# Data Analytical Model on Water Quality Assurance Based On the Hybrid Classification and Data Mining Techniques

A THESIS

SUBMITTED TO THE

**TILAK MAHARASHTRA VIDYAPEETH, PUNE**

FOR THE DEGREE OF

**DOCTOR OF PHILOSOPHY**

**IN COMPUTER SCIENCE**

**UNDER THE BOARD OF MODERN SCIENCES & PROFESSIONAL SKILLS**

BY

**Swapnali Dattatray Mahadik**

(25315008500)

UNDER THE GUIDANCE OF

**Dr. Anup Girdhar**

**DEPARTMENT OF COMPUTER SCIENCE**

**July 2023**

# CERTIFICATE

It is certified that the work titled "Data Analytical Model on Water Quality Assurance Based On the Hybrid Classification and Data Mining Techniques" is an original research work done by Swapnali Dattatray Mahadik, Under my supervision for the degree of Doctor of Philosophy in Computer Science Must be awarded by Tilak Maharashtra Vidyapeeth, Pune. To the best of my knowledge, this Thesis embodies the candidate's work and has not formed the basis for the award of any degree or similar title of this or any other University or examining body upon her.

Signature

**Dr. Anup Girdhar**

Research Guide

**Place:** Pune

**Date:**

# DECLARATION

I **Swapnali Dattatray Mahadik**, **PRN Number 25315008500** a Ph.D scholar of the Tilak Maharashtra Vidyapeeth in Computer Science Subject, thesis entitled "Data Analytical Model on Water Quality Assurance Based on the Hybrid Classification and Data Mining Techniques" under the supervision of Dr. Anup Girdhar. Solemnly affirm that the thesis submitted by me is my own work and has not been formed on the basis of the award of any degree or other similar title or any other university.

Signature

**Swapnali Dattatray Mahadik**

Research Scholar, TMV

(25315008500)

**Place:**

**Date:**

# Acknowledgment

The experience of carrying out this research has changed my life, and it would not have been possible without the assistance and direction I received. I am incredibly appreciative of everyone who has invested so much time, love, and energy in me.

First and foremost I am extremely grateful to my Guide, Dr.Anup Girdhar for his invaluable advice, continuous support, and patience during the study. His immense knowledge and plentiful experience have encouraged me in all the time of my research. He has supported me in every possible way since the beginning of my research. His invaluable assistance had a significant impact on how I conducted my experiment and evaluated the outcomes. Without his constant motivation and encouragement, I would not have been in a position to complete the research work.

I wish to express my sincere gratitude to Dr. Sunanda Yadav, Head PhD Department, TMV, and Dr. Asmita Namjoshi, Head of the Computer Science Department, TMV for her cooperation and assistance throughout the journey.

I am really grateful that I was given the opportunity to study at the esteemed Tilak Maharashtra Vidyapeeth, Pune. I had the chance to work with some of the most talented people. Thanks for this experience and for the opportunity.

It was an immense pleasure for me to do my research work at DES's NMITD, Mumbai with great facilities, helpful peers, and a pleasant environment.

My heartfelt thanks to my family members, in-laws, relatives, and friends for their cooperation and support in completing this journey.

I owe a debt of gratitude to my Parents Shri.Dattatray E.Mahadik and Sou.Kusum D. Mahadik for their unwavering encouragement and support, which served as a major motivator for me.

Words are not enough to express my gratitude to my husband Mr.Mahesh Jagdale , without his firm support in my life, I could not have gone so far.

Special thanks to my daughter Chi. Prisha, you are my inspiration to achieve greatness. Without you, I would not be where I am today.

**Swapnali Dattatray Mahadik**

**Research Scholar, TMV**

# List of Tables


| Sr. No | Description | Page No |
|---|---|---|
| 1.1 | Water Quality Criteria | 3 |
| 1.2 | Classification of Water Quality Index for Surface Water | 5 |
| 1.3 | Major Water Quality Parameter | 6 |
| 1.4 | Source of Water Supply | 16 |
| 2.1 | Comparison of some Common Data Mining Methods Used in Water Quality Analysis | 36 |
| 2.2 | Quality of Water | 38 |
| 2.3 | Comparative Analysis of Research Papers | 38 |
| 2.4 | Classification Hybrid Classification | 40 |
| 2.5 | Clustering Technique | 41 |
| 2.6 | Regression Analysis | 43 |
| 2.7 | Support Vector Machine and Support Vector Regression | 45 |
| 2.8 | K Nearest Neighbor (KNN) Technique for Water Quality Prediction | 46 |
| 2.9 | Reviewed Papers on K-Nearest Neighbor and RMSE Technique for Water Quality Prediction | 48 |
| 2.10 | Root Mean Square Error (RMSE) technique | 50 |
| 3.1 | Difference between the Five Models | 59 |
| 3.2 | Correlation Based Values | 61 |
| 3.3 | Scales Used for Data Measurements | 63 |
| 3.4 | Detailed Demography Of Survey Questionnaire | 64 |
| 3.5 | Reliability Statistics | 65 |
| 3.6 | Count of Locations of Respondents | 69 |
| 3.7 | Water Intake per Day in Respondents Organizations | 70 |
| 3.8 | Count of Origin of The Source of Drinking Water | 71 |
| 3.9 | Percentage of Sick Leave | 72 |
| 3.10 | Count of TDS Level Identified in the Organizations | 73 |

| | | |
|---|---|---|
| 3.11 | Opted Filtered Water in the Organizations | 74 |
| 3.12 | PH Level of Drinking Water Within the Organization | 75 |
| 3.13 | Water Source with the Highest Quality | 76 |
| 3.14 | The budget Allotted for Water Purification | 77 |
| 3.15 | Water Analyzed from Laboratory | 78 |
| 3.16 | Process To Improve Water Quality | 79 |
| 3.17 | Responsible for Drinking Water Quality | 80 |
| 3.18 | The Process to Preserve Drinking Water | 81 |
| 3.19 | Interviewer Organization Details | 85 |
| 3.20 | Type of Organization | 86 |
| 3.21 | Classification of Type Of Water as per the use In Organizations | 87 |
| 3.22 | Trust Factor Analysis | 88 |
| 3.23 | Idea about pH level of Drinking Water | 89 |
| 3.24 | Idea about TDS level of Drinking Water | 89 |
| 3.25 | Count for Reasons for Sick Leave | 90 |
| 3.26 | Frequency of Testing the Water Quality in Your Organization | 91 |
| 3.27 | Overall Budget Allocated in Company towards Purified Drinking Water | 92 |
| 3.28 | Effective Complaint Mechanisms | 93 |
| 3.29 | Water Sample Details Of Organizations | 95 |
| 3.30 | Statistical Details of Regression | 97 |
| 4.1 | Sample Dataset 1 | 112 |
| 4.2 | Statistical Algorithm Implementation | 113 |
| 4.3 | Accuracy Matrix | 124 |
| 4.4 | Water Quality Parameters | 126 |
| 4.5 | Water Parametric Values | 130 |
| 4.6 | Accuracy Matrix of Water Sample Stored in Stainless Steel | 132 |
| 4.7 | Accuracy Matrix of Water Sample Stored in Mud Matka | 133 |
| 4.8 | Accuracy Matrix of Water Sample Stored in Plastic | 134 |
| 4.9 | Accuracy Matrix of Water Sample Stored in Copper Vessels | 134 |

| | | |
|---|---|---|
| 4.10 | Accuracy Matrix of Tap Water | 135 |
| 4.11 | Scaling Factor of Severity | 137 |
| 4.12 | Mapping of Water Sample with Parameters | 140 |
| 5.1 | Comparative Analysis of Accuracy Score | 152 |

# List of Figures


| Sr. No | Description | Page No |
|---|---|---|
| 1.1 | Maharashtra River Basin Map | 2 |
| 1.2 | Sources of Water Pollution | 3 |
| 1.3 | The average frequency of various WQI categories across WQMS in the corresponding sub-basins of Maharashtra | 5 |
| 1.4 | Map of Mumbai | 21 |
| 3.1 | Location of organizations | 69 |
| 3.2 | Water Intake Per day | 70 |
| 3.3 | Source of water | 71 |
| 3.4 | Percentage of sick leave | 72 |
| 3.5 | TDS Level Identified in the Organizations | 73 |
| 3.6 | Opted Filtered Water in the Organizations | 74 |
| 3.7 | PH level of Drinking Water | 75 |
| 3.8 | Water Source with the Highest Quality | 76 |
| 3.9 | The budget allotted for Water purification | 77 |
| 3.10 | Water analyzed from Laboratory | 78 |
| 3.11 | The Process to Improve Water Quality | 79 |
| 3.12 | Responsibility for Drinking Water Quality | 80 |
| 3.13 | The Process to Preserve Drinking Water | 81 |
| 3.14 | Type of Organizations Covered During Interview | 86 |
| 3.15 | Classification of Type of Water as per the Use in Organizations | 87 |
| 3.16 | Trust Factor Analysis | 88 |
| 3.17 | The idea about pH level of Drinking Water | 89 |
| 3.18 | The idea about TDS level of Drinking Water | 90 |
| 3.19 | Count for Reasons for Sick Leave | 91 |
| 3.20 | Frequency of Testing the Water Quality in Your Organization | 92 |
| 3.21 | Overall Budget Allocated in the Company towards Purification | 93 |

| | | |
|---|---|---|
| 3.22 | Effective Complaint Mechanisms | 93 |
| 3.23 | Line Fit Plot | 96 |
| 3.24 | Prediction Interval | 96 |
| 3.25 | Residual Plot | 97 |
| 3.26 | Distribution Graph significance level(α) | 98 |
| 3.27 | Residual Histogram | 99 |
| 3.28 | Residual QQ Plot | 99 |
| 3.29 | Sample Water Analysis Report | 101 |
| 3.30 | Proposed NavJal model | 106 |
| 4.1 | System Architecture | 109 |
| 4.2 | Logistic Regression | 115 |
| 4.3 | Decision tree | 116 |
| 4.4 | Random forest | 118 |
| 4.5 | XGBoost Classifier | 119 |
| 4.6 | Using K nearest Neighbour Classifier | 120 |
| 4.7 | Support Vector Machine | 122 |
| 4.8 | Adaboost Implementation | 123 |
| 4.9 | comparative Analysis of all techniques according to Accuracy Score | 124 |
| 4.10 | Water Collection in Steel Vessel | 128 |
| 4.11 | Water Collection in Plastic | 128 |
| 4.12 | Water Collection in Copper | 128 |
| 4.13 | Water Collection in Mud Matka | 129 |
| 4.14 | Local Brands of Drinking Water Available in the Mumbai Region | 129 |
| 4.15 | Mumbai Suburbs Location-wise Water Samples | 129 |
| 4.16 | Analysis of Two parameters (pH and TDS) Based on Storage patterns | 131 |
| 4.17 | Comparative Analysis of Water Sample Stored in Stainless Steel | 132 |
| 4.18 | Comparative analysis of Water Sample stored in Mud Matka | 133 |
| 4.19 | Comparative Analysis of Water Sample Stored in Plastic | 134 |
| 4.20 | Comparative analysis of Water Sample stored in Copper Vessels | 135 |
| 4.21 | Comparative Analysis of Tap Water | 136 |

| | | |
|---|---|---|
| 5.1 | Parametric Representation | 148 |
| 5.2 | Correlation Matrix 1 | 149 |
| 5.3 | Potability Representation 1 | 150 |
| 5.4 | Potability Representation 2 | 150 |
| 5.5 | Correlation Matrix 2 | 151 |
| 5.6 | Comparative Analysis of Accuracy Score | 153 |

# List of Abbreviations

| Abbreviation | Explanation |
|---|---|
| AI | Artificial Intelligence |
| ANFIS | Adaptive Neural Fuzzy Inference System |
| ANN | Artificial Neural Network |
| ARIMA | Auto-Regressive Integrated Moving Average |
| As | Arsenic |
| BOD | Biological Oxygen Demand |
| $C_6H_5OH$ | Phenolic Compounds |
| Ca | Calcium |
| Cd | Cadmium |
| CGWB | Central Ground Water Board |
| Cl | Chloride |
| CN | Cyanide |
| COD | Chemical Oxygen Demand |
| COO | Chief Operating Officer |
| CPCB | Central Pollution Control Board |
| Cu | Copper |
| DO | Dissolved Oxygen |
| EC | Electrical Conductivity |
| ESR | Environmental Status Report |
| F | Fluoride |
| Fe | Iron |
| FIS | Fuzzy Inference System |
| GDWQ | Guidelines for Drinking-Water Quality |
| GOV | Government |
| GP | Genetic Programming |
| Hg | Mercury |
| HOD | Head of the Department |

| | |
|---|---|
| IOT | Internet of Things |
| KNN | K-Nearest Neighbor |
| MCDM | Multi-Criteria Decision-Making |
| MCGM | Municipal Corporation of Greater Mumbai |
| Mn | Manganese |
| MPCB | Maharashtra Pollution Control Board |
| $NO_3$ | Nitrate |
| NSF | National Science Foundation |
| NSF | National Sanitation Foundation |
| NSFWQI | National Sanitation Foundation Water Quality Index |
| NTU | Nephelometric Turbidity Units |
| Pb | Lead |
| pH | potential of Hydrogen |
| RBDT | Rule-Based Decision trees |
| RMSE | Root Mean Square Error |
| Se | Selenium |
| $SO_4$ | Sulfate |
| SVM | Support Vector Machines |
| TA | Total Alkalinity |
| TDS | Total Dissolved Solids |
| TH | Total Hardness |
| WHO | World Health Organization |
| WQC | Water Quality Classification |
| WQI | Water Quality Index |
| WQMS | Water Quality Monitoring Station |
| Zn | Zinc |

# INDEX


| Sr. No | | Particulars | Page No. |
|---|---|---|---|
| **Chapter 1: Introduction** | | | |
| 1.1 | | Introduction | 1 |
| 1.2 | | Overview | 2 |
| | 1.2.1 | Overall Study | 2 |
| | 1.2.2 | Water Quality Standards | 5 |
| | 1.2.3 | Indian Standards For Drinking Water | 6 |
| 1.3 | | Research Background | 9 |
| | 1.3.1 | Data Mining | 11 |
| | 1.3.2 | Data Mining Algorithms | 12 |
| | 1.3.3 | Types of Data Mining Techniques | 13 |
| 1.4 | | Proposed Methodology | 14 |
| 1.5 | | Problem Statement | 15 |
| 1.6 | | Objectives of the Study | 17 |
| 1.7 | | Research gap Analysis | 18 |
| 1.8 | | Opportunities and Challenges | 18 |
| 1.9 | | Need and Benefits of Study | 20 |
| 1.10 | | Study Area | 21 |
| 1.11 | | Chapter Scheme | 22 |
| | | | |
| **Chapter 2: Literature Review** | | | |
| 2.1 | | Introduction | 24 |
| 2.2 | | Challenges and an Evaluation of the Quality | 25 |
| 2.3 | | Water Quality Evaluation | 26 |
| 2.4 | | Water Quality Forecasting | 27 |
| 2.5 | | The Hydrology of Artificial Intelligence and Its Applications | 32 |

| | | | |
|---|---|---|---|
| 2.6 | | Essential data mining methods consist of categorization, regression, and AIML. | 34 |
| 2.7 | | Water Quality Index | 37 |
| 2.8 | | Technology-wise Comparative analysis of research papers | 38 |
| 2.9 | | Gap Analysis from the Literature Review | 50 |
| | | | |
| **Chapter 3:Research Methodology** | | | |
| 3.1 | | Introduction | 52 |
| | 3.1.1 | The Objective of the Research Study | 52 |
| | 3.1.2 | Significance of Research Study | 53 |
| | 3.1.3 | Need of Research Work | 54 |
| | 3.1.4 | Benefits of Research Work | 55 |
| 3.2 | | Research Methodology | 55 |
| | 3.2.1 | Type of Research | 56 |
| 3.3 | | Research Population and Sample | 60 |
| | 3.3.1 | Sampling Technique | 60 |
| | 3.3.2 | Calculating Sample size | 61 |
| 3.4 | | Questionnaire Formation | 64 |
| 3.5 | | Reliability and Validity Test of Questionnaire | 64 |
| 3.6 | | Sample Questionnaire on Drinking Water Quality | 65 |
| 3.7 | | Analysis Report of Data Collected through Survey | 69 |
| 3.8 | | Key Understandings of Survey | 82 |
| 3.9 | | Sample Questionnaire for Interviewer | 83 |
| 3.10 | | Analysis Report of Interviews | 84 |
| 3.11 | | Key Understandings of Interviews | 94 |
| 3.12 | | Linear Regression for two variables pH and TDS | 94 |
| 3.13 | | Lab Test Analysis | 101 |
| 3.14 | | Methods and Simulation Tools used for Analysis | 102 |
| 3.15 | | Proposed Framework for Water Quality Assurance | 103 |

| | | | |
|---|---|---|---|
| 3.16 | | Data Accuracy, Parameter selections, Computation Time, and Results | 107 |
| **Chapter 4: Analysis and Interpretation** | | | |
| 4.1 | | Introduction | 108 |
| 4.2 | | System Architecture | 108 |
| 4.3 | | Evaluation of the proposed framework | 111 |
| | 4.3.1 | Logistic regression | 114 |
| | 4.3.2 | Decision Tree Classifier | 115 |
| | 4.3.3 | Random Forest | 117 |
| | 4.3.4 | XGBoost Classifier | 118 |
| | 4.3.5 | K nearest Neighbor | 119 |
| | 4.3.6 | Support Vector Machine | 121 |
| | 4.3.7 | Using AdaBoost Classifier | 122 |
| 4.4 | | Hypothesis Testing | 125 |
| 4.5 | | Conclusion | 141 |
| **Chapter 5:Conclusion, Findings and Recommendations** | | | |
| 5.1 | | Introduction | 143 |
| 5.2 | | Key Findings | 143 |
| 5.3 | | Proposed Framework: NavJal Model Pros and Cons | 144 |
| 5.4 | | Suggestions on the Proposed Framework | 145 |
| 5.5 | | Scope for Future Work | 145 |
| 5.6 | | Findings of the Implementation of NavJal Model | 147 |
| 5.7 | | Validation of proposed Framework | 152 |
| 5.8 | | Research Contribution | 153 |
| 5.9 | | Limitations of The Research | 154 |
| | | | |
| | | **References** | 156 |
| | | | |
| | | **Questionnaire** | 163 |

# Abstract

**“When the Well is dry, we know the worth of water”**

**Benjamin Franklin**

Every cell in the human body contains water, which is a component that is needed on a daily basis. Water makes up about 60% of the human body and is the basic liquid of life. Water is essential for all kinds of life and the foundation of all existence, according to the online desk of India Today. There wouldn't be any life on Earth without it; it is a necessary nutrient. Water is one of the most precious and necessary resources, and it is vital to the functioning of the human body. We can go for up to a few weeks without eating, but only a few days without drinking.

Water Pollution is becoming an important issue in India nowadays. According to Maharashtra Pollution Control Board (MPCB), there are some predefined water quality criteria that define the quality, Such as ph, total coliforms Organism, Dissolved oxygen, etc. At present MPCB monitors 48 surface water and 25 ground water locations in Maharashtra. Similarly, The Central Pollution Control Board (CPCB) is the body in the field of water quality management in India, which provides data so it is necessary to build a data model to monitor and analyze the water quality based on some important specific parameters. The quality of water may be described in terms of the concentration and dissolved state of some organic and inorganic materials present in the water. On the basis of that criteria, the Proposed Model collects and analyzes the data which will help to monitor the quality of water.

Proposed Framework NavJal is developed after going through the Literature Review and analysis of Surveys and questionnaires. This model is divided into 3 modules namely Data Gathering and Data Preprocessing, Data Training, and Model Creation i.e. NavJal. Throughout the process, Data mining plays an important role.

The proposed research pertains to the development of a data analytical model, which will be functional in the collection of water quality factors or parameters and to analyze them for giving a better quality of water. So the major benefit of the Proposed Data Model is to overcome the various issues till date researchers have applied and provided the best model to analyze the water quality based on the Data Mining techniques.

# Chapter 1

# Introduction

# Chapter 1: Introduction

## 1.1 Introduction

The World Health Organization (WHO) defines water contamination as any alteration to the physical, chemical, or biological characteristics of water that has an adverse effect on living things. Despite the fact that water covers about 70% of the earth's surface, water pollution is a major worldwide issue because fresh water is so rare. Pollution-related deterioration of the water quality from these sources will cause water shortages, ecosystem degradation, and adverse effects on human health.

Water pollution refers to the contamination of water bodies, such as rivers, lakes, oceans, and groundwater, by harmful substances or pathogens. These contaminants can be in the form of chemicals, microorganisms, or other pollutants, and can have negative impacts on human health, the environment, and wildlife.

Sources of water pollution can include industrial and agricultural activities, improper disposal of waste and sewage, oil and chemical spills, and atmospheric deposition. When these pollutants enter the water bodies, they can cause a variety of problems, such as reduced oxygen levels, harmful algal blooms, and the accumulation of toxic substances in fish and other aquatic animals.

Water pollution can have significant consequences for human health, including the spread of waterborne diseases such as cholera, typhoid, and dysentery. It can also impact the economy by reducing the availability of clean water for agricultural, industrial, and domestic use, and by affecting the tourism industry in areas affected by water pollution.

Preventing water pollution requires a combination of regulatory measures, pollution control technologies, and public education and awareness. Efforts to reduce water pollution include the treatment of wastewater before discharge, the implementation of best management practices in agriculture and industry, and the development of sustainable infrastructure and land-use practices. By working together to reduce water pollution, we can protect our water resources and ensure a healthy environment for future generations.

## 1.2 Overview

### 1.2.1 Overall Study

Water is the most important natural resource for a state or a country and the entire humanity. A nation's wealth is largely dependent on the wise use of this resource. Thus, it can be stated that the main wealth of the nation is water, which flows in rivers and streams.

India's rivers, which are highly revered by its citizens on a religious level, are among its most distinctive characteristics and in Maharashtra it is more.

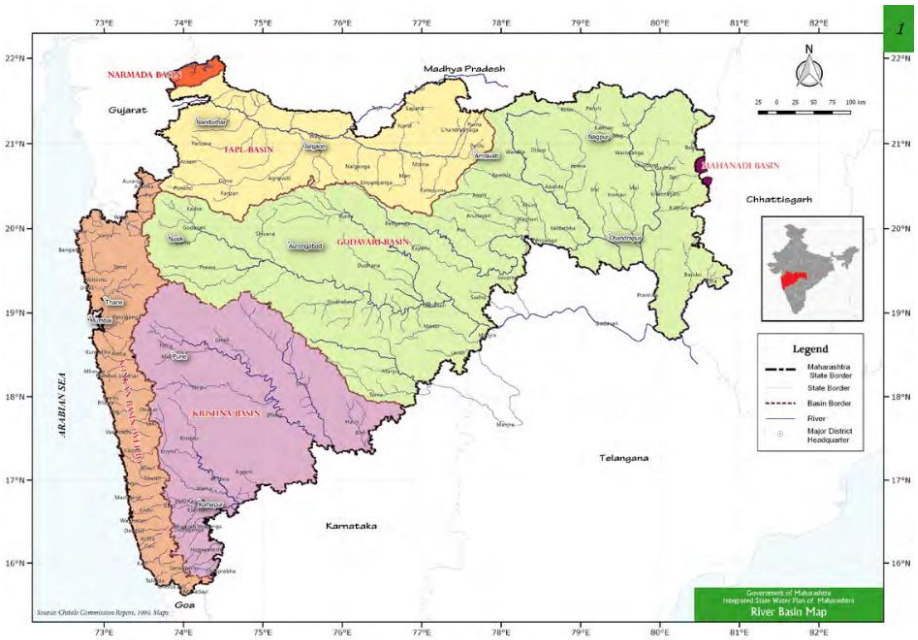


**Figure 1.1: Maharashtra River Basin Map**
**[**Source: Maps Maharashtra Government Maps**]**

Maharashtra rivers, which span a huge area and they have contributed significantly to the states rural prosperity. Its numerous rivers are extremely valuable to Maharashtra as well as India because of their broader significance in the development of culture, economy, geography, and religion.

The socio-economic growth of any region may severely depends on unavailability of safe water. Keeping this in mind maintenance of water quality is very important factor.

In Maharashtra because of higher pollution and having a metro cities like Mumbai ,Pune ,Nasik etc water pollution chances are more. so Consuming such polluted water can spread illnesses like Polio, Malaria, Typhoid, Dysentery, Diarrhea, and Cholera.

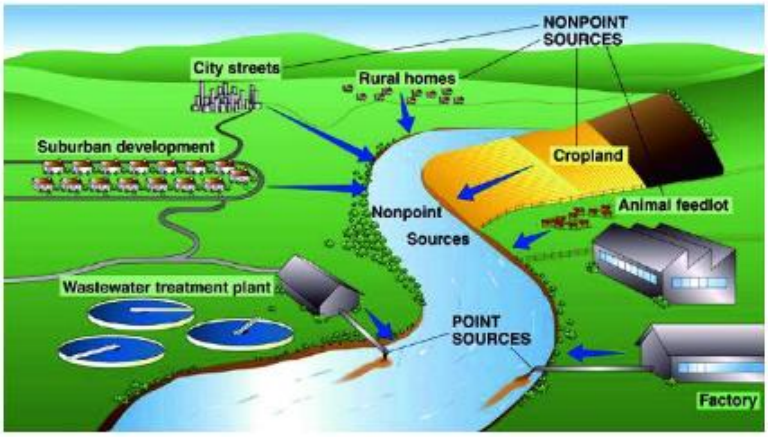


**Figure 1.2: Sources of Water Pollution**

[Source: MPCB Water Quality Status of Maharashtra 2020-21]

According to Central Pollution Control Board (CPCB) Water Quality Criteria's are defined as shown in the table 1.1 on the basis of their designated Best use.

**Table 1.1 Water Quality Criteria**

[Source: https://cpcb.nic.in/wqm/Designated_Best_Use_Water_Quality_Criteria.pdf]

| Designated-Best-Use | Class of water | Criteria |
|---|---|---|
| Drinking WaterSource without conventional treatment but after disinfection | A | Total Coliforms Organism MPN/100ml shall be 50 or less |
| | | pH between 6.5 and 8.5 |
| | | Dissolved Oxygen 6mg/l or more |
| | | Biochemical Oxygen Demand 5 days 20C 2mg/l or less |
| Outdoor bathing (Organised) | B | Total Coliforms Organism MPN/100ml shall be 500 or less pH between 6.5 and 8.5 Dissolved Oxygen 5mg/l or more |
| | | Biochemical Oxygen Demand 5 days 20C 3mg/l or less |

| | | |
|---|---|---|
| Drinking water source after conventional treatment and disinfection | C | Total Coliforms Organism MPN/100ml shall be 5000 or less pH between 6 to 9 Dissolved Oxygen 4mg/l or more |
| | | Biochemical Oxygen Demand 5 days 20C 3mg/l or less |
| Propagation of Wild life and Fisheries | D | pH between 6.5 to 8.5 Dissolved Oxygen 4mg/l or more |
| | | Free Ammonia (as N) 1.2 mg/l or less |
| Irrigation, Industrial Cooling, Controlled Waste disposal | E | pH betwwn 6.0 to 8.5 |
| | | Electrical Conductivity at 25C micro mhos/cm Max.2250 |
| | | Sodium absorption Ratio Max. 26 |
| | | Boron Max. 2mg/l |
| | Below-E | Not Meeting A, B, C, D & E Criteria |

As shown in above table water has been divided into five different classes. Accordingly, the parametric values are different. These five different classes is as below:

- Drinking water source without conventional treatment and disinfection
- Outdoor Bathing (Organized)
- Drinking water source after conventional treatment and disinfection
- Propagation of Wild Life Fisheries
- Irrigation, Industrial Cooling, Controlled Waste Disposal

MPCB has taken into account a number of factors when testing and analyzing surface water samples in the years 2020–2021. However for calculating WQI four parameters namely pH, Dissolved Oxygen, BOD and FC are taken into consideration. This helps in conveying about water quality status effectively.

The range of WQI is been defined by MPCB is as shown in table 1.2. Also figure 1.3 shows the comparison between the WQI recorded at a monitoring site and the typical frequency with which the WQI was of a given category across all the WQMS in that basin.

**Table 1.2 Classification of Water Quality Index for Surface Water**

[source: https://www.mpcb.gov.in/]

| WQI | Quality Classification | Remarks |
|---|---|---|
| 63 to 100 | Good to Excellent | Non-Polluted |
| 50 to 63 | Medium to Good | Non-Polluted |
| 38 to 50 | Bad | Polluted |
| 38 and less | Bad to very Bad | Heavily Polluted |

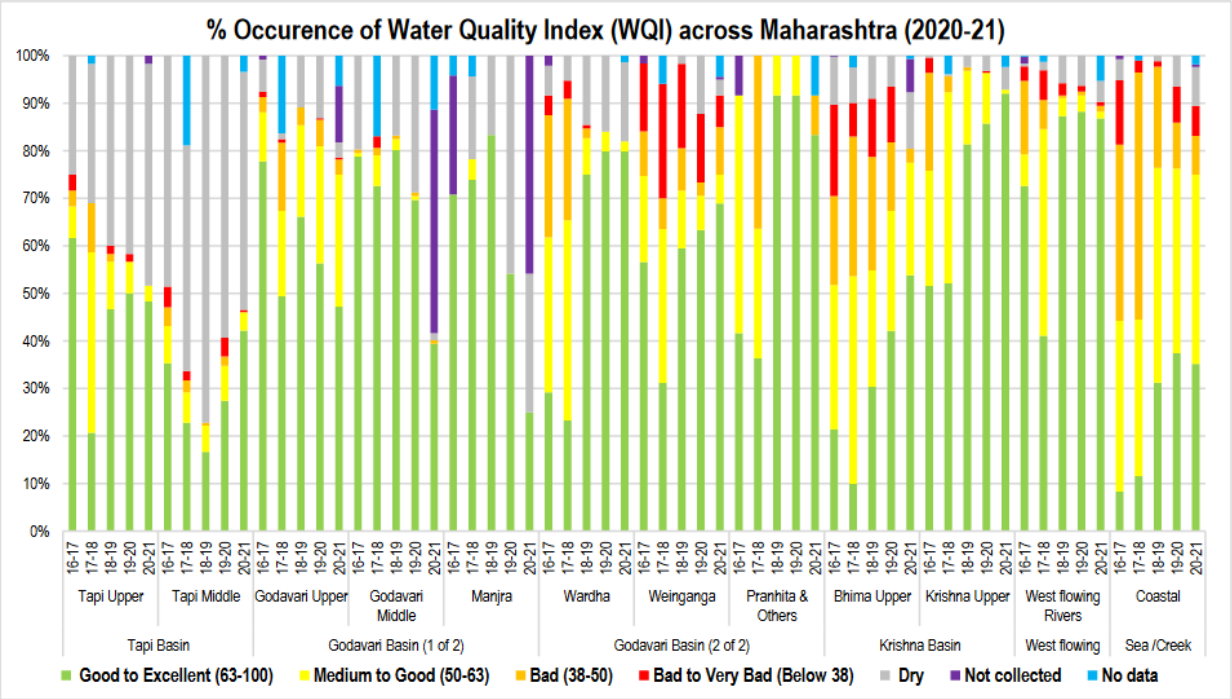


**Figure 1.3 Average frequency of various WQI categories across WQMS in the corresponding sub-basins of Maharashtra**

[Source: https://www.mpcb.gov.in/sites/default/files/wqr_report_04052022.pdf]

### 1.2.2 Water Quality Standards

The World Health Organization (WHO) provides a standard for drinking water quality around the world in Guidelines for Drinking-water Quality which is currently in its fourth edition. This publication provides guidelines for a vast array of chemical, microbial, and

radiological contaminants commonly found in drinking water. The WHO has health-based targets for many of the contaminants. A health-based target is defined as, “measurable health, water quality or performance objectives that are established based on a judgment of safety and on risk assessments of waterborne hazards.” Guidelines for Drinking-water Quality also provide a framework for achieving safe drinking water by implementing health-based targets, creating a water safety plan, and maintaining water surveillance.

The World Health Organization's (WHO) Guidelines for Drinking-Water Quality (GDWQ) fourth edition builds on more than 50 years of WHO guidance on drinking-water quality, which has served as a reliable foundation for the adoption of national water safety regulations and standards in support of public health.(WHO Report4th Edition]

### 1.2.3 Indian Standards for drinking water

The Bureau of Indian Standards prescribes the quality of drinking water in its BIS 10500-1991 standards. It lists three major quality parameters- Physical, Chemical and Biological.

SPECIFICATION - ( BIS 10500 : 1991 )

**Table 1.3 Major Water Quality Parameter**
**[Source: Drinking Water-Specification**
**http://cgwb.gov.in/documents/wq-standards.pdf]**

| Sr No | Substance or Characteristic | Requirement (Desirable Limit) | In the lack of an alternative source, the permissible limit |
|---|---|---|---|
| **Essential characteristics** | | | |
| 1 | Color, ( Hazen units, Max ) | 5 | 25 |
| 2 | Odor | Unobjectionable | Unobjectionable |
| 3 | Taste | Agreeable | Agreeable |
| 4 | Turbidity ( NTU, Max) | 5 | 10 |

| | | | |
|---|---|---|---|
| 5 | pH Value | 6.5 to 8.5 | No Relaxation |
| 6 | Total Hardness (as $CaCo_3$) mg/lit.,Max | 300 | 600 |
| 7 | Iron (as Fe) mg/lit,Max | 0.3 | 1 |
| 8 | Chlorides (as Cl) mg/lit,Max. | 250 | 1000 |
| 9 | Residual,free chlorine,mg/lit,Min | 0.2 | -- |
| **Desirable Characteristics** | | | |
| 10 | Dissolved solids mg/lit,Max | 500 | 2000 |
| 11 | Calcium (as Ca) mg/lit,Max | 75 | 200 |
| 12 | Copper (as Cu) mg/lit,Max | 0.05 | 1.5 |
| 13 | Manganese (as Mn)mg/lit,Max | 0.1 | 0.3 |
| 14 | Sulfate (as $SO_4$) mg/lit,Max | 200 | 400 |
| 15 | Nitrate (as $NO_3$) mg/lit,Max | 45 | 100 |
| 16 | Fluoride (as F) mg/lit,Max | 1.9 | 1.5 |
| 17 | Phenolic Compounds (as $C_6H_5OH$) mg/lit, Max. | 0.001 | 0.002 |
| 18 | Mercury (as Hg)mg/lit,Max | 0.001 | No relaxation |

| | | | |
|---|---|---|---|
| 19 | Cadmium (as Cd)mg/lit,Max | 0.01 | No relaxation |
| 20 | Selenium (as Se)mg/lit,Max | 0.01 | No relaxation |
| 21 | Arsenic (as As) mg/lit,Max | 0.05 | No relaxation |
| 22 | Cyanide (as CN) mg/lit,Max | 0.05 | No relaxation |
| 23 | Lead (as Pb) mg/lit,Max | 0.05 | No relaxation |
| 24 | Zinc (as Zn) mg/lit,Max | 5 | 15 |
| 25 | Anionic detergents (as MBAS) mg/lit,Max | 0.2 | 1 |
| 26 | Chromium (as $Cr^{6+}$) mg/lit,Max | 0.05 | No relaxation |
| 27 | Polynuclear aromatic hydro carbons (as PAH) g/lit,Max | -- | -- |
| 28 | Mineral Oil mg/lit,Max | 0.01 | 0.03 |
| 29 | Pesticides mg/l, Max | Absent | 0.001 |
| 30 | Radioactive Materials | | |
| | i. Alpha emitters Bq/l,Max | -- | 0.1 |
| | ii. Beta emitters pci/l,Max | -- | 1 |
| 31 | Alkalinity mg/lit.Max | 200 | 600 |
| 32 | Aluminium (as Al) mg/l,Max | 0.03 | 0.2 |
| 33 | Boron mg/lit,Max | 1 | 5 |

## 1.3 Research Background

Hydrology is a critical field of study that focuses on understanding the occurrence, distribution, movement, and properties of water in the environment. This includes the study of water resources, water quality, and the impact of water on the environment and human activities.

