"I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of the degree of Bachelor of Electrical and Electronics Engineering And Computer Science And Engineering

Signature	:	…….....
Name of Supervisor	:	ASST. PROF. DR. IFTEKHARUL MOBIN
Date	:	……..

Fault Detection in Waste Water Treatment Plant Using Statistical Analysis & Machine Learning

MD. EZAZUL HAQUE - 12121049
MD. MAZED UL ISLAM - 12321046
NOOR A ELAHI RAHAT-12321012
MD. MAHMUDUL AMIN- 10321016

A thesis submitted in fulfilment of the
requirements for the award of the degree of

Bachelor of Electrical and Electronics Engineering and Computer Science Department of
Computer Science and Engineering and Electrical and Electronics Engineering
BRAC University Bangladesh

December 2019

Declaration

This is to certify that this final thesis report of '***Fault detection in wastewater treatment plant using Statistical analysis and machine learning'*** is submitted by the authors for the purpose of obtaining a degree in Bachelors of Science in Computer Science. We hereby declare all the work presented in this thesis paper are authentic and any inspiration of the work have been accredited with proper referencing.

Signature of Supervisor

Dr. IFTEKHARUL MOBIN

Signature of Authors

MD. MAZED UL ISLAM

NOOR E ELAHI RAHAT

MD. EZAZUL HAQUE

MD. MAHMUDUL AMIN CHOWDHURY

Acknowledgement

First and foremost, all praises to Almighty God, the most high and merciful.

Our most sincere gratitude to our thesis supervisor Dr. Iftekharul Mobin. Without his assistance and dedicated involvement in every step throughout the process, completing this thesis might have proved to be an impossible task.

Also, we want to thank our family, friends and well-wishers who inspired, encouraged and fully supported us through every trial that came our way, who helped us not only financially, but morally and spiritually.

Finally, we are very thankful to BRAC University, Bangladesh for giving us a chance to complete our B.Sc. degree in Electrical and Electronics Engineering.

Abstract

.

A suitable model needs to be developed for detecting fault of wastewater treatment plant in order to monitor, predict plant performance and for reducing environment pollutions. Main objective of this study is to introduce time and cost effective data science & machine learning technique to monitor WWTP's performance and detect plant's fault instead of manual, laboratory based time consuming, costly, difficult methods. One year of unsupervised data of WWTP collected and convert into supervised data in order to visualized plant's fault using python. Moreover four model has been created based on water quality standard parameters(Ph, BOD, COD, suspended solid) and we applied different machine learning algorithm's to take decision by machine itself after identifying normal or faulty data. Machine learning technique in case of finding fault, taking decision gives satisfactory result but different algorithms shows best accuracy for different model. However, machine-learning method will be accurate automatic solution for detecting fault of wastewater treatment plant and reducing environment pollution.

Table of Contents

Chapter 1

Introduction

Wastewater treatment plant is an engineering process to convert waste water into fresh water. This treatment plant is divided into three parts pre-treatment, primary, secondary. It's a dynamic and complex system widely used in many chemical, textile, medicine, leather industries etc. Principle objective of the system is to remove toxic element from industrial waste water and make it usable. In many countries establishing WWTP or ETP in industry is mandatory. As many industries set up their WWTP but it is difficult to maintain this large dynamic and complex system. If the system gets damage or doesn't work properly it will be more difficult to find out the fault and it will take so much time & cost. That's why we were looking for a simple, time saving, cost saving, automatic solution of this problem. In this study we want to analyze the waste water treatment plant to determine whether the plant is working properly or not. Occasionally plant may experience disruption and may not work efficiently. Hence Multivariate Statistical Techniques will be used to identify faults. Then Machine Learning technique will be used to predict the fault pattern of the treatment plant. Our result will demonstrate the overall performance of the waste water treatment plant and will be helpful in reducing environment pollution.

Chapter 2

Literature review

Historical Background

The release of wastewater to situations made unfriendly condition and this drove the advancement of concentrated strategies for sewage treatment. Sedimentation and synthetic precipitation were one of the primary procedures utilized for wastewater treatment. In 1865, early examinations on microbiology of slime assimilation were led in England [2]. In 1868, early tests on irregular filtration of wastewater were led, while in 1870 early examinations on discontinuous sand filtration were made in England. In 1882, first tests on air circulation occurred in England. Joined States was the first to utilize bar racks in 1884. In United States first substance precipitation treatment plant was introduced in 1887. In 1889, filtration in contact beds was attempted at the Lawrence Experiment station, Massachusetts. In 1891, the technique for muck processing in tidal ponds was created in Germany.

In 1895, methane gas was collected from the septic tanks and used for plant lighting in England. The first rotary sprinklers for rotary filters were developed in 1898. The first grit chambers were developed in the United States in 1904. The hostile character of the slop created by sedimentation prompted the utilization of septic tanks in which the solids were rendered pretty much innocuous, however troubles of different sorts prompted the general adaption of Travis two-story septic tank in England in 1904 and the Imh off tank which was licensed in Germany in 1904. In 1912-13 air circulation of wastewater in tanks containing slate was done at Lawrence Experiment station. In 1914, tests were led by Arden and Lockett that prompted the improvement of the actuated slop process, wherein a high level of decontamination is

accomplished. The procedure was first connected in a city plant for treating sewage at San Marces, Tex in 1916 [2]. In 1925, contact aerators were produced by Buswell in United States. In 1912-13 aeration of wastewater in tanks containing slate was carried out at Lawrence Experiment station. In 1914, experiments were conducted by Arden and Lockett that led to the development of the activated sludge process, wherein a high degree of purification is achieved. The process was first applied in a municipal plant for treating sewage at San Marces, Tex in 1916. In 1925, contact aerators were developed by Buswell in the United States. The changing qualities of wastewater, because of the release of numerous contaminants are in charge of the numerous progressions that are occurring today in the wastewater treatment. Wastewater or sewage treatment is one such option, wherein numerous procedures are composed and worked keeping in mind the end goal to copy the normal treatment procedures to diminish contamination load to a level that nature can deal with. In such a manner, exceptional consideration is important to survey the natural effects of existing wastewater treatment offices.

2.1 Objectives to the study

Objective of this study is to developed a model for detecting fault of wastewater treatment plant in order to

- Monitor, predict plant performance and for reducing environment pollutions.
- Introduce time and cost effective data science & machine learning technique to monitor WWTP's performance
- Determine the specific controlling process to perform.

2.2 Important parameters

1. **Ph:** Ph indicates that water is acidic or basic. Standards 0-14.from 0 to below 7 its acid. Above 7 it's basic.
2. **BOD:** Biological or biochemical oxygen demand means the standard oxygen need for biodegradation of microorganisms. A BOD level of 1-2 ppm is considered very good. More than 100 ppm it is consider as heavily polluted.
3. **COD:** Chemical oxygen demand measurement process is fast and reliable.

