Using random forest to estimate the efficiency of an aerated lagoon at a pulp and paper mill

Lucas M. Almeida*, Brenner B. Silva, Karla O. Esquerre
Polytechnic School
Federal University of Bahia
Prof. Aristídes Novis St, 2 - Federação, Salvador - BA, Brazil
(lucasmascalmeida,brennerbiasiss,karla.esquerre)@gmail.com

May 25, 2019

Abstract

The treatment and discharge process of effluent at an industry is a process very important to attend the environmental requirements. The mainly aim of wastewater treatment systems is reduce the amount of pollutants that the effluent has for the wastewater be able to discharge adequately. Legally at Brazil, by the resolution 430/2011 of Environmental National Council (Conselho Nacional do Meio Ambiente - Conama), the effluents can be discharge after the correct treatment that follow some conditions, one condition is the charge of biochemical oxygen demand (BOD) should be removed at least 60% to totally charge. The advance of modern industry generated larger volumes of data and this fact provided a necessity for understand what information these data carries, a good solution for attend this demand are the empirical models. In this work was made the BOD and chemical oxygen demand (COD) prediction by the Random Forest method. The final product was the effluent classification for final discharge and the estimate the efficiency of an aerated lagoon at a pulp and paper mill in relation the BOD remove.

1 Introduction

The treatment and discharge process of effluent at an industry is a process very important to attend the environmental requirements.The mainly aim of wastewater treatment systems is reduce the amount of pollutants that the effluent has for the wastewater be able to discharge adequately [1].

Legally at Brazil, by the resolution 430/2011 of Environmental National Council (Conselho Nacional do Meio Ambiente - Conama), the effluents can be discharge after the correct treatment that follow some conditions, one condition

*

is the charge of biochemical oxygen demand (BOD) should be removed at least 60% to totally charge [2].

Beyond the environmental requirements, the changes on market also can create important tech changes in pulp and paper mill industry [3]. The advance of modern industry generated larger volumes of data and this fact provided a necessity for understand what information these data carries, a good solution for attend this demand are the empirical models [4].

Of the many process that can be supervised in an Wastewater Treatment Plant (WWTP), the effluent parameters are very important because they translate the principal proprieties of effluent [5]. The principal effluent parameters of pulp and paper mill industry are BOD and chemical oxygen demand (COD) [1]. However, the obtain of these parameters can consume much time, mainly for BOD that demand five days per analyze [6].

In history many models were created for predicted of theses parameters (BOD and COD), some of theses models were made with artificial neural networks [1], because there is a notability complexity in bio process modeling [7]. So, in a searching for solutions less hard and with results betters, in the last years, the data mining and machine learning methods are growing and they are incorporating various sectors of engineer [8][9].

The aim of this work is applier the Random Forest method to prediction of treatment system parameters BOD_{in} and BOD_{out}, in order to promote the classification for final discharge and the estimate the efficiency of an aerated lagoon at a pulp and paper mill in relation the BOD remove.

2 Materials and Methods

2.1 Data set

This work was made on base in information of monitoring pulp and paper mill industry data set. The data are referent the window time from September 1998 to July 2000, with the total of 1537 observation and 23 variables.

The data are diary level, however the $\mathrm{BOD}_{5\,days}$ need to a period of five days for incubation and only after that can be analyse.

The Table 1 shows the variables descriptions that are in this work. Of the statistics showed in Table 1 the NA (%) is a highlight, for values more than 35% the variable was excluded of developed model.

2.1.1 R software

The developed modeling was made by the R software. The mainly packages are caret, DMwR, e1071 and the tidyverse family.

2.2 Random Forest

A general random forests algorithm for a tree-based model can be implemented as shown in Algorithm 1 [10].

