“Water Quality Analysis Of Major Rivers Of India Using Machine Learning”

A PROJECT REPORT

SUBMITTED IN THE PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE AWARD OF DEGREE OF

MASTER OF TECHNOLOGY

IN

SOFTWARE ENGINEERING

Submitted By

ASHISH KUMAR SINGH

(2K2/SWE/07)

Under the supervision of

Mr. Sanjay Patidar
Assistant Professor
Department of Software Engineering
Delhi Technological University, Delhi

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)Bawana
Road, Delhi-110042

AUGUST, 2023

DECLARATION

I, **Ashish Kumar Singh, 2K21/SWE/07** student of **M.Tech (SWE)**, hereby declare that the project entitled **“Water quality analysis of major rivers of India using Machine Learning”** is submitted by me to the Department of Software Engineering, **Delhi Technological University**, Shahbad Daulatpur , Delhi. I have done my project in partial fulfilment of the requirement for the award of the degree of Master of Technology in Software Engineering and it has not been previously formed the basis for any fulfilment of the requirement in any degree or other similar title or recognition.

This report is an authentic record of my work carried out during my degree under the guidance of **Mr. Sanjay Patidar.**

Place: Delhi   Ashish Kumar Singh

Date: 22nd August, 2023   **(2K19/SWE/07)**

CERTIFICATE

I hereby certify that the project entitled **“Water quality analysis of major rivers of India using Machine Learning”** which is submitted by **Ashish Kumar Singh (2K21/SWE/07)** to the Department of Software Engineering, Delhi Technological University, Shahbad Daulatpur , Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology in Software Engineering, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any degree or diploma to this university or elsewhere.

Place :Delhi

Date:1st June,2023

Mr. Sanjay Patidar

SUPERVISOR

Dept. of software engineering

ACKNOWLEDGEMENT

I am very thankful to **Mr Sanjay Patidar**(Assistant professor, Department of Software Engineering) and all the faculty members of the Department of Software Engineering at DTU. They all provided me with immense support and guidance for the project.

I would also like to thank the university for providing the laboratory, infrastructure, test facilities and environment so that I can work without obstacles.

I would also like to thank our lab assistants, seniors, and peer groups for providing me with all the knowledge on various topics.

Ashish Kumar Singh

(2K21/SWE/07)

ABSTRACT

Water is one of the most important natural resource after air. In-spite of the fact that most of the surface of of earth is composed of water there is only small part of that water which is usable. This vital natural resource must therefore be used very carefully. Because this river's water is used for drinking, domestic purposes, irrigation, and aquatic life including fish and fisheries, river quality may be a major concern. In-order to understand the quality of water that whether it clean or not we have to study and analyse various water quality parameters like BOD (Biochemical Oxygen Demand), temperature , pH(Potential of Hydrogen),DO (Dissolved Oxygen), and conductivity and to understand about the quality of water of various rivers of India that whether it is clean or not ,a classification model using three different classifier is presented in the study.The water quality data were classified using the J48, LMT, and Nave Bayes classification algorithm.The WEKA tool was used for analyse the collected data of various rivers then classify as clean or not clean. In this work we have studied 15 papers from various publishers and created a summary to study about how and why the water is classified as clean or not clean using various machine learning algorithms.

CONTENTS

List of Figures

List of Tables

List of symbols and abbreviations

Abbreviations	Full Form
ML	Machine Learning
AI	Artificial Intelligence
ROC Curve	Receiver Operating Characteristic Curve
AUC	Area Under ROC Curve
R2	R-Square Score
MAE	Mean Absolute Error
MSE	Mean Square Error
RMSE	Root Mean Square Error
VIF	Variance Inflation Factor
LMT	Logistic Model Tree

CHAPTER 1

INTRODUCTION

1.1 Machine Learning

A significant area of artificial intelligence (AI) called machine learning focuses on creating models and algorithms that can learn on their own by generating predictions or decisions based on data. It is used to improve to help the computer system to learn from the past experience without any explicit programming.

Fundamentally, machine learning is training a model using a dataset made up of a set of input and output variables. The data model hence developed tries learns patterns, relationships, and dependencies in the data by adjusting its internal parameters through a process called training. The primary aim of machine learning is to is to help the developed model so as to make some accurate predictions on the unseen data.

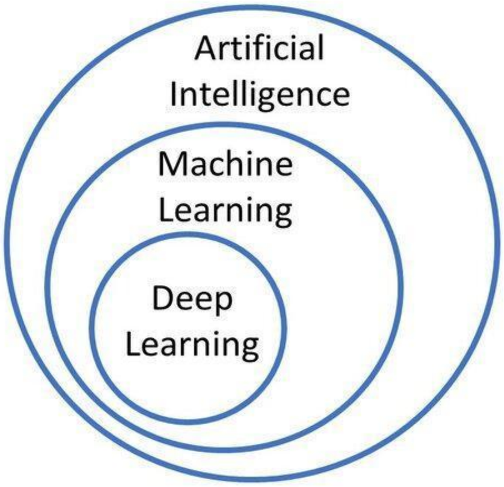


Fig.1: AI, Machine Learning, Deep Learning

1.2 Types of Machine Learning

1.2.1 Supervised Learning:

In supervised machine learning, the model learns from the data which is labeled ,in such cases the input data is associated with its correct output. It trains and learns to generalize from these examples given to the model during testing phase so as to to determine the output for the new unseen inputs.

We can observe the following diagram to understand the the supervised machine learning easily.

Two categories of problems are further separated from the supervised learning:

a.Regression

When there is a particular kind of relationship between the input variable and the output variable, the regression algorithm is employed. The basic goal is to establish a mathematical relationship between the input and output variables. In this case, the output is some sort of real value.

Examples include rupees, weight, earnings, etc.Some of the basic regression algorithms are :

- Linear Regression
- Regression Trees
- Non-Linear regression
- Bayesian Linear Regression
- Polynomial Regression

b.Classification

The classification algorithm is used when we have some kind of categorical output.ie the output variable is category. For example red/blue, disease/no disease ,Yes/no, true/false, male/female etc.

- Random Forest
- Decision Trees
- Logistic Regression
- Support vector machin

1.2.2 Unsupervised Learning:

The contrary of supervised machine learning, where the model does not need supervision, is unsupervised machine learning, where the model analyses and learns by observing the patterns and structures in the unlabeled data. Without any external direction of any kind, the model just observes the relationship inside the data. Problems like dimensionality reduction, grouping, and anomaly detection benefit from this kind of machine learning.

The unsupervised learning is divided into two types of problems:

a. Clustering: Using this technique, items are sorted into various groups. This is done is such a way that the object with similar properties are put in one group.

b. Association: It is used in large datasets to find relationship between variables.For instance, persons who buy X (let's say a T-shirt) item also frequently buy Y (let's say jeans).

Of course, supervised learning data is labeled, but unsupervised learning data is unlabeled.

1.2.3 Semi-supervised learning:

When only a tiny percentage of a sizable dataset is labeled, a learning challenge known as a "semi-supervised learning problem" arises. Both supervised and unsupervised learning experience these issues. Consider a photo collection of dogs, cats, people, etc. where some pictures have labels while the majority don't.

1.2.4 Reinforcement learning:

An agent that interacts with its environment and learns to make choices or execute actions that maximise a reward signal is the subject of this type of machine learning. The agent in this scenario receives feedback in the form of rewards or penalties based on its behaviours, enabling it to discover the best methods through experimenting.

1.3 Some Terminologies for machine learning:

a. Model:

A model is a particular representation, or hypothesis. When machine learning algorithms are used, the model learns from the application and produces the desired results.

b. Features:

Features are the specific, quantifiable characteristics of the data. Feature vectors can be used to quickly describe a number of numerical features. The model receives the feature vector as input. For instance, you can make predictions about fruits based on their colour, aroma, and flavour.

c. Target: Target variables, also known as labels, are the values that the model projected for us. Suppose, for instance, that we need to discover the names of fruits that is stated in the Features Section, making Fruit the label that specifies the fruit's value.

d. Training: We already have the input data, features, target (label), and necessary outputs for this mapping process. Thus, we use all the data to build a model that connects newly acquired information with previously learned classifications.

e. Forecast: When the model is ready, you can enter a set of inputs that provide predictive output (labels).

1.4 Objectives of the Project

Determination of water quality of major Indian rivers using machine learning algorithms to classify as clean or not clean.The algorithm which give determines the class of water as clean or not clean with highest accuracy is used for further implications.

The procedure used to achieved this is mentioned below:

- Generating a custom dataset for training and validating the ML model.
- Exploratory data analysis of data collected.
- Training and validating the performance of ML mode.
- Comparing the performance of different ML models available.

CHAPTER 2

REVIEW OF LITERATURE

Sakshi Khullar, et.al in 2023 [1],published an article that suggested a hybrid machine learning methodology for classifying river water quality.They proposed model for water quality classification of river Yamuna. The accuracy and the performance of the classification technique used in the study was compared and analysed with respect to some of the other basic classification techniques that include the svm(Support Vector Machine), Naïve Bayes classifier and the bagged and boosted tree classifiers. The proposed approach in the paper by them obtains an overall accuracy of 99.65% . In their work, Abhishek Bajpai et al. [2] offered a real-time approach that will aid in classifying the river ganga's water quality by looking at the information gathered from the mehendi ghat and the information from kannauj.They proposed a paper in which they wanted to identify and analyze the water quality of ganga that whether it is pure and clean or not.They identified the water quality of river Ganges that whether the water is healthy and portable. Their major objective was to classify the water of river ganga on some of the major water quality parameters using different classification algorithms of machine learning. The accuracy of the proposed model is found to be around 99 percent, which is far ahead in terms of accuracy in compared to other approaches used for water quality prediction. In 2015 Preeti Chawla, et.al [3] proposed a paper where they predicted the Pollution Potential of several Rivers in India with the help of empirical equation which consisted of water quality parameters. In 2014 Shailesh Jaloree , et.al [4] presented a paper where they studied the approach of decision tree to study the water quality. To analyse the water quality data of the Narmada River at a site in the Harda district, the research provided a classification model using a decision tree approach.They used the WEKA software to implement the data model. So as to classify the water quality data decision tree was applied which identify the class of water. The factors that affect the water quality, such as nitrogen content, temperature, pH, COD, and BOD, are particularly significant. A review study for evaluating water quality characteristics was given by S.P. Gorde et al. in 2013[5]. A study looked at the variables needed to evaluate the water's quality.Salinity, phosphates, turbidity, temperature, pH, and nitrates are among the primary factors that determine the quality of water. In January 2019 Bhatta Mahesh[6] proposed a review on various machine algorithms.As these machine leaning algorithms are used for various purposes which reduces the human effort and saves time. In 2022,using supervised machine learning, Preeti Nanjundan,et.al[7] presented a procedure for calculating and assessing the water quality..In order to further employ the Water Quality Index (WQI) to determine the water quality, the paper examined its analysis and calculation. In 2019 Dziri Jalal,et.al [8] proposed a paper where they analyzed the performance i.e. the accuracy of various machine learning algorithms for a system which determines the water quality. Some of the machine learning algorithms include tree and

SVM. Their performance was measured and evaluated further. In 2017] Anoop Kumar Shukla, et.al [9] proposed a paper in-order to analyze the water quality of surface water of river ganga using a methodology called as index mapping.They developed a system that can combine the Geographic Information System (GIS) and the Water Quality Index (WQI) in order to accurately estimate the water quality of the Ganga River.The Varanasi station had the poorest water quality, while they saw very poor water quality in the river's lower reaches. In 2021 Jitha P Nair, et.al[10] proposed an examination of predictive models for analysing the water quality of rivers using different machine learning approaches.In this study, the efficacy of various big data and machine learning-based prediction models for water quality is evaluated. In 2023 D.Kavitha,et.al [11] proposed a survey on water quality prediction. The objective is to study various machine learning methodology and approaches so as to predict the waste water quality with most accuracy.In 2015 preeti chawla et.al[12]offered a method for creating an empirical equation to forecast the pollution potential of river water, noting that it is consistent with the current contamination potential of Indian rivers.In 2022 Anoop Abraham et.al [13] conducted an experimental study in which they used various machine learning algorithms to decide the river water quality of san antonio river.In 2023 D.Brindha et.al[15] conducted a research whose main objective is to water quality using machine learning technique.In 2021H.choi et.al [17]proposed a paper to look into how quickly growing urbanization affects water pollution..In 2018 R.Roy [18] published a paper to understand the water quality analysis in which inorder to verify whether the source of the water is acceptable, a number of water quality characteristics are evaluated and compared to their standard levels. In 2019 M.Jaiswal et.al[19] presented a paper on Comprehensive evaluation of water quality status for entire stretch of Yamuna River India where they presented the study on physicochemical water quality of yamuna river.In 2019 Feng-Jen Yang[21] proposed a paper on an extended idea about decision trees in which he demonstrated two experiments to understand that fundamental theory of decision tree can be extended to go beyond boolean decision.Neha Radhakrishnan et al. [22]SVM, Decision Tree, and Naive Bayes were three different machine learning algorithms that were evaluated and compared in the study "Comparison of Water Quality Classification Models Using Machine Learning" that was presented in 2020. They discovered that the Decision Tree method generated the most precise classification results.In 2015 Salisu Yusuf Muhammad et.al[23] proposed a paper on “Classification Model for Water Quality using Machine Learning Techniques” to propose a suitable classification model to classify water quality based on machine learning algorithms.They analysed several classification models and algorithms.In 2016 Onder Gursoy[24] proposed a paper to understand about the most appropriate classification method for water quality.A study on the application of machine learning to forecast water quality assessments was presented by Suma S[25].

