UNIVERSAL WISER
PUBLISHER

Machine Learning by Data Mining REPTree and M5P for Predicating Novel Information for PM_{10}

Yas A. Alsultanny
Arab German Academy for Science and Technology, Germany
E-mail: alsultanny@hotmail.com

Abstract: We examined data mining as a technique to extract knowledge from database to predicate PM_{10} concentration related to meteorological parameters. The purpose of this paper is to compare between the two types of machine learning by data mining decision tree algorithms Reduced Error Pruning Tree (REPTree) and divide and conquer M5P to predicate Particular Matter 10 (PM_{10}) concentration depending on meteorological parameters. The results of the analysis showed M5P tree gave higher correlation compared with REPTree, moreover lower errors, and higher number of rules, the elapsed time for processing REPTree is less than the time processing of M5P. Both of these trees proved that humidity absorbed PM_{10}. The paper recommends REPTree and M5P for predicting PM_{10} and other pollution gases.
Keywords: data mining, machine learning, meteorological, air quality, decision trees, gas concentration, climate change

1. Introduction

Data scientists analyzed data sensors like environmental and meteorological data to get useful information for social benefits. For example, environmental control processes such as improving air quality, and reducing the levels of pollution [1].

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis using sophisticated mathematical algorithms to segment the data and evaluate the probability of future events [2, 3]. The era of big data has accelerated the use of data mining. Data mining methods, with their power and automaticity, have the ability to cope with huge amounts of data and extract value [4, 5].

The oil consumption increased from 10 million barrels daily during the fifties, to 96.5 million barrels per day in 2017, as reported by the International Energy Agency (IEA) [6]. Evidently the issue of global warming was thought to be connected that lead to negative effects on environment. Particular Matter 10 (PM_{10}) is one of the air pollute, that needs monitoring and discovering the rules, that relate with meteorological parameters. The air monitoring station measure the PM_{10} at the rate of one reading per five minute interval. The average of the readings within one hour is used in this paper as an hourly reading.

The importance of this paper lies in using the data mining trees to predicate the rules related PM_{10} and meteorological parameters. Because PM_{10} is one of the important pollution parameters, especially in the Arabic region that have very limited rate of rain.

The paper consists of five sections. After the introduction, the theory and research background introduced. Next, data collection and main research findings are provided in the third section. Finally, research results are discussed.

2. Theory and research background

Information is the key factor that drives the modern world, enabling activities from checking the weather to making complex decisions based on data from weather monitoring stations.

The amount of data in the world is huge, and it grows on an annual basis of 50% of its original size [7]. Collecting data and using it ethically is one of the important issues in data analysis to take accurate decisions [8]. Big data could be used as a useful tool that could enhance decision making [1]. The approach that was used to discover the relation between data resulting from observation is called Knowledge Discovery in Databases (KDD) [9].

Machine learning by data mining methods sometimes are more suited than others to transparent interpretation. For example, decision trees are human friendly for results explanation. Decision trees can be used for classification, estimation,

DOI: https://doi.org/10.37256/ccds.112020418

or prediction [10].

Classification assigns items in a collection to target categories or classes. There are popular machine learning data mining classification algorithms like; REPTree, M5P, C4.5, k-NN, J48, SVM, Naïve Bayes, RandomTree, and Logistical Model Trees [11, 12].

The Reduced Error Pruning Tree (REPTree) is a fast decision tree used with numeric attributes, and it builds a decision tree based on the information by increasing or reducing the variance. It is a decision tree learner, which builds a decision or regression tree using information gained as the splitting criterion, and prunes it, by using reduced error pruning. It deals with missing values by splitting instances into pieces. we can set the minimum number of instances per leaf, maximum tree depth (useful when boosting trees), minimum proportion of training set variance for a split (numeric classes only), and number of folds for pruning [13, 14].