The water cycle, or hydrologic cycle, is a fundamental process in hydrology that describes the continuous movement of water between the atmosphere, land, rivers, groundwater, and oceans. The water cycle includes processes such as evaporation, precipitation, infiltration, runoff, and groundwater recharge. Understanding the water cycle is crucial for managing water resources, predicting floods and droughts, and mitigating the impacts of climate change.

Hydrological studies are essential for managing water resources sustainably and protecting the environment. These studies help to identify the sources of water and the patterns of water movement, which can inform decisions about water allocation and management. Hydrological studies also help to assess the impacts of human activities on the water cycle and to develop strategies for mitigating these impacts.

Today, hydrology is more important than ever, as climate change, population growth, and increasing demands for water resources are putting pressure on our water systems. Hydrological studies are critical for ensuring that we have access to clean and safe water, protecting the environment, and supporting sustainable development.

And for that Water pollution testing methods and techniques used to determine the quality of water and the presence of contaminants in it.

There are several methods used for water pollution testing, including

1. **Physical testing**: This method involves the use of physical parameters, such as temperature, color, and odor, to identify the presence of contaminants in water.
2. **Chemical testing**: This method involves the use of chemical reagents to identify the presence and concentration of specific chemicals or pollutants in water.

3. **Biological testing:** This method involves the use of living organisms, such as fish, to determine the toxicity of the water and the presence of harmful substances.
4. **Microbiological testing**: This method involves the use of microbiological tests to detect the presence of harmful bacteria and viruses in water.
5. **Remote sensing:** This method involves the use of satellite images and other remote sensing techniques to detect water pollution in large bodies of water.
6. **Online monitoring:** This method involves the use of sensors and other online monitoring devices to continuously measure the quality of water and detect changes in water quality.

The selection of the appropriate testing method depends on the nature and severity of the pollution, the type of water being tested, and the desired outcomes. A combination of different testing methods is often used to provide a comprehensive assessment of water quality.

Water quality assurance is critical for ensuring public health and environmental sustainability. Developing an effective analytical model for water quality assurance is a complex task that requires a combination of classification and data mining techniques.

One approach to building such a model is to use a hybrid classification and data mining framework. In this framework, the data is first classified based on specific water quality parameters, such as pH, temperature, and dissolved oxygen. The data is then mined using techniques such as clustering, association rule mining, and decision trees to identify patterns and relationships between the different parameters.

The model should also consider factors that may impact water quality, such as weather, land use, and sources of pollution. This can be achieved through the use of geographic information systems (GIS) and spatial analysis techniques.

Once the model is built, it can be used to identify areas of concern for water quality and to develop strategies for improving water quality. The model can also be used to predict changes in water quality based on changes in weather patterns, land use, and other factors.

Overall, a hybrid classification and data mining framework can be an effective approach to developing a data analytical model for water quality assurance. By incorporating multiple techniques and factors, the model can provide a comprehensive understanding of water quality and support the development of effective strategies for ensuring clean and safe water for all.

### 1.3.1 Data Mining

Data mining is the process of discovering patterns, relationships, and insights from large datasets. It involves using statistical and computational methods to analyze large datasets and uncover hidden patterns and trends that may not be apparent through traditional analytical methods.

Data mining techniques can be used in a wide range of fields, including finance, marketing, healthcare, and environmental science. In environmental science, for example, data mining techniques can be used to analyze data related to water quality, air quality, climate change, and biodiversity, among other things. Evaluation of the data mining methods reflected the high accuracy and performance of these methods in classifying the water quality.( Sattari, Mohmmd, Feizi,2021)

Some of the common techniques used in data mining include association rule mining, clustering, classification, and prediction. Association rule mining involves identifying relationships between different variables in a dataset, while clustering involves grouping data points based on their similarities. Classification involves categorizing data into predefined groups or classes, while prediction involves using past data to predict future outcomes. Classification is an instance of supervised learning that includes a training phase to create a model. Its task is to predict the class of items in a data set using a certain model of a classifier.( Consolata Gakii and Jeniffer Jepkoech,June 2019)

Data mining can provide valuable insights and inform decision-making in a wide range of applications. For example, in the context of water quality management, data mining techniques can be used to identify sources of water pollution, predict the likelihood of water

contamination, and inform the development of effective water quality management strategies.

However, it is important to note that data mining is not a one-size-fits-all approach and that the selection of appropriate techniques will depend on the specific nature of the data being analyzed and the goals of the analysis.

### 1.3.2 Data Mining Algorithm

The selection of the appropriate algorithm should be based on the type and size of the data, the complexity of the problem, and the desired outcome. However, some of the commonly used and popular data mining algorithms are:

1. **Decision Trees**: This algorithm is used for classification and prediction problems. It is simple to understand and interpret and can handle both numerical and categorical data.
2. **Random Forest**: This algorithm is an ensemble of decision trees and is used for classification and regression problems. It can handle large datasets and is less prone to overfitting.
3. **K-Means Clustering**: This algorithm is used for unsupervised learning problems and is used to group similar data points together. It is fast and efficient and can handle large datasets.
4. **Support Vector Machines** (SVM): This algorithm is used for classification and regression problems. It is effective in high-dimensional spaces and can handle non-linear data.
5. **Neural Networks:** This algorithm is used for classification and prediction problems. It is effective in handling complex and non-linear data and can learn from large datasets.

The selection of the best algorithm for a specific problem depends on the specific data and the problem at hand. It is important to try different algorithms and compare their results to select the best one for a specific task.

### 1.3.3 Types of Data Mining Techniques

There are several data mining techniques that can be used to analyze water quality data. Some of the most commonly used techniques are:

1. **Association Rule Mining**: This technique is used to find patterns and relationships among different water quality parameters. It helps to identify how the presence or absence of one parameter may affect the presence or absence of another parameter.
2. **Clustering**: This technique is used to group water quality parameters that have similar characteristics.It he huge data into a group of similar objects. (S. Brintha Rajakumari, C. Nalini, 2016) It helps to identify different groups of parameters that can be used to describe the water quality of a particular region.
3. **Classification**: A classification algorithm is used to assign predefined classes to test instances for evaluation or future instances to an application. ( Consolata Gakii and Jeniffer Jepkoech,June 2019)
4. **Regression**: This technique is used to predict the values of one water quality parameter based on the values of other parameters. It helps to identify the relationship between different parameters and how they affect the overall water quality.
5. **Decision Trees**: Decision tree is a predictive self-learning model that determines the target value of a new sample based on various attribute values of the available data.( Consolata Gakii and Jeniffer Jepkoech,June 2019) It helps to identify the most important parameters and their optimal range to maintain good water quality.
6. **Neural Networks**: This technique is used to simulate the human brain to identify patterns in large datasets. It helps to identify complex relationships between different water quality parameters and predict future trends.

The proposed data analytical model can use a combination of these techniques to provide a comprehensive analysis of water quality data and help to improve water quality management.

## 1.4 Proposed Methodology

The development of a data analytical model for water quality analysis based on data mining techniques can offer several benefits over existing approaches. By utilizing advanced analytical methods, the proposed model can provide more accurate and comprehensive insights into water quality parameters and their relationships.

This can be particularly useful in addressing the limitations of traditional water quality monitoring methods, which may not be able to capture the full range of contaminants and pollutants present in water bodies. The proposed model can help to identify and monitor a wider range of contaminants, including emerging contaminants such as pharmaceuticals, personal care products, and endocrine disruptors.

Furthermore, the proposed model can provide a more cost-effective and efficient approach to water quality monitoring, by automating the process of data collection and analysis. This can help to reduce the time and resources required for manual monitoring and allow for more frequent and comprehensive monitoring of water quality parameters.

In addition, the proposed data model can be useful in identifying and predicting the sources and causes of water pollution, which can help in the development of effective strategies for pollution control and prevention. This can be particularly useful in areas where water quality is a significant concern, such as in urban and industrial areas.

The development of a data analytical model for water quality analysis based on data mining techniques can provide significant benefits in terms of accuracy, efficiency, and effectiveness, and can support the goal of ensuring clean and safe water for all.

For the proposed model, samples of minerals and different types of metals are evaluated for various locations and were analyzed for determination of the degree of quality with respect to some of the following physicochemical parameters and heavy metals opted for investigation.

i. Ph
ii. Turbidity

iii. Hardness
iv. Total dissolved Solids
v. Conductivity
vi. Chloramine
vii. Sulfate
viii. Carbon
ix. Trihalomethane
x. Calcium
xi. Fluroid
xii. Mineral Oil

And Heavy Metals are:

i. Iron
ii. Cadmium
iii. Copper
iv. Zinc
v. Manganese
vi. Lead

## 1.5 Problem Statement

The commercial center of India is considered Mumbai. A wide range of businesses and a significant amount of job opportunities have been created in the city because of the ongoing growth of its infrastructure, industries, and commercial units. Many people came to Mumbai in pursuit of work from western Maharashtra, the Konkan, and other states. Because of ongoing migration, the city is now too crowded.

Greater Mumbai as a whole is an urban region. All of the significant surface water reservoirs are used for water supply in the neighboring districts that are based on significant rivers. These rivers include the Vaitarna and Ulhas, which have their source in the Konkan area, along with the Patalganga and Amba.

Although four of these rivers do not pass through Greater Mumbai, their basins still provide the majority of the surface water used for household and industrial purposes in Greater Mumbai.

Six lakes—Tulsi, Vihar, Tansa, Upper Vaiterna, Bhatsa, and Mumbai III—provide water to Mumbai City.. (Ground Water Information report of Government of India ministry of water resources central ground water board, Year 2013)

As per the environment status report of 2021-22 of Mumbai, there is 3850 MLD water supply for Mumbai. (Environment Status report 2020-21)

**Table 1.4 Source of Water Supply**

**[source: Mumbai Corporation Environment Status report 2020-21]**

| Sr No | Source | Year | Yield in MLD | | Distance from City in KM | Remarks |
|---|---|---|---|---|---|---|
| | | | | Commutative | | |
| 1 | Vihar | 1860 | 90 | 90 | Within City | Present Source |
| 2 | Tulsi | 1872 | 18 | 108 | Within City | |
| 3 | Tansa | 1892-1945 | 500 | 608 | 106 | |
| 4 | Lower Vaitarna | 1594 | 455 | 1063 | 119 | |
| 5 | Upper Vaitarna | 1972 | 635 | 1698 | 163 | |
| 6 | Bhatsa | 1980-2007 | 2020 | 3718 | 102 | |
| 7 | Middle Vaitarna | 2014 | 455 | 4173 | 150 | |
| 8 | Gargai | 2022-2027 | 440 | 4613 | 180 | Future Source |
| 9 | Pinjal | 2024-2025 | 865 | 5478 | 195 | |
| 10 | Damanganga | 2029-2030 | 1586 | 7064 | | |

The problem statement is a clear, concise, and specific statement that describes the issue or challenge, which may face because of water quality. In the case of the proposed data analytical model for water quality assurance based on the hybrid classification and data mining techniques, a possible problem statement could be:

"Water quality monitoring and management is becoming an increasingly complex and challenging task due to the large number of pollutants and environmental factors that affect it.

Current data mining techniques used in water quality assessment often lack the ability to accurately classify and predict water quality based on multiple parameters. As a result, there is a need for an advanced data analytical model that can effectively analyze and classify water quality parameters using hybrid classification and data mining techniques to ensure the provision of clean and safe water to the population."

This problem statement clearly highlights the need for a more effective and accurate approach to water quality analysis and classification, which can be achieved through the proposed data analytical model.

### 1.6 Objective of the Study:

Research Objectives provides a detailed description of what the research is attempting to accomplish. This research outcome work as a guideline to the overall research

1. Identify and study a set of water parameters to make an inventory to check water quality.
2. To study the impact of external resources on water quality
3. To design a model which carries out Data Mining techniques on water parameters to verify the water quality
4. To identify the factors which improve water purity and make water a healthy energy drink.
5. To improve the efficiency and effectiveness of the water using Data Mining and Classification techniques for better outcomes.

## 1.7 Research gap Analysis

Overall Research Review suggests that Data Mining Techniques are useful to segregate the water quality parameters and to analyze it for water quality by applying different algorithms like Clustering, Classification, etc. But the majority of the pre-existing techniques have certain restrictions and problems.

The identified research gap in the current study is that the existing data mining techniques used for water quality analysis have certain limitations and problems. Some of these limitations may include:

1. **Lack of accuracy:** The existing techniques may not provide accurate results due to the complex nature of water quality parameters and the presence of multiple pollutants.
2. **Limited data sets:** The existing techniques may be based on limited data sets, which may not provide a comprehensive understanding of water quality.
3. **Single algorithm approach:** Some of the existing techniques may be based on a single algorithm approach, which may not be sufficient to accurately analyze and classify water quality parameters.
4. **Limited focus:** The existing techniques may be focused on specific water quality parameters, which may not provide a complete understanding of the overall water quality.

Therefore, there is a need to develop an advanced data analytical model that can overcome these limitations and provide accurate and reliable water quality analysis. The proposed model aims to address these limitations by using a hybrid classification and data mining approach to analyze multiple water quality parameters and provide a comprehensive understanding of water quality.

## 1.8 Opportunities and Challenges

Monitoring water quality in the 21st century presents both opportunities and challenges. On one hand, advances in technology and data analytics provide new opportunities for more efficient and effective monitoring of water quality parameters. For example, the use

of remote sensing, sensor networks, and data mining techniques can help to identify patterns and trends in water quality data that may not be apparent through traditional monitoring methods.

However, there are also significant challenges associated with water quality monitoring. One of the main challenges is the large number of chemicals and pollutants that are used in our everyday lives. Monitoring all of these pollutants and their sources can be a daunting task, particularly given the limited resources available for monitoring.

Another challenge is the need to manage nutrients across the water and achieve related goals. Nutrient pollution from sources such as agricultural runoff, wastewater treatment plants, and urban stormwater can have significant impacts on water quality and aquatic ecosystems. Managing nutrient pollution requires effective monitoring and management strategies that can be challenging to implement on various scales of water, from local streams to large rivers and lakes.

Another challenge is the cost of water quality monitoring. While there are many effective monitoring programs in place, such as the U.S. Geological Survey monitoring sites, budget reductions can lead to fewer monitoring stations, which can reduce the overall effectiveness of monitoring efforts. To address this challenge, it may be necessary to develop new, more cost-effective monitoring strategies, such as the use of low-cost sensors and citizen science initiatives.

In summary, while there are challenges associated with water quality monitoring, there are also opportunities for new and innovative approaches to monitoring that can help to address these challenges and ensure clean and safe water for all.

Environmental conditions such as temperature, precipitation, and other factors can have a significant impact on water quality, and these factors need to be taken into account when interpreting water quality data. This is because water quality can vary over time and in different locations, and environmental factors can affect the behaviour and concentration of pollutants in water.

One of the main benefits of the proposed data analytical model is that it can help to monitor and analyze water quality based on multiple parameters, including environmental

conditions. By collecting and analyzing data on a wide range of parameters, the model can provide a more comprehensive and accurate picture of water quality, which can help to inform decision-making and improve management strategies.

### 1.9 Need and Benefits of Study:

The study of water quality assurance using data mining and hybrid classification techniques is essential for a number of reasons.

Firstly, water is a vital resource for human survival and a key component of many ecosystems. Ensuring that our water resources are of high quality and free from harmful contaminants is essential for protecting public health and supporting a wide range of economic and social activities.

Secondly, water quality is affected by a wide range of factors, including industrial and agricultural activities, urbanization, and climate change. Effective monitoring and management of water quality is therefore a complex and challenging task that requires a multifaceted approach.

The proposed data analytical model can help to address these challenges by providing a more comprehensive and accurate picture of water quality based on multiple parameters. By analyzing large volumes of data using data mining and hybrid classification techniques, the model can help to identify patterns and trends in water quality data that may not be apparent through traditional monitoring methods. This can help to inform management strategies and enable more effective decision-making.

In addition, the proposed model has the potential to reduce the workload on monitoring personnel and increase the efficiency and consistency of data collection and analysis. This can help to reduce costs and improve the overall effectiveness of water quality monitoring programs.

Overall, the study of water quality assurance using data mining and hybrid classification techniques is essential for ensuring the sustainability and resilience of our water resources and has the potential to bring significant benefits to society and the environment.

## 1.10 Study Area

Greater Mumbai is situated on Maharashtra State's westernmost border. Seven islands, including Bombay, Mazgaon, Matunga, Mahim, Worli, Soyster Rode, and Old Woman's Island(little Colaba), made up the city's initial foundation.

Over time, these islands were connected to the mainland by a number of reclamation techniques, such as the filling of small creeks. By including suburban regions in the north and along Thane District's perimeter, the area was further expanded.( Maps_Maharashtra Governmet Maps)

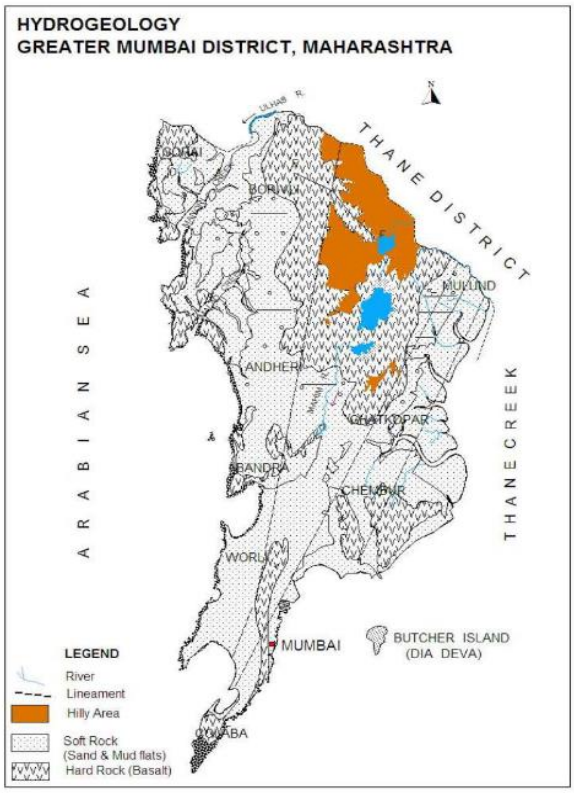


**Figure 1.4: Map of Mumbai**

**[Source**: https://cgwb.gov.in/District_Profile/Maharashtra/Greater%20Mumbai.pdf]

### 1.11 Chapter Scheme

The proposed chapter scheme for the research synopsis seems reasonable and includes most of the essential components that are typically included in Overall research. Here's a brief overview of each chapter:

The chapter scheme provided includes the following five chapters:

**Chapter 1**: Introduction - This chapter typically provides an overview of the research problem and objectives. It also includes a brief background of the research topic and an explanation of the significance of the research. Also it provides Research Gap analysis , needs and benefits of analysis.

**Chapter 2**: Literature Review - This chapter typically provides a comprehensive review of the existing literature on the research topic. It includes a discussion of the theories and concepts related to the research problem, as well as a review of previous studies that have been conducted on the topic.

**Chapter 3:** Research Methodology - This chapter typically describes the research methods and techniques used to collect and analyze data for the study. It includes a discussion of the sampling strategy, data collection methods, and data analysis techniques used in the study. This chapter also provides detailed outline of Proposed Model.

**Chapter 4:** Implementation of NavJal Model - This chapter typically presents the results of the study and provides a detailed analysis of the data collected. It includes tables, figures, and other visual aids to help present the data in a clear and understandable manner. Also, it gives a brief experimental analysis of seven algorithms i.e. Logistic Regression, Decision Tree Classifier, Random Forest, K nearest neighbor, XGBoost Classifier, Ada Boost Classifier and Support Vector Machine.

**Chapter 5:** Findings, Conclusion, and Future Scope- This chapter typically provides a summary of the research findings and conclusions drawn from the study. This chapter gives a brief review of the validation of the proposed model. It also includes a discussion of the implications of the research, recommendations for future research, and limitations of the study.

# Chapter 2

# Literature Review

# Chapter 2: Literature Review

## 2.1 Introduction

Maharashtra, India's third-largest state by area, is also the world's third most populous country. Mumbai, the capital of Maharashtra City, is the most populous city in Maharashtra City, with several rivers nearby. Mumbai is situated between 18o 53' and 19o 16' North Latitude and between 72o and 72o 59' East Longitude on the western seacoast of India. Mumbai experiences a tropical savanna environment. In general, the South-West monsoon comes in the city in June and departs in September. The temperature and climate directly/indirectly affect the water quality. In addition, Mumbai, one of the major cities in the globe, is also known for having the highest population density. Mumbai's populace is continuing to increase. The environment is severely impacted by the inverse relationship between area and people. (MCGM ESR English 2020 – 2021 Report).

Early in the 1990s, Maharashtra City saw the slow start of the industrial revolution, and new industries are still expanding across the nation, especially in the Mumbai region.

**Review of the Literature**

Numerous studies have been conducted around the world employing the principles of hydrodynamic analysis, water quality analysis, environmental flow assessment, flow augmentation, and their combinations. For the linked river network of Maharashtra, the idea of hydrodynamic analysis, water quality analysis, creating flow scenarios, and conducting environmental flow assessments is new.

Because of this, this research concentrates on indicators of water quality like dissolved oxygen (DO) and biological oxygen demand. (BOD). The two most important water quality measures needed to gauge how well coastal water can absorb waste are BOD and DO. DO is depleted by the waste inflow, especially organic particle matter undergoing organic breakdown. BOD is used as a general estimation of the effluent's capacity for oxygen demand.

According to the literature study conducted by Singh and Mishra (2021), data analytical models have been extensively used for water quality assurance. The study provides a comprehensive overview of various methods such as statistical analysis, machine learning,

and data mining that have been employed for ensuring water quality. The authors highlight the importance of a hybrid approach, which combines multiple techniques to enhance the accuracy and effectiveness of the models. The advantages of using hybrid models include improved prediction accuracy and reduced false-positive rates. The study also examines several computational techniques used in feature-based and part-based models, along with machine learning and data mining techniques such as support vector machines, decision trees, and neural networks.

The literature review concludes by emphasizing the need for more accurate and comprehensive datasets, the development of advanced hybrid models, and the integration of real-time monitoring systems for effective certification of water quality. Overall, the study provides a valuable resource for researchers and practitioners seeking to ensure the quality of drinking water. (Singh & Mishra, 2021)

## 2.2 Challenges and an Evaluation of the Quality

The quality of the water supply is an important factor that affects both human health and the environment (Huang et al., 2021; Sharma et al., 2020). With the increase in urbanization and population growth, water supply contamination has become a global problem (Kumar et al., 2020). To address this issue, researchers and environmental managers have focused on the assessment and prediction of water quality (Chen et al., 2019; Khan et al., 2018).

Data mining techniques, such as decision trees, neural networks, clustering, and association rule mining, have been used to analyze water quality data (Alizadeh et al., 2021; Chen et al., 2019). These techniques enable researchers to identify patterns and relationships in the data that can help improve their understanding of the factors that affect water quality (Kumar et al., 2020; Wang et al., 2018).

In order to assess information on the state's water supply, researchers in the Indian state of Maharashtra frequently employ specific water quality indicators that were created by the Maharashtra Pollution Control Board (MPCB). Many physical, chemical, and biological characteristics, including pH, dissolved oxygen, biochemical oxygen demand, total dissolved solids, and faecal coliform count, are measured by these tests.

By applying data mining techniques to analyse water quality data, researchers can determine the most important elements that significantly affect the water quality in a range of locations. With the use of this knowledge, it is feasible to develop workable solutions for improving water quality, such as the implementation of waste management reforms and strategies for lowering pollution.

### 2.3 Water Quality Evaluation

The many various data mining methods that can be utilised to assess the quality of water were discussed by Shoba G and Dr. Shobha G. The Bayes method, as described by Li and Wang, was one of the methods that these researchers examined. According to some, this approach is straightforward, practical, and simple to comprehend in terms of calculation. The method was deemed useful by the authors for analysing the conditions of the water quality in numerous distinct areas, and they discovered that three of the locations had higher concentrations of ammonia nitrogen.

The authors also explored the use of the Fuzzy Comprehensive Evaluation Method for analyzing water quality data in real-time (Ji-hong et al., 2020). This method, which is based on fuzzy logic, allows for the simultaneous evaluation of multiple parameters and was found to be effective in assessing water quality (Ji-hong et al., 2020; Biao et al., 2019).

In order to enhance the performance of a combined water quality assessment model, Yang Zengchuen created an artificial neural network (ANN) using the shuffling frog leaping technique (SFLA). The ANN is a form of machine learning approach that may be used for pattern recognition, whereas the SFLA is an evolutionary algorithm that may be used for optimization. These two algorithms were created through natural selection. It shows that the authors considered this method to be a useful way to assess the calibre of the water they were testing.

Back propagation is one of the ways that is used most frequently, according to the review that Shoba G and Dr. Shobha G conducted. An ANN type known as back propagation is frequently used to carry out supervised learning tasks like classification and prediction. Even if more details are not given, it is likely that the authors found this method to be quite helpful for figuring out the water quality.

Examining how the indices of the water quality in Egypt's Lake Burullus are responding to climate change Shalby and associates In this study, the Egyptian lake Burullus's water quality is examined in relation to climate change. The researchers looked into a range of elements, such as temperature, pH, dissolved oxygen levels, total alkalinity, and nutrient concentrations, among others, that affect the lake's water quality. The results show that higher temperatures and nutrient concentrations have an impact on the lake's ecosystem, which in turn has an impact on Lake Burullus's water quality. To recognise and address the implications of this condition, the authors contend that regular monitoring of the lake's water quality is necessary.

In 2009 saw the publication of "Water Quality Study of Yamuna River Using WQI in the National Capital Territory of India (2000-2009)," written by D. Sharma and A. Kansal. The Yamuna River in India's National Capital Region's water quality was analysed by the authors of this study using the Water Quality Index. The procedure of assessing the water quality makes use of the WQI approach. The total coliform count, pH, dissolved oxygen, and biochemical oxygen demand are only a few of the variables that are taken into account by this approach. The Yamuna River in the NCT, according to the study's authors, has poor water quality and did not meet requirements for aquatic life or drinking water. The study's conclusions indicate that in order to improve water quality, effective management and control of water resources are necessary.

"Lower Danube water quality measured through WQI and multivariate analysis," C. Iticescu et al. The focus of this study is on the Lower Danube, which passes through Romania. The authors used the WQI to assess the water quality of the river, using multivariate analysis to identify the main sources of contamination. The results showed that human activities, such as industrial and agricultural operations, had a significant impact on the water quality of the Lower Danube. The study emphasizes how important it is to identify pollution sources when developing plans for managing water resources.

### 2.4. Water Quality Forecasting

An article titled "Improving prediction of water quality indices using novel hybrid machine-learning algorithms" was written by D. T. Bui and his colleagues. The authors of this article introduced some brand-new hybrid machine learning algorithms to forecast

water quality measures. The researchers used a range of machine learning algorithms, including decision trees, random forests, artificial neural networks, and support vector regression, to build hybrid models. The results showed that when it comes to forecasting water quality metrics, the hybrid models outperformed the component models significantly. The study's conclusions suggest that machine learning algorithms could be quite good at anticipating indicators of water quality.

In the study by Z. M. Yaseen et al., the water quality index was determined using hybrid adaptive neuro-fuzzy models.

To obtain an accurate reading of the WQI, the authors of this study advised employing hybrid adaptive neuro-fuzzy models. The authors used a number of water quality parameters as input variables and the WQI as the target variable. pH, dissolved oxygen, and total dissolved solids are some of these parameters. The findings showed that hybrid models outperformed traditional fuzzy models when it comes to computing the WQI. In the study, the importance of creating water-specific hybrid machine learning models is addressed.

Machine learning should be used to evaluate the possibility of nitrate contamination in groundwater, according to Awais et al. The indicators of groundwater quality in Pakistan's Pothohar region were analysed and assessed using data mining technologies. The findings showed that the suggested strategy may be useful in identifying regions with a high risk of nitrate groundwater pollution.

In their comprehensive analysis of communicable diseases, Tulchinsky and Varavikova looked at the factors that contribute to their occurrence as well as treatments and preventative strategies. Many communicable illnesses have had their epidemiology, etiology, therapies, and impacts on public health investigated.

Research on the health risks associated with the consumption of wastewater for crop irrigation was done by Khalid et al. The research concentrated on high- and low-income nations. The study investigated the potential health concerns linked to the use of wastewater. These dangers included the potential for bacterial, heavy metal, and other toxic

interactions. The authors offer a range of suggestions for reducing the harmful health impacts associated with consuming wastewater.

The idea was summarized, its ramifications were examined, and the impact it would have on the environment was rated by Chu and Karr. The study examined various approaches for assessing environmental impact as well as the challenges involved in doing so. The authors offered many potential solutions for reducing the harm that human activity causes to the ecosystem. Their actions included pollution prevention and resource conservation.

In order to ensure global food security, Kopittke et al. stressed the importance of soil in connection to the intensification of agricultural practices. Throughout the course of the study project, it was emphasized both the challenges provided by soil contamination and deterioration as well as the significance of soil in providing crops with essential nutrients and water. The application of organic fertilisers and the adoption of sustainable agricultural practices were the authors' two suggested ways for improving soil quality.

Hameed et al. used artificial intelligence-based algorithms to forecast the water quality index in a tropical area of Malaysia. After using data mining techniques to study and assess the numerous factors that affect water quality, the research's findings demonstrated that the suggested method has the capacity to predict the water quality index with accuracy.

Bournaris et al. suggested a decision-support system to include ecologically sound irrigation water usage practices into the planning of agricultural production. The results indicate that the suggested technique has a good likelihood of facilitating agricultural production planning that emphasizes ecologically responsible irrigation water use. The study researched, analyzed, and evaluated a number of elements of irrigation water quality using various data mining techniques.

Rufino and associates constructed an integrated system in the Agro-Aversano area in southern Italy to evaluate the potential of using urban groundwater resources for agriculture and drinking water requirements. To satisfy the needs of the area, this was done. The study's conclusions showed how well the method for determining whether urban groundwater resources are suitable for drinking and irrigation can be applied. According

to the paper, indicators of groundwater quality were analysed and assessed with the aid of data mining methods.

Vadiati et al. suggested using fuzzy logic to inform decision-making as a way to assess the calibre of groundwater. In their research, Vadiati and associates created a fuzzy logic-based decision-making system for assessing the quality of groundwater (2016). The suggested technique divides groundwater quality into four groups based on the six input factors it takes into account. These six parameters include pH, electrical conductivity, total dissolved solids, sodium adsorption ratio, magnesium adsorption ratio, and chloride concentration. The degree to which the input parameters correspond to each of the four categories is determined by the Fuzzy Inference System (FIS) using the fuzzy membership functions. This choice is determined in light of the provided parameters. When the technique was used on a dataset containing data on the groundwater quality in an Iranian agricultural region, it was shown to be effective, and the results supported the hypothesis. The authors came to the conclusion that the fuzzy-logic-based decision-making approach can be a useful tool for determining the quality of groundwater, particularly in areas where there is a shortage of pertinent data and subject-matter expertise. Throughout the research, this was one of the writers' main grounds of dispute. However, they emphasized that by using more sophisticated fuzzy inference techniques and more input parameters, the strategy may be improved even more. The study provides a significant contribution to the field of groundwater quality assessment and emphasises the utility of fuzzy logic in environmental management decision-making.

In their essay titled "River Water Quality Modeling Using Artificial Intelligence (AI) Methods from 2000 to 2020," T. M. Tung and Z. M. Yaseen give a thorough description of the research that was done on the subject during those two years. The study looks into a wide range of topics related to river water quality, including the forecasting of water quality and the localization of pollution sources.

Data mining techniques are used in a case study by K. Khosravi, L. Mao, O. Kisi, Z. M. Yaseen, and S. Shahid to anticipate hourly suspended sediment load in a glacierized Andes watershed in Chile. A number of data mining techniques, including random forest, support vector machines, and artificial neural networks, are tested throughout the project (RF).

Study employing hybrid data-intelligence models and meteorological data mining on reference evapotranspiration (ET0) simulation in Iraq by K. Khosravi, P. Daggupati, M. T. Alami, S. M. Awadh, M. I. Ghareb, and M. Panahi. Writers include K. Khosravi, M. T. Alami, S. M. Awadh, and P. Daggupati. In the research, a hybrid model for ET0 simulation is suggested. To operate, this model combines data mining and AI methods.

A stochastic modeling approach is described by authors K. Khosravi, R. Barzegar, S. Miraki, J. Adamowski, P. Daggupati, M. R. Alizadeh, and others for predicting groundwater fluoride contamination. The performance of several types of lazy learners, including K-nearest neighbour (KNN) and radial basis function (RBF) networks, is examined in this study, and their outcomes are contrasted with those of more conventional models.

