International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Analysis of air quality of Jabalpur city and prediction by using WEKA software

Anil Kumar Patel[1], *Shailza Verma*[2]

Jabalpur Engineering College, Environmental Engineering

ABSTRACT:

In today's rapidly increasing industrial environment, air pollution is becoming a reason to worry. The air quality is deteriorating day by day due to careless approach, so causing a serious impact on human health. A need has aroused to predict the level of future air pollutants by entering past data. The development of data mining application such as classification, clustering, prediction has shown the necessity for machine learning algorithm to be applied to environment data. In this Thesis the framework is proposed using data mining to study the existing pattern of air pollution data and to predict future pattern of it. The data mining tool WEKA (Waikato Environment for Knowledge Analysis) is used to compare various prediction techniques. It is an open source data mining software and it consist of machine learning algorithms. The aim of this Thesis is to study the performance of different prediction method such as Linear Regression, Gaussian, IBK(KNN), on air pollution data of Jabalpur city.

Keywords: Data Mining, Air Pollution, WEKA, Bagging, Linear algorithm, Regression Analysis, Time Series analysis

Introduction:

The pollution is playing important role in our world in our world as it affects human beings, animals, our planet and all living things. Pollution can lead to unstable climate change which can disrupt the ecosystem. Air pollutant is one of the world's biggest killers. Pollution was directly involved in 4.2 million deaths in 2021. Furthermore, pollution was the cause of 19% of all cardiovascular deaths worldwide, 21% of stroke death and 23% for lung cancer deaths. Because of this, the authoritative predicting technique is needed to lead us to an important role in the danger of crisis response and emergency plans (Wang, H et al, 2016).

This study looks for predicting air pollutant using an artificial K nearest neighbors. Air pollutant prediction will support analyzing the changing patterns of type of pollutants. It will also help in arranging the protective measures in cases of pollution problems or disasters and managing them.

Background of the study:

Jabalpur is one of the major cities of Madhya Pradesh located in the central region of India. The city is having a major presence in agricultural production as well as industrial production, focused around many types of commercial crops such as cotton or sugar cane, which are distributed both domestically as well as internationally. It has a sizeable population of over 1.26 million inhabitants, a figure taken from a 2011 census and thus will have grown exponentially since then. With a rapidly growing population, alongside large amounts of infrastructure and urban development taking place, Jabalpur would subsequently be subject to some poor levels of air quality, something that is commonplace in many cities throughout India that are undergoing their own rapid growth in all areas. In early 2021, Jabalpur (formerly known as Jabalpur.) saw PM2.5 levels as high as 47.2 μg/m³ being recorded in late February

Materials and methods:

The methodology involves data collection about the air quality parameters and further analysis of the data have been done.The required data have been downloaded from Madhya Pradesh pollution control website, a good source of data for analysing air pollutants and. A dataset with several missing values are used to analyse and forecast pollutants using the WEKA forecast tool. The complete dataset is the pollution database of Jabalpur city for the period of 2019-2022 period. WEKA is used to pre-process data, remove unused attributes, and duplicate instance using filters, and the resulting dataset is used for the analysis. WEKA is used to extract data for analysis, with no missing data are considered after cleaning and removing extra columns in the data. After pre-processing, the resulting dataset includes predicted value of considered city.

The present work focusses use of Data Mining to study the existing pattern of air pollution data and to predict future pattern of it. A study about the technology identified is conducted and then implemented model is suggested. Data mining tool WEKA 3.8.1 is chosen for the analysis. Different analyses are done on air pollution data of Jabalpur city and regression algorithms are used to understand the data behaviour pattern. Predictive analysis is used to evaluate the future trend of data using ibk (KNN), linear regression, and Gaussian method of prediction and compare all these methods. This three method compare with actual data of air pollutant and mean absolute error is calculated. The method which gives least mean absolute error is more accurate to calculate future data of air pollution.

Data collection:

The MPPCB dataset contains a total of 28 months, spread across the years 2019-2022, as detailed in below Table 1. Note that there is a substantial data collection gap between 1st of September 2019 to the 31st of December 2021, during which time the DOAS system was non-operational due to an essential maintenance repair.

