417	42.5563	1	1	2.1214	1
Dindugal	0.2719	30.0604	47.1738	34.2858	03447	23.5206	39.0004	32.7776	1	1	1.4443	1
Gummidipoondi	0.048	41.7698	34.2266	49.0402	0.3472	33.2538	27.0393	40.9231	1	1	0.7142	1
Kanchipuram	0.0017	15.4544	44.3225	19.6591	0.2104	10.5676	38.0419	17.5411	1	1	2.3654	1
Ooty	0.0338	20.8951	40.4602	23.0332	0.406	14.1556	26.2172	18.0833	1	1	1.8028	1
Puducherry	7.00E-04	23.7632	38.0411	33.2306	0.2831	19.3033	26.9665	29.0862	1	1	1.5794	1
Pudukkottai	0.3005	19.1145	32.1174	23.6014	0.3112	18.3586	31.4286	23.5047	1	1	1.574	1
Ramanathapuram	0.1719	13.358	40.3917	18.3174	0.3039	11.8307	36.0393	16.8014	1	1	2.7162	1
Ranipet	0.1187	29.9684	55.3354	34.8891	0.6263	17.0478	29.8867	23.1021	1	1	1.7148	1
Salem	0.0022	33.6434	47.4878	36.4714	0.2764	25.4238	36.3523	31.1647	1	1	1.3311	1
Thanjavur	0.0478	10.358	19.7385	18.4546	0.347	6.3529	9.7447	16.8418	1	1	2.2852	1
Thoothukodi	0.0203	17.9686	45.415	22.9177	0.0852	16.7531	40.4632	22.2403	1	1	2.321	1
Tiruchirapalli	0.1404	17.4808	49.7005	21.2054	0.3198	14.4198	36.8227	19.1081	1	1	2.473	1
Tirunelveli	0.2021	7.2481	22.1518	11.5543	0.1895	6.3871	17.5857	11.9636	1	1	3.2566	1
Tirupur	0.2117	19.4273	22.4994	24.1917	0.3044	15.0972	18.0499	23.1422	1	1	1.144	1
Vellore	0.0027	29.3374	49.617	32.9394	0.4638	18.6742	33.2829	24.3694	1	1	1.4251	1

Place	k-Nearest Neighbor				Decision Tree				Random Forest			
	R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE
Ariyalurr	0.8181	7.944	22.6181	10.4121	0.6169	11.3359	33.7738	14.6391	0.9052	5.5871	15.0636	7.7389
Chengalpattu	0.6822	10.2415	16.1218	15.0127	0.1777	16.3823	24.5833	23.8341	0.8865	7.3372	11.0601	10.3666
Chennai	0.6875	16.2145	18.0249	21.1919	0.5789	18.0802	20.9902	24.292	0.8412	11.7486	12.8957	15.5052
Coimbatore	0.8288	10.388	15.3578	13.4578	0.4186	18.2232	28.6814	23.2123	0.8993	7.7367	11.2281	10.5303
Cuddalore	0.7543	16.7005	28.8795	23.916	0.2645	26.808	46.6983	39.4416	0.8541	12.1834	21.441	19.1117
Dindugal	0.7953	14.8325	22.7096	18.9847	0.5325	20.7172	32.4605	27.4719	0.9019	10.2937	14.4101	13.524
Gumidipoondi	0.6601	24.4838	19.5172	30.0774	0.5481	28.9924	22.5538	33.7872	0.8226	18.2722	14.4841	22.3026
Kanchipuram	0.4876	9.1752	32.3868	14.2327	0.3688	9.5896	37.003	15.6328	0.0338	6.1833	19.111	9.341
Ooty	0.7714	9.4402	17.7424	11.5305	0.4181	14.262	27.7585	17.8756	0.8795	6.978	12.868	8.5629

