International Journal of Recent Development in Computer Technology & Software Applications
http://eurekajournals.com/IJRDCTSA.html
ISSN: 2581-6276

Proactive Prediction of Air Quality Index using Machine Learning Techniques to Detect Lung Cancer

Drashti Shah, Sharmistha Mondal

Abstract

Air pollution play a critical position when it comes to the effect, it has on the surroundings in flip affecting public health, socio-economics, politics and agriculture. In the usage of a system mastering algorithms, we record the pollutant concentrations using air high-quality index (API) in India over the term of (January 1970 – January 2015). Delhi is one of the most polluted towns globally, particularly due to vehicle pollution. Factors that implement the machine learning algorithm at the input are meteorological parameters, pollutant concentrations and timestamp. This paper proposes a machine learning technique to predict lung cancer due to air pollution with determination of preventing health concerns, like respiratory infections, asthma, pneumonia, cardiovascular problems and lung cancers. The proposed techniques can be used by the health department of urban to check air quality, physicians to estimate spatial–temporal profile of air pollution and air excellent indices. Further research is to study the efficiency and potency ofML with geometric, computational and statistical models.

Keywords: AQI, Smog, Analysis, COPD, NSCLC.

Introduction

The difficulty Air pollutants and its prevention have constantly challenged scientists over the last decadeis particularly because of the most important health outcomes they have induced. It continues to remain a big worldwide hassle due to the effect they have at the human respiration and cardiovascular structures which leads them to being the motive of an increase in mortality fee and associated illnesses all over the globe [2].

The state authorities have been making efforts, so that it will understand and are expecting the AQI [Air Quality Index][3] values aimed at improving public health and with the largest and most drastic changes inside the area of studies and development with recognize to artificial intelligence and machine learning over the last few years, this record pursuits to bring the electricity of artificial intelligence into light [7]. The technology of AI is in which the gadget makes the choice on its very own as an alternative on conventionally taking the orders from a programmer within the shape of an application which has step by step

commenced influencing all aspects of our existence. Starting from early-stage, startup corporations ending at large platform companies, for all of them, artificial intelligence and its element system learning have end up the principal attention place [8].

Machine learning is an area where the system which implements artificial intelligence gathers statistics from sensors in an environment and learns a way to act. One of the reasons why we select machine getting to know to are expecting air pleasant index, become this potential of adapting of gadget mastering (ML) algorithms [9]. Every day measured values of the parameters of air high-quality are in many instances above the limit values which might be taken into consideration safe for human fitness. In the larger urban areas the state ofaffairs is urgent [5]. Some of the sports to lower the air pollution are undertaken through the neighborhood authorities, a few with the aid of the state government. This project represents our effort at the medical level to make contributions in handling this trouble by means of analyzing and predicting the AQI (Air Quality Index) of the destiny to result in cognizance to the community with regards to how the degradingair best additionally causes lung most cancers [6].

Lung cancer is a form of cancer that starts off evolved in lungs, the NSCLC- Non-Small Cellular lung cancer accounts roughly 80 to 85 percent of lung in most cases. Other styles of lung cancer encompass small cellular lung most cancers and mesothelioma [2]. While smoking absolutely stays the major cause in the back of this deadly sickness, not only smoking will cause this sickness. Even if people inhale poisonous substances in air, passive smoker as well expose to radioactive gases can increase the chance of lung cancer.

Living and breathing in a polluted metropolis is also a major cause for different health consequences, the health department should handle and address this awareness to the public [2].

Check pollution degrees earlier, before stepping out of the house

- Avoid areas with a high population density
- Avoid working outdoors especially when there are high levels of pollution.
- Avoid burning wood or trash

The air quality index provides a way to determine the air pollutant in the living environment and maintain the better health and fitness concerns of individual in the society as well it creates awareness in the society to be away from breathing contaminated air.

Table1: Air Quality Index

AQI	Zone
301-500	Hazardous
201-300	Very Unhealthy
151-200	Unhealthy
101-150	Unhealthy for Sensitive Groups
51-100	Moderate
0-50	Good

The table1 shows the range of values for air quality index in ppm that can be measured in the enrolment, if it ranges from 0 to 50 ppm; it is good for human health and if it is 51ppm to 100 ppm then its moderate for the human beings. If it is above the range as shown in table, then it is very harm for the human; If the range is among 51 ppm to 100 ppm, then it leads to respiratory issues, needs to take some measures.

