International Journal of Engineering & Technology, 7 (2.32) (2018) 550-553

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A new approach for wastewater treatment using predictive data mining - A comparative study

P. Pandi Selvi *

Assistant Professor, Department of Computer Science, Dr. Umayal Ramanathan College for Women, Karaikudi, Tamil Nadu. India
**Corresponding author E-mail: selvim11215@gmail.com*

Abstract

In the field of science data mining plays a major role in solving complex real world problems. The proposed method uses the predictive approach to determine the quality of water. To carry out the work, waste water samples were collected from textile industries and a dataset was created. Initially, preprocessing of the sample dataset was carried out. Classification is performed with, Random forest and Random Trees. Mean square error and the mean absolute error values were computed and the results are tabulated. Based on this, decision can be made regarding the recycling of the treated water. With the result it is well evident that the proposed method is able to predict the quality in a better way.

Keywords: *Predictive Data Mining; Preprocessing; Random Forest; Random Trees; Wastewater Treatment.*

1. Introduction

Water plays an important role in our day to day life. It is the elixir of our life. Without water no one can live. On considering the importance of water it must be preserved for future generations. In recent days, there is a scarcity of water in most of the areas; various measures have to be taken to recycle even the waste water. Waste waters from textile industries, machineries, home, food scraps, chemicals, washing machines, dish water, some business places can be recycled and transferred again to the environment for useful purpose. Accumulation of waste water and effluents may create a bad impact on human beings as well as the surroundings. The process of recycling is helpful to living and non – living organisms [20]. Before transmitting the contaminated water to the environment, it has to be recycled.

In this broad area of waste water treatment, data mining plays an important role in the prediction of water quality. Various models are available such as, artificial neural network, Support vector machines, K-nearest neighbor, trees, ANFIS, etc.

The technique of data mining is mainly used to discover knowledgeable data from data repository. Among various types of mining techniques predictive data mining is used in majority of real world application problems [18]. Predictive approach compares past successes and failures and uses those results to predict future outcomes [19]. In the proposed method classification process is carried out with Random Forest and Random Trees.

2. Literature review

In 2017, Bartosz Szeląg, et al [1] ., developed a new model using data mining methods for predicting waste water quality indicators at the inflow to the treatment plant. The models were developed using multivariate adaptive regression spline method, artificial neural network combined with the SOM classification model and the cascade neural network. With their approach, the lowest value of absolute and relative errors was obtained using artificial neural network and the SOM model. Higher error values were obtained through multivariate adaptive regression spline method. The results obtained prove the efficiency of their method.

In 2017, Bharat B. Gulyani and Arshia Fathima[2], introduced a bagging model, to predict the performance of WWTP. Ensemble models used stabilize the base classifier and it avoided over fitting of the data. They used bagging to predict the performance of individual units and the global plant performance. The predicted performance of individual units computed was also used as inputs to predict the global performance. Upon application of their model to the WWTP dataset, it performed better than ANN or SVM for the prediction.

In 2012, Carlos Marquez-Vera et al [3], proposed a genetic programming algorithm for solving the challenges due to the number of factors that affect the low performance of students and their 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 were 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 2017, Corominas Li, et al [4]., described various computer based techniques for the analysis to improve waste water treatment. In their analysis most frequently used techniques were identified and some are artificial neural network, Principal component analysis, fuzzy logic, clustering, independent component analysis and partial least squares regression. On the other side, they also discussed about the limitations of these methods.

In 2013, Daniel Ribeiro, et al [5]., presented an approach to predict the performance of waste water treatment plant located in northern Portugal. They used support vector machine method for prediction. They have considered two parameters Biochemical Oxygen

Demand and the Total Suspended Solids for their work. They also described the dataset created by them, for their work and the results were tabulated.

In 2008, Davut Hanbay, et al [6]., developed a waste water treatment plant model to predict its performance. Their model is based on wavelet packet decomposition, entropy and neural network. They retrieved datasets for their work from WWTP in Malatya, Turkey. They used wavelet packets and NN for feature extraction and classification in intelligent modeling. With the results they proved their ability to design a new intelligence model.

