Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
(Print copies up to 2016)
e-ISSN: 2395-3454

Vol. 22 | No. 2 | pp. 853-860 | 2023

Original Research Paper

https://doi.org/10.46488/NEPT.2023.v22i02.027

Open Access Journal

Comparison of Machine Learning Models in the Prediction of Accumulation of Heavy Metals in the Tree Species in Kanchipuram, Tamilnadu

R. Sumathi*† and G. Sriram**

*Department of Civil and Structural Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Enathur, Kanchipuram-631 561, Tamilnadu, India

**Department of Mechanical Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Enathur, Kanchipuram-631 561, Tamilnadu, India

†Corresponding author: R. Sumathi; sumathiram72@gmail.com

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com

Received: 14-09-2022
Revised: 03-11-2022
Accepted: 10-11-2022

Key Words:
Machine learning
Heavy metals
Accumulation
Plant species
Prediction

ABSTRACT

Arsenic, aluminum, iron, lead, chromium, copper, zinc, manganese, and cadmium are some of the heavy metal pollutants in the air that cause severe impacts on the biotic and abiotic environment. This study intended to find the accumulation capacity of the heavy metals on the leaves of tree species such as *Terminalia catappa*, *Syzygium cumini*, *Saraca asoca*, *Pongamia glabra*, and *Ficus religiosa* and predict their accuracy by comparing different machine learning (ML) models. The samples were collected at six different locations (likely Vellagate, Cancer Institute, CSI hospital area, Moongilmandapam, Collectorate, and Pallavarmedu) and distributed in a manner within Kanchipuram town, Tamil Nadu, in February and March of 2018 and 2019, respectively. Six ML methods were selected, such as KStar (K*), Lazy IKB, Logistic Regression Algorithm (LR), LogitBoost Classifier (LB), Meta Randomizable Filtered Classifier (MRFC), and Random Tree (RT), for prediction and to compare the efficiency of their predictions. Out of six models, Logistic functions perform well in terms of TP rate when compared to other classifiers (93.21%-99.81% TPR– 0.93–0.99) and Logitboost attained a low TP rate that ranged from 0.76 to 0.82. This study indicates the feasibility of different ML methods in the prediction of species capabilities toward the accumulation of heavy metals.

INTRODUCTION

Air pollution is one of the significant problems in the environment and causes serious health hazards to human beings and has a severe impact on non-living things. Industrialization, urbanization, and an increased number of vehicles lead to the emission of various gases, particulate matter, and heavy metals. Out of this, heavy metals pose complexity on human health, and sometimes losses are inexpressible, so it is essential to quantify them and also reduce their concentration in the atmosphere. Monitoring by equipment presented many challenges, including the high cost of establishing sampling stations, confined sampling, and extensive labor (Norouzi 2016, Gaza 2018). All these difficulties are overcome with biomonitoring techniques by using a variety of vegetation since plants are available and distributed in remote areas, which makes sampling and monitoring very easy and economical (Sharma et al. 2015). Bio-monitoring measures the pollutant levels in the atmosphere both quantitatively and qualitatively by analyzing the accumulation, deposition, and distribution rates in the environment (Ozturk et al. 2017). Plants like mosses and lichens are perfect biomonitors, but due to their unavailability in industrial and urban areas, higher vascular plants are now used (Chang 2016, Ojiodu 2018, and Aasawari Tak 2017). Heavy metals deposited on the leaves of the trees directly quantify the pollution level in the atmosphere (Maghakyan 2016). The percentage of dust deposited on the leaves of roadside trees is high compared with the trees away from roads (Ahmed 2016). Heavy metals deposited on the trees available in the local areas are an effective tool for measuring the air quality and developing these trees was used to maintain the greenbelt (Hajizadeh 2019).

The results obtained from the analysis were validated by using machine learning tools, which provided more accuracy (Akiladevi 2020). Particulate matters 2.5 in the air were effectively predicted by using auto-regression and logistic regression models (Aditya 2018). Machine learning models such as the linear support vector machine and boosted trees were used to predict the PM 2.5 level in the air based on six climatic factors (Deters 2017). In the ML approach, root

mean square error and mean absolute error were taken as the scales to compare the accuracy of the various regression models (Saba Ameer et al. 2017).

In this study, the heavy metals such as Al, As, Cd, Cr, Pb, Mn, Fe, Cu, and Zn deposited on the leaves of *Saraca asoca* (Ashoka), *Terminalia catappa* (Badam), *Ficus religiosa* (Pupil), *Pongamia glabra* (Pongam), and *Syzygium cumini* (Jamun) were experimentally analyzed. The accuracy of the results was predicted by a machine learning approach using various algorithms, namely KStar (K*), Lazy IKB, Logistic Regression Algorithm (LR), LogitBoost Classifier (LB), Meta Randomizable Filtered Classifier (MRFC), and Random Tree (RT).

MATERIALS AND METHODS

Sampling Species

In this present study, five trees, such as *S. asoca*, *T. catappa*, *F. religiosa*, *P. glabra*, and *S. cumini*, were selected based on their easy availability and generally found in all the selected sites. Out of these five trees *S. asoca* and *S. cumini* are evergreen trees, *T. catappa* and *P. glabra* are deciduous trees, and *F. religiosa* comes under both categories. Leaves from all six selected tree species were sampled in the early hours of the day, from 6 a.m. to 8 a.m., during February and March of 2018 and 2019. Samples were collected at two various heights, viz., less than 1.8 m and above 2.4 m, to identify the pollutants at any level from different sources. The collected samples were stored carefully in zip-lock covers to prevent any addition or deletion of pollutants and brought to the laboratories for analysis. The experiment was carried out by using inductively coupled plasma mass spectrometry after the closed microwave digestion process.

Study Area

Kanchipuram is a district in the northeast of Tamil Nadu, adjacent to the Bay of Bengal. It is bounded in the west by Vellore and Thiruvannamalai districts, in the north by Thiruvallur District and Chennai District, and the south by Villupuram District. It lies between 11°00' and 12°00' north latitudes and 77°28' and 78°50' east longitudes. The district has a total geographical area of 4,432 km^2 and a coastline of 57 km. The town of Kanchipuram is the district headquarters. The maximum and minimum temperatures range from 28.0°C to 45.0°C and 14.0°C to 21.0°C, respectively. Fig. 1 shows the index map of Kanchipuram with the site location of the sample collection. Samples were collected at six different locations in the distributed way in Kanchipuram town, such as the residential area (Pallavarmedu), commercial area (Collectorate), sensitive area (CSI hospital), institutional area (Cancer Institute), industrial area (Vella gate), and traffic area (Moongil Mandapam).

Dataset Construction

The baseline factors were gathered from the actual dataset and each factor was converted to be convenient for the

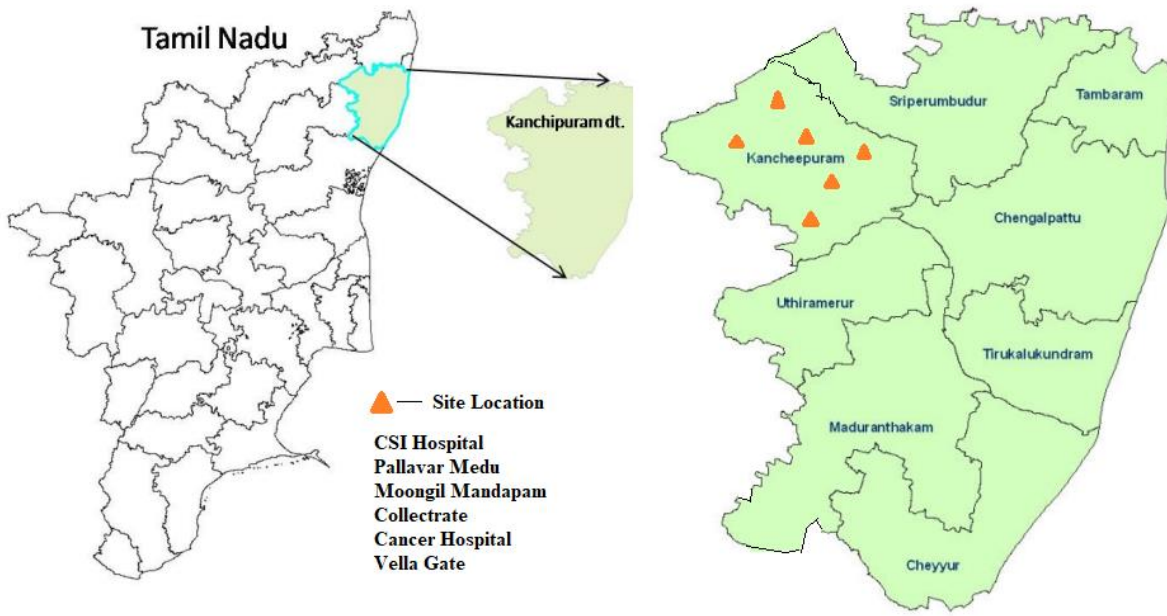


Fig. 1: Study Area – Kanchipuram.

cluster analysis to the scale between 0 and 1 and analyzed by weka.clusterers SimpleKMeans. After framing the dataset, the program code for reference was prepared. The factors that were fed as input data were latitude, longitude, nine heavy metals (Al, As, Fe, Cu, Cd, Cr, Pb, Zn and Mn), site, site name and sample name, totaling 14 attributes and 30 instances, such as five species at six sites.

RESULTS AND DISCUSSION

The results obtained from the experiments were taken for statistical analysis. Mean, median and standard deviation were calculated for 2018 and 2019 (Tables 1 and 2).

Heavy metal deposition on the leaves of the tree species at both the high point and low point for the two consecutive years 2018 and 2019 were taken for the analysis, totaling thirty instances and fourteen attributes. The information or results obtained from the experiment were run by using Weka classifiers, including various classifiers such as KStar (K*), Lazy IKB, Logistic Regression Algorithm (LR), LogitBoost Classifier (LB), Meta Randomizable Filtered Classifier (MRFC) and Random Tree (RT). The time taken

Table 1: Statistical analysis of pollutants deposited at six sites on five species during 2018.

Parameters	Max	Min	Mean	Median	STD
S. asoca- 2018					
Fe	124.8	59.1	93.89	98	23.78
Cu	2.2	0	0.56	0	0.93
Zn	21.15	0	8.69	7.675	9.01
Al	75.65	26.05	48.15	44.45	18.35
As	1.15	0	0.19	0	0.47
Mn	9.8	0	4.45	3.875	4.92
T. catappa					
Fe	86.05	21.7	39.12	33.125	26.15
Cu	1.6	0	0.27	0	0.65
Zn	11.3	0	4.66	4.01	5.22
Al	54.25	5	21.42	15.875	18.07
As	0	0	0.00	0	0.00
Mn	10.1	0	4.47	4.125	4.93
F. religiosa					
Fe	44.5	15	27.61	28.38	11.26
Cu	1.35	0	0.41	0.00	0.64
Zn	8.05	5.95	6.46	6.18	0.79
Al	28.9	14.2	18.26	17.13	5.47
Mn	9.65	6.05	7.38	6.85	1.48
P. glabra					
Fe	130.2	17.5	50.95	31.90	45.40
Cu	2.75	0	0.82	0.00	1.28
Zn	14.45	1.15	6.88	6.15	6.80
Al	109.75	7.85	36.64	21.50	39.29
Mn	14.585	0	6.16	4.63	6.97
S. cumini					
Fe	69.15	7.85	34.27	23.40	25.50
Cu	1.05	0	0.18	0.00	0.43
Zn	10.35	0	3.66	2.63	4.35
Al	43.9	9.4	20.87	13.40	14.50
Mn	14.4	0	4.04	1.71	5.70

Table 2: Statistical analysis of pollutants deposited at six sites on five species during 2019.

Parameters	Max	Min	Mean	Median	STD
S. asoca-2019					
Fe	149.195	10.8	87.55	94.67	64.22
Zn	35.95	1.225	9.38	3.1775	13.55
Al	131.095	3.2365	35.43	14.8275	49.63
Cr	1.185	0	0.20	0	0.48
Mn	37.975	0	8.11	2.323	14.73
T. catappa					
Fe	41.84	5.995	20.46	16.7025	15.83
Zn	31.335	0	6.20	1.3525	12.34
Al	66.23	0	22.47	17.23	24.29
Mn	2.765	0	0.62	0	1.12
F. religiosa					
Fe	52.25	3.425	29.00	25.22	19.08
Zn	8.67	0	3.78	4.06	3.05
Al	61.25	3.985	23.43	15.14	21.32
Mn	5.43	0	1.14	0.00	2.18
P. glabra					
Fe	108.99	18.21	40.73	25.45	34.39
Zn	25.56	0	5.49	2.20	9.91
Al	94.58	4.09	37.99	34.72	34.69
Mn	4.37	0	1.45	0.96	1.79
S. cumini					
Fe	54.255	11.915	30.40	29.95	14.97
Zn	3.39	0	1.62	1.38	1.16
Al	47.68	4.535	24.99	25.23	13.82
Mn	3.23	0	1.52	1.46	1.04

to test the model on the training data was 0.01 seconds. From the evaluation of the training set, results such as correctly classified instances, incorrectly classified instances, mean absolute error, root mean squared error, relative absolute error, and root relative squared error were obtained. The true positive rate is taken as the correctly classified instances. After completing the analysis, the correlation between the various classifiers was calculated, and the performance of the machine

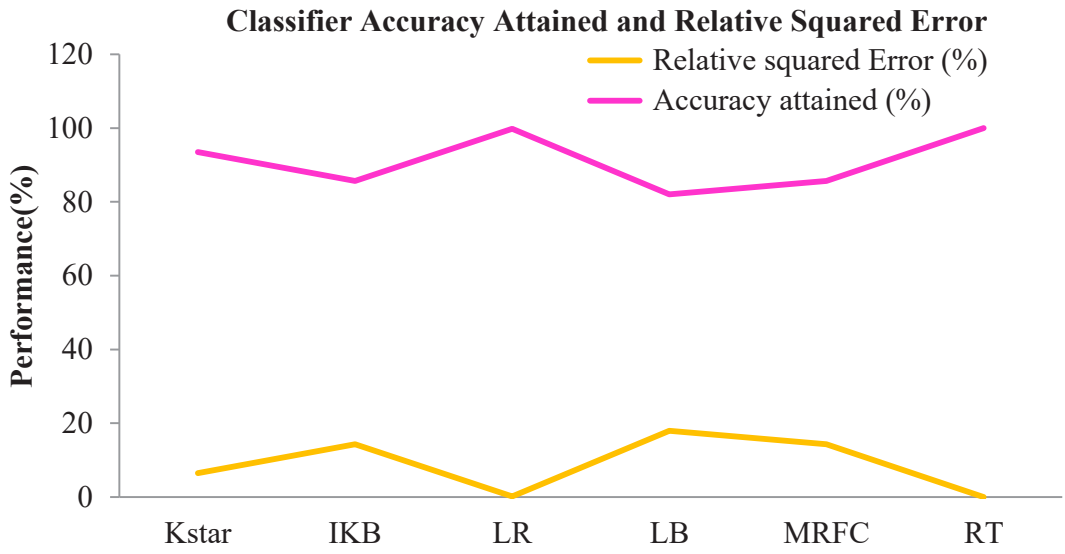


Fig. 2: Performance of the machine learning tools for the prediction of the pollutants at the high point during 2018 (Deposition).

learning tools for the prediction of the pollutants at the high point in 2018 and 2019 respectively is shown in Figs. 2 and 3.

Hyejin Park & Kim (2019) predicted the impact of heavy metals, namely cadmium, mercury, and lead, on hypercholesterolemia (HC) in the population and compared the accuracy of various five machine learning algorithms based on the data received from the Korea National Health Department. Vijayarani & Muthulakshmi (2013) analyzed the performance of two classifiers, namely Lazy and Bayesian, by considering various factors and proved that the Bayesian classifier is less efficient than the Lazy classifier.

Figs. 4 and 5 show the performance of the machine learning tools for the prediction of the pollutants at the high point and low point during 2018 and 2019 respectively.

The accuracy measures of different classifiers for high and low points for the deposition during 2018 and 2019 were listed in Tables 3 and 4. and it is observed that logistic functions perform well in terms of TP rate when compared to other classifiers.

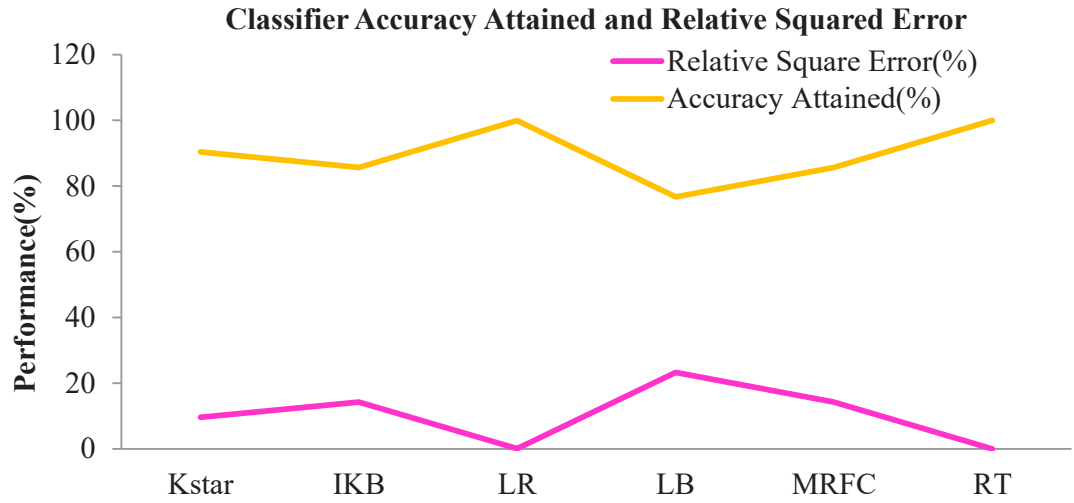


Fig. 3: Performance of the machine learning tools for the prediction of the pollutants at the low point during 2018 (Deposition).

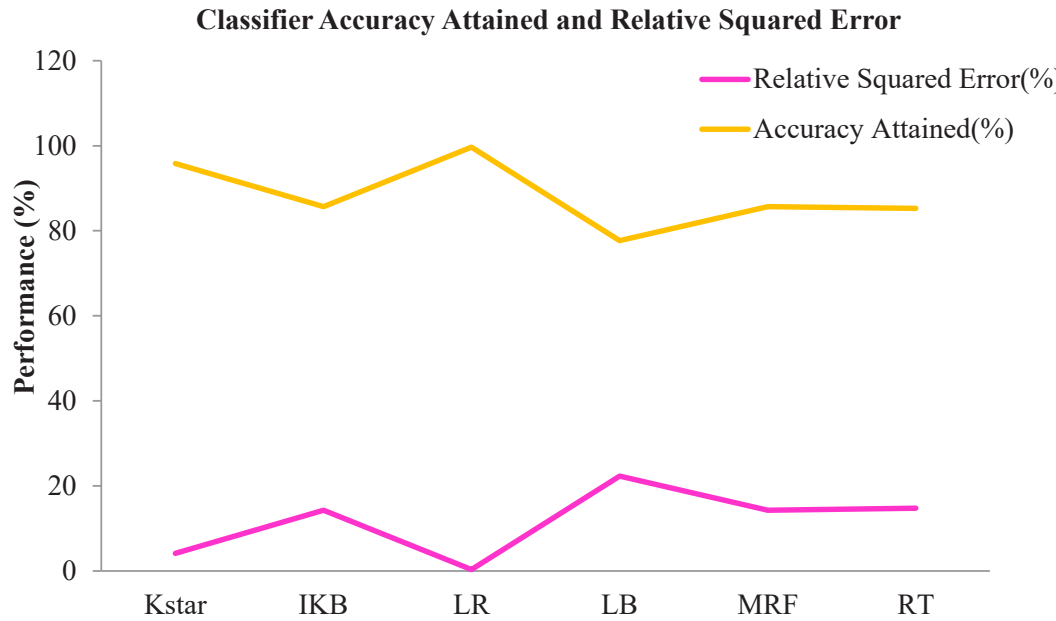


Fig. 4: Performance of the machine learning tools for the prediction of the pollutants at the high point during 2019 (Deposition).

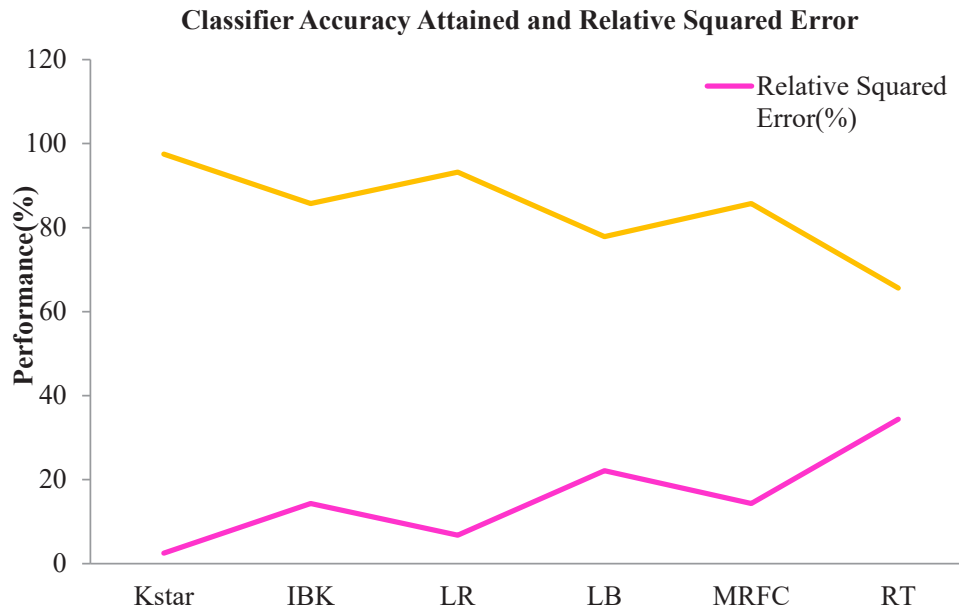


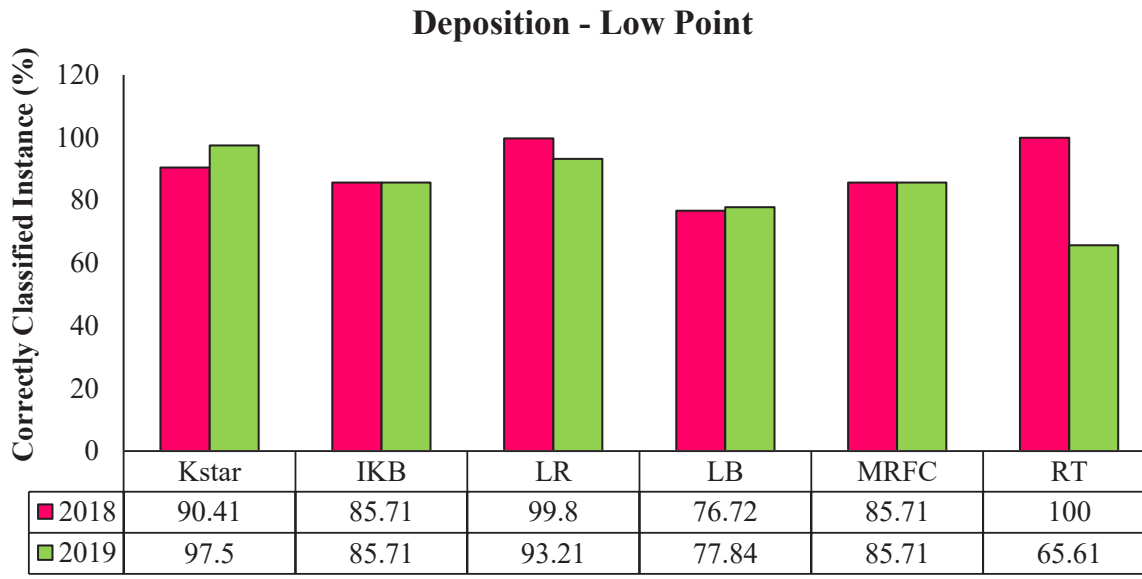
Fig. 5: Performance of the machine learning tools for the prediction of the pollutants at the low point during 2019 (Deposition).

Table 3: Efficiency analysis of each data classification model- High Point (Deposition).

Performance Error	Different Model Algorithms											
	Kstar		Lazy IKB		Functions-Logistic		Logit Boost		Meta Randomizable Filtered classifier		Trees-Random Tree	
Year	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019
Number of Selected Attributes	14	14	14	14	14	14	14	14	14	14	14	14
Correctly Classified Instances(%)	93.52	95.84	85.71	85.71	99.81	99.67	82.04	77.69	85.71	85.71	100	85.27
Incorrectly Classified Instances%)	6.48	4.16	14.29	14.29	0.19	0.33	17.96	22.31	14.29	14.29	0	14.73
TR Rate(True Positive Rate)	0.93	0.95	0.85	0.85	0.99	0.99	0.82	0.77	0.85	0.85	1	0.85

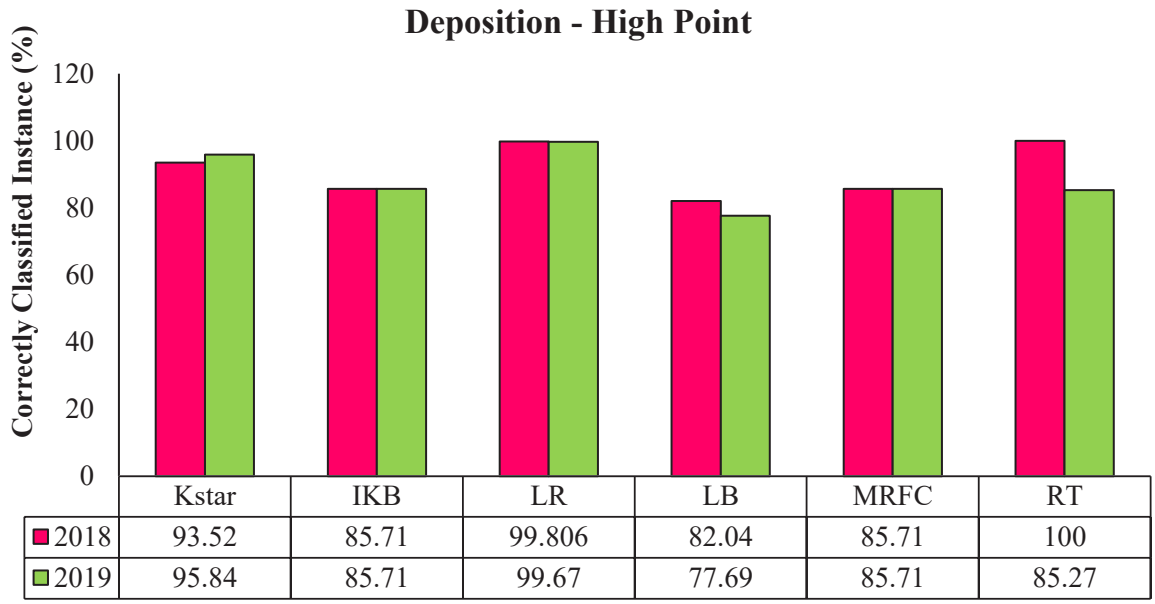
Table 4: Efficiency analysis of each data classification model- Low Point (Deposition).

Performance Error	Different Model Algorithms											
	Kstar		Lazy IKB		Functions-Logistic		Logit Boost		Meta Randomizable Filtered classifier		Trees-RandomTree	
Year	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019
Number of Selected Attributes	14	14	14	14	14	14	14	14	14	14	14	14
Correctly Classified Instances(%)	90.41	97.50	85.71	85.71	99.80	93.21	76.72	77.84	85.71	85.71	100	65.61
Incorrectly Classified Instances%)	9.59	2.50	14.29	14.29	0.20	6.79	23.28	22.16	14.29	14.29	0	34.39
TR Rate(True Positive Rate)	0.90	0.97	0.85	0.85	0.99	0.93	0.76	0.77	0.85	0.85	1	0.65



	Kstar	IKB	LR	LB	MRFC	RT
2018	90.41	85.71	99.8	76.72	85.71	100
2019	97.5	85.71	93.21	77.84	85.71	65.61

Fig. 6: Comparison between different classifier for predicting pollutants measured at low point of the tree (Deposition).



	Kstar	IKB	LR	LB	MRFC	RT
2018	93.52	85.71	99.806	82.04	85.71	100
2019	95.84	85.71	99.67	77.69	85.71	85.27

Fig. 7: Comparison between different classifier for predicting pollutants measured at high point of the tree (Deposition).

Figs. 6 and 7 show the comparison between the classifiers' correctly classified instances of a low point and high point during the deposition of heavy metals on the leaves from 2018 to 2019, respectively.

CONCLUSIONS

It is well known that machine learning techniques using various prediction models are widely used in various applications and in different fields with strong abilities. In this research paper, the deposition of nine heavy metals on the leaves of five tree species was predicted using six ML classifiers and compared with each other. The results obtained from the study indicated that the performance of the machine learning tools for the prediction of the pollutants at the high point and low point during 2018 and 2019 was good. The deposition of heavy metals was analyzed both in 2018 and 2019 at high and low points, with logistic boost methods

showing lower performance and logistics functions showing better performance with higher accuracy and a true positive rate. Based on these findings, ML may serve as an alternate method for predicting the deposition of heavy metals from the atmosphere.

REFERENCES

Aasawari Tak, A. and Umesh Kakde, B. 2017. Assessment of air pollution tolerance index of plants: A comparative study. Inter. J. of Phar. and Pharma. Sci., 9(7): 83-89. http://dx.doi.org/10.22159/ijpps.2017v9i7.18447

Aditya, C R. Chandana Deshmukh, R. Nayana, D K. and Praveen Gandhi Vidyavastu. 2018. Detection and prediction of air pollution using machine learning models. Inter. J. of Engg. Trends and Tech., 59(4): 204-207.

Ahmed, S.S. Jabeen, R. Johar, S. Hameed, M. and Irfan, S. 2016. Effects of roadside dust pollution on fruit trees of Miyyaghundi (Quetta) and Ghanjdori (Mastung), Pakistan. Inter. J. of Basic and App. Sci., 5(1): 38-44. http://dx.doi.org/10.14419/ijbas.v5i1.5477

Akiladevi, R. Nandhini Devi, B. Nivesh Karthick, V. and Nivetha, P. 2020. Prediction and Analysis of Pollutant using Supervised Machine Learning. Inter. J. of Recent Tech. and Engg., 9(2): 50-54. http://dx.doi.org/10.35940/ijrte.A2837.079220

Ameer, S., Shah, M.A., Khan, A., Song, H., Maple, C., Islam, S.U. and Asghar, M.N., 2019. Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access, 7: 128325-128338.

Chang, L., Peng, Z. and Yanming, F. 2016. Monitoring airborne heavy metal using mosses in the city of Xuzhou, China. Bull. of Environ. Contamin. and Toxi., 96: 638-644.

Deters, J.K. Zalakeviciute, R. MarioGonzalez, R. and Rybarczyk, Y. 2017. Modeling PM2.5 urban pollution using machine learning and selected meteorological parameters. Hind. J. of Electri. and Comp. Engg., pp. 1-14. https://doi.org/10.1155/2017/5106045

Etim, N.E.O. Anthony, O. E. Mbom-Obong, N.E. and Dodeye, O. 2015. Heavy metal levels in Pine (*Pinus caribeae* Morelet) tree barks as indicators of atmospheric pollution Calabar Municipality, South Eastern Nigeria. J. of Environ. and Earth Sci., 5(22): 30- 32.

Gaza, T. and Kugara, J. 2018. Study of heavy metal air pollution, using a Moss (Grimmia dissimulate) biomonitoring technique. Univer. J. of Chem., 6(1): 1-13. https://doi.org/10.13189/ujc.2018.060101

Hajizadeh, Y. Mokhtari, M. Faraji, M. Abdolahnejad, A. and Mohammadi, A. 2019. Biomonitoring of airborne metals using tree leaves: Protocol for biomonitor selection and spatial trend. MethodsX. 6: 1694–1700. https://doi.org/10.1016/j.mex.2019.07.019

Hyejin Park, H. and Kim, K. 2019. Comparisons among machine learning models for the prediction of hypercholestrolemia associated with exposure to lead, mercury, and cadmium. Inter. J. of Environ. Res. and Pub. Health., 16: 1-8. https://doi.org/10.3390/ijerph16152666

Maghakyan, N. Tepanosyan, G. Belyaeva, O. Sahakyan, L. and Saghatelyan, A. 2016. Assessment of pollution levels and human health risk of heavy metals in dust deposited on Yerevan's tree leaves (Armenia). Acta Geochimica., 36(1): 16- 26. https://doi.org/10.1007/s11631-016-0122-6

Mohammad, B. Md. Venna, E.P.R. Pallepogu, C.P. and Madhu Babu Redapongala, B.M. 2020. Predictive modelling of air pollution using machine learning models and neural networks. Inter. J. of Scien. & Tech. Res., 9(6): 623-631.

Norouzi, S. Khademi, H. Cano, A.F. and Acosta, J.A. 2016. Biomagnetic monitoring of heavy metals contamination in deposited atmospheric dust, a case study from Isfahan, Iran J. of Environ. Manage., 173: 55-64. http://doi.org/10.1016/j.jenvman.2016.02.035

Ojiodu, C.C. Olumayede, E.G. and Okuo, J.M. 2018. The level of heavy metals in the atmosphere of a macro environment, Lagos State, Southwestern - Nigeria using Moss plant (*Dicranium scorparium*) as bioindicator. Sci. World J., 13(4): 69-74.

Ozturk, A. Yarci, C. and Ozyigit, I.I. 2017. Assessment of heavy metal pollution in Istanbul using plant (Celtis australis L.) and soil assays. Biotech. & Biotechnological Equip., 31(5): 948- 954. https://doi.org/10.1080/13102818.2017.1353922

Sharma, A. and Uniyal, S.K. 2015. Heavy metal accumulation in *Pyrrosia flocculosa* (D. Don) Ching growing in sites located along a vehicular disturbance gradient. Environ. Moni. and Assess., 188(547): 1-12. https://doi.org/10.1007/s10661- 016-5561-3

Vijayarani, S. and Muthulakshmi, M. 2013. Comparative analysis of Bayes and Lazy classification algorithms. Inter. J. of Advan. Res. in Comp. and Communi. Engg. 2(8): 3118-3124.