   Permissible **value** of **COD** is 250 mg/l.
4. **Conductivity:** Conductivity of water determines electrons flow capacity through water or electricity passing ability of the water. Conductivity standard level 0-800 is suitable for human consumption and 800-2500 is suitable for all livestock. Over 10000 the level is not suitable for both human and irrigation.
5. **Sediments:** It defines the clearness of water or how much clarity present in water. Filtering is the reducing process of sediments.
6. **Suspended Solids:** plant, leaves, degraded organic materials and many other things known as suspended solid of water which can be remove by filtering.

Chapter 3

Experiments

An experiment has been done for detecting waste water treatment fault in this research. Data driven technique applied to an online data of a plant in this process, we have found that output depends on the quality and quantity of data. Data driven technique has two part univariate technique and multivariate technique. The multivariate technique we applied because through univariate technique a large dataset where each variable depends on other can't analyze on the opposite side multivariate technique can handle a large number of inter related variables. There are various multivariate technique (factor analysis, cluster analysis, multidimensional scaling, principle component analysis) [3]. There is two type of learning, one is supervised and the other one is unsupervised. In our research, we have collected an unsupervised dataset that's why, we use cluster analysis for identifying the good data or bad data of a wastewater treatment plant. Before cluster analysis, we have generated some graph which represents some correlation among the parameters and there we have noticed little flocculated point, those points or data we consider as faulty data which generated because of disturbance of plant machinery.

3.1 Fault Detection

In this research, we are detecting a fault of a waste water treatment plant where we are analyzing a large number of variable using the multivariate technique (cluster analysis). As we know multivariate technique provides a primary variable list which reflects another variable of the process.

3.2 Dataset

The research aim is to determine the condition of the waste water treatment plant by using statistical analysis and machine learning. A dataset is collected from UCI machine learning respiratory this dataset created by Manel Poah, Unitat d'Enginyeria Quinica Universitat Autonoma de Bercelona [3]. This dataset consists the variables record for 527 days. 3500 m^3 of domestic and industrial waste water is treated daily by this waste water treatment plant. This plant has 3 parts (pre treatment, primary treatment, secondary treatment). This dataset comes from the daily measures of sensors in an urban waste water treatment plant. The objective of this work is to classify the state of the plants on which they operate, in order to predict faults through the state variables of the plant at each of the stages of the treatment process. This domain has been stated as an ill-structured domain and consist 38 attributes.

Table-1: WWTP's thirty eight attributes

Attributes	Description
Q-E	Input flow to plant
ZN-E	Input zinc to plant
PH-E	Input ph to plant
DBO-E	Input biological oxygen demand
DQO-E	Input chemical demand of O_2
SS-E	Input suspended solid to plant
SSV-E	input suspended solids to plant
SED-E	input volatile suspended solids to plant
COND-E	input sediments to plant
PH-P	input conductivity to plant
DBO-P	input pH to primary settler
SS-P	input Biological demand of oxygen to primary settler
SSV-P	input suspended solids to primary settler
SED-P	input volatile suspended solids to primary settler
COND-P	input sediments to primary settler

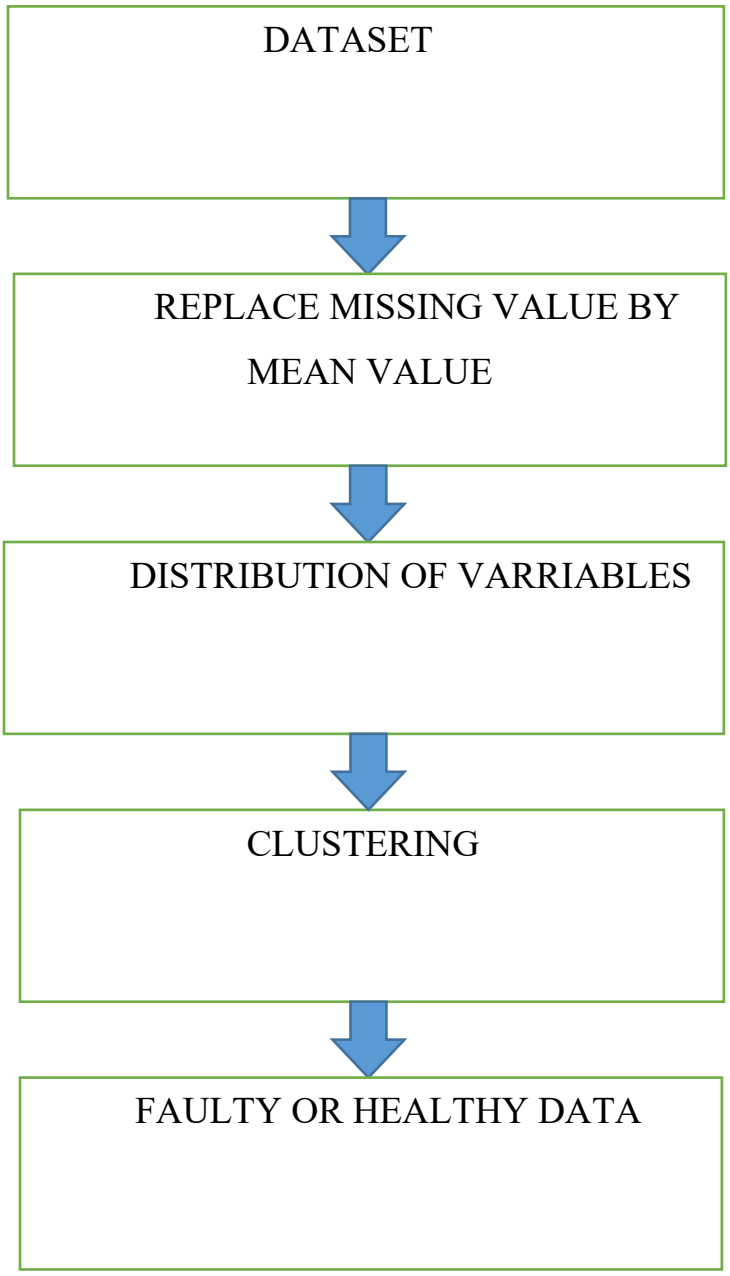
PH-D	input conductivity to primary settler
DBO-D	input Biological demand of oxygen to secondary settler
DQO-D	input chemical demand of oxygen to secondary settler
SS-D	input suspended solids to secondary settler
SSV-D	input volatile suspended solids to secondary settler
SED-D	input sediments to secondary settler
COND-D	input conductivity to secondary settler
PH-S	output pH
DBO-S	output Biological demand of oxygen
DQO-S	output chemical demand of oxygen
SS-S	output suspended solids
SSV-S	output volatile suspended solids
SED-S	output sediments
COND-S	output conductivity
RD-DBO-P	performance input Biological demand of oxygen in primary settler
RD-SS-P	performance input suspended solids to primary settler
RD-SED-P	performance input sediments to primary settler
RD-DBO-S	performance input Biological demand of oxygen to secondary settler
RD-DQO-S	oxygen to secondary settler
RD-DBO-G	global performance input Biological demand of O_2
RD-DQO-G	global performance input chemical demand of O_2
RD-SS-G	global performance input suspended solids
RD-SED-G	global performance input sediments

Table-3.1: WWTP's thirty eight attributes, source: [3]

In this dataset, there are some missing values which are replaced by each parameter mean or average value. The dimensional dataset is reduced because a large dataset is very hard to

classify. From this dataset, using variable distribution and Neural Network technique, we identify healthy or faulty data.