Literature Review

According to medical doctors of the Ganga Ram Hospital, it may be the poisonous air and high-degree of air pollution inside the metropolis which brought on the degree four most cancers within the female. As per the inferences from the doctors of Ganga Ram Hospital, Ghazipur metropolis predicted that the air quality index in the city is above the threshold level, which intern may affect the people with breathing troubles and advised the people to reduce the exposure to outside [2].

It has been observed the theses kind of air pollution even affect the young around the age of 20 to 25.Dr Arvind Kumar, a chest surgeon said, “I suspect the purpose at the back of its miles polluted and poisonous air in Delhi. Polluted air also contains factors observed in cigarettes. It isn't always an isolated case”. He similarly brought, “I even have reported such cases earlier too. On average, I even have seen 2-three lung cancer instances each month in non-smoking individuals at the beginning of their 30's. But thatis the primary case within the 20's".

Proposed Method

In ML, the point of interest is on gaining knowledge from facts. This is possibly better illustrated the useof a easy analogy. As kids we typically study what's “proper” or “exact” behavior by means of not to do or punished for doing that.

The basic idea in machine learning model is that; we build the model by training followed by validation.If the validation results are up to the threshold the model goes for testing the real data. The errors will be reduced during the validation process.

In ML, price functions are used to estimate how badly functions are performing. Put genuinely, a cost feature is a measure of ways wrong the version is in terms of its capability to estimate the connection between X and y. This is commonly expressed as a difference or among the expected value and the real value. Cost function is calculated iteratively by running the model and compared the estimated value.

The Naive Forecasting Technique that is estimating method in which the closing length's actual is used as the era's forecast, without altering them or attempting to set up causal elements. It’s used for evaluation with the forecasts generated by using the higher (sophisticated) techniques. Now the naive forecast may bevery noisy as it does not filter any noise by any means. That makes it very nervous but additionally responsive to changes in call for. Nervous way we've got a totally risky forecast. However, the naive method may be used as a benchmark to greater complicated methods and it tells you whether or not a greater

complicated technique is virtually advanced.

For naïve forecasts, all forecasts to be the value of the closing remark. If the old data be represented by y1,…,yT, y1,…,yT, then we can write the forecasts as.

$$\hat{y}_{T+h|T} = \bar{y} = (y_1 + \cdots + y_T)/T.$$

The notation $\hat{y}_{T+h|T}$ is a short- estimate ofbased on the data y1,…,yT.

This approach works better for many financial time series data. naive(y, h) rwf(y, h) # Equivalent alternative

The naïve forecast is finest, if the data follow a random walk, these are also called random walkforecasts.

Results and Discussion

	sampling_date	state	si	ni	rpi	spi	AQI
0	February - M021990	Andhra Pradesh	6.000	21.750	0.0	0.0	21.750
1	February - M021990	Andhra Pradesh	3.875	8.750	0.0	0.0	8.750
2	February - M021990	Andhra Pradesh	7.750	35.625	0.0	0.0	35.625
3	March - M031990	Andhra Pradesh	7.875	18.375	0.0	0.0	18.375
4	March - M031990	Andhra Pradesh	5.875	9.375	0.0	0.0	9.375

Table1: Shows the air quality index values

The table1 shows the air quality index values in Andhra Pradesh during February and March for 1990.