For decision making trees the predicted values on the test instances are $p_1, p_2, \ldots, p_n$; the actual values are $a_1, a_2, \ldots, a_n$. The p_i refers to the numeric value of the prediction for the i^{th} test instance. Mean Absolute Error (MAE) is the average of the magnitude of the individual errors without taking account of their sign calculated by the equation (1) [14].

$$\text{Mean absolute error} = \frac{|p_1 - \overline{a}_1| + \ldots + |p_n - \overline{a}_n|}{n} \quad (1)$$

Where:

p: are predicted values

a: are actual values

$\overline{a}$: actual mean values

n: number of variables

The Root Mean-Squared Error (RMSE) calculated by equation (2) to reduce the figure to have the same dimensionality as the quantity being predicted.

$$\text{Root Mean-Squared Error} = \sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}} \quad (2)$$

The relative errors are important in showing the relative error according to the actual values. The errors are normalized by the error of the simple predictor that predicts average values. The relative absolute error is the total absolute error. Calculated by equation (3).

$$\text{Relative absolute error} = \frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \overline{a}| + \ldots + |a_n - \overline{a}|} \quad (3)$$

The root relative squared error calculated by equation (4).

$$\text{Root relative squared error} = \sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(a_1 - \overline{a})^2 + \ldots + (a_n - \overline{a})_2}} \quad (4)$$

The correlation coefficient measures the statistical correlation between the a's and the p's. They are rated between -1 and +1. The positive values mean positive correlation, while the negative values mean negative correlation. Table 1 shows the scale of 5 levels used to measure the strength of the correlation [15].

Table 1. Interpreting strengths of correlations

N	Correlation value	Interpretation
1	± .70 or higher	Very strong correlation
2	± < .70 to ± .40	Strong correlation
3	± < .40 to ± .30	Moderate correlation
4	± < .30 to ± .20	Weak correlation
5	± < .20 to ± .01	No or negligible correlation

Source: www.quinnipiac.edu [15]

The absolute error and root mean square error are used in the REPTree and M5P algorithms. The RMSE tends to be higher than MAE as the distribution of error magnitudes becomes more variable. The MAE measures the average of the errors in a set of forecasts and creates a linear score which means that all the individual differences are weighted equally in the average. The RMSE measures the average of the errors. The RMSE gives a relatively high weight to large errors. Both MAE and RMSE can be used to diagnose the variation in the error in a set of predications. Both the MAE and RMSE can range from 0 to ∞ [16].

REPTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects the best one from all generated trees [17]. Dragomir in 2016 [18], utilized the REPTree to forecast air quality by generating decision trees to extracts heuristic predictive rules. Vitkar in 2017 [19], applied the REPTree to explore the pattern of air pollution data and to predict the air pollution parameters.

In Weka M5 algorithm called M5P, where 'P' stands for 'Prime' This algorithm uses "divide and conquer" to generate decision lists and sets of if-then rules for regression problems. The M5P algorithm generates accurate classifiers, particularly when most of the attributes are numeric. The M5P algorithm measure both MAE and MASE to evaluate the proposed model [20]. The M5P has been used in different fields such as the environment for predicting daily pollution concentrations [21]. Dragomir in 2016 used the M5P algorithm to forecast PM_{10} and concentrations of air pollutants [18].

3. Data collection and research findings

Weka (Waikato Environment for Knowledge Analysis), is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. It was developed at the University of Waikato in New Zealand. Weka version 3.8 was used in this paper in implement the decision trees, it is available at http://www.cs.waikato.ac.nz/ml/weka [22]. Weka allows users access to its sophisticated data mining routines through a graphical user interface designed for productive data analysis [23]. It contains a collection of algorithms for data mining tasks, including data preprocessing, association mining, classification, regression, clustering, and visualization [24, 25].

The data available for this paper was collected from air pollution monitoring station at Arabian Gulf region in 2017. These data are an hourly readings. To implement the decision trees, the data size for each parameter is 12 × 24= 288 readings (instances), each day have 24 hourly readings, and for 12 months. These data were selected to reflect the effect of the meteorological parameters on thePM_{10}.