3.3 Workflow



Chapter 4

Analysis

4.1 Analysis & Discussion

In this research work we want to find out whether there are faulty machineries inside the water treatment plants or not. We eager to investigate what is the impact on the plant output in different stage of the plant. To make it more precise our focus is to analyze dataset and observe different parameters (PH, Dissolved Oxygen, Sediments, and solid particles) [4] and identify which parameter or set of attributes show abnormal behaviour. Our goal is to precisely estimate -

1. Whether the water plant has any fault or does it operating properly.
2. Observe abnormal behavior of certain parameters in the dataset and predict which machines or parts of plants are having problem/may have problem.

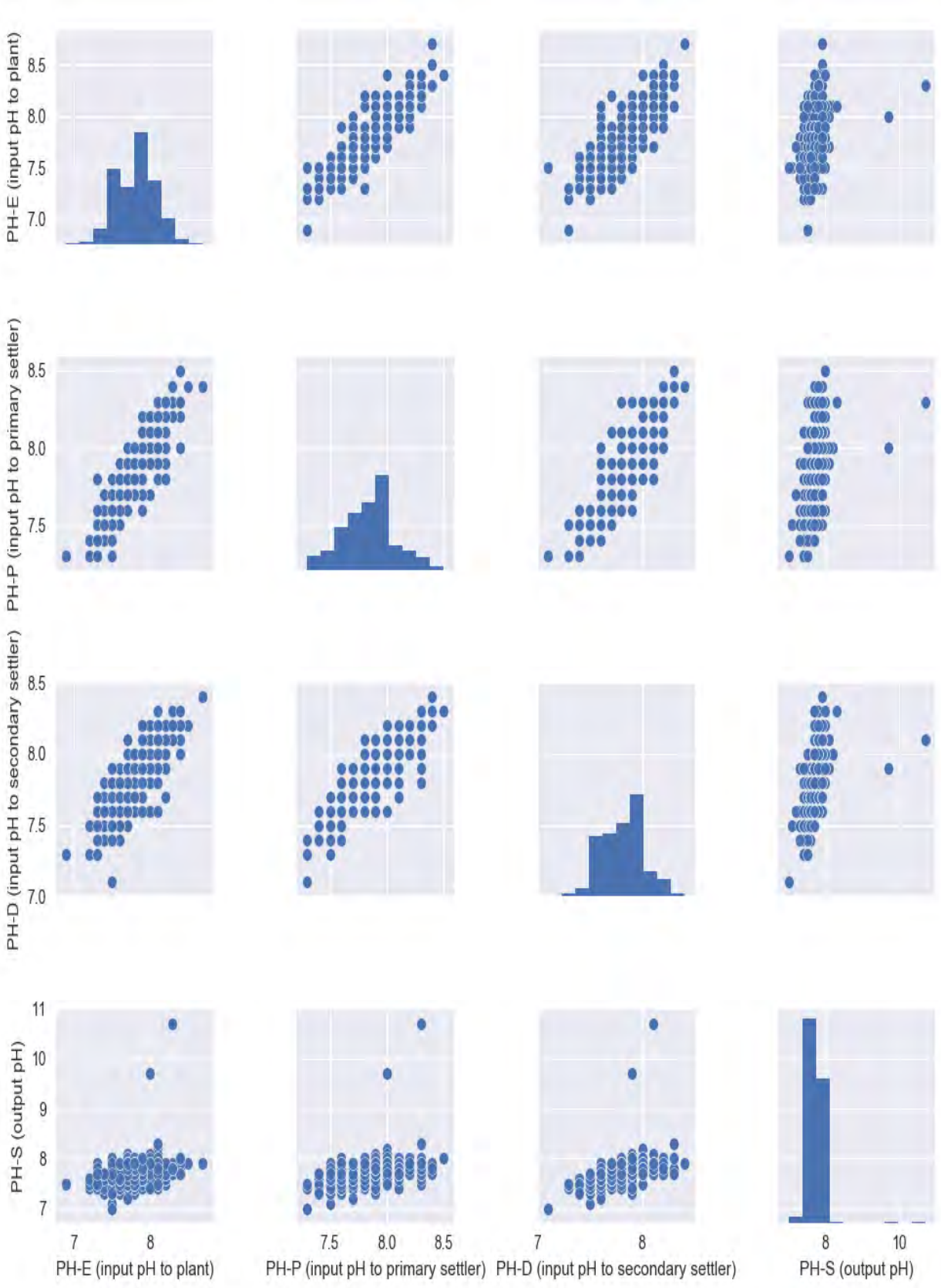
4.2 PH Analysis

The following picture represents a scatter plot matrix which depicts variable relationships among each other. In the picture it shows how 'PH-E [4] (input pH to plant)','PH-P [4] (input pH to primary settler)', 'PH-D (input pH to secondary settler)', 'PH-S [4] (output pH)' these four variables' values are distributed. Do they show any kind of patterns? Do the charts 'PH-E (input pH to plant)','PH-P (input pH to primary settler) show any specific behaviors? Here every single charts represents two variables' ('PH-E (input pH to plant)','PH-P (input pH to primary settler)') relationships or values distributions we can observe.

That means here we can see 16 charts because it is a representation of 4x4 matrix. The histogram chart (A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data – for an example it will show how many students have got GPA A or B or C in a class) is shown whenever we are getting same variables on X and Y matrix. Look at bottom right corner, there the X axis is PH-S (output pH) and if you look at bottom left you can see Y axis is also PH-S (output pH).