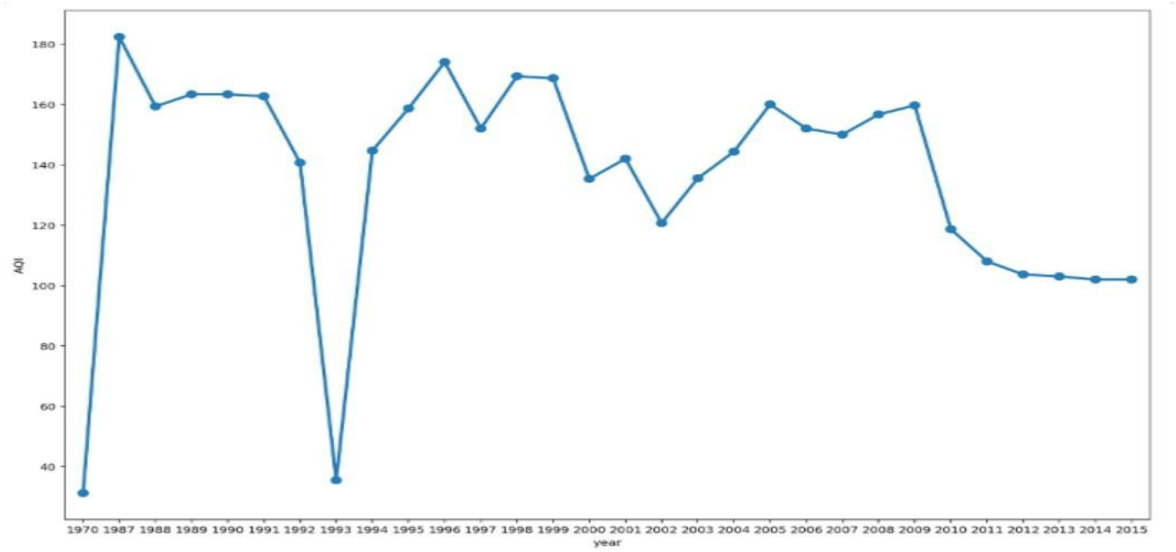

Figure 1: AQI values from 1970 to 2015

The figure1 shows the air quality index values from the year 1970 to 2015.

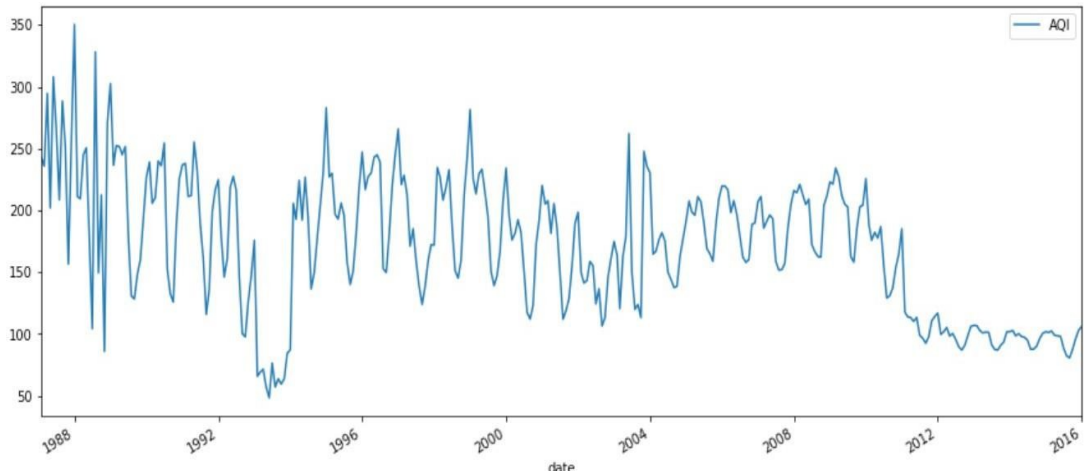

Figure 2: Shows the training data.

