WASTE WATER TREATMENT - AN APPLICATION OF PREDICTIVE DATA MINING

[1]P.Pandi selvi
Assistant Professor
Department of Computer Science
Dr.Umayal Ramanathan College for Women, Karaikudi.

Abstract: In order to predict future results in a better way, predictive data mining techniques are used. In this paper, the predictive analytic approach is used in waste water treatment. The parameter values were calculated first. After computing, if the range value of these parameters were at its desired level, relevant to the predictive value, the water was in its purified form. It can be predicted to be used in the future. Waste water samples were collected from industries, and sample dataset was first created containing 500 records. Preprocessing operation is then carried out to remove any unwanted information in the data set. In order to perform classification between the predictive data and the current data three models were utilized. They were Linear regression model, multilayer perceptron model and SMOreg model. The results of the three models were then compared. With the results, it is evident that linear regression and multilayer perceptron models prove to be better in predicting the data.

Index Terms : Predictive data mining, preprocessing, Linear regression model, Waste water treatment.

I.INTRODUCTION

Data mining is the process of finding relevant data in a heterogeneous data repository. In this, predictive analytic approach is nothing but the use of previous samples to predict or forecast results [14]. They just compare past successes and failures and use those results to predict future outcomes. Most probably, predictive models were used in business to assess the risk with certain conditions in order to make decision making [15]. Linear regression model is the most commonly used model for predictive analysis. They are mainly used in examining, whether the specified set of attributes, were efficient in predicting the future outcome in a better way [12]. Linear regression models find the statistical relationship between predictive and dependent attributes.

With the advent of various technologies in the present day world, many industries were established to improve the status. This creates various side effects in the environment. The water from these industries was very much polluted and must undergo various treatment processes to get purified. In this paper, predictive analysis is used in the application of waste water treatment [15]. It is the process of converting water that is of no use, back into the environment for any useful purpose [10]. Water is the Elixir of life. It is the sole responsibility of each and every individual to save water. At the same time, waste water can be recycled and transferred to the environment for useful purposes.

II. LITERATURE REVIEW

In 2012, Carlos Marquez-Vera et al [1], proposed a genetic programming algorithm for solving the challenges due to number of factors that affect the low performance of students and the imbalanced nature. Various methods involved are, Data Gathering, Pre-Processing, Data Mining, and Interpretation. Interpretable Classification Rule Mining (ICRM) and SMOTE (Synthetic Minority Over-sampling Technique) algorithms are used. As a first step they collected student's data set. WEKA tool is used by them for implementing their work. Accuracy, True positive rate, True negative rate and Geometric mean were the parameters used by them for performance measurement. Their experimental results proved to be accurate and it achieved the best predictions of student failure (98.7 %).

In 2007, M.Dixon et al [2], specified the way data mining is used in the anaerobic waste water treatment process. They presented their future direction of work in four ways; their experience in data mining area, the use of confidence and prediction intervals, generalization over different sizes and types of anaerobic digester and the relationship to overall supervision.

In 2014, Manel Poch et al [6], presented a review in waste water treatment. They analyzed some of the main tools that are applied to obtain information and knowledge from raw data. The authors were trying to manage two important specific problems in their work. They were, sludge bulking and greenhouse gas emissions.

In 2016, Festim halili et al [4], proposed a predictive modeling using data mining regression technique applied in a prototype. In this paper, they have applied the regression model in their prototype and analyzed its performance. As a future work, they are about to do other analysis with predictive models.

In 2016, V.Kavya et al [5], made a review on predictive analytics in data mining. In their study, they mainly focused towards, predictive analytics, regression techniques and forecasting. With this survey they analyzed that, the predictive analytic approach is more efficient in marketing and other social media.

In 2018, Shakuntala Jatav et al [8], proposed an algorithm for predictive data mining approach in medical diagnosis. In this paper, they analyzed the prediction systems for diabetes, kidney and liver disease. They used a combination of two classification techniques namely, support vector machine and random forest. The performances of the techniques were compared

based on precision, recall, accuracy, f_measure and time. The experimental results show that the accuracy was in the range of 99.35%, 99.37% and 99.14%.