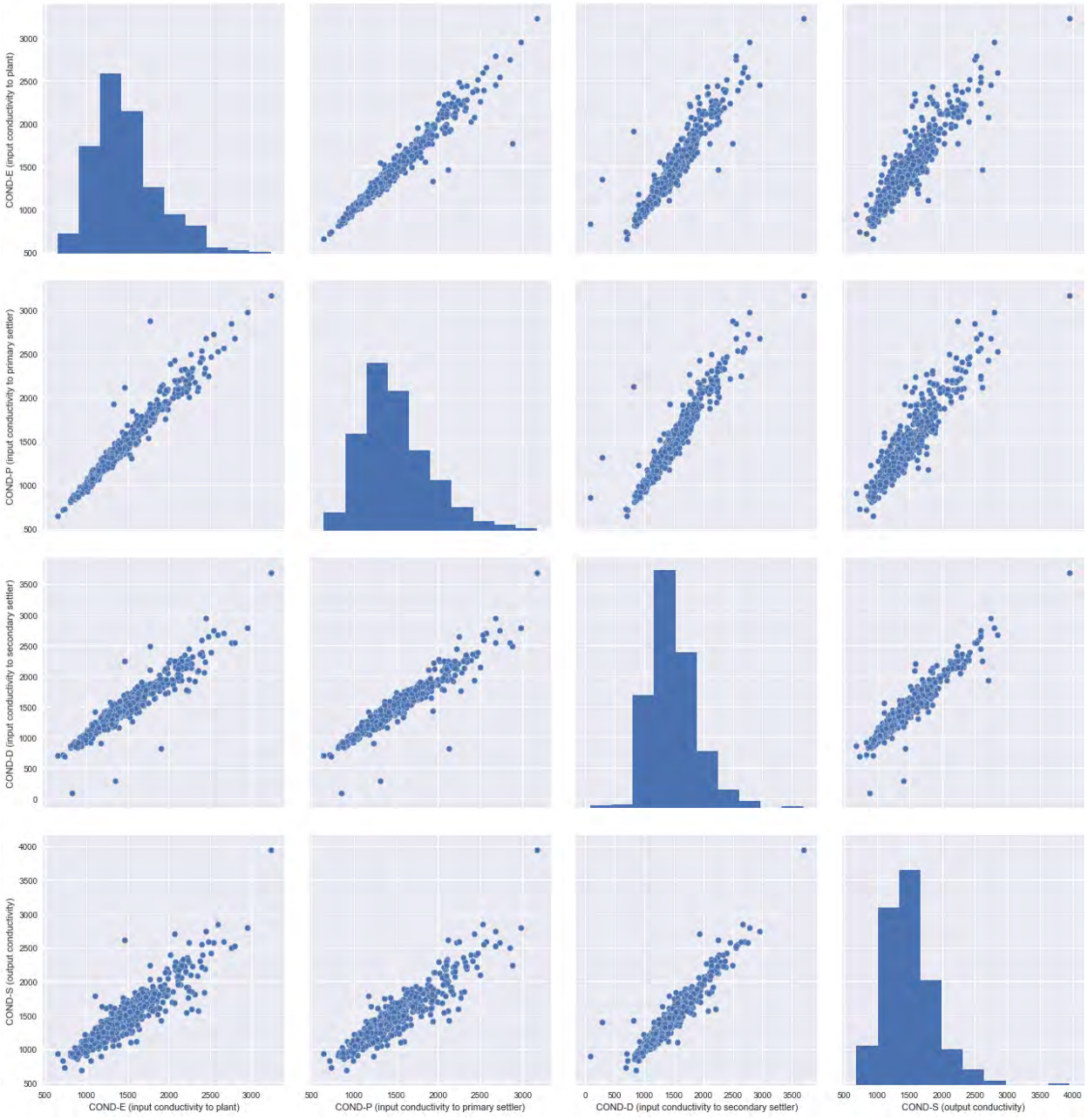
Now other charts represents one variable against another variable such as: Look at top right corner on the graph here X axis is PH-S (output pH) and left top corner represents Y axis which is PH-E (input pH to plant). From this graph distribution we can clearly visualize that most of the pH values are within 7-8 range except very few. We are assuming these few values are faulty data and these data are generated because of machineries fault during the peak hour.

Figure-1: Following graph represents distribution of 'PH-E (input pH to plant)','PH-P (input pH to primary settler)', 'PH-D (input pH to secondary settler)', 'PH-S (output pH)'.


PH-E (input pH to plant)
PH-P (input pH to primary settler)
PH-D (input pH to secondary settler)
PH-S (output pH)
8.5
8.0
7.5
7.0
11
10
9
8
7
7
8
7.5
8.0
8.5
7
8
8
10

4.3 Conductivity analysis

Figure-2: Following graph represents distribution of 'COND-E (input conductivity to plant)', 'COND-P (input conductivity to primary settler)', 'COND-D (input conductivity to secondary settler)', 'COND-S (output conductivity)'.