The meteorological parameters: temperature (Temp), humidity (Hum), wind speed (WS), and wind direction (WD) were used to predicate PM_{10}, the total size of data 288 × 5 attribute (four meteorological parameters and PM_{10}) =1440 readings. This limited size of data used to visualize a simple tree, and to discover the rules that relate to the PM_{10} with the meteorological parameters. The trees are implemented with 10 folds-cross validation to minimize any bias in the process and improve the effect of the process, as recommended by Juncal-Martínez et al. [13] to gain the best classification model with the training dataset [26].

Figure 1 shows the results of implementing REPTree. The size of this tree is 25 with the number of nodes being 12 and 13 leaves. The leaves of the tree represent the predictive rules of the tree. The process time of building this tree is 0.03 seconds. The correlation coefficient is 0.6702, which indicates a strong correlation between PM_{10} and the four meteorological parameters. The Mean Absolute Error value is 18.9028 and the Root Mean Squared Error value is 24.7308. These two types of errors represent the differences between the real values and predicated values. The following show the run information of this tree:

Instances: 288
Attributes: 5
PM_{10}
Temp
WD
WS
Hum
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
REPTree
==========
Hum < 23
| WS < 2.45: 117.18 (7/253.67) [4/2749.4]
| WS > = 2.45
| | Hum < 16
| | | Temp < 42.5
| | | | Temp <38: 85 (8/448) [6/2440.67]
| | | | Temp > = 38
| | | | | Temp <40.5: 172 (2/64) [1/0]
| | | | | Temp > = 40.5: 126 (2/6.25) [1/2.25]
| | | Temp > = 42.5
| | | | WD <202.5: 109 (3/130.67) [1/2704]
| | | | WD > = 202.5: 61 (3/254) [0/0]
| | Hum. > = 16: 59 (5/248.4) [2/144.5]
Hum > = 23
| WS < 1.75
| | Temp < 31
| | | WD < 80.5: 45.33 (4/226.19) [2/127.81]
| | | WD > = 80.5
| | | | Hum < 76
| | | | | Temp < 21: 56.88 (5/10.4) [3/119]
| | | | | Temp > = 21: 84.33 (2/90.25) [1/1482.25]
| | | | Hum > = 76: 71.3 (5/90.8) [5/479.8]
| | Temp > = 31: 110.14 (6/149.56) [1/711.11]
| WS > = 1.75: 50.66 (140/561.88) [69/710.04]
Size of the tree: 25
Time taken to build model: 0.03 seconds
=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.6702
Mean absolute error	18.9028
Root mean squared error	24.7308
Relative absolute error	73.4864 %
Root relative squared error	74.6384 %
Total Number of Instances	288

