PREDICTION OF BIOCHEMICAL OXYGEN DEMAND IN MEXICAN SURFACE WATERS USING MACHINE LEARNING

Maximiliano Guzmán-Fernández, Misael Zambrano-de la Torre, Claudia Sifuentes-Gallardo, Oscar Cruz-Dominguez, Carlos Bautista-Capetillo, Juan Badillo-de Loera, Efrén González-Ramírez, Héctor Durán-Muñoz

Maximiliano Guzmán-Fernández

Academic Unit of Electrical Engineering, Autonomous University of Zacatecas, Mexico

4 - 5 AUGUST 2021

CONTENTS

Introduction

Water Security - Water Quality

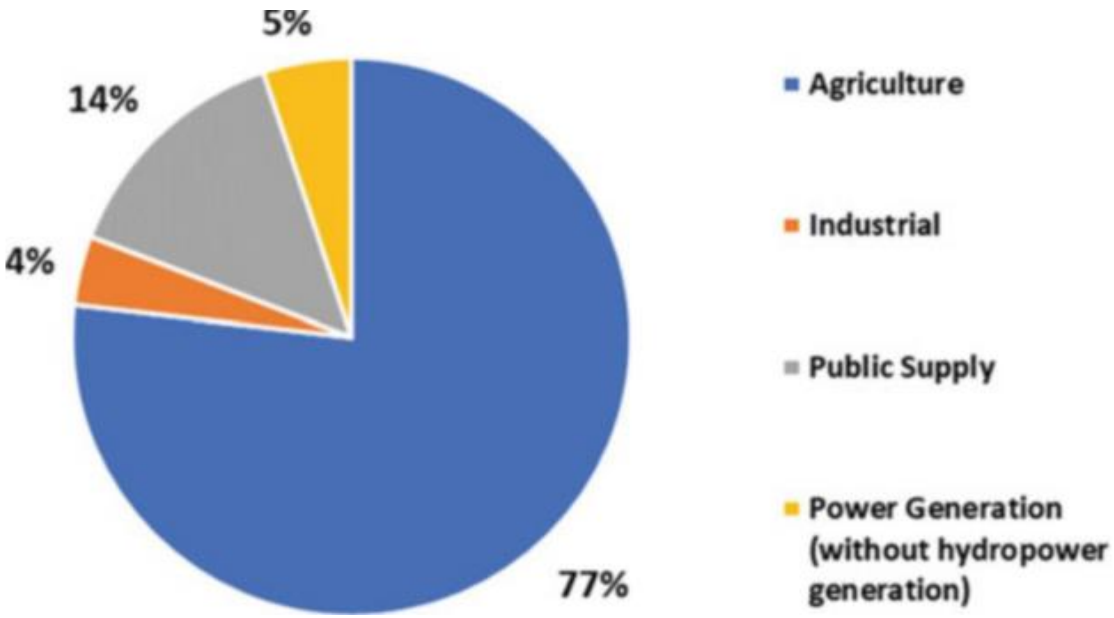


Figure 1: Uses of Water in Mexico.
M. E. Raynal Gutierrez. (2020). Water use and consumption: industrial and domestic in water resources of Mexico. Springer Nature. doi: 10.1007/978-3-030-40686-8.