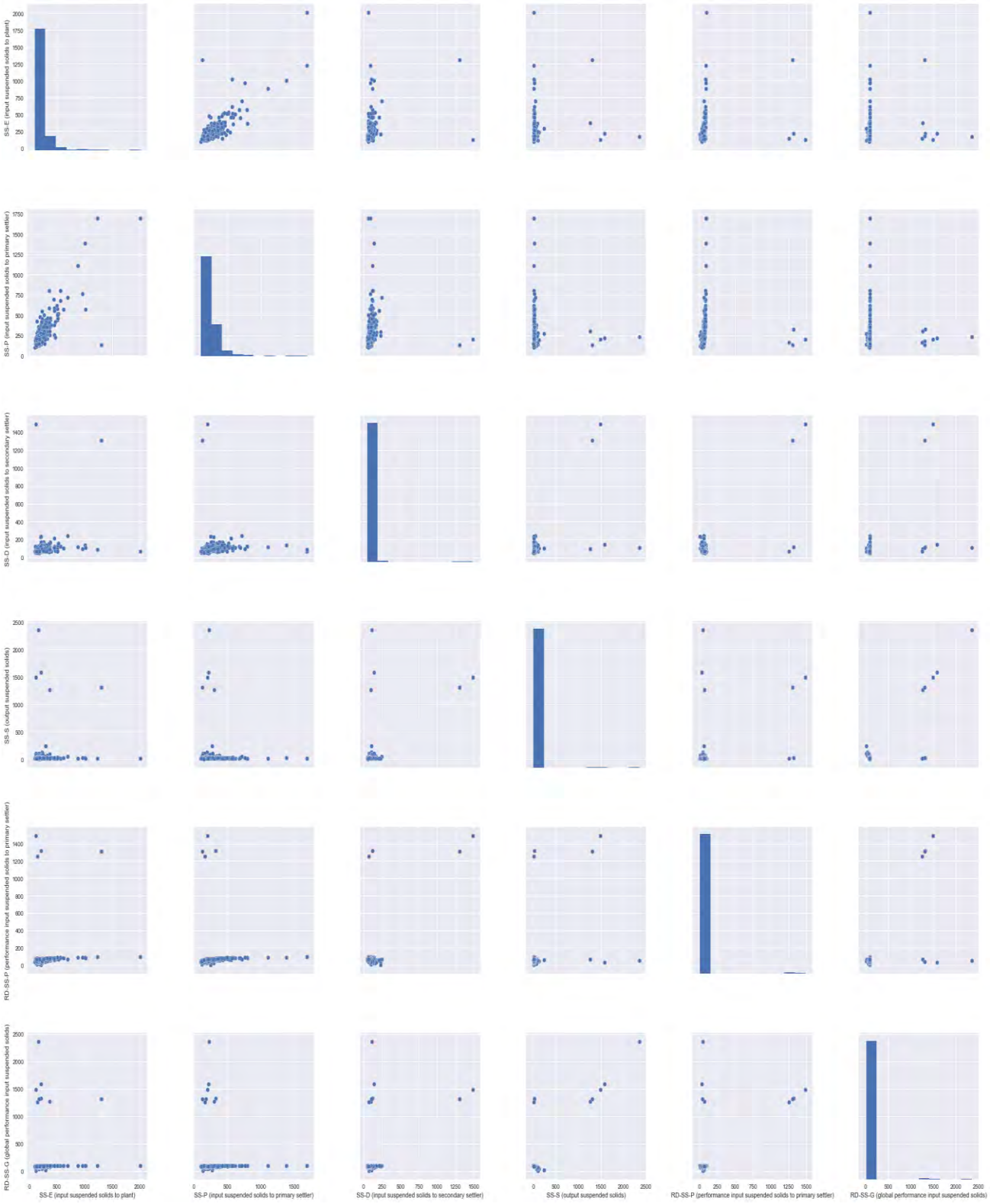


The above picture represents a scatter plot matrix which depicts variable relationships among each other. In the above picture it shows how 'COND-E (input conductivity to plant)', 'COND-P (input conductivity to primary settler)', 'COND-D (input conductivity to secondary settler)', 'COND-S (output conductivity)' these four variables are distributed. Here every single charts represents two variables'. here we can see 16 charts because it is a representation of 4x4 matrix.

Now other charts represent one variable against another variable such as: Look at top right corner on the graph here X axis is 'COND -S (output conductivity)' and left top corner represents Y axis which is 'COND -E (input conductivity to plant)'. From this graph distribution we can clearly visualize that the conductivity values are within 700-3000 range and most of them are distributed among 1000-2000 range except very few which are distributed out of 3000 range. We are assuming these few values are faulty data and these data are generated because of machineries fault during the peak hour.

4.4 Suspended Solid Analysis

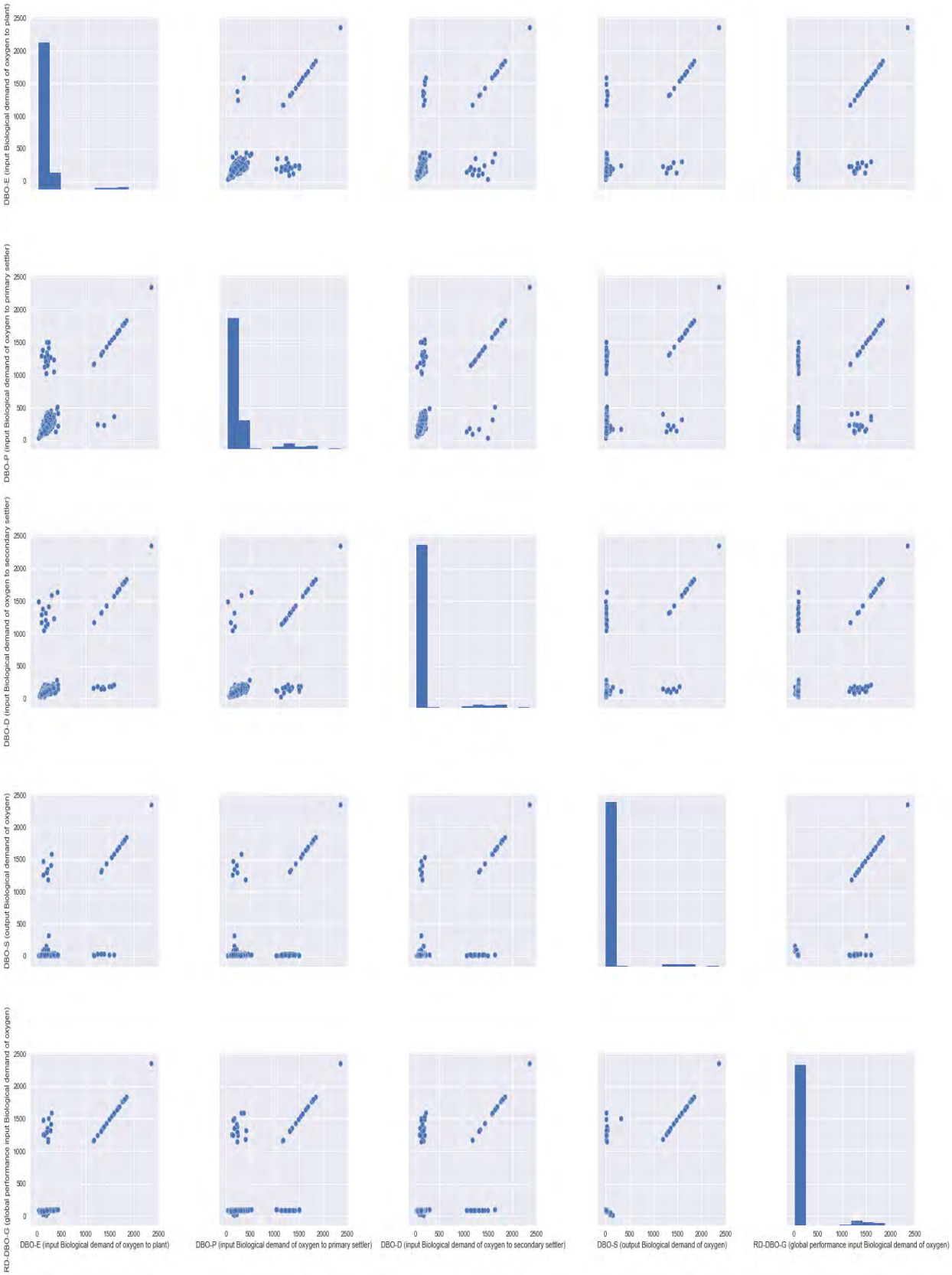
Figure-3: Following graph represents distribution of 'SS-E (input suspended solids to plant)','SS-P (input suspended solids to primary settler)', 'SS-D (input suspended solids to secondary settler)', 'SS-S (output suspended solids)", 'RD-SS-G (global performance input suspended solids)'


SS-E (input suspended solids to plant)
SS-P (input suspended solids to primary settler)
SS-D (input suspended solids to secondary settler)
SS-S (output suspended solids)
RD-SS-P (performance input suspended solids to primary settler)
RD-SS-G (global performance input suspended solids)
SS-E (input suspended solids to plant)
SS-P (input suspended solids to primary settler)
SS-D (input suspended solids to secondary settler)
SS-S (output suspended solids)
RD-SS-P (performance input suspended solids to primary settler)
RD-SS-G (global performance input suspended solids)

Now other charts represent one variable against another variable such as: Look at top right corner on the graph here X axis is 'RD-SS-G (global performance input suspended solids)' and left top corner represents Y axis which is 'SS-E (input suspended solids to plant)'. From this graph distribution we can clearly visualize that the output global performance input suspended solids values are within range and most of them are distributed among 0-300 range on the other side input values are within the range of 100-500 except very few which are distributed out of range. We are assuming these few values are faulty data and these data are generated because of machineries fault during the peak hour.

4.5 Biological oxygen demand analysis

Figure-2: Following graph represents distribution 'DBO-E (input Biological demand of oxygen to plant)','DBO-P (input Biological demand of oxygen to primary settler)', 'DBO-D (input Biological demand of oxygen to secondary settler)', 'DBO-S (output Biological demand of oxygen)','RD-DBO-G (global performance input Biological demand of oxygen)'.


DBO-E (input Biological demand of oxygen to plant)
DBO-P (input Biological demand of oxygen to primary settler)
DBO-D (input Biological demand of oxygen to secondary settler)
DBO-S (output Biological demand of oxygen)
RD-DBO-G (global performance input Biological demand of oxygen)
0
500
1000
1500
2000
2500

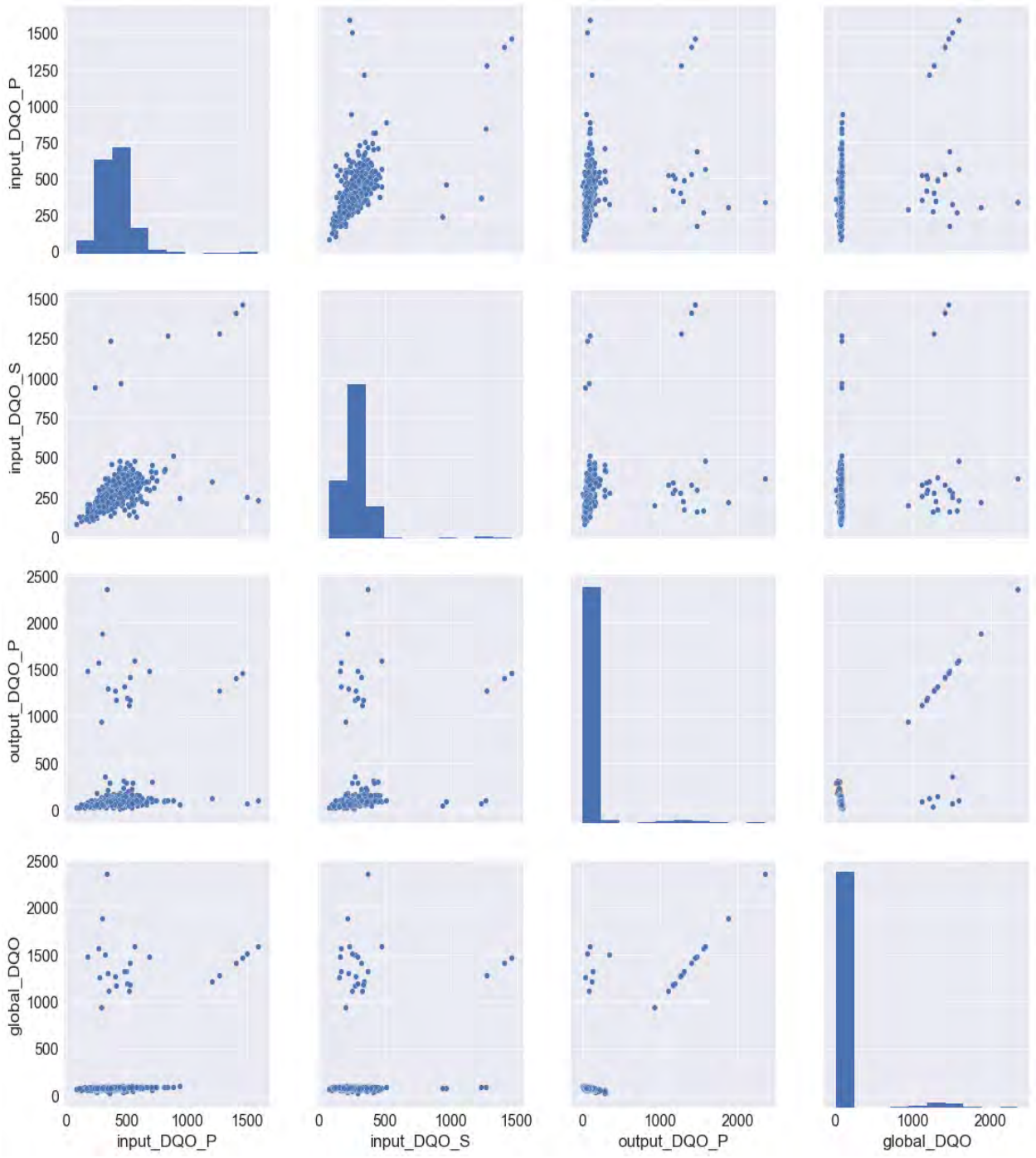
The above picture represents a scatter plot matrix which depicts variable relationships among each other. In the above picture it shows how 'DBO-E (input Biological demand of oxygen to plant)','DBO-P (input Biological demand of oxygen to primary settler)', 'DBO-D (input Biological demand of oxygen to secondary settler)', 'DBO-S (output Biological demand of oxygen)', 'RD-DBO-P (performance input Biological demand of oxygen in primary settler)','RD-DBO-S (performance input Biological demand of oxygen to secondary settler)','RD-DBO-G (global performance input Biological demand of oxygen)' these seven variables are distributed. Here every single charts represents two variables'. Here we can see 49 charts because it is a representation of 7x7 matrix.

Now other charts represent one variable against another variable such as: Look at top right corner on the graph here X axis is 'RD-DBO-G (global performance input Biological demand of oxygen)' and left top corner represents Y axis which is 'DBO-E (input Biological demand of oxygen to plant)'. From this graph distribution we can clearly visualize that the input of biological oxygen demand is within the range 0-500 and 1000-2000.However output & global performance input Biological oxygen demand values are within range of 0-250 and 1000-1800 most of them are distributed among the range except very few which are distributed out of range. We are assuming these few values are faulty data and these data are generated because of machineries fault during the peak hour.