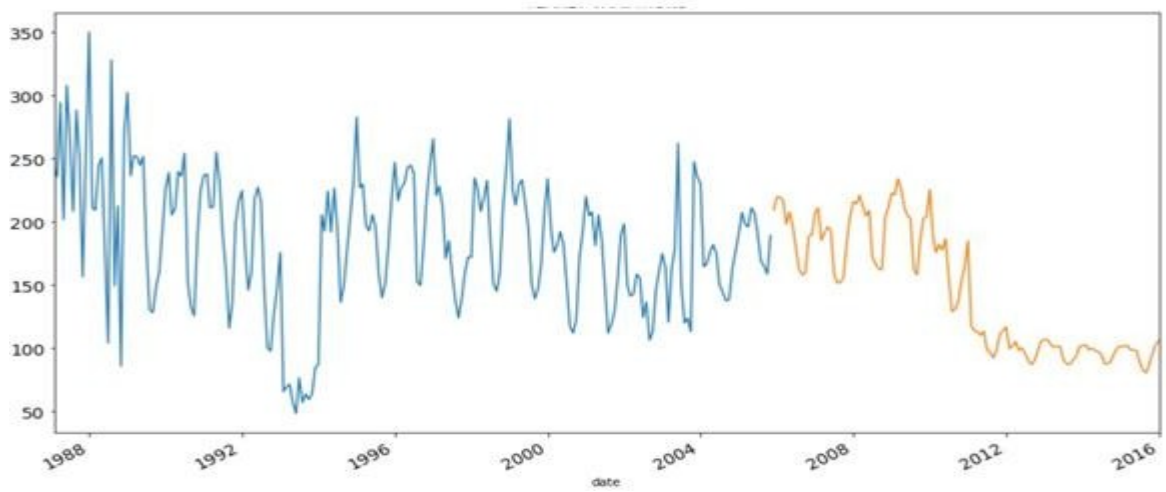

Figure 3: Shows the training and testing values of AQI.

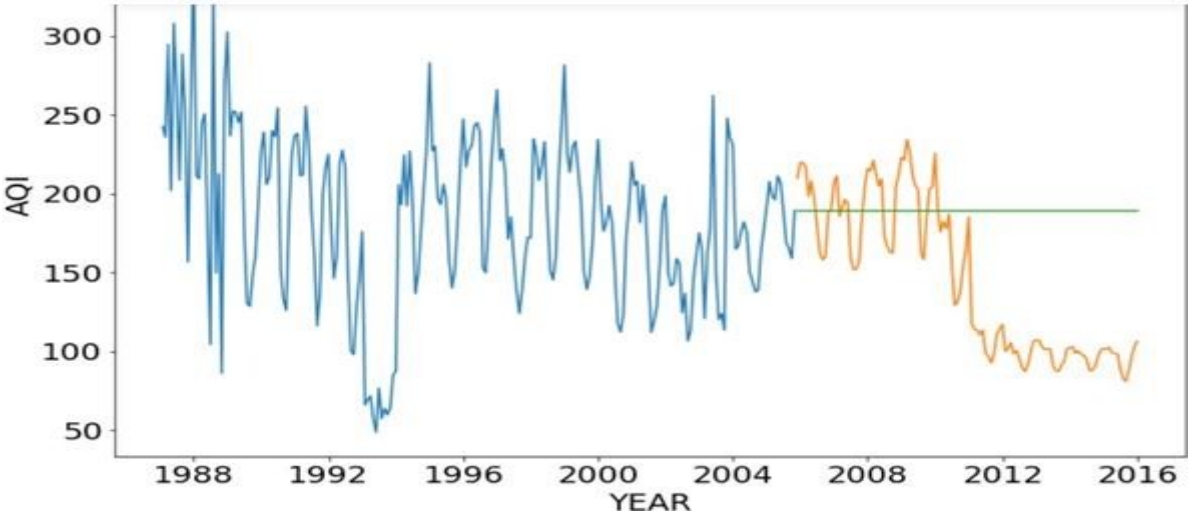

Figure 4: Using Naïve Forecasting analysis to train the data using mean value

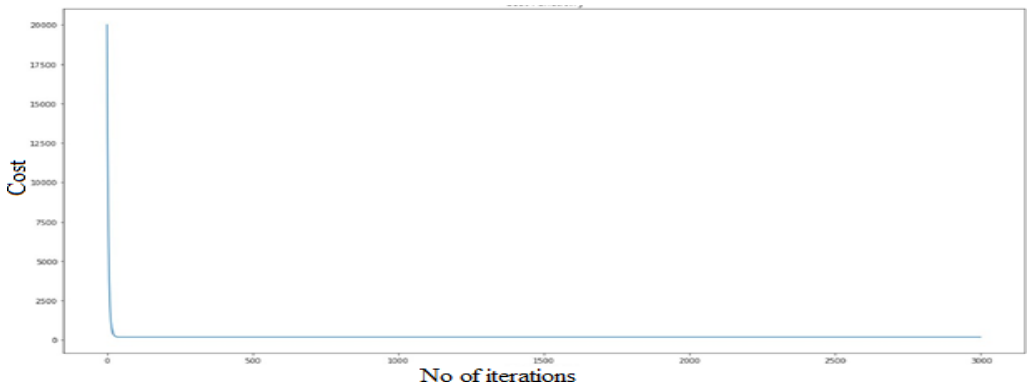

Figure 5: Cost Function graph to measure error in optimisation values due to gradient decent

The optimization function used in the proposed approach is gradient descent to identify the value of coefficients. These parameters minimize a value feature (price) of the function. Gradient descent is exceptional used, if the parameters cannot be calculated systematically and have to be searched as per the optimization rules.

The procedure starts with initial values for the coefficient of the characteristic. These can be 0.0 ora small random cost.

Coefficient = 0.0

The price of the coefficients is evaluated by way of plugging them into the function and calculatingthe value.

Fee = f(coefficient);Or

Price = compare(f(coefficient));

The value of cost is calculated. The spinoff is a notion from calculus and refers back to the slope of the characteristic at given factor. We want to recognise the slope in order to recognise the route (sign) to move the coefficient values with a view to get a lower fee on the subsequent generation.

Delta = derivative (cost)

Now that we understand from the by-product which r

oute is downhill, we will now replace the coefficient values. A studying price parameter (alpha) must be distinct that controls how much the coefficients can exchange on every update.

Coefficient = coefficient – (alpha * delta)

This procedure is repeated until the price of the coefficients (price) is 0.0 close sufficient to zero to be desirable sufficient. Batch gradient descent is where we calculate the derivative after taking all the training facts as input right before calculating an update.

Table2: Shows the actual and predicted values 1998 to 2010

	year	AQI	Actual	Predicted
23	2010	188.283360	188.283360	198.008667
21	2008	214.378174	214.378174	198.477794
20	2007	211.160807	211.160807	198.712357
19	2006	219.623267	219.623267	198.946920
18	2005	207.546049	207.546049	199.181484
17	2004	164.661496	164.661496	199.416047
16	2003	163.875510	163.875510	199.650610
14	2001	205.138247	205.138247	200.119736
13	2000	195.772377	195.772377	200.354300
10	1997	220.903571	220.903571	201.057989
9	1996	216.850189	216.850189	201.292553
7	1994	205.636343	205.636343	201.761679
5	1992	177.485106	177.485106	202.230806
1	1988	211.076502	211.076502	203.169058

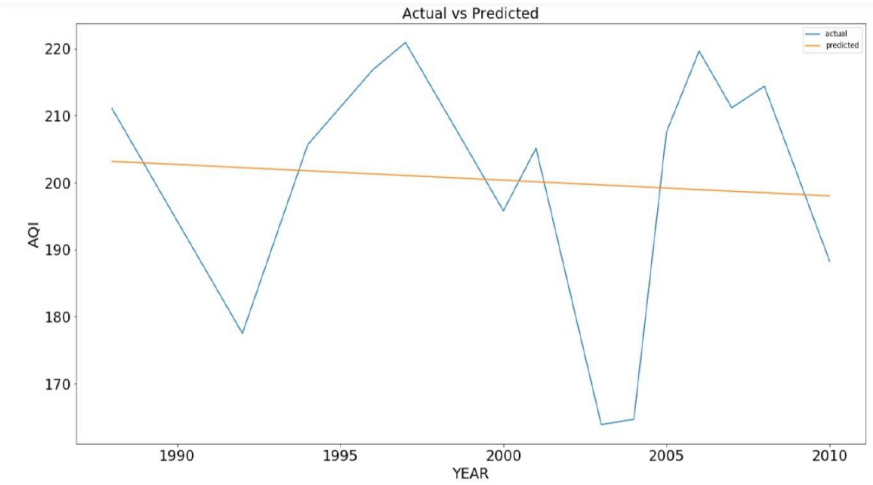

Figure 6: First cycle of predicted values after training, testing and optimising

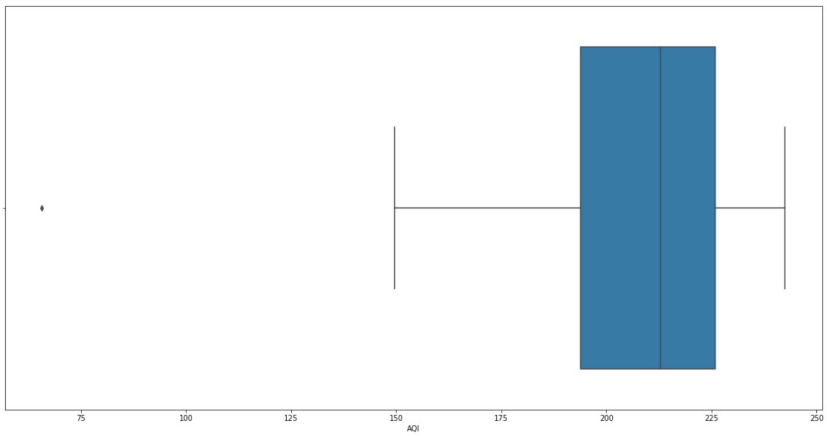

Figure 7: Box plot to identify and remove outlier

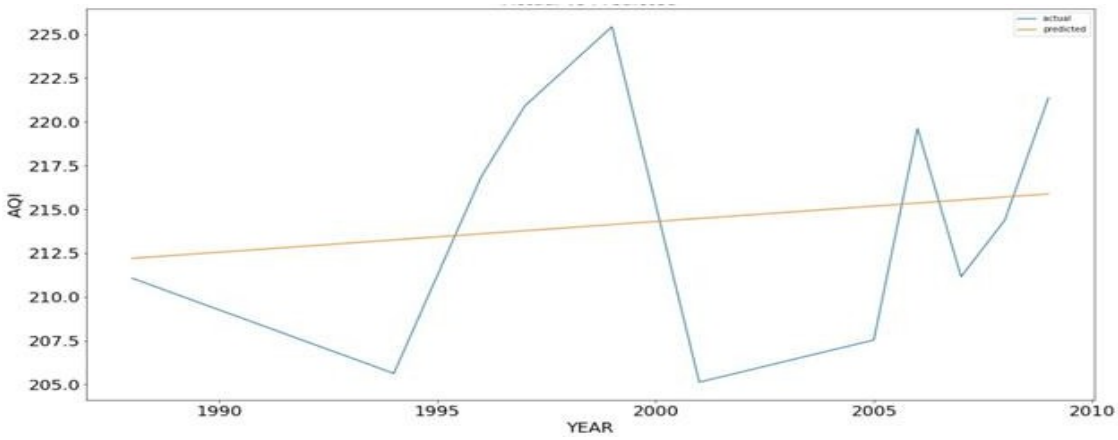

Figure 8: Second cycle of predicted values after training, testing and optimising

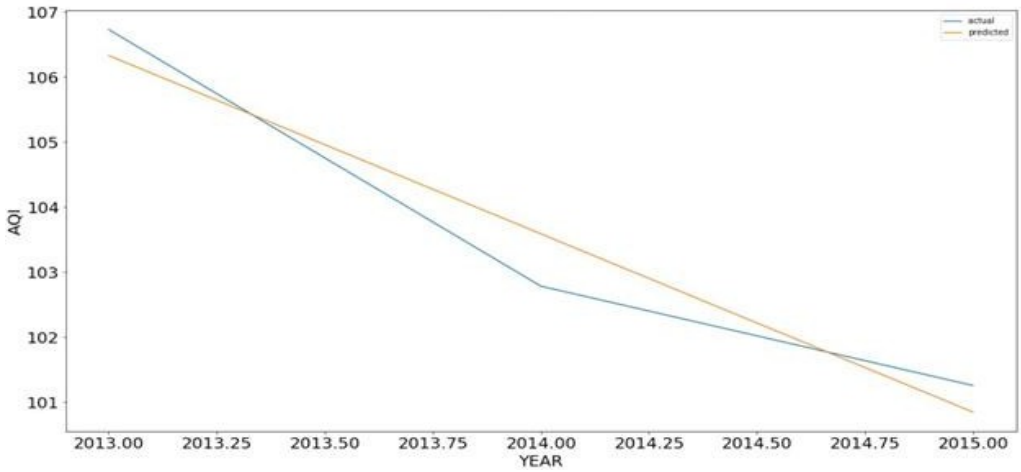

Figure 9: Third cycle of predicted values after training, testing and optimising

By using Machine Learning to optimize the AQI values by training the dataset, we further predicted the future values using naïve analysis. A rough estimate of the AQI value in the year 2020would be 106.33.

Conclusions

In the proposed method the predicted AQI value of 106.33 falls in the range 101 to 150 in the AQI table which stands for how it would affect sensitive groups in the community article and brings to light how our atmosphere greatly impacts our health. With all the ongoing research and constant development in the field of machine learning and artificial intelligence, soon AI and ML will be able to make more accurate predictions and analysis to aid mankind.

References

1. https://towardsdatascience.com/machine-learning-fundamentals-via-linear-regression-41a5d11f5220
2. https://app.cpcbccr.com/ccr_docs/How_AQI_Calculated.pdf
3. Agirre-Basurko, E., Ibarra-Berastegi, G. and Madariaga, I. (2006). Regression and Multilayer Perceptron-based Models to Forecast Hourly O3 and NO2 Levels in the

Bilbao Area. Environ. Modell. Softw. 21: 430–446.

4. Baawain, M.S. and Al-Serihi, A.S. (2014). Systematic Approach for the Prediction of Ground-Level Air Pollution (around an Industrial Port) Using an Artificial Neural Network. Aerosol Air Qual. Res. 14: 124–134.
5. Bhaskar, B.V. and Mehta, V.M. (2010). Atmospheric Particulate Pollutants and Their Rekha G., Bhanu Sravanthi D., Ramasubbareddy S., Govinda K, "Prediction of stock market using neural network strategies", Journal of [6]
6. Computational and Theoretical Nanoscience, Vol.16, No.9,pp.2333 -2336,(2019)..
7. Cairncross, E.K., John, J. and Zunckel, M. (2007). A Novel Air Pollution Index based on the Relative Risk of Daily Mortality Associated with Short- Term Exposure to Common Air Pollutants. Atmos. Environ. 41: 8442–8454.
8. Deelip M.S., Govinda K., Ramasubbareddy S., Swetha E., Aditya Sai Srinivas T, "Analysis of twitter data for prediction of iPhone X reviews", Journal of Computational and Theoretical Nanoscience, Vol.16, No.9,pp.2050-2054,(2019),.
9. Govinda K., Srikanth Deelip.M, "Big Data Analytics using Hadoop over Cloud", International Journal of Pure and Applied Mathem atics, Vol.118,No.9,(2018)..
10. Govinda K., Narendra.B, "Opinion Mining using Classification Techniques", International Journal of Pure and Applied Mathematics, Vol.118,No.9,(2018).
11. Cai, M., Yin, Y. and Xie, M. (2009). Prediction of Hourly Air Pollutant Concentrations near Urban Arterials Using Artificial Neural Network Approach. Transp. Res. Part D 14: 32–41.
12. K.Govinda, Shruthi Hiremath, "Rainfall Prediction Using Artificial Neural Network", International Journal of Applied Engineering Research,Vol:9,No:23, pp: 21231-21241..
13. Venugopal V.K., Naveen A., Rajkumar R., Govinda K., Masih J, "Low cost audio based intelligent guidance system for visually impaired people",, International Journal of Psychosocial Rehabilitation, *Vol:24, Issue: 3, Pg.No(515-520), DOI: 10.37200/IJPR/V24I3/PR200809.*.
14. Nalluri S., Vijaya Saraswathi R., Ramasubbareddy S., Govinda K., Swetha E, Chronic Heart Disease Prediction Using Data Mining Techniques, Advances in Intelligent Systems and Computing, *Vol:1079*, *Pg.No(903-912), DOI: 10.1007/978-981-15-1097-7_76.*
15. de Gennaro, G., Trizio, L., Di Gilio, A., Pey, J., Pérez, N., Cusack, M., Alastuey, A. and Querol, X. (2013). Neural Network Model for the Prediction of PM10 Daily Concentrations in Two Sites in the Western Mediterranean. Sci. Total Environ. 463–464: 875–883.