Summary Table

AUTHOR	YEAR	METHODOLOGY	Dataset
Shakshi kullar, et.al[1]	2022	Hybrid ML technique	Delhi based CPCB(Central Pollution Control Board)
Abhishek Bajpai, et.al [2]	2022	Random forrest	CPCB
Preeti Chawla, et.al[3]	2015	Emperical equation	Real time data
Shailesh Jaloree,et.al[4]	2014	Decision tree	Real time data
S.P.Gorde, et.al[5]	2013	Analysis	Real time data
Bhatta Mahesh[6]	2019	ML algo	Real time data
Preethi Nanjundan, et.al[7]	2022	Unsupervised ML	Environment Information System
Dziri Jalal, et.al[8]	2019	Classification algo	Water treatment station “Ghadir El Golla” of Tunisia.
Anoop kumsar shukla et.al[9]	2017	Index mapping	CPCB,UPPCB,NIH,CWC,ASTER,DEM,LPD,AAC
Jitha P nair,et.al[10]	2021	ML and big data	Real time data
D kaviha,et.al[11]	2023	Machine learning	Real time data
Preeti chawla,et.al[12]	2015	Machine learning	Real time data
Anoop abraham,et.al[13]	2022	Machine learning,data analysis	kaggle
Sago dzeroski,et.al[14]	1995	Analysis and classification	Real time data
D.brindha et.al[15]	2023	Machine learning	Real time data
Z. Kılıç,et.al[16]	2020	analysis	Real time data
H. Choi,et.al[17]	2021	Multivariate statistical technique	Real time data
R. Roy,et.al[18]	2018	analysis	Real time data
M. Jaiswal,et.al[19]	2019	analysis	Real time data
Swapan Shakhari,et.al[20]	2020	Predictive analysis	Real time data
Feng-Jen Yang,et.al[21]	2019	Decision tree	Real time data
Neha Radhakrishnan,et.al[22]	2020	Predictive analysis	Real time data
Salisu Yusuf Muhammad,et.al[23]	2015	weka	Department of Environment(DOE),Malaysia
Onder Gursoy[24]	2016	classification	Real time data
Suma S,et.al[25]	2023	weka	Remote sensing,laboratory,field measurements

Table 1:Summary table

CHAPTER 3

THEORITICAL CONCEPT

Six significant Indian rivers' water quality data are being collected for this investigation. The variables under investigation are conductivity, temperature, BOD, and pH. The goal of this study is to use several categorization techniques to categorise the water quality data of India's six main rivers as clean or unclean. Beas, Godavari, Sutlej, Ganga, Yamuna, and Brahmaputra are the six important rivers that are being examined for water quality.

Most studies that are associated with the analysis of the water quality use many water quality parameters such temperature, pH level, alkalinity, acidity, BOD, DO, physical phenomenon, TDS etc. Water quality analysis of stream could be an immense field and critically vital further. Ganga Water Mission is one amongst the biggest water quality analysis done on any major stream in India.

3.1 Why Water quality quality analysis?

Water is considered to be one of the most important natural resource for all living organisms on earth.The existence of life without water is not possible.Water is obtained from various sources like rivers , lakes etc.[18] Most Indians depend heavily on their rivers for their daily needs that are living their lives in some diverse and different parts of the country as this water from various rivers across the country is utilized and consumed for various kinds of purposes like in domestic work,for the purpose of irrigation, potable water, manufacturing of electricity etc. that contribute to humans, creatures and industries [3]. We need this water to be at a specified purity level because going over or under these limitations might be very damaging for the life of all different sorts of living things on Earth. It has been determined that the majority of Indian rivers are extremely polluted and unfit for regular use based on the analysis of several surveys on these waterways. [16]

The contamination of river can be because of man-made activity or because of some natural activity. The domestic sewage discharged into the water bodies cause various kinds of water-borne diseases. The animal dunghills also contaminate the water sources .[20]Huge amount of heavy metals are added into the ground water source by various activities like mining and construction. The fertilizers and pesticides used by farmers in the fields contaminate the ground water. The harmful chemicals released from the industries also contaminate the water. Rapid industrial development is the major cause for degrading the water quality of all the rivers[1].

Some of the recent studies have shown that approximately 15-16 lakhs people die each and every year due to the consumption of such polluted water. World Health Organization(WHO) recently conducted a study and the results were alarming ,it found that nearly 37% of people in the urban region and 64% people that are living in rural parts of the country are living without access to pure water. About 80 percent of the illness and diseases are waterborne ,which have resulted in death of about fifty lacs people and 250 crore infirmities[6].

The analysis of the quality of water is method where we have to identify the quality of water that whether it is clean or not by studying the value of different water quality parameters like temperature ,BOD etc. and analyzing that whether these values lie withing the prescribed values already set by the government. This complete process becomes very tedious and time consuming when it is done manually thus now machine learning is used to solve such classification problems by means of regression and forecasting. The water quality data containing several parameters of major rivers are collected in which the parameters that are investigated are BOD, pH, DO, temperature and conductivity. Eventually this water quality data obtained is classified using different classification techniques as clean or not clean [3].

Calculating the required water quality of water using a number of links between different criteria results in a standard index known as the Water Quality Index (WQI)[9]. With the help of the National Water Quality Monitoring Program, India's Central Pollution Control Board (CPCB) is in charge of eliminating water pollution in rivers.[19]

Machine learning is one of the most efficient method in-order to evaluate huge dataset which is collected from a range of different sources in order to find patterns in the data or to draw some conclusions about the data which was collected. Some suitable procedures must be chosen and the models need to be trained and validated so that we could apply various machine learning algorithms on the collected data.The choice of algorithms that needs to be applied on various data becomes important as one algorithm may give poor performance on some kind of data whereas the same algorithm might perform excellently on other data. The machine learning can also be further classified into two major sub-categories that include the supervised and unsupervised learning. The labeling of datasets is the main distinction between these two categories[14].

3.2 METHODOLOGY

3.1.1 Data Analysis Tool

Exploratory data analysis of the data collected is done using jupyter notebook which is an open-source environment to analyse and visualize the data.It can also be used for machine learning.The data gathered in this survey is analysed using the WEKA tool.Essentially, it is a piece of open source software that is utilised for the purpose of machine learning.It offers a variety of machine learning algorithms and other pertinent tools that are useful for data classification, preprocessing, clustering, regression, and other tasks.This programme was created at Waikato University in New Zealand. [4]

Weka supports a diverse range of data formats and allows users to manipulate and preprocess their datasets efficiently. It offers a wide array of data preprocessing techniques, such as attribute selection, normalization, and missing value handling, which are essential for preparing data before applying machine learning algorithms.One of Weka's key advantages is the vast library of machine learning algorithms it offers. It covers some well-known machine learning techniques, such as random forests, naive Bayes, support vector machines, and k-nearest neighbors (k-NN). This wide range of algorithms addresses different machine learning problems, allowing customers to explore and select the best solution for their unique requirements. [25]

3.1.2 Datasets

In this study, data from the Central Pollution Control Board, which is supported by the Ministry of Environment and Forests, Government of India, was used to classify and build the decision tree.

3.1.3 Water Quality Parameters Used

Any water source's water quality can be assessed using a variety of water quality metrics.In this study, we employed five main water quality parameters to examine and assess the water quality of several Indian rivers. The parameters which were used to determine whether the water is clean is pH level, Temperature, Conductivity, BOD and DO.

3.2 WORKFLOW

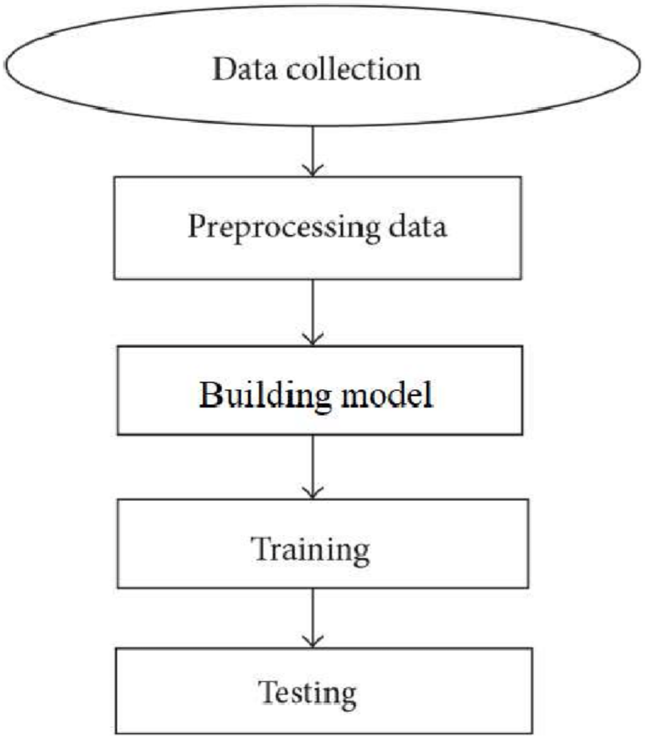


Fig.2 :Workflow

a. Data Collection

In this work, data from the Central Pollution Control Board, which is supported by the Ministry of Environment and Forests, Government of India, were used to classify and build the decision tree.

b. Data Preprocessing

Before the raw data is supplied to the algorithms for analysis, it needs to be further processed. Data integration, data reduction, data cleansing, and data transformation are all part of this step. [25]

c. Building Model

To discover patterns in data, machine learning is utilised. Either supervised or unsupervised learning can be done. Regression, categorization, regression, and forecasting are only a few of the machine learning techniques. This stage involves selecting and training one machine learning algorithm that generates a prediction based on the supplied data. [25]

d. Training

The generated model is trained using the data collection. A fraction of the complete dataset is used to train the model at this step.

e. Testing

Here, the model that was developed using training data is put to the test using test data. So, during this stage, the model's accuracy is checked.

3.4 Proposed work

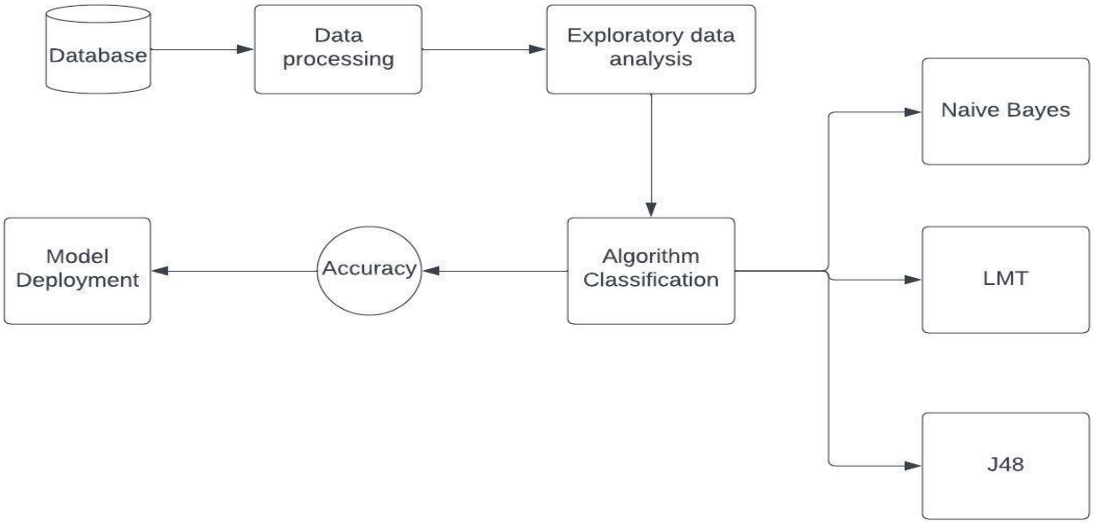


Fig:3:Architecture diagram of water quality classification

a.Database

In this work, data from the Central Pollution Control Board, which is supported by the Ministry of Environment and Forests, Government of India, were used to classify and build the decision tree.

b.Data preprocessing

The goal of data preparation is to make raw data usable for a machine learning model. Therefore, while building a machine-learning model, data preprocessing is the most important stage.

Unfortunately, because real-world data is compiled from various sources utilizing data mining and warehousing techniques, it is full of errors, noise, missing values, and incomplete information. Machine learning must employ data preprocessing in order to address all of these.

The need for data preprocessing is described below:

1.Improving Data Quality: Data preparation is essential for enhancing the quality of the data by correcting inconsistencies, inaccuracies, and errors in order to enable accurate and robust analytics.

2.Dealing with Missing Values: Data preparation techniques like imputation are crucial for successfully handling missing data because datasets commonly contain missing values. Machine learning models can be significantly less effective when missing values are present.

3.Scaling and normalizing features: Normalizing or scaling characteristics with the use of data preparation is essential for algorithms that rely on the scale of the input. Many machine learning methods need that all of the attributes be scaled similarly in order to work properly.

4.Dimensionality reduction:Principal Component Analysis (PCA), among other data preprocessing techniques, is used to decrease the number of input attributes.. This not only enhances the performance of models but also makes the dataset more manageable and computationally effective.

5. Handling Outliers:The accurate identification and control of outliers are made possible by data preparation. Because outliers can skew results and have a disproportionately negative impact on the modeling process, this is important.

c.Exploratory Data Analysis

Exploratory data analysis (EDA) is a technique for looking at data to find patterns, trends, or to verify data assumptions. It makes use of graphs and statistical summaries.

d.Algorithms

The algorithms that are used for classification and comparison of water quality for various rivers across India are already discussed in the next section.

e.Accuracy

The result and the accuracy obtained on the dataset for various rivers are discussed in the next section in detail.

3.3 ALGORITHMS USED

a. Naïve Bayes- One of the most well-known machine learning algorithms is naive bayes.A type of probabilistic classifier is the naive bayes.The "naive" assumption is that the features used for classification are neither related to or independent of one another. This method is based on the Bayes theorem. Even though the technique is straightforward, it nonetheless works well for a wide range of applications and objectives.For instance, text classification and spam filtering. The likelihood of a data point belonging to each class is determined by the naive bayes algorithm. The conditional probabilities of the characteristics given each class and all prior knowledge of the class distribution are used to achieve this.The probabilities are estimated using the training dataset. [13]

Naive Bayes learns the probabilities by counting how many times each feature is appearing in each class and then computing the appropriate probabilities during the training phase, when the model is provided with training data to train the machine learning model. The prior probability of each class is also estimated and predicted using the training data's frequency of occurrence. The algorithm uses the probabilities it learned during the testing phase to categorise fresh, unobserved data points in the testing phase, which involves testing the trained model with test data. The nave Bayes classifications employ the Bayes' theorem to determine the posterior probability of each class given a set of data.

b. The Logistic Model Tree (LMT)- A extremely effective classification model is produced by the Logistic Model Tree (LMT) algorithm, a machine learning technique that blends decision trees and logistic regression. The LMT method is mostly recognised for its capacity to handle numerical and categorical features in a very efficient manner. The creation of decision trees is the foundation of the LMT algorithm. By repeatedly splitting the data based on feature conditions that maximise the separation of the target classes, decision trees are created. This procedure keeps on until a stopping criterion is satisfied, such as when the depth is reached or a specific minimum number of samples are present in each leaf node.

The majority class of training data samples in each leaf node determines the projected class in conventional decision tree techniques. But in LMT, the leaf node predictions are improved through the application of logistic regression. Logistic regression provides probabilities to each class rather than relying merely on majority voting, producing a more complex and probabilistic forecast. [25]

Both the decision tree and the logistic regression models are constructed concurrently at the leaf during the training phase of our LMT approach, where the model is trained with training data. Within each leaf node, relationships between the input features and the target variable are captured using logistic regression models. A fresh data point is run through the decision tree during the testing phase to identify the proper leaf node. The class probabilities are then determined using the logistic regression model linked to that leaf node, and the final class prediction is then made using those results. The LMT algorithm has a number of advantages. First, it combines the advantages of logistic regression and decision trees, resulting in a more precise and understandable classification model. The LMT technique is adaptable for a variety of datasets since it can handle both numerical and categorical information. Its prediction power is increased by its capacity to recognise nonlinear correlations and interactions between features.

c. J48-- The J48 algorithm, additionally called C4.5.It is one of the most well-known machine learning algorithms out there and is based on the categorization task-specific decision tree. The J48 algorithm is created by further extending the ID3 algorithm. J48 creates a decision tree using a technique that involves iteratively dividing the data based on the characteristic that results in the greatest information gain or impurity reduction. Making a tree that maximises the separation of the target classes is the goal here. Usual criteria for determining a node's impurity include entropy and the Gini index. The root node serves as the starting point for the J48 algorithm because it essentially reflects the complete dataset. The best feature is then chosen and selected, and the data is divided depending on information gain or impurity reduction.This procedure recursively continues node until a stopping condition is satisfied for each child, such as reaching a predetermined depth or having a minimum number of samples in each leaf node. The J48 algorithm's capacity to handle both categorical and numerical information is one of its advantages [25].

d. Decision Tree- Another well-known and popular machine learning approach that may be applied to both classification and regression applications is the decision tree. It is shown as a kind of tree structure in the form of a flowchart, where each internal node in the tree essentially reflects a characteristic or attribute of the data, each branch denotes a set of rules for making decisions, and each leaf node shows the result or prediction. Recursively splitting the data based on several feature values to produce nodes and branches is how the decision tree is built. The main goal is to identify the optimum splits that minimise prediction error while maximising the separation of the target classes.To divide the data based on a certain criterion, such as information gain, Gini index, or mean squared error, a feature is selected at each node of the decision tree. The feature that gives the largest post-split improvement in class purity or forecast accuracy is the one that is selected. The procedure iteratively repeats itself until a stopping condition is met, such as reaching a maximum depth or having a minimum number of samples in each leaf node.[21]Because of how simple it is to visualise the resulting tree structure, the decision tree model is incredibly simple to grasp and comprehend. Each decision rule represents a condition that directs the classification or regression process along a path going from the root to a leaf node. The ultimate forecasts or results are found in the leaf nodes. [1]

The decision tree method learns the best splits during the training phase, when the model is trained using training data, by examining the training data and identifying the feature values that produce the best separation or prediction. In the case of classification, each leaf node is given the majority class of the training samples that reach that node. Regression assigns the mean or median value of the target variable to the leaf nodes. New data points are categorised or forecasted by traversing the decision tree based on their feature values while the model is being tested using the testing data. The final prediction or result is provided by the data point as it follows the decision rules from the root to a particular leaf node. [25]

3.5: Metrics for performance (Regression)

Regression falls under supervise learning, so we will map the actual value to the anticipated value to see how well our model was constructed. The following performance measures are used to assess the regression model:

1. MAE

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Fig 4: MAE formula

Regression analysis typically utilizes the Mean Absolute Error (MAE) metric to assess a prediction model's accuracy. The average difference between anticipated or expected values and actual values is calculated. Before determining MAE, we first calculate the absolute difference between each anticipated value and its corresponding actual value. Then, by averaging these absolute differences, the MAE value is determined.

The mathematical expression of the MAE formula is as follows:

MAE = [y_pred - y_actual] * (1/n)

In this case, the variables "MAE" and "n" stand for the total number of data points, "" stands for the summation symbol, "y_pred" stands for the predicted value, and "y_actual" stands for the actual value. The average magnitude of the model's predicted mistakes or errors can be understood using MAE. Given that it is unaffected by the errors' sign (positive or negative), it is a reliable metric. Regardless of the error's indication, each one is handled equally.

Better model performance is indicated by a lower MAE number, which signals that the predictions are more accurate. A higher MAE number, on the other hand, denotes greater differences between the expected and actual values. [20]

2. MSE

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Fig.5: MSE formula

Mean Squared Error (MSE) is a statistic that is frequently used in regression analysis to assess how well a prediction model is doing. The average squared difference between the expected and actual values is what is measured. The squared difference between each predicted value and its matching actual value is first computed in order to determine MSE. The MSE value is then calculated by averaging these squared differences.

The mathematical expression of the MSE formula is as follows:

MSE = (y_pred - y_actual) * (1/n)

In this equation, the terms "MSE" and "n" stand for the total number of data points, "" stands for the summation symbol, "y_pred" stands for the predicted value, and "y_actual" stands for the actual value. The average magnitude of the squared errors caused by the model's predictions is measured by MSE. The impact of greater errors is amplified by squaring the differences, which makes MSE particularly sensitive to outliers or extreme errors in the predictions.

Better model performance is indicated by a lower MSE value, which means that the predictions are more accurate. A higher MSE value, on the other hand, denotes greater differences between the predicted and actual values[20].

3. RMSE

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Fig.6:RMSE formula

In regression analysis, the Root Mean Squared Error (RMSE) is a commonly used statistic to assess how well a predictive model performs. The average squared difference between the predicted values and the actual values is what is measured.

We compute the squared difference between each projected value and its matching actual value before calculating RMSE. The average of these squared differences is then calculated. In order to determine the RMSE value, we finally square the average.

The mathematical expression of the RMSE formula is as follows:

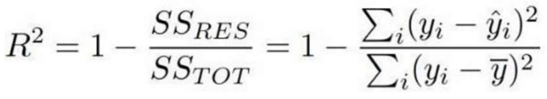
RMSE is equal to sqrt((1/n)*y_pred-y_actual)/2)

In this equation, the terms "RMSE" and "n" stand for the total number of data points, "" stands for the summation symbol, "y_pred" stands for the predicted value, and "y_actual" stands for the actual value.

The RMSE offers a measurement of the predictions' errors. The error values are returned to the target variable's original scale by calculating the square root of the average squared differences. This makes RMSE simple to understand and directly compared to the target variable.

Better model performance is indicated by a lower RMSE value, which means that the predictions are more accurate. A higher RMSE value, on the other hand, denotes more differences between the expected and actual values [20].

4. R-Squared



$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Fig.7:R-Squared

A statistical metric called R-squared, commonly referred to as the coefficient of determination, is used to evaluate the goodness-of-fit of a regression model. It calculates the percentage of the dependent variable's overall variation that the model's independent variables can account for.

R-squared runs from 0 to 1, where a value of 0 means that the model explains no variability in the dependent variable and a value of 1 means that the model perfectly fits the data and accounts for all variability. R-squared numbers can be negative, nevertheless, if the model performs worse than the dependent variable's simple average.

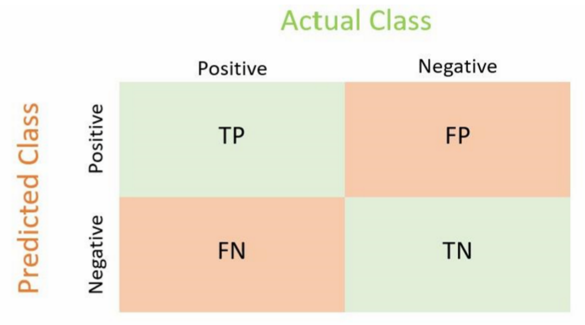
The total sum of squares (SST), which represents the entire variation of the dependent variable, is compared to the sum of squared residuals (SSR), which represents the unexplained variation of the dependent variable, to determine Rsquared. R-squared is calculated as follows:

(SSR / SST) - R2 = 1

In this equation, "R2" stands for the R-squared value, "SSR" for the squared residual sum, and "SST" for the sum of all squares R-squared shows how well the regression model matches the collected data. A higher R-squared value denotes that the independent variables in the model account for a bigger part of the variance in the dependent variable. A lower R-squared value, on the other hand, indicates that the model does not adequately account for the variability

3.6: Metrics for performance (Classification)

The following metrics are used for analyzing the performance of the models that are calculated using confusion matrix (as shown in Fig. 9) generated on the dataset:



		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Fig.8: Confusion matrix

1. Accuracy

Accuracy is a vital measure for evaluating the performance of a classification model. It determines the proportion of examples in the dataset that were correctly classified out of all the instances. We divide the total number of cases by the number of occurrences that were correctly classified (true positives and true negatives) in order to calculate accuracy.

The following is the accuracy formula:

Accuracy is calculated as follows: (Number of Instances Correctly Classified) / (Total Number of Instances)

A straightforward and understandable metric for gauging how accurately a classification algorithm predicts the right class labels is accuracy. It is frequently stated as a percentage, with a range of 0% to 100%. A better-performing model with a higher percentage of cases accurately predicted is indicated by a higher accuracy value. [8]

2. AUC (Area Under ROC Curve)

The Area Under the Curve (AUC) statistic is widely used to assess a classification model's efficacy in binary classification tasks. It determines the overall accuracy of the model's predictions by calculating the area under the Receiver Operating Characteristic (ROC) curve.

3. Precision

In classification problems, especially binary classification, precision is a commonly used parameter to assess how well a model predicts the future. Out of all instances projected as positive, it quantifies the percentage of correctly predicted positive instances.

The number of true positive (TP) predictions is divided by the total number of true positives and false positives (FP) to determine precision. The following is the precision formula:

Precision is equal to TP/(TP + FP).

Precision sheds light on the model's capacity to identify true positives and prevent false positives. It concentrates on the accuracy of positive forecasts and ignores situations where positive outcomes were actually projected to be negative [13].

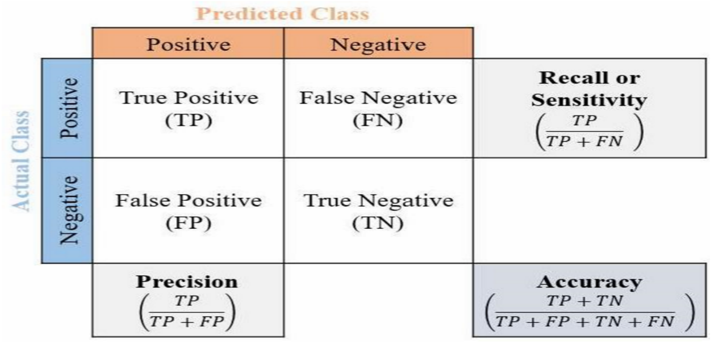
4. Recall

Recall is a frequently used statistic in classification tasks, particularly binary classification, to assess the model's accuracy in recognizing positive cases. It is also known as sensitivity or the true positive rate. It determines the proportion of true positive events that the model correctly discovered out of all actual positive instances. The sum of true positives and false negatives is divided by the total number of true positive predictions (TP), which is how recall is calculated.The following is the recall formula:

Recall is TP / (TP + FN).

Recall is concerned with how well the model can identify and categorise positive cases, regardless of the quantity of false negatives. It gives information on how well the model performs in terms of locating the pertinent positive cases [13].

The following figure (Fig. 11) represents the confusion matrix along with the metrics calculations:



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN)	Recall or Sensitivity $\left(\frac{TP}{TP+FN}\right)$
	Negative	False Positive (FP)	True Negative (TN)	
		Precision $\left(\frac{TP}{TP+FP}\right)$		Accuracy $\left(\frac{TP+TN}{TP+FP+TN+FN}\right)$

Fig.9: confusion matrix

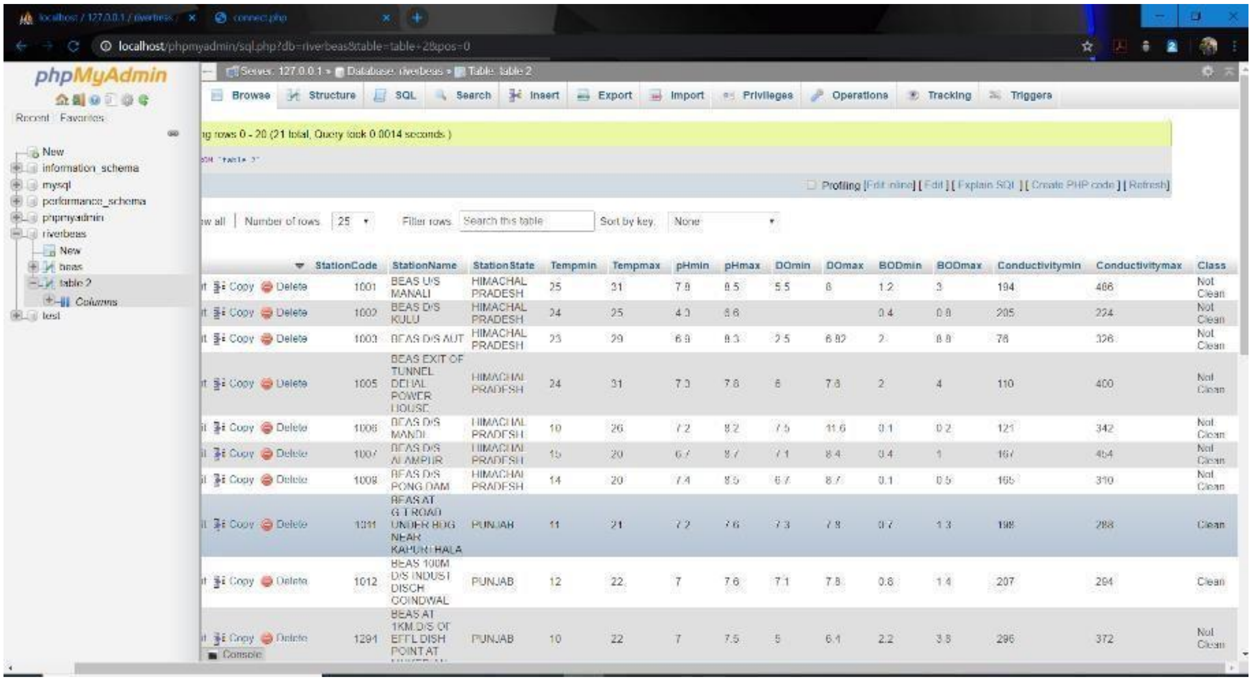
CHAPTER 4

Experimental Setup

4.1 Datasets:

In this Project, we collected the data from Central Pollution Control Board(CPCB) which is sponsored by Ministry of Environment and Forests, Government of India.

The Data table is given below:



ng rows 0 - 20 (21 total, Query took 0.0014 seconds.)

StationCode	StationName	StationState	Tempmin	Tempmax	pHmin	pHmax	DOmin	DOmax	BODmin	BODmax	Conductivitymin	Conductivitymax	Class
1001	BEAS U/S MANALI	HIMACHAL PRADESH	25	31	7.9	8.5	5.5	8	1.2	3	194	406	Not Clean
1002	BEAS D/S KULU	HIMACHAL PRADESH	24	25	4.3	8.6			0.4	0.9	205	224	Not Clean
1003	BEAS D/S AUT	HIMACHAL PRADESH	23	29	6.9	8.3	2.5	6.82	2	8.8	78	326	Not Clean
1005	BEAS EXIT OF TUNNEL DEHAL POWER HOUSE	HIMACHAL PRADESH	24	31	7.3	7.8	6	7.8	2	4	110	400	Not Clean
1006	BEAS D/S MANDI	HIMACHAL PRADESH	10	26	7.2	8.2	7.5	11.6	0.1	0.2	121	342	Not Clean
1007	BEAS D/S ALAMPUR	HIMACHAL PRADESH	15	20	6.7	8.7	7.1	8.4	0.4	1	167	454	Not Clean
1008	BEAS D/S PONG DAM	HIMACHAL PRADESH	14	20	7.4	8.5	6.7	8.7	0.1	0.5	165	310	Not Clean
1011	BEAS AT G.T.ROAD UNDER BDG NEAR KAPURTHALA	PUNJAB	11	21	7.2	7.6	7.3	7.8	0.7	1.3	198	288	Clean
1012	BEAS 100M D/S INDUST DISCH GOINDWAL	PUNJAB	12	22	7	7.6	7.1	7.8	0.8	1.4	207	294	Clean
1291	BEAS AT 1KM.D/S OF EFFL DISH POINT AT	PUNJAB	10	22	7	7.5	5	6.1	2.2	3.8	296	372	Not Clean

Fig 10:snapshot of the Original Datasets

Some of the water quality data which is observed in own dataset are:

1.pH

You can tell if water is acidic or alkaline by looking at its pH level. Using a logarithmic scale, the amount of hydrogen ions (H+) in the water is calculated. 0 to 14 make up the pH scale, with 7 being classified as neutral.Higher levels indicate greater alkalinity, while lower values indicate stronger acidity, pH levels below 7 suggest acidity while those above 7 indicate alkalinity.[25]

The negative logarithm of the hydrogen ion concentration serves as the basis for the pH scale. In order to determine the pH, the molar concentration of hydrogen ions in the water is multiplied by the negative logarithm (base 10).[22]

2.Dissolved Oxygen(DO)

Dissolved oxygen (DO), a crucial indicator of water quality, measures how much oxygen gas is dissolved in water. It is an essential indicator of the health and vitality of aquatic ecosystems since it directly affects the survival and behavior of aquatic organisms.All aquatic animals, including fish, invertebrates, and microbes, require oxygen to breathe. A number of variables, including temperature, water flow, air pressure, and the presence of photosynthetic organisms, affect the DO levels in water. [22]

3.Biochemical oxygen demand(BOD)

Among the several indicators an indicator which is used to determine the quality of the water is the biochemical oxygen demand (BOD), which quantifies how much oxygen is used by microbes in the biological breakdown of organic matter in the water. It offers details on the degree of organic pollution and any possible effects on the wellbeing of aquatic environments. In connection to the presence of organic pollutants such sewage, industrial effluents, and agricultural runoff, BOD is a crucial indication of water quality. These organic materials are used as food by bacteria, who then break them down through biological processes while exhaling oxygen.

4.Temperature

In order to evaluate the quality of the water and comprehend the peculiarities of aquatic ecosystems, temperature is a crucial element. It alludes to the measuring of the level of water's hotness or coldness. Numerous physical, chemical, and biological processes in aquatic settings are influenced by temperature.[11]

Aquatic ecosystems can be significantly impacted by changes in water temperature, both in terms of overall health and functionality. It immediately affects the water's dissolved oxygen levels. The solubility of oxygen declines with temperature, which has an impact on aquatic species' ability to breathe and survive. Aquatic organisms' metabolic rates are likewise impacted by temperature. In general, many species respond to higher temperatures by increasing their metabolic activities, growth rates, and reproduction rates. On the other hand, colder temperatures can slow down metabolic activities and affect the dynamics of the entire ecosystem.

5.Conductivity

Since conductivity measures how well water can carry an electrical current, it is a key indicator of water quality. It provides details on the total number of dissolved ions as well as the salinity or mineral makeup of the water.The quantity and types of dissolved materials in water, such as salts, minerals, and organic matter, affect conductivity. These dissolved substances, when ionized, allow the flow of electrical charges through the water.Ions in the water serve as the charge carriers when an electric current is run through it. The conductivity of the water increases with the number of ions present. As a result, conductivity and the total amount of ions in the water are directly connected.

4.2Data Cleaning

A dataset's flaws, inconsistencies, and inaccuracies are found and fixed through the process of data cleaning, sometimes referred to as data cleansing or data preparation. The procedure makes sure that the data utilized for analysis is trustworthy, accurate, and appropriate for modeling and deriving valuable insights.

The following steps are commonly included in data cleaning:

a. **Handling Missing Values**: Missing values can result in biased analysis and incorrect results, thus they must be found and dealt with.Techniques like imputation (replacing missing values with estimated ones) or deletion of rows/columns with missing values are utilized, depending on the type and amount of missing data.

b. **How to handle outliers**: Outliers are data points that differ noticeably from other observations and might skew analyses. It is important to carefully analyze the context and domain knowledge of the data while identifying outliers and decide whether to discard, transform, or keep them.

c. **Handling Duplicates**: Duplicate entries might cause unnecessary information and bias analytical results. Data integrity and bias are both ensured by locating and eliminating duplicate records.

d. **Handling Duplicates**: Duplicate entries might cause unnecessary information and bias analytical results. Data integrity and bias are both ensured by locating and eliminating duplicate records.

e. **Dealing with unnecessary Data:** The performance of the model and the computational load can both be improved by removing redundant or unnecessary characteristics or variables from the dataset.

4.3 Exploratory data analysis

Data science's primary exploratory data analysis (EDA) approach involves visually and statistically analyzing a dataset to obtain knowledge, identify patterns, and comprehend its underlying structure. Data scientists and analysts can develop hypotheses, spot outliers, and decide which data pretreatment procedures might be required for additional analysis and modeling with the aid of EDA.

Lets first analyse the **bar plot** in the figure 11 shown below.It represents the number of instances which are classified as clean(represented by 1) or not clean(represented by 0)

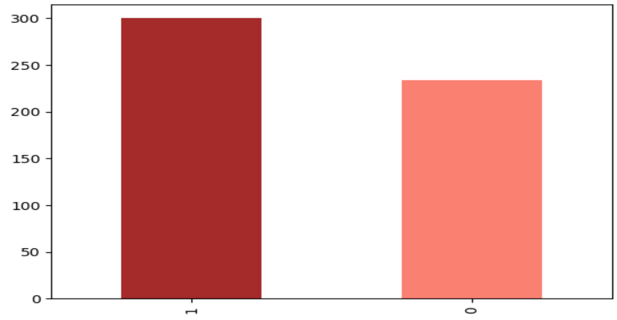


Fig.11: Classification

Lets have a look on the **heatmap** (fig :11) which describes the correlation between all the available parameters in the dataset.A value closer to 1 indicate strong correlation between the parameter. A heat-map plot is a graphic depiction of a matrix in which each element is colored to show the patterns and relationships in the data. It is frequently used to show correlations between variables or the relationship between the frequencies of two categorical variables.The cell with lighter shade of colour indicate a stronger correlation as compared to the cell of darker colour.

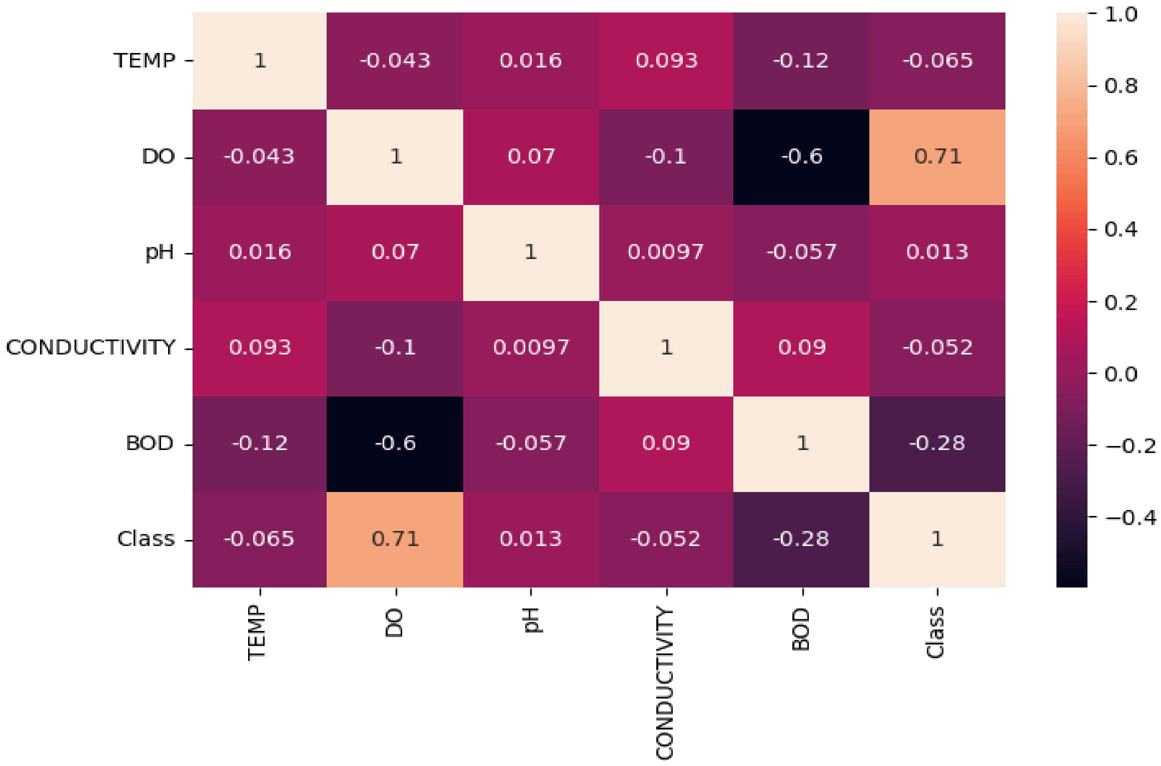


Fig.12: Heat-map of all the attributes

The correlation among the different variables can also be studied by observing the scatterplot shown below. Data points are shown as dots on a two-dimensional plane in a scatter plot, a sort of data visualization. The values of two variables, one on the x-axis and the other on the y-axis, define where each dot on the plot reflects a single observation. The relationship or correlation between these two variables is depicted using **scatter plots.** (fig:12)[25]

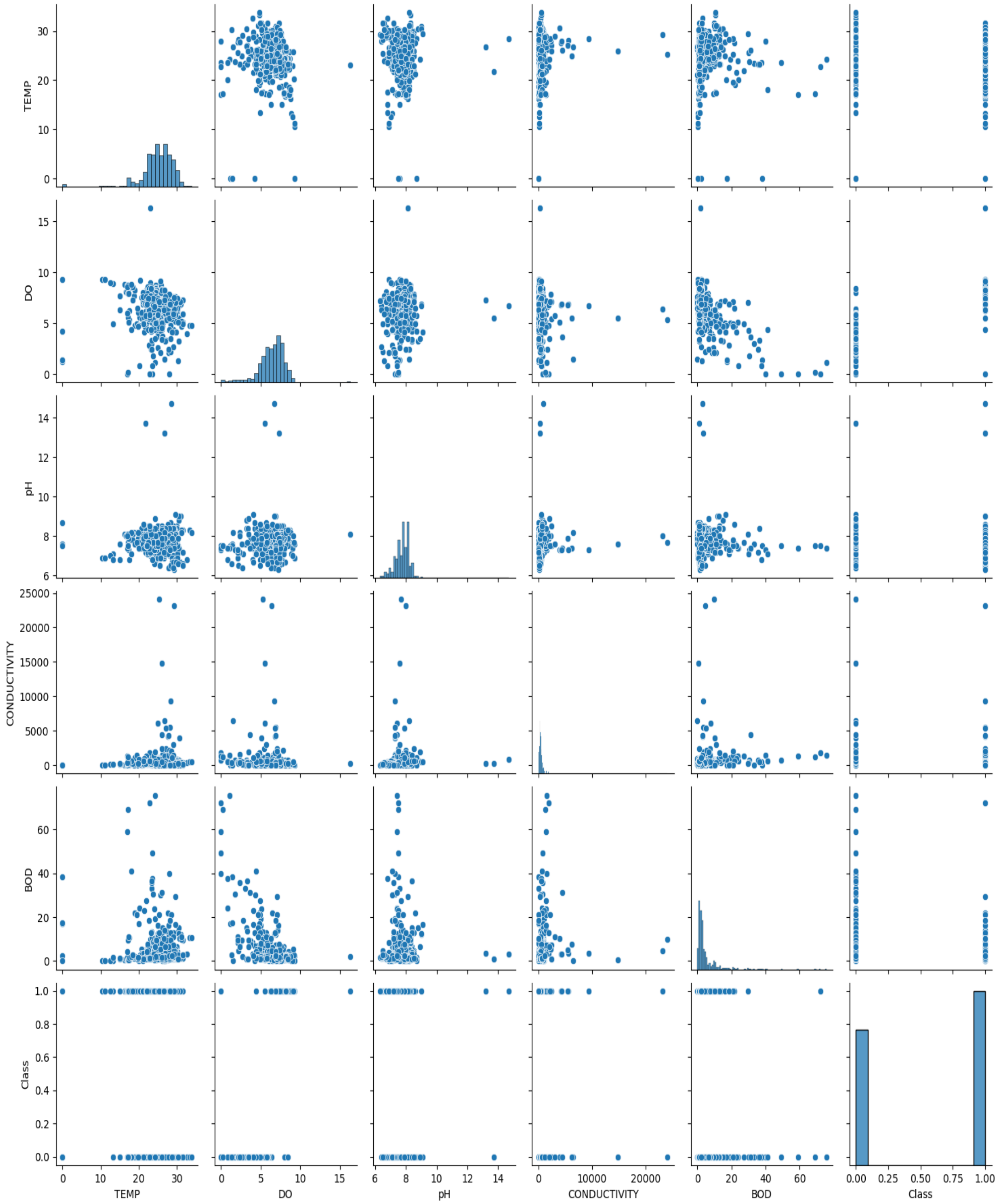


Fig.13: Scatter-plot for all the attributes

Outliers-data points that dramatically depart from the majority of observations—can be found using scatter plots. Outliers may be the result of measurement errors or they may indicate significant and intriguing data points that demand additional research.

A **pair plot** is a method for exploring the pairwise correlations between several variables in a dataset by generating a grid of scatter plots. It offers a thorough and effective way to depict the relationships between various variables, making it ideal for preliminary data exploration and locating potential correlations.

A pair plot creates a matrix of scatter plots by pairing each variable in the dataset with every other variable, including itself. Histograms or kernel density plots of each variable's distribution are typically shown on the matrix's diagonal. The link between two variables is represented by each scatter plot in the grid, where data points are depicted as dots on a two-dimensional plane in the fig:13 below.

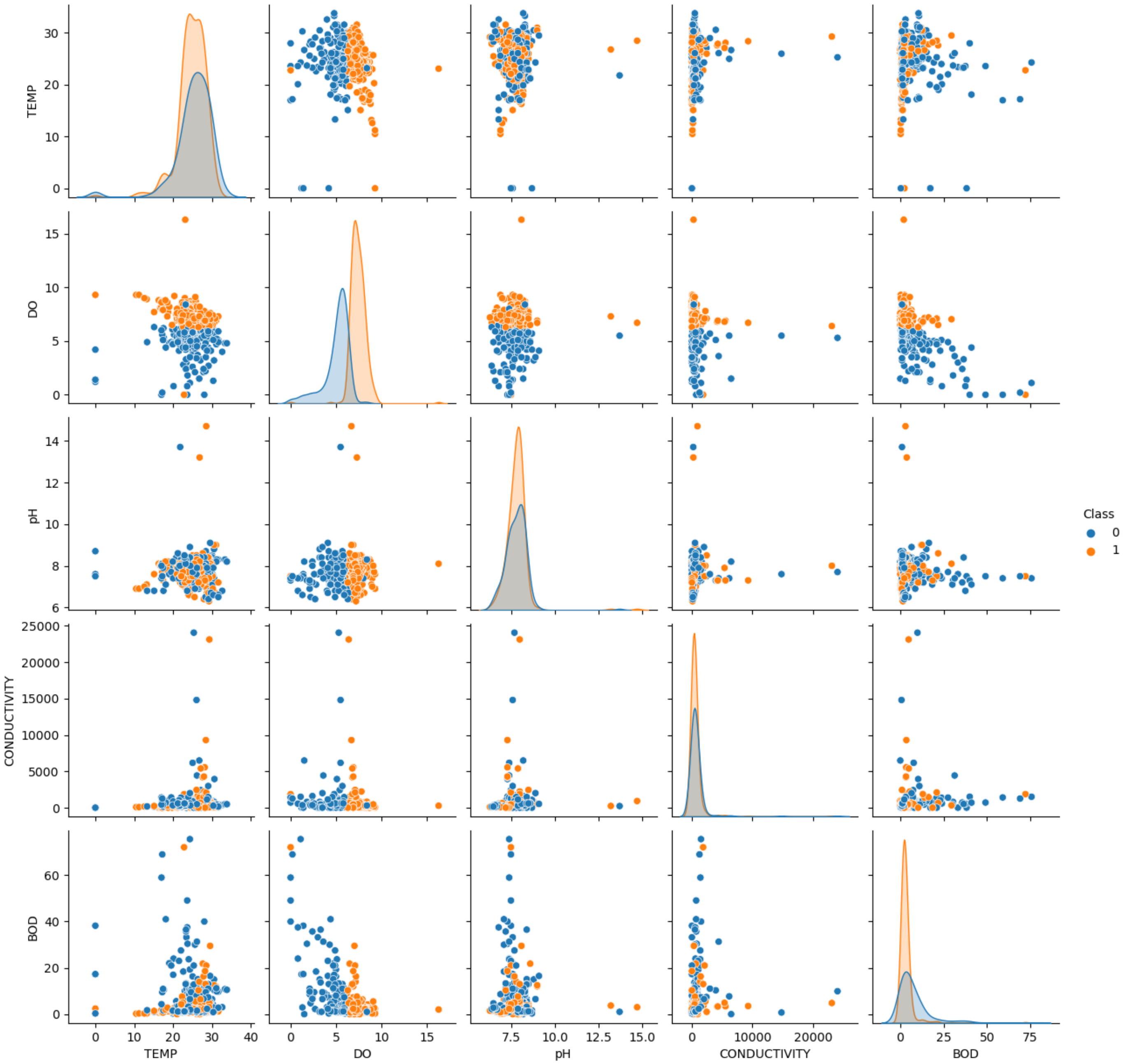


Fig.14: Pair-plot

One **histogram** will be automatically generated by the data.hist() function for each numerical column in the DataFrame. The frequency distribution of values for each histogram is displayed for the relevant column in the fig:14 shown below.The x axis here shows the values of the parameters and the y axis shows the respective frequencies of the parameter values.

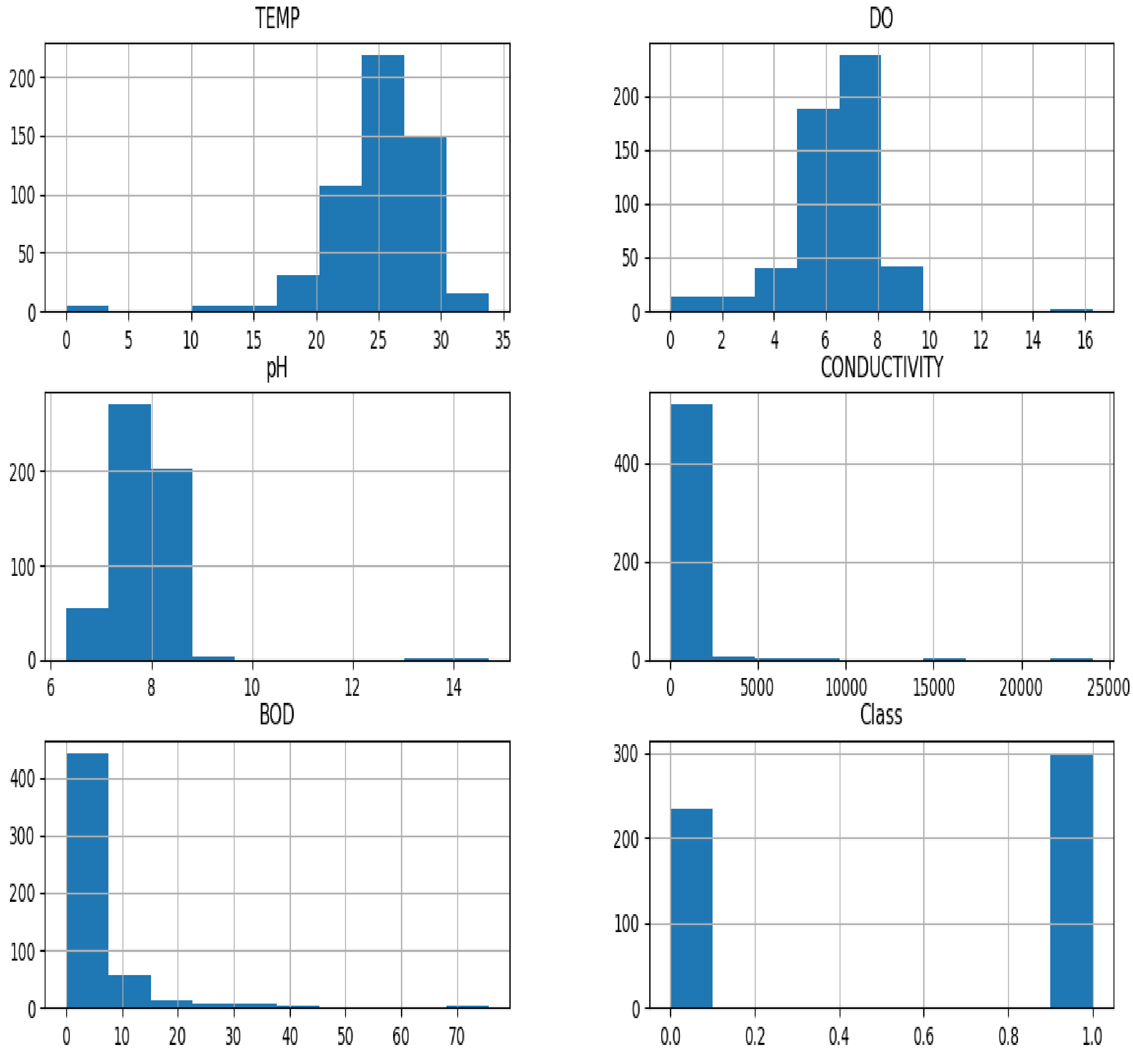


Fig.15: Histogram

The in fig 15 **outliers** in the data is observed from the box-plot diagram shown below. Data points known as outliers differ greatly from the majority of observations in a dataset. These data deviate greatly from the norm and can significantly affect statistical calculations and machine learning models. Analyzing and comprehending outliers is a crucial step in data pretreatment and analysis.

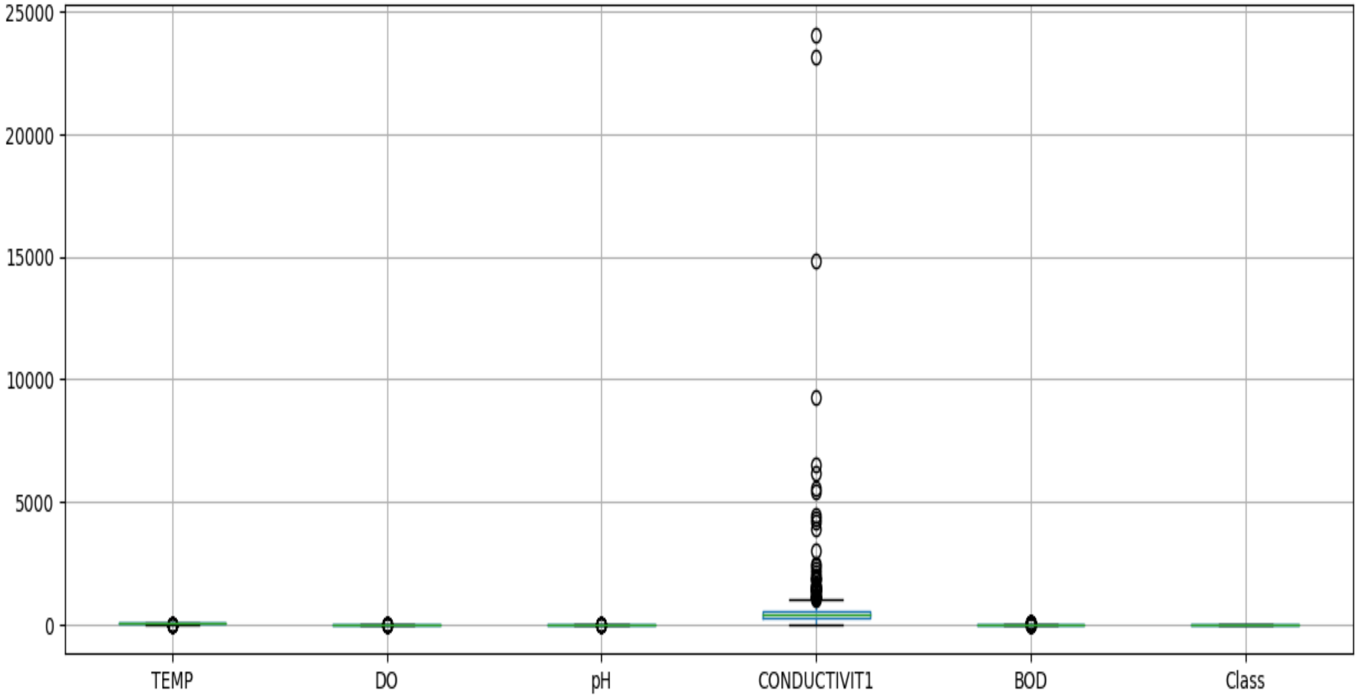


Fig.16: Box-plot for all the attributes

CHAPTER 5

RESULTS

The data of various major rivers of India is tested with three classification algorithms i.e Naive Bayes,J48 and LMT,we then observe that which algorithm is performing best and giving highest accuracy on the collected data.

1.River Beas

The data of river Beas was tested with three classification algorithm – Naïve Bayes, J48 and LMT. Firstly 20% of data was tested with these algorithms but the accuracy was not on the higher side so the complete data was tested with the above-mentioned algorithm and results were on higher side. Naïve Bayes gave the highest accuracy of 90.47%. [1]

S. No.	Algorithm	Accuracy	Time Taken to build the model
1.	Naïve Bayes	90.47%	0.01 seconds
2.	J48	76.19%	0.01 seconds
3.	LMT	85.71%	0.13 seconds

Table 2: Summary of Results of Beas River

In the decision tree given below obtained from J48 algorithm, we can observe that if the pH(Max) is greater than to 7.8 then water is not clean but if pH(Max) is less than or equal to 7.8 then if D.O.(Min) is less than or equal to 6 mg/l then water is not clean else clean.

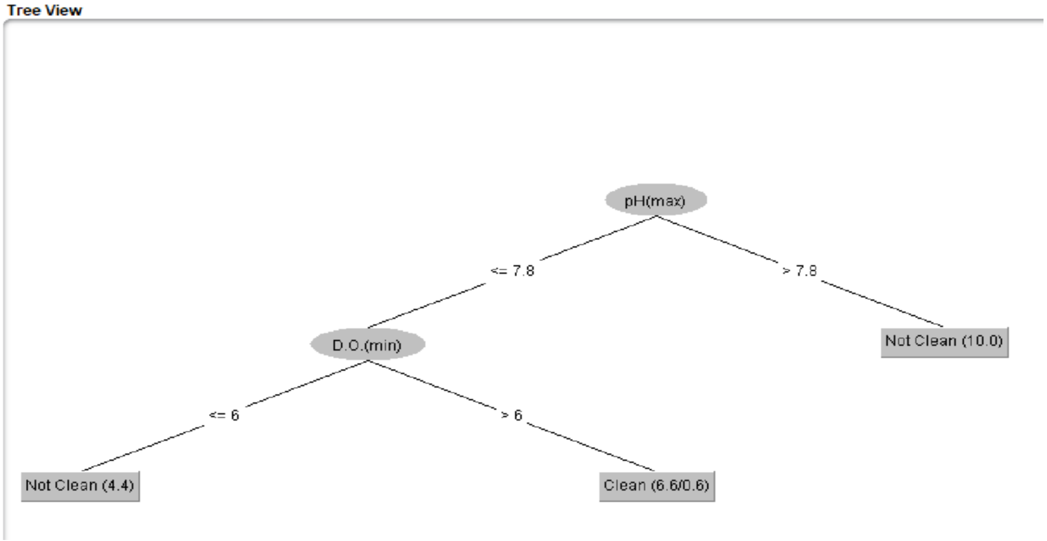


Fig.17: Decision Tree by J48

2.River Godavari

The data of river Godavari was tested with three classification algorithm – Naïve Bayes, J48 and LMT. Firstly 20% of data was tested with these algorithms but the accuracy was not on the higher side so the complete data was tested with the above-mentioned algorithm and results were on higher side. Highest accuracy was obtained by using J48 and LMT[23].

S. No.	Algorithm	Accuracy	Time Taken to build the model
1.	Naïve Bayes	94.16%	0.01 seconds
2.	J48	96.35%	0.00 seconds
3.	LMT	96.35%	0.52 seconds

Table 3: Summary of Results of Godavari River

In the decision tree given below obtained from J48 algorithm, we can observe that if the D.O.(Max) is less than or equal to 6.4 mg/l then water is not clean but if D.O.(Max) is greater than 6.4 mg/l then if Conductivity(Max) is less than 514µmho/cm then water is clean else not clean.

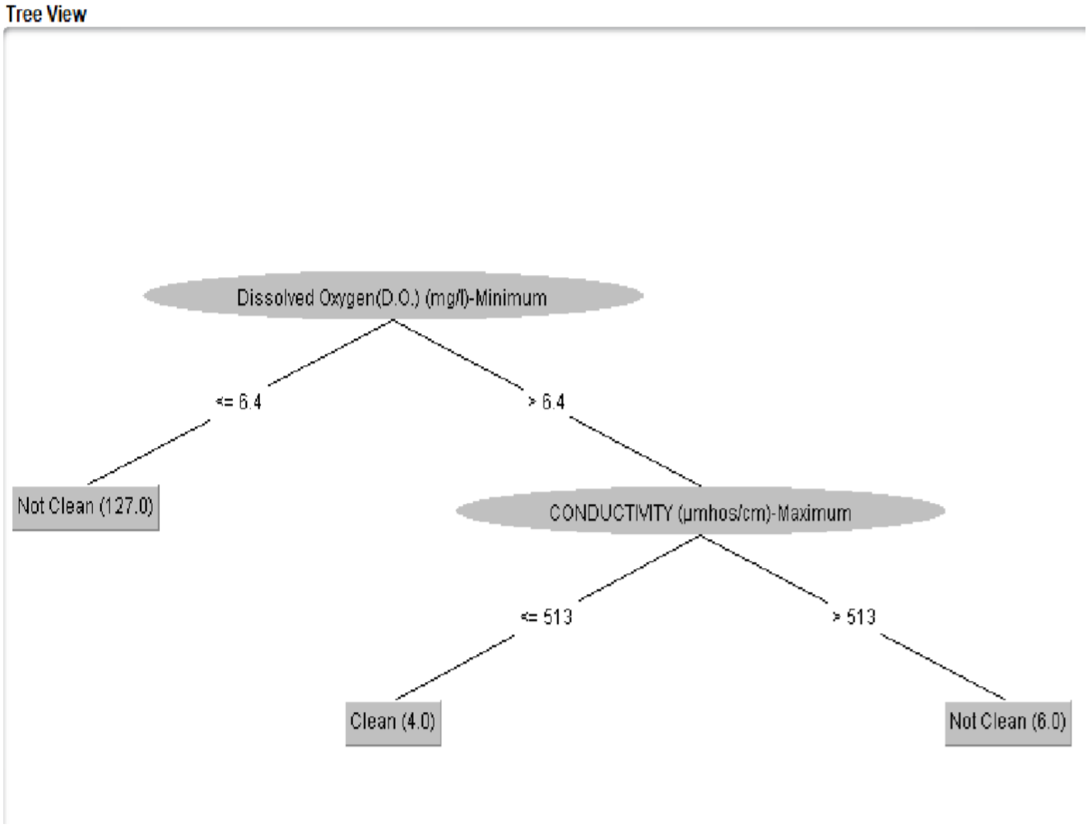


Fig.18:Decision Tree by J48

3.River Ganga

The data of river Ganga was tested with three classification algorithm – Naïve Bayes, J48 and LMT. Firstly 20% of data was tested with these algorithms but the accuracy was not on the higher side so the complete data was tested with the above-mentioned algorithm and results were on higher side. Highest accuracy was obtained by using LMT[23].

S. No.	Algorithm	Accuracy	Time Taken to build the model
1.	Naïve Bayes	83.92%	0.01 seconds
2.	J48	85.71%	0.01 seconds
3.	LMT	94.64%	0.04 seconds

Table 4: Summary of Results of Ganga River

In the decision tree given below obtained from LMT algorithm, we can observe that if the D.O.(Max) is greater than 9.0 mg/l then water is not clean but if D.O.(Max) is less than or equal to 9.0 mg/l then if Conductivity(Max) is less than 550µmho/cm then water is clean else not clean.

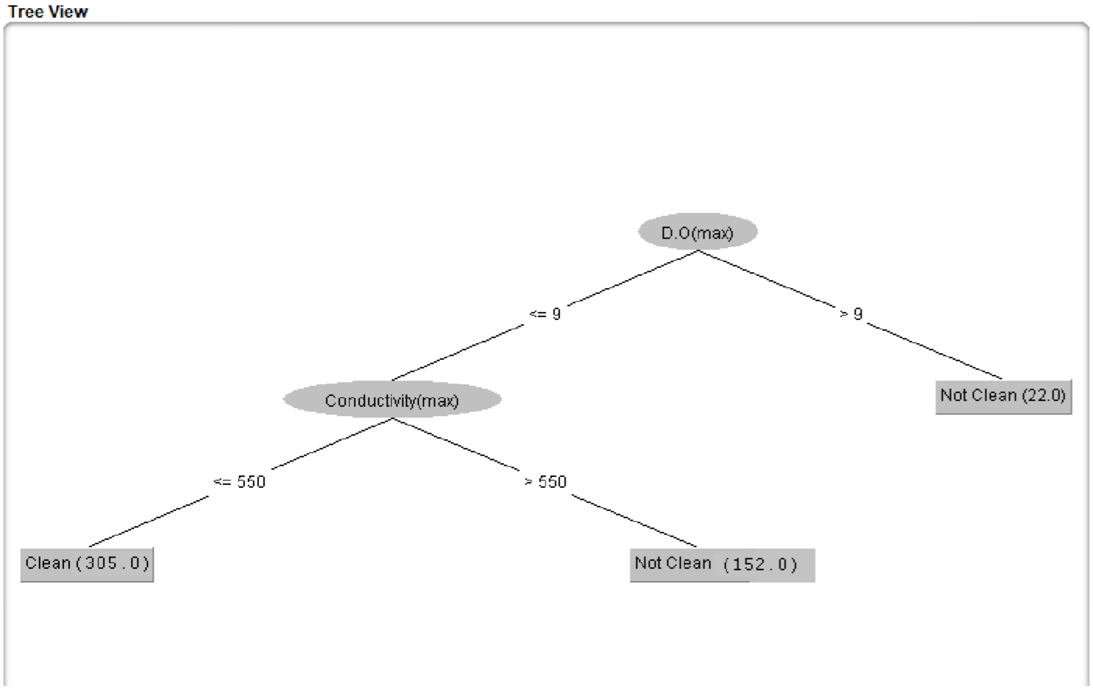


Fig.19 : Decision Tree by J48

4.River Yamuna

The data of river Yamuna was tested with three classification algorithm – Naïve Bayes, J48 and LMT. Firstly 20% of data was tested with these algorithms but the accuracy was not on the higher side so the complete data was tested with the above-mentioned algorithm and results were on higher side. Highest accuracy was obtained by using J48 [23].

S. No.	Algorithm	Accuracy	Time Taken to build the model
1.	Naïve Bayes	62.5%	0.01 seconds
2.	J48	91.66%	0.00 seconds
3.	LMT	79.16%	0.07 seconds

Table 5: Summary of Results of Yamuna River

In the decision tree given below obtained from J48 algorithm, we can observe that if the D.O.(Min) is less than or equal to 8.0 mg/l then water is not clean but if D.O.(Max) is greater 8.0 mg/l then water is clean.

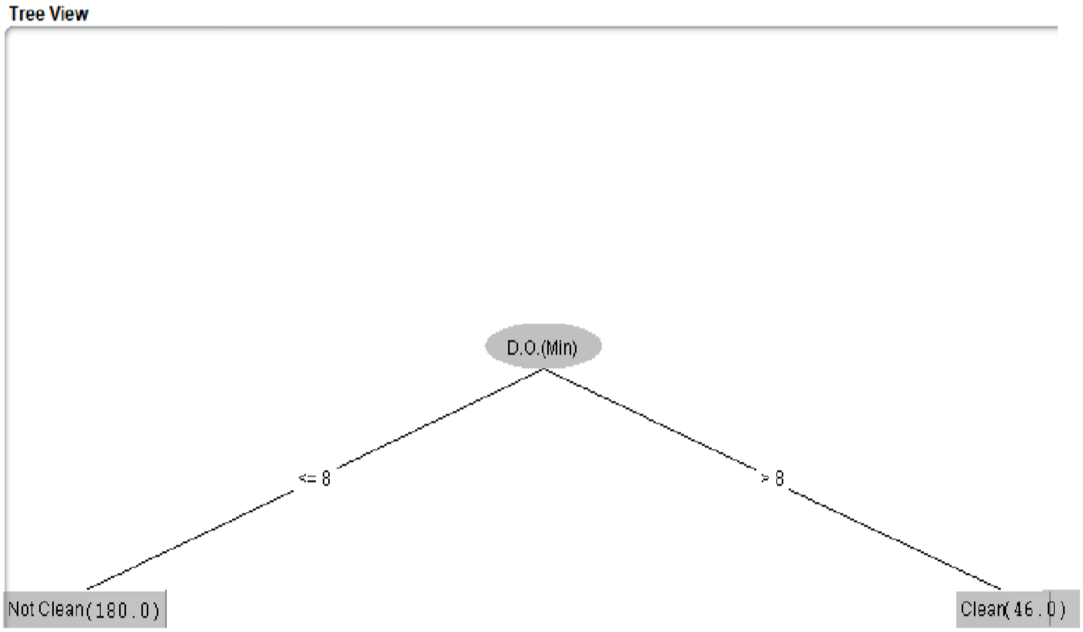


Fig.20: Decision Tree by J48

5. River Sutlej

The data of river Sutlej was tested with three classification algorithm – Naïve Bayes, J48 and LMT. Firstly 20% of data was tested with these algorithms but the accuracy was not on the higher side so the complete data was tested with the above-mentioned algorithm and results were on higher side. Highest accuracy was obtained by using J48 [23].

S. No.	Algorithm	Accuracy	Time Taken to build the model
1.	Naïve Bayes	84.37%	0.01 seconds
2.	J48	93.75%	0.01 seconds
3.	LMT	81.25%	0.01 seconds

Table 6: Summary of Results of sutlej River

In the decision tree given below obtained from J48 algorithm, we can observe that if the Temp(min) is less than or equal to 12.0 ºC then water is not clean but if Temp(min) is greater than 12.0 ºC as well as D.O.(Min) is greater than 7.2mg/l then water is clean and if Temp(min) is greater than 12.0 ºC as well as D.O.(Min) is less than or equal to 7.2mg/l then water is not clean.

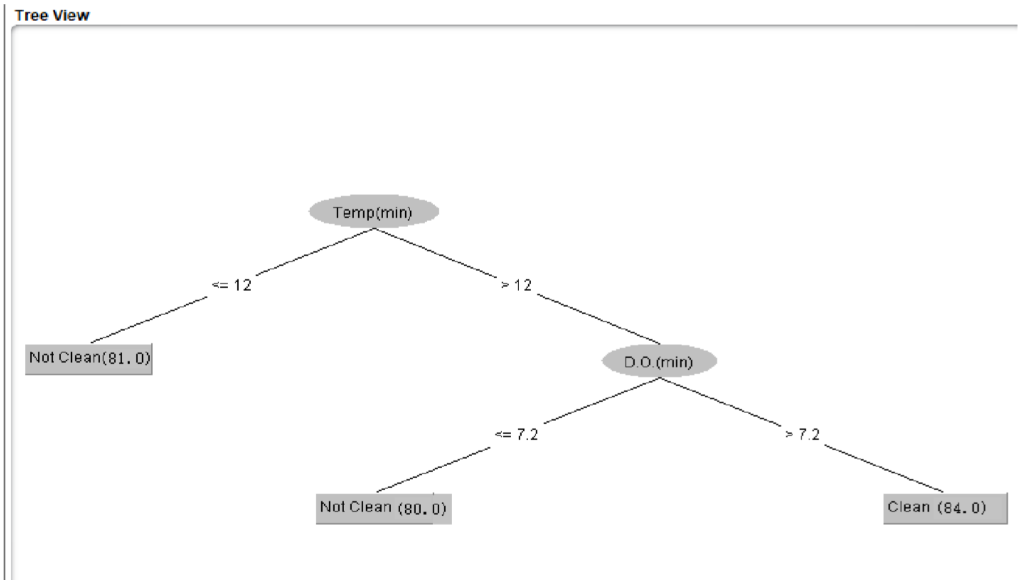


Fig.21: Decision Tree by J48

6. River Brahmaputra

The data of river Sutlej was tested with three classification algorithm – Naïve Bayes, J48 and LMT. Firstly 20% of data was tested with these algorithms but the accuracy was not on the higher side so the complete data was tested with the above-mentioned algorithm and results were on higher side. Highest accuracy was obtained by using LMT [23].

S. No.	Algorithm	Accuracy	Time Taken to build the model
1.	Naïve Bayes	70.0%	0.01 seconds
2.	J48	70.0%	0.01 seconds
3.	LMT	83.33%	0.01 seconds

Table 7 : Summary of Results of Brahmaputra River

In the decision tree given below obtained from LMT algorithm, we can observe that if the D.O.(Min) is less than or equal to 6.4 mg/l then water is clean but if D.O.(Max) is greater 8.0 mg/l then water is not clean.

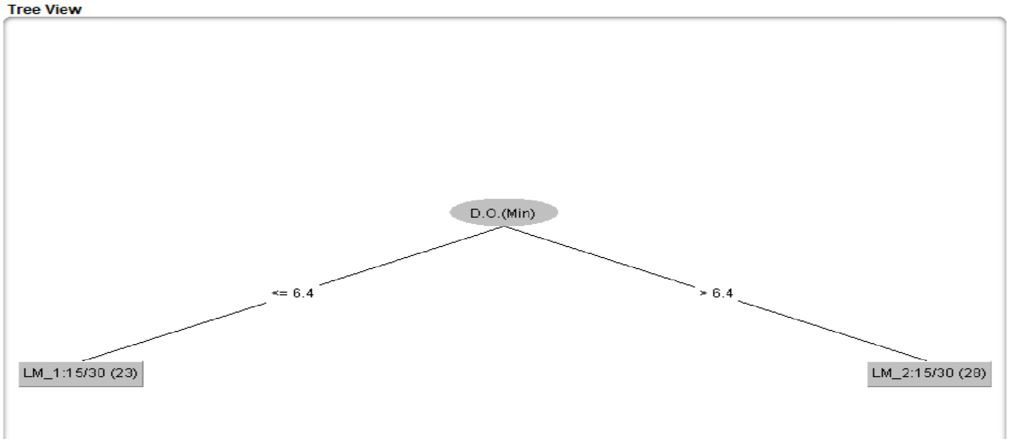


Fig.22: Decision tree by LMT

Summary of each river

Asummary of algorithms which gives best performance with respect to the data-set of each river is shown below in table:7.

S. No.	River	Classifier	Accuracy	Time Taken to Build the model
1.	Beas	Naïve Bayes	90.47%	0.01 seconds
2.	Godavari	J48	96.35%	0.52 seconds
3.	Ganga	LMT	94.64%	0.04 seconds
4.	Yamuna	J48	91.66%	0.01 seconds
5.	Sutlej	J48	93.75%	0.01 seconds
6.	Brahmaputra	LMT	83.33%	0.01 seconds

Table 8: Summary of each river

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

Decision tree technique is used in this study to implement the water quality model. The analysis of water temperature, pH level, BOD, DO and conductivity can play a major role in assessing water quality. Three different classifiers which are Naïve Bayes, J48 and LMT were used to build the model and the accuracy compared. These three classifiers were used against each river water quality data and the one with highest accuracy was used for further implications and results. It was observed that Naive Bayes algorithm gave the highest accuracy of 90.47% for data of river beas, the J48 algorithm gives highest accuracy of 96.35% for river godavai,for river Ganga LMT algorithm shows maximum accuracy which is 94.64%,for yamuna river j48 algorithm shows maximum accuracy which is 91.66%,for sutlej river J48 algorithm shows maximum accuracy which is 93.75% and for Brahmaputra river LMT shows maximum accuracy which is 83.33%.

5.2 Future Directions

Pollution stems from a wide range of sources, encompassing domestic sewage, agricultural practices, industrial activities, mining, quarrying, and cooling processes. The pollutants themselves can take various forms, such as natural organic substances, living organisms, plant nutrients, organic and inorganic chemicals, sediments, and heat.

It is crucial for the people of India to take proactive measures to safeguard and preserve the invaluable rivers of the country. In this regard, the government plays a vital role by enforcing stricter regulations on industries and factories, ensuring compliance, and preventing further pollution of rivers. By implementing and reinforcing these rules and regulations, the government can work towards preserving the health and integrity of the rivers across India.

REFERENCES

[1] Sakshi khullar,Nanhay Singh, River Water Quality Classification using a Hybrid Machine Learning Technique,9th International conference on computing for sustainable global development,IEEE,nov 20,2022.

[2] Abhishek Bajpai,Srishti Chaubey,Bdk Patro,Abhineet Verma,A real time approach to classify the water quality of the river ganga at mehandi ghat,Kannuaj,International Conference on artificial Intelligence in Engineering and technology,IEEE 2022.

[3] Preeti Chawla,Nitasha Hasteer,Prediction of Pollution potential of indian rivers using empirical equation consisting of water quality parameters,IEEE,2015,International conference on technological Innovation in ICT for agriculture and rursl development.

[4] Shailesh Jaloree,Anil Rajput,Sanjeev Gour,Decision tree approach to build a model for water quality. Binary Journal of Data Mining & Networking,vol. 4, pp. 25-28, 2014.

[5] S. P. Gorde, M. V. Jadhav, Assessment of Water Quality, S. P. Gorde et al Int. Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp.2029-2035,2019**.**

[6] Batta Mahesh,Machine Learning Algorithms ,International Journal of science and research(IJSR),Vol 9,A Review 2019.

[7] Preethi Nanjundan, Jossy P George, Aabhas Vij, A Reliable Method of Predicting Water Quality Using Supervised Machine Learning Model, IEEE International Conference on Data Science and Information System (ICDSIS),2022.

[8] Dziri Jalal and Tahar Ezzedine, Performance analysis of machine learning algorithms for water quality monitoring system,2019,IEEE.

[9] Anoop Kumar Shukla, C. S. P. Ojha and R. D. Garg, SURFACE WATER QUALITY ASSESSMENT OF GANGA RIVER BASIN, INDIA USING INDEX MAPPING,2017,IEEE.

[10] Jitha P Nair, Vijaya M S, Predictive Models for River Water Quality using Machine Learning and big data techniques-A survey, Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7,IEEE 2021.

[11] D.Kavitha, Gayathri.T.R, Dhamini Devaraj Hasitha.V, Survey on Water Quality Prediction,2023,IEEE.

[12] Preeti Chawla,Nitasha Hasteer,Prediction of Pollution potential of indian rivers using empirical equation consisting of water quality parameters,International conference on technological Innovation in ICT for agriculture and rural development,2015,IEEE.

[13] Anoop Abraham, Daniel Livingston, Izabella Guerra, Jeong Yang, Exploring the Application of Machine Learning Algorithms to Water Quality Analysis.7TH international conference on big data,cloud computing and data science(BCD),2022 IEEE.

[14] Sago Dzeroski, Jasna Grbovic, Knowledge discovery in a water quality database.

[15] D.Brindha,Vishwanath Puli,Bala Karthik Sobula,Vamsi Stephen Mittakandala,Guru Dinesh Nanneboina,Water quality analysis and prediction using machine learning,2023,IEEE

[16]Z. Kılıç, "The importance of water and conscious use of water", International Journal of Hydrology, vol. 4, no. 5, pp. 239-241, 2020.

[17]H. Choi, Y.-C. Cho, S.-H. Kin, S.-J. Yu, Y.-S. Kim and J.-K. Im, "Water Quality Assessment and Potential Source Contribution Using Multivariate Statistical Techniques in Jinwi River Watershed South Korea", Water (Switzerland), vol. 13, no. 21, pp. 2976, 2021.

[18]R. Roy, "An Introduction to water quality analysis", ESSENCE-International Journal for Environmental Rehabilitation and Conservation, pp. 94-100, 2018.

[19]M. Jaiswal, J. Hussain, S. K. Gupta, M. Nasr and A. K. Nema, "Comprehensive evaluation of water quality status for entire stretch of Yamuna River India", Environ. Monit. Assess., vol. 191, no. 4, 2019.

[20]Swapan Shakhari, Aayush Kumar Verma, Indrajit Banerjee “Remote Location Water Quality Prediction of the Indian River Ganga: Regression and Error Analysis”, 17th International Conference on ICT and Knowledge Engineering (ICT&KE),20-22 Nov.,2019.

[21]Feng-Jen Yang, “An Extended Idea about Decision Tree”,International Conference on Computational Science and Computational Intelligence (CSCI)IEEE,2019.

[22]Neha Radhakrishnan;Anju S Pillai, “Comparison of water quality classification models using machine learning”,5th International Conference on Communication and Electronics Systems (ICCES),IEEE,2020.

[23]Salisu Yusuf Muhammad, Mokhairi Makhtar, Azilawati Rozaimee, Azwa Abdul Aziz and Azrul Amri Jamal, "Classification model for water quality using machine learning techniques", International Journal of software engineering and its applications, vol. 9, no. 6, pp. 45-52, 2015.

[24]Onder Gursoy, "Determining the most appropriate classification methods for water quality", Earth and Environmental Sciences, vol. 44, 2016.

[25]Suma S;Rohit Moon;Mohammed Umer;K. Srujan Raju;Nuthanakanti Bhaskar;Rakshita Okali, “A Prediction of Water Quality Analysis Using Machine Learning”, International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE),IEEE,2023