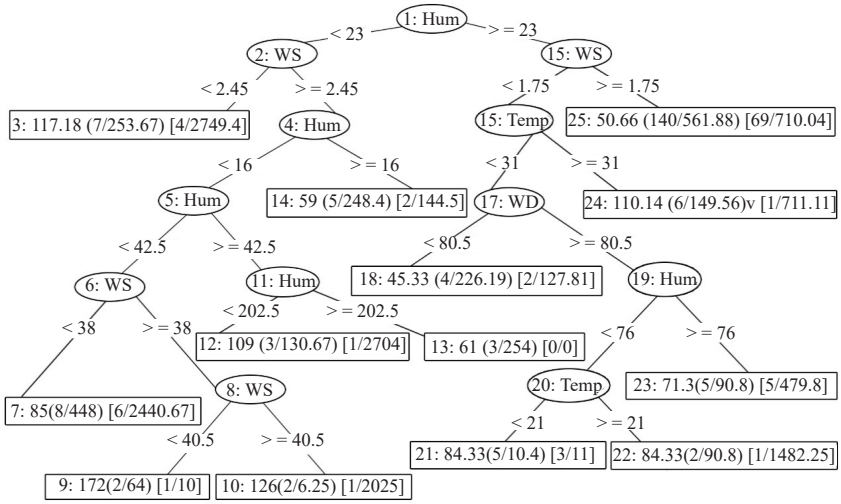

Figure 1. REPTree for predicting PM_{10} by using meteorological parameters

Table 2 presents the 13 predictive rules of IF-THEN. For example, the highest value of PM_{10} is in rule number 4, which stated that: if humidity is less than 23% and wind speed is greater than or equal to 2.45 m / s (8.82 km / h), humidity is less than 16%, temperature is less than 42.5°C and temperature greater than or equal to 38°C, temperature less than 40.5°C and then the predicated value of PM_{10} is 172ppm. It means when the temperature is between 38°C to 42°C with a wind speed of about 9 km / h and humidity less than 23%, the predicted value of PM_{10} is greater than 172ppm. The lowest value of PM_{10} is declared in rule number 10. Which clearly shows that if humidity is greater than or equal to 23%, wind speed is less than 1.75m / s (6.3km / h), temperature is less than 31°C, and wind direction is less than 80.5 degree then the predicated value of PM_{10} is 45.33ppm. This means when the humidity is higher than 23% with a wind speed less than 6.3km/ h and the temperature is less than 31°C, the PM_{10} concentration is 45.33, which is minimal.

Table 2. REPTree for predictive PM_{10} rules by using meteorological parameters

Rule No.	Predictive Rules
1	If Hum < 23 and WS < 2.45 then PM_{10}_next = 117.18
2	If Hum < 23 and WS > = 2.45 and Hum > = 16 then PM_{10}_next = 59
3	If Hum < 23 and WS > = 2.45 and Hum < 16 and Temp < 38 then PM_{10}_next = 85
(4)	If Hum < 23 and WS > = 2.45 and Hum < 16 and Temp < 42.5 and Temp > = 38 and Temp < 40.5 then PM_{10}_next = 172
5	If Hum < 23 and WS > = 2.45 and Hum < 16 and Temp < 42.5 and Temp > = 38 and Temp > = 40.5 then PM_{10}_next = 126
6	If Hum < 23 and WS > = 2.45 and Hum < 16 and Temp > = 42.5 and WD < 202.5 then PM_{10}_next = 109
7	If Hum < 23 and WS > = 2.45 and Hum < 16 and Temp > = 42.5 and WD > = 202.5 then PM_{10}_next = 61
8	If Hum > = 23 and WS > = 1.75 then PM_{10}_next = 50.66
9	If Hum > = 23 and WS < 1.75 and Temp > = 31 then PM_{10}_next = 110.14
(10)	If Hum > = 23 and WS < 1.75 and Temp < 31 and WD < 80.5 then PM_{10}_next = 45.33
11	If Hum > = 23 and WS < 1.75 and Temp < 31 and WD > = 80.5 and Hum > = 76 then PM_{10}_next = 71.3
12	If Hum > = 23 and WS < 1.75 and Temp < 31 and WD > = 80.5 and Hum < 76 and Temp < 21 then PM_{10}_next = 56.88
13	If Hum > = 23 and WS < 1.75 and Temp < 31 and WD > = 80.5 and Hum < 76 and Temp < 21 then PM_{10}_next = 84.33

Dragomir et al. in 2016 stated in their study that, the REPTree can be used to predict the next PM_{10} concentrations by using readings of the meteorological parameters: temperature, and relative humidity [18]. Furthermore, Moghadam and Ravanmehr in 2017 proved that the REPTree algorithm achieved the significant performance to knowledge discovery by the meteorological parameters [27].

Figure 2 shows the results of implementing M5P. The size of the tree is 31 with a number of nodes as 15 and 16 leaves. These leaves represent Machine Learning rules. The leaves of the tree represent the predictive rules of the tree. The process time of building this tree is 0.44 seconds. The correlation coefficient is 0.686, which indicates a strong correlation between PM_{10} and the four meteorological parameters. The Mean Absolute Error value is 18.055 and Root Mean Squared Error value is 24.0868. The following show the run information of this tree:

Instances: 288
Attributes: 5
PM_{10}
Temp
WD
WS
Hum
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
M5 pruned model tree:
(using smoothed linear models)
Hum < = 31.5:
| Temp < = 27: LM1 (13/33.983%)
| Temp > 27:
| | Temp < = 36.5: LM2 (25/29.467%)
| | Temp > 36.5:
| | | Hum < = 12.5: LM3 (11/64.558%)
| | | Hum > 12.5:
| | | | WD < = 217: LM4 (6/48.779%)
| | | | WD > 217: LM5 (6/35.59%)
Hum. > 31.5:
| WD < = 284.5:
| | Hum < = 88:
| | | Temp < = 27.5:
| | | | WD < = 192: LM6 (23/34.185%)
| | | | WD > 192:
| | | | | Hum < = 65.5: LM7 (11/22.824%)
| | | | | Hum > 65.5:
| | | | | | Temp < = 11.5: LM8 (4/15.46%)
| | | | | | Temp > 11.5: LM9 (3/15.059%)
| | | Temp >27.5: LM10 (29/52.945%)
| | Hum > 88:
| | | Temp < = 29.5:
| | | | WD < = 271.5: LM11 (16/36.703%)
| | | | WD > 271.5: LM12 (6/43.074%)
| | | Temp > 29.5:
| | | | WD < = 64: LM13 (6/27.66%)
| | | | WD > 64: LM14 (19/32.78%)
| WD > 284.5:
| | WS < = 4.85: LM15 (83/43.506%)
| | WS > 4.85: LM16 (27/47.155%)
LM num1: PM_{10} = 0.9299 * Temp + 0.0729 * WD - 6.9352 * WS - 1.1822 * Hum + 63.8746
LM num2: PM_{10} = 1.5851 * Temp + 0.0876 * WD - 12.783 * WS - 1.5856 * Hum + 88.1574
LM num3: PM_{10} = -2.7027 * Temp - 0.1714 * WD - 5.9489 * WS - 1.9865 * Hum + 300.1471
LM num4: PM_{10} = -0.8848 * Temp + 0.0225 * WD - 5.9489 * WS - 1.9532 * Hum + 173.5939
LM num5: PM_{10} = -1.3404 * Temp - 0.0261 * WD - 5.9489 * WS - 1.9532 * Hum + 206.0747
LM num6: PM_{10} = -0.3575 * Temp + 0.0289 * WD - 0.3785 * WS + 0.0029 * Hum + 59.5517
LM num7: PM_{10} = 0.6031 * Temp + 0.0314 * WD - 0.3785 * WS + 0.1114 * Hum + 41.2158
LM num8: PM_{10} = 1.312 * Temp + 0.0216 * WD - 0.3785 * WS + 0.2407 * Hum + 31.011
LM num9: PM_{10} = 1.3454 * Temp + 0.0314 * WD - 0.3785 * WS + 0.2407 * Hum + 28.5949
LM num10: PM_{10} = 0.2918 * Temp - 0.0302 * WD - 0.3785 * WS + 0.0292 * Hum + 69.808
LM num11: PM_{10} = -0.4128 * Temp + 0.051 * WD - 2.6123 * WS + 1.431 * Hum - 78.3223

LM num12: PM_{10} = -0.4128 * Temp + 0.073 * WD + 2.4368 * WS + 1.0277 * Hum - 51.4867
LM num13: PM_{10} = -1.396 * Temp + 0.0047 * WD - 1.2777 * WS - 0.1129 * Hum + 101.9175
LM num14: PM_{10} = -0.3787 * Temp + 0.0047 * WD - 1.2777 * WS - 0.1129 * Hum + 64.8528
LM num15: PM_{10} = 0.0417 * Temp - 0.0335 * WD - 0.854 * WS - 0.0421 * Hum + 64.371
LM num16: PM_{10} = 0.0417 * Temp - 0.8291 * WD - 1.473 * WS - 0.0421 * Hum + 299.4608
Number of Rules: 16
Time taken to build model: 0.44 seconds
=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.686
Mean absolute error	18.055
Root mean squared error	24.0868
Relative absolute error	70.1905 %
Root relative squared error	72.6947 %
Total Number of Instances	288

Where LM: Learning Machine.
num: number of the learning machine.