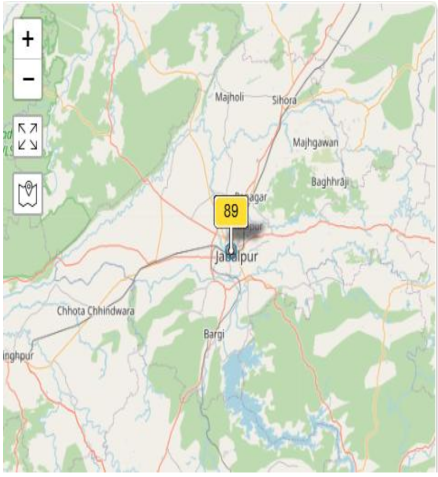


Fig1: the DOAS instrument installed on the premises of Jabalpur

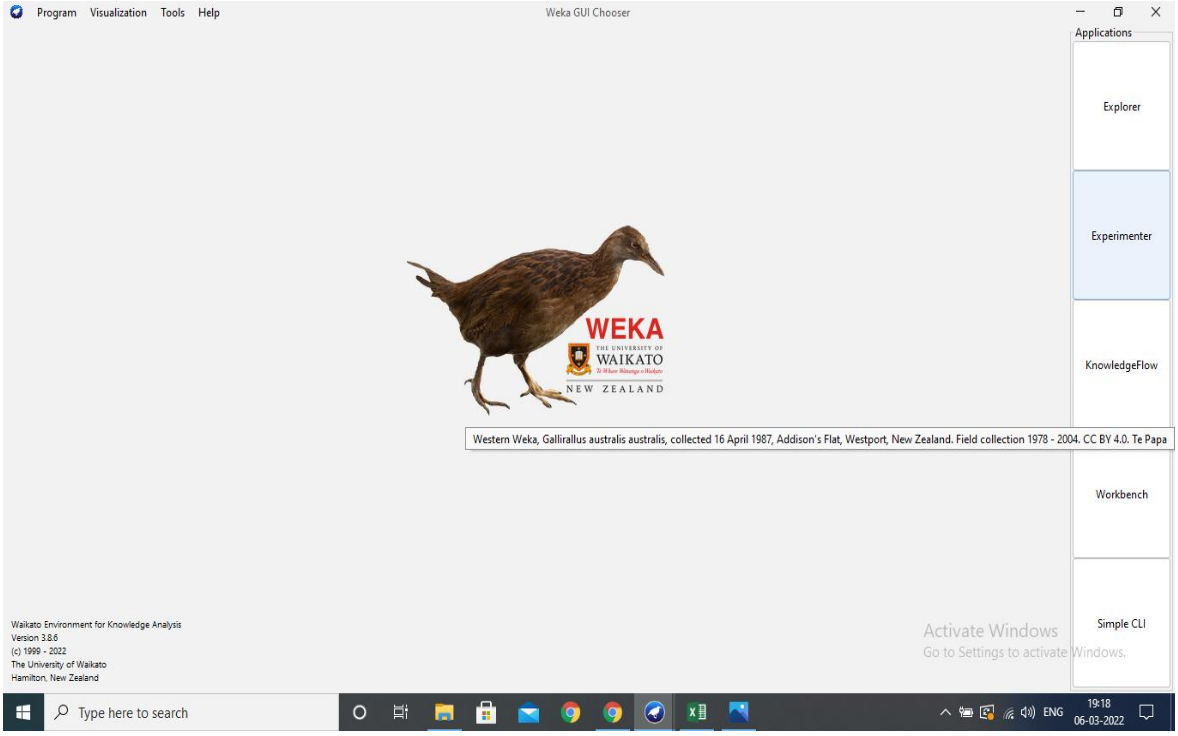
The MPPCB dataset is a sequence of measurements presented in a timeseries. The measurements include concentration values of four gases measuredin µgm^{-3} and two particulate matter measured in µgm^{-3} and CO measured in (mg/m3).