4.6 Chemical oxygen demand analysis

Figure-2: Following graph represents 'DQO-E (input chemical demand of oxygen to plant)','DQO-D (input chemical demand of oxygen to secondary settler)','DQO-S (output

chemical demand of oxygen)', 'RD-DQO-G (global performance input chemical demand of oxygen)'.



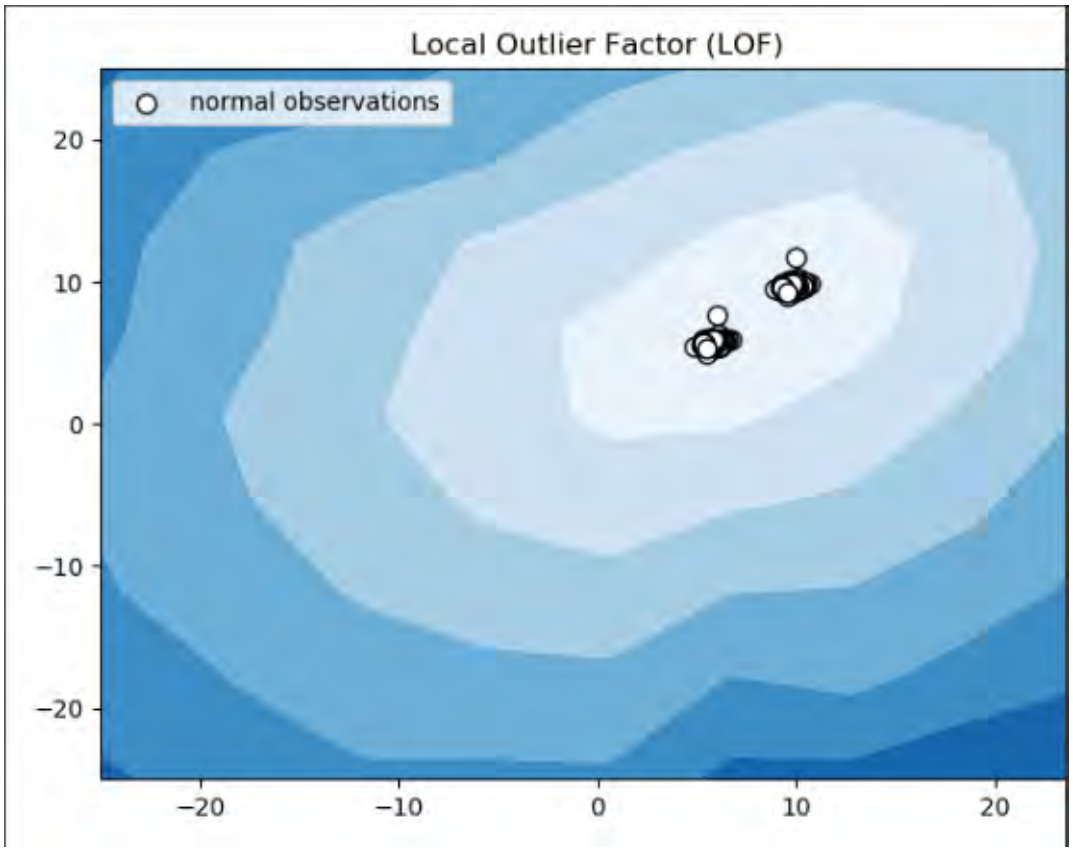
The above picture represents a scatter plot matrix which depicts variable relationships among each other. In the above picture it shows how represents 'DQO-P (input chemical demand of oxygen to plant)','DQO-S (input chemical demand of oxygen to secondary settler)', 'DQO-S (output chemical demand of oxygen)', 'RD-DQO-S (performance input chemical demand of oxygen to secondary settler)', 'RD-DQO-G (global performance input chemical demand of oxygen)' these five variables are distributed. Here every single charts represents two variables'. Here we can see 25 charts because it is a representation of 5x5 matrix.

Now other charts represent one variable against another variable such as: Look at top right corner on the graph here X axis is 'RD-DQO-G (global performance input chemical demand of oxygen)' and left top corner represents Y axis which 'DQO-E (input chemical demand of oxygen to plant)'. From this graph distribution we can clearly visualize that most of the output and global performance input chemical demand of oxygen values are within range of 0-250 whereas most of the input chemical demand oxygen values are within the range of 100-700 except very few which are distributed out of range. We are assuming these few values are faulty data and these data are generated because of machineries fault during the peak hour.