In 2016, S.B.Soumya et al [9], described a data mining system with predictive analytics for financial applications. Their basic idea is to apply patterns on available data and generate new assumptions and behavior using predictive analysis. It can be applied in various application areas like, surveillance and warning systems, predicting abnormal stock market returns, corporate bankruptcies, financial distress, management fraud. In financial services, the approach is used to segment customers and predict cross-selling promotions. They classify customers, who respond to offers for additional products and services.

In 2017, Fatimetou Zahra et al [3], made a study on the different application areas of predictive analytics and how it is used to solve various problems in industries. On looking into the benefits part, it reduced and prevented risk, saved time, cost and management of resources. The challenges include, get real, sufficient and clean data, which were developed to test the models. Weakness in the research area includes, focus on the development of models only, the wrong choice of models variables and algorithms affecting the results of predictions.

In 2015, Sakshi rungta et al [7], presented a system to analyze user stories incorporating the data of energy and health demands of four countries for the past 30 years and finally to predict future trend of the parameters. The correlations between the entities were found using pearson's coefficient. They have predicted the emerging trends in the form of power view charts. The future direction for improving the user in the loop workflow for predictive analytics was also presented.

III.PROPOSED METHOD

The various steps involved in the proposed method is as follows,

(i) Data Collection.

(ii) Data cleaning.

(iii) Classification- Regression analysis, Multilayer Perceptron, SMOreg

The structure of work flow is as shown in the below figure 1.

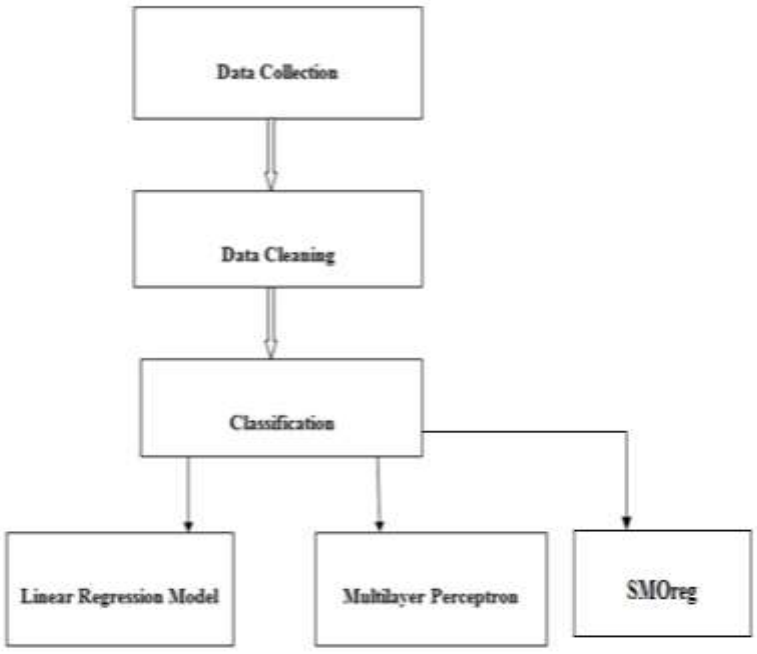

Figure: 1 Data Flow Diagram.

(i) Data Collection: This is the first step in any data mining process. Sample datasets were collected from industries and a sample database was created containing 500 records. The following figure 2 shows a sample database used for the research.