Table 1: Attributesofthe monthly (sep2019- dec2021)Dataset. From MPPCB

PM10 (µg/m3)	PM2.5 (µg/m3)	NOx(ppb)	CO (mg/m3)	SO2 (µg/m3)	NH3 (µg/m3)	Ozone (µg/m3)	
25	11	7.93	0.84	3.43	7	41.2	
135.1	68.15	20.75	1.04	9.2	18.66	41.35	
177.25	93.25	35.29	1.17	9.41	21.52	43.39	
192	111.41	44.25	1.32	12.5	23.04	36	
148.08	78.91	40.67	1.19	7.5	14.6	30.49	
154.74	66.64	42.71	1.02	9.63	12.78	40.66	
95.22	35.78	21.07	0.63	7.74	9.84	50.21	
85.17	35.16	12.95	0.5	6.94	9.89	64.15	
86.55	28.77	14.95	0.53	6.94	8.7	73.65	
56.68	19.32	17.51	0.68	6.62	10.53	55.55	
48.86	20.49	15.66	0.66	13.79	8.67	36.02	
36.53	14.54	14.76	0.69	13.66	6.9	23.68	
56.15	20.72	21.24	0.78	13.48	10.11	32.89	
122.45	49.13	21.8	1.08	10.39	12.88	44.61	
	97.58	33.85	1.22	13.34	20.11	49.41	
202.49	98.25	55.11	2.19	13.89	24.67	42.17	
207.73	106.74	46.74	1.47	17.22	22.97	42.35	
178.3	66.7	52.03	1.2	20.04	23.07	47.17	
156	47.7	38.29	1.04	15.49	20.01	47.99	
150.99	50.79	25.16	1.15	17.02	15.37	61.04	
74.01	27.1	20.69	0.73	16.12	12.99	63.66	
70.32	24.07	22.22	0.87	16	10.4	57.99	
46.83	12.89	14.94	0.74	10.69	11.35	27.16	
68.95	20.51	15.76	0.75	8.44	10.85	24.31	
47.28	16.52	14.37	0.78	8.34	10.75	29.57	
121.79	38.42	26.23	1.04	6.89	11.53	40.03	
214.13	99.14	45.08	1.69	16.89	25.62	56.01	
196.39	91.88	36.48	1.65	14.81	28.79	46.51	

Prediction of air quality by Weka Tool software:

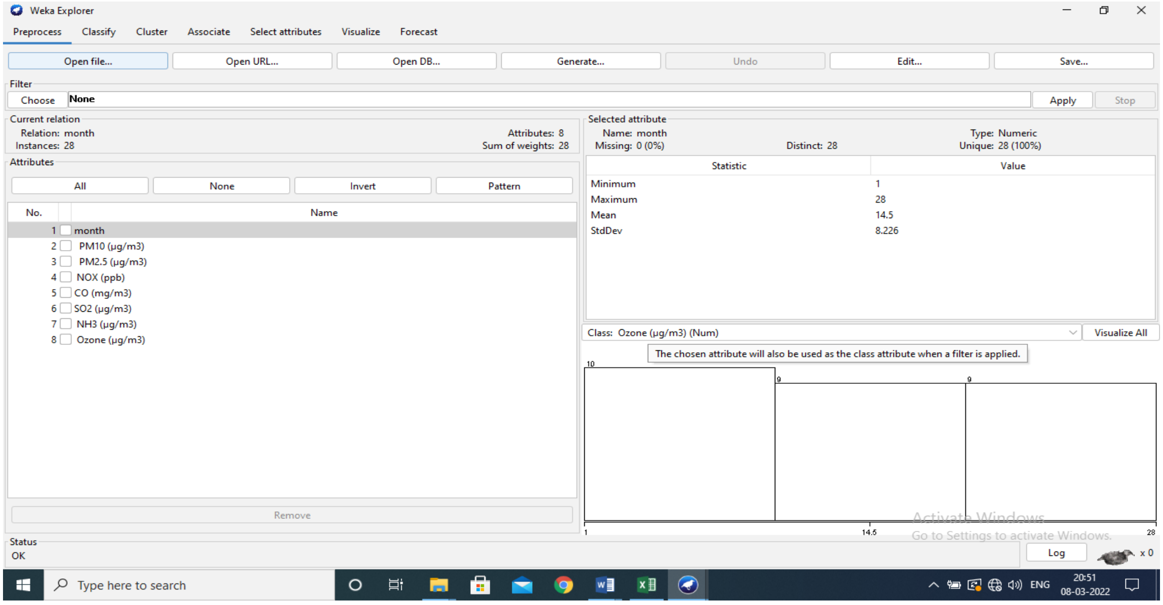
Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, forecasting, regression, clustering, association rules mining, and visualization.