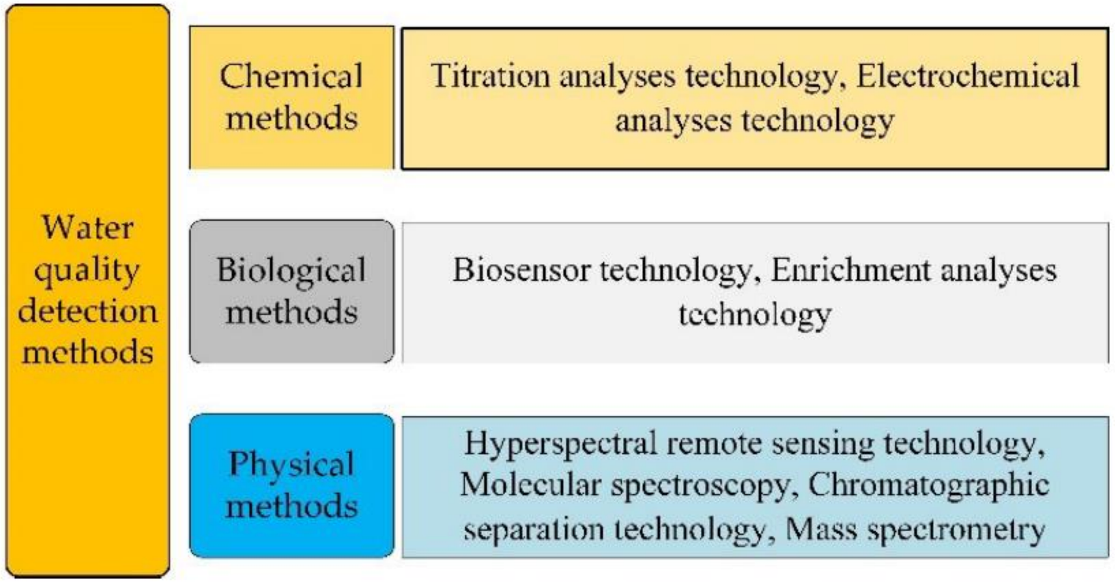


Figure 2: Methods for Determining Water Quality.
Y. Guo, C. Liu, R. Ye, and Q. Duan. (2020). Advances on Water Quality Detection by UV-Vis Spectroscopy, Appl. Sci.,10(19), doi: 10.3390/app10196874.

Introduction

Biochemical Oxygen Demand

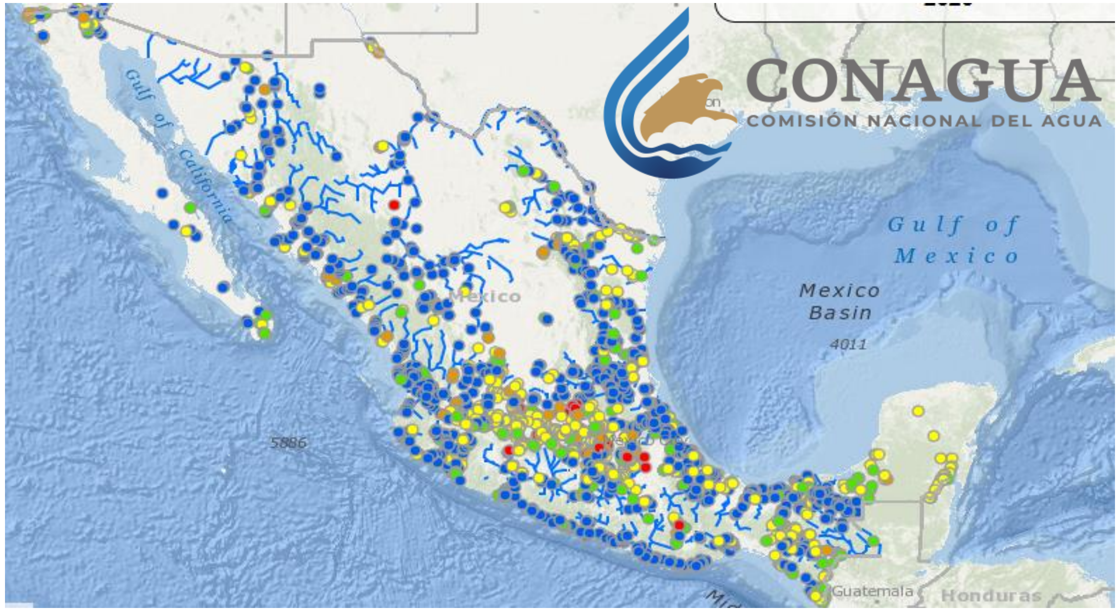


Figure 3: BOD5 levels at surface water monitoring stations in Mexico. CONAGUA (2020).

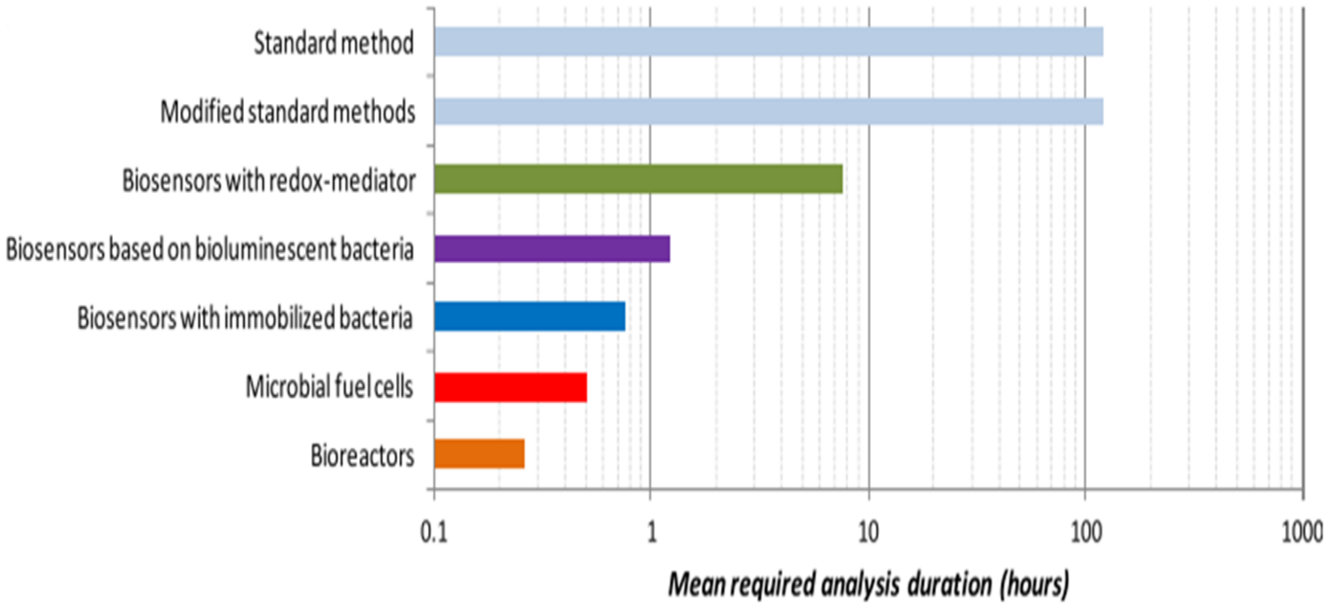


Figure 4: Required Analysis Duration to Estimate the BOD5. S. Jouanneau et al. (2014). Methods for assessing biochemical oxygen demand (BOD): A review, Water Res., 49(1), doi: 10.1016/j.watres.2013.10.066.

Objectives

1. Identify water quality parameters that can be used to predict biochemical oxygen demand.
2. Form two groups of water quality parameters that satisfy the following:

   A: If it is possible to transfer the sample to a laboratory, a group of parameters that are obtained quicker than determining biochemical oxygen demand.

   B: A group of parameters that can be measured in the study area.

3. Predict biochemical oxygen demand in surface waters using the two groups of parameters as input to machine learning algorithms (multiple linear regression, ridge regression, random forest and elastic net).

Literature Review

BOD Prediction

$$BOD = f\left(Ca^{2+}, Na^{+}, Mg^{2+}, NO_2^{-}, NO_3^{-}, PO_4^{3-}, EC, pH, Turbidity\right)$$

Table 1: Statistical appraisal of proposed models at testing stage.

Model	*R*	RMSE	MAE	OI
LS-SVM-RBF	0.83	5.725	3.959	0.761
LS-SVM-Poly	0.85	5.463	4.508	0.778
MARS	0.79	6.719	5.399	0.688
ANN	0.74	6.946	5.940	0.671
ANFIS	0.81	6.118	4.727	0.733
MLR	0.78	15.775	14.52	− 0.391
MNLR	0.59	20.871	15.40	− 1.333

M. Najafzadeh and A. Ghaemi. (2019). Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods," Environ. Monit. Assess.,191(6). doi: 10.1007/s10661-019-7446-8.

Table 2: Statistical appraisal of proposed models at testing stage.

Model	*R*	RMSE	MAPE	NSE
EPR	0.84	5.60	32.08	0.709
GEP	0.86	5.388	31.53	0.731
MT	0.76	6.803	40.76	0.571

M. Najafzadeh, A. Ghaemi, and S. Emamgholizadeh. (2019). Prediction of water quality parameters using evolutionary computing-based formulations. Int. J. Environ. Sci. Technol., 16(10). doi: 10.1007/s13762-018-2049-4.

Methodology

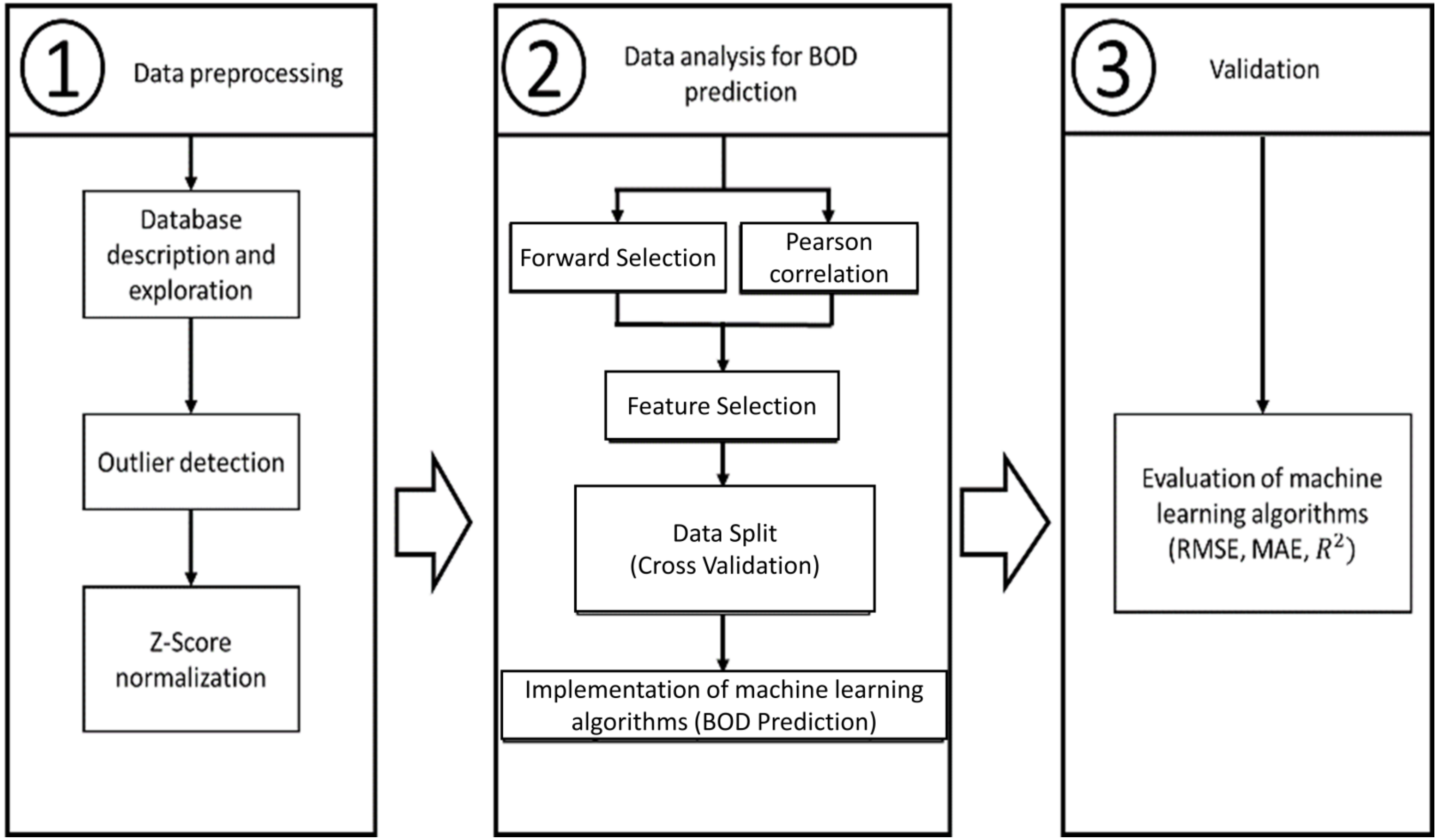


Figure 5: Stages of the methodology: (1) Data preprocessing, (2) Data analysis for prediction of biochemical oxygen demand and (3) Validation.

Methodology

Stage 1: Data preprocessing.

(A)

CONAGUA
COMISIÓN NACIONAL DEL AGUA

2764 surface water monitoring stations on the period of time 2012-2019.

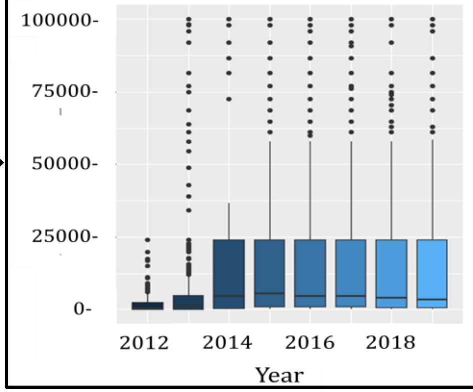
Database	Observations
177 Parameters	110827

Cleaning of the data.

Dataset	Observations
31 Parameters	58824

(B)

Outliers.



Parameter	Min	Mean	Max
Parameter1	###	####	###
Parameter2	###	####	###

(C)

Z-Score normalization of the parameters.

$$z - score = \frac{(x - \mu)}{\sigma} \quad (1)$$

where x represents the parameter value, μ is the mean of the parameter and σ is the standard deviation.

Figure 6: Steps of Stage 1: (A) Database description and exploration, (B) Outlier detection and (C) Z-Score normalization.

Methodology

Stage 2: Data analysis for prediction of biochemical oxygen demand.

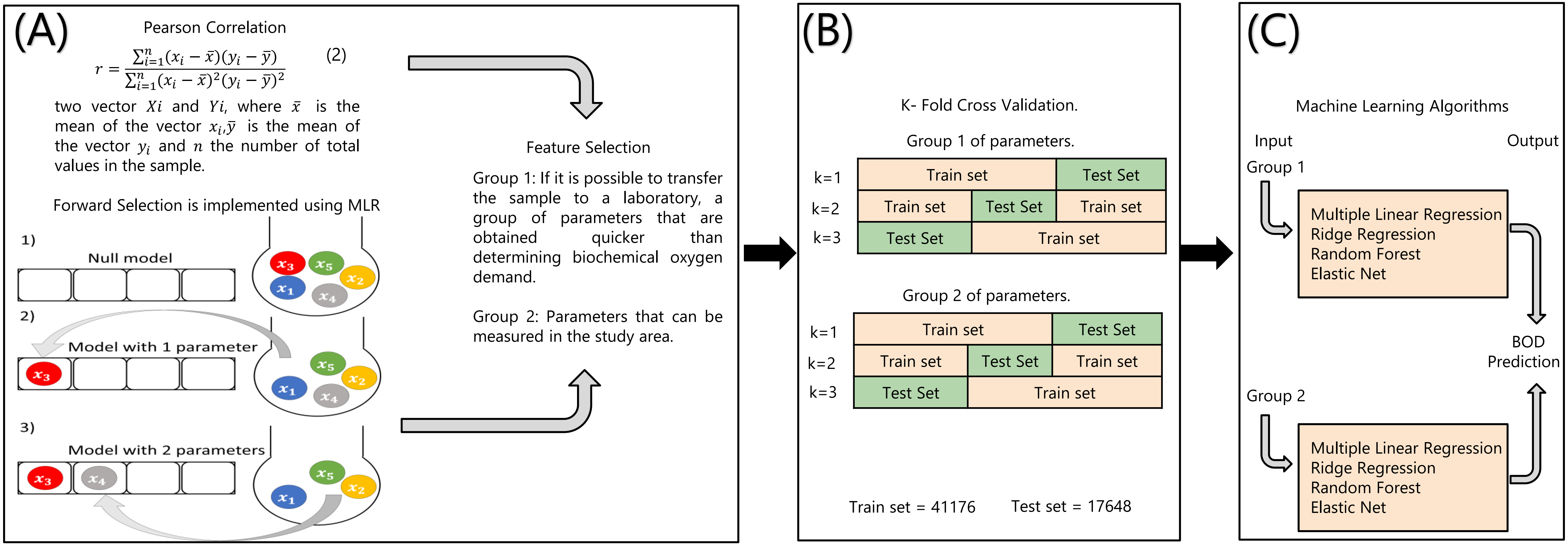


Figure 7: Steps of Stage 2: (A) Feature Selection, (B) Data Split and (C) Implementation of ML Algorithms.

Methodology

Stage 3: Validation.

Goodness- of-Fit Statistics

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2} \quad (4)$$

$$MAE = \frac{1}{n}\|y_i - y_i'\| \quad (5)$$

where y_i are the actual values, y_i' are the predictions, $\bar{y}$ represents the mean of the values and n the number of total values in the sample.

Figure 8: Evaluation of Machine Learning Algorithms.

Results and Discussions

Outlier Detection.

Table 3: Basic statistics of the parameters.

Parameter (Units)	Min	Mean	Max	Parameter (Units)	Min	Mean	Max
Fecal Coliform (NMP/100mL)	1	55772	24196000	Total Suspended Solids (mg/L)	0.1	105	20812
Escherichia Coli (NMP/100mL)	1	46459	24196000	Turbidity (UNT)	0.01	75	21500
Biochemical Oxygen Demand (mg/L)	0.1	23.2	7667	Arsenic (mg/L)	0.0001	0.006	1
Chemical Oxygen Demand (mg/L)	0.9	77.7	14489	Cadmium (mg/L)	0.00002	0.0002	0.1
Phosphorus (mg/L)	0.001	1.3	95.2	Chromium (mg/L)	0.0002	0.01	76.5
Organic Nitrogen (mg/L)	0	2.5	827.8	Mercury (mg/L)	0.00001	0.0003	0.5
True Color (U Pt/Co)	2.5	55.2	8000	Nickel (mg/L)	0	0.005	7.3
UV Absorbance (U Abs/cm)	0.002	0.17	17	Lead (mg/L)	0.001	0.003	1.8
Total Dissolved Solids (mg/L)	2.4	354.5	159520	Hardness (mg/L)	3.8	295.2	37965
Electrical Conductivity(uS/cm)	3.8	1056	199400	Temperature (°C)	-6	27.6	51
PH (UpH)	2.9	7.8	11.8	Water Temperature (°C)	4	24.9	62
% Dissolved Oxygen (% Saturation)	0.6	73.2	1113.3	Total Organic Carbon (mg/L)	0.06	12.8	2490
Dissolved Oxygen (mg/L)	0.05	5.7	762	Nitrogen (mg/L)	0.008	7.4	1244.1
Ammoniacal Nitrogen (mg/L)	0.003	3.7	497	Kjeldahl Nitrogen (mg/L)	0.003	6.34	1239.8
Nitrogen Dioxide (mg/L)	0.0005	0.1	21.84	Orto-Phosphate (mg/L)	0.0005	0.87	144.4
Nitrate Nitrogen (mg/L)	0.0004	1	336.2				

Table 4: Basic statistics of the parameters after assigning a limit value.

Parameter (Units)	Min	Mean	Max	Parameter (Units)	Min	Mean	Max
Biochemical Oxygen Demand (mg/L)	0.1	14.7	120	Total Suspended Solids (mg/L)	0.1	66.8	400
Chemical Oxygen Demand (mg/L)	0.9	55.2	250	Phosphorus (mg/L)	0.001	1.2	20
Dissolved Oxygen (mg/L)	0.05	5.7	40	Temperature (°C)	-6	27.6	51
True Color (U Pt/Co)	2.5	45.1	200	Turbidity (UNT)	0.01	49.2	500
UV Absorbance (U Abs/cm)	0.002	0.17	2	Water Temperature (°C)	4	24.9	62
Ammoniacal Nitrogen (mg/L)	0.003	3.7	200	Kjeldahl Nitrogen (mg/L)	0.003	6.3	400
Electrical Conductivity(uS/cm)	3.8	900	5000	Total Dissolved Solids (mg/L)	2.4	455.4	1000
Total Organic Carbon (mg/L)	0.06	12.5	1000				

Results and Discussions

Feature Selection.

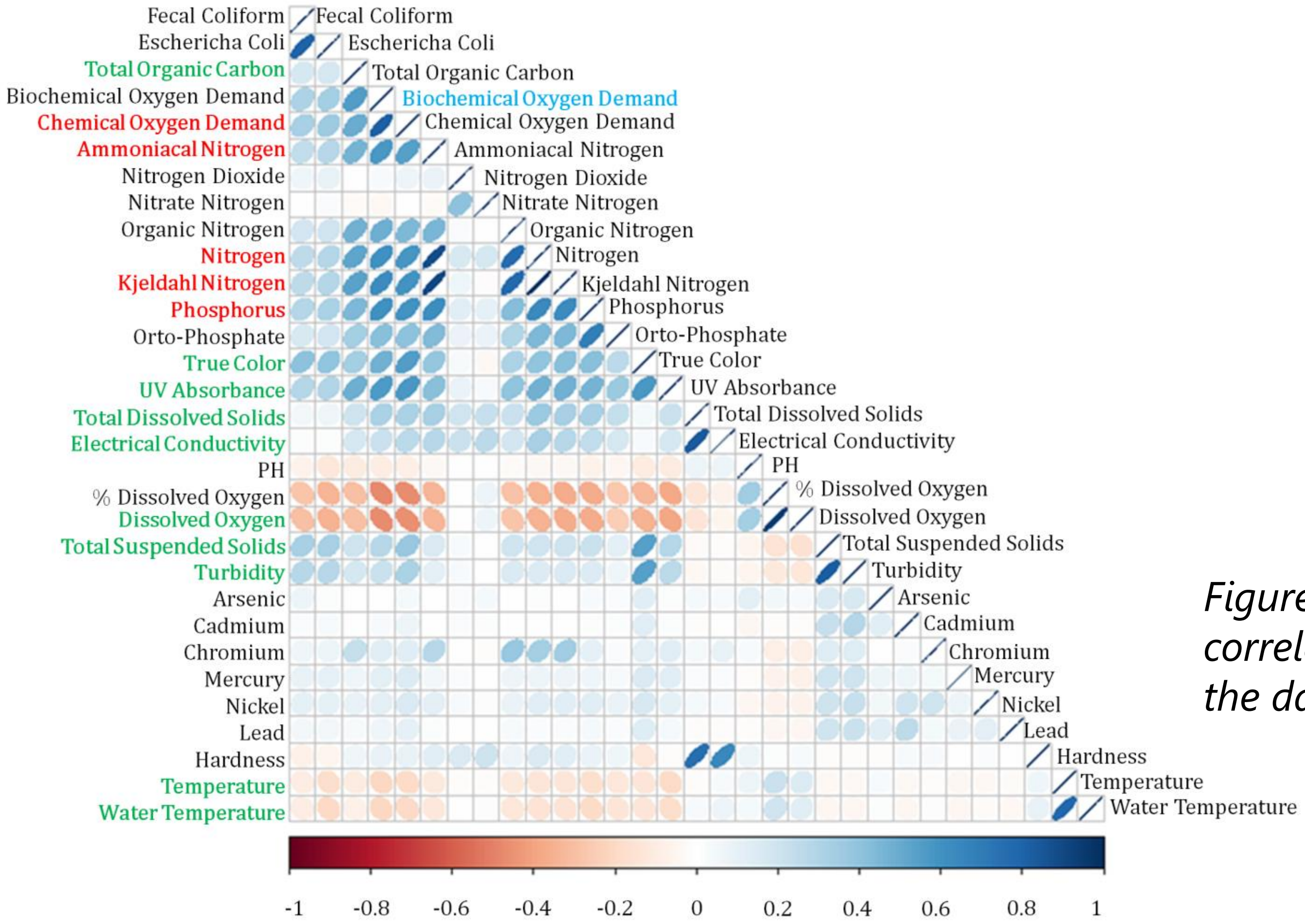


Figure 9: Heat map of Pearson's correlation for the 31 parameters in the dataset.

Results and Discussions

Feature Selection.

Table 5: Identification of individual parameters with the highest coefficient of determination when applying Forward Selection.

Parameter	Coefficient of Determination [0-1]	
	Training	Testing
Fecal Coliform	0.09	0.08
Escherichia Coli	0.11	0.09
Biochemical Oxygen Demand	1	1
Chemical Oxygen Demand	0.66	0.66
Total Suspended Solids	0.07	0.06
Total Dissolved Solids	0.1	0.1
Phosphorus	0.36	0.32
True Color	0.22	0.21
UV Absorbance	0.31	0.3
Electrical Conductivity	0.04	0.04
PH	0.008	0.008
% Dissolved Oxygen	0.22	0.23
Dissolved Oxygen	0.21	0.22
Turbidity	0.04	0.04
Arsenic	0.00001	0.00003
Cadmium	0.001	0.002
Chromium	0.016	0.019
Mercury	0.01	0.009
Nickel	0.013	0.025
Lead	0.004	0.001
Hardness	0.01	0.01
Temperature	0.04	0.04
Water Temperature	0.04	0.04
Total Organic Carbon	0.3	0.2
Ammoniacal Nitrogen	0.3	0.29
Nitrogen Dioxide	0.001	0.003
Nitrate Nitrogen	0.001	0.001
Organic Nitrogen	0.24	0.19
Nitrogen	0.37	0.33
Kjeldahl Nitrogen	0.39	0.34
Orto-Phosphate	0.17	0.17

Table 6: Coefficient of determination when combining the parameters by Forward Selection into two sets.