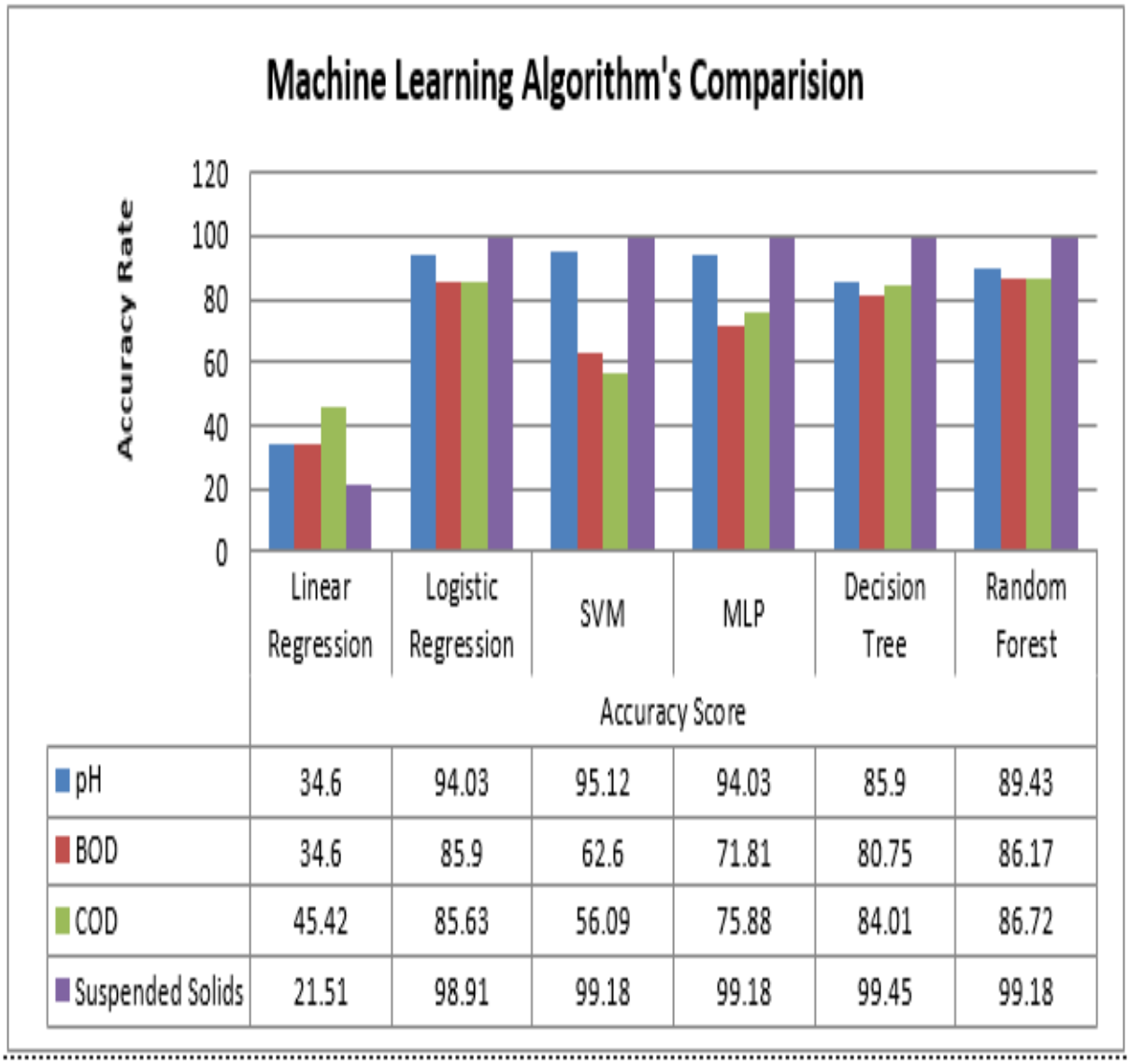
Chapter 5

Results

Here in this picture we have received the effects of pH which we can easily describe the waste water treatment plant output pH density. Here the highly dense area represents good data or its near standard pH value, on the other hand the value is away from dense area represents defected data or pH value is high or low than the standard value of pH.



Algorithms Comparison



	Linear Regression	Logistic Regression	SVM	MLP	Decision Tree	Random Forest
pH	34.6	94.03	95.12	94.03	85.9	89.43
BOD	34.6	85.9	62.6	71.81	80.75	86.17
COD	45.42	85.63	56.09	75.88	84.01	86.72
Suspended Solids	21.51	98.91	99.18	99.18	99.45	99.18

Table-2: Algorithm Comparison

From the algorithms Comparison we can identify that different algorithm shows best accuracy result for different parameters. For ph SVM(support vector machine) algorithm is best suited. On the other hand for BOD and COD random forest algorithm gives the best result. In case of suspended solids three algorithm SVM, MLP and random forest shows same accuracy rate. In future we are considering to develop a single algorithm which will be best fitted for every different parameter.

Conclusion

In this research we are classifying different parameter sets. Different parameters have also some sub category. For example: PH, DBO, DQO, SS these parameters have sub category like input PH, primary settler PH, secondary settler PH, output PH. So we have applied different algorithm for different parameters. And at the end, we found that there is no such classifier which is best every single scenario. We found that different classifier is best for different set of parameters. So later on, to find out an accurate result from a single classifier we have applied Voting classifier method. The technique of voting classifier is it will always choose the best classifier which can give us the best accuracy for a certain data set. That's why we don't need to rely on a particular single classifier. So we can work with multiple classifier and whoever give us the best accuracy, Voting classifier will choose that classifier. That's how we can get our most optimal efficient classification algorithm. Identifying machineries fault of a waste water treatment plant by statistical analysis and machine learning is the principle objective of this research. Moreover analyzing a large data sets is very complex for human being which can be easily done by clustering and machine learning technique. There is a disadvantage of this technique because it is totally dependent on data's. So accuracy might not be satisfactory if any abnormality of data or any values are missing. To overcome this problem we replaced missing values with each parameter's average value. This experiment can be done by using various programming language. In our thesis we use Python. A high correlation coefficient between input and output variable indicates the result of this study. In future other technique and algorithm can be used for developing accuracy of this model.

References

[1] Benazzi, F., Gernaey, K. V., Jeppsson, U., Katebi, R., 2007. On-line estimation and detectionOf abnormal substrate concentrations in WWTPS using a software sensor: A benchmark Study. Environmental Technology 28 (8), 871–882.

[2] Bernholt, T., Fried, R., Gather, U., Wegener, I., 2006. Modified repeated median filters. Statistics and Computing 16 (2), 177–192.

[3] Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

[4] Carstensen, J., Harremo¨es, P., Strube, R., 1996. Software sensors based on the grey-box Modelling approach. Water Science and Technology 33 (1), 117–126.

[5] Carstensen, J., Madsen, H., Poulsen, N. K., Nielsen, M. K., 1994. Identification of wastewater treatment processes for nutrient removal on a full-scale WWTP by statistical methods Water Research 28 (10), 2055–2066.

[6] Cecil, D., Kozlowska, M., 2010. Software sensors are a real alternative to true sensors. Environmental Modelling& Software 25 (5), 622–625.

[7] C´er´eghino, R., Park, Y.-S., 2009. Review of the Self-Organizing Map (SOM) approach in Water resources: Commentary. Environmental Modelling & Software 24 (8), 945–947.

[8] Ravi Kumar, P, Liza Britta Pinto, Somashekar, R.K. (November 2010), Assessment of the Efficiency Of Sewage Treatment Plants: A comparative study between Nagasandra and Mailasandra Sewage Treatment Plants, Kathmandu University Journal of Science, Engineering and Technology Vol. 6, No. II, pp 115-125.

[9] Victor Chipofya, Andrzej Kraslawski and YuryAvramenko (June 2010), Comparison of pollutant levels in effluent from wastewater treatment plants in Blantyre

[10] MalawiInternational Journal of Water Resources and Environmental Engineering Vol. 2(4), pp. 79-86

[11]Nobel Francisco Rovirosa Morell 1997, Performance Evaluation of an on-site domestic sewage treatment plant for individual residences.

[12]Majed M. Ghannam 2006, Performance Evaluation of Gaza Waste water Treatment Plant, Islamic University-Gaza

[13] E. Awuah& K. A. Abrokwa 2008, Performance evaluation of the USAB sewage treatment plant at James Town (Mudor), Accra, 33rd WEDC International Conference, Accra, Ghana.

[14]Sushil Kumar Shah Teli, (December 2008), Performance Evaluation of Central Wastewater Treatment Plant: a Case Study of Hetauda Industrial District, Nepal, 36/Environment and Natural Resources Journal Vol.6, No.2.

[15] Al-Zboon, Kamel and Al-Ananzeh, Nada (August 2008), Performance of wastewater treatment plants in Jordan and suitability for reuse, African Journal of Biotechnology Vol. 7 (15), pp. 2621-2629.

[16] YahayaMijinyawa and Nurudeen Samuel Lawal (June 2008), Treatment efficiency and economic benefit of Zartech poultry slaughter house waste water treatment plant, Ibadan, Nigeria, Scientific Research and Essay Vol. 3 (6), pp. 219-223.