Table 1: Description and some statistics for the predicted and predictor variables

Parameter	Description	Mean	S.D	NA (%)
BOD_{in}	Inlet wastewater BOD (mg/L)	245.12	46.35	12.8
BOD_{out}	Outlet wastewater BOD (mg/L)	85.19	25.45	12.6
COD_{in}	Inlet wastewater COD (mg/L)	561.39	104.16	12.8
COD_{out}	Outlet wastewater COD (mg/L)	315.44	73.55	12.5
pH	pH	7.45	1.21	10.4
NAm	Inlet ammonia concentration (mg/L)	2.45	1.77	57.1
NN	Inlet nitrate concentration (mg/L)	1.43	0.88	81.8
P	-	1.41	1.00	83.1
Col	Color (mg/L)	464.40	123.52	10.3
T	Temperature (° C)	45.45	3.07	37.3
Cond	Conductivity (μS/cm at 25 ° C)	1530.46	378.03	10.6
RF	Rainfall (mm/day)	4.82	11.50	23.6
Pulp	Pulp production (t/day)	884.74	159.26	13.7
Pap	Paper production (t/day)	1042.71	94.09	12.9
FR	Inlet flow rate (m 3 /day)	67358.61	11592.81	7.0
SS	inlet total suspended solids (mg/L)	149.20	85.74	63.0

Algorithm 1: Random Forest Algorithm

```
Select the number of models to build, m
for i = 1 to m do
    Generate a bootstrap sample of the original data
    Train a tree model on this sample
    for each split do
        Randomly select k (< P) of the original predictors
        Select the best predictor among
        the k predictors and partition the data
    end
    Use typical tree model stopping criteria to determine
    when a tree is complete (but do not prune)
end
```

For the number models parameter the value of 128 trees there is no more significant difference between the forests using 256, 512, 1024, 2048 and 4096 trees. Therefore, it is possible to suggest a range between 64 and 128 trees in a forest [11]. For this work the quantity of trees is 128 trees.

2.3 Efficiency

The Equation 1 shows the efficiency calculation for this work. After the calculation every observation that showed a efficiency more than 60% is classified how be able to discharged (positive class) and for efficiency less than 60% how not be able to discharged (negative class).

$$Eff = \frac{(BOD_{in} - BOD_{out})}{BOD_{in}} \cdot 100 \quad (1)$$

2.4 Model evaluation

In order to evaluate the produced models some indexes were utilized. For prediction models the root-mean-square error (RMSE) was utilized for choose the best model (mtry) of Random Forest models and the R^2 was utilized for evaluate the quality of model selected.

The indexes to evaluate the classification model utilized are Accuracy, Precision, Sensitivity and Specificity. Accuracy is the quotient between the number of all correct predictions and the total number of the data set. Precision is the quotient between the number of correct positive predictions and the total number of positive predictions. Sensitivity is the quotient between the number of correct positive predictions and the total number of positives. Specificity is the quotient between the number of correct negative predictions and the total number of negatives.

3 Results and Discussion

The variables SS (63.0 %), NAm (57.1 %), NN (81.8 %), P (83.1 %) e T(37.3%) show high values of missing data (NA) and they were excluded of developed models.

3.1 Inlet BOD Model

The Figure 1 exhibits the checking graph for Inlet BOD model. The model presents a RMSE = 31.09 and $R^2 = 0.514$.

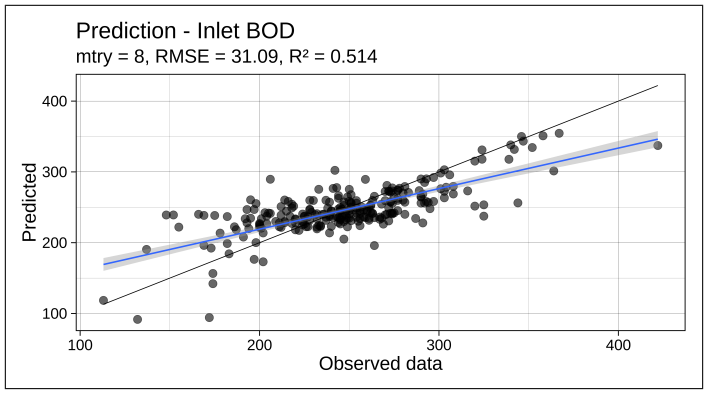

Figure 1: Checking Graph - Inlet BOD Model

The Figure 2 presents the mainly parameters to model develop. The positive highlight is the COD_{in}.

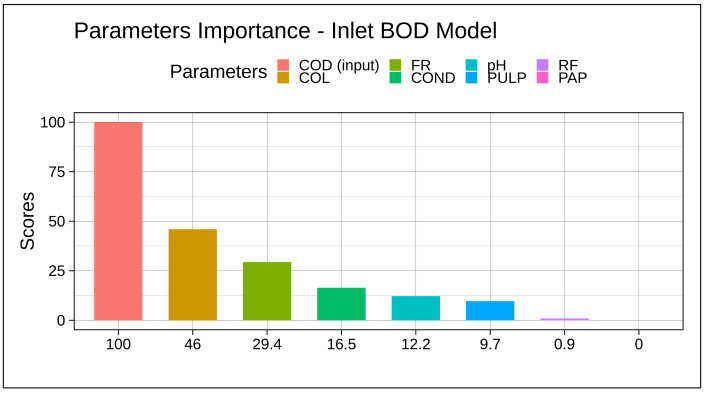

Figure 2: Importance Graph - Inlet BOD Model

The Figure 3 presents the comparison between the observed data and the prediction data (COD_{in} vs BOD_{in} observed/predicted).

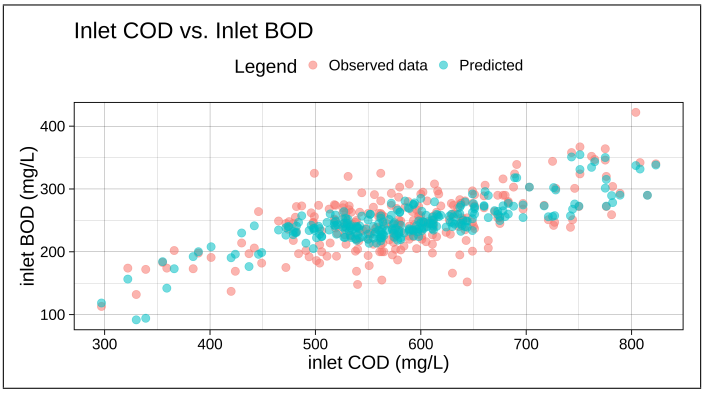

Figure 3: Observed/Predicted - Inlet BOD Model

3.2 Outlet BOD Model

In order to verify the importance of COD_{out} for BOD_{out} model, two models were made, the first without the COD_{out} how predictor variable and the second with the COD_{out} how predictor variable.

The Table 2 shows the comparison between the models. The RMSE of second model is substantially less than the firs model RMSE, and the R^2 value of the second is the double of first. These results suggest a necessity to put the COD_{out} how predictor variable. So a model for COD_{out} was made to be a predictor variable in BOD_{out} final model.

Table 2: Verify the nessecity of COD_{out} how predictor variable

Model	RMSE	R^2
Without COD_{out}	19.70	0.314
With COD_{out}	14.32	0.633

The Figure 4 exhibits the checking graph for Outlet COD model. The model presents a RMSE = 54.08 and $R^2 = 0.406$.

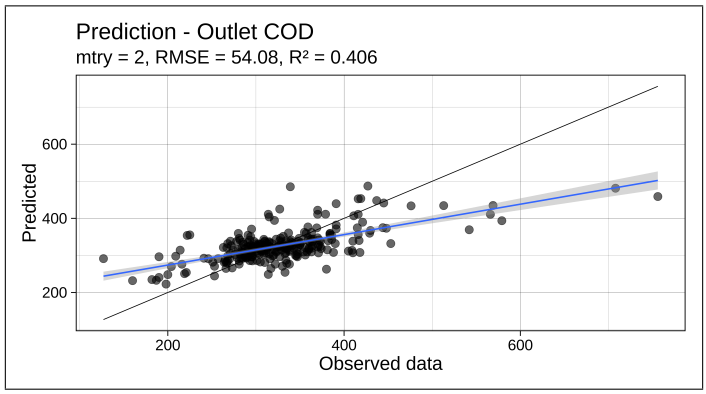

Figure 4: Checking Graph - Outlet COD Model

Even though the appear low performance of model The Figure 5 presents the checking graph for Outlet BOD model, that was made with Outlet COD model how predictor variable, with a satisfactory performance, the RMSE is 16.93 and R^2 is 0.493.

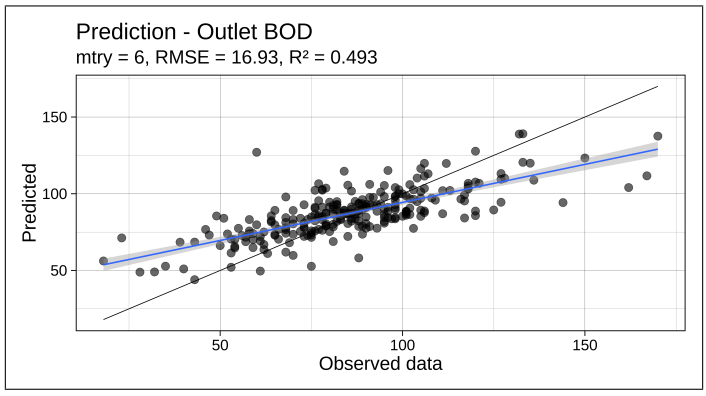

Figure 5: Checking Graph - Outlet BOD Model

The Figure 6 presents the mainly parameters to model develop. The positive

highlight is the COD_{out}, it is prove the importance of Outlet COD model.

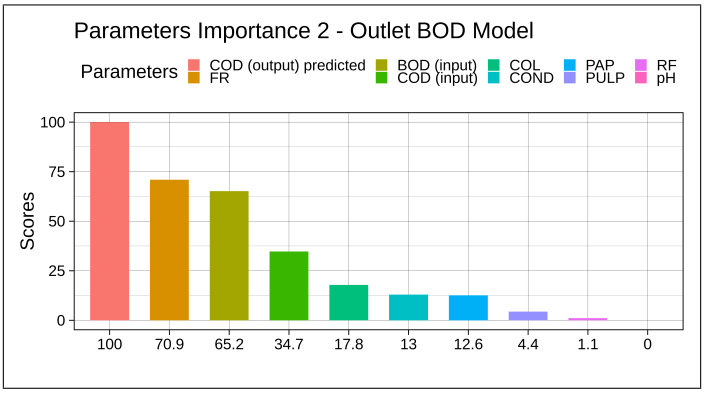

Figure 6: Importance Graph - Outlet BOD Model

The Figure 7 presents the comparison between the observed data and the prediction data (COD_{out} predicted vs BOD_{out} observed/predicted).

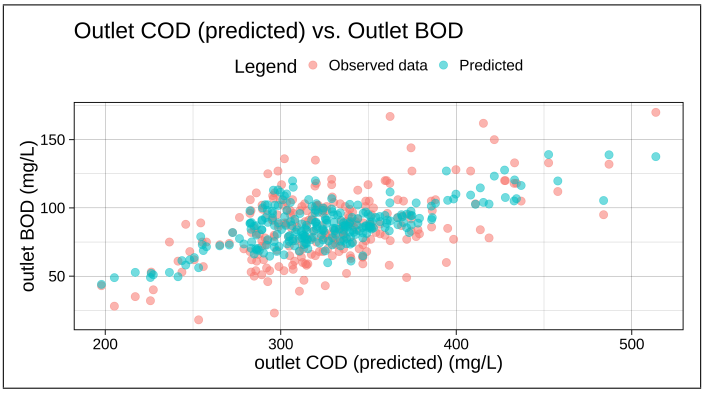

Figure 7: Observed/Predicted - Outlet BOD Model

3.3 Efficiency

The Table 3 shows some statistics for efficiency predicted and efficiency observed. Beyond of the values of the mean, median and standard deviation suggest that the prediction presents a good performance, the statistics show that the WWTP performance is satisfactory too. The mean value is more than Conama limit of 60%.

The Figure 8 presents the indexes of accuracy, precision, sensitivity and sensibility how results of confusion matrix of effluent classification. The values of the indexes show that the model performance is satisfactory, mainly when the

aim is analyze the true positives, because the precision and the sensitivity show high values.

Table 3: Some statistics for the predicted and observed efficiency

Efficiency	Mean	S.D.	Median	Range
Observed	64.65	6.68	64.49	71.12
Predicted	64.58	9.66	65.13	119.25

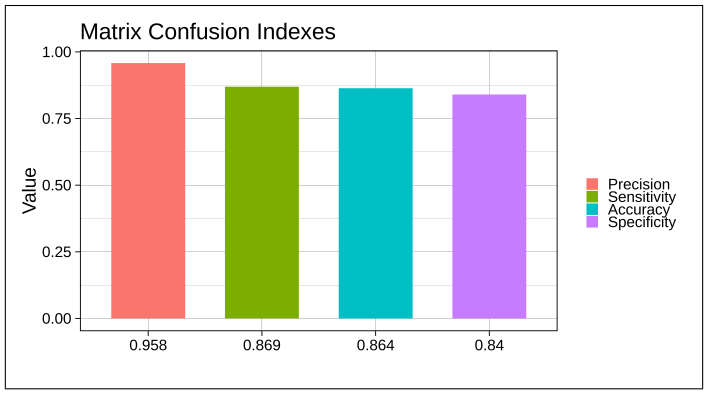

Figure 8: Classifier performance

4 Conclusion

Although of the difficulties in relation the data missing problems and the complexity of modeling a bio process vulnerable to external factors, how rains and temperature variation, the prediction models of Inlet BOD (RMSE = 31.09 and $R^2 = 0.514$) and Outlet BOD (RMSE = 16.93 and $R^2 = 0.493$) and the effluent classification model (Accuracy = 0.864, Precision = 0.958, Sensitivity = 0.869 and Specificity = 0.840) presentation a satisfactory performance, and suggest a possibility of grow in application of machine learning methods for modeling industry process like this.

Acknowledgments

The present work was supported by the "Conselho Nacional de Desenvolvimento Científico e Tecnológico" (CNPq), the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES), the "Programa de Pós-Graduação em Engenharia Industrial" (PEI) and the GAMMA (Growing with Applied Multivariate Analysis) group of the Federal University of Bahia.

References

[1] Morais, J. T. G. (2011) Análise de Componentes Principais Integrada a Redes Neurais Artificiais Para Predição de Matéria Orgânica. [s.l.] Universidade Federal da Bahia.

[2] Brasil. Resolução No 430. [s.l.] Ministério do Meio Ambiente.

[3] Oliveira-Esquerre, K. P. (2004) Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part I. Linear approaches. *Chemical Engineering Journal, v. 104, n. 1–3*, pp. 73–81.

[4] Ge, Z. (2017) Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access, v. 5*, pp. 20590–20616.

[5] Von Sperling, M. (2014) Introdução à qualidade das águas e ao tratamento de esgotos. Edição 4 ed. Belo Horizonte MG: UFMG.

[6] Yu, P. (2019) A Real-time BOD Estimation Method in Wastewater Treatment Process Based on an Optimized Extreme Learning Machine. *Applied Sciences, v. 9, n. 3*, pp. 523.

[7] Oliveira-Esquerre, K. P. (2004) Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part II. Nonlinear approaches. *Chemical Engineering Journal, v. 105, n. 1–2*, pp. 61–69.

[8] Asadi, A. (2017) Wastewater treatment aeration process optimization: A data mining approach. *Journal of Environmental Management, v. 203*, pp. 630–639.

[9] Qiu, Y. (2017) A Feasible Data-Driven Mining System to Optimize Wastewater Treatment Process Design and Operation. *Water, v. 10, n. 10*, pp. 1342.

[10] Kuhn, M. & Johnson, K. (2013) *Applied Predictive Modeling.* New York: Springer.

[11] Mayumi, O. T., Santoro, P. P. & Baranauskas, J. A. (2012) How Many Trees in a Random Forest? *Lecture notes in computer science. v. 7376*, pp. 154–168.