In 2015, Djeddou. M,Achour. B [8], developed ANNs, for the prediction of sludge volume index using influent quality parameters (TSS (mg/L), COD (mg/L), BOD (mg/L), temperature (°C), pH, conductivity (μS/cm), NH+4 (mg/L), NO3 (mg/L), P (mg/L), and operating parameters, TSS efficiency (%), COD efficiency (%), BOD efficiency (%), Ammonia efficiency (N-NH+4) (%), Nitrate efficiency (N-NO3) (%), Phosphorus efficiency (P-PO4). They have selected Batna wastewater treatment plant, from the time period 2011-2014 for their analysis. The training of their neural networks was achieved by Levenberg-Marquardt algorithm. The results obtained by their method were satisfactory. Consequently, their models can also be used for the prediction of SVI. Their model for SVI prediction is composed of one input layer with fifteen input variables, one hidden layer with thirteen nodes and one output layer with one output variable with R= 0.8784, MAE = 0.186, RMSE = 0.443 and MAPE = 10.98%. The modeling approaches used in their study had better prediction power.

In 2017, Fatimetou Zahra et al [9], 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. The weakness in their research is that, they focus on the development of models only, the wrong choice of models variables and algorithms affecting the results of predictions.

In 2016, Festim halili et al [10], proposed a predictive model 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 different predictive models.

In 2018, P.Pandi selvi [13], proposed a predictive analytic approach in the area of waste water treatment. Dataset was first created with about 500 records. In order to carry out the classification process three models were used and are compared. The models were Linear regression model, Multilayer perceptron and SMOreg model. The results obtained was tabulated and the efficiency of each method was proved.

In 2015, Sakshi rungta et al [14], 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 by them.

In 2018, Shakuntala Jatav et al [15], proposed an algorithm for predictive data mining approach in medical diagnosis. In this paper, they analyzed the prediction system for diabetes, kidney and liver disease. They used a combination of two classification techniques namely, support vector machine and random forest. The performance 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 [16], 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, etc. In financial services, their approach is used to segment customers and predict cross-selling promotions. They classified customers, who respond to offers for additional products and services.

In 2018, C. Victoria Priscilla, A. Anusuya [17], made a survey of various techniques like Data Mining, Machine learning, Principal Component Analysis, Support Vector Machine (SVM), Regression Trees (RT) and provided the estimation of the wastewater quality characteristics. They analysed Wastewater Treatment Plant using various techniques at different locations. Their study shows that when the waste water is recycled, their quantity is reduced in urban area. Their future plan is to analyse the data set from UCI Machine Learning Repository and implement the water quality indicator to evaluate various data mining algorithms.

3. Proposed method

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

1) Data Collection.
2) Data cleaning.
3) Classification- Random Forest, Random Trees.

The structure of the proposed method is as shown in the below figure 1[13].

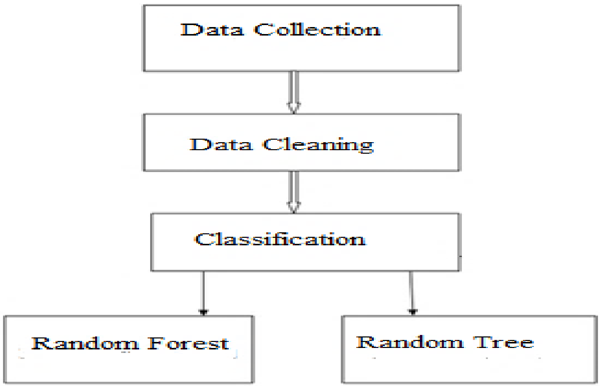
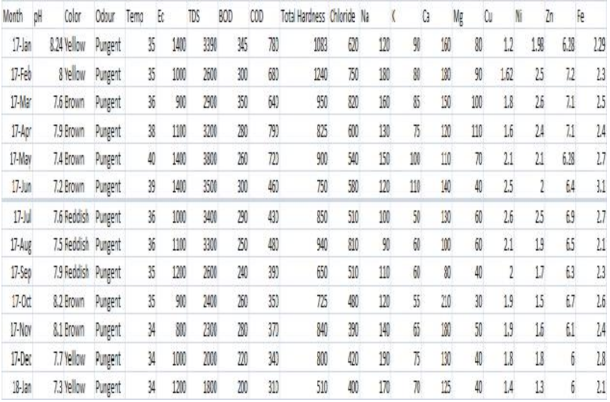