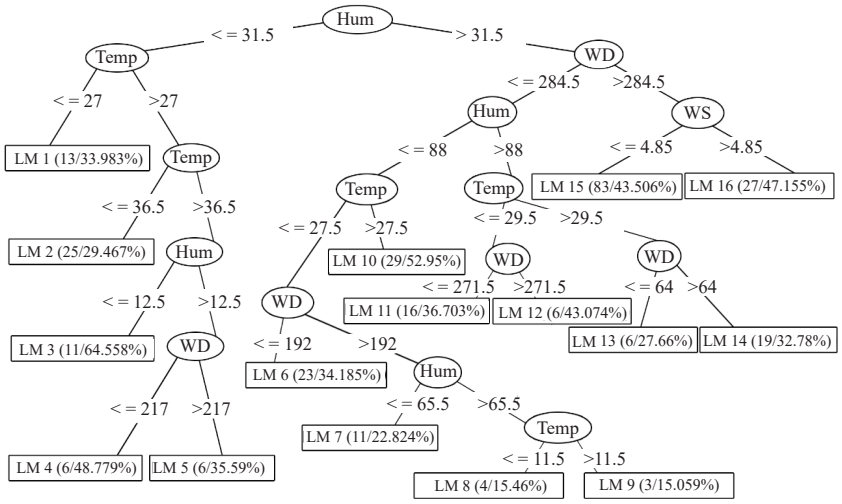

Figure 2. M5P for predicting PM_{10} by using meteorological parameters

Table 3 presents the 16 predictive rules of IF-THEN. For example, the highest value of PM_{10} is in the rule LM 3. which states that: if the humidity is less than or equal to 12.5% to less than or equal 13.5%, and temperature is greater than 27°C and greater than 36.5°C then the predicate value of PM_{10} is represented by the linear model LM3. This model is: PM_{10} = -2.7027 * Temp - 0.1714 * WD - 5.9489 * WS - 1.9865 * Hum + 300.1471. The wind speed has the highest coefficient followed by temperature, which means their effect were higher than the other parameters.

The lowest value of PM_{10} is declared in the rule ML 9. which states that: if humidity is greater than 31.5% and less than or equal to 88%, wind direction is greater than 192 degree (SW) to less than or equal 284.5 degree (NW) and the temperature is greater than 11.5°C less than or equal to 27.5°C then the predicate value of PM_{10} is represented in the linear model LM9: PM_{10}=1.3454 * Temp + 0.0314 * WD - 0.3785 * WS + 0.2407 * Hum + 28.5949. The temperature and wind speed have the highest coefficients, which indicate the main cause of increasing the concentration of PM_{10} is temperature and wind speed.

Dragomir et al. in 2016 used M5P to predict PM_{10} concentrations by using readings for meteorological temperature, and relative humidity [18]. Furthermore, Voukantsis et al. in 2010 in their study applied M5P to consider the effect of $PM_{2.5}$ and other gases on CO_2 [28].

Table 3. Testing M5P tree for predictive PM_{10} by using meteorological parameters