Sets of parameters used as input to multiple linear regression algorithm	Coefficient of Determination [0-1]	
	Training	Testing
Chemical Oxygen Demand, Ammoniacal Nitrogen	0.69	0.69
(Set 1) Chemical Oxygen Demand, Ammoniacal Nitrogen, Kjeldahl Nitrogen, Phosphorus.	0.70	0.70
Total Organic Carbon, True Color, UV Absorbance, Total Dissolved Solids, Electrical Conductivity.	0.48	0.46
(Set 2) Total Organic Carbon, True Color, UV Absorbance, Total Dissolved Solids, Electrical Conductivity, Total Suspended Solids, Turbidity, Dissolved Oxygen, Temperature, Water Temperature.	0.53	0.51

Results and Discussions

Feature Selection.

Group 1

Parameter (Units)
Chemical Oxygen Demand (mg/L)
Ammoniacal Nitrogen (mg/L)
Kjeldahl Nitrogen (mg/L)
Phosphorus (mg/L)

Group 2

Parameter (Units)
Total Organic Carbon (mg/L)
True Color (U Pt/Co)
UV Absorbance (U Abs/cm)
Total Dissolved Solids (mg/L)
Electrical Conductivity(uS/cm)
Total Suspended Solids (mg/L)
Turbidity (UNT)
Dissolved Oxygen (mg/L)
Temperature (°C)
Water Temperature (°C)

Figure 10: Groups of water quality parameters selected for BOD prediction.

Results and Discussions

Validation of Machine Learning Algorithms.

Table 7: Results in the testing stage of the algorithms using group 1 as input.

Parameter (Units)	Algorithm	Goodness of fit Statistics		
		Root Mean Square Error	Coefficient of Determination	Mean Absolute Error
Chemical Oxygen Demand (mg/L) Ammoniacal Nitrogen (mg/L) Kjeldahl Nitrogen (mg/L) Phosphorus (mg/L)	Multiple Linear Regression	0.53	0.7	0.30
	Ridge Regression	0.53	0.7	0.30
	Random Forest	**0.48**	**0.76**	**0.23**
	Elastic Net	0.53	0.7	0.30

Table 8: Results in the testing stage of the algorithms using group 2 as input.

Parameter (Units)	Algorithm	Goodness of fit Statistics		
		Root Mean Square Error	Coefficient of Determination	Mean Absolute Error
Total Organic Carbon (mg/L) True Color (U Pt/Co) UV Absorbance (U Abs/cm) Total Dissolved Solids (mg/L) Electrical Conductivity(uS/cm) Total Suspended Solids (mg/L) Turbidity (UNT) Dissolved Oxygen (mg/L) Temperature (°C) Water Temperature (°C)	Multiple Linear Regression	0.67	0.52	0.42
	Ridge Regression	0.67	0.52	0.42
	Random Forest	**0.48**	**0.75**	**0.24**
	Elastic Net	0.67	0.52	0.42

Conclusions

Water quality is essential for the human life development. Through the present work it was possible to identify the best algorithm that can predict the biochemical oxygen demand in surface waters of Mexico. Also, the parameters that have the most influence. Random forest showed flexibility when implemented in the prediction of biochemical oxygen demand by obtaining 0.48 RMSE, 0.76 R^2 and 0.23 MAE using the parameters Chemical Oxygen Demand, Ammoniacal Nitrogen, Kjeldahl Nitrogen and Phosphorus. In addition, 0.48 RMSE, 0.75 R^2 and 0.24 MAE were obtained using the parameters Total Organic Carbon, True Color, UV Absorption, Total Dissolved Solids, Electrical Conductivity, Total Suspended Solids, Turbidity, Dissolved Oxygen, Water Temperature and Temperature. This indicates that based on the local conditions and the study area, the biochemical oxygen demand can be obtained in a similar way and diagnose water contamination in Mexico in a relatively short time. As a future work, it is proposed to design and develop a real-time electronic monitoring device to measure the parameters of group 2 obtained in this work.

THANK YOU