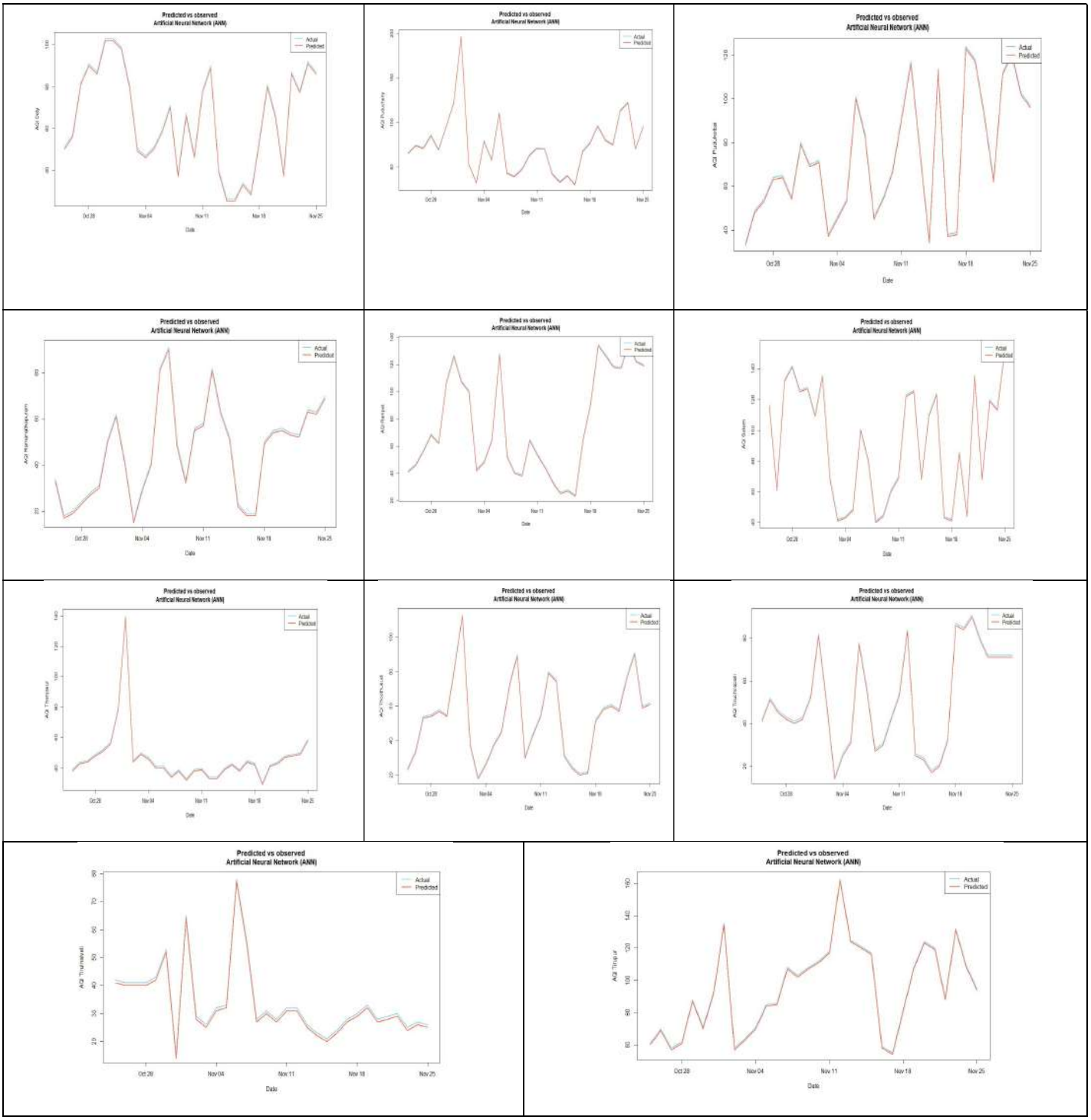
Puducherry	0.5762	15.2022	22.1998	21.9718	0.3528	20.0039	28.2463	26.7426	0.8352	11.4607	17.0702	15.4705
Pudukkottai	0.5115	15.2245	24.8074	20.031	0.4221	17.4219	28.6287	21.4528	0.7944	10.5738	16.8482	14.0591
Ramnathpuram	0.7833	7.9585	21.8306	10.0124	0.3656	13.1094	38.4285	16.0324	0.8982	5.3877	15.5803	7.4035
Ranipet	0.8577	10.4165	16.5945	14.44	0.613	18.6849	31.94	23.1196	0.9149	8.1594	12.7211	11.4712
Salem	0.6151	19.6863	26.4213	23.264	0.3835	23.6228	33.4836	28.6664	0.7794	15.4385	20.3485	18.2373
Thanjavur	0.5781	5.6422	10.1696	12.3689	0.2473	8.9705	16.6484	16.408	0.8598	4.2261	8.0428	8.1126
Thoothukodi	0.6563	11.4852	25.3894	14.3749	0.1589	16.675	39.5268	21.2347	0.8663	7.9047	18.0406	10.2777
Tiruchirapalli	0.7602	8.1902	20.7172	11.8066	0.5239	11.8229	30.2382	15.7811	0.8789	5.9957	15.1426	8.6574
Tirunelveli	0.4776	5.4242	17.1349	9.401	0.3164	6.5833	20.7229	10.6946	0.7856	3.5606	11.4215	6.3183
Tirupur	0.6981	11.006	12.5592	15.2774	0.3065	16.5845	19.6166	22.6909	0.8327	8.6456	9.7773	11.7883
Vellore	0.7668	12.374	18.2409	16.2547	0.6267	14.3962	24.1866	20.1526	0.8806	9.0793	13.8221	12.0018

The criteria for all mentioned metrics should be High R^2 and low MAE,RMSE,MAPE. According to the table, maximum R^2 value was 1 for all the cities for ANN model,next to that Random Forest shows 0.9019 for Dindugal city. There is an exception of Puducherry with beyond the limits of R^2. MAE and RMSE has least value 1 for all the cities with ANN model, next to that Random Forest for Tirunelveli(3.5606) for MAE and 6.3183 for RMSE. It shows that after ANN, Random Forest exhibits high performance. The lowest value of MAPE is 0.7142(Gummidipoondi), 1.144(Tirupur), 1.311(Salem), 1.4251(Vellore) with ANN.

The experimental results showed that best model was Artificial Neural Network satisfying the criteria of the metrics when compared to other models. It demonstrates that ANN is a good substitute for analyzing model. The following chart shows the actual and predicted values for all the Cities which had low MAPE followed by ANN meeting with other criteria metrics:



Continue ……



In this work, we had used several models like Linear Regression, Support Vector Regression, Artificial Neural Networks, k-Nearest Neighbor, Decision Trees, and Random Forest to analyze the AQI for 20 cities in Tamil Nadu. Performance Metrics like Coefficient of Determination, MAE,MAPE, and RMSE was used to evaluate performance. The criteria for all mentioned metrics should be High R^2 and low MAE,RMSE,MAPE. The experimental results showed that best model was Artificial Neural Network satisfying the criteria of the metrics when compared to other models. It demonstrates that ANN is a good substitute for analyzing model.

CONCLUSIONS

Accurate air quality forecasting plays a vital role for humanity. The study helps Government to take preventive measure in the emergency scenario. In this study, we built Regression models and also machine learning models to predict air indicators. The results show that ANN can fetch good results when compared to Linear Regression, SVM,KNN, Decision Tree, Random Forest. The metrics showed that Coefficient of Determination has high value for all the cities in all the models. Among 20 cities,

Gummidipoondi had low metrics with ANN model. The existing model is not sufficient to tackle health related issues. Mostly software and hardware approaches are used to identify and analyze values. The present study visualizes making alerts on quality of the air. The potential of algorithms used in the present study may be valuable to develop real time AQI prediction system uplifts other models. In this concern, policymakers need up-to-date and accurate information to implement effective measures for improvising AQI.

REFERENCES

Mohammed Sarmadi, Sajjad Rahimi, Mina Rezaei, Darryoush Sanaei, Mostafa Dianatinasab, A quality Index variation before and after the onset of Covid 19 Pandemic: a comprehensive study on 87 capital, industrial and polluted cities of the word". Environmental Sciences, 33:134(2021) 1-17.

Shivani Nigam, B.P.S. Rao, N. Kumar, V.A. Mhaisalkar, "Air Quality Index – A Comparative study of Accessing the status of Air Quality", Research J. Engineering and Tech, 6(2), 2015,1-8.

Anil Kender, P. Gayal, "Forecasting of Daily air quality Index in Delhi",Science of Total Environment, 409(24), 2011,5517-5523.

Ligia T.Silva, Jose F.G. Mendes, City Noise-Air: An Environment quality Index for cities", Elsevier, 4(1), 2012, 1-11.

Fabio Muren, "Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naple", Atmospheric Environment, 38(36), 2004,6195-6202.

George Kyrkilis, Arrhontoula Chaloulakou,Pavious A. Kassomenos," Development of an aggregate Air Quality Index for an urban Mediterranean agglomeration: Relation to potential health effects", Environmental International, 33(5), 2007,670-676.

Huixiang Liu, Qing Li, Dongbing Yu and Yu Gu," Air Quality Index and Air pollutant concentration Prediction based on Machine learning Algorithms", Applied Sciences, 9(19),2019,1-9.

Huan Li, Hong Fan, Feiyue Mao, "A Visualization approach to Air Pollution Data Exploration- A case study of Air Quality Index($PM_{2.5}$) in Beijing, China", Atmosphere, 7(3), 2016, 1-35.

Biswanath Bishai, Amit Prakash, V.K. Jain, "A comparative study of Air Quality Index based on Factor Analysis and US-EPA methods for an urban environment", Aerosol and Air Quality Research, 9(1),2009,1-17.

Antonella Plaia, Mariantonietta Ruggieri, "Air Quality Indices: a review », Reviews in Environmental Science and Biotechnology, 10(2), 2011, 165-179.

Dongsheng Zhan, Mei-Po Kwan, Wenzhong Zhang, Xiaoften Yu, BinMeng,Qianqian Liu, The driving factors of Air Quality Index in China", Journal of cleaner production, 197, 2018, 1342-1351.

Mohammed Hossein Sowlat, Hamed Gharilal, Masud Yunesian, Maryam Tayefeh Mahmoudi Saeedeh Lotfi, "A novel, fuzzy-based air quality index(FAQI) for air quality assessment", Atmospheric Environment, 45(12), 2011, 2050-2059.

Shwetha Kumari, Manish Kumar Jain, "A critical review on Air Quality Index", Environmental Pollution: select proceedings of ICWEES-2016,87-102,2017.

Manju Mohan, Anurag Kandya, "An analysis of the annual and seasonal trends of air quality index of Delhi", Environmental Monitoring and assessment, 131(1), 2007, 267-277.

Kanchan, Amit Kumar Gorai, Pramila Goyal,"A review on air quality indexing system", Asin Journal of Atmospheric Environment, 9(2), 2015, 101-113.

Prashant Kumar, "A critical evaluation of air quality index models(1960-2021)", Environmental Monitoring and assessment, 194(5), 2022, 1-45.

Prabhat K Swamee, Aditya Tyagi, Formation of an Air Pollutant Index", Journal of the Air & Waste Management Association, 49(1), 1999,88-91.

Rohit Sharma, Raghavendra Kumar, Devendra Kumar Sharma, Le Hoang son, Ishaani Priyadarshini, Binh Thai Pham, Dieu Tien Bui, Sakshi Rai, "Inferring air pollution from air quality index by different Geographical areas: Case study in India, Air Quality, atmosphere& Health, 12(11), 2019, 1347-1357.

Zbigniew Bagienski, "Traffic Air Quality Index", Science of total environment, 505, 2015, 606-614.

Lyndon R Babcock Jr, "A combined pollution Index for measurement of total air pollution", Journal of the Air Pollution control association, 20(10), 1970, 653-659.

****