Rule No.	Predictive Rules
1	If Hum < = 13.5 and Temp < = 27 then LM1
2	If Hum < = 13.5 and Temp > 27 and Temp < = 36.5 then LM2
(3)	If Hum < = 13.5 and Temp > 27 and Temp > 36.5 and Hum < = 12.5 then LM3
4	If Hum < = 13.5 and Temp > 27 and Temp > 36.5 and Hum > 12.5 and WD < = 217 then LM4
5	If Hum < = 13.5 and Temp > 27 and Temp > 36.5 and Hum > 12.5 and WD > 217 then LM5
6	If Hum > 31.5 and WD < = 284.5 and Hum < = 88 and Temp < = 27.5 and WD < = 192 then LM6
7	If Hum > 31.5 and WD < = 284.5 and Hum < = 88 and Temp < = 27.5 and WD > 192 and Hum < = 65.5 then LM7
8	If Hum > 31.5 and WD < = 284.5 and Hum < = 88 and Temp < = 27.5 and WD > 192 and Hum > 65.5 and Temp < = 11.5 then LM8
(9)	If Hum > 31.5 and WD < = 284.5 and Hum < = 88 and Temp < = 27.5 and WD > 192 and Hum > 65.5 and Temp > 11.5 then LM9
10	If Hum > 31.5 and WD < = 284.5 and Hum < = 88 and Temp >27.5 then LM10
11	If Hum > 31.5 and WD < = 284.5 and Hum > 88 and Temp < = 29.5 and WD < = 271.5 then LM11
12	If Hum > 31.5 and WD < = 284.5 and Hum >88 and Temp < = 29.5 and WD > 271.5 then LM12
13	If Hum > 31.5 and WD < = 284.5 and Hum > 88 and Temp > 29.5 and WD < = 64 then LM13
14	If Hum > 31.5 and WD < = 284.5 and Hum > 88 and Temp > 29.5 and WD > 64 then LM14
15	If Hum > 13.5 and WD > 284.5 and WS < = 4.85 then LM15
16	If Hum > 13.5 and WD > 284.5 and WS > 4.85 then LM16

4. Research findings

Machine leaning by data mining played an important role in decision making prediction, especially the decision trees algorithms. Table 4 summaries the results of implementing the two decision trees algorithms, to predict the concentration value for PM_{10} by using meteorological parameters. The two decision tree algorithms used the same predictor meteorological parameters and data. As a finding for this paper, humidity is an important parameter related to PM_{10}, because humidity absorbs PM_{10}, therefore the concentration of PM_{10} reduced with increasing the relative humidity, and this is one of the innovative facts proved in this paper.

The M5P algorithm has the higher correlation coefficient value of 0.686, while for REPTree model value is 0.6702. The lowest MAE value is for M5P model with a value of 18.055, for REPTree model the value is 18.902. The lowest RMSE is for M5P with a value of 24.086, for REPTree the value is 24.730. The M5P has 31 rules, the REPTree has 25 rules. REPTree processing time is 0.03 seconds, and M5P consumed 0.44 seconds. This indicates that the M5P and the REPTree inductive models were suitable for predicting PM_{10} concentration. As a future study, we recommend develop an application that can send a notification for any environmental disasters.

Table 4. Comparison between REPTree and M5P decision trees

Statistical parameter	REPTree	M5P
Predictor	Humidity	Humidity
Correlation R	0.670	0.686
MAE	18.9028	18.055
RMSE	24.7308	24.0868
No. of rules	13	16
Size of the tree	25	31
Time taken to build tree	0.030	0.440
Predicted by	Value	Rule

Conflict of interest

The author declares there is no conflict of interest.

References

[1] Shumway, R. *One Solution for Air Pollution Big Data*. 2014. Available from: http://www.deseretnews.com/article/865617771/One-solution-for-air-pollution-Big-data.html.

[2] Alsultanny, Y. Selecting a suitable method of data mining for successful forecasting, *Journal of Targeting, Measurement and Analysis for Marketing*. 2011; 19(3/4): 207-225.
[3] Oracle. *Data Mining Concepts*. 2017. Available from: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON002.
[4] Shmueli, G., Bruce, C., Yahav, I., Patel, R., Lichtendahl, C. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. USA, New Jersey: John Wiley & Sons; 2017.
[5] Wang, L., Wang, G., Alexander, A. Big data and visualization: methods, challenges and technology progress. *Digital Technologies*. 2015; 1(1): 33-38.
[6] IEA. Oil 2017 Analysis and forecast to 2022, market report series. *International Energy Agency*. Available from: https://www.iea.org/Textbase/npsum/oil2017MRSsum.pdf [Accessed 2017].
[7] Gantz, J., Reinsel, D. Extracting value from Chaos. *IDC iview*. Available from: https://www.emcgrandprix.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf [Accessed 2011].
[8] Alsultanny, Y. Evaluating the effect of studying computer ethics and computer ethics rules and regulations on computer ethics at work. *Journal of Cloud Computing and Data Science*. 2020; 1(1): 21-30.
[9] Fayyad, U., Uthurusamy, R. Evolving data into mining solutions for insights. *Communications of the ACM*. 2002; 45(8): 28-31.
[10] Larose, D. *Discovering Knowledge in Data an Introduction to Data Mining*. John Wiley & Sons, Inc., Publication; 2005.
[11] Wu, X., Kumar, V., Quinlan, R., Ghosh, J., Yang, Q., Motoda, H., Philip, Y. Top 10 algorithms in data mining. *Knowledge and Information Systems*. 2008; 14(1): 1-37.
[12] Han, J., Kamber, M. *Data Mining Concepts and Techniques 3rd Edition*. Amsterdam, Netherlands: Elsevier; 2011.
[13] Srinivasan, B., Mekala, P. Mining social networking data for classification using REPTree. *International Journal of Advance Research in Computer Science and Management Studies*. 2014; 2(10): 155-160.
[14] Witten, H., Frank, E., Hall, A., Pal, J. *Data Mining: Practical Machine Learning Tools and Techniques 4th Edition*. Amsterdam, Netherlands: Elsevier; 2016.
[15] www.quinnipiac.edu. *Pearson's r Correlation (Modified from Instructor's Resource Guide for the Text)*. Available from: http://faculty. Quinnipiac.edu/libarts/polsci/Statistics.html.
[16] Willmott, C., Matsuura, K. Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in assessing average model performance. *Climate Research JSTOR*. 2005; 30(1): 79-82.
[17] Nagar, A. A Comparative study of data mining algorithms for decision tree approaches using WEKA tool. *American-Eurasian Network for Scientific Information*. 2017; 11(9): 230-241.
[18] Dragomir, E., Oprea, M., Popescu, M., Mihalache, S. Particulate matter air pollutants forecasting using inductive learning approach. *Revista de Chimie*. 2016; 67(10): 2075-2081.
[19] Vitkar, S. Comparative analysis of various data mining prediction algorithms demonstrated using air pollution data of Navi Mumbai. *Research Journal of Chemical and Environmental Sciences*. 2017; 5(1): 79-85.
[20] Chi, H. *Improving M5 model tree by evolutionary algorithm*. MSc thesis, Department of Computer Science, Ostfold University, Norway. 2015.
[21] Csépe, Z., Makra, L., Voukantsis, D., Matyasovszky, I., Tusnády, G., Karatzas, K., Thibaudon, M. Predicting daily ragweed pollen concentrations using computational intelligence techniques over two heavily polluted areas in Europe. *Science of the Total Environment*. 2014; 476(1): 542-552.
[22] Weka. Available from: http://www.cs.waikato.ac.nz/ml/weka [Accessed 2020].
[23] Dean, J. *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. John Wiley & Sons, Inc., Hoboken, New Jersey; 2014.
[24] Han, J., Kamber, M., Pei, J. *Data Mining: Concepts and Techniques.* 3rd ed. Morgan Kaufmann; 2012.
[25] Alsultanny, Y. Data mining and visualization: meteorological parameters and gas concentration use case. *19th International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID)*. Russia: Moscow State University; 2017. p.10-13.
[26] Juncal-Martínez, J., Alvarez-López, T., Gavilanes, F., Costa-Montenegro, E., González-Castano, J. GTI at SemEval-2016 Task 4: Training a Naive Bayes Classifier using Features of an Unsupervised System. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. USA: California; 2016. p.115-119.
[27] Moghadam, A., Ravanmehr, R. Multi-agent distributed data mining approach for classifying meteorology data: case study on iran's synoptic weather stations. *International Journal of Environmental Science and Technology*. 2017; 1(1): 1-10.
[28] Voukantsis, D., Niska, H., Karatzas, K., Riga, M., Damialis, A., Vokou, D. Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece. *Atmospheric Environment*. 2010; 44(39): 5101-5111.