It appears that Dr. Akepati S. Reddy, Siddhartha Sharma, Neetu Arora, and Amarpreet Singh Arora collaborated on a study to raise the standard of the water in the River Satluj. The research endeavor included long-term river monitoring, and significant discriminant features that led data on water quality to cluster were also identified. Data clustering is a type of unsupervised learning strategy that involves putting similar data items in groups based on their shared properties.(Neetu Arrora,Amarpreet Arrora,et.al 2015) Data clustering can be used to identify patterns in the data pertaining to the quality of the water that might be connected to certain environmental situations or sources of pollution when it comes to the examination of water quality.

It appears that the researchers that conducted this study used a method called data clustering to find distinctive collections of the Satluj River's water quality data. After that, we examined the discriminant variables to determine which ones were most responsible for the data's grouping. They were able to do this successfully in order to identify the key elements affecting the river's water quality. With this knowledge, plans might be made to put different water treatment techniques into place or to reduce the quantity of pollution coming from particular sources, thereby improving the water quality of the River Satluj. The writers were able to better grasp the numerous variables that affect the river's water quality and come up with more practical proposals for how it may be improved by focusing on the data before anything else.

By using a technique known as a Decision Tree, Shailesh Jaloree, Anil Rajput, and Sanjeev Gour created a model for figuring out the water quality of the Narmada River. The software programme Weka was used to bring the idea into practise.

The decision tree is a supervised learning technique that can be used in a variety of applications for classification or regression analysis. Decision trees can be used to categorise water samples according to the pollutant class in the context of a water quality assessment. This can help in identifying the key variables that have the biggest impact on the status of the water.

In order to forecast the pollutant class of the water in this study, important variables like NH3 N, NO3 N, pH, BOD, and COD were used in conjunction with decision tree classification. The authors were successful in identifying the pollutant class of the water samples that were taken from the Narmada River by using these variables as inputs into the Decision Tree model. These traits are regarded to be important elements with the capacity to affect the quality of the water.

### 2.5 The Hydrology of Artificial Intelligence and it's Applications

Because of their capacity to predict hydrological variables, model complicated water systems, and evaluate water resource management, technologies based on artificial intelligence (AI) have recently become widely used in the field of hydrology. This assessment of the relevant literature presents a discussion of recent research publications that investigate the application of AI modelling techniques in the field of hydrology.

The authors, T.M. Tung and Z.M. Yaseen, conduct an in-depth analysis of how artificial intelligence models can be utilised in the modelling of river water quality. The authors address the creation of decision trees, fuzzy logic models, artificial neural networks (ANNs), and support vector machines (SVMs), with the purpose of detecting the quality of the water, and they review the most recent discoveries made from research in this field.

In December 2018, a study authored by K. Khosravi and colleagues and titled "Quantifying hourly suspended sediment load using data mining models: Case study of a glacierized Andes watershed in Chile" was published in the Journal of Hydrology. The purpose of this project is to conduct a case study on the application of data mining strategies for the

purpose of estimating the hourly loads of suspended sediment that are present in a watershed in Chile that has been affected by glaciation.

The performance of several artificial intelligence models was evaluated and compared to the performance of conventional regression models. ANNs, random forests, support vector machines, and Gaussian process regression were some of the models that were used. They discovered that AI models performed far better than traditional models and produced estimates that were more precise.

A case study in Iraq for reference evaporation modelling using hybrid data-intelligence models and meteorological data mining, by K. Khosravi et al. This research article explains how a hybrid data-intelligence strategy was utilised to analyse the results of a reference evaporation simulation case study that was conducted in Iraq. The case study was concluded with Iraq as the location. Data mining strategies such as artificial neural networks (ANNs), decision trees, and support vector machines were utilised by the authors of the study in order to develop hybrid models. These models mix AI models with physical models (SVMs). They discovered that hybrid models outperform traditional approaches in terms of delivering predictions that are more accurate than those produced by conventional methods.

K. Khosravi and his colleagues are responsible for the writing of the study titled "Stochastic Modeling of Groundwater Fluoride Contamination: Incorporating Lazy Learners." Lazy learners are a category of machine learning algorithms, and this article presents an original method for stochastic modelling of groundwater fluoride contamination utilising those techniques. When it comes to performance, lazy learning algorithms perform far better than typical machine learning algorithms. It's possible for these algorithms to work with data that's quite precise and high-dimensional, and they could produce predictions that are more accurate.

You might be able to locate a study here by K. Khosravi and his colleagues titled "A Comparative Assessment of Flood Susceptibility Modeling Utilizing Multi-Criteria Decision-Making Analysis and Machine Learning Techniques." In this particular piece of research, an investigation of the usefulness of modelling flood susceptibility using machine learning algorithms and multi-criteria decision-making (MCDM) assessments is carried

out. A number of different approaches to machine learning, such as ANN, SVM, decision trees, and random forests, were evaluated alongside the effectiveness of the MCDM analysis to determine which was the most effective. The findings of A. Sharafati and colleagues, which were published in an article with the title "The potential of current data mining algorithms for global solar radiation projection," revealed that predictions made using machine learning approaches produced more accurate findings than those made using MCDM analysis. Using the help of data mining techniques such as artificial neural networks (ANNs), support vector machines (SVMs), and decision trees, the purpose of this investigation is to establish whether or not it is possible to estimate the total amount of solar energy that the world would get. As compared to the results obtained by traditional models, the authors found that the predictions obtained through the use of data mining techniques yielded a significantly higher level of precision.

### 2.6 Essential data mining methods consist of categorization, regression, and AI/ML.

A research project with the working title "A Comparative Investigation of Data Mining Methods for Water Quality Analysis" is carried out by T. Chakraborty, D. Chakraborty, and P. Mukherjee. The writers of this research paper conduct an investigation into and examination of a variety of data mining strategies that can be applied to the process of analyzing water quality. The authors use a dataset that contains information about water quality in order to evaluate the effectiveness of decision trees, artificial neural networks (ANN), and support vector machines (SVM). According to the findings of the study, each of the three methodologies has the ability to accurately determine the quality of the water; however, the ANN and SVM methodologies perform noticeably better than decision trees do in terms of both accuracy and predictive capacity. The authors also examine the computational difficulty of the various approaches and find that decision trees, artificial neural networks, and support vector machines (SVMs) have the lowest levels of computational complexity among all of the different methods. This study highlights the usefulness of data mining tools for analyzing water quality and exhibits the benefits and negatives of various strategies. [Citation needed] [Citation needed] The importance of data mining technologies is highlighted throughout the entirety of this research as a result of the findings. The authors provide insights into which methods perform best for specific kinds

of water quality data and applications by comparing and analyzing various procedures. These procedures are used to determine water quality. These realizations can be used to the process of decision-making to improve it.

The piece of writing titled "Machine Learning Techniques for Water Quality Analysis: A Review," which was written by S. Sharma, M. Arora, and A. Bhardwaj, offers a comprehensive investigation into a variety of machine learning strategies that have been utilized in the context of water quality analysis. By utilizing datasets concerned with water quality, the authors assess the efficacy of a variety of approaches, including decision trees, artificial neural networks (ANN), fuzzy logic systems, and others.

The findings of the study indicate that all of the strategies have been successfully applied to the task of analysing the water quality, and the consequences have been encouraging. According to the findings of the study, the water quality analysis task was successful. Although ANN and fuzzy logic systems are capable of capturing more complicated and non-linear correlations between the water quality measurements, the benefit of utilizing decision trees is that they are simple to perceive and comprehend. This is despite the fact that these systems are capable of capturing these correlations.

The authors focus a significant amount of stress on the significance of feature selection as well as data pretreatment when it comes to the creation of reliable forecasts. They explore the benefits and drawbacks of a number of preprocessing methods, such as feature scaling and dimensionality reduction, and they emphasise the importance of carefully cleaning and filtering the data.

The main significance of this work is that it gives a helpful overview of several machine learning algorithms for analysing water quality and highlights the benefits and downsides of each technique. This is an important contribution to the field of research. The authors offer some pointers on how to select the methodology that is going to be the most useful for analysing a particular water quality dataset and application. This can be accomplished by taking into account the myriad of elements that have the potential to affect the effectiveness of these methods. Some examples of such aspects include the standard of the data and the level of preprocessing that is applied to the data.

Neetu Arora and her colleagues concentrated their emphasis on a particular pollutant cluster using clustering methodologies in order to validate the water's quality. This allowed them to determine whether or not the water was safe to drink. In a study that was somewhat comparable to this one, Shailesh Jaloree and colleagues utilised the Weka tool to apply decision tree classification. However, in order to characterise the quality of the water, they restricted themselves to using only five quality parameters.

However, it is imperative to keep in mind that the applicability of certain water quality tests may vary greatly depending on the precise location of the measurement as well as the conditions that were present at the time of the measurement. Because of this, it is absolutely necessary for researchers, prior to making their picks, to carry out in-depth evaluations of the factors that are most pertinent to their research.

In a general sense, the application of data mining methods can be a useful instrument for analysing water quality. Researchers are able to gather accurate and insightful information on the quality of the water and make decisions that are defendable for the management and conservation of water resources if they narrow their emphasis to a restricted number of criteria and focus their attention on those elements.

**Table 2.1 Comparison of some common data mining methods used in water quality analysis**

| **Data Mining Method** | **Description** | **Advantages** | **Disadvantages** |
|---|---|---|---|
| Decision Trees | Builds a tree-like model to classify data based on a set of rules | Easy to interpret and visualize, can handle both numerical and categorical data | Prone to overfitting, can be sensitive to small changes in data, may not work well for complex datasets |
| Artificial Neural Networks | Models the human brain to identify patterns in data and make predictions | Can handle non-linear relationships, can work with noisy data | Black box model, difficult to interpret, requires a large amount of data |

| | | | |
|---|---|---|---|
| Support Vector Machines | Separates data into classes based on a boundary called a hyperplane | Can handle high-dimensional data, good at handling noise and outliers | Computationally expensive, may require a lot of parameter tuning |
| Random Forests | Ensemble learning method that constructs multiple decision trees to improve accuracy | Robust to noise and overfitting, handles missing data well | Can be computationally expensive, difficult to interpret |
| Clustering | Identifies groups of similar data points within a dataset | Useful for identifying patterns and outliers | May not work well for complex data, requires careful selection of distance measures and clustering algorithms |

It is worth noting that the advantages and disadvantages of each method may vary depending on the specific dataset and context. Additionally, other data mining methods may be more appropriate for certain types of data or research questions.

**2.7 Water Quality Index:**

The water quality Index is a traditional method which was originally developed by the National Sanitation Foundation (NSF) that provides a single grade which expresses the overall water quality at a certain location on various patterns. This technique gives the exact value of water quality index but on restricted parameters.

National Sanitation Foundation Water Quality Index:

$$\text{WQI} \sum_{i=1}^{n} (wi\ X\ li) \quad \text{.....(NSF-WQI Method)}$$

Where

Wi = Weight parameter untill i, scale 0-1.0

Ii = Value Sub Index for ith

n = The number of water quality parameters

It uses a scale from 0 to 100 to rate the quality of water,100 will be the highest score which shows the excellence of water.

**Table 2.2 Quality of Water(scales) (source: NSF-WQI)**

| Range | Quality of Water |
|---|---|
| 91.1-99.9 | Excellent |
| 70.9-90 | Good |
| 50.9-70 | Medium |
| 26-50 | Fair |
| 0-25 | Poor |

## 2.8 Technology-wise Comparative Analysis of research papers

**Table 2.3 Comparative Analysis of Research Papers**

| Sr. No | Title | Author | Year | Objective | Technique |
|---|---|---|---|---|---|
| 1 | Total coliforms and data mining as a tool in water quality monitoring | Ivana D. Radojević, Dušan M. Stefanović, Ljiljana R. Č, Aleksandar M. Ostojić, Marina D. Topuzović, Nenad | 2012 | Analysis of water quality based on physiochemical and biological parameters and data mining techniques | Total coliforms, reservoir, water quality, data mining, cluster analysis, classification |
| 2 | Assessment of water quality parameters : A review | S.P.Gorde , M.V.Jadhav | 2013 | To access aquatic macro invertebrates | Water Quality Index |
| 3 | Ground water quality assessment using Data Mining techniques | Kamakshaiah.Kolli, R. Seshadri | 2013 | To emphasize on water contamination, variation of groundwater quality and its | Water quality Index |

| | | | | | |
|---|---|---|---|---|---|
| | | | | spatial distribution | |
| **4** | Detection and Mapping of Water Quality Variation in the Godavari River Using Water Quality Index | Indrani Gupta , Awkash Kumar , Chandrakant Singh , Rakesh Kumar | 2015 | To develop a tool for planning and managing the water quality of River Godavari | Water Quality Index, ArcGIS |
| **5** | Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China | Xiaoping Wang, Fei Zhang | 2017 | To establish a model for estimating and assessing the WQI | RMSE, Water Quality Index |
| **6** | Big Data, Knowledge Mapping for Sustainable Development: A Water Quality Index Case Study | Lei Zhang | 2019 | How to keep track of water quality through Big Data | Water Quality Index |
| **7** | A review of water quality index models and their use for assessing surface water quality | Md.Galal Uddin stephenNash Agnieziska | 2021 | Reviewed multiple WQI Models | Water Quality Index |

The reviewed publications in Table 2.4 deal with the use of data mining methods and the Water Quality Index (WQI) in monitoring and evaluating water quality. The papers seek to identify the variables impacting water quality and examine the physiochemical and

biological characteristics of water quality. Data mining methods are used to categorise water quality data and create models for water quality evaluation, including cluster analysis, decision trees, and support vector machines. The papers' main goal is the creation of a tool for the management and planning of water quality, particularly in rivers, reservoirs, and groundwater. Data mining, classification, hybrid classification, and the usage of ArcGIS and the Weka tool for analysis are some of the techniques used.

**Table 2.4 Classification Hybrid Classification**

| Sr No | Title | Author | Year | Objective | Technique |
|---|---|---|---|---|---|
| **1** | Total coliforms and data mining as a tool in water quality monitoring | Ivana D. Radojević, Dušan M. Stefanović, Ljiljana R. Č, Aleksandar M. Ostojić, Marina D. Topuzović, Nenad | 2012 | Analysis of water quality based on physiochemical and biological parameters and data mining techniques | Total coliforms, reservoir, water quality, data mining, cluster analysis, classification |
| **2** | Decision tree approach to build a model for water quality | Shailesh Jaloree, Anil Rajput, Sanjeev Gour | 2014 | To access 5 attributes of water quality data which affects quality | Data Mining, classification model, Decision tree, Weka Tool |
| **3** | Construction of a Novel Water Quality Classifier: A case study in the Sebou region | Mariam Elkhechafi,Hanaa Hachimi,Aouatif Amine And Youssfi El Kettani | 2017 | To apply hybrid classification techniques | Principal Components Analysis (PCA), Support vector Machine |
| **4** | Ground Water Quality Data Analysis Using Classification Techniques | Devi, S.V.S. Ganga, | 2018 | Quality Data analysis | Classification |
| **5** | A Classification Model For Water Quality Analysis | Consolata Gakii and Jeniffer Jepkoech | 2019 | Analyse water quality usig Decision tree | Classification and Decision Tress |

| | | | | | |
|---|---|---|---|---|---|
| | Using Decision Tree | | | and classification | |
| **6** | Surface water quality classification using data mining approaches: Irrigation along the Aladag River | Sattari,Mohmmd,Feizi | 2021 | Surface Water Quality Classification | Data Mining and Classification |

**Table 2.5 Clustering Technique**

| Sr. No | Title | Author | Year | Objective | Technique |
|---|---|---|---|---|---|
| **1** | Applying data mining and HPC for water quality assessment and prediction | Ruijian Zhang , Hairong Zhao, Yong Piao | 2011 | Water quality prediction by using simulation model | Clustering |
| **2** | Water quality event detection and customer complaint clustering analysis in distribution systems | Stephen Mounce, John Machell, Joby Boxall | 2012 | To analyze different software component | water quality; online monitoring; event detection; data analysis, Clustering |
| **3** | Assessment of Water Quality of Euphrates River Using Cluster Analysis | Emad A. Mohammad Salah, Ahmed M. Turki, Eethar M. Al-Othman | 2012 | To access temporal and spatial variations in water quality | Cluster Analysis; Surface Water; Euphrates River |
| **4** | Application of Factor and cluster analysis in the | Twaha B,Kassim S, | 2013 | To access the water quality using cluster | Factor analysis, underground |

| | | | | | |
|---|---|---|---|---|---|
| | assessment of source contaminants in Borehole water in Tanzania | C.Mary,Paul S | | and factor analysis | water, Cluster analysis |
| **5** | Water Quality Prediction Using Data Mining Techniques | Shoba G, Dr. Shobha G. | 2014 | Predicting water quality | Decision Tree, Cluster Analysis, Factor Analysis, GIS |
| **6** | Estimation of Water Quality Parameters Using the Regression Model with Fuzzy K-Means Clustering | Muntadher A. SHAREEF*, Abdelmalek TOUMI*, Ali KHENCHAF * | 2014 | to find contaminants and determine the Water Quality Parameters (WQPs) based on theories of the texture analysis. | Gray Level Co-occurrence Matrix (GLCM), K means Clustering |
| **7** | An Enhanced K-Means Algorithm for Water Quality Analysis of The Haihe River in China | HuiZou,ZhihongZou,and Xiaojing Wang | 2015 | To presents a modified fast clustering algorithm for water quality analysis | K-Means Cluster Algorithm |
| **8** | Use of Cluster Analysis-A Data mining tool for improved water quality monitoring of River Satluj | Neetu Arora, Amarpreet Singh Arora, Siddhartha Sharma, Dr. Akepati S. Reddy | 2015 | To obtain natural clusters | Cluster analysis, Data mining, river water quality |
| **9** | Data Clustering Approach to Industrial Process Monitoring, Fault Detection and Isolation | Kiran Jyoti, Dr. Satyaveer Singh | 2015 | To simulate the use of K means clustering | K- means Clustering, Fuzzy Based Fault Detection and Isolation |
| **10** | Analysis of lead in river water quality data using clustering | S. Brintha Rajakumari, C. Nalini | 2016 | Lead analysis in water | Clustering Technique |

| | | | | | |
|---|---|---|---|---|---|
| **11** | Groundwater Quality Assessment: An Improved Approach to K-Means Clustering, Principal Component Analysis and Spatial Analysis: A Case Study | Ana Elizabeth,Diego Cruz,Elena,et al | 2018 | Ground water quality assessment | K-Means Clustering |
| **12** | Evaluation of river water quality by using hierarchical clustering analysis | B Warsito et al | 2021 | Water Analysis by implementing Clustering | Hierarchical Clustering |

**Table 2.6 Regression Analysis**

| Sr. No | Title | Author | Year | Objective | Technique |
|---|---|---|---|---|---|
| **1** | A hybrid approach of SVR with genetic algorithm optimization for aquaculture WQ prediction | ShuangyinLiuabHaijiangTaiaQishengDingaDaoliangLi | 2013 | To build effective hybrid model | Real Value genetic algorithm ,Support Vector Regression |
| **2** | Estimation of Water Quality Parameters Using the Regression Model with Fuzzy K-Means Clustering | Muntadher A. Shareef*, Abdelmalek Toumi*, Ali Khenchaf* | 2014 | to find contaminants and determine the Water Quality Parameters (WQPs) based on theories of | Gray Level Co-occurrence Matrix (GLCM), Regression |

| | | | | | |
|---|---|---|---|---|---|
| | | | | the texture analysis. | |
| **3** | A Hybrid Prediction Model for Monitoring of River Water Quality in the USN System | Hoontae Kim, Minsoo Kim | 2015 | To predicts salinity level of groundwater | Statistical Regression Model |
| **4** | Estimation of Water Quality Parameters With Data-Driven Model | Sattari, Mohammad Taghi; Joudi, Ali Rezazadeh; Kusiak, Andrew, | 2016 | Using Electric Conductivity and TDS obtaining the water quality. | Support vector regression, K nearest neighbor |
| **5** | Data for the evaluation of groundwater quality using WQI and regression analysis in parts of Nalgonda district, Telangana, South India | K.saikrishna D.Purushotham V.Sunitha | 2020 | evaluate the groundwater quality through drinking WQI and regression analysis | regression analysis |
| **6** | Water Quality Modeling and Regression Analysis of Vembanad Lake and its Primary Inflow | Athira K R , Varsha Murali , Aswin C R , Devaprabha B, Aneesh P C | 2021 | Water Quality Analysis through Regression | Regression |
| **7** | Water quality index estimation using linear regression model | Venkaraman, Nikitha, Rishika,et al | 2022 | WQI through Linear regression | Linear regression |

**Table 2.7 Support Vector machine and Support Vector Regression**

| Sr. No | Title | Author | Year | Objective | Technique |
|---|---|---|---|---|---|
| **1** | A hybrid approach of SVR with genetic algorithm optimization for aquaculture water quality prediction | Shuangyin Liuab Haijiang Taia Qisheng Dinga Daoliang Li | 2013 | To build effective hybrid model | Real Value genetic algorithm ,Support Vector Regression |
| **2** | Estimation of Water Quality Parameters With Data-Driven Model | Sattari, Mohammad Taghi; Joudi, Ali Rezazadeh; Kusiak, Andrew, | 2016 | Using Electric Conductivity and TDS obtaining the water quality. | Support vector regression, K nearest neighbor |
| **3** | Fault Diagnosis of Water Quality Monitoring Devices Based on Multiclass Support Vector Machines and Rule-Based Decision Trees | Shuangyin Liu , Longqin Xu, Qiucheng Li, Xuehua Zhao, Daoliang Li | 2017 | Fault Diagnosis | Multiclass support vector Machine, Rule Based Decision Tree |
| **4** | Construction of a Novel Water Quality Classifier: A case study in the Sebou region | Mariam Elkhechafi,Hanaa Hachimi,Aouatif Amine And Youssfi El Kettani | 2017 | To apply hybrid classification techniques | Principal Components Analysis (PCA), Support vector Machine |
| **5** | A Study on Ground Water Quality Prediction using R Tool | S. Vijay | 2018 | to assess the overall groundwater quality | Association rules, Naïve Bayes, ANN and SVM |
| **6** | Fault diagnosis of water quality monitoring devices based on | shuangyinliu, longqin xu1, | 2019 | To diagnose a fault of water | rule-based decision trees (RBDT), |

| | | | | | |
|---|---|---|---|---|---|
| | multicast svm and RBDT | qiucheng li, xuehua zhao et.al | | quality monitoring | Support vector regression |
| **7** | Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM) | Wei Cong Leong, Aleriza,Jie,Ahmed | 2021 | Prediction of WQI | using support vector machine (SVM) and least square-support vector machine (LS-SVM) |

**Table 2.8 K Nearest Neighbor (KNN) Technique for Water Quality Prediction**

| Sr. No. | Title | Author | Year | Objective | Technique |
|---|---|---|---|---|---|
| **1** | Modeling of Water Quality Parameters Using Data-Driven Models | H. Orouji, O. Bozorg Haddad, E. Fallah-Mehdipour, and M. A. Mariño | 2013 | To predict and simulate water quality | Adaptive neural fuzzy inference system (ANFIS) and genetic programming (GP) |
| **2** | Water quality Management using statistical analysis and time series prediction model | Kulwinder Singh Parmar, Rashmi Bhardwaj | 2014 | Prediction based on monthly variations | -Prediction Model |

| | | | | | |
|---|---|---|---|---|---|
| **3** | Predicting Urban Water Quality with Ubiquitous Data | Ye Liu, Yuxuan Liang, Shuming Liu, David S. Rosenblum | 2016 | Using data-driven approach forecast water quality of multiple station | Multi-View and Multi-Task Learning |
| **4** | Estimation of Water Quality Parameters With Data-Driven Model | Sattari, Mohammad Taghi; Joudi, Ali Rezazadeh; Kusiak, Andrew | 2016 | Using Electric Conductivity and TDS obtaining the water quality | Support vector regression, K nearest neighbor |
| **5** | Review on Data Mining Techniques for Prediction of Water Quality | Priya Sing, Pankaj Kaur | 2017 | By applying various mining technique check water quality | Native Bayes, Artificial Neural Network, K-Nearest Neighbor, Back Propagation Algorithm |
| **6** | The Classification Status of River Water Quality in Riau Province Using Modified K-Nearest Neighbor Algorithm with STORET Modeling and Water Pollution Index | Mustakim, Rosdina, Dian Ramad | 2020 | Classify the status of water quality | K-Nearest Neighbor Algorithm |

| | | | | | |
|---|---|---|---|---|---|
| **7** | Water Quality Prediction Using KNN Imputer and Multilayer Perceptron | Afaq Juna , Muhammad Umer , Saima Sadiq,et al | 2022 | Water quality prediction | KNN model |

**Note:** The values of RMSE are approximate and may vary depending on the specific implementation and dataset used in each study.

**Table 2.9 Reviewed Papers on K-Nearest Neighbor and RMSE Technique for Water Quality Prediction**

| Sr. No | Title | Author | Year | Objective | Technique |
|---|---|---|---|---|---|
| **1** | Modeling of Water Quality Parameters Using Data-Driven Models | H. Orouji et al. | 2013 | To predict and simulate water quality | ANFIS and GP |
| **2** | Water Quality Management using statistical analysis and time series prediction model | Kulwinder Singh Parmar and Rashmi Bhardwaj | 2014 | Prediction based on monthly variations | KNN |
| **3** | Predicting Urban Water Quality with Ubiquitous Data | Ye Liu et al. | 2016 | Using data driven approach forecast water quality of multiple station | Multi-View and Multi-Task Learning |

| | | | | | |
|---|---|---|---|---|---|
| **4** | Water Quality Analysis and Prediction Using Hybrid Time Series and Neural Network Models | L. Zhang et al. | 2016 | To predict water quality by Ecological and Environmental Monitoring | Hybrid ARIMA |
| **5** | Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China | Xiaoping Wang and Fei Zhang | 2017 | To establish a model for estimating and assessing the WQI | KNN |

Note: ANFIS stands for Adaptive Neural Fuzzy Inference System, GP stands for Genetic Programming, and ARIMA stands for Auto Regressive Integrated Moving Average.

**The Root Mean Square Error (RMSE) technique**

The Root Mean Square Error (RMSE) technique is a commonly used statistical method for evaluating the accuracy of predictive models. It is particularly useful in the field of data analysis and is often applied in various scientific fields including engineering, finance, and meteorology.

The following table summarizes some of the key findings from a literature review on the RMSE technique:

**Table 2.10 Root Mean Square Error (RMSE) technique**

| Study | Key Findings |
|---|---|
| Liu et al. (2018) | RMSE is an effective method for evaluating the performance of machine learning models. The technique is widely used in the field of finance to predict stock prices and analyze market trends. |
| Karabulut and Baykara (2019) | RMSE can be used to assess the accuracy of rainfall prediction models. The study found that the technique can help improve the accuracy of short-term rainfall predictions. |
| Alfares and El-Sayed (2019) | RMSE is a useful tool for evaluating the accuracy of wind speed forecasting models. The technique can help improve the efficiency of wind energy systems by providing more accurate predictions of wind speed. |
| Zhang et al. (2020) | RMSE is a common technique for evaluating the accuracy of machine learning models used in predicting medical outcomes. The study found that the technique can help improve the accuracy of models used in predicting patient outcomes. |

Overall, the literature review suggests that the RMSE technique is a versatile and effective method for evaluating the accuracy of predictive models across a wide range of scientific fields. By providing a standardized measure of error, the technique can help researchers and practitioners make more informed decisions based on the performance of their models.

## 2.9 Gap Analysis from the Literature Review

The literature survey on "Data Analytical Model on Water Quality Assurance Based on the Hybrid Classification and Data Mining Techniques" identifies several gaps and limitations in the existing research on water quality assurance using data analytical models. These gaps and limitations include:

1. **Lack of comprehensive and accurate datasets:** The survey identifies a lack of comprehensive and accurate datasets as one of the major limitations in the existing research on water quality assurance. The survey suggests the need for more comprehensive and accurate datasets to improve the accuracy and effectiveness of the models.
2. **Limited research on advanced hybrid models:** The review notes that there is limited research on advanced hybrid models that can effectively combine different techniques for water quality assurance. The literature review suggests that future research should focus on the development of advanced hybrid models that can provide more accurate and reliable results.
3. **Inadequate integration of real-time monitoring systems:** The survey highlights the need for the integration of real-time monitoring systems into water quality assurance models to improve the effectiveness of the models. The survey suggests that future research should focus on developing models that can incorporate real-time data and monitoring systems for improved accuracy and reliability.
4. **Research based on Environmental Factors:** The review points out that there has been little research on the effects of external elements, such as weather and climate change, on models for ensuring water quality. The study recommends that future research concentrate on comprehending how these outside influences affect water quality assurance models and creating models that can take these elements into consideration.

# Chapter 3

# Research Methodology

# Chapter 3: Research Methodology

## 3.1 Introduction

This chapter focuses on a specific research methodology and describes a novel method for use in research. A research proposal follows an overall approach in regard to the overall purpose, and it may also involve experiments, surveys, models, and case studies. Several data-gathering methods can be applied given the research approach, and quantitative and qualitative methods can be employed to process the complete investigation.

As a result, the overall study is a Pragmatic approach to research (Mixed Methods) that was chosen to meet the research objectives. As a result, pragmatic research allows for the use of any of the methods, techniques, and procedures traditionally connected with quantitative or qualitative research. It has also been acknowledged that each method has limits and that various ways can be complementary.

The chosen analysis approach was employed to collect research data and provide a solution to the problem formulation. Throughout the thesis, we will adhere to pragmatism-based research philosophy and emphasis on practical facts. When examining the water quality in drinking water and quality assurance, both subjective and objective viewpoints will be taken into account.

### 3.1.1 Objective of the Research Study

Research Objectives provides a detailed description of what the research is attempting to accomplish. This research outcome work as a guideline to the overall research

The primary objective of this research study is to develop a Water quality model using hybrid classification and data mining technique for water quality assurance. Humans and other organisms derive this objective from the need to improve the overall quality of water and ensure that it is safe for consumption. The research study aims to develop a technique that can classify water quality based on various parameters and identify the factors that contribute to poor water quality.

The objective of this study is to provide a more accurate and efficient method for monitoring water quality and identifying potential problems.

1. Identify and study a set of water parameters to make an inventory to check water quality.
2. To study an impact of external resources on water quality
3. To design a model which carry out Data Mining techniques on water parameters to verify the water quality
4. To identify the factors which improves water purity and make the water as energy drink.
5. To improve the efficiency and effectiveness of the Data Mining and Classification techniques for better outcome.

### 3.1.2 Significance of the Research Study

The significance of this research study lies in its potential to improve the quality of water and protect public health. Poor water quality can have serious consequences, including the spread of waterborne diseases and environmental degradation.

This research study aims to provide a more effective method for monitoring water quality and identifying potential issues before they become significant problems. The development of a hybrid classification and data mining technique for water quality assurance can help to identify the factors that contribute to poor water quality, such as pollution and environmental degradation. By identifying these factors, steps can be taken into mitigate their impact and improve the overall quality of water.

In addition, the development of this technique can have significant economic benefits. Water is a valuable resource, and poor water quality can have a significant impact on industries that rely on it, such as agriculture and manufacturing. By improving water quality, the economic impact of poor water quality can be reduced, and the overall value of water as a resource can be enhanced.

Overall, the significance of this research study lies in its potential to improve the quality of water, protect public health, and provide economic benefits.

### 3.1.3 Need of research work

The need for this research work arises due to the growing concerns over the quality of water. The pollution of water bodies is a major environmental issue, and it can have serious implications on human health. Therefore, there is a need to develop a technique that can help in ensuring the quality of water. In addition to the concerns over the pollution of water bodies and its impact on human health, there are several other reasons why this research work is needed.

1. **Limited availability of clean water:** Clean water is a scarce resource, and its availability is becoming increasingly limited due to the growing population, urbanization, and climate change. Therefore, there is a need to ensure that the available water resources are of good quality and safe for consumption.
2. **Inadequacy of traditional water quality monitoring methods:** Traditional water quality monitoring methods, such as manual sampling and laboratory analysis, are time-consuming, expensive, and may not provide real-time data. Therefore, there is a need for a more efficient and cost-effective technique that can provide real-time data on water quality.
3. **Increasing complexity of water quality parameters:** Water quality is affected by several factors, such as temperature, pH, dissolved oxygen, turbidity, and the presence of contaminants. Therefore, there is a need for a technique that can analyze multiple parameters simultaneously and provide accurate results.
4. **Importance of water quality in various industries:** Water quality is essential for various industries, such as agriculture, food and beverage, and pharmaceuticals. Poor water quality can have significant economic consequences, including decreased productivity and increased costs. Therefore, there is a need for a technique that can ensure the quality of water for these industries.

In summary, the need for this research work arises from the limited availability of clean water, the inadequacy of traditional water quality monitoring methods, the increasing complexity of water quality parameters, and the importance of water quality in various industries.

### 3.1.4 Benefits of Research work

The research work can provide several benefits, including the identification of factors affecting water quality, prediction of water quality, and development of a technique for water quality assurance. The research work can also help in the prevention of waterborne diseases and improvement of public health. In addition to the benefits mentioned above, this research work can have several other benefits, such as:

1. **Environmental benefits:** The development of a technique for water quality assurance can help in reducing the pollution of water bodies, which can have positive environmental impacts. This can help in preserving aquatic life and maintaining the ecological balance.
2. **Economic benefits:** The technique developed through this research work can help in reducing the economic impact of poor water quality. This can be achieved by improving the quality of water for industries that rely on it, such as agriculture and manufacturing.
3. **Technological benefits:** The development of a hybrid classification and data mining technique for water quality assurance can contribute to the advancement of technology in this field. This can have a positive impact on other areas of research that rely on similar techniques.
4. **Social benefits:** The development of a technique for water quality assurance can have a positive impact on public health and can contribute to the overall well-being of society. This can help in reducing the spread of waterborne diseases and improving the quality of life for individuals and communities.

Overall, the benefits of this research work are numerous and can have a significant impact on various aspects of society.

## 3.2 Research Methodology

The research methodology used in this research work consists of a primary data-based questionnaire survey. After a rigorous literature review the analysis has been done. The questionnaire survey was conducted and responses were collected from all type of people.

This survey has conducted to know the details about people what they know about their drinking water. Based on the analysis, this research contains experimental based proposed framework NavJal. This proposed model is validated using a water sample collected and it is detailed explain in further chapters.

### 3.2.1 Type of Research

Study approaches are research plans and procedures that range from general assumptions to detailed data collecting, analysis, and interpretation methodologies. There are three research approaches, which are as follows:

1. Quantitative
2. Qualitative, and
3. Pragmatic (Mixed methods)

**Quantitative research:**

The process of gathering and analyzing numerical data is known as quantitative research. It can be used to identify trends and averages, formulate hypotheses, examine causality, and extrapolate findings to larger populations.

There are five main types of quantitative research designs and they are:

- Descriptive Research Design.
- Survey Research.
- Correlational Research Design.
- Quasi-experimental Research Design.
- Experimental Research Design.

These five research categories describe five methods for gathering quantitative data. These five approaches also differ owing to the many procedures that each type does. Each method aids in the collection of meaningful data that can be used to make confident decisions.

**1. Descriptive Research Design**

Descriptive research is used to gain a better understanding of a phenomenon, situation, or population. Unlike experimental research, descriptive research does not include the manipulation of specific variables. This approach is employed in quantitative research when attempting to find features, categories, and trends. Case studies, observations, and surveys are the most frequent techniques of gathering descriptive research. Offline survey tools can be used to perform observations or face-to-face surveys.

**2. Survey Research**

Surveys are the most popular and widely used type of data collection. The act of gathering data from a sample of people by asking them questions is known as survey study. Data analysis is also straightforward because modern survey analytics software is available.

**3. Correlational Research Design**

Correlational research is a non-experimental research strategy for determining a link between two variables with no effect from any other variable.

The direction and/or strength of the link will be reflected in the correlation between the two variables.

- A positive correlation indicates that both variables change in the same way.
- A negative correlation indicates that the variables are changing in opposing directions.

A zero correlation indicates that there is no relationship between the variables under consideration.

**4. Quasi-Experimental Research Design**

Quasi-experimental research, like experimental research, seeks to establish a cause-and-effect link between two variables: independent and dependent variables. A quasi-experiment, on the other hand, involves individuals being randomized to groups based on non-random factors.

This approach is commonly used in quantitative research when actual experiments cannot be carried out for practical or ethical reasons. Because it frequently incorporates real-world interventions rather than an artificial laboratory setting, this design has more external validity than most actual experiments.

**5. Experimental Research Design**

The goal of experimental research, also known as real experimentation, is to use the scientific method to evaluate the effect of one or more independent factors on one or more dependent variables. This is accomplished by altering the independent variable in order to investigate its impact on the dependent variable.

The experimental study design entails carrying out a series of procedures in order to evaluate the hypothesis. Subjects in experimental study are randomized to groups at random rather than based on non-random criteria.

**Qualitative Research:**

Data from qualitative studies is not quantifiable. Answers are descriptive text or statements rather than options from a predetermined response set due to the exploratory nature of the question. In order to obtain insights, qualitative research gathers non-numerical data. It is done to acquire a deeper understanding of a subject, problem, or issue from a particular viewpoint.

This is of 5 types: [https://measuringu.com/qual-methods/]

- Ethnographic: This research is the most well-known and widely used sort of qualitative approach. To comprehend the goals, cultures, problems, motivations, and themes that arise through ethnography, the researcher must immerse himself or herself in the context of the target participants.
- Narrative: To build a cohesive story, the narrative technique knits together a sequence of events, usually from just one or two individuals. Conducting in-depth interviews, reading documents, and looking for themes are all essential

- Phenomenological: When describing an event, action, or phenomenon, the aptly titled phenomenological investigation is a suitable qualitative method. In a phenomenological study, it is required to employ a variety of approaches to explore the meaning people place on whatever is being studied, such as conducting interviews, reading papers, watching movies, or visiting places and events. The results rely on the opinions of the participants to provide insight into their reasons.
- Theoretical Foundations: A phenomenological investigation attempts to depict the core of an action or event, whereas grounded theory attempts to provide an explanation or theory for the happenings. It is preferable to construct a theory based on evidence using primarily interviews and existing materials. These studies frequently use bigger sample sizes—between 20 and 60—to further support a theory.
- Case Analysis: Case studies are particularly useful in explaining an organisation, institution, company, or event to quantitative researchers. A case study requires a thorough comprehension of many sorts of data sources. Case studies might be explanatory, exploratory, or event-based.
- The table below summarizes the differences between this methods.

**Table 3.1 Difference between the Five Models**

| **Method** | **Focus** | **Data Collection** |
|---|---|---|
| Ethnography | Context or culture | Observation & interviews |
| Narrative | Individual experience & sequence | Individual and document stories |
| Phenomenological | Individuals who have witnessed a phenomenon | Interviews |
| Grounded Theory | Create a theory based on field facts. | Interviews, followed by open and axial coding |

| | | |
|---|---|---|
| Case Study | Organization, entity, person, or event | Documents, interviews, reports, and observations |

### 3.3 Research Population and Sample

The total group of individuals and objects has referred to as a population or Target population. Individuals in the population have certain features in common. The target audience for this study is all types of people obviously because everyone needs water. It would have been impractical to conduct this study on the entire target population and therefore narrow down the population.

#### 3.3.1 Sampling Technique

Here in this research, the sampling method is used in the survey is a **Simple Random Stratified Sampling Technique**. With this approach, a random sample is chosen from each stratum after splitting the population into smaller groups or strata. When the population is heterogeneous and you want to make sure that the sample is representative of several subgroups, this strategy can be helpful.

Both primary and secondary data is used in the analysis in order to evaluate the objectives and their attainment. Primary data are gathered through filling out questionnaires and interviewing respondents, including both working professionals and other people.

Sample is a portion of a larger population. The sample is a designated group of certain elements drawn from the entire population. The analysis of the sample can reveal anything about the total population. Sampling is employed when a population is too large to survey everyone. Sampling is used to gauge a presumed population characteristic.

The method, which is used to collect the data, was based on the parameters like Sample Population Size, Confidence Level, and Margin of Error in terms of percentage.

The Correlation based on the Value Increased or Decreased are mentioned below:

**Table 3.2 Correlation-Based Values**

| **Parameters** | **Value increased** | **Value decreased** |
|---|---|---|
| **Population size** | Accuracy decreases | Accuracy increases |
| **Sample size** | Accuracy increases | Accuracy decreases |
| **Confidence level** | Accuracy increases | Accuracy decreases |
| **Margin of error** | Accuracy decreases | Accuracy increases |

### 3.3.2 Calculating Sample Size:

The following formula used for calculating the necessary sample size-

Formula:

$$ss = \frac{(Zscore)^2 * PoP(1-PoP)}{(moe)^2}$$

Note: ss is the sample size, Zscore is the z-value, PoP is the proportion of population (generally taken as 0.5), moe is the margin of error

| Desired confidence level | zscore |
|---|---|
| 80% | 1.28 |
| 85% | 1.44 |
| 90% | 1.65 |
| 95% | 1.96 |
| 99% | 2.58 . |

……(Source:https://www.geopoll.com/blog/sample-size-research/)

Where z-scores for the most common confidence levels:

$$\text{Necessary Sample Size} = \frac{(1.96)^2 \text{ X } 0.5 \text{ X } (1-0.5)}{0.05 \text{ X } 0.05} = 384.16 = 385$$

**Our sample size= 413 (survey respondents)**

**Interview respondents = 27**

**Total Sample Size = 413 + 27 = 440**

The list of procedures used for Data Collection are as follows:

a. Surveys
b. Key Informants Interviews

**Water sample Collection:**

The data is collected using various methods. The data was collected from different sources, including government agencies, academic institutions, and non-governmental organizations. The type of water considered for data collection is depends on the objective of the study.

The proposed water dataset includes the data collected from different types of drinking water depending on the location and purpose of the study. Regarding water quality specifically, the parameters measured include physical parameters such as temperature, pH, turbidity, and conductivity, as well as chemical parameters such as dissolved oxygen, nutrients, metals, and organic compounds. The specific parameters measured depend on the objectives of the study and the type of water body being studied.

The study is focused on drinking water quality, the parameters measured are different from those measured in a study focused on assessing the ecological health of a river. Similarly,

if the parameters measured on marine ecosystems can be different than those measured in a study focused on freshwater ecosystems.

Season and climate should be taken into account when gathering data because they can have a big effect on water quality. For instance, increased runoff and erosion during the wet season may lead to higher amounts of suspended solids, nutrients, and other pollutants in the water. Water levels may drop during the dry season, which may lead to a rise in the amount of pollutants in the water.

**Scales used for water Data Measurements:**

Various scales are used to measure the data collected for the proposed dataset. The physical parameters of the water bodies, such as temperature, pH, and turbidity, will be measured using standard units of measurement. The chemical parameters, such as dissolved oxygen, nutrients, and heavy metals, will be measured using milligrams per liter (mg/L) or parts per million (ppm). The biological parameters, such as the presence of microorganisms, will be measured using the number of organisms per milliliter (ml) or colony-forming units (CFU).

**Table 3.3 Scales used for Data measurement**

| Data Type | Scale | Unit of Measurement |
|---|---|---|
| Physical parameters | Temperature | Celsius (°C) |
| | pH | pH units |
| | Turbidity | Nephelometric Turbidity Units (NTU) |
| Chemical parameters | Dissolved oxygen | Milligrams per liter (mg/L) |
| | Nutrients (e.g., Nitrogen, Phosphorus) | Parts per million (ppm) |
| | Heavy metals (e.g., lead, mercury) | Parts per million (ppm) |

## 3.4 Questionnaire Formation

The survey questionnaire is designed to analyze the user's awareness about drinking water. Respondents are given survey questionnaire covering all aspects of research objectives. There were total of 18 questions in the survey for collecting general data as well as factors related to awareness of water quality and its impact.

**Table 3.4: Detailed Demography of Survey Questionnaire**

| **Details** | **Remarks** |
|---|---|
| Demographic and Authentication | Questions Related to the Background of Respondents |
| Authentication and Authorization | Recognizing user identity and authority |
| Integrity factors | Questions related correctness |
| Confidentiality | Preventing user's information from sharing or disclosure to the third party |
| Factors for trust | Questions related to trust |

## 3.5 Reliability and validity Test of Questionnaire:

### a. Reliability of questionnaire using Cronbach's alpha test

Reliability was measured using Cronbach alpha test. The reliability coefficient between 0.6 – 0.7 is considered acceptable.

- below .60 unacceptable
- between .60 and .65 undesirable
- between .65 and .70 minimally acceptable
- between .70 and .80 respectable
- between .80 and .90 very good

**Table 3.5 Reliability Statistics**

| Reliability Statistics(Survey) | | |
|---|---|---|
| **Cronbach's Alpha** | **Cronbach's Alpha Based on Standardized Items** | **No of Items** |
| 0.81 | 0.84 | 18 |
| **Reliability Statistics (Interviewer)** | | |
| **Cronbach's Alpha** | **Cronbach's Alpha Based on Standardized Items** | **No of Items** |
| 0.81 | 0.98 | 07 |

**b. Validation of questionnaire after discussion of domain expert**

Validity describes how accurately a system does the calculations it is designed to perform. The validity of the research questionnaire was assessed by a conversation with subject-matter specialists. Studies with a high level of validity offer conclusions that relate to the characteristics, traits, and variances seen in the real world.

## 3.6 Sample Questionnaire on Drinking Water Quality

### 3.6.1 Sample Questionnaire for Survey

This survey has been taken for understanding the awareness of water quality in different areas. Kindly answer all the following questions mentioned in the given survey.

**Section I. Demographic Questions**

1. Name

……..

2. Email ID

……..

3. Phone Number

……..

4. Designation

……..

5. Organization Name

……..

6. Location

……..

**Section II. Drinking Water and its Quality**

7. How much water do you drink per day?

- 0.5 L
- 1 L
- 1.5 L
- 2 L
- 3 L
- More than 3L

8. Which is the source of water intake on daily basis?

- Tap water
- Filtered Water
- Packed Water
- Outsource
- Other

9. Percentage of sick leaves in your organization?

- 5% - 15%
- 15% - 25%
- 25% - 35%
- More than 35 %

10. What is the pH level of drinking water in your organization

- No idea
- <1
- 1 to 4
- 4 to 6.5
- 6.5 to 8.5

- 8.5 to 10
- 10 to 12
- 12 to 14

11. What is the TDS level of drinking water in your organization?

- Excellent (less than 300mg/L)
- Good (300mg/L-600mg/L)
- Fair (600mg/L-900mg/L)
- Poor (900mg/L-1200mg/L)
- Unacceptable (greater than 1200mg/L)
- No idea about TDS

12. Which Filtered water do you prefer in your organization?

- Bisleri
- Kinley
- Aquafina
- Bailley
- RO
- Other/Local Brands

13. According to you , where is the water of highest quality found?

- From wells
- From packed water
- From tap water
- RO
- No Idea

14. What percentage of overall budget is allocated in your company towards purified water for drinking water?

- 0 to 20%

- 20 to 40%
- 40 to 60%
- 60 to 80%
- 80 to 100%

15. Have you analyzed your drinking water quality in laboratory?

- Yes, Frequently
- Never
- Yes,Once upon a time
- No Idea

16. In your opinion, who should be responsible for drinking water quality

- Water Neutralization
- Filteration of water
- Chlorination
- Other water purification Techniques

17. In your opinion , who should be responsible for drinking water quality?

- Society for conversation /Pollution of water
- Municipalities
- Water suppliers
- In person
- Everyone as mentioned

18. How do you preserve drinking water in your home/Organization?

- Water tanks
- Earthen pots
- Plastic bottles
- Steel vessels
- Others

### 3.7 Analysis Report of Data Collected through Survey

#### 3.7.1 Location of organizations

Outsourcing of Water supply works across the country, However the following cities covered majorly to collect the input for versatile results. Some respondents were given online questionnaires to fill, some filled them in hard copy, printed formats. Some values were taken during interviews.

**Table 3.6: Count of Locations of Respondents**

| **Location** | **Count** |
|---|---|
| Mumbai | 177 |
| Pune | 98 |
| Bengaluru | 73 |
| Delhi | 29 |
| Sholapur | 13 |
| Hyderabad | 10 |
| Other Cities | 13 |
| **Total** | **413** |

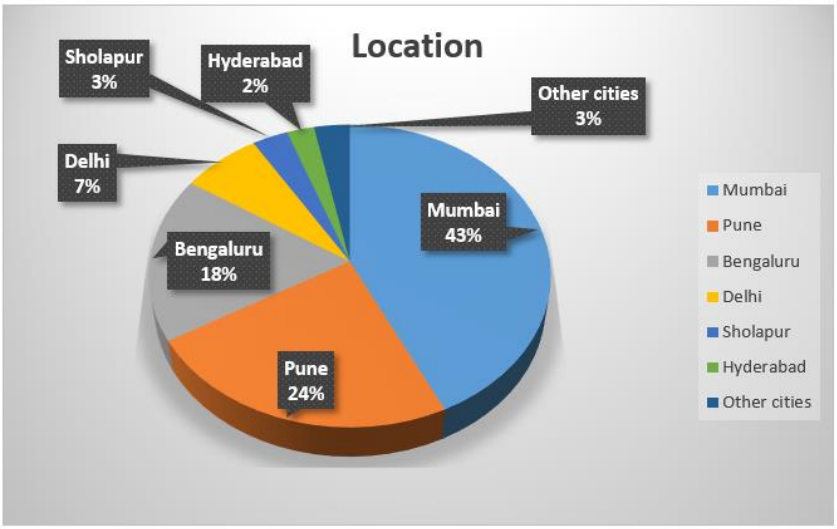


**Figure.3.1 Location of Organizations**

### 3.7.2 Water Intake Per Day

The questionnaire was initiated from the consumption the water on a per-day basis in person who is the representative for their respective Organizations. Based on the analysis it has been identified that out of these heads ,44% people drink more than 3 Liters of water per day, 16% around 3litres,15% 2 liters, 14% 1.5 liters, 9% 1 liter, and 2% 0.5 liters in the entire day. As per WHO guidelines 3 to 4 liters of water should be taken to hydrate the body is mandatory, however, it depends on the weight and other parameters as well.

**Table 3.7: Water Intake per Day in Respondents Organizations**

| Water Intake per Day | Count |
|---|---|
| 0.5 Litres | 9 |
| 1.0 Litres | 38 |
| 1.5 Litres | 56 |
| 2 Litres | 61 |
| 3 Litres | 66 |
| More than 3 Litres | 183 |
| **Total** | **413** |

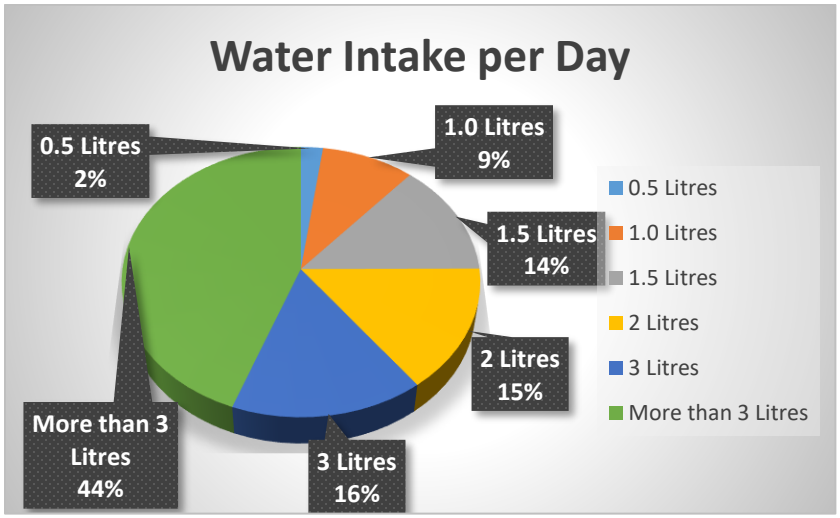


**Figure 3.2: Water Intake Per day**

### 3.7.3 Source of Drinking Water

According to analysis, 53% of people started using Filtered Water as compared to other resources.

**Table 3.8: Count of Origin of the Source of Drinking Water**

| **Source of drinking water** | **Count** |
|---|---|
| Tap water | 15 |
| Filtered Water | 219 |
| Packed Water | 96 |
| Outsource | 78 |
| Other | 5 |
| **Total** | **413** |

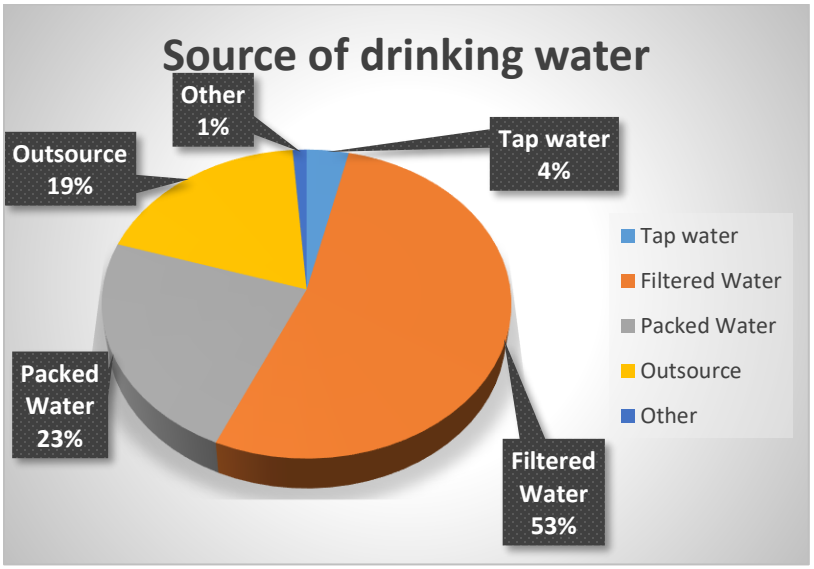


**Figure 3.3: Source of water**

### 3.7.4 Percentage of Sick Leave on Monthly basis

The short term sick leave is mostly due to indigestion, fever or so with in the organization. It has been identified that 5 to 15 % people are more in the organizations who take sick leave monthly. 25% to 35% take less sick leave where 57 in count answered that they don't take any sick leave as they are health conscious and mostly bring their own water bottle from home.

**Table 3.9: Percentage of Sick Leave**

| % of Sick Leaves | Count |
|---|---|
| >35% | 36 |
| 25-35% | 30 |
| 15-25% | 122 |
| 5-15% | 168 |
| No Sick Leave | 57 |
| **Total** | **413** |

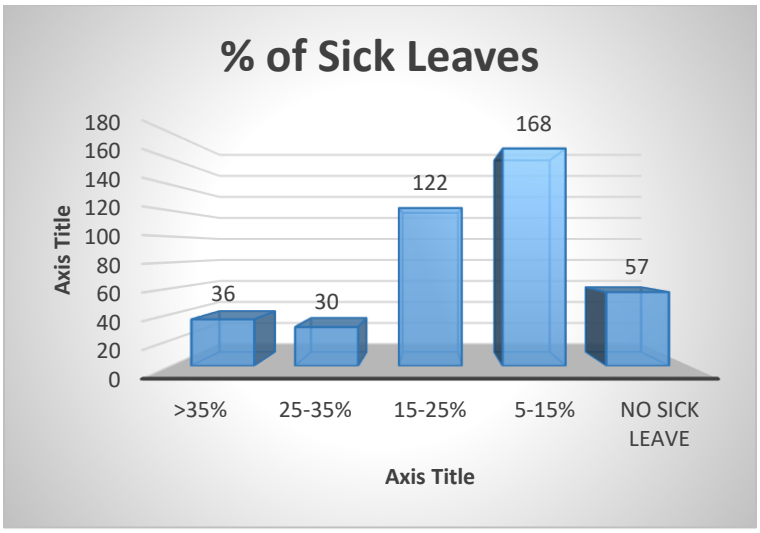


**Figure.3.4: Percentage of sick leave**

### 3.7.5 Level of TDS in drinking water

TDS level is the initial phrase of water quality testing used commonly. It has been identified that most of the organizations were aware about TDS and they maintain the level of TDS as per the standards. However, the count values identified for Excellent level of water for drinking were maintained which should be less than 300mg/L. It has also been noticed that 15 count were identified for unacceptable values and 73 in count were no aware about TDS value of so.

**Table 3.10: Count of TDS Level Identified in the Organizations**

| TDS Level | Count |
|---|---|
| Excellent (<300mg/L) | 163 |
| Good (300mg/L to 600mg/L) | 64 |
| Fair (600mg/L to 900mg/L) | 53 |
| Poor (900mg/L to 1200mg/L) | 45 |
| Unacceptable (>1200mg/L) | 15 |
| No idea about TDS | 73 |
| **Total** | **413** |

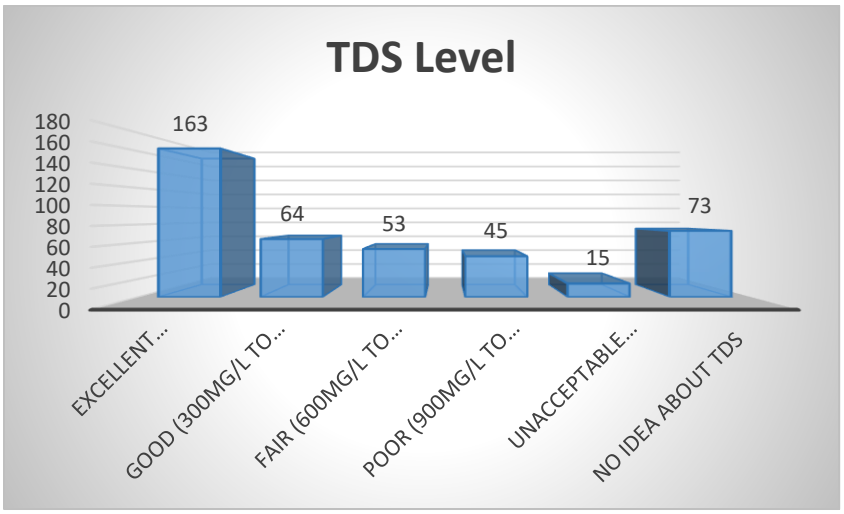


**Figure 3.5: TDS Level Identified in the Organizations**

### 3.7.6 Filtered/Packed water bottle opted in the Organization

46% of the organizations use other/local brand for drinking water, it is cheaper than other brands as well. However, organizations only faith to their local vendors. Moreover, few organizations installed one time RO which is of 25% in the organization and they maintain it by themselves.

**Table 3.11: Opted Filtered Water in the Organizations**

| Opted Filtered water | Count |
|---|---|
| Bisleri | 45 |
| Kinley | 32 |
| Aqufina | 28 |
| Bailley | 15 |
| RO | 103 |
| Other/Local Brand | 190 |
| **Total** | **413** |

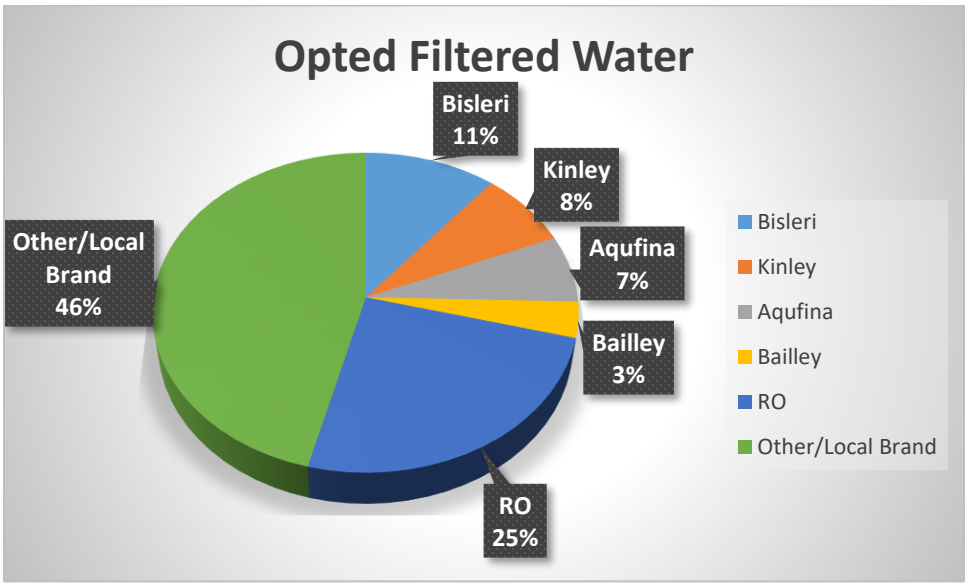


**Figure 3.6: Opted Filtered Water in the Organizations**

### 3.7.7 PH level of drinking water

PH level is another way to check the drinking water quality. The best PH level is 6.5 to 8.5 which is suitable for drinking water. However, 83 count in person do not have any idea about PH level and they never tested. Here 2 in count organizations identified that the drinking water used is highly acidic and the value was <1 which was very harmful to health. Where 5 organization identified use to drink highly alkaline water and that to even not good for health. Out of 413 only 107 people are thinking they are maintaining the PH level.

**Table 3.12: PH Level of Drinking Water within the Organization**

| PH level of drinking water | Count |
|---|---|
| No Idea | 83 |
| <1 | 2 |
| 1 to 4 | 25 |
| 4. to 6.5 | 92 |
| 6.5 to 8.5 | 107 |
| 8.5 to 10 | 74 |
| 10 to 12 | 25 |
| 12 to 14 | 5 |
| Total | **413** |

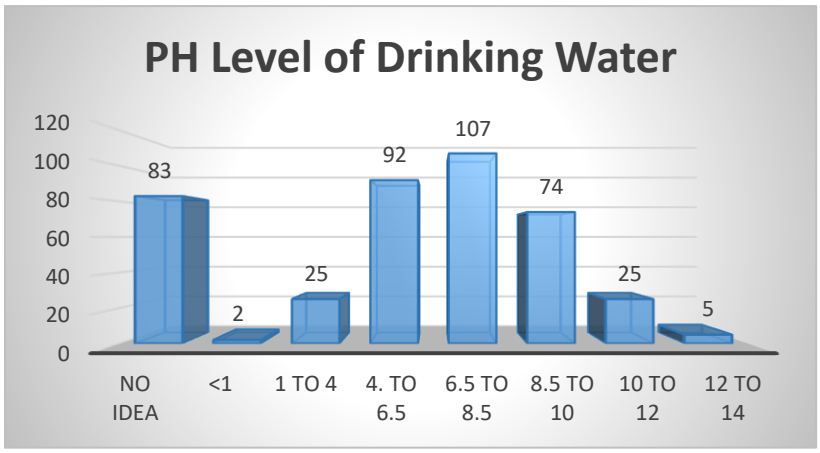


**Figure 3.7: PH level of Drinking Water**

### 3.7.8 Water source with the highest quality

Out of the total respondents, 35% of people feel that packed water carries the highest quality and 26% of people feels RO water is quality water as compared to various other common resources available.

**Table 3.13: Water Source with the Highest Quality**

| **Water source with the highest quality** | **Count** |
|---|---|
| Wells/ underground water | 63 |
| Packed Water | 146 |
| RO | 106 |
| Tap Water | 96 |
| No idea | 2 |
| **Total** | **413** |

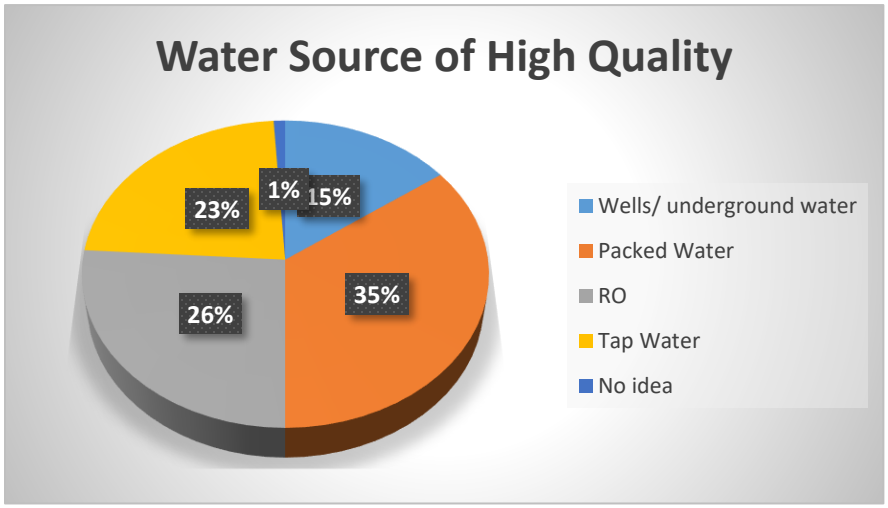


**Figure 3.8: Water Source with the Highest Quality**

### 3.7.9 Budget allotted for Water purification

Out of the total number of respondents, 113 count says that now the organization take care of water purification and they have a proper budget for the same, However, 60 in count employees are disheartened because of the improper amenities. Overall, it has been identified that the organizations who outsource the water from local vendors have less budget and employees are not satisfied due to such inconvenience.

**Table 3.14: The Budget Allotted for Water Purification**

| The budget allotted for Water purification | Count |
|---|---|
| 0 to 20% | 60 |
| 20-40% | 67 |
| 40-60% | 93 |
| 60-80% | 113 |
| 80%-100% | 80 |
| **Total** | **413** |

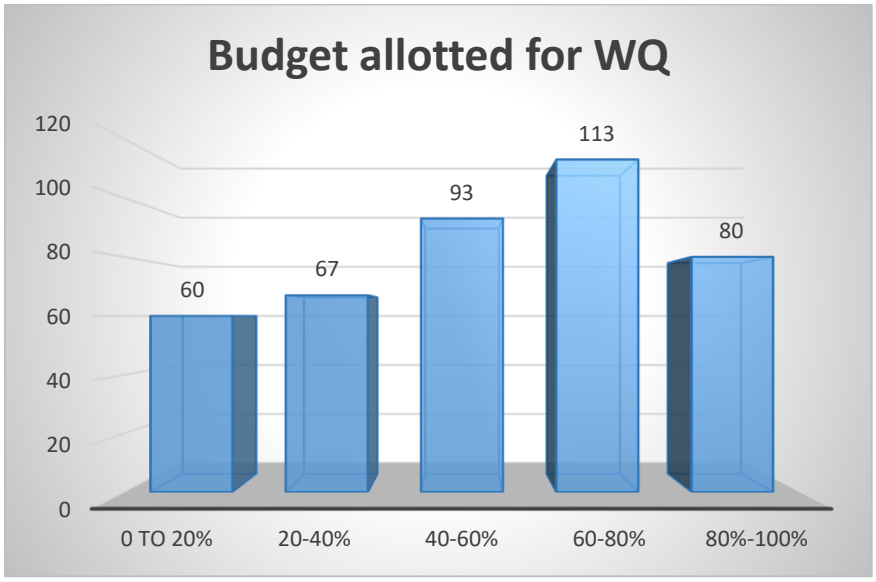


**Figure 3.9: Budget allotted for Water purification**

### 3.7.10 Water analyzed from Laboratory

Out of the total number of respondents, 53% say that they never went to any laboratory for water testing/analysis. Only 3% of organization frequently send the water sample for Laboratory test.

**Table 3.15: Water Analyzed from Laboratory**

| **Water analysed from Laboratory** | **Count** |
|---|---|
| Yes, Frequently | 12 |
| Yes, Once upon a time | 60 |
| Never | 219 |
| No Idea | 122 |
| **Total** | **413** |

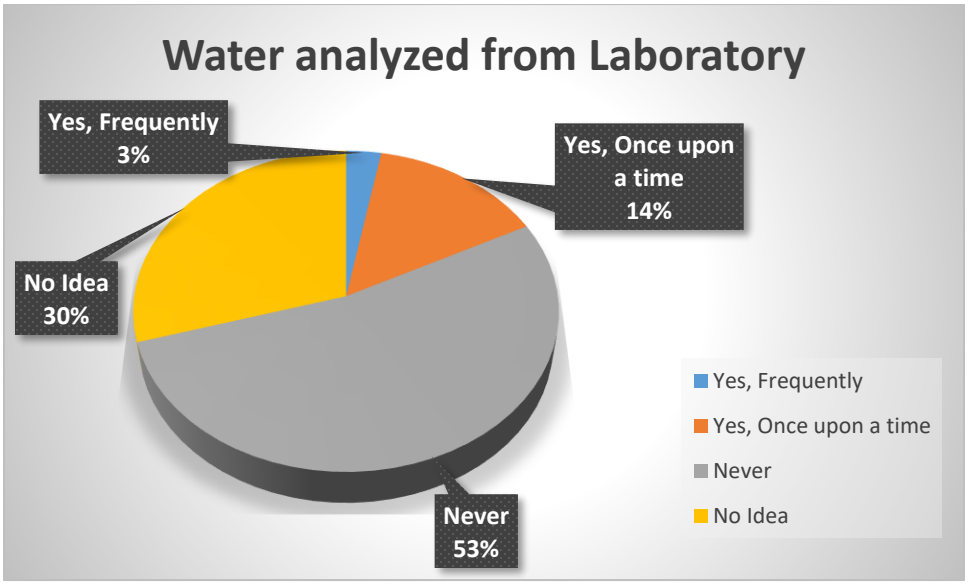


**Figure 3.10: Water analyzed from Laboratory**

### 3.7.11 Process followed to improve water quality

Out of the total number of respondents, 53% say they filter the drinking water, whereas 34% of people chlorinate the drinking water and 9% of people use other techniques.

**Table 3.16: Process to Improve Water Quality**

| The process to improve water quality | Count |
|---|---|
| Water Neutralization | 15 |
| Filtration of Water | 217 |
| Chlorination | 142 |
| Other techniques | 39 |
| **Total** | **413** |

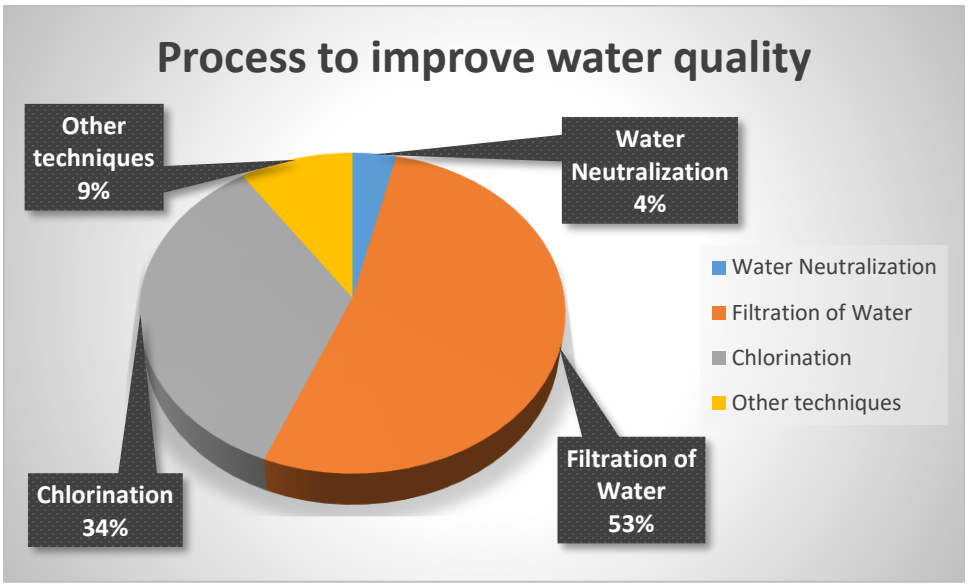


**Figure 3.11: Process to Improve Water Quality**

### 3.7.12 Responsible for drinking water quality

Out of the total number of respondents, 29% Municipalities and 27% Water suppliers are responsible, where 34% people chlorinate the drinking water.

**Table 3.17: Responsible for Drinking Water Quality**

| **Responsible for drinking water quality** | **Count** |
|---|---|
| Water | 72 |
| Water suppliers | 112 |
| Municipalities | 121 |
| In person | 60 |
| Everyone as mentioned | 48 |
| **Total** | **413** |

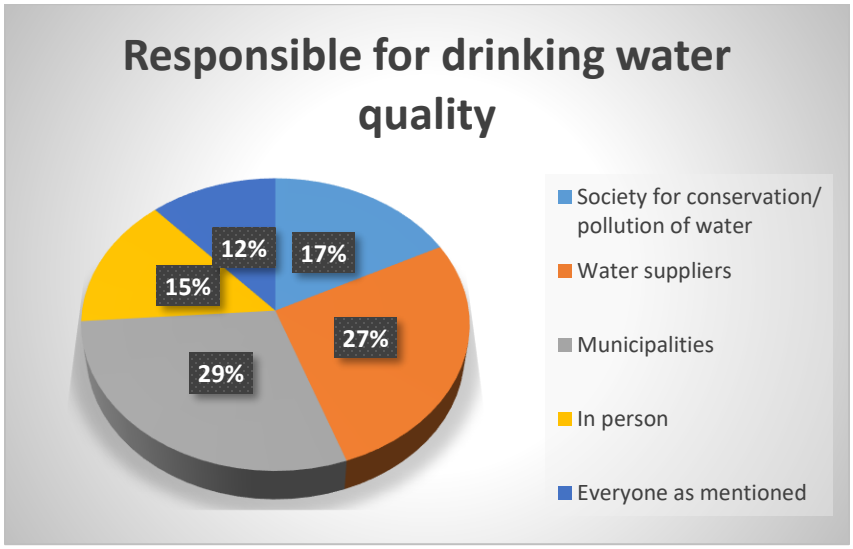


**Figure 3.12: Responsibility of Drinking Water Quality**

### 3.7.13 Process to preserve drinking water

Out of the total number of respondents, 33% of people uses the water tank to preserve drinking water, whereas 6% of people use Earthen Pots, and the majority of people are uses Plastic Bottles to preserve the king water which is not good for health.

**Table 3.18: The Process to Preserve Drinking Water**

| **The process to preserve Drinking water** | **Count** |
|---|---|
| Water Tanks | 138 |
| Earthen Pots | 24 |
| Plastic Bottles | 152 |
| Steel Vessels | 56 |
| Others | 43 |
| **Total** | **413** |

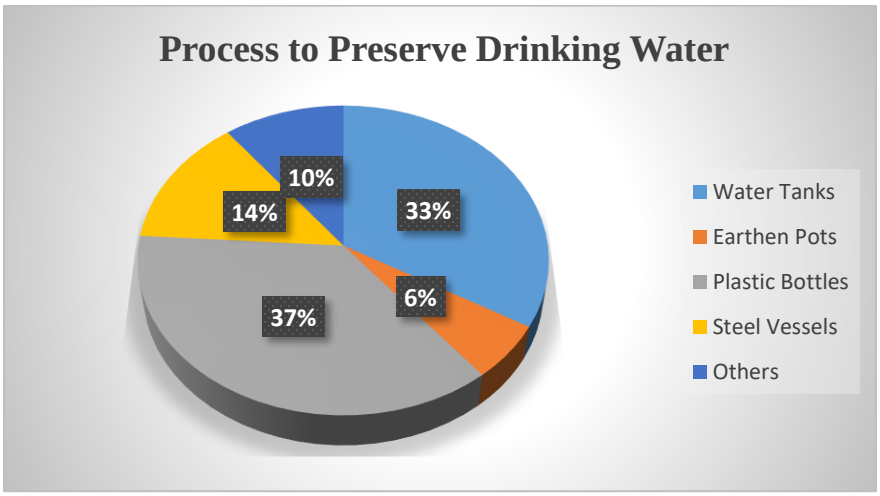


**Figure 3.13: The process to Preserve Drinking Water**

### 3.8 Key Understandings

The key understandings that we have got after analysis of the collected data are as follows:

1) The survey has been conducted on seven different types of Organizations.
2) As per water intake per day analysis it has been identified that out of these heads, 44% are people drink more than 3 Litres of water per day, 16% around 3litres,15% 2 liters, 14% 1.5 liters, 9% 1 liter, and 2% 0.5 liters in the entire day. As per WHO guidelines 3 to 4 liters of water should be taken to hydrate the body is mandatory, however, it depends on the weight and other parameters as well.
3) As per the outcome, 53% of people started using Filtered Water as compared to other resources.
4) Percentage of Sick Leave on a Monthly basis is due to indigestion, fever or so within the organization. It has been identified that 5 to 15 % of people are more in the organizations who take sick leave monthly. 25% to 35% take less sick leave where 57 in count answered that they don't take any sick leave as they are health conscious and mostly bring their own water bottle from home.
5) TDS level is the initial phase of water quality testing used commonly. It has been identified that most of the organizations were aware about TDS and they maintain the level of TDS as per the standards. However, the count values identified for an Excellent level of water for drinking were maintained which should be less than 300mg/L. It has also been noticed that 15 in count were identified for unacceptable values and 73 in count were no aware about TDS value of so.
6) As per the comparison in branded water supplier, 46% of the organizations use other/local brand for drinking water, it is cheaper than other brands as well. However, organizations only faith to their local vendors. Moreover, few organizations installed one-time RO which is of 25% in the organization and they maintain it by themselves.
7) PH level is the another way to check the drinking water quality. The best PH level is 6.5 to 8.5 which is suitable for drinking water. However, 83 count in person do not have any idea about PH level and they never tested. Here 2 in count organization identified that the drinking water used is highly acidic and the value was <1 which

was very harmful for the health. Where 5 count in people identified use to drink highly alkaline water and that to even not good for health. Out of 413 only 107 organizations are maintaining the PH level.

8) As per the source of high water quality, Out of the total respondents, 35% people feel that packed water carries the highest quality as compared to various other common resources available.
9) As per budget is concern, out of the total number of respondents, 113 count says that now the organization take care of water purification and they have proper budget for the same, However, 60 in count employees are dishearten because of the improper amenities. Overall, it has been identified that the organizations who outsource the water from local vendors have less budget and employees are not satisfied due to such inconvenience.
10) Out of the total number of respondents, 53% say that they never went to any laboratory for water testing/ analyzing. Only 3% organization frequently send the water sample for Laboratory test.
11) Out of the total number of respondents, 53% say they filter the drinking water, where 34% people chlorinate the drinking water.
12) Out of the total number of respondents, 29% Municipalities and 27% Water suppliers are responsible, where 34% people chlorinate the drinking water.

### 3.9 Sample questionnaire for Interviewers

**General Questions:**

Name of the Employee:_____

Name of The Organization:_____

Type of the Company:_____

Designation:_____

**Water Quality Related Questions:**

1. Which filtered water do you prefer in your organization

   a. Water Purifier
   b. Outsourced /Packaged Drinking
   c. Tap Water
   d. Plastic Bottle
   e. Bring your own Water (BYOW)
2. Do you trust the Water Supply of your organization?
   a. Yes
   b. No
3. What are the pH and TDS levels of Drinking water in your organization?
   a. PH Level
   b. TDS Level
4. What are the reasons for sick leave in your Organization
   (Climate change, Water (water retention/dehydration), Food Infection, Stress)
5. How frequently do you test the Water Quality
   a. Quarterly
   b. Six Monthly
   c. Annually
   d. Ignore
6. Is there any overall budget is allocated in your company towards purified drinking water?
   a. Yes
   b. No
7. Are there effective complaint mechanisms at the level of the water service provider?
   a. Yes
   b. No
   c. In progress

### 3.10 Analysis Report of Interviews:

The following table shows the companies covered for interviews overall throughout Mumbai location.

**Table 3.19 Interviewer Organization Details**

| Sr.No. | Organization Name |
|---|---|
| 1 | Water Filtration Plant |
| 2 | 900MLD Water Treatment Plant |
| 3 | Aqua Fresh Water Technology |
| 4 | IRCTC Water Treatment Plant |
| 5 | APX Water Treatment Plant |
| 6 | Viksara Complete Water Solutions |
| 7 | Krupasindhu Consulting Engineers |
| 8 | Monitor finance |
| 9 | Shriram Finance Limited |
| 10 | AAA Finserve - A Finance Company Mumbai |
| 11 | DES's NMITD |
| 12 | Mumbai Educational Trust |
| 13 | B.N.Bandodkar College |
| 14 | Vikas College |
| 15 | Hinduja Hospital |
| 16 | Nanavati Hospital |
| 17 | Bothras Hospital |
| 18 | High Court Mumbai |
| 19 | Post Office |
| 20 | SBI Bank |
| 21 | Connectwise |
| 22 | Predoole analytics |
| 23 | Iraje Softwares |
| 24 | Mobicule Technologies Pvt. Ltd. |
| 25 | BNP Paribas |
| 26 | Siddhnath Patpedhi |
| 27 | KVAR Technologies |

### 3.10.1 Type of Organizations

Among all the respondents, the majorly targeted companies are Water Treatment Plants, Outsource Companies and Financial/Banking services. The following details collected by the personal interaction with the senior officers of their respective companies.

**Table 3.20 Type of Organization**

| Sr.No | Type of the Organization | Count |
|---|---|---|
| 1 | WTP | 7 |
| 2 | Finance | 3 |
| 3 | Education | 4 |
| 4 | Healthcare | 3 |
| 5 | Government | 3 |
| 6 | Outsource Companies | 5 |
| 7 | Other | 2 |
| | **Total** | **27** |

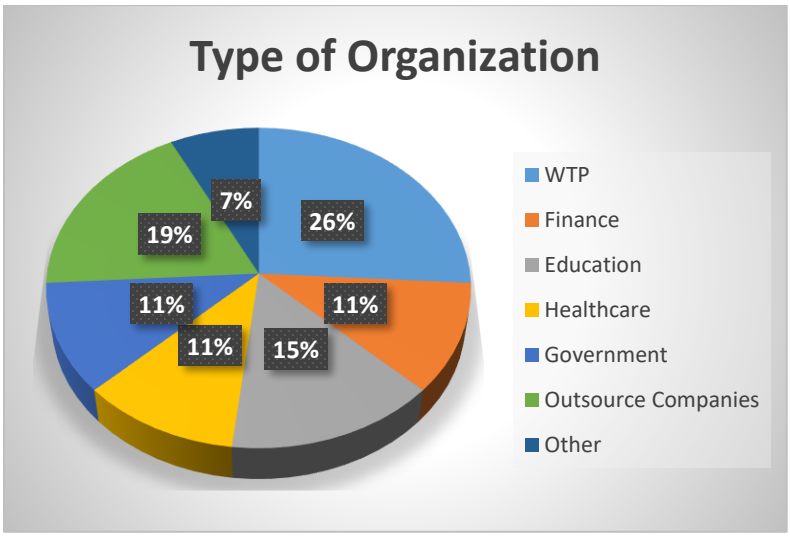


**Figure 3.14: Type of Organizations Covered During Interview**

### 3.10.2 What Type of filtered water do you prefer in your organization?

According to the interviewers, 37% of people are using Water purifiers in their organizations to give good quality water to the employees. Only 3 % of people preferred Tap water for drinking water.

**Table 3.21: Classification of Type of Water as per the Use In Organizations**

| Sr. No | Which filtered water do you prefer in your organization? | Count |
|---|---|---|
| 1 | Water Purifier | 10 |
| 2 | Outsourced /Packaged Drinking | 4 |
| 3 | Tap Water | 3 |
| 4 | Plastic Bottle | 5 |
| 5 | Bring your own Water (BYOW) | 5 |
| | **Total** | **27** |

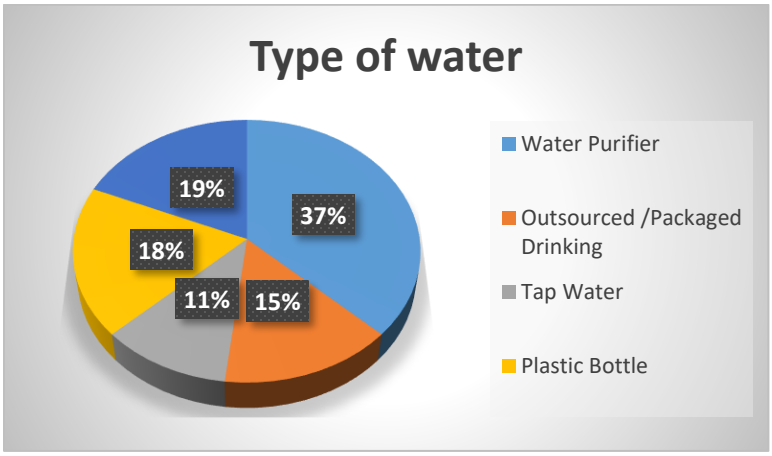


**Figure 3.15 : Classification of Type of Water as per the Use in Organizations**

### 3.10.3 Do you Trust the water supply of your Organization

According to the survey, 78% trust their water supply of organization. It means the quality of water which is provided by the government is trustable as per them. But some of the organizations have problems with the method of water distribution, water pipe lekage ,etc .

**Table 3.22: Trust Factor Analysis**

| Sr. No | Do you trust the Water Supply of your organization? | Count |
|---|---|---|
| 1 | Yes | 21 |
| 2 | No | 6 |
| | **Total** | **27** |

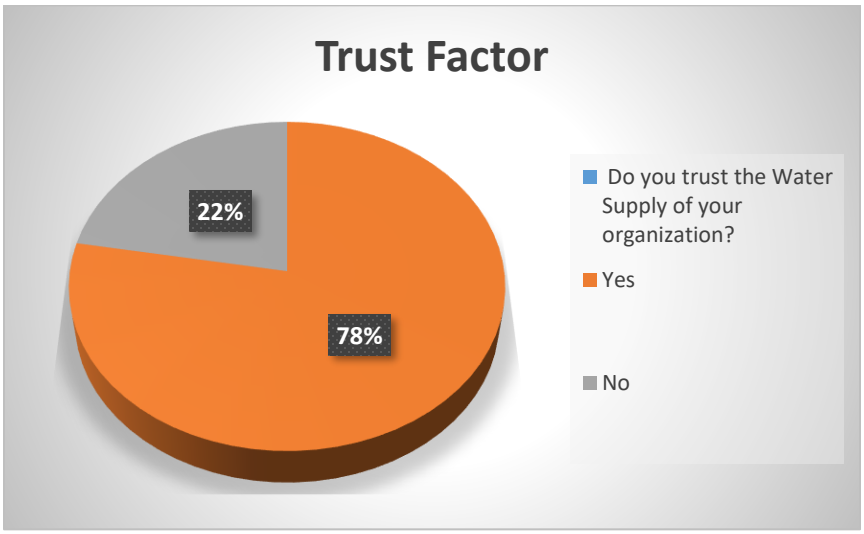


**Figure 3.16 : Trust Factor Analysis**

### 3.10.4 What are the pH and TDS levels of Drinking water in your organization?

As per survey, 33% of people does not know about the pH level and TDS level of their drinking water quality. Only 67% people knows about their pH level and TDS level of Drinking Water.

**Table 3.23: Idea about pH level of Drinking Water**

| Sr. No | Idea about pH level of Drinking water in your organization | Count |
|---|---|---|
| 1 | No idea | 9 |
| 2 | Knows about pH value(Tentative | 18 |
| | **Total** | **27** |

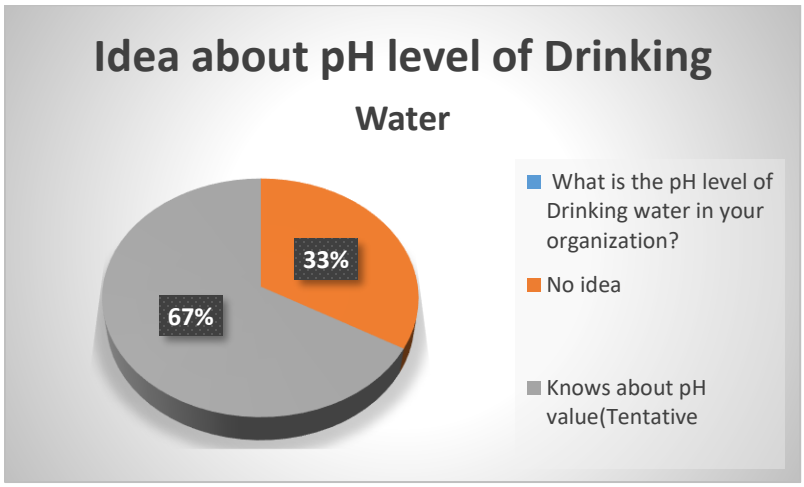


**Figure 3.17: Idea about pH level of Drinking Water**

**Table 3.24: Idea about TDS level of Drinking Water**

| Sr. No | What is the TDS level of Drinking water in your organization? | Count |
|---|---|---|
| 1 | No Idea | 12 |
| 2 | Knows about TDS value(Tentative) | 15 |
| | **Total** | **27** |

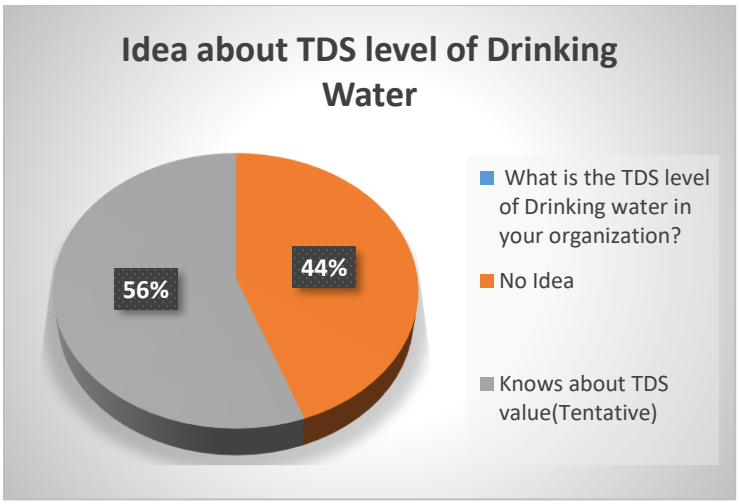


**Figure 3.18: Idea about TDS level of Drinking Water**

### 3.10.5 What are the reasons for sick leave in your Organization

According to the survey, 48% of people feels because of water quality they get problem and they take sick leaves. Even because of climate change, people may face certain health issues and because of that, they take sick lives.

**Table 3.25: Count for Reasons for Sick Leave**

| Sr. No | What are the reasons for sick leave in your Organization | Count |
|---|---|---|
| 1 | Climate Change | 5 |
| 2 | Water | 13 |
| 3 | Stress | 5 |
| 4 | Food | 4 |
| | **Total** | **27** |

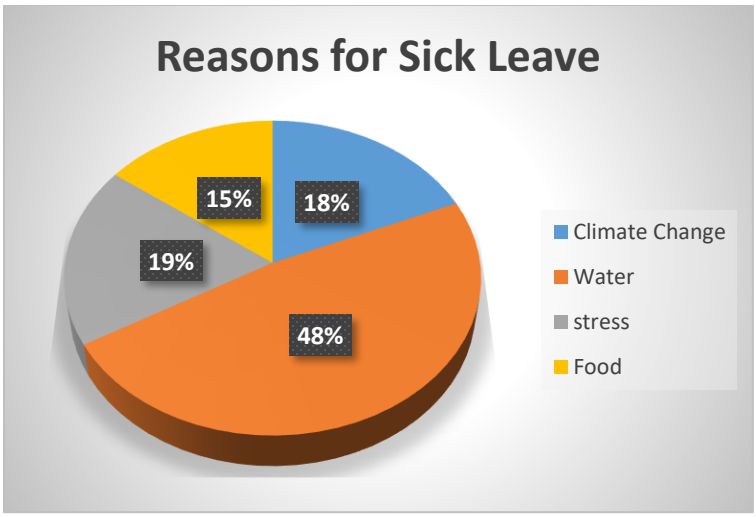


**Figure 3.19: Count for Reasons for Sick Leave**

### 3.10.6 How frequently do you test the Water Quality in your Organization?

The survey clearly shows that 41% of organizations test the water quality on an annual basis but almost 33% of organizations never do this type of testing nor another effort they take for water quality testing.

**Table 3.26: Frequency of Testing the Water Quality in Your Organization**

| Sr. No | How frequently do you test the Water Quality in your Organization | Count |
|---|---|---|
| 1 | Quarterly | 3 |
| 2 | six Monthly | 4 |
| 3 | Annually | 11 |
| 4 | Ignore | 9 |
| | **Total** | **27** |

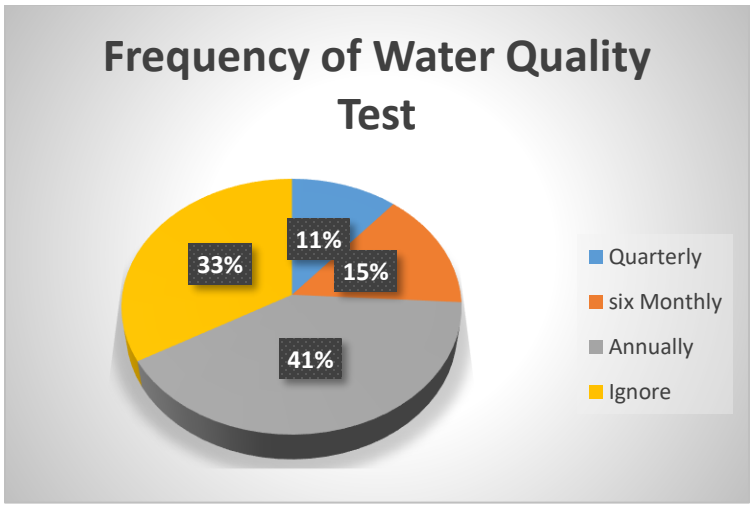


**Figure 3.20: Frequency of Testing the Water Quality in Your Organization**

### 3.10.7 Is there any overall budget is allocated in your company towards purified drinking water?

Overall analysis shows that 89% of organizations having budget allocated for water quality analysis. Only 3% organizations are not having separate provision of budget for water quality.

**Table 3.27: Overall Budget Allocated In Company towards Purified Drinking Water**

| Sr. No | Is there any overall budget allocated in your company towards purified drinking water? | Count |
|---|---|---|
| 1 | Yes | 24 |
| 2 | No | 3 |
| | **Total** | **27** |

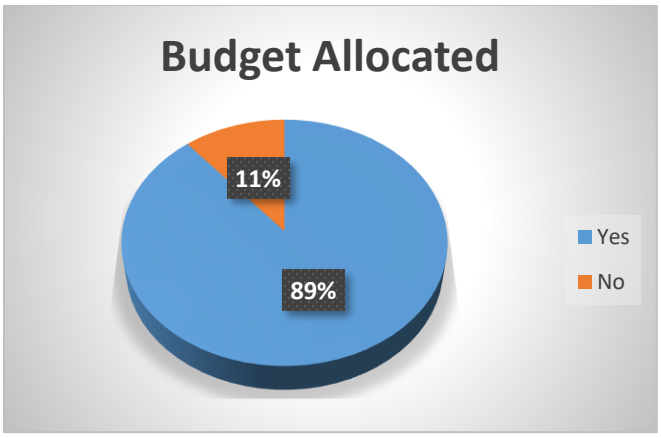


**Figure 3.21: Overall Budget Allocated in the Company towards Purification**

### 3.10.8 Are there effective complaint mechanisms at the level of the water service provider?

There is 56% of interviews are saying they are having Complaint mechanism in their organizations but others do not have.

**Table 3.28: Effective Complaint Mechanisms**

| Sr. No | Are there effective complaint mechanisms at the level of the water service provider? | Count |
|---|---|---|
| 1 | Yes | 12 |
| 2 | No | 15 |
| | **Total** | **27** |

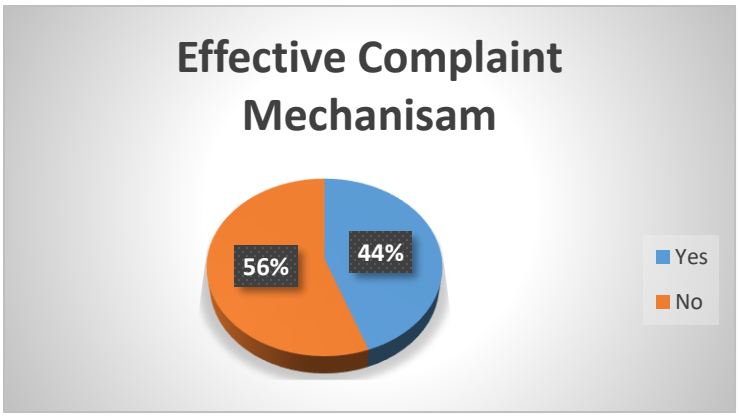


**Figure 3.22: Effective Complaint Mechanisms**

### 3.11 Key Understandings

The key understandings that we have got after analysis of the collected data are as follows:

- The survey has been conducted all over Mumbai with various types of organizations to get an idea about what they know about their drinking water quality.
- Total of 27 employees Interview taken who belong to different types of organizations.
- As per the outcome total of 27% water treatment plants, 11% of Finance, 15% of Education, 11% of Healthcare, 11% of Government,19% of Outsource companies, and 2% other are considered.
- Percentage of people who are using water purifiers i.e. 37% is more than the other resources. In the discussion, it has come to know that because of Covid this change has happened
- pH and TDS level is another way to identify drinking water quality but almost 35% to 45% of people do not have any idea about *it*.
- It has been noticed that the Percentage of sick leave is around 48% because of water.
- Still there are 33% of organizations, who does not even test the water quality annually or quarterly.
- Even 89% of organizations are having budget allocated for quality testing and maintenance but still ignorance is more. In addition, because of that it has been clear that water quality is very important anywhere whether it is in an organization or home.
- So the Nav Jal model can be implemented to have better water quality.

### 3.12 Linear Regression for two variable pH and TDS

Regression is a method for determining how independent features or variables relate to a dependent feature or result. It is a technique for machine learning predictive modelling, where an algorithm is used to forecast continuous outcomes. The linear equation that best fits the points is the linear regression. There are several methods for selecting the best-

fitting ting line, but the ordinary least squares (OLS) method is the most popular. The dependent variable (Y) and independent variables' (X) relationship is described via linear regression. By minimizing the sum of squares of the discrepancies between the observed dependent variables (Y) and the estimated value by the linear regression (), the ordinary least squares approach selects the line parameters.

This linear regression helps

- To predict the dependent variable (Y)
- Estimate the effect of each independent variable (X) on the dependent variable (Y).
- Calculate the correlation between the dependent variable and the independent variables.
- Test the linear model significance level.

Here in these study considered two variables pH and TDS to show the regression between these variables. Total twenty-seven readings of pH and TDS are considered which was collected from different places all over Mumbai.

**Table 3.29 : Water Sample Details of Organizations**

| Sample No | pH | TDS |
|---|---|---|
| 1 | 7 | 450 |
| 2 | 7.4 | 520 |
| 3 | 7.9 | 95 |
| 4 | 6.5 | 754 |
| 5 | 6.2 | 523 |
| 6 | 7.8 | 55 |
| 7 | 7.1 | 850 |
| 8 | 7 | 41 |
| 9 | 8 | 55 |
| 10 | 7.1 | 59 |
| 11 | 6.8 | 45 |
| 12 | 6.9 | 452 |
| 13 | 6.9 | 48 |
| 14 | 7.3 | 652 |
| 15 | 7.3 | 458 |

| | | |
|---|---|---|
| 16 | 6.8 | 54 |
| 17 | 8 | 62 |
| 18 | 7.5 | 45 |
| 19 | 7.5 | 41 |
| 20 | 7.5 | 352 |
| 21 | 7.3 | 158 |
| 22 | 7.1 | 42 |
| 23 | 6.9 | 42 |
| 24 | 6.8 | 56 |
| 25 | 7.2 | 82 |
| 26 | 7.3 | 74 |
| 27 | 7.1 | 54 |

**Line Fit Plot**

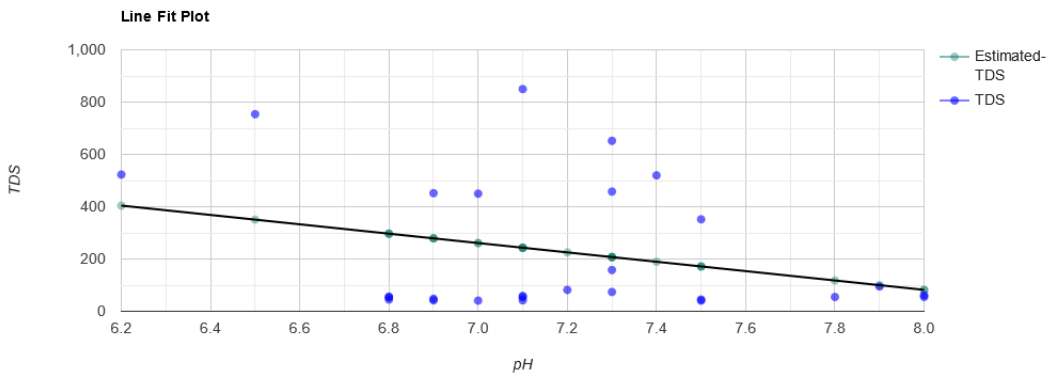


**Figure 3.23: Line Fit Plot**

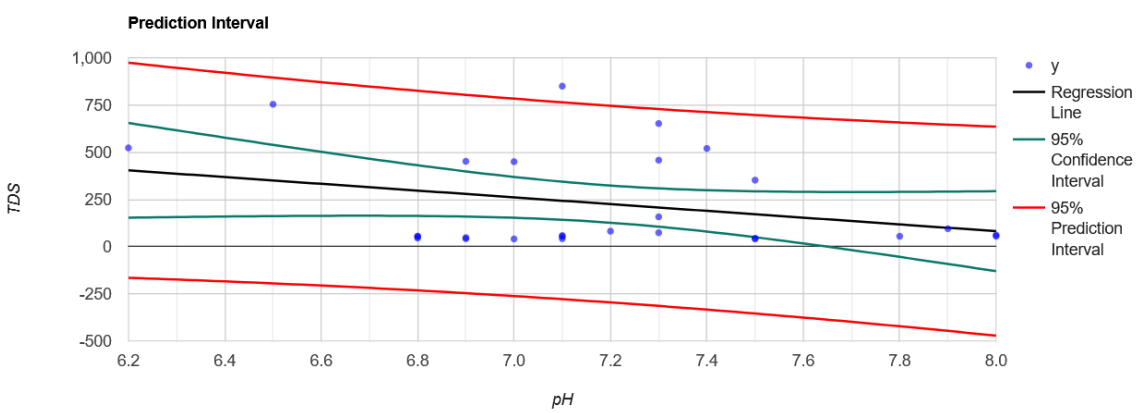


**Figure 3.24: Prediction Interval**

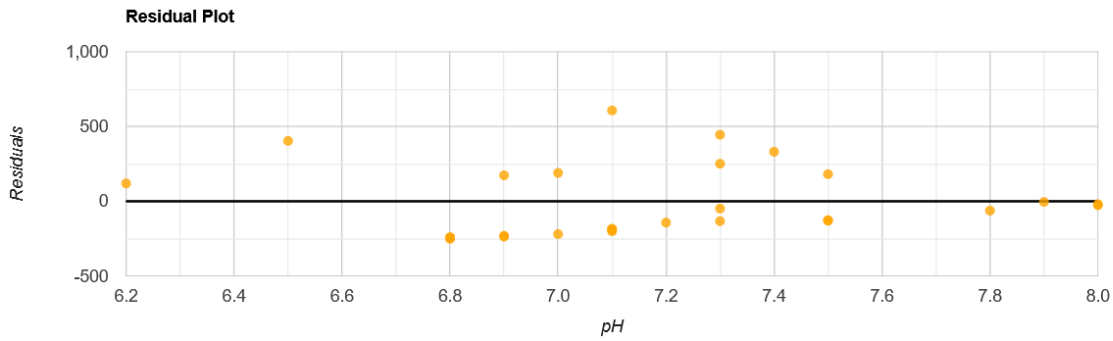


**Figure 3.25: Residual Plot**

**Table 3.30: Statistical Details of Regression**

| Source | Degree of Freedom | Sum of Square | Mean square | F Statistic (df1,df2) | P-value |
|---|---|---|---|---|---|
| Regression | 1 | 155023.547 | 155023.55 | 2.5158 (1,25) | 0.1253 |
| Residual | 25 | 1540526.75 | 61621.07 | | |
| Total | 26 | 1695550.3 | 65213.473 | | |

**1. Relationship of pH and TDS**

- R-Squared ($R^2$) equals **0.09143.** This means that 9.1% of the variability of TDS is explained by pH.
- Correlation (R) equals **-0.3024**. This means that there is a weak inverse relationship between pH and TDS.
- The slope: $b_1$=**-178.9957** CI[-411.4183, 53.4269] means that when you increase pH by 1, the value of TDS decreases by 178.9957.
- The y-intercept: $b_0$=**1514.0729** CI[-160.541, 3188.6868] means that when pH equals 0, the prediction of TDS's value is 1514.0729.
- The x-intercept equals 8.4587.

**2. Goodness of fit**

- Overall regression: right-tailed, F(1,25) = 2.5158, p-value = 0.1253. Since p-value ≥ α (0.05), we accept $H_0$.
- The linear regression model, $\mathbf{Y = b_0 + b_1X + \varepsilon}$, doesn't provide a better fit than the model without the independent variable resulting in $\mathbf{Y = b_0 + \varepsilon}$.
- The slope ($b_1$): two-tailed, T(25)=-1.5861, p-value = 0.1253. For one predictor it is the same as the p-value for the overall model.
- The y-intercept ($b_0$): two-tailed, T(25) = 1.8621, p-value = 0.07438. Hence, $b_0$ is not significantly different from zero. It is still most likely recommended not to force $b_0$ to be zero.

**3. Residual normality**

The linear regression model assumes normality for residual errors. The Shapiro-Wilk p-value equals 0.004109. It is assumed that the data is not normally distributed, You should consider transformation, or taking a bigger sample.

**4. Outliers**

The data does not contain any outliers.

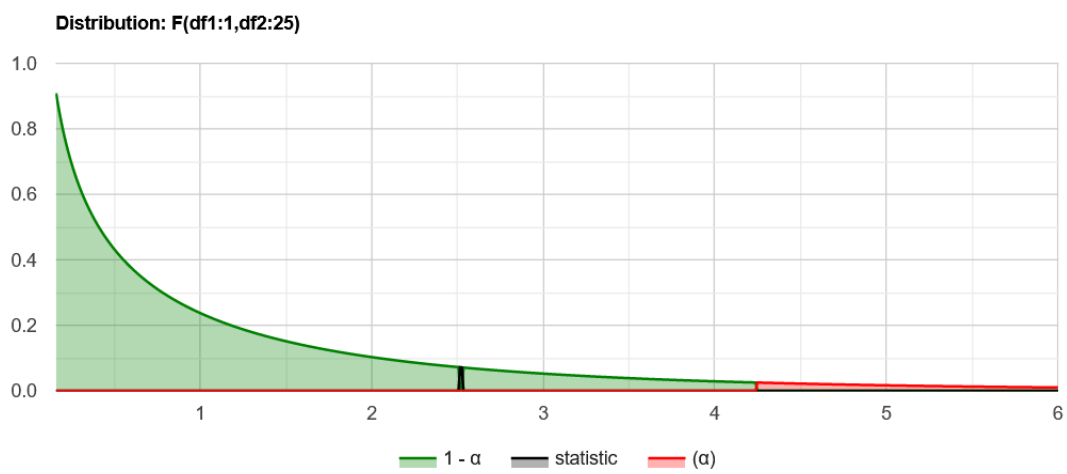


**Figure 3.26: Distribution Graph significance level(α)**

**Residual Normality**

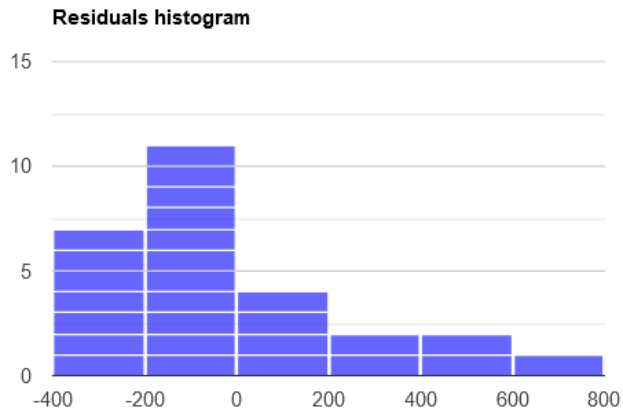


**Figure 3.27: Residual Histogram**

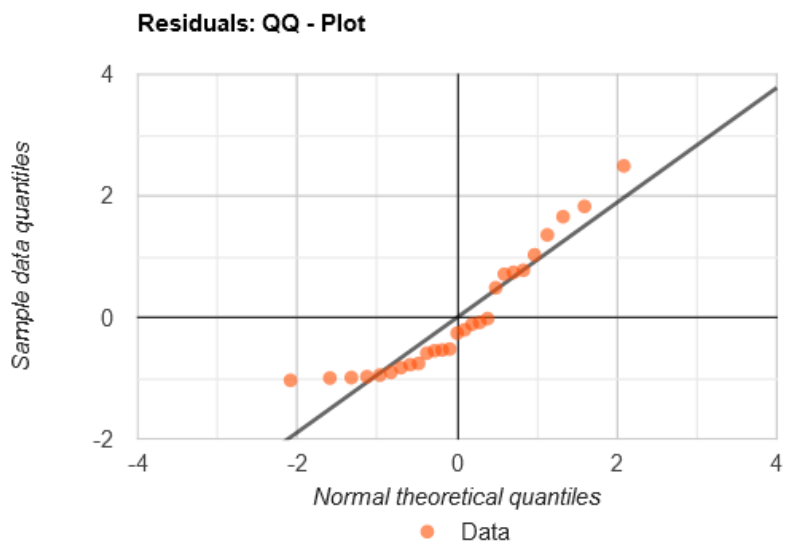


**Figure 3.28: Residual QQ Plot**

**Solution**

$$\hat{Y} = b_0 + b_1 X$$

$$b_1 = \frac{SP_{xy}}{SS_x} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}$$

$$b_1 = \frac{-866.0741}{4.8385} = -178.9957$$

$b_0 = \bar{y} - b_1\bar{x}$
$\bar{x} = 7.1926$
$\bar{y} = 226.6296$
$b_0 = 226.6296+178.9957*7.1926 = 1514.0729$

$$R^2 = \frac{SS_{Regression}}{SS_{total}} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{155023.5468}{1695550.296} = 0.09143$$

The standard deviation of the residuals is:

$$MS_{residual} = S^2_{res} = \frac{\Sigma(y_i - \hat{y})^2}{n - 2}$$

**Residual Outliers**

$S_{res} = \sqrt{MSE} = \sqrt{61621.07} = 248.2359$.

The average of the residuals is always zero.

The thresholds used to calculate the outliers are: $\pm k*S_{res}$.

In this case, the thresholds are $\pm 3*248.2359 = \pm 744.7077$.

Following code is the R code to which produce the similar results:

```
rm(list = ls())
if(!"car" %in% installed.packages()){install.packages("car")}
library(car)
x1 <- c(7,7.4,7.9,6.5,6.2,7.8,7.1,7,8,7.1,6.8,6.9,6.9,7.3,7.3,6.8,8,7.5,7.5,7.5,7.3,7.1,6.9,6.8,7.2,7.3,7.1)
y1 <- c(450,520,95,754,523,55,850,41,55,59,45,452,48,652,458,54,62,45,41,352,158,42,42,56,82,74,54)
model1 = lm(y1~x1)
summary(model1)
```

### 3.13 Lab Test Analysis

Water parametric values play an important role in this study. After conducting the survey and a few interviews water samples has been collected from different sources. This water sample has been tested in the lab and the results are produced in the below given format.

The entire analysis is completely based on the test report, which is used in further chapters 4 and 5 for analysis. Once the lab has completed the testing of water, will receive a report in the below given format Figure 3.5

**TEST REPORT**

**Water Sample Analysis**

**Report Code: W-290119-04** **Issue Date: 01/02/2019**

**Issued To** :

| | | |
|---|---|---|
| **Sample Received On** | : | **29/01/2019** |
| **Sample Drawn By** | : | |
| **Sample Description** | : | **School Kitchen** |
| **Sample Drawn On** | : | **29/01/2019** |
| **Sample Quantity** | : | **1.0 Lt.** |
| **Weather Conditions** | : | **Normal** |
| **Analysis Duration** | : | **29/01/2019 To 01/02/2019** |

**RESULTS**
**Water Test as per IS:10500:2012**

| S.No. | Parameter | Test Method | Results | Units | Acceptable Limit | Permissible Limit |
|---|---|---|---|---|---|---|
| 1. | pH | IS:3025 (Part-11) | 7.21 | - | 6.5 – 8.5 | - |
| 2. | Colour | IS:3025 (Part-4) | <5.0 | Hazen | 5 | 25 |
| 3. | Odour | IS:3025 (Part-5) | Agreeable | - | - | - |
| 4. | Taste | IS:3025 (Part-8) | Agreeable | - | - | - |
| 5. | Turbidity | IS:3025 (Part-10) | <1.0 | NTU | 1 | 10 |
| 6. | Total Hardness (as CaCO3) | IS:3025 (Part-21) | 24.0 | mg/l | 200 | 600 |
| 7. | Chloride (as Cl) | IS:3025 (Part-32) | 18.9 | mg/l | 250 | 1000 |
| 8. | Residual Free Chlorine | IS:3025 (Part-26) | <0.1 | mg/l | 0.2 Max | - |
| 9. | Iron (as Fe) | IS:3025 (Part-52) | <0.01 | mg/l | 0.3 | - |
| 10. | Fluoride (as F) | IS:3025 (Part-60) | <0.1 | mg/l | 1 | 1.5 |
| 11. | Total Dissolved Solid | IS:3025 (Part-16) | 127 | mg/l | 500 | 2000 |
| 12. | Total Coli form | IS:1622 | Absent | MPN/ 100ml | Absent | 10 max |
| 13. | E. Coli | IS: 1622 | Absent | MPN/ 100ml | Absent | Absent |

**Remarks – Based on above tested parameter, water sample meet as per IS-10500 in Acceptable limit**

**Figure 3.29 Sample Water Analysis Report**

It lists the contaminants that were examined, provides their concentrations, and, in certain cases, highlights any problematic chemicals. The units used to measure the amount of

contaminants in the water are a significant aspect of the report. For contaminants like metals , milligrams per liter (mg/l) of water are utilized. One part per million (ppm), or one part pollutant to one million parts water, is also equal to one milligrams per liter.

### 3.14 Methods and Simulation Tools used for analysis

There are several simulation tools that can be used for data analytical models on water quality assurance based on hybrid classification and data mining techniques. Some of the popular simulation tools include:

1. **WEKA**: WEKA (Waikato Environment for Knowledge Analysis) is a popular open-source software tool for data mining and machine learning. It provides a collection of algorithms for data preprocessing, classification, clustering, association rules, and visualization.
2. **RapidMiner**: RapidMiner is another popular open-source data mining tool that offers an integrated environment for data preparation, modeling, evaluation, and deployment.
3. **MATLAB**: MATLAB is a programming language and simulation tool widely used in engineering and scientific research. It provides a range of toolboxes for data analysis, machine learning, and signal processing.
4. **R**: R is a popular open-source statistical programming language and environment for data analysis and graphics. It offers a vast collection of packages for data mining, machine learning, and statistical analysis.
5. **Orange**: Orange is an open-source data visualization and analysis tool that offers a wide range of algorithms for data mining, machine learning, and feature selection.
6. **KNIME**: KNIME is an open-source data analytics platform that provides a visual programming interface for data preprocessing, modeling, and visualization.

These simulation tools can be used to implement and evaluate various data analytical models for water quality assurance based on hybrid classification and data mining techniques

**MATLAB**

MATLAB is a high-level programming language and simulation tool developed by MathWorks. It is widely used in engineering, scientific research, and data analysis. MATLAB provides a range of toolboxes for data analysis, machine learning, and signal processing, making it a popular choice for developing analytical models and algorithms.

Some of the key features of MATLAB include:

1. Matrix manipulation: MATLAB provides powerful matrix manipulation capabilities, making it easy to work with large datasets and perform operations on them.
2. Visualization: MATLAB provides a range of visualization tools for creating 2D and 3D plots, graphs, and animations.
3. Toolboxes: MATLAB provides a wide range of toolboxes for various applications such as statistics, machine learning, signal processing, image processing, and control systems.
4. Interactivity: MATLAB provides an interactive environment for data analysis and modeling, allowing users to modify and experiment with models in real-time.
5. Programming language: MATLAB is a high-level programming language that allows users to write scripts and functions for automating tasks and performing complex computations.

MATLAB is widely used in various industries, including finance, aerospace, automotive, and healthcare, to name a few. It is also used in academic research for developing models and algorithms for various applications.

By using MATLAB toolboxes for data analysis, machine learning, and signal processing, it is possible to develop a data analytical model on water quality assurance based on hybrid classification and data mining techniques. So Matlab is used for parameter simulation in this study.

### 3.15 Proposed Framework for Water Quality Assurance

Water quality is an essential parameter for human health and the environment. With the increase in population and industrialization, the water quality has been degraded globally.

Therefore, it is necessary to monitor the water quality continuously to ensure that it meets it meets the required standards. Based on the research gap here the model is proposed for water quality assurance based on hybrid classification and data mining techniques. And also it can boost the energy at any particular human being without any side effects.

The Proposed model **NavJal** is developed to Replenish the body with bioavailable electrolytes. It provides five essential vitamins (B1, B2, B5, B12 and C) and five essential electrolytes (Calcium, Magnesium, Phosphorus, Potassium and Sodium) added to improve the quality of drinking water. When mixed with water, it's a perfect any-time hydration option for everyone.

So for this the proposed model, dataset includes data collected data from various resources and the framework is being used for data analytical technique to monitor and ensure the water quality.

The first step of proposed model is to collection of data and the segregate the data according to the use and for those following steps can be taken:

1. **Sample collection and preprocessing**: The first step is to collect the water quality data and preprocess it to remove any outliers or missing values.
2. **Sample**: The sample for the water quality data is depend on the specific research question and study design. Generally, a larger sample size is preferable for obtaining more representative and reliable data. A statistical power analysis can help determine the appropriate sample size based on the desired level of precision and confidence
3. **Sample location**: The specific locations where the water quality data is collected from Mumbai, Maharashtra .Some factors to consider when selecting sample locations include the type of water body, the proximity to pollution sources, and the accessibility of the site.
4. **Data collection**: The data collection process will involve taking physical, chemical, and biological measurements of the water quality parameters using appropriate equipment and techniques. The equipment used will depend on the specific parameters being measured, but could include thermometers, pH meters, turbidity meters, dissolved oxygen meters, nutrient analyzers, and microbial sampling

devices. The data is collected at regular intervals and at consistent locations within the sample site to minimize variability.

5. **Data preprocessing**: Once the data is collected, it will be preprocessed to remove any outliers or missing values. Outliers can be identified using statistical methods such as the z-score or interquartile range (IQR), and removed or corrected as needed. Missing values can be imputed using techniques such as mean imputation, regression imputation, or multiple imputation. The cleaned data can be analyzed to answer the research question or hypothesis.
6. **Feature extraction:** The next step is to extract relevant features from the preprocessed data. This can be done using various feature extraction techniques available in MATLAB toolboxes such as Wavelet Toolbox or Signal Processing Toolbox.
7. **Hybrid classification:** The next step is to use a hybrid classification technique such as decision tree or support vector machine (SVM) to classify the water quality data. MATLAB offers several toolboxes for machine learning and classification such as Statistics and Machine Learning Toolbox and Neural Network Toolbox.
8. **Data mining:** The final step is to use data mining techniques such as association rule mining or clustering to identify patterns or trends in the classified data. MATLAB provides several toolboxes for data mining such as Data Mining Toolbox and Bioinformatics Toolbox.

**Proposed Framework: NavJal Model**

Proposed Framework NavJal is developed after going through the Literature Review and analysis of Surveys and questionnaires. This model is divided into 3 modules namely Data Gathering and Data Preprocessing, Data Training and Model Creation i.e. NavJal.

Throughout the process Data mining plays an important role in the NavJal Model .

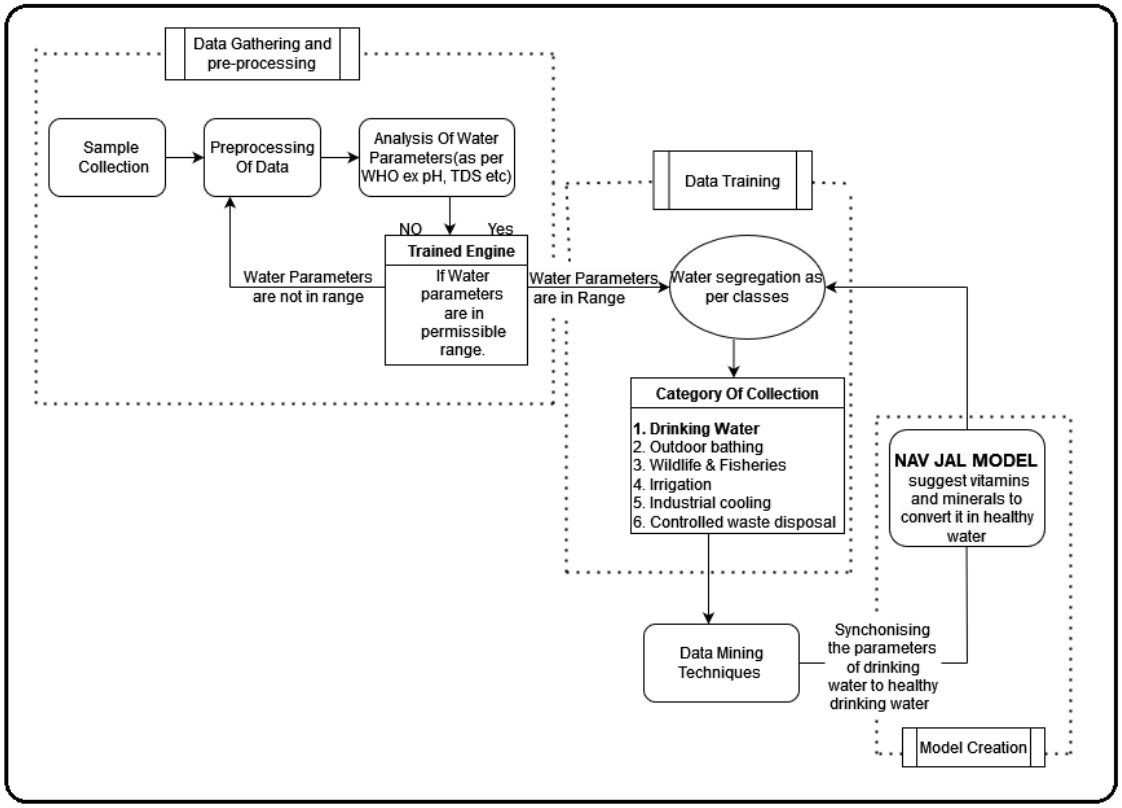


**Figure 3.30: Proposed NavJal model**

Detailed process is described as follow:

**1. First Module: Data Gathering and Pre-processing**

Here the water samples has been collected from different locations. After the collection of sample data is gone under preprocessing to remove the noise from the data. Then this data is analyzed according the World Health Organization (WHO) and CPCB standards for generalize parameters.

**2. Second Module: Training Data set**

Once the data comes under the training dataset , it is checked that the parametric values are in permissible range or not. If the Values are in permissible range it will be forward to further process i.e. Segregation of type of water.(Such as 1. Drinking Water, 2.Outdoor Bathing, 3. Wildlife and Fisheries 4. Irrigation, Industrial Cooling, and Controlled waste disposal).If the parameters are not in range it will exit the process. Then the various analysis techniques are implemented to check the accuracy of water.

**3.Model Creation: NavJal**

After the application of various analysis techniques, the model helps to convert the water into healthy drinking water by suggesting the deficient parametric values to convert the normal drinking water to healthy water.

Overall this model will help to produce new healthy water with purity and energy.

**3.16 Data Accuracy, Parameter selections, Computation Time, and Results**

Here is an overview of each of these factors:

1. **Data accuracy**: The accuracy of the data used for water quality analysis is critical to ensuring the validity of the results. A range of factors, including sampling methods, laboratory analysis procedures, and instrumentation, can affect data accuracy. It is important to use appropriate procedures to ensure the accuracy of the data, including using certified reference materials and replicates and adhering to standard operating procedures.
2. **Parameter selection**: The selection of water quality parameters to measure can have a significant impact on the accuracy and relevance of the results. It is important to select parameters that are relevant and that can be measured accurately and reliably. The selection of parameters may also depend on the availability of data and resources for analysis.
3. **Computation time**: The computation time required for water quality analysis can vary depending on the complexity of the analysis and the amount of data being analyzed. In addition, it is important to consider the practical implications of computation time, such as the availability of resources and the need for timely results.
4. **Results**: The results of water quality analysis can be accurate, reliable, and relevant to the research objective.

# Chapter 4

# Analysis And Interpretation

# Chapter 4 :NavJal Model Implementation

## 4.1 Introduction

Many pollutants have been threatening water quality in recent years. As a result, modeling and prediction of water quality have become critical in the control of water pollution. An advanced Model is created in this work to predict the water quality and water quality classification (WQC). Furthermore, for Water Quality analysis Machine Learning techniques, namely support vector machine (SVM), K-nearest neighbor (K-NN), etc, were applied.

In order to report measurements of water quality to those who are not environmental specialists, the water quality index, also known as the WQI, is utilized. Each country and category has its own unique collection of WQI, which can include public indices, indices that are specialized to consumption, and planning indices.

The National Sanitation Foundation International (NSF), which is based in the United States and is a certification organization, is responsible for the development of the National Sanitation Foundation Water Quality Index (NSFWQI), which is among the most well-known WQIs.

The National Science Foundation Water Quality Index (NSF-WQI) produces values from nine water quality characteristics with varying weights assigned to each parameter. (NSF Report)

Here the proposed model, has used eighteen relevant parameters, and they were tested under this model. According to the results, the suggested models can accurately forecast Water Quality and suggest the deficiency of water quality based on greater resilience.

## 4.2 System Architecture

Water Quality Prediction Analysis was implemented on seven different models for the purpose of evaluation.

This system architecture elaborates the overall process of Data handling.

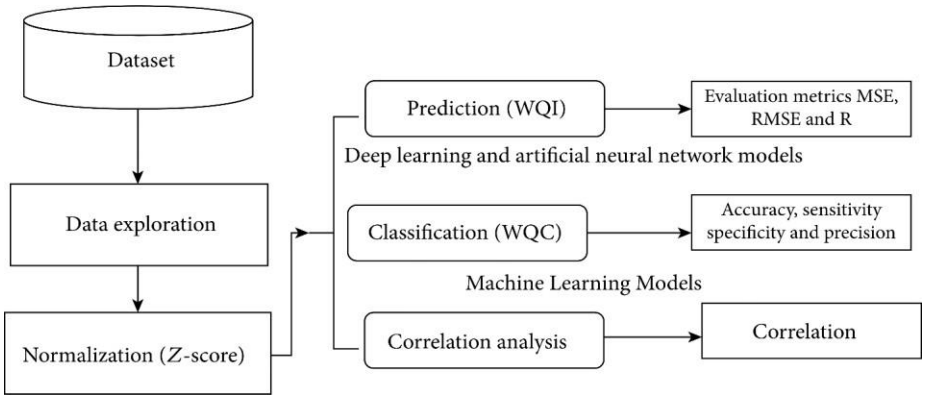


**Figure 4.1 System Architecture**

- As shown in the figure 4.1 at very first step Dataset has to be considered for Data Exploration.
- In normalization Z score is the value to understand how far the data point is from the mean.
- Prediction pointer will help the data analysis and will proceed to calculate the WQI to predict the Water Quality. It may help with the evaluation metrics of RMSE and MSE
- Classification helps to maintain the accuracy, sensitivity specificity , and precision
- The correlation analysis help to correlate or to maintain the statistical relationship.

**Parameter simulations on MATLAB for water quality,:**

i. Define the parameters: Determine the water quality parameters that you want to simulate, such as pH, dissolved oxygen, turbidity, and conductivity.

ii. Set up the simulation: Create a MATLAB script that sets up the simulation, including defining the time period for the simulation, the initial values for the parameters, and any equations or models that will be used to simulate the parameters.

iii. Create a table: Define a table in MATLAB to store the simulation results. The table should include columns for the time period and each of the water quality parameters being simulated.

iv. Run the simulation: Run the MATLAB script to perform the simulation and populate the table with the simulation results.

v. Visualize the results: Use MATLAB's plotting and visualization tools to create graphs and charts that show the simulation results over time for each of the water quality parameters being simulated.

**MATLAB code to create a table parameter simulation for water quality:**

```
% Define simulation parameters

time_period = 24; % hours

pH_initial = 7.0;

dissolved_oxygen_initial = 5.0;

turbidity_initial = 10.0;

conductivity_initial = 500.0;

% Define simulation equations

pH = @(t, pH_initial) pH_initial + 0.1 * t;

dissolved_oxygen = @(t, dissolved_oxygen_initial) dissolved_oxygen_initial - 0.2 * t;

turbidity = @(t, turbidity_initial) turbidity_initial + 0.05 * t;

conductivity = @(t, conductivity_initial) conductivity_initial - 0.1 * t;

% Create table to store simulation results

sim_table = table('Size', [time_period, 5], 'VariableTypes', {'double', 'double', 'double', 'double',
'double'}, 'VariableNames', {'Time', 'pH', 'DissolvedOxygen', 'Turbidity', 'Conductivity'});

% Populate table with simulation results

for t = 1:time_period

    sim_table(t, 1) = t;
```

```
    sim_table(t, 2) = pH(t, pH_initial);
    sim_table(t, 3) = dissolved_oxygen(t, dissolved_oxygen_initial);
    sim_table(t, 4) = turbidity(t, turbidity_initial);
    sim_table(t, 5) = conductivity(t, conductivity_initial);
end
% Visualize simulation results
plot(sim_table.Time, sim_table.pH);
hold on;
plot(sim_table.Time, sim_table.DissolvedOxygen);
plot(sim_table.Time, sim_table.Turbidity);
plot(sim_table.Time, sim_table.Conductivity);
legend('pH', 'Dissolved Oxygen', 'Turbidity', 'Conductivity');
xlabel('Time (hours)');
ylabel('Parameter value');
title('Water Quality Parameter Simulation');
```

This MATLAB code simulates the changes in pH, dissolved oxygen, turbidity, and conductivity over a 24-hour period using predefined equations. The simulation results are stored in a table, and a plot is created to visualize the changes in each parameter over time.

### 4.3 Evaluation of the proposed framework

Here for the experiment total of eighteen parameters have been considered. The experiment has been implemented on Python 3.x series with various scientific libraries.

- **Dataset**

The dataset used in this study is collected from certain different locations of Maharashtra. The dataset has total nine significant parameters, namely, pH value, Hardness, Solids(TDS), Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity.

- **Data Preprocessing**

The Pre-processing phase is very important in data analysis to improve the data quality. In this phase, the WQI has been calculated from the most significant parameters of the dataset. Then, water samples have been classified on the basis of the WQI values. For obtaining superior accuracy, the z-score method has been used as a data normalization technique.

The Sample dataset for a few parameters has been mentioned below.

**Table 4.1. Sample Dataset 1**

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | NaN | 204.890455 | 20791.318981 | 7.300212 | 368.516441 | 564.308654 | 10.379783 | 86.990970 | 2.963135 | 0 |
| 1 | 3.716080 | 129.422921 | 18630.057858 | 6.635246 | NaN | 592.885359 | 15.180013 | 56.329076 | 4.500656 | 0 |
| 2 | 8.099124 | 224.236259 | 19909.541732 | 9.275884 | NaN | 418.606213 | 16.868637 | 66.420093 | 3.055934 | 0 |
| 3 | 8.316766 | 214.373394 | 22018.417441 | 8.059332 | 356.886136 | 363.266516 | 18.436524 | 100.341674 | 4.628771 | 0 |
| 4 | 9.092223 | 181.101509 | 17978.986339 | 6.546600 | 310.135738 | 398.410813 | 11.558279 | 31.997993 | 4.075075 | 0 |

**Table 4.2 Statistical Algorithm Implementation**

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 2785.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 2495.000000 | 3276.000000 | 3276.000000 | 3114.000000 | 3276.000000 | 3276.000000 |
| mean | 7.080795 | 196.369496 | 22014.092526 | 7.122277 | 333.775777 | 426.205111 | 14.284970 | 66.396293 | 3.966786 | 0.390110 |
| std | 1.594320 | 32.879761 | 8768.570828 | 1.583085 | 41.416840 | 80.824064 | 3.308162 | 16.175008 | 0.780382 | 0.487849 |
| min | 0.000000 | 47.432000 | 320.942611 | 0.352000 | 129.000000 | 181.483754 | 2.200000 | 0.738000 | 1.450000 | 0.000000 |
| 25% | 6.093092 | 176.850538 | 15666.690297 | 6.127421 | 307.699498 | 365.734414 | 12.065801 | 55.844536 | 3.439711 | 0.000000 |
| 50% | 7.036752 | 196.967627 | 20927.833607 | 7.130299 | 333.073546 | 421.884968 | 14.218338 | 66.622485 | 3.955028 | 0.000000 |
| 75% | 8.062066 | 216.667456 | 27332.762127 | 8.114887 | 359.950170 | 481.792304 | 16.557652 | 77.337473 | 4.500320 | 1.000000 |
| max | 14.000000 | 323.124000 | 61227.196008 | 13.127000 | 481.030642 | 753.342620 | 28.300000 | 124.000000 | 6.739000 | 1.000000 |

Once the water parameters have been validated it can be forwarded to further process to check the accuracy. And this accuracy is measured by implementing the following list of algorithms.

The List of the Algorithm used for the result comparison is as follows:

1. Logistic Regression
2. Decision Tree
3. Random Forest

4. XGBoost
5. KNeighbours
6. SVM
7. AdaBoost

A statistical technique known as regression analysis is used to simulate the relationship between a dependent and independent variables using one or more independent variables. So with the help of sample data Maschine learning algorithm creates a mathematical model with the aid of Historical data or training data that helps in making predictions without being explicitly Programmed.

### 4.3.1 Logistic regression

Predictive analytics and classification are both done using this technique. Logistic regression calculates the likelihood that an occurrence will occur. A logit transformation is given to the odds in logistic regression, which is the probability of success divided by the probability of failure.

The sample code with the outcome is given below

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, classification_repo
rt

# Creating model object
model_lg = LogisticRegression(max_iter=120,random_state=0, n_jobs=20)

# Training Model
model_lg.fit(X_train, y_train)

LogisticRegression(max_iter=120, n_jobs=20, random_state=0)

# Making Prediction
pred_lg = model_lg.predict(X_test)

# Calculating Accuracy Score
lg = accuracy_score(y_test, pred_lg)
print(lg)
```

```
0.6284658040665434
```

```
print(classification_report(y_test,pred_lg))

              precision    recall  f1-score   support

           0       0.63      1.00      0.77       680
           1       0.00      0.00      0.00       402

    accuracy                           0.63      1082
   macro avg       0.31      0.50      0.39      1082
weighted avg       0.39      0.63      0.49      1082

# confusion Maxtrix
cm1 = confusion_matrix(y_test, pred_lg)
sns.heatmap(cm1/np.sum(cm1), annot = True, fmt=  '0.2%', cmap = 'Reds')
<AxesSubplot:>
```

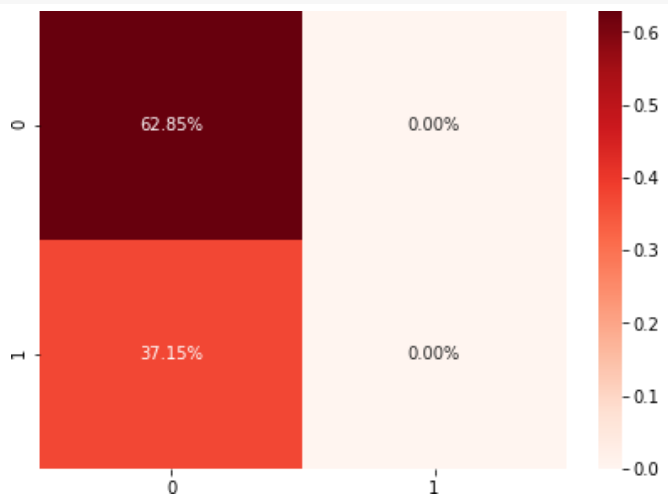


**Figure: 4.2 Logistic Regression**

Using logistic regression on the data set it produces the accuracy 0.6284658040665434

### 4.3.2 Decision Tree Classifier

While Decision Tree is a supervised learning technique that can be used to handle both classification and regression problems, it is most frequently applied to classification problems.

The Decision Node and the Leaf Node are the two elements that make up a decision tree. A choice node is used to make decisions and has many branches, whereas a Leaf node is the outcome of that choice and has no other branches.

```
from sklearn.tree import DecisionTreeClassifier
linkcode
# Creating model object
model_dt = DecisionTreeClassifier( max_depth=4, random_state=42)
```

```
# Training Model
model_dt.fit(X_train,y_train)
```

```
DecisionTreeClassifier(max_depth=4, random_state=42)
# Making Prediction
pred_dt = model_dt.predict(X_test)
# Calculating Accuracy Score
dt = accuracy_score(y_test, pred_dt)
print(dt)
```

```
0.6451016635859519
```

```
print(classification_report(y_test,pred_dt))
   precision    recall  f1-score   support

           0       0.66      0.90      0.76       680
           1       0.56      0.22      0.32       402

    accuracy                           0.65      1082
   macro avg       0.61      0.56      0.54      1082
weighted avg       0.62      0.65      0.60      1082
```

```
# confusion Maxtrix
cm2 = confusion_matrix(y_test, pred_dt)
sns.heatmap(cm2/np.sum(cm2), annot = True, fmt=  '0.2%', cmap = 'Reds')
<AxesSubplot:>
```

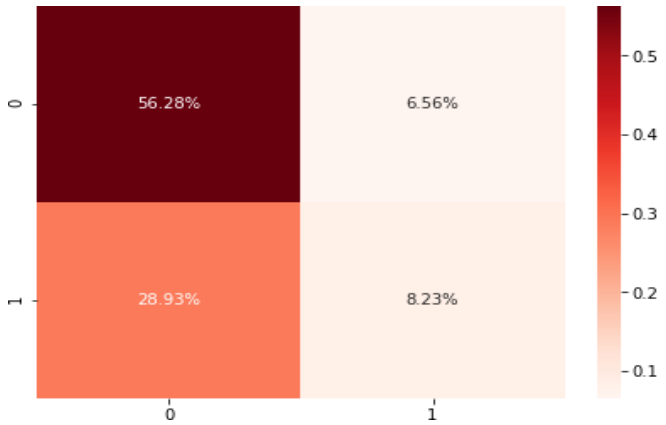


**Figure 4.3 : Decision tree**

Using Decision Tree on the dataset it produces the accuracy `0.6451016635859519`

### 4.3.3 Random Forest

Random forest, a supervised machine learning method, is a well-liked strategy for problems with classification and regression. It uses a variety of samples to construct decision trees, relying on their consensus for classification and average for regression.

One of the most important characteristics of the Random Forest Algorithm is its ability to handle data sets with continuous factors, as in regression, and categorical variables, as in classification. It functions better for classification and regression jobs.

```
from sklearn.ensemble import RandomForestClassifier
# Creating model object
model_rf = RandomForestClassifier(n_estimators=300,min_samples_leaf=0.16, random_
state=42)
```

```
# Training Model
model_rf.fit(X_train, y_train)
```

```
RandomForestClassifier(min_samples_leaf=0.16, n_estimators=300, random_state=42)
```

```
# Making Prediction
pred_rf = model_rf.predict(X_test)
```

```
# Calculating Accuracy Score
rf = accuracy_score(y_test, pred_rf)
print(rf)
```

```
0.6284658040665434
```

```
print(classification_report(y_test,pred_rf))
              precision    recall  f1-score   support

           0       0.63      1.00      0.77       680
           1       0.00      0.00      0.00       402

    accuracy                           0.63      1082
   macro avg       0.31      0.50      0.39      1082
weighted avg       0.39      0.63      0.49      1082
```

```
# confusion Maxtrix
cm3 = confusion_matrix(y_test, pred_rf)
sns.heatmap(cm3/np.sum(cm3), annot = True, fmt=  '0.2%', cmap = 'Reds')
<AxesSubplot:>
```

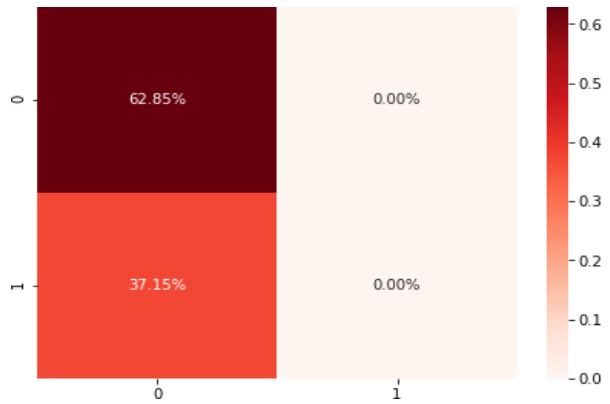


**Figure 4.4: Random forest**

Using Random forest on the dataset it produces the accuracy `0.6284658040665434`

### 4.3.4 XGBoost Classifier

A distributed, scalable gradient-boosted decision tree (GBDT) machine learning system is called Extreme Gradient Boosting (XGBoost). It provides a parallel tree boosting and is the best machine-learning tool for regression, classification, and ranking problems.

Gradient Boosting Decision Trees (GBDT), a decision tree ensemble learning algorithm similar to the random forest, are used for classification and regression. Ensemble learning algorithms combine various machine learning algorithms to produce a more accurate model.

```
from xgboost import XGBClassifier
# Creating model object
model_xgb = XGBClassifier(max_depth= 8, n_estimators= 125, random_state= 0,
learning_rate= 0.03, n_jobs=5)
# Training Model
model_xgb.fit(X_train, y_train)
[08:10:23] WARNING: ...../src/learner.cc:1095: Starting in XGBoost 1.3.0, the d
efault evaluation metric used with the objective 'binary:logistic' was chang
ed from 'error' to 'logloss'. Explicitly set eval_metric if you'd like to re
store the old behavior.

GBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
              importance_type='gain', interaction_constraints='',
              learning_rate=0.03, max_delta_step=0, max_depth=8,
              min_child_weight=1, missing=nan, monotone_constraints='()',
              n_estimators=125, n_jobs=5, num_parallel_tree=1, random_state0
,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
              tree_method='exact', validate_parameters=1, verbosity=None)
```

```
# Making Prediction
pred_xgb = model_xgb.predict(X_test)
# Calculating Accuracy Score
xgb = accuracy_score(y_test, pred_xgb)
print(xgb)
0.6709796672828097
print(classification_report(y_test,pred_xgb))
              precision    recall  f1-score   support

           0       0.68      0.89      0.77       680
           1       0.61      0.31      0.41       402

    accuracy                           0.67      1082
   macro avg       0.65      0.60      0.59      1082
weighted avg       0.66      0.67      0.64      1082
# confusion Maxtrix
    cm4 = confusion_matrix(y_test, pred_xgb)
sns.heatmap(cm4/np.sum(cm4), annot = True, fmt=  '0.2%', cmap = 'Reds')
<AxesSubplot:>
```

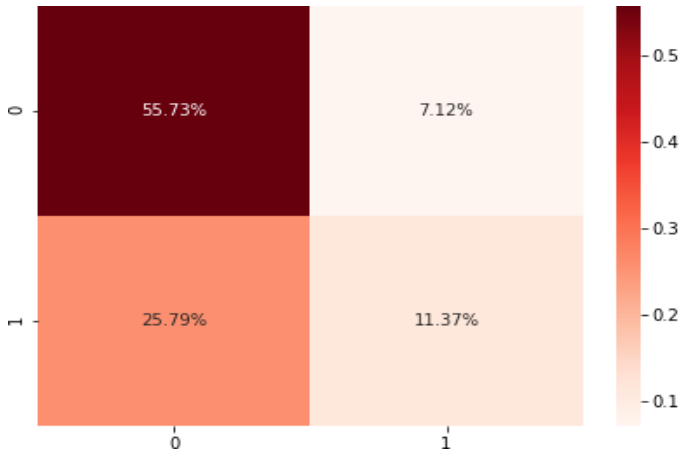


**Figure 4.5 Using XGBoost Classifier**

Using XG Boost Classifier on the dataset it produces the accuracy `0.6709796672828097`

### 4.3.5 K nearest Neighbour

The k-nearest neighbours algorithm is a non-parametric supervised learning technique in statistics that was created in 1951 by Evelyn Fix and Joseph Hodges and subsequently improved by Thomas Cover. Regression and categorization are two uses for it. The input in both situations comprises of a data set's k closest training examples.

```
from sklearn.neighbors import KNeighborsClassifier
# Creating model object
model_kn = KNeighborsClassifier(n_neighbors=9, leaf_size=20)
# Training Model
model_kn.fit(X_train, y_train)

KNeighborsClassifier(leaf_size=20, n_neighbors=9)

# Making Prediction
pred_kn = model_kn.predict(X_test)

# Calculating Accuracy Score
kn = accuracy_score(y_test, pred_kn)
print(kn)

0.6534195933456562
print(classification_report(y_test,pred_kn))

              precision    recall  f1-score   support

           0       0.69      0.82      0.75       680
           1       0.55      0.37      0.44       402

    accuracy                           0.65      1082
   macro avg       0.62      0.60      0.59      1082
weighted avg       0.64      0.65      0.63      1082

# confusion Maxtrix
cm5 = confusion_matrix(y_test, pred_kn)
sns.heatmap(cm5/np.sum(cm5), annot = True, fmt=  '0.2%', cmap = 'Reds')

<AxesSubplot:>
```

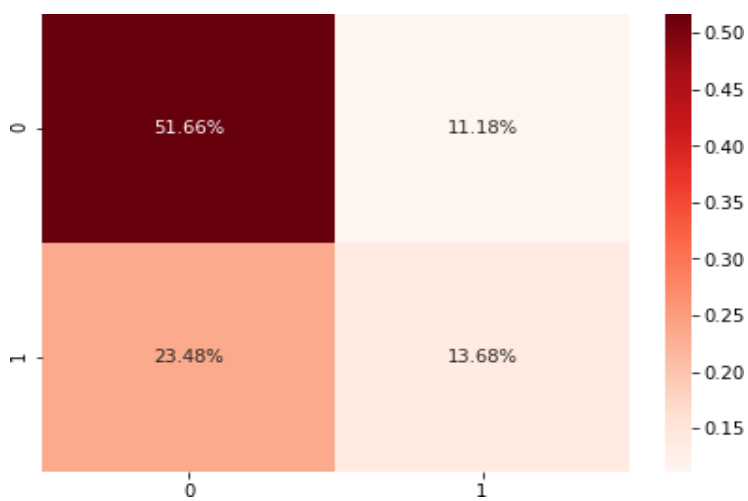


**Figure 4.6 Using K nearest Neighbour Classifier**

Using K nearest neighbour Classifier on the dataset it produces the accuracy 0.65341959 33456562

### 4.3.6 Support Vector Machine

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression issues. However, it is mainly employed in Machine Learning Classification issues.

The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into groups, allowing us to quickly classify new data points in the future.

```
from sklearn.svm import SVC, LinearSVC
model_svm = SVC(kernel='rbf', random_state = 42)
model_svm.fit(X_train, y_train)
```

```
SVC(random_state=42)
```

```
# Making Prediction
pred_svm = model_svm.predict(X_test)
```

```
# Calculating Accuracy Score
sv = accuracy_score(y_test, pred_svm)
print(sv)
```

```
0.6885397412199631
```

```
print(classification_report(y_test,pred_kn))
```

```
              precision    recall  f1-score   support

           0       0.69      0.82      0.75       680
           1       0.55      0.37      0.44       402

    accuracy                           0.65      1082
   macro avg       0.62      0.60      0.59      1082
weighted avg       0.64      0.65      0.63      1082
```

```
# confusion Maxtrix
cm6 = confusion_matrix(y_test, pred_svm)
sns.heatmap(cm6/np.sum(cm6), annot = True, fmt=  '0.2%', cmap = 'Reds')
```

```
<AxesSubplot:>
```

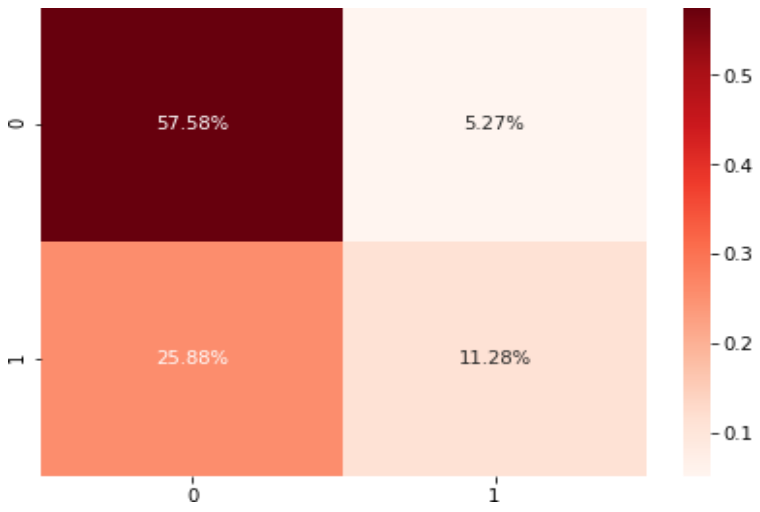


**Figure 4.7 Using SVM**

Using SVM Classifier on the dataset it produces the accuracy `0.6885397412199631`

### 4.3.7 Using AdaBoost Classifier

AdaBoost, also known as Adaptive Boosting, is a machine learning method that is applied as an ensemble. Decision trees with one level, or Decision trees with only one split, are the most popular predictor used with AdaBoost. Another name for these plants is Decision Stumps.

This algorithm creates a model while giving each data element the same weight. Then, it gives points that were incorrectly classified as greater weights.

```
from sklearn.ensemble import AdaBoostClassifier

model_ada = AdaBoostClassifier(learning_rate= 0.002,n_estimators= 205,random_stat
e=42)
model_ada.fit(X_train, y_train)
AdaBoostClassifier(learning_rate=0.002, n_estimators=205, random_state=42)
#Making Prediction
pred_ada = model_ada.predict(X_test)

# Calculating Accuracy Score
ada = accuracy_score(y_test, pred_ada)
print(ada)

0.634011090573013
```

```
print(classification_report(y_test,pred_ada))
```

```
              precision    recall  f1-score   support

           0       0.63      0.99      0.77       680
           1       0.62      0.04      0.07       402

    accuracy                           0.63      1082
   macro avg       0.62      0.51      0.42      1082
weighted avg       0.63      0.63      0.51      1082
```

```
# confusion Maxtrix
cm7 = confusion_matrix(y_test, pred_ada)
sns.heatmap(cm7/np.sum(cm7), annot = True, fmt=  '0.2%', cmap = 'Reds')
```

```
<AxesSubplot:>
```

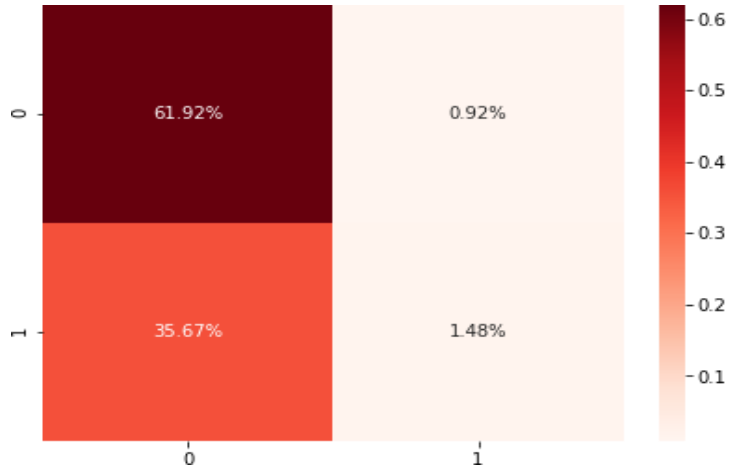


**Figure 4.8 Adaboost Implementation**

Using AdaBoost Classifier on the dataset it produces the accuracy 0.634011090573013

**Observations:**

After the implementation of all the seven algorithms is clear that SVM is providing the best accuracy of the dataset.

```
models = pd.DataFrame({
    'Model':['Logistic Regression', 'Decision Tree', 'Random Forest', 'XGBoost',
'KNeighbours', 'SVM', 'AdaBoost'],
    'Accuracy_score':[lg, dt, rf, xgb, kn, sv, ada]
})
models
sns.barplot(x='Accuracy_score', y='Model', data=models)

models.sort_values(by='Accuracy_score', ascending=False)
```

**Table 4.3 Accuracy Matrix**

| Sr. No | Model | Accuracy_score |
|---|---|---|
| 1 | SVM | 0.688540 |
| 2 | XGBoost | 0.670980 |
| 3 | KNeighbours | 0.653420 |
| 4 | Decision Tree | 0.645102 |
| 5 | AdaBoost | 0.634011 |
| 6 | Logistic Regression | 0.628466 |
| 7 | Random Forest | 0.628466 |

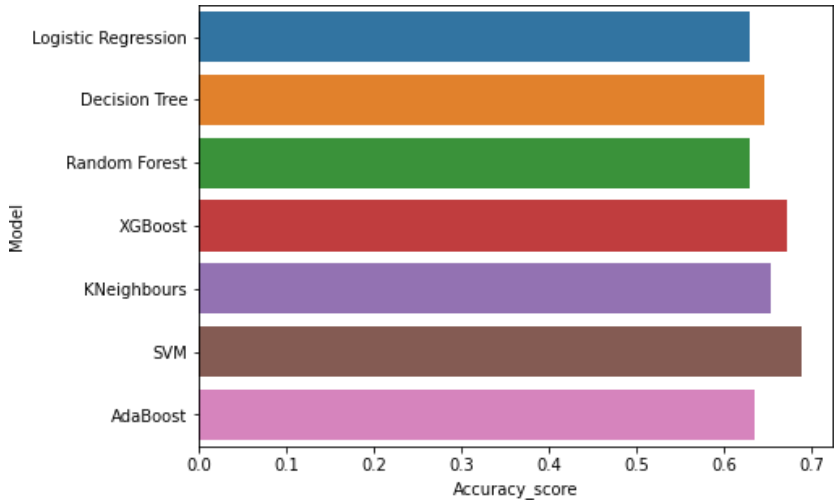


**Figure 4.9: Comparative Analysis of all techniques according to Accuracy Score**

Finally, this implementation of all seven algorithms' comparative analysis proves that SVM achieved the highest accuracy.

### 4.4 Hypothetical Analysis

In hypothesis testing, a hypothesis is formulated and tested against a sample of data. The goal is to determine whether the evidence supports the hypothesis or not.

To perform hypothesis testing on this study, we would need access to the data used in the study and the statistical methods applied. It would also be important to consider the study's limitations and potential sources of bias.

The study likely involves analyzing a dataset related to water quality parameters, such as pH, turbidity, dissolved oxygen, etc., and applying data mining techniques.

This type of study is relevant in the field of water quality management and can assist in identifying potential issues with water quality, such as contamination or other anomalies that may require further investigation. Additionally, by using data analytical models to monitor water quality parameters, it is possible to identify trends and patterns in water quality data that may help to improve overall water quality management strategies.

An experiment can be classified into different factors:

1. **H1: Sample collection and water parameter compilation**
    a. **Introduction:**

For preprocessing, collected various water parameters based on the Maharashtra Pollution Control Board policy and compiled them as per the requirement of the Proposed Model. And also few samples have been tested by the lab.

There is no restriction on the collection and compilation of quality parameters. As previous researchers have their limited quality parameters to identify the water quality.

    b. **Objective:**

Identify and study a set of water parameters to make an inventory to check water quality.

    c. **Related Study**

Few of the government bodies like Central Pollution Control Board provide data sets for analysis in the raw format. They provide different kind of resources.

Such as:

For measurement of temperature: Thermometer

For analysis of pH: pH meter is needed

For Electrical conductivity: Conductivity meter

For all measurements there is no need of different instruments, there is a single instrument with different probes which measure all three parameters.

For sampling few things has been considered:

I. Date and time of collection
II. Sampling location
III. Parameters
IV. Weather conditions

Here in this study total of 18 parameters have been considered including heavy metals. And the parameters which are within a range that is only considered for further process.

**Table 4.4 Water Quality Parameters**

**[Source: Drinking Water-Specification**

**http://cgwb.gov.in/documents/wq-standards.pdf]**

| Sr.no | Substances | Requirement | In the lack of alternative sources, The permissible Limit |
|---|---|---|---|
| 1 | pH Value | 6.5 to 8.5 | No Relaxation |
| 2 | Turbidity ( NTU, Max) | 5 | 10 |
| 3 | Total Hardness (as $CaCo_3$) mg/lit.,Max | 300 | 600 |
| 4 | Dissolved solids mg/lit,Max | 500 | 2000 |
| 5 | Conductivity | less than 1000 | |
| 6 | Chloramin | 1 | 4 |
| 7 | Sulfate (as $SO_4$) mg/lit,Max | 200 | 400 |
| 8 | Organic Carbon | 5 | 5000 |
| 9 | Trihalomthane | upto 80 | |

| | | | |
|---|---|---|---|
| 10 | Calcium (as Ca) mg/lit,Max | 75 | 200 |
| 11 | Fluoride (as F) mg/lit,Max | 1.9 | 1.5 |
| 12 | Mineral Oil mg/lit,Max | 0.01 | 0.03 |
| 13 | Iron (as Fe) mg/lit,Max | 0.3 | 1 |
| 14 | Cadmium (as Cd)mg/lit,Max | 0.01 | No Relaxation |
| 15 | Copper (as Cu) mg/lit,Max | 0.05 | 1.5 |
| 16 | Zinc (as Zn) mg/lit,Max | 5 | 15 |
| 17 | Manganese (as Mn)mg/lit,Max | 0.1 | 0.3 |
| 18 | Lead (as Pb) mg/lit,Max | 0.05 | No Relaxation |

**d. Conclusion:**

After collecting a huge dataset interpretation of data will be easier to get a better monitoring system. With this study, relevant background information can also be analyzed for future aspects.

**2. H2 : Impact of external objects on water quality**

**a. Introduction:**

There are various objects which may affect directly or indirectly to the water quality. Especially the storage patterns which are generally used that affect the parametric values.

**b. Objective**:

To study the impact of external resources on water quality.

**c. Related Study:**

Here in this study five storage patterns (Stainless steel, mud matka, copper, plastic, tap water) have been considered and analyzed it based on different parameters.

Following figures shows few samples which was collected from Local areas.

**Figure 4.10 Water Collection in Steel Vessel**

**Figure 4.11 Water Collection in Plastic**

**Figure 4.12 Water Collection in Copper**

**Figure 4.13 Water Collection in Mud Matka**

**Figure 4.14 Local Brands of Drinking Water Available in Mumbai Region**

**Figure 4.15 Mumbai Suburbs Location wise Water Samples**

Below given table shows eighteen water parametric values of five different storage patterns, and how it affects the water quality.

**Table 4.5 Water Parametric Values**

| Sr. No | Parameters | Stainless steel | Mud Matka | Plastic | Copper | Tap Water |
|---|---|---|---|---|---|---|
| 1 | pH Value | 7.7 | 7.7 | 7.8 | 7.4 | 7.8 |
| 2 | Turbidity ( NTU, Max) | 6.2 | 6.5 | 5.6 | 7.9 | 4.9 |
| 3 | Total Hardness (as $CaCo_3$) mg/lit., Max | 482 | 412 | 561 | 355 | 256 |
| 4 | Dissolved solids mg/lit, Max | 45 | 47 | 52 | 58 | 55 |
| 5 | Conductivity | 580 | 560 | 400 | 890 | 700 |
| 6 | Chloramine | 1.3 | 2.1 | 2.3 | 3.1 | 4.9 |
| 7 | Sulfate (as $SO_4$) mg/lit, Max | 222 | 423 | 425 | 395 | 256 |
| 8 | Organic Carbon | 120 | 1520 | 890 | 2556 | 2541 |
| 9 | Trihalomethane | 65 | 58 | 52 | 74 | 62 |
| 10 | Calcium (as Ca) mg/lit, Max | 102 | 153 | 112 | 159 | 60 |
| 11 | Fluoride (as F) mg/lit, Max | 1.8 | 1.7 | 1.5 | 1.8 | 1.84 |
| 12 | Mineral Oil mg/lit, Max | 0.02 | 0.01 | 0 | 0.01 | 0.02 |
| 13 | Iron (as Fe) mg/lit, Max | 0.2 | 0.2 | 0.3 | 0.3 | 0.2 |
| 14 | Cadmium (as Cd)mg/lit, Max | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| 15 | Copper (as Cu) mg/lit, Max | 0.05 | 0.09 | 0.3 | 1.3 | 0.02 |
| 16 | Zinc (as Zn) mg/lit, Max | 6 | 6 | 8 | 5 | 6 |
| 17 | Manganese (as Mn)mg/lit, Max | 0.1 | 0.2 | 0.1 | 0.2 | 0.1 |
| 18 | Lead (as Pb) mg/lit, Max | 0.05 | 0.05 | 0.02 | 0.05 | 0.02 |

As the above table shows different parametric values after the same water sample has been stored for a specific amount of time it affects the water quality. After the implementation of linear regression, it showed a better impact of copper vessels on the quality of water.

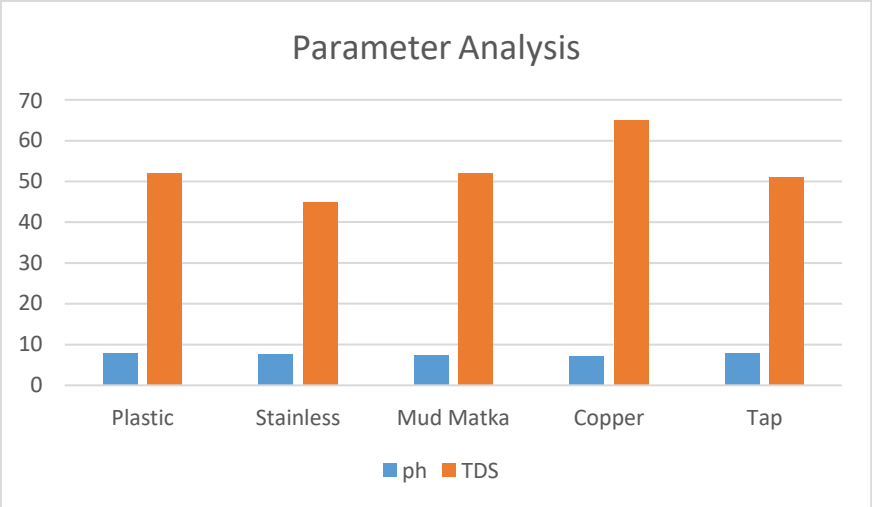


**Figure 4.16 Analysis of Two Parameters (Ph And TDS) Based on Storage Patterns**

**d. Conclusion:**

The results obtained from the linear regression technique show that the storage of water in any particular type of vessel affects the quality of water depending on the storage pattern.

**H3 : Implementation of Various mining techniques to verify the quality of water on the overall dataset**

**a. Introduction:**

Here in this study, a total of eighteen parameters have been considered including heavy metals for overall analysis. The proposed model considered seven algorithms for this testing.

i. Logistic Regression
ii. Decision Tree
iii. Random Forest
iv. XGBoost
v. KNeighbours
vi. SVM
vii. AdaBoost

**b. Objective:**

To Design a model which carries out Data Mining Techniques on water parameters to verify the water quality.

### c. Implementation:

Here above mentioned water parameters have been tested under the stated seven algorithms.

From these algorithms, SVM shows better accuracy results on every storage pattern. Below given table shows the accuracy level of Stainless Steel stored water samples after and before the implementation of NavJal Model.

**Table 4.6 Accuracy Matrix of Water Sample Stored in Stainless Steel**

| Algorithms Used | Accuracy Score (Before) | Accuracy Score (After the Implementation of Nav Jal Model) |
|---|---|---|
| Logistic Regression | 0.624212858 | 0.62456585 |
| Decision Tree | 0.640520193 | 0.64565629 |
| Random Forest | 0.647975739 | 0.66458978 |
| KNN | 0.642837533 | 0.65124589 |
| SVM | 0.644201352 | 0.69555456 |
| XGBoost | 0.621615605 | 0.63614545 |
| AdaBoost | 0.665775337 | 0.67212312 |

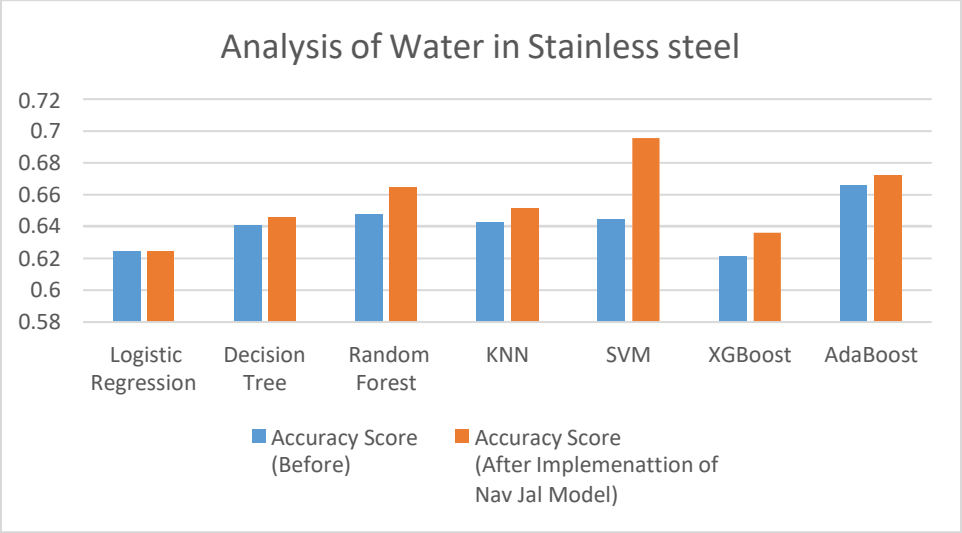


**Figure 4.17 Comparative Analysis of Water Sample Stored in Stainless Steel**

**Table 4.7 Accuracy Matrix of water sample stored in Mud Matka**

| Algorithms Used | Accuracy Score (Before) | Accuracy Score (After the Implementation of Nav Jal Model) |
|---|---|---|
| Logistic Regression | 0.685344528 | 0.693457878 |
| Decision Tree | 0.671561071 | 0.683454457 |
| Random Forest | 0.620166095 | 0.620124545 |
| KNN | 0.681684347 | 0.652256565 |
| SVM | 0.652325853 | 0.712894552 |
| XGBoost | 0.672502433 | 0.698748754 |
| AdaBoost | 0.64961248 | 0.658977415 |

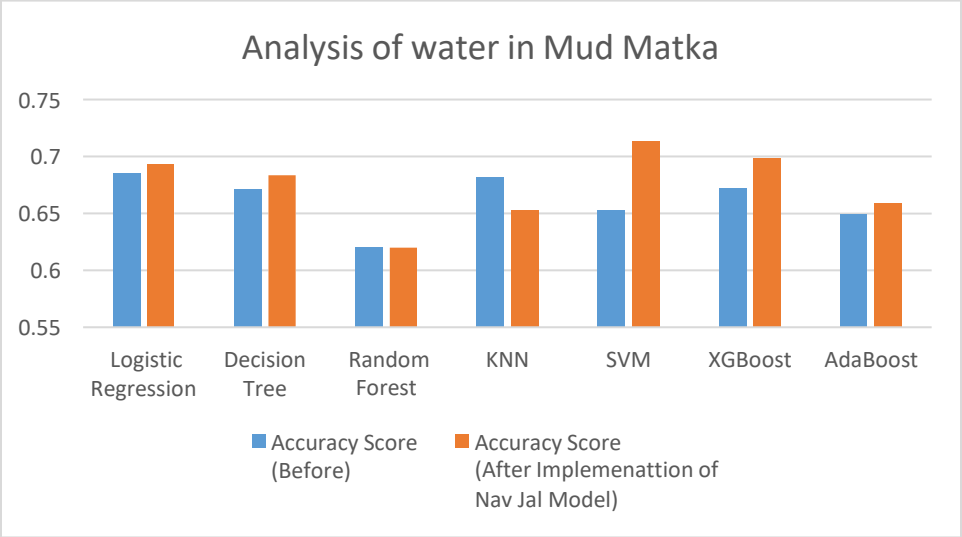


**Figure 4.18 Comparative analysis of water sample stored in Mud Matka**

**Table 4.8 Accuracy Matrix of water sample stored in Plastic**

| Algorithms Used | Accuracy Score (Before) | Accuracy Score (After the Implementation of Nav Jal Model) |
|---|---|---|
| Logistic Regression | 0.638169426 | 0.64748595 |
| Decision Tree | 0.673691636 | 0.67369156 |
| Random Forest | 0.616450435 | 0.61650548 |
| KNN | 0.610019331 | 0.620214578 |
| SVM | 0.669778487 | 0.689998989 |
| XGBoost | 0.626545449 | 0.62859587 |
| AdaBoost | 0.652137379 | 0.684512326 |

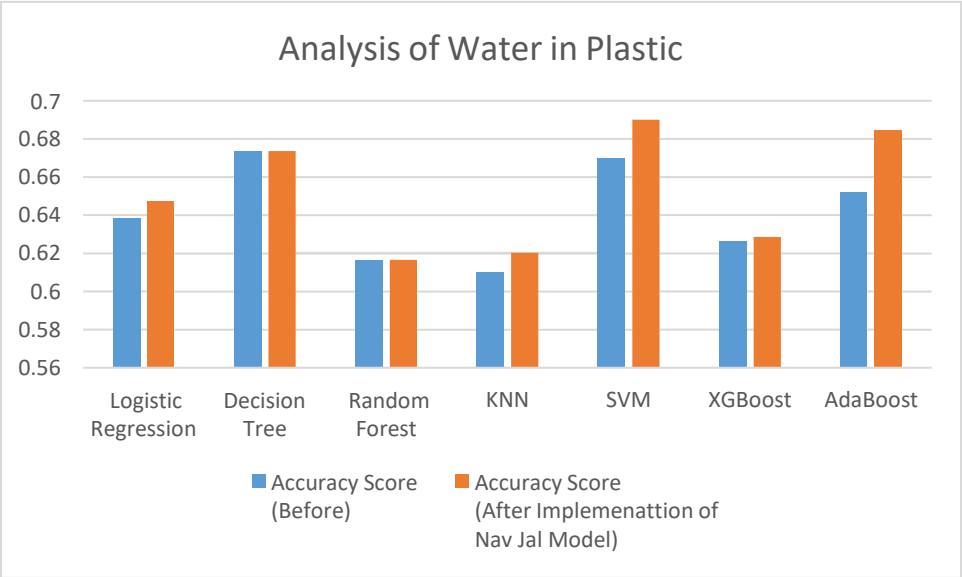


**Figure 4.19 Comparative Analysis of Water Sample Stored in Plastic**

**Table 4.9 Accuracy Matrix of water sample stored in Copper Vessels**

| Algorithms Used | Accuracy Score (Before) | Accuracy Score (After Implementation of NavJal Model) |
|---|---|---|
| Logistic Regression | 0.626062514 | 0.63858985 |

| | | |
|---|---|---|
| Decision Tree | 0.620151142 | 0.635845124 |
| Random Forest | 0.632111793 | 0.64332256 |
| KNN | 0.640137748 | 0.65859542 |
| SVM | 0.614494564 | 0.655245959 |
| XGBoost | 0.655301777 | 0.668542628 |
| AdaBoost | 0.661852722 | 0.689863695 |

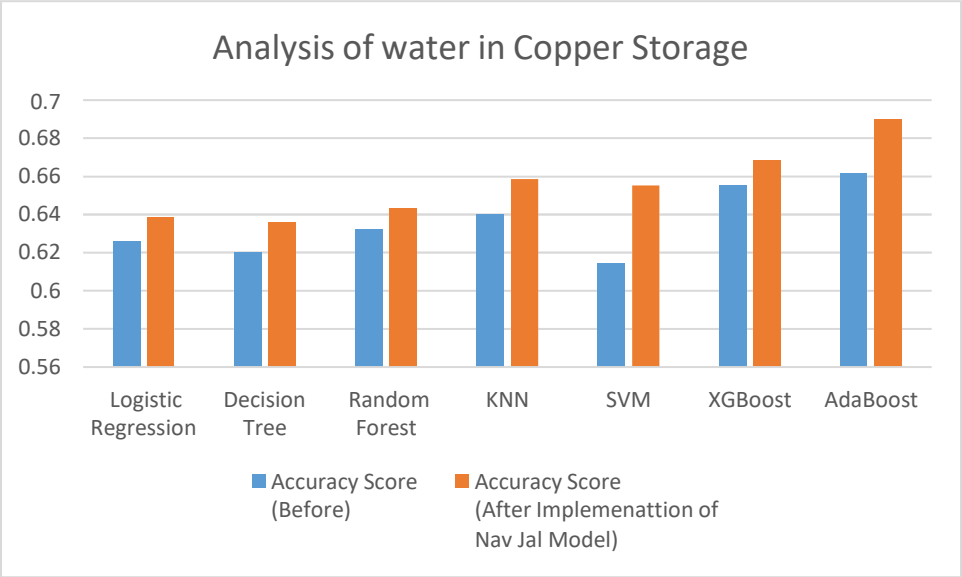


**Figure 4.20 Comparative analysis of water sample stored in Copper Vessels**

**Table 4.10 Accuracy Matrix of Tap Water**

| **Algorithms Used** | **Accuracy Score (Before)** | **Accuracy Score (After the Implementation of Nav Jal Model)** |
|---|---|---|
| Logistic Regression | 0.686937866 | 0.698515356 |
| Decision Tree | 0.663930631 | 0.678912354 |
| Random Forest | 0.680540645 | 0.690532154 |
| KNN | 0.657575761 | 0.668975739 |
| SVM | 0.644034261 | 0.668264719 |
| XGBoost | 0.614272612 | 0.621834262 |
| AdaBoost | 0.679211756 | 0.686545128 |

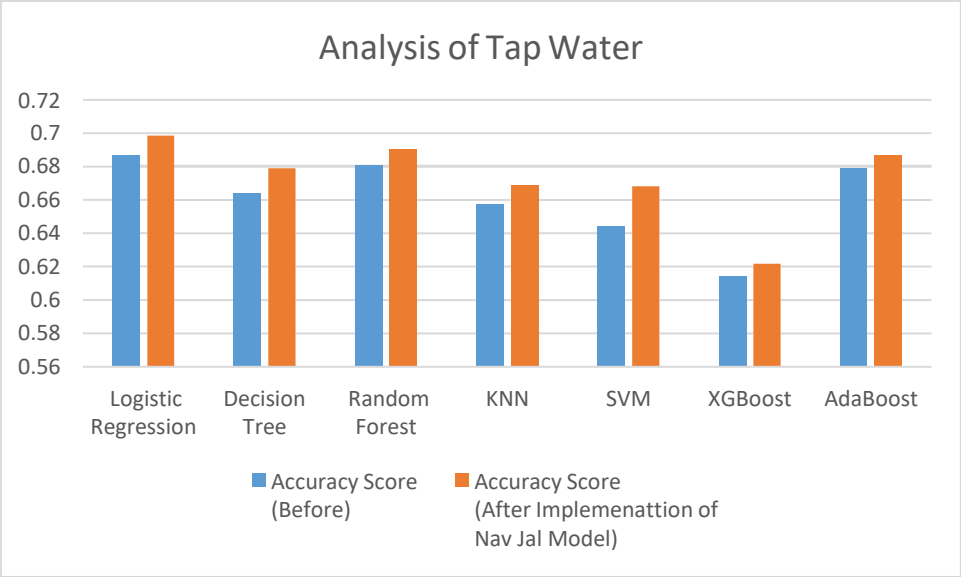


**Figure 4.21 Comparative Analysis of Tap Water**

**d. Conclusion**

The proposed model works on three different modules that carry these seven algorithms and give the best accuracy through Support Vector Machine Technique.

**H4 : Identify the water parameters which are not in permissible range.**

**a. Introduction:**

According to WHO and CPCB, every water parameter has its permissible range to fit specific criteria. If the water parameter is not in that range it can be considered that the water sample is not suitable for drinking water. So it is important to check whether the parameters are within a range or not which has already been defined by WHO/CBCP.

**b. Objectives:**

To identify the factors which improve water purity and make water a healthy energy drink.

**c. Related study:**

Implementation of the algorithm shows that if the water is in the permissible range it will proceed. If not it suggests the deficit value of lacking parameter and the severity so that by adding additional vitamins and minerals can produce a healthy energy drink. Here is the sample of code through which range of water parameter can be validate.

The proposed model first check whether the parameters are in permissible range or not. If values are not in the range the model supposed to give the deficit value of the parameter so that severity of sample can be analyzed.

**Table 4.11 Scaling Factor of Severity**

| **Level** | **Deficit value** |
|---|---|
| High | 60% to 100 % |
| Medium | 40% to 60% |
| Low | Less than 40% |

```
def deficit(value,minimum,maximum):
    if value <= minimum:
        difference = abs(value - minimum)
        if (difference / minimum) * 100 >= 60:
            severity = "High"
        elif (difference / minimum) * 100 >=40:
            severity = "Medium"
        else:
            severity = "Low"
        return {"Deficit":difference,"Severity":severity}
    elif value >= maximum:
        difference = abs(value - minimum)
        if (difference / minimum) * 100 >= 60:
            severity = "High"
        elif (difference / minimum) * 100 >=40:
            severity = "Medium"
        else:
            severity = "Low"
        return {"Deficit":difference,"Severity":severity}
    else:
        return {"Deficit":0,"Severity":"Low"}
```

After the calculation of deficient value if the severity of any parameter is high then it will exit the test engine. But if the deficient value is less it will continue the process as severity is considered to be low or medium. So the test engine will return deficient value for the betterment of water quality.

The below-given code snippet shows the same procedure for total of eighteen parameters.

```
def checker(df):
    if deficit(df['ph'],6.5,8.5)['Severity'] in ['High']:
        return False
    elif deficit(df['Hardness'],300,600)['Severity'] in ['High']:
        return False
    elif deficit(df['Solids'],500,2000)['Severity'] in ['High']:
        return False
    elif deficit(df['Chloramines'],1,4)['Severity'] in ['High']:
        return False
    elif deficit(df['Sulfate'],200,400)['Severity'] in ['High']:
        return False
    elif deficit(df['Conductivity'],0,1000)['Severity'] in ['High']:
        return False
    elif deficit(df['Organic_carbon'],5,5000)['Severity'] in ['High']:
        return False
    elif deficit(df['Trihalomethanes'],0,80)['Severity'] in ['High']:
        return False
    elif deficit(df['Turbidity'],5,10)['Severity'] in ['High']:
        return False
```

```
elif deficit(df['Calcium'],75,200)['Severity'] in ['High']:
    return False
elif deficit(df['Fluroide'],1.5,1.9)['Severity'] in ['High']:
    return False
elif deficit(df['Mineral Oil'],0.01,0.03)['Severity'] in ['High']:
    return False
elif deficit(df['Iron'],0.3,1)['Severity'] in ['High']:
    return False
elif deficit(df['cadmium'],0,0.01)['Severity'] in ['High']:
    return False
elif deficit(df['Copper'],0.05,1.5)['Severity'] in ['High']:
    return False
elif deficit(df['Zinc'],5,15)['Severity'] in ['High']:
    return False
elif deficit(df['Manganese'],0.1,0.3)['Severity'] in ['High']:
    return False
elif deficit(df['Lead'],0,0.05)['Severity'] in ['High']:
    return False
else:
    return True
```

#### d. Conclusion:

This study shows that the model helps to convert the water into a healthy energy drink by suggesting the deficiencies of parametric values to convert normal drinking water into energized drinkable water.

### 5. H5. Improve the efficiency and effectiveness of mining techniques to improve the quality of water

#### a. Introduction:

This study shows that after the implementation of this algorithms it gives the accuracy which predict the water is drinkable or not.

#### b. Objectives:

To improve the efficiency and effectiveness of the water quality using Data Mining and Classification techniques for better outcomes.

**c. Related study:**

This study shows the implementation of these seven algorithms gives an accuracy of water samples and also it provides the deficit value of water parameters which can be improvised after the implementation of the proposed NavJal Model. So the created water sample will be the energized drinkable water with vitamins and minerals.

This can be mapped in the following format for one sample with n number of parameters.

**Table 4.12 Mapping of Water Sample with Parameters**

| Sr. No | Substance | Reading (Before) | The permissible range of Water | Deficient value | Recommendations | Reading (After) | Post Treatment |
|---|---|---|---|---|---|---|---|
| 1 | pH Value | 7.8 | 6.5 to 8.5 | nil | 2.5 to 5 gram | 7.1 | Drinkable |
| 2 | Turbidity ( NTU, Max) | 4.9 | 5 to 10 | 0.1 | 2.5 to 5 gram | 5.3 | Drinkable |
| 3 | Total Hardness (as $CaCo_3$) mg/lit.,Max | 256 | 300 to 600 | 44 | 2.5 to 5 gram | 398 | Drinkable |
| 4 | Dissolved solids mg/lit, Max | 55 | 500 to 2000 | 445 | 2.5 to 5 gram | 657 | Drinkable |
| 5 | Conductivity | 700 | less than 1000 | nil | 2.5 to 5 gram | 758 | Drinkable |
| 6 | Chloramine | 4.9 | 1 to 4 | Additional 0.9 | 2.5 to 5 gram | 4.1 | Non Drinkable |
| 7 | Sulfate (as $SO_4$) mg/lit,Max | 256 | 200 to 400 | nil | 2.5 to 5 gram | 213 | Drinkable |
| 8 | Organic Carbon | 2541 | 5 to 5000 | nil | 2.5 to 5 gram | 2612 | Drinkable |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9 | Trihalomthane | 62 | upto 80 | nil | 2.5 to 5 gram | 68 | Drinkable |
| 10 | Calcium (as Ca) mg/lit,Max | 60 | 75 to 200 | 15 | 2.5 to 5 gram | 78 | Drinkable |
| 11 | Fluoride (as F) mg/lit,Max | 1.84 | 1.5 to 1.9 | Additional 0.11 | 2.5 to 5 gram | 2.01 | non Drinkable |
| 12 | Mineral Oil mg/lit,Max | 0.02 | 0.01 to 0.03 | nil | 2.5 to 5 gram | 0.02 | Drinkable |
| 13 | Iron (as Fe) mg/lit,Max | 0.2 | 0.3 to 1 | 0.1 | 2.5 to 5 gram | 0.31 | Drinkable |
| 14 | Cadmium (as Cd)mg/lit,Max | 0.01 | 0.01 | nil | 2.5 to 5 gram | 0.01 | Drinkable |
| 15 | Copper (as Cu) mg/lit,Max | 0.02 | 0.05 to 1.5 | 0.03 | 2.5 to 5 gram | 0.9 | Drinkable |
| 16 | Zinc (as Zn) mg/lit,Max | 6 | 5 to 15 | nil | 2.5 to 5 gram | 6.8 | Drinkable |
| 17 | Manganese (as Mn)mg/lit,Max | 0.1 | 0.1 to 0.3 | nil | 2.5 to 5 gram | 0.38 | Drinkable |
| 18 | Lead (as Pb) mg/lit,Max | 0.02 | 0.05 | 0.03 | 2.5 to 5 gram | 0.052 | Drinkable |

**Conclusion:**

The proposed model improves the efficiency and effectiveness of the mining algorithm and classification techniques for better results.

### 4.5 Overall Conclusion

The overall analysis of experiments of hypothesis clearly represents that

- After collecting a huge dataset interpretation of data will be easier to get a better monitoring system.

- The storage of water in any particular type of vessel affects the quality of water depending on the storage pattern.
- The model works on three different modules that carry these above mentioned seven algorithms and give the best accuracy through Support Vector Machine Technique.
- The model helps to convert the water into a healthy energy drink by suggesting the deficiencies of parametric values
- The model improves the efficiency and effectiveness of the various algorithm and classification techniques for better results.

# Chapter 5

# Conclusion, Findings and Recommendation

# Chapter 5:Conclusion, Findings and Recommendation

## 5.1 Introduction

The information presented in the parts that came before this one makes it abundantly evident that the quality of the water is a big problem that affects not only the natural world but also the health of the general population.

As a consequence of accelerating urbanization and rising numbers of people, water pollution has emerged as one of the most pressing concerns on a global scale. Technology that can effectively test and monitor water quality is essential if we are to fulfill our responsibility to make sure that water can be consumed by humans and other organisms without risk.

Methods for assuring water quality can be produced through the application of a number of distinct research approaches, including mixed methods, qualitative research, quantitative research, and the testing of hypotheses.

Researchers are able to identify the factors that contribute to poor water quality and develop actionable strategies for monitoring and improving water quality by conducting in-depth analyses of the published literature and datasets that are now at their disposal.

The method that is offered for maintaining water quality is a data mining approach, and it has the potential to improve water quality, protect public health, and have a good effect on the economy.

By clearly defining water quality based on multiple metrics and identifying the reasons of poor water quality, such as pollution and the destruction of the environment, it is feasible to take steps to decrease their effects and improve the overall quality of the water.

This can be accomplished by taking action. It is of the utmost importance to continue research and development of effective methods for water quality certification in order to maintain public health, the environment, and the long-term viability of water supplies for future generations.

## 5.2 Key Findings

The following are the main key findings reached after analyzing the data gathered:

- The proposed NavJal Model is an effective method for water quality assurance which includes data collection, data preprocessing, classification, data mining.
- The classification and data mining techniques are important methods for monitoring water quality and identifying the factors that contribute to poor water quality.
- By accurately classifying water quality and identifying the factors that contribute to poor water quality, effective measures can be taken to mitigate their impact and improve the overall quality of water.

**Overall Analysis**

- Currently in the Market there is not a single system that provides water quality analysis based on 18 parameters. But here the Proposed NavJal Model helps to analyze the water with 18 parameters with heavy metals. i.e. not only ph, TDS, DO, and Turbidity is been considered but other parameters are also taken into consideration.
- The proposed model can be further improved by incorporating more advanced data mining techniques such as deep learning and natural language processing.
- Future work in water quality assurance includes the development of more sophisticated and accurate methods for water quality monitoring and the implementation of policies and regulations to protect water resources and promote sustainable water use.
- The limitations of the research include the availability and quality of data, the complexity of the water quality parameters, and the generalizability of the proposed framework.

### 5.3 Proposed Framework : NavJaL model Pros and Cons

**Pros:**

- Generated sample water can boost the Immunity
- Based on overall analysis SVM gives the best accuracy
- Proposed model can easily predict the deficiency of missing water parameters or parameters which are not in a permissible range.
- Good Taste of water so water consumption is more.

- It lowers the chances of getting sick
- Water samples generated by this proposed model can be beneficial to any age group including senior citizens, Pregnant women, and kids also.
- No human intervention in the process so no errors and accuracy can be maintained

**Cons:**

- NavJal-generated water samples can be expensive because of the minute level of analysis and add-ins of external minerals
- It may be a time-consuming process but which results in better
- A Large number of samples are required for wide scope of analysis.

### 5.4 Suggestions on Proposed Framework

The proposed model is a comprehensive approach to water quality assurance that leverages the power of advanced statistical and Data Mining techniques to provide a more accurate and efficient method for monitoring water quality and identifying potential problems. The proposed model can be further improved by incorporating more advanced data mining techniques such as deep learning and natural language processing.

So the suggestions are as below:

- Here in the proposed data model only drinking water is considered but the model can also be customized to meet the different classes of water with the different purposes of use.
- Also proposed model can be modified in a way with the specific needs of different water bodies.
- Additionally, the model can be extended to include different parameters which may especially treat for various factors.
- Make sure the parametric values should not affect any age group with any basic medical history

### 5.5 Scope for Future Work

Future water quality assurance research will cover a wide range of topics. Future study in this area may focus on a variety of topics, including:

- Creation of new IOT based Device: There is possibility to create a device which can measure maximum parameters of water quality and auto generate the reports which concludes he deficiencies of water and suggest the changes through NavJal Model,
- Creation of cutting-edge water quality monitoring methods: There is room to create more complex and precise systems for monitoring water quality given the speed at which technology is growing. These might include methods like real-time monitoring systems, machine learning, and remote sensing.
- Combining environmental data with information about water quality: The effects of land use, biodiversity, and climate change all interact with the effects of water quality on the environment. A more thorough understanding of the effects of water quality on the environment can be provided by integrating water quality data with other environmental data, which can aid in the development of efficient environmental management plans.
- Development of new and sustainable water treatment solutions: With the rising demand for clean water, innovative and sustainable water treatment solutions are required. This might entail the application of bioremediation methods, nanotechnology, and natural treatment systems.
- Implementing policies and regulations to protect water resources: To prevent pollution and deterioration of water resources, it is necessary to implement policies and laws that are effective. This could involve actions like trading on water quality, managing watersheds, and providing incentives to businesses to switch to more environmentally friendly production methods.
- Cross-disciplinary cooperation in water quality research: Environmental science, public health, engineering, and the social sciences are just a few of the disciplines that are involved. The skills and knowledge from several sectors can be combined through cross-disciplinary collaboration to provide practical solutions for ensuring the quality of the water.
- Future research in ensuring water quality will include a wide range of topics and necessitate a multidisciplinary strategy. We can assure the sustainability of water supplies and safeguard the environment by continuing to develop new and

sustainable solutions, laws, and regulations. The hybrid classification and data mining strategy combines two distinct methods for categorizing and analyzing data on water quality. The first strategy is classification, which entails grouping data on water quality into various classes in accordance with predetermined criteria or traits. The second strategy is known as data mining, and it entails the use of sophisticated analytical methods to find patterns, connections, and trends in the data.

### 5.6 Findings of the Implementation of NavJal Model

A table displaying correlation coefficients between variables is called a correlation matrix. The correlation between the two variables is displayed in each cell of the table. Data are summarized using correlation matrices, which are also used as inputs for more sophisticated studies and as diagnostics for such analyses.

A correlation matrix can be used to summarize a big data collection, find trends, and guide decisions. Additionally, it can show which variable is more connected with which variable and visualize the outcomes.

The demonstration is given as below:

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

```
print(df.nunique())
ph                 2785
Hardness           3276
Solids             3276
Chloramines        3276
Sulfate            2495
Conductivity       3276
Organic_carbon     3276
Trihalomethanes    3114
Turbidity          3276
Potability            2
dtype: int64
```

```
In [10]:
print(df.isnull().sum())
ph                 491
Hardness             0
Solids               0
Chloramines          0
Sulfate            781
Conductivity         0
Organic_carbon       0
Trihalomethanes    162
Turbidity            0
Potability           0
dtype: int64
```

```
In [11]:
df.dtypes
Out[11]:
ph                 float64
Hardness           float64
Solids             float64
Chloramines        float64
Sulfate            float64
Conductivity       float64
Organic_carbon     float64
Trihalomethanes    float64
Turbidity          float64
Potability           int64
dtype: object
```

```
In [12]:
sns.heatmap(df.isnull())
Out[12]:
<AxesSubplot:>
```

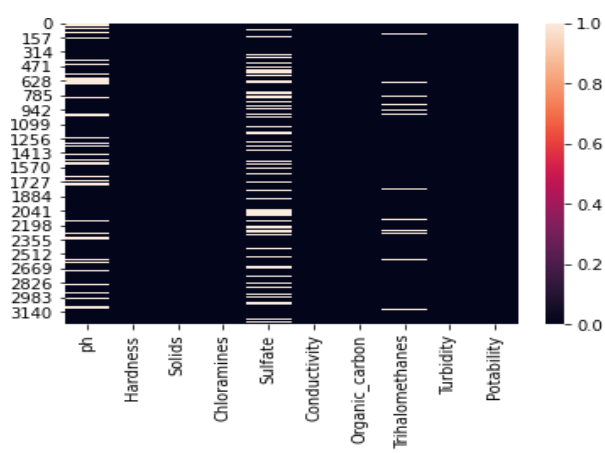


**Figure 5.1 Parametric Representation**

```
In [13]:
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(), annot= True, cmap='coolwarm')

Out[13]:
<AxesSubplot:>
```

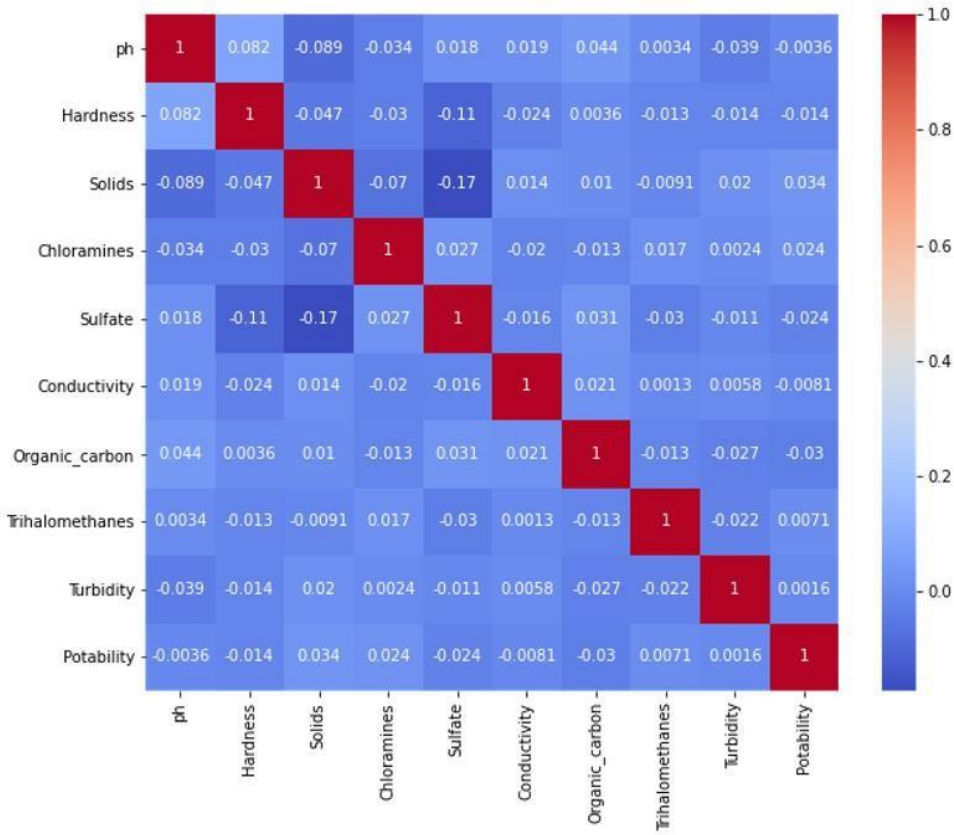


**Figure 5.2 Correlation Matrix 1**

```
In [14]:
# Unstacking the correlation matrix to see the values more clearly.
corr = df.corr()
c1 = corr.abs().unstack()
c1.sort_values(ascending = False)[12:24:2]

Out[14]:
Hardness  Sulfate          0.106923
ph        Solids           0.089288
Hardness  ph               0.082096
Solids    Chloramines      0.070148
Hardness  Solids           0.046899
ph        Organic_carbon   0.043503
dtype: float64
```

```
ax = sns.countplot(x = "Potability",data= df, saturation=0.8)
plt.xticks(ticks=[0, 1], labels = ["Not Potable", "Potable"])
plt.show()
```

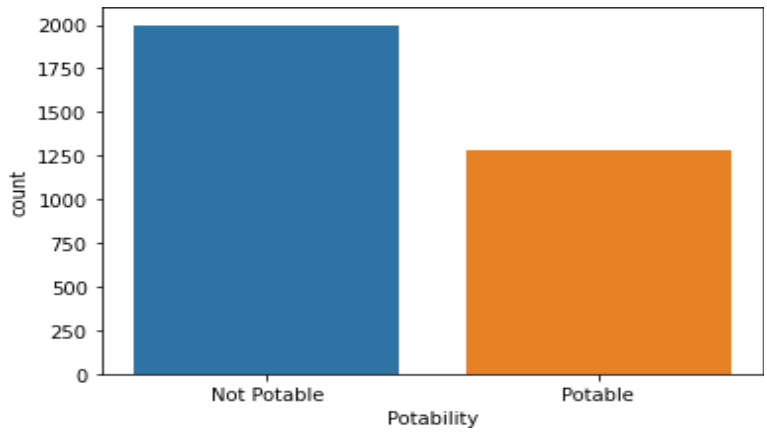


**Figure 5.3: Potability Representation 1**

```
x = df.Potability.value_counts()
labels = [0,1]
print(x)
0    1998
1    1278
Name: Potability, dtype: int64
sns.violinplot(x='Potability', y='ph', data=df, palette='rocket')
<AxesSubplot:xlabel='Potability', ylabel='ph'>
```

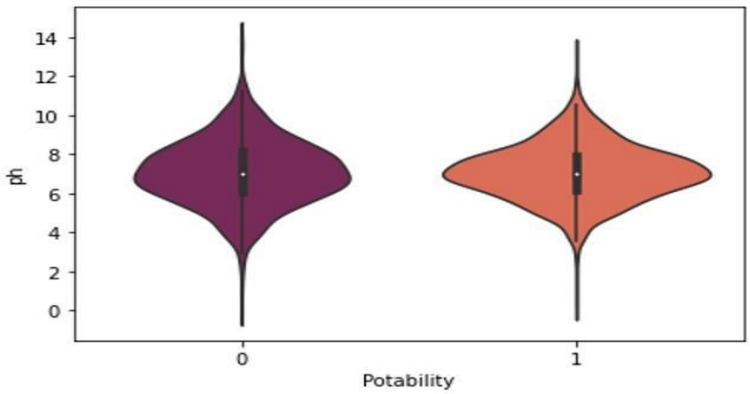


**Figure 5.4: Potability Representation 2**

This is done by implementing the model first with different nine parameters which show whether the water is potable or not as an outcome. But out of eighteen remaining nine

parameters have been considered to find more correlation between them. And the other nine parameters are calcium, fluoride, Mineral oil, Iron, Copper, Zinc, Manganese, Cadmium, and Lead. So the created correlation matrix of these nine parameters found the relation of portability with the remaining elements.

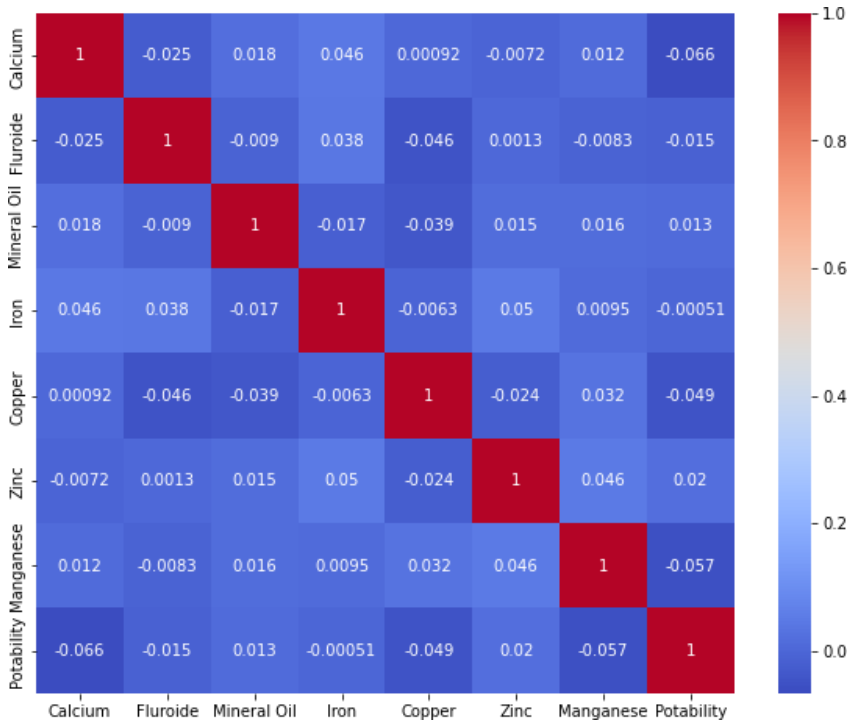


**Figure 5.5 Correlation Matrix 2**

It was clear from the matrix that the correlation is not good as expected and doing the prediction on this may lead to underfitting of the model.

So along with that proposed model can check the parameters which are not in the specified range.

So proposed model extracts the data according to the permissible limits of the water of each element. If a particular data satisfies all the permissible limits then only it is sent to the model for prediction of water quality.

## 5.7 Validation of Proposed Framework

This section constitutes the final stage of assessment of the Proposed Model NavJal. The purpose is to verify whether the proposed framework is viable and acceptable in the system. The validation process considered the dataset after implementation of NavJal Model. It shows new data and testing the model's ability to accurately classify water quality and identify the factors that contribute to poor water quality.

This provides valuable insights into the model's effectiveness and practicality and helps refine the proposed model to better address the challenges of water quality assurance.

For validation purposes, the same dataset has been tested after the NavJal model implementation with the same parameters but with enhanced values.

The Proposed NavJal Model generates the new water sample, and this new sample has been tested for accuracy. The implementation is shown as below

**NavJal Sample Testing :**

```
from sklearn.svm import SVC, LinearSVC
model_svm = SVC(kernel='rbf', random_state = 42)
model_svm.fit(X_train, y_train)

SVC(random_state=42)

# Making Prediction
pred_svm = model_svm.predict(X_test)

# Calculating Accuracy Score
sv = accuracy_score(y_test, pred_svm)
print(sv)
```

**Table 5.1 Comparative analysis of Accuracy Score with proposed model**

| Sr. No | Model | Accuracy _score | NewJal Model Accuracy_Score |
|---|---|---|---|
| 1 | SVM | 0.688540 | 0.72501 |
| 2 | XGBoost | 0.670980 | 0.68052 |
| 3 | Kneighbours | 0.653420 | 0.71420 |

| Sr. No | Model | Accuracy _score | NewJal Model Accuracy_Score |
|---|---|---|---|
| 4 | Decision Tree | 0.645102 | 0.68125 |
| 5 | AdaBoost | 0.634011 | 0.634021 |
| 6 | Logistic Regression | 0.628466 | 0.628852 |
| 7 | Random Forest | 0.628466 | 0.637455 |

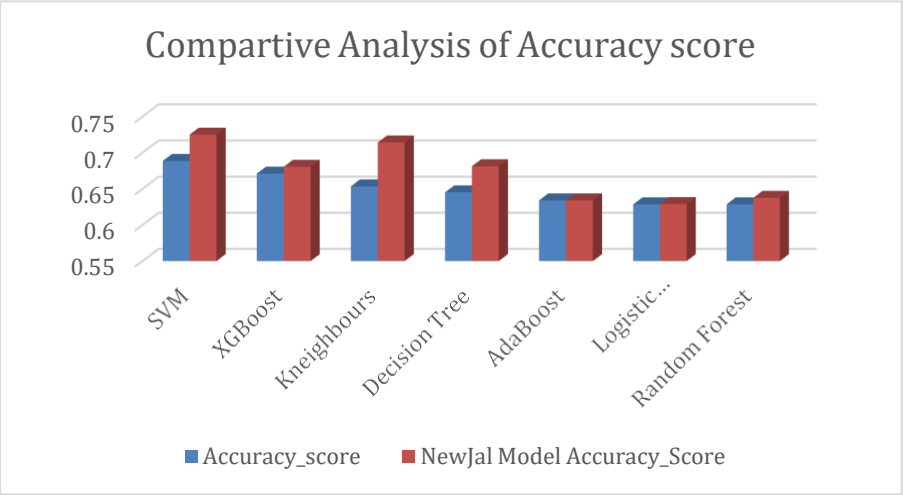


**Figure: 5.6 Comparative Analysis of Accuracy Score**

Conclusion: Here water sample which was created by NavJal model has produced the highest accuracy by SVM model that the traditional model.

### 5.8 Research Contribution

This study promises to contribute in multiple aspects

- NavJal Model will help to standardize the system and improves the quality of Drinking water.
- The research and its findings are useful for future researchers aimed at creating a more enhanced model/framework for the betterment of any category of water quality

- The findings of the study will give insights to academic researchers, Industry experts, and the government for water quality enhancement techniques.
- Also this outcome and findings will not only help to improve the drinking water quality but can be used for another category of water also.

### 5.9 Limitations of the Research

The research has some limitations which includes

- **Availability and quality of data:** One of the major limitations of the research is the availability and quality of data. The proposed model relies heavily on water quality data to monitor and assess water quality. However, there may be limitations in the availability of this data, especially in developing countries or remote regions.
- **Data Complexity:** The complexity of the water quality parameters is a major limitation of this model. Water quality is influenced by a multitude of physical, chemical, and biological factors that can interact in complex ways. As a result, it can be challenging to accurately model and predict water quality using traditional statistical and machine learning techniques.
- **Type of Water:** The study is limited to a specific class of water, the current model doesn't suitable to other types of water bodies.
- In addition, the quality of the data may be inconsistent, incomplete, or inaccurate, which can impact the accuracy and effectiveness of the proposed Model.
- **Generalizability of the proposed model:** The proposed model is developed based on water quality data from specific regions and water bodies. As such, its generalizability to other regions and water bodies may be limited. The effectiveness and accuracy of the framework may be influenced by factors such as geography, climate, and human activities.

- **Ethical and social considerations:** The proposed model is primarily focused on the technical aspects of water quality assessment and may not fully consider the ethical and social implications of water quality management. For example, the model may not fully address issues related to social equity and access to clean water, which are important considerations in developing countries.
- **Integration with existing water quality management systems:** The proposed framework may not be fully integrated with existing water quality management systems, which can create challenges in implementing and scaling the model. Effective implementation of the model may require significant investment in data infrastructure, capacity building, and stakeholder engagement.

# References

# References


- Afaq Juna , Muhammad Umer , Saima Sadiq,et al , “Water Quality Prediction Using KNN Imputer and Multilayer Perceptron” Water 2022, 14, 2592. https://doi.org/10.3390/w14172592
- Akumtoshi L, Singh MR, Puro N (2020) Assessment of water quality status of Doyang River, Nagaland, India, using Water Quality Index. Appl Water Sci