Month	pH	Color	Odour	Temp	Ec	TDS	BOD	COD	Total Hardness	Chloride	Na	K	Ca	Mg	Cu	Ni	Zn	Fe
17-Jan	8.24	Yellow	Pungent	35	1400	3390	345	780	1083	620	120	90	160	80	1.2	1.98	6.28	2.29
17-Feb	8	Yellow	Pungent	35	1000	2600	300	680	1240	750	180	80	180	90	1.62	2.5	7.2	2.3
17-Mar	7.6	Brown	Pungent	36	900	2900	350	640	950	820	160	85	150	100	1.8	2.6	7.1	2.5
17-Apr	7.9	Brown	Pungent	38	1100	3200	280	790	825	600	130	75	120	110	1.6	2.4	7.1	2.4
17-May	7.4	Brown	Pungent	40	1400	3800	260	720	900	540	150	100	110	70	2.1	2.1	6.28	2.7
17-Jun	7.2	Brown	Pungent	39	1400	3500	300	460	750	580	120	110	140	40	2.5	2	6.4	3.1
17-Jul	7.6	Reddish	Pungent	36	1000	3400	290	430	850	510	100	50	130	60	2.6	2.5	6.9	2.7
17-Aug	7.5	Reddish	Pungent	36	1100	3300	250	480	940	810	90	60	100	60	2.1	1.9	6.5	2.1
17-Sep	7.9	Reddish	Pungent	35	1200	2600	240	390	650	510	110	60	80	40	2	1.7	6.3	2.3
17-Oct	8.2	Brown	Pungent	35	900	2400	260	350	725	480	120	55	210	30	1.9	1.5	6.7	2.6
17-Nov	8.1	Brown	Pungent	34	800	2300	280	370	840	390	140	65	180	50	1.9	1.6	6.1	2.4
17-Dec	7.7	Yellow	Pungent	34	1000	2000	220	340	800	420	190	75	130	40	1.8	1.8	6	2.8
18-Jan	7.3	Yellow	Pungent	34	1200	1800	200	310	510	400	170	70	125	40	1.4	1.3	6	2.1

Figure 2. Sample Dataset.

(ii) Data Cleaning: Cleaning process is carried out to remove any unwanted information in the database.

(iii) Classification: In this case, the results of three different classifiers were compared. They were linear regression, multilayer perceptron and SMOreg. The sample database was fed as input into the system. The system was first trained with the given dataset.

The maximum, minimum, mean and standard deviation of, COD and BOD parameters in the water sample is as follows,

The values of BOD,

Statistic	Value
Minimum	200
Maximum	350
Mean	275
StdDev	43.78

The values of COD,

Statistic	Value
Minimum	310
Maximum	790
Mean	518.462
StdDev	177.616

If the range of the values were at its desired level, the water will be in its purified level and can be predicted to be used in the environment. Otherwise, they have to undergo further treatment for purification.

IV.RESULTS AND DISCUSSION

In the proposed method, classification process is carried out with three different classifiers and the results are compared. In the linear regression model, it just models the relationship between the variables by fitting linear equation to the observed data. Among the variables one is an explanatory variable and the other is a dependent variable [13]. The result obtained from a linear regression model is as shown in figure 3.

4.1 Results of Linear Regression Model

```
=== Run information ===
Relation:     Dataset
Instances:    13
Attributes:   19
Test mode:    evaluate on training data
=== Classifier model (full training set) ===
Linear Regression Model
Time taken to build model: 0.02 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correlation coefficient                  1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               13
```

Figure: 3 The output of a linear regression model

4.2 Results of Multilayer Perceptron

The multilayer perceptron network [17] follows a supervised learning technique for training. They were a class of feed forward artificial neural network in which, they were organized into layers. These networks have the capability to learn and generalize from previous samples to new solutions. Apart from this, they have the capability to extract needed information from various inputs containing irrelevant data. Their functionality is similar to that of a human brain. Just as a human being could recognize past events, the network could produce the desired output that is learned from training.

```
=== Run information ===
Relation:     Dataset    Instances:    13        Attributes:   19
Test mode:    evaluate on training data
=== Classifier model (full training set) ===
Linear Node 0
    Inputs    Weights
    Threshold    0.036646157935600304
    Node 1    0.04410965369794418                Node 9    -0.3290319313209849
    Node 2    -0.7184156484665677                Node 10    0.554224144680224
    Node 3    0.20078912935232962                Node 11    -0.6136349990351804
    Node 4    0.04037093940985022                Node 12    0.1817643321045038
    Node 5    -0.263748188158538                 Node 13    -0.39730175215394864
    Node 6    0.2610068648432821                 Node 14    -0.4708124943327938
    Node 7    -0.2181345488053821                Node 15    0.8759615397895395
    Node 8    0.384137026794254                  Node 16    -0.04690883950936131
Sigmoid Node 1
    Inputs    Weights                Threshold    -0.12347056975507993
Time taken to build model: 0.19 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correlation coefficient                  1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               13
```