Fig. 1: Data Flow Diagram.



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

Fig. 2: Sample Dataset.

1) Data Collection

In order to carry out the mining process, waste water samples were collected from industries. A sample database was created containing 500 records. The following figure 2 [13]shows a sample dataset.

2) Data Cleaning

The dataset that is used for the mining process must be free from noise. Hence cleaning is carried out to remove any unwanted information in the dataset.

3) Classification

The process is carried out with Random forest and Random trees. Random forest is a supervised learning algorithm. It builds multiple decision trees and finally merges them together to obtain precise and firm prediction. In most of the machine learning systems, it is used for solving both classification and regression problems. The algorithm searches for the best feature from a random subset of features. It easily measures the importance of each feature on prediction. It is a flexible tool to carry out the predictive task [21]. Random trees were nothing but the decision trees built on a random subset of columns.

The proposed method is implemented in weka tool. From the sample dataset, the parameters considered from the waste water samples were as follows, maximum, minimum, mean and standard deviation of, COD and BOD. Among them COD and BOD were the two important parameters in determining the quality of water. Their values were as in Table 1 and Table 2[13]. The random forest for the given sample dataset is as shown in figure 3

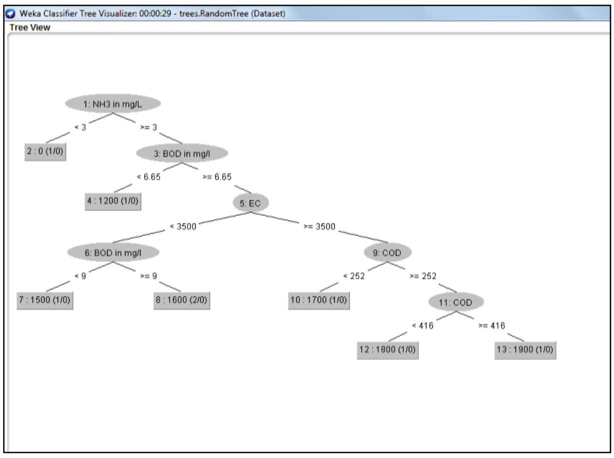
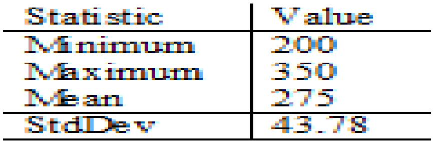


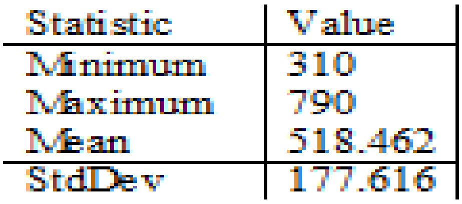
Fig. 3: Random Forest.

Table 1: The values of BOD



Statistic	Value
Minimum	200
Maximum	350
Mean	275
StdDev	43.78

Table 2: The Values of COD



Statistic	Value
Minimum	310
Maximum	790
Mean	518.462
StdDev	177.616

The range values for the parameters COD and BOD was identified first. In the treated sample, if the range of the values were at its desired level, the water is considered to be in its purified form. The recycled water can then be fed into the environment. Otherwise, the water has to undergo further treatment for purification.

In the proposed method, classification process is carried out with Random Forest and Random trees and their results are tabulated. The results of mean absolute error and mean square error obtained from the models are as shown in Table 3.