Western Weka, Gallirallus australis australis, collected 16 April 1987, Addison's Flat, Westport, New Zealand. Field collection 1978 - 2004. CC BY 4.0. Te Papa

Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

Fig2: Screen shot of Weka software



Filter: None

Current relation
Relation: month
Instances: 28
Attributes: 8
Sum of weights: 28

Attributes

No.	Name
1	month
2	PM10 (µg/m3)
3	PM2.5 (µg/m3)
4	NOX (ppb)
5	CO (mg/m3)
6	SO2 (µg/m3)
7	NH3 (µg/m3)
8	Ozone (µg/m3)

Selected attribute
Name: month
Missing: 0 (0%)
Distinct: 28
Type: Numeric
Unique: 28 (100%)

Statistic	Value
Minimum	1
Maximum	28
Mean	14.5
StdDev	8.226

Class: Ozone (µg/m3) (Num)

The chosen attribute will also be used as the class attribute when a filter is applied.

Status
OK

Fig3.: Weka software application show various attribute

Flow chart:

flow chart of the step taken to prediction of air pollutant parameter by using Weka tool**.**

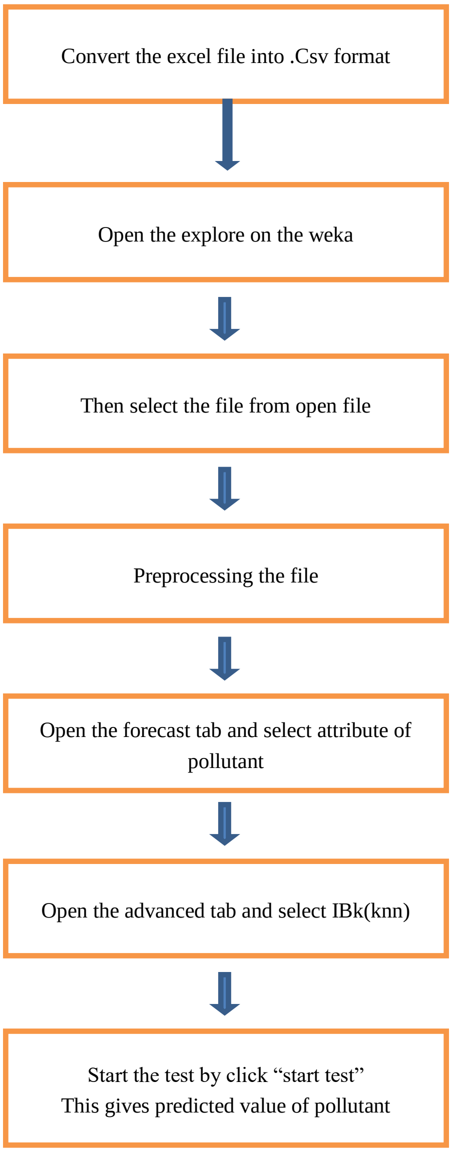

Convert the excel file into .Csv format
Open the explore on the weka
Then select the file from open file
Preprocessing the file
Open the forecast tab and select attribute of pollutant
Open the advanced tab and select IBk(knn)
Start the test by click "start test"
This gives predicted value of pollutant

Fig 4: flow chart of Weka tool application

Result and Discussions:

The result and discussion chapter uses the data analysis Weka tool. For the establishment of urban air quality assessment, the dependent variable is the pollution level, and the independent variable is the characteristics factor of air pollution. According to the standard, the six characteristics factors of CO, O3, NOx, PM10, PM2.5, SO2, NH3 ,are used the decision characteristic variables in the prediction of air pollutant .In this chapter we analyse the different air pollution and based on this data we are apply three method of prediction like k-nearest neighbor , linear regression , and Gaussian method Which is used for future prediction . These three method predict the pollutant concentration for the future. But this is accurate or not we calculate mean absolute error. So we conclude the result of actual data and predicted data is close or not. This chapter is helpful to forecast future data of air pollution.

The framework is proposed using Data Mining to study the existing pattern of air pollution data and to predict future pattern of it. A study about the technology identified is conducted and then implemented model is suggested. Data mining tool WEKA 3.8.1 is chosen for the analysis. Different analyses are done on air pollution data of Jabalpur city and regression algorithms are used to understand the data behaviour pattern. Predictive analysis is used to evaluate the future trend of data using ibk (KNN), linear regression, and Gaussian.

Table shows the predicted value of year 2022 of 2 months (Jan-Feb).The .csv file is imported from Weka Explorer- Process module. Being time series data, different regression algorithms are used for forecasting. And values for the same air pollutants (PM10, PM2.5, CO, $NO_{X,}$ SO_2, $NH_{3,}$ O_3) for the year 2022

Prediction BY KNN:

Table 2: prediction of air pollutant by KNN

Month	PM10	PM2.5	NOx	CO
January	162.02	90.51	28.92	1.10
February	158.46	92.13	31.79	1.08

Month	SO_2	NH_3	O_3	
January	8.60	24.64	22.46	
February	8.37	23.75	37.23	

Prediction BY linear regression:

Table 3: prediction of air pollutant by linear regression

Month	PM10	PM2.5	NOx	CO
January	151.19	66.75	29.07	1.09
February	153.04	63.00	30.51	1.13

Month	SO_2	NH_3	O_3	
January	12.69	28.04	30.37	
February	11.98	27.76	32.51	

Prediction BY linear Gaussian:

Table 4: prediction of air pollutant by Gaussian

Month	PM10	PM2.5	NOx	CO
January	174.86	81.86	34.43	1.43
February	196.799	92.35	41.54	1.64

Month	SO_2	NH_3	O_3	
January	14.53	30.07	28.38	
February	14.08	33.08	28.46	

Actual data of Jan and Feb (2022)

Table5: actual data of Jan and Feb (2022)

Month	PM10	PM2.5	NOx	CO	SO_2	NH_3	O_3
January	175.98	83.32	32.53	1.5	9.86	26.19	29.87
February	159.75	85.57	27.94	1.21	10.85	28.19	32.81

Mean Absolute Erroras compare predicted value of air pollutant and actual value of air pollutant and mean absolute error is calculated.

For KNN:

Table 6: mean absolute for KNN

PM10	PM2.5	NOx	CO	SO2	NH3	O3
6.335	17.47	3.73	.135	1.125	2.225	6.315

For linear regression:

Table 7: mean absolute for linear regression

PM10	PM2.5	NOx	CO	SO2	NH3	O3
9.04	12	3.015	.165	.92	1.5	6.315

For Gaussian:

Table 8: mean absolute for Gaussian

PM10	PM2.5	NOx	CO	SO2	NH3	O3
19.08	19.12	5.85	.25	1.84	.58	1.815

When we compared mean absolute error of all method use for prediction linear regression method given least mean absolute as compared to another two Method that is used. As result show the linear regression method is most nearest to the actual value as compared to all air pollutant parameter.

Prediction of air pollution for the year 2022 to 2025

Table 9: Prediction of air pollution for the year 2022 to 2025

year	PM10	PM2.5	NOx	CO	SO_2	NH_3	O_3
2022	150.78	66.91	27.13	1.18	13.42	28.19	30.85
2023	151.03	70.60	26.17	1.22	12.95	27.77	31.06
2024	151.33	75.34	24.93	1.26	13.42	25.98	31.72
2025	148.38	78.45	21.96	1.24	13.22	26.13	32.07

Conclusion:

The present work shows that the prediction of different air pollutant parameter by using KNN, linear regression and Gaussian method. And compare these predicted value to actual value and calculate the how these value close to the actual value of air pollutant. Linear regression shows more close to the actual value of air pollutant. This thesis proved that Data mining techniques are valuable tool that could be used for environmental monitoring and natural resources science field, and are thus of interest to NGO'S, Municipal Corporations etc. This work is a better starting point for implementation of data mining for real world examples. The use of open-source data mining tools is the main advantages of this study. WEKA is a very alternative over other existing tools available. The current value of the air pollutants is provided to different algorithms such as ibk (KNN), linear regression, and Gaussian.

References:

[1] W. Wang, C. Men, and W. Lu, "Online prediction model based on support vector machine," Neurocomputing, vol. 71, no. 4–6, pp. 550–558, Jan. 2008

[2] A. S. Luna, M. L. L. Paredes, G. C. G. de Oliveira, and S. M. Corrêa, "Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil," Atmospheric Environment, vol. 98, pp. 98–104, 2014.

[3] N. Jiang and M. L. Riley, "Exploring the Utility of the Random Forest Method for Forecasting Ozone Pollution in SYDNEY," Journal of Environment Protection and Sustainable Development, vol. 1, no. 5, pp. 245–254, 2015.

[4] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," International Journal of Electrical Power & Energy Systems, vol. 60, pp. 126–140, Sep. 2014.

[5] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," Atmospheric Environment, vol. 80, pp. 426–437, Dec. 2013.

[6] A. J. Cannon and E. R. Lord, "Forecasting Summertime Surface-Level Ozone Concentrations in the Lower Fraser Valley of British Columbia: An Ensemble Neural Network Approach," Journal of the Air & Waste Management Association, vol. 50, no. 3, pp. 322–339, Mar. 2000.

[7] U.S. Environmental Protection Agency, "Guidelines for Developing an Air Quality (ozone and PM2.5) Forecasting Program," 2003.

[8] U. Schlink, S. Dorling, E. Pelikan, G. Nunnari, G. Cawley, H. Junninen, A. Greig, R. Foxall, K. Eben, T. Chatterton, J. Vondracek, M. Richter, M. Dostal, L. Bertucco, M. Kolehmainen, and M. Doyle, "A rigorous inter-comparison of ground-level ozone predictions," Atmospheric Environment, vol. 37, no. 23, pp. 3237–3253, Jul. 2003.

[9] J. Gomez-Sanchis, J. Martın-Guerrero, E. Soria-Olivas, J. Vila-France´s, J. Carrasco, and S. Valle-Tasco´n, "Neural networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration," Atmospheric Environment, vol. 40, no. 32, pp. 6173–6180, Oct. 2006.

[10] Department of the Environment and Heritage, "Ground-level ozone (O3) - Air quality fact sheet," Commonwealth of Australia, 2005. [Online]. Available: http://www.environment.gov.au/protection/publications/factsheet-ground-level- ozone-o3. [Accessed: 05-Nov-2016].

[11] WEKA; the University of Waikato, "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/index.html. [Accessed: 27-Feb-2015].

[12] Kurt, A., Gulbagci, B., Karaca, F., & Alagha, O. (2008). An online air pollution forecasting system using neural networks. Environment International, 34(5), 592-598.

[13] Cai, M., Yin, Y., & Xie, M. (2009). Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. TransportationResearch Part D:Transport and Environment, 14(1), 32-41.

[14] Wan, F., & Lei, K. S. (2008). Adaptive Neuro-Fuzzy Approach to Air Pollution Prediction in Macau.

[15] Pérez, P., Trier, A., & Reyes, J. (2000). Prediction of PM2. 5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment, 34(8), 1189-1196

[16] Kolehmainen, M., Martikainen, H., & Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting. Atmospheric Environment, 35(5), 815-825.