Figure: 4 The output of Multilayer Perceptron

4.3 Results of SMOreg

The output of the SMOreg regression model is as shown below.

```
=== Run information ===
Test mode:    evaluate on training data
=== Classifier model (full training set) ===
SMOreg
Number of kernel evaluations: 91 (91.495% cached)
Time taken to build model: 0.02 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correlation coefficient          1
Mean absolute error              0.0011
Root mean squared error          0.0012
Relative absolute error          0.4837 %
Root relative squared error      0.4404 %
Total Number of Instances        13
```

Figure: 5 The output of SMOreg

CONCLUSION

From the results, it is well evident that among all the three methods, linear regression model and multilayer perceptron proved to achieve better results in predicting the data. Hence the range of values were identified and predicted from the whole dataset. It is well evident from the output based on error value, that the parameters were at its predicted desired level and the treated water can be used in the environment. As a future work, a dataset was about to be created with more than thousands of records that can be utilized for research purpose.

REFERENCES

[1] Carlos Marquez-Vera, Alberto Cano, Cristobal Romero, Sebastian Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data". Springer Science Business Media, LLC 2012.

[2] M.Dixon, J.R.Gallop, S.C.Lambert, J.V.Healy, "Experience with Data Mining for the Anaerobic Wastewater Treatment Process". Environmental Modelling and Software 22 (2007) 315-322.

[3] FatimetouZahra Mohamed Mahmoud, "The Application of Predictive Analutics: Benefits, Challenges and How it can be Improved". International Journal of Scientific and Research Publications, Volume 7, Issue 5, May 2017. ISSN: 2250-3153.

[4] Festim Halili, Avni Rustemi, "Predictive Modeling: Data Mining Regression Technique Applied in a Prototype". International Journal of Computer Science and Mobile Computing, Vol.5 Issue 8, August -2016, Pg:207-215, ISSN: 2320-088x. www.ijcsmc.com.

[5] V.Kavya, S.Arumugam, "A Review on Predictive Analytics in Data Mining". International Journal of Chaos, Control, Modelling and Simulation (IJCCMS) Vol.5, No.1/2/3, September 2016.

[6] Manel Poch, Joaquim Comas,Jose Porro, Manel Garrido-Baserba Lluis Corominas, Maite Pijuan, "Where are we in Wastewater Treatment Plants Data Management? A Review and a Proposal". International Environmental Modelling and Software Society (IEMSS), 7th Intl Congress on Env. Modelling and Software, San Diego, CA, USA, Daniel P.Ames, Nigel W.T.Quinn and Andrea E. Rizzoli(Eds). https://www.iemss.org/society/index.php/iemss-2014-proceedings.

[7] Sakshi Rungta, Vanita Jain, Akanksha Utreja, "Data Mining Engine using Predictive Analytics". International Journal of Computer Applications (0975 – 8887). Volume 121 – No.5, July 2015.

[8] Shakuntala Jatav, Vivek Sharma, "An Algorithm for Predictive Data Mining Approach in Medical Diagnosis". International Journal of Computer Science and Information Technology (IJCSIT) Vol 10, No 1, February 2018.

[9] S.B.Soumya, N.Deepika, "Data Mining With Predictive Analytics for Financial Applications". International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-1, January 2016. ISSN:2395-3470. www.ijseas.com.

[10] https://www.conserve-energy-future.com/process-of-wastewater-treatment.php

[11] http://www.marketingprofs.com/articles/2010/3567/the-nine-most-common-data-mining-techniques-used-in-predictive-analytics

[12] http://www.statisticssolutions.com/what-is-linear-regression/

[13] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[14] https://www.techopedia.com/definition/30597/predictive-data-mining

[15] https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86

[16] https://en.wikipedia.org/wiki/Linear_regression

[17] https://en.wikipedia.org/wiki/Multilayer_perceptron