Table 3: The Values of Mean Absolute Error and Mean Square Error

Algorithm	MAE	MSE	Correlation Coefficient
Random forest	3.1	19.98	0.92
Random tree	1.6	10.99	0.98

4. Conclusion

The proposed method used random forest and random trees to predict the dissolved effluents in waste water. The range of BOD and COD values were identified from the sample dataset. Based on the level obtained from the treated water, the range is identified. Once the treated water was in its desired range, it can be fed into the environment for use. The mean square error and the mean absolute error values were computed and tabulated. The output indicates that the approach was able to predict the effluent parameters in a better way. It means that the effluent in water was reduced and it is at its desirable level, hence it can be used in the environment.

References

[1] Bartosz Szeląg, Krzysztof Barbusiński, Jan Studziński, Lidia Bartkiewicz,” Prediction of wastewater quality indicators at the inflow to the wastewater treatment plant using data mining methods”. E3S Web of Conferences 22, 00174 (2017) https://doi.org/10.1051/e3sconf/20172200174.

[2] Bharat B. Gulyani and Arshia Fathima,” Introducing Ensemble Methods to Predict the Performance of Waste Water Treatment Plants (WWTP)”. International Journal of Environmental Science and Development, Vol. 8, No. 7, July 2017 https://doi.org/10.18178/ijesd.2017.8.7.1004.

[3] 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. https://doi.org/10.1007/s10489-012-0374-8.

[4] Corominas Li, Garrido-Baserba.M, Olsson G, Cortes.U, Poch.M,” Transforming data into knowledge for improved waste water treatment operation: A critical review of techniques”. Environmental Modelling and Software, Elsevier. Scopus:85044678780. DOI: 10.1016/j.envsoft.2017.11.023. Publishing date: 2017-12-08. https://doi.org/10.1016/j.envsoft.2017.11.023.

[5] Daniel Ribeiro, Antonio Sanfins, Orlando Belo,”Wastewater treatment plant performance prediction with Support Vector Machines”. ICDM’13 Proceedings of the 13th International conference on Advances in Data Mining: Applications and theoretical aspects, pages 99-111. Springer-verlag Berlin, Heidelberg 2013. ISBN: 978-3-642-39735-6, https://doi.org/10.1007/978-3-642-39736-3_8.

[6] Davut Hanbay , Ibrahim Turkoglu , Yakup Demir, “Prediction of wastewater treatment plant performance based on wavelet packet decomposition and neural networks”. Expert Systems with Applications 34 (2008) 1038–1043. Available online at www.sciencedirect.com. https://doi.org/10.1016/j.eswa.2006.10.030.

[7] 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. https://doi.org/10.1016/j.envsoft.2005.07.031.

[8] Djeddou. M,Achour. B,” The Use Of A Neural Network Technique For The Prediction Of Sludge Volume Index In Municipal Wastewater Treatment Plant “Larhyss Journal, ISSN 1112-3680, 24, Décember 2015, pp. 351-370 © 2015 All rights reserved, Legal Deposit 1266-2002.

[9] 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.

[10] 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. https://doi.org/10.5121/ijccms.2016.5301.

[11] 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.

[12] 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.

[13] P.Pandi Selvi,” Waste water treatment – An application of Predictive data mining”. Journal of Emerging Technologies and Innovative Research. Volume 5, Issue 9, September 2018, ISSN: 2349-5162. www.jetir.org.

[14] 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. https://doi.org/10.5120/21537-4545.

[15] 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. https://doi.org/10.5121/ijcsit.2018.10102.

[16] 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.

[17] C. Victoria Priscilla, A. Anusuya," A Survey On Wastewater Treatment (Wwt) Analysis Using Various Techniques". International Journal of Advanced Research in Computer Science (ISSN: 0976-5697). Volume 9, Special Issue No. 1, 2018. www.ijarcs.info.

[18] http://www.statsoft.com/textbook/data-mining-techniques.

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

[20] https://water.usgs.gov/edu/wuww.html.

[21] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd.