Influent And Effluent Parameters Prediction In A Wastewater Treatment Plant

A

Project Report

Submitted for the partial fulfilment

of B.Tech. Degree

in

COMPUTER SCIENCE & ENGINEERING

by

Aditya Chaudhary	**1805210004**
Nitish Kr. Chaudhary	**1805210031**
Priyanshu Sharma	**1805210039**
Anand Keshari	**1900520109002**

Under the supervision of

Dr. Parul Yadav

Mr. Sandeep Yadav

Department of Computer Science and Engineering

Institute of Engineering and Technology

Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh.

June, 2022

Contents

Declaration

We hereby declare that this submission is our own work and that, to the best of our belief and knowledge, it contains no material previously published or written by another person or material which to a substantial error has been accepted for the award of any degree or diploma of university or other institute of higher learning, except where the acknowledgement has been made in the text. The project has not been submitted by us at any other institute for requirement of any other degree.

Submitted by: - Date: 20 June 2022

(1) Name: Aditya Chaudhary
Roll No.: 1805210004
Branch: Computer Science and Engineering
Signature:

(2) Name: Nitish Kumar Chaudhary
Roll No.: 1805210031
Branch: Computer Science and Engineering
Signature:

(3) Name: Priyanshu Sharma
Roll No.: 1805210039
Branch: Computer Science and Engineering
Signature:

(4) Name: Anand Keshari
Roll No.: 1900520109002
Branch: Computer Science and Engineering
Signature:

Certificate

This is to certify that the project report entitled "**Influent and Effluent Parameters Prediction In A Wastewater Treatment Plant**" presented by *Aditya Chaudhary*, *Nitish Kumar Chaudhary*, *Priyanshu Sharma* and *Anand Keshari* in the partial fulfillment for the award of Bachelor of Technology in Computer Science and Engineering, is a record of work carried out by them under my supervision and guidance at the Department of Computer Science and Engineering at Institute of Engineering and Technology, Lucknow.

It is also certified that this project has not been submitted at any other Institute for the award of any other degrees to the best of my knowledge.

Dr. Parul Yadav
Department of Computer Science and Engineering
Institute of Engineering and Technology, Lucknow

Mr. Sandeep Yadav
Department of Computer Science and Engineering
Institute of Engineering and Technology, Lucknow

Acknowledgement

We would like to express our sincere gratitude and appreciation to all those who gave us the opportunity to complete this project. First, we wish to express our sincere gratitude to our supervisors, Dr. Parul Yadav and Mr. Sandeep Yadav, for their insightful comments, helpful information and practical advice that have helped us tremendously at all times in my research and writing of this project. Without their support and guidance, this project would not have been possible. We would like to express gratitude to all our friends who motivated us and helped us at each step in completing this project work. We would also like to express our sincere regards to all the Authors of all the references and other literary work referred to in this project work.

(1) Name: Aditya Chaudhary
Roll No.: 1805210004
Branch: Computer Science and Engineering
Signature:

(2) Name: Nitish Kumar Chaudhary
Roll No.: 1805210031
Branch: Computer Science and Engineering
Signature:

(3) Name: Priyanshu Sharma
Roll No.: 1805210039
Branch: Computer Science and Engineering
Signature:

(4) Name: Anand Keshari
Roll No.: 1900520109002
Branch: Computer Science and Engineering
Signature:

Abstract

A rise in the population of a region implies an increase in water consumption and such a continuous increase in the usage of water worsens wastewater treatment in the region. This escalation in wastewater (influent) requires the Wastewater Treatment Plants (WWTPs) to operate efficiently in order to process the demand for sewage disposal (effluent). This project is based upon visualizing, analyzing and building prediction models for the parameters of influent like COD, BOD, TSS, pH, MPN and also, the parameters of effluent like COD, BOD, DO, pH and MPN of 345 MLD UASB-based Bharwara STP/WWTP situated in Lucknow, Uttar Pradesh, India which is the largest UASB-based wastewater treatment plant in Asia.

We have designed and implemented some models using the machine learning based techniques to analyze as well as predict the parameters of influent and effluent of the WWTP. Model Performance is measured using Mean Squared Error (MSE) and Correlation Coefficient (R). For analyzing and designing the model, the parameters of influent and effluent have been collected over a period of 38 months on a daily basis covering the variations between seasons and climate. As a result, the model shall provide a better quality of effluent along with consuming the plant resources in an efficient manner. We had initially conducted our research using Linear regression algorithm and we have found that the model was presenting adequate results. For further improvements, we have incorporated KNN(K-Nearest Neighbours) and ANN (Artificial Neural Network) models for effluent parameter prediction and ARIMA(Auto Regressive Integrated Moving Average) for influent flow in our study which can find more robust relationships between parameters and time series analysis.

List of Figures

List of Tables

Chapter 1
Introduction

Wastewater Treatment Plants (WWTPs) play a crucial role in shaping the urban and rural environments as they are used for processing sewage water and removal of various particles and chemicals which are harmful for the water hydrosphere and the organisms which are dependent on it. An increase in the population of a region implies an increase in water consumption and such a continuous increase in the usage of water results in an increase in the wastewater generated by the region **[1]**. This increase in influent requires the wastewater treatment plants to operate efficiently in order to process the demand for effluent (sewage disposal) **[2, 3, 4]**.

Besides increase in influent, another more challenging issue in a wastewater treatment plant is the fluctuating or uncertain behaviour of various parameters of the influent in the plant which can be due to varying environmental factors also **[5]**. To maintain the effluent parameters within the standard range, the wastewater treatment plants need to operate and process on the influent coping up with its varying parameters. On the other side, the wastewater treatment plants require to do optimum utilization of resources during the treatment of influent. Consequently, this uncertain nature of influent parameters demands to find insights and hidden patterns by applying visualization and analytics on the real time historical/ recorded data which in turn shall help to provide/estimate better and efficient (optimized) utilization of resources at wastewater treatment plants. Further knowing the flow and parameters of influent and parameters of effluent in advance shall reduce operational cost of the wastewater treatment plants.

This project is based upon designing and implementing machine learning-based models for analyzing and predicting flow and quality parameters of influent like COD, BOD, TSS, pH, MPN and also, the parameters of effluent like COD, BOD, DO, pH and MPN of Bharwara WWTP situated in Lucknow, India. The designed model shall provide support to centrally monitor processes and operations of wastewater treatment plants. This project shall improve operational efficiency and provide cost-effective utilization of various resources at wastewater treatment plants by knowing the influent and effluent parameters in advance.

Our work demonstrates techniques by which we can monitor the concentration of influent and effluent particles, find the relation between influent and effluent particles, determine the factors which are the cause of varying efficiency of the plant, and propose a model which will provide us a better estimation based on effluent concentration.

Initially, we had conducted our research using Linear regression algorithm and we have found that the model was presenting adequate results. For further improvements, we are trying to incorporate KNN (K-Nearest Neighbours) and ANN (Artificial Neural Network) models in our study which can find more robust relationships between parameters and shall give us a better estimate than Linear Regression based model. Surveys suggested that almost 70% of WWTPs failed because of Influent fluctuations. For monitoring the dynamic changes in the influent quality and quantity we have used ARIMA as a forecasting method for influent flow in our study which can find more robust relationships between parameters and time series analysis.

With this objective, we have collected and recorded the water parameters for over 38 months (April 2019 to May 2022) from Bharwara Wastewater Treatment Plant situated in Lucknow district which is the largest UASB based wastewater treatment plant in Asia as it can operate and process an average flow rate of 345 MLD (Million Litres per Day) with the ability to handle a peak load of 517 MLD of sewage daily. However, the implemented model shall be applicable for any UASB based wastewater treatment plant or any wastewater treatment plant after a specific training part or maybe after minor model refinements.

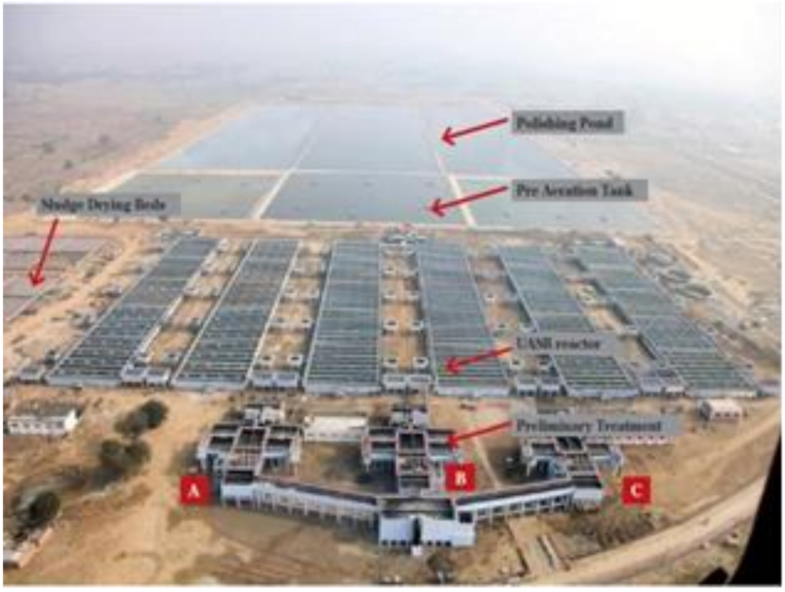

Fig 1.1: *Aerial photograph of Bharwara WWTP*

Project Objectives

- Studying existing tools, methods for predicting wastewater parameters in wastewater treatment plant (WWTP)
- Analysis and visualisation of data from a WWTP
- Design and implementation of model to predict effluent parameters in WWTP
- Design and implementation of model to predict influent flow in WWTP

Challenges

- Data collection
- Irregularities in data
- Analyzing hidden details

Chapter 2
Literature Review

We have carried out the literature survey in the line of the project under two dimensions. The first dimension of the study is in line with predicting effluent parameters for wastewater treatment plants outside and inside India, and the second dimension of the study is in line with analyzing and predicting influent parameters for wastewater treatment plants outside and inside India. We shall discuss both dimensions one by one.

2.a International Status-

Work done for Predicting Effluent Parameters:

i. Konya Wastewater treatment plant [Konya, Turkey]:

In **[2]**, an artificial neural network was used to propose a model for the prediction of Total Suspended solids based on the input parameters COD, BOD, TSS. Model performance was evaluated via Mean Squared Error and Correlation Coefficient (R) for the Konya Wastewater treatment plant. Neural Networks of various hidden layers were used and the correlation coefficient in the training set reached up to 0.99, a satisfactory result from the proposed model.

ii. Wastewater treatment plant in Korea:

In **[3]**, ANN and SVM models were proposed to predict the Total Nitrogen (T-N) concentration in the plant. For evaluation of the model, Coefficient of Determination(R^2), Nash-Sutcliff efficiency, and relative efficiency criteria were used. A sensitivity analysis was done using a pattern search algorithm and Latin Hypercube One factor At a Time (LH-OAT) **[4]** which showed that the ANN model gave the superior result as compared to the SVM model.

iii. Wastewater treatment plant in Italy

In **[6]**, a study was conducted on stormwater discharge and a model was proposed for estimation of COD, BOD, TSS, and TDS in the wastewater. Support Vector Regression and Regression Tree algorithm were used for modeling and Coefficient of determination(R^2) and Root Mean Squared Error (RMSE) were the performance evaluators. For COD, TSS and TDS, the SVR model performed better than the Regression tree while for BOD, Regression Trees Gave better results than SVR.

iv. Wastewater treatment plant in Hong-Kong

In **[7]**, it was shown that wastewater quality can be monitored online. UV/VIS spectrometry and a turbidimeter were used to monitor COD, TSS, and O&G concentrations. Sensor fusion technique was used to fuse the signals from the two sensors. Boosting-Partial Least Squares (Boosting-PLS) method was used to make the model and predict the wastewater quality based on the fused information.

Work done for Analyzing/ Predicting Influent Parameters:

v. Gongxian Wastewater Treatment Plant in Yibin, China

In **[8]**, four machine learning methods of Linear Regression, Ridge, ElasticNet, and Lasso were used for predicting the influent quality. For influent parameter predictions, different methods showed high accuracy for different parameters. The results published in the reference used these models as warning modules for assisting in the daily operations of WWTP.

2.b National Status –

Work done for Predicting Effluent Parameters:

i. Wastewater treatment plant in Mangalore:

In this study, an Artificial Intelligence-based model was used to predict the performance of a treatment plant for the removal of effluent nitrogen particles. Three different models, SVM, ANFIS trapezoidal MF model and ANFIS Gbell MF model were made in matlab. Influent parameters taken were pH, ammonia nitrogen, free ammonia, and Kjeldahl nitrogen. Performance evaluation was done by RMSE, NSE, and Correlation Coefficient(R). networks SVM model gave satisfactory results.**[9]**

Work done for Analyzing/ Predicting Influent Parameters:

ii. 345 UASB Bharwara STP

This study focused on the working performance of STP and upgrading UASB reactor technology. The removal efficiency of COD, BOD, and TSS was measured and the relation between pH and influent parameters was determined.**[10]**

iii. Sewage Treatment Plant in Delhi

This study focused on the monitoring of inlet and outlet parameters and measuring the effectiveness of STP. The cluster Analysis approach was performed to find any relation between the current site and other sites, aiming to find similar sites. Sulfate, Nitrates, Chloride and Phosphate, and Bi-carbonates concentrations were measured and the results showed that STP efficiency was not up to the mark.**[11]**

Chapter 3

Methodology

Our project can be broadly divided into three parts considering the completion of our three objectives:

- Pre-processing, visualisation and analysis of the collected real time data.
- Design and implementation of model to predict effluent parameters in WWTP.
- Design and implementation of model to predict influent parameters in WWTP.

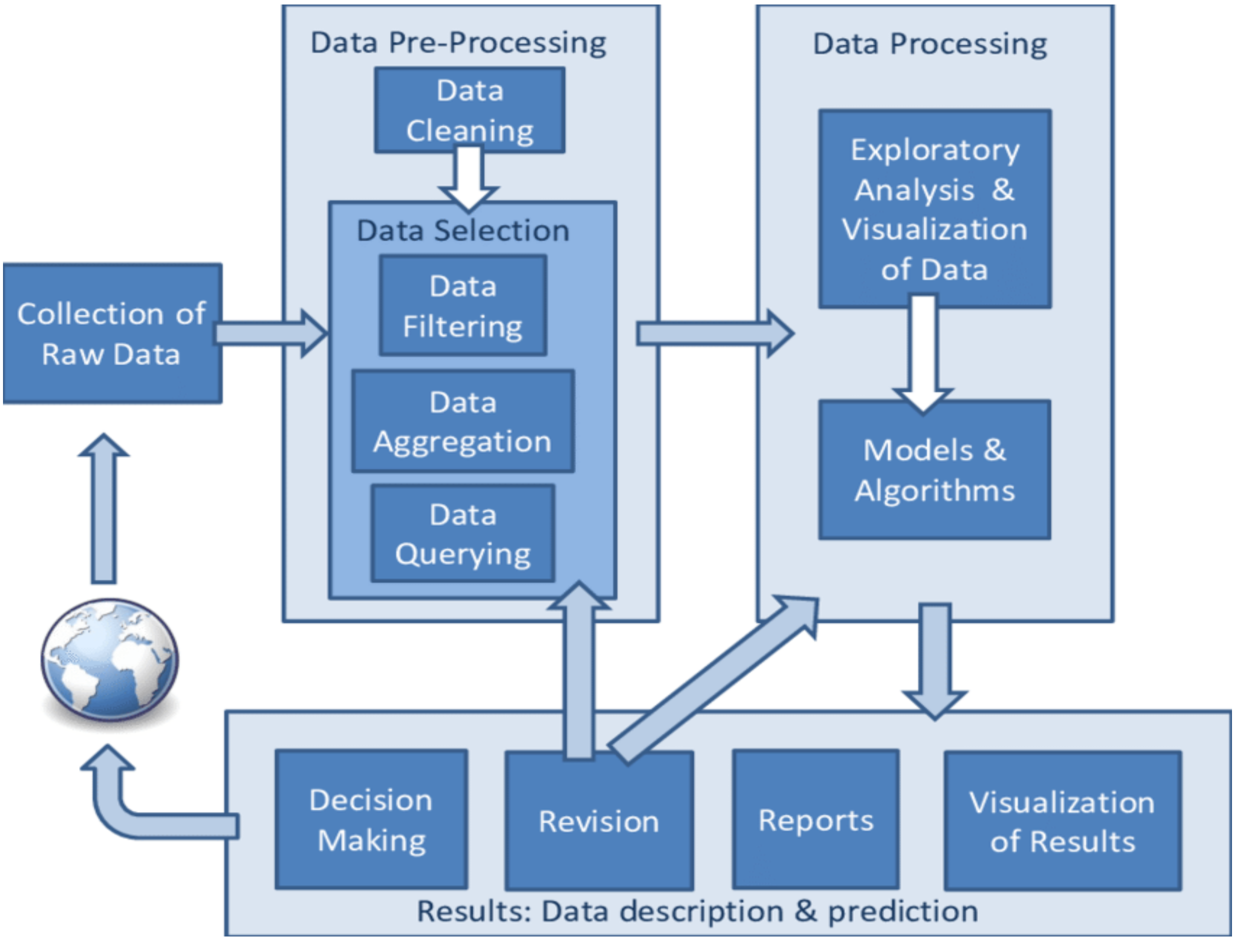

Fig 3.1 *Data Collection and Processing*

3.1 Collecting and Processing the data

We have designed machine learning based models to predict parameters of influent and effluent which shall provide efficient utilization of chemical resources during treatment process ensuring the desired level of quality indicators in effluent. We have collected a real time dataset of 345 MLD UASB-based Bharwara STP using manual process for data analysis. The methodology for the model is briefed using the following four steps:

Table 3.1: Locations and Measuring Parameters

Location	**Parameters**
Inlet Chamber	BOD, pH, Suspended Solids, Temperature, COD, oil, flow, Phosphorous, DO
Outlet of UASB Reactor	BOD, Suspended Solids, pH, COD
Polishing pond	Dissolved Oxygen, pH
Outlet of Chlorine contact Tank	BOD, Suspended solids, pH, COD, Fecal Coliform, Residual Chlorine, Dissolved Oxygen.
Primary sludge	pH, Total Solids, Volatile solids.

1. **Identification of locations and water parameters to be captured at Plant:**

We along with supporting staff at 345 MLD UASB-based Bharwara STP identified five locations where the water parameters are to be captured. The placing of various locations in the plant are shown in Figure 3.2. At each location, we identified and listed the water parameters like BOD, COD, DO, SS, temperature, pH, Residual Chlorine etc. be measured. The basis of identifying water parameters at a particular location in the plant is the process/ treatment/ chemical reactions taking place at these locations. These identified locations and respective parameters to be measured at these locations are listed in Table 3.1.

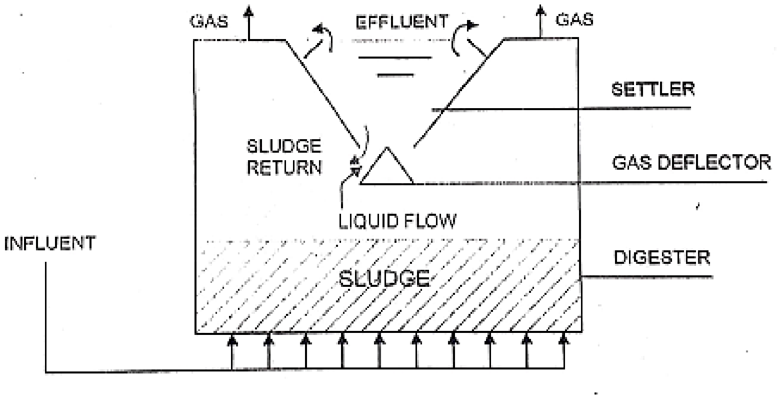

Fig 3.2: *Schematic of a UASB Reactor*

2. **Data Collection:** We collected a real-time data set of the 38 months (April 2019 to May 2022) from the plant. In the data set, selected parameters of influent and effluent are collected/ captured and recorded using manual process adopted at the plant.

Table 3.2 Description of Dataset

Data Quality Parameters	**Units**	**Range (Influent)**	**Range (Effluent)**
pH	No.	6-8	7-9
Dissolved Oxygen (DO)	mg/l	0	>4
Total Suspended Solids (TSS)	mg/l	300-600	<50
Chemical Oxygen Demand (COD)	mg/l	200-500	<100
Biological Oxygen Demand (BOD)	mg/l	150-250	<30
Most probable number (MPN)	No./100ml	106 - 109	106 - 109
Flow Rate	Millions of Litre per Day	250-400	

3. **Data Preprocessing:** We have done pre-processing on the recorded data set. For preprocessing, we treat missing values and outliers using standard procedures and kNN, and further normalized the data set. The outlier treatment is performed using statistical techniques i.e., calculating interquartile range and neglecting the values above lower limit and upper limit **[12]**. The normalization of the data set is performed using the following formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x’ is the normalized value, x is the original value, and min(x) and max(x) respectively are the minimum and maximum values. The data is normalized in the range between 0 and 1.

TABLE 3.3 SUMMARY OF DATASET BEFORE PREPROCESSING

	MLD	INFLUENT PH	EFFLUENT PH	INFLUENT TSS	EFFLUENT TSS	INFLUENT COD	EFFLUENT COD	INFLUENT BOD	EFFLUENT BOD	INFLUENT MPN	EFFLUENT DO
COUNT	1138	1138	1138	1138	1138	1138	1138	1127	1127	927	1138
MEAN	329.95	7.34	7.57	227.48	41.75	277.42	72.53	141.15	25.39	416937.5	4.53
STD	46.542	0.120	0.102	38.16	3.92	46.87	9.86	22.37	2.26	712074.1	1.178
MIN	115.16	6.93	7.23	139	4	30	6	75	18	1.4	4
25%	309.69	7.26	7.52	204	39	248	64	128	24	14	4.3
50%	339.05	7.34	7.59	225	42	272	72	140	26	25.7	4.4
75%	357.76	7.42	7.65	251	45	304	80	155	27	910000	4.7
MAX	460.06	7.8	7.87	556	72	528	96	450	29	14000000	43

4. **Discovering Unknown Patterns:** We discover various patterns or relations within the collected data sets. We visualize the patterns in the data set. We design and implement a machine learning-based model to analyze and predict the parameters of influent/ effluent in the wastewater treatment Plant.

The raw data (Table 3.3) has thus been treated and pre-processed to remove outliers and impute missing values which will help in the performance of our predictionary models. We can see the summary of the pre-processed data in Table 3.4.

TABLE 3.4 SUMMARY OF DATASET AFTER PREPROCESSING

	MLD	INFLUENT PH	EFFLUENT PH	INFLUENT TSS	EFFLUENT TSS	INFLUENT COD	EFFLUENT COD	INFLUENT BOD	EFFLUENT BOD	INFLUENT MPN	EFFLUENT MPN	EFFLUENT DO
COUNT	1126	1126	1126	1126	1126	1126	1126	1126	1126	1126	1126	1126
MEAN	0.549	0.49	7.58	0.48	41.77	0.51	72.68	0.49	25.562	0.389	0.067	4.49
STD	0.171	0.17	0.093	0.188	3.64	0.219	9.53	0.18	1.985	0.190	0.013	0.286
MIN	0	0	7.33	0	30	0	40	0	20	0	0.035	4
25%	0.56	0.37	7.53	0.367	39	0.388	64	0.39	24	0.239	0.06	4.3
50%	0.594	0.5	7.59	0.479	42	0.49	72	0.485	26	0.317	0.068	4.4
75%	0.62	0.609	7.65	0.635	45	0.63	80	0.62	27	0.494	0.078	4.6
MAX	1	1	7.82	1	49	1	96	1	29	1	0.1	5.3

3.2 Model for influent flow prediction

We have designed and implemented a machine learning model to predict the flow rate of influent water in the wastewater treatment plant. Fig 3.3 shows overview of process flow which was carried out to develop the influent flow prediction model.

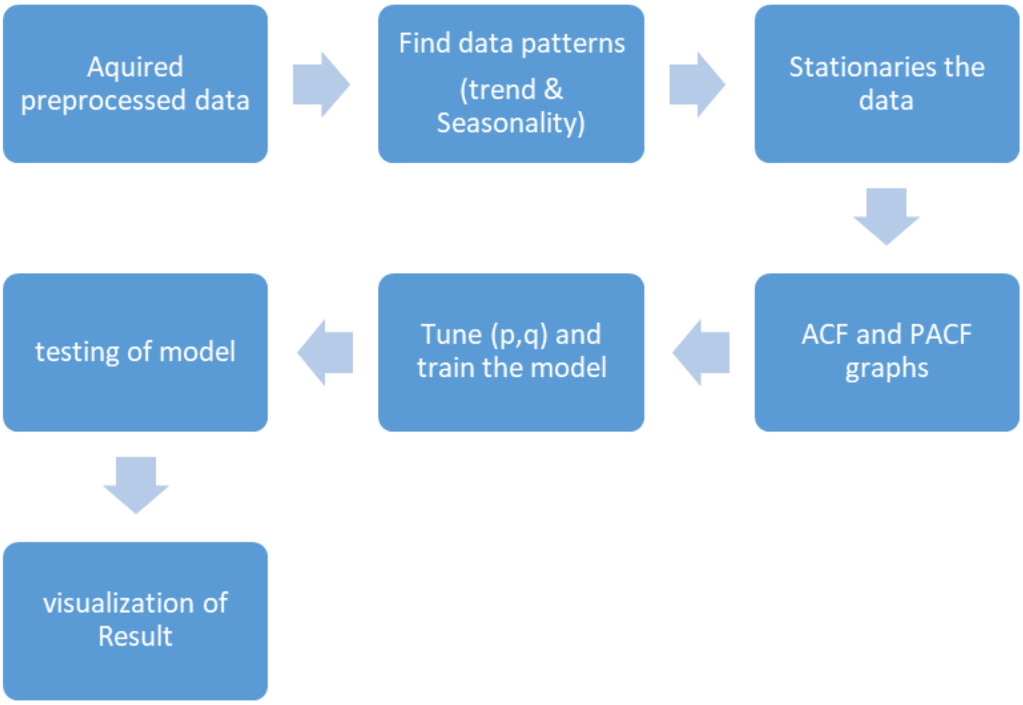
Aquired preprocessed data
Find data patterns (trend & Seasonality)
Stationaries the data
ACF and PACF graphs
Tune (p,q) and train the model
testing of model
visualization of Result

Fig 3.3: *Process flow for Influent Flow Prediction Model*

Fig. 3.4 shows the flow rate (in MLD) of influent with respect to month. In Fig. 3.5, we can clearly observe the inconsistency in the flow. Fig. 3.6 shows day wise flow of influent in the month of January of 2020 and 2021. For the month of February 2020, the flow is around the 7000 million litres but it rises too nearly 12000 million litres in the month of July. Also, the same months of different years have shown the major differences in the data which can be clearly observed in Fig. 3.6.

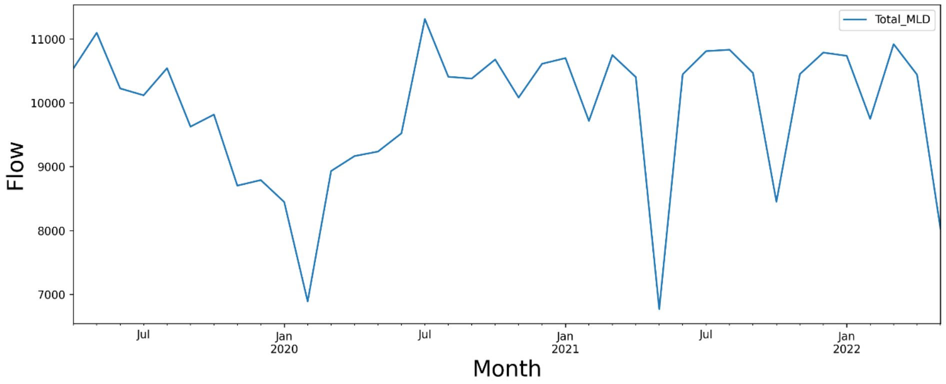

Fig. 3.4: *Flow Rate of Influent by Month*

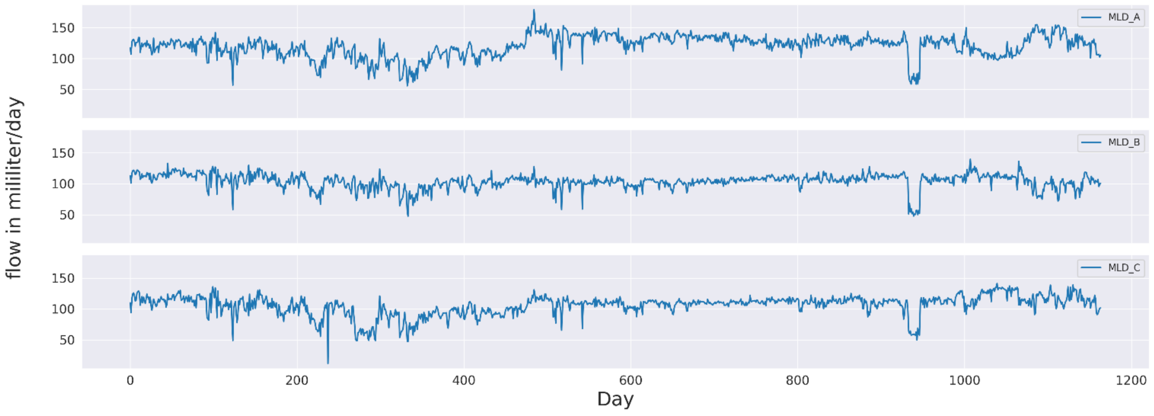

Fig. 3.5: *Flow rate of Influent by day*

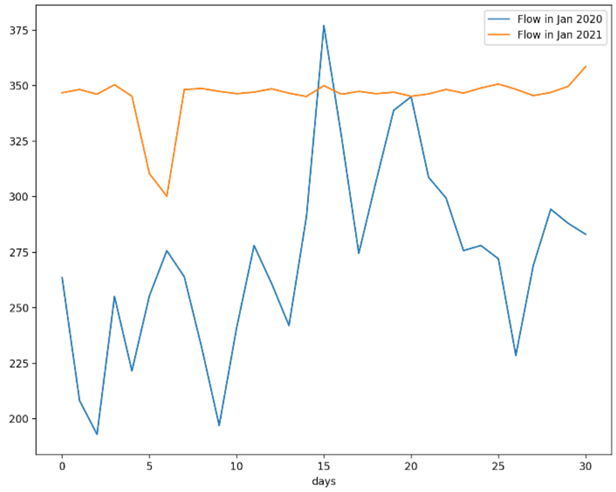

Fig. 3.6: *Comparison of flow rate between Jan 2020 and Jan 2021 by day*

This inconsistent flow of wastewater into the plant is the cause of inefficient treatment of the waste contained within the water and thus its prediction can help WWTP so that the plant is ready to handle peaks and troughs of wastewater generation. To predict such an inconsistent flow, use of time series analysis algorithms is necessary. Our machine learning based prediction model is based on SARIMA (Seasonal Auto Regressive Integrated Moving Average).

Time Series: A time series is a successive order of data points occur over a period of time. The purpose of using the time series analysis is to find the patterns of the historical data and to analyze the dependency of the future observations on the previously recorded data and make predictions for the future events. Shortly the time series patterns can describe two primary components of the data that depends on time are trends and seasonality.

Trend: Trend is the pattern in a time series data that represents the movement if a series to relatively higher or lower values over a long period of time. In general, the trend is an observed increase or decrease in the data over a period of time.

Seasonality: The almost repetition of a section of the long time series data happened historically which can recur in a calendar year.

Auto Regressive Integrated Moving Average (ARIMA):

ARIMA is a statistical machine learning model which is uses the time series data to understand the data and predict the future trends based on the historical data for (t+1) time. Time series data is the data which follows the dependency with the time means it changes with time change (ex. - stock prices). The ARIMA model is the generalization of an Autoregressive Moving Average (ARMA) model where we add the integration part for making the time series data stationary, we will discuss about it in detail further.

An ARIMA model can be explained by outlining all its three components:

- Autoregressive model (AR)
- Integration(I)
- Moving Average model (MA)

Autoregressive model (AR) –

The AR component of the ARIMA indicates the changing variable of time series is regressed (linearly) on its own lagged (previous) values.

An Autoregressive model of order p in short AR (p) or ARIMA (p, 0, 0). The equation for the AR (p) model is in equation 1.

$$Z_t = \mu + Ø_1 Z_{t-1} + Ø_2 Z_{t-2} + ... + Ø_p Z_{t-p} + a_t \quad .... (1)$$

Where,

Z_t = stationary time series (observed value)
μ = constant
Z_{t-p} = independent variable (p units lagged value)
$Ø_p$ = coefficient of the autoregressive p
a_t = error value at time t

Moving Average model (MA)–

The MA component of the ARIMA model describes the effect of the regression error of lagged values on the observed time series data. A Moving Average model of order q in short MA (q) or ARIMA (0, 0, q). The equation for the MA (q) model is in equation 2.

$$Z_t = \mu + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} -.... - \theta_q a_{t-q} \quad .... (2)$$

Where,

Z_t = stationary time series (observed value)

μ = constant

a_{t-q} = independent variable (q units error value at time t)

$Ø_q$ = coefficient of the moving average q

a_t = error value at time t

ARMA –

The Autoregressive Moving Average model (ARMA) is a combined model of Autoregressive (AR) model and the moving average (MA) model. The ARMA model works on the assumption that the current data depends on the previous data.

An ARMA model of autoregressive order p and moving average order q or ARMA (p, q). The general equation for ARMA (p, q) or ARIMA (p, 0, q) is shown in equation 3.

$$Z_t = \mu + Ø_1 Z_{t-1} + Ø_2 Z_{t-2} +..... + Ø_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} -..... - \theta_q a_{t-q} \quad .... (3)$$

Integration (I) –

The ARMA model does not care about the stationarity of the data. If the time series data shows the trend, then the ARMA model fails in accuracy. So, the integrated ARMA model (ARIMA) is used based on the assumption that the time series data used must be stationary means the average variation in data must be constant. The accuracy of the prediction drops when the data is not stationary. To get rid of the unstable data, differencing of the time series data is done to make data stationary this process is called integration.

A difference of order one means that the each observed value is subtracted with its previous value to get a new time series data. Hence “d” is referred as the order of the differencing.

$Y_t = Z_t – Z_{t-1}$ .... (4)

A combination of AR (p), MA (q) and I (d) models is called an ARIMA (p, d, q).

The Process of Training an ARIMA model

ARIMA can be used for forecasting of the Stationary as well as non-stationary data, so the first step is the analysis of the data patterns in time series data. The purpose of this analysis is to find out if the data is stationary or not. If the data is not stationary then it will be transformed before training the model.

Identification of data patterns to choose the order of “d” can be done using Augmented Dickey-Fuller (ADF) test. Stationary data has the ADF absolute value higher than the test critical values. If the ADF value < critical value, then do the differencing once and do the ADF test again. If the test shows that the data is stationary then the order of the integration can be set to one else repeat the process to get the stationary results and the value of d.

After getting the stationary data we analyze the patterns in the time series data for predicting the p and q values. The tuning of the p and q value can be done using the Auto Correlation (AC) and Partial Auto Correlation (PAC) graphs.

Auto Correlation –

Auto Correlation Function calculates the correlation between the current observed value and its lagged values or it calculates the correlation between t and (t-k) period. It includes all the lagged values between this time periods. Auto correlation considers all the previous values irrespective of the effect on the present or future time period.

Partial Auto Correlation –

PACF determines the partial correlation between the current and some lagged value of the time series means the correlation between t and (t-k) time periods without taking middle values in consideration. It signifies that the current forecast can have the dependency on the 3 day or 4 day prior data and not on the yesterday’s data.

3.3 Model for Effluent quality parameters

We have designed and compared multiple machine learning models in order to select the best performing model to be implemented to predict the quality parameters of effluent water in the wastewater treatment plant. Fig 3.7 shows overview of process flow which was carried out to develop the effluent parameter prediction model.

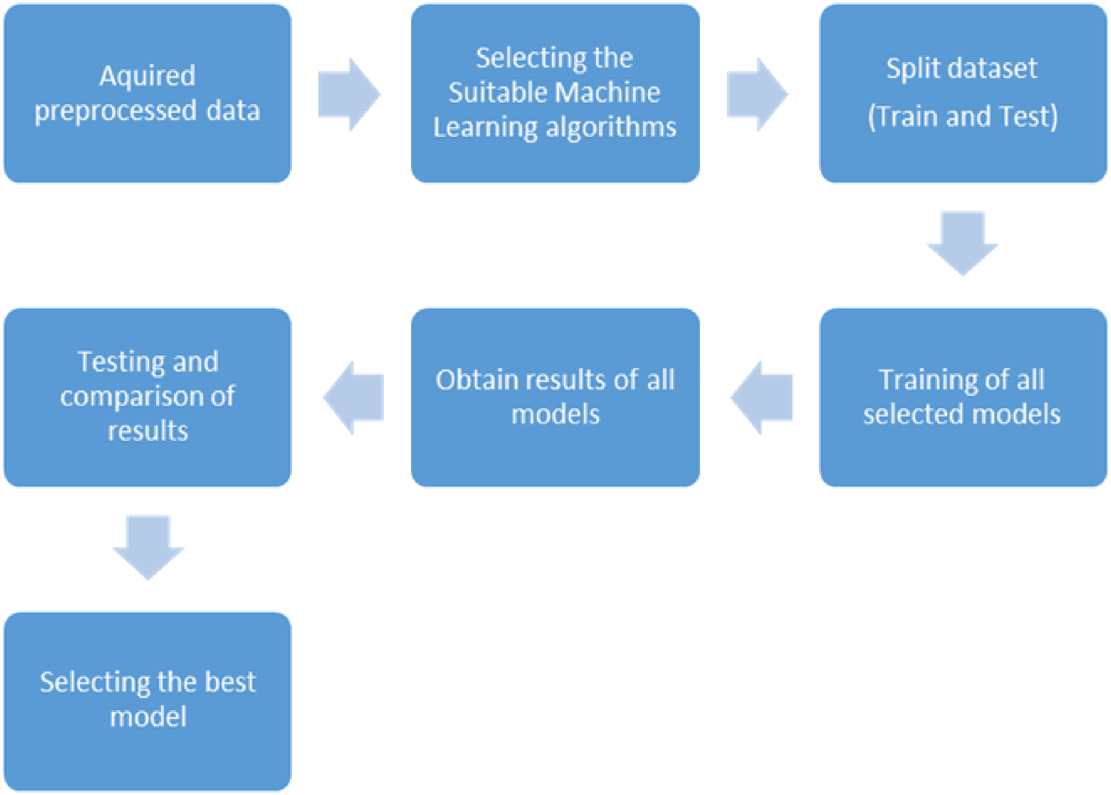
Aquired preprocessed data
Selecting the Suitable Machine Learning algorithms
Split dataset (Train and Test)
Training of all selected models
Obtain results of all models
Testing and comparison of results
Selecting the best model

Fig 3.7: *Process flow for Effluent Parameter Prediction*

Linear Regression

We used Linear Regression to design the preliminary prediction model. Linear regression **[13]** is a statistical tool for the prediction of a dependent variable from an independent variable. It establishes a linear relationship between the independent (input) and dependent (output) variables. Linear Regression is a modelling technique where a dependent variable is predicted based on the independent variables. Linear Regression is the most widely used technique among all statistical techniques. The linear regression model is designed on Google Colab using python 3.7.12 for performing analysis.

Let us discuss the dependent variable, independent variable, line of regression, data pre-processing, model properties for the linear regression model.

Dependent variable: It is a variable that depends on other factors (independent variables) that are measured.

Independent variable: It is the variable **[14]** that is stable and unaffected by another variable which we are trying to measure Independent variables (predictors) are used to predict the value of the dependent variable (target variable).

Line of regression model: It is the relationship between independent and dependent variables.

Model Properties: We implemented the initial model using Linear Regression in Python Implementation environment for the model is given in Table 3.5.

Table 3.5: Implementation Environment

Language	Python (version 3.7.12)
Tool	Google Colaboratary
Libraries	NumPy, Pandas, Matplotlib, Scikit Learn, SciPy and Seaborn

The model Properties are as follows:

- Model inputs: Inlet COD, BOD, PH, TSS, MLD and MPN
- Model outputs: Outlet COD, BOD, PH, TSS, DO, MPN
- Training - Test split: 80/20
- Estimator Function: Mean Square Error

In order to measure the performance of the model, Mean Square Error (MSE) is used. Formula for MSE is given as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

KNN

Regression method which predicts the effluent by taking influents as the predicting variables by using the minimum distance between nearest Neighbours.

Here, optimal nearest Neighbour was found to be 14 and the model performs comparatively better than linear regression

Gradient boosting tree

Regression method which predicts the effluent by taking influents as the predicting variables by using ensemble of several different decision trees where output of one layer serves as an input to other layer.

Random forest regression

Regression method which predicts the effluent by taking influents as the predicting variables by using ensemble of several different decision trees prediction effluent parallelly.

ANN

The term "Artificial Neural Network" is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons interconnected to one another, artificial neural networks also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes.

Dendrites from Biological Neural Network represent inputs in Artificial Neural Networks, cell nucleus represents Nodes, synapse represents Weights, and Axon represents Output.

Optimiser function : Adam

Learning Rate : 0.001

Cost : MeanSquaredError

Activation function: Sigmoid, Tanh, ReLU

Chapter 4

Experimental Results

This section discusses about the ways in which we have successfully computed the results and determined the effectiveness of our project.

4.1 Data Analysis/Visualisation

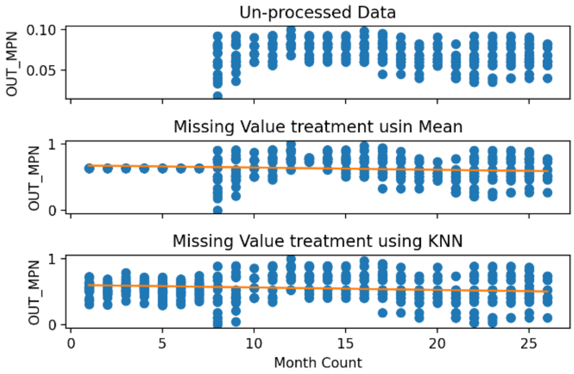

Fig. 4.1: *Preprocessing of OUT_MPN*
a) on raw data
b) Replaced missing values using mean
c) Replaced missing values using KNN

We collected the raw data from 345 MLD UASB-based Bharwara STP during April 2019 to May 2022. The summary of the collected data, analyzed firstly. Based upon the initial analysis, it is found that the data has some missing facts/ details/ values and the outliers under few variables. Therefore, we applied Mean method and KNN to treat missing values in the given dataset. Fig. 3.2 shows the results after treating missing values on OUT_MPN. However, the similar results are obtained for the other variables (columns) with missing values in the dataset.

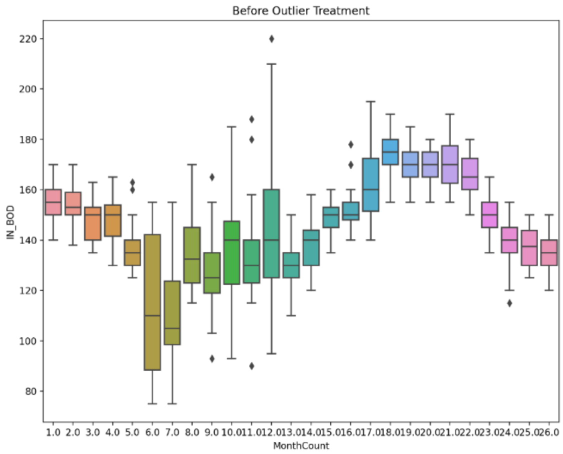

Fig. 4.2: *Before Outlier Treatment*

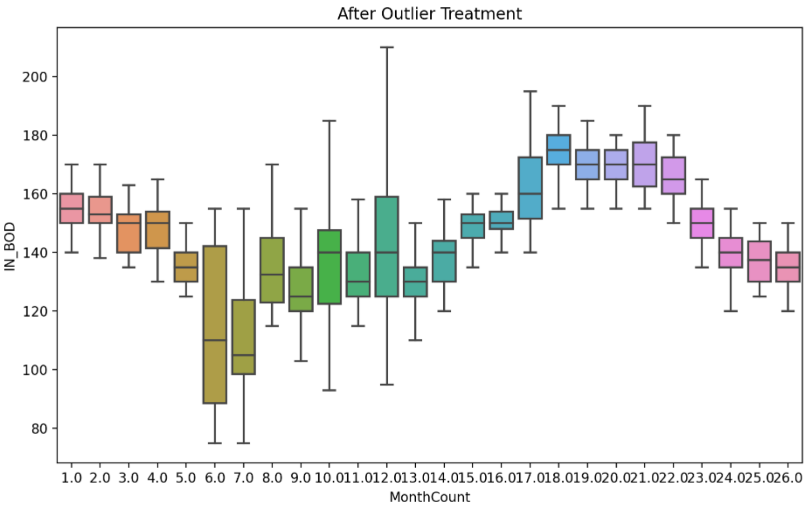

Fig. 4.3: *After Outlier Treatment*

The raw data contained the outliers which were impacting/decreasing the efficiency of the model(s) created. So, the treatment for removing the outliers from the data is done during pre-processing. The Fig. 3.3 shows the graph for inlet BOD before the outlier treatment and Fig. 3.4 shows the graph for inlet BOD after outliers' treatment. Similarly, we did the outlier treatment for other variables (columns) in the data set.

Fig. 10 signifies the effectiveness of WWTP by showing the relationship between influent (untreated water) and Effluent (Treated and processed) water. The parameters include Total Suspended Solids (TSS), Chemical Oxygen Demand (COD), Biological Oxygen Demand (BOD), Most Probable Number (MPN), Dissolved Oxygen (DO) and pH.

The graph in Fig. 10, show that there is a great fluctuation/ variation in the influent parameters which are the main factors affecting the efficiency of the plant. Therefore, a prediction by a Machine Learning model shall greatly help in managing and enhancing the quality and effectiveness of waste water treatment processes used in the plant.

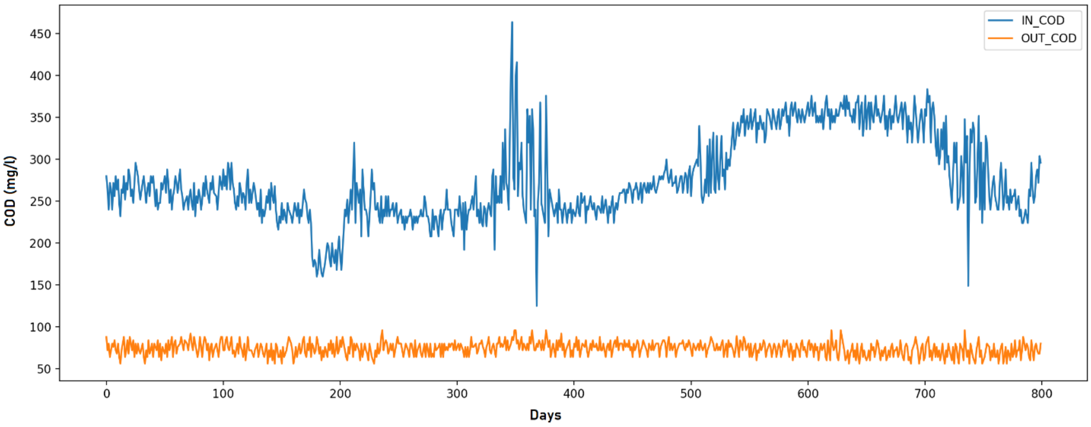

Fig. 4.4: *Relationship between values of an influent and effluent parameter*

The relationship between parameters can be analysed by the correlation coefficient. It can be used to obtain the effectiveness of the relationship among the parameters and can be used for further analysis and modelling. The positive correlation signifies that if one value increases another also increases, higher value shows the stronger correlation.

In Fig. 3.8, heatmap shows this relation with the intensity of the colour used; darker colour shows the stronger relationship. The colour turning to blue shows the negative relationship means the increase in one value will lead to the decrease in another.

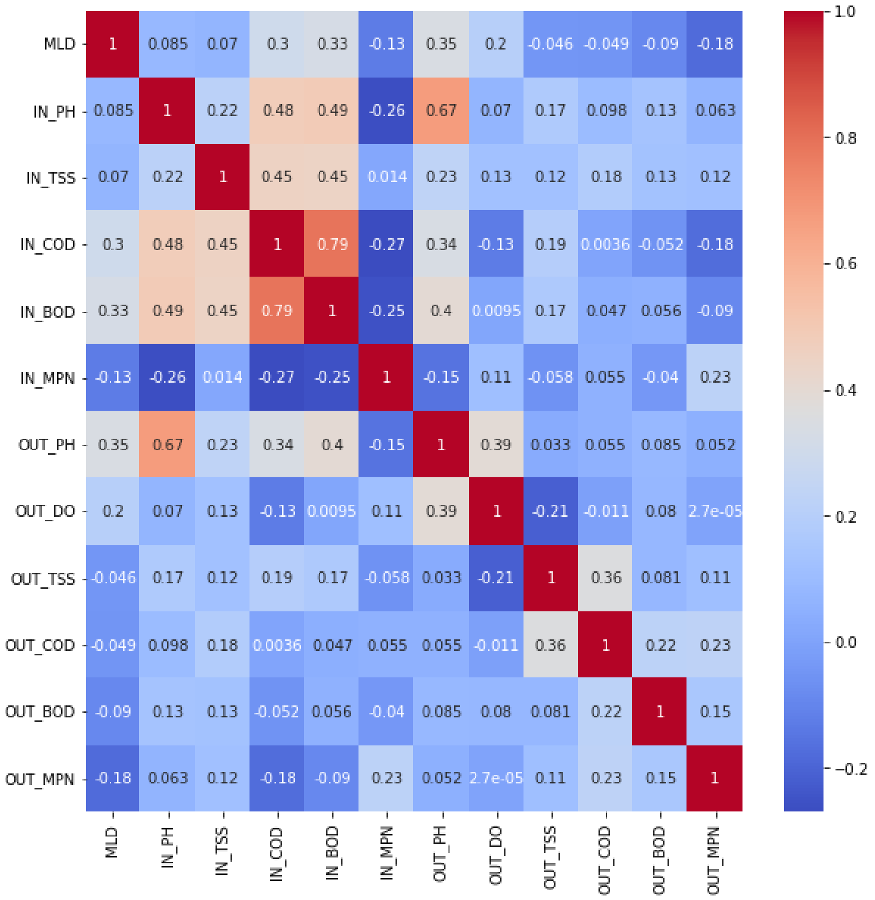

Fig. 4.5: *Heatmap for Correlation*

4.2 Results of Influent flow prediction model

To define the p and q value for the AC and PAC graph we will consider then both for the AR or p process, we expect that the ACF plot will gradually decrease and simultaneously the PACF should have a sharp drop after p significant lags. To define a MA process, we expect the opposite from the ACF and PACF plots. Then after selecting the p and q value we fine tune with the nearby range of the p and q value to get the best result.

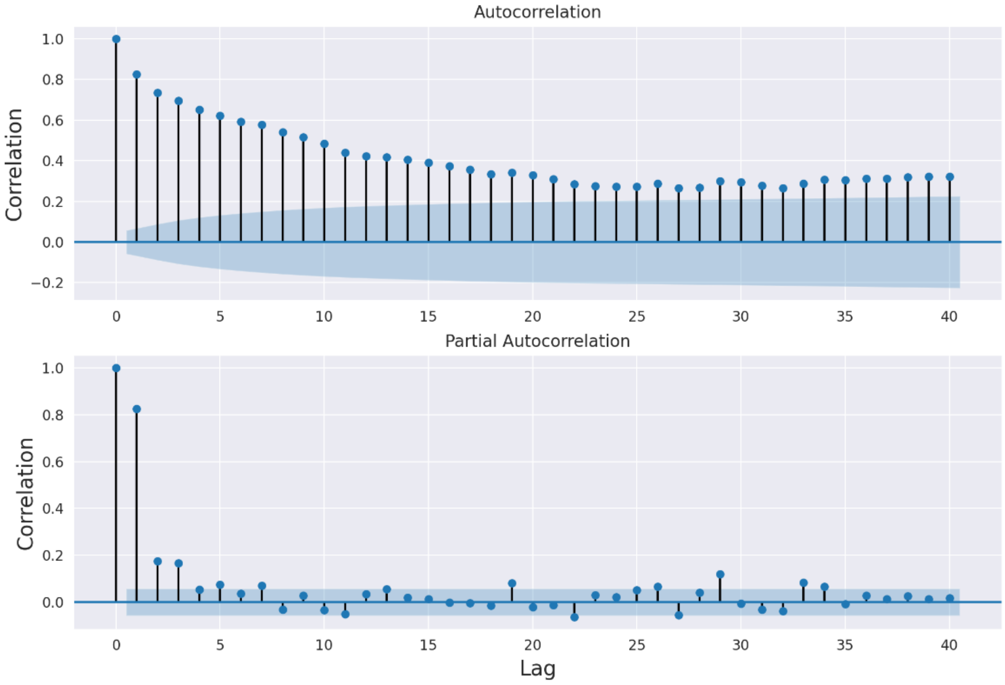

Fig. 4.6: *Auto Correlation and Partial AC graph*

The Fig 3.7 shows seasonal decomposition of influent flow data. We can see that trend of the data set changes from low to high in middle months and stays at constant pace afterwards.

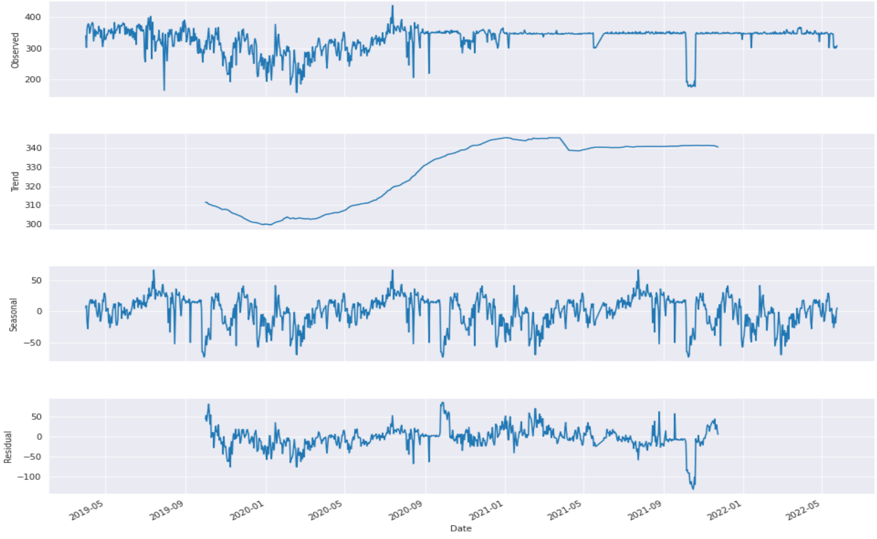

Fig. 4.7: *Trend and Seasonality pattern of flow*

After training the ARIMA (p,d,q), we analyse the accuracy of the model using the Mean Absolute Percentage Error (MAPE) and Symmetric Mean Absolute Percentage Error(SMAPE).

MAPE and SMAPE:

MAPE and SMAPE are the most popular metrics for checking the forecasting performances.

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{(A_t - F_t)}{A_t} \right|$$

$$SMAPE = (100\%)/n \sum_{(t=1)}^{n} \left| \frac{|F_t - A_t|}{\frac{(A_t + F_t)}{2}} \right|$$

The graph (Fig. 4.8) shows the Rolling standard of the data set we can clearly see the rolling standard is not in alignment with the original data so we had to apply the integration.

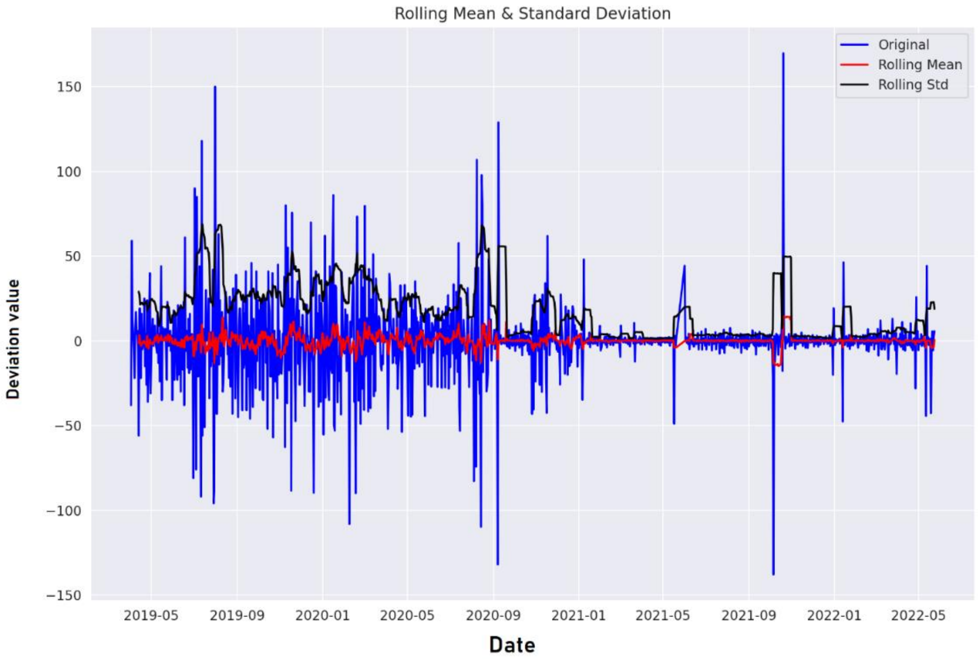

Fig 4.8: *Rolling Mean and Standard Deviation of Influent flow data*

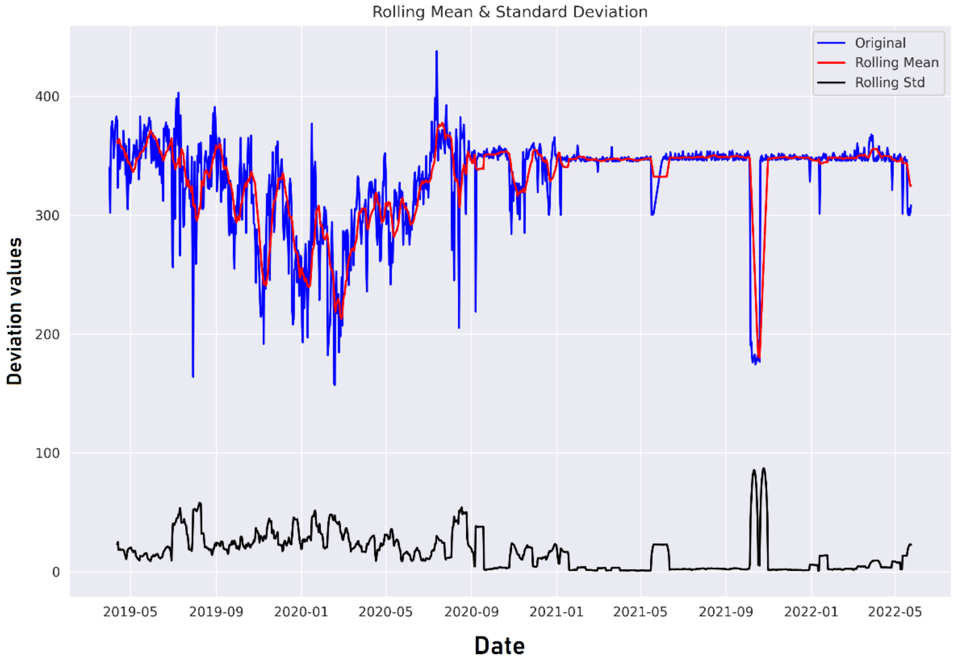

Fig 4.9: *Rolling Mean and Standard Deviation of Influent flow data after Integration*

After applying the integration in the dataset (formula mentioned in methodology) once we achieved the stationarity. Cleary seen in Fig. 4.9 that the rolling standard is in alignment with the integrated dataset. So, we need not to further apply the integration and we achieved stationarity with d order of 1.

Performance of ARIMA model was found to be good as the MAPE error and SMAPE error were within the range of 0-5% which is considered an error rate which is highly acceptable for time series future value predictions.

Our ARIMA model with additional handling of seasonal data was able to predict values at an error rate of:

MAPE – 2.67%

SMAPE – 2.59%

The final prediction graph can be seen in Fig 4.10.

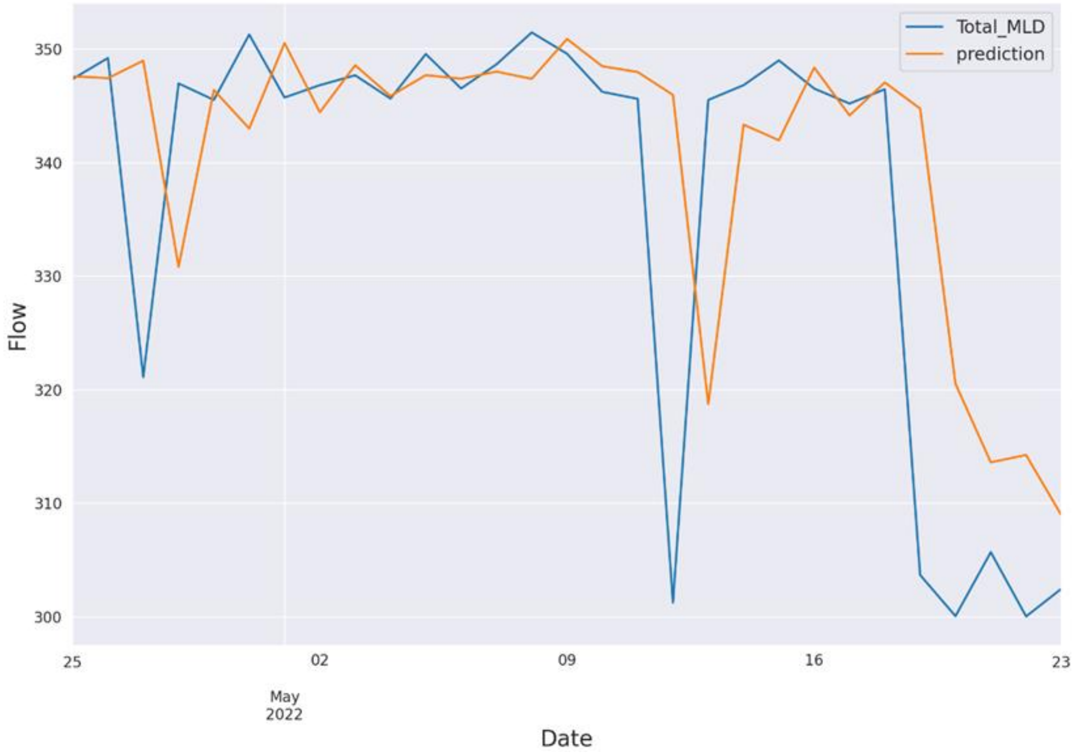

Fig. 4.10: *Actual and predicted value of flow*

The residual error graph (Fig 4.11) was also generated in order to get an idea of the predictions and its accuracy.

Residual Error = Actual value – predicted value.

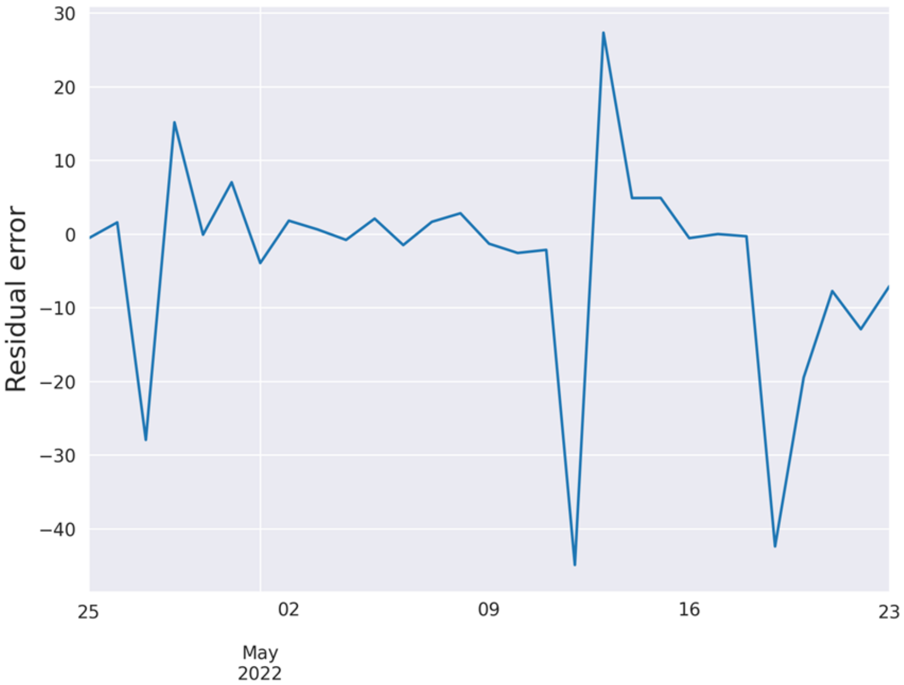

Fig 4.11: *Residual error graph of predicted and actual data*

4.3 Results of Effluent parameter quality prediction models

Various models were used to predict the effluent parameters and the results produced by them are:

Table 4.1 Results of Effluent prediction

	OUT_PH	OUT_DO	OUT_TSS	OUT_COD	OUT_BOD	OUT_MPN
Linear Regression	71.23	12.38	11.06	3.09	7.52	2.21
KNN	73.45	15.48	6.63	4.86	8.40	4.42
Gradient Boosting Regression	70.79	14.60	8.40	3.53	9.29	3.53
Random Forest Regression	71.68	11.50	6.63	3.98	10.61	4.42

Out of the four models of machine learning KNN Model performs the best out of the basic algorithms.

ANN predictions

- Currently we have implemented several machine learning models for Influent and effluent prediction
- For Effluent Parameters we have used Gradient Boosted Regression, KNN, Random Forest Regression Model.
- The Highest Score Achieved in KNN is **0.45(Out_PH)**
- The Highest Score Achieved in Gradient Boosted Regression is **0.5(Out_PH).**
- The Minimum Cost Achieved in Artificial Neural Networks is around **3e-3+5.**
- The best Model is **Artificial Neural Network** which predicts more than **50%** for each of the effluent correctly
- After Comparing the efficiency of the above mentioned models, we have concluded that **ANN** Model is best for our use case

Table 4.2 Results of Artificial Neural Network

	OUT_PH	OUT_DO	OUT_TSS	OUT_COD	OUT_BOD	OUT_MPN
Artificial Neural Network	81.02	53.05	77.58	79.80	7.52	67.81

Chapter 5

Conclusions

5.1 Conclusions

We have analyzed the flow and quality parameters like COD, BOD, TSS, DO, pH, Temperature, Ammonia, Phosphorous and oil content, etc. in influent, and also parameters like COD, BOD, DO, pH, etc. of effluent in the WWTP.

Data visualization depicts the fluctuating and varying nature of influent parameters in 345 MLD UASB-based Bharwara STP. The use of this real time data is detrimental in producing a model which would be helpful for further study.

Introducing Machine Learning Models in STP Power Plants will reduce human efforts in calculating the influent and effluent parameters, it will predict these parameters for later works.

We have predicted influent flow quantity values with a minimal error of 2.67% which is in an acceptable prediction error range (0-5% error is considered good). This influent prediction model can be used by any UASB based WWTP to get a dependable prediction.

Our Effluent parameter prediction model is also providing us with satisfactory results upon the use of ANN for the prediction algorithm.

5.2 Future Works

As we are in the era of Machines, Artificial Intelligence has emerged as the zenith of constantly evolving world. We need more advance machine learning models which can automate the works done by human efficiently, for easier and faster work completion.

We can provide a Graphical User Interface to the Engineers present in various Waste Water treatment plants to help them easily use our prediction tools in order to tweak the plants distributed control systems.

Access of more wastewater treatment plant data will also help us to further tweak and perfect our prediction models as data was a constraint in this project.

References

[1] Gautam, Rajneesh & Verma, Saumya & Islamuddin, Islamuddin & More, N.. (2018). Sewage Generation and Treatment Status for the Capital City of Uttar Pradesh, India. Avicenna Journal of Medicine. 5. 8-14. 10.15171/ajehe.2018.02.

[2] Tümer, Abdullah & Edebali, Serpil. (2015). An Artificial Neural Network Model for the Wastewater Treatment Plant of Konya. International Journal of Intelligent Systems and Applications in Engineering. 3. 131. 10.18201/ijisae.65358.

[3] Guo Hong, Jeong Kwanho, Lim Jiyeon, Jo Jeongwon, Kim Young, Park Jong-pyo, Kim Joon Ha, Cho Kyung, "Prediction of effluent concentration in a wastewater treatment plant using machine learning models", Journal of Environmental Sciences, vol. 32, pp. 90-101, 2015

[4] Matala Anna., "Sample size requirement for Monte Carlo simulations using Latin hypercube sampling.", Helsinki University of Technology, Department of Engineering Physics and Mathematics, Systems Analysis Laboratory, 2008

[5] Sin, Gürkan, Krist V. Gernaey, Marc B. Neumann, Mark CM van Loosdrecht, and Willi Gujer. "Uncertainty analysis in WWTP model applications: a critical discussion using an example from design.", Water Research, vol.43, no. 11, pp. 2894-2906, 2009.

[6] Granata Francesco, Papirio Stefano, Esposito Giovanni, Gargano Rudy, DE MARINIS, Giovanni., "Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators", Water, vol. 9, no. 2, pp. 105, 2017.

[7] Qin Xusong, Gao Furong, Chen Guohua, "Wastewater quality monitoring system using sensor fusion and machine learning techniques", Water research, vol. 46, no. 4, pp. 1133-1144.

[8] Wang Rui, Pan Zhicheng, Chen Yangwu, Tan Zhouling, Zhang J. "Influent Quality and Quantity Prediction in Wastewater Treatment Plant: Model Construction and Evaluation", Polish Journal of Environmental Studies, vol. 30, no. 5, 20

[9] D S, Manu, Thalla Arun, "Artificial Intelligence Models for Predicting the Performance of Biological Wastewater Treatment Plant in the removal of Kjeldahl Nitrogen from Wastewater", Applied Water Science, vol. 7, no. 7, pp. 3783-3791, 2017.

[10] Banerjee, Arif Siddiquie1 Rajiv. "Performance Evaluation & Upgradation of UASB Technology used for the Treatment of Sewage Generated from Lucknow City.", 2016

[11] Gautam, Sandeep Kumar, Divya Sharma, Jayant Kumar Tripathi, Saroj Ahirwar, and Sudhir Kumar Singh. "A study of the effectiveness of sewage treatment plants in Delhi region." Applied Water Science, vol. 3, no. 1, pp. 57-65, 2013.

[12] Alnaa, Samuel, Ahiakpor, Ferdinand, "ARIMA (autoregressive integrated moving average) approach to predicting inflation in Ghana", Journal of Economics and International Finance, vol.3, pp. 328-336, 2011

[13] Cramer, Duncan, and Dennis Laurence Howitt. The Sage dictionary of statistics: a practical resource for students in the social sciences, Sage, 2004.

[14] Available at: https://www.datavedas.com/model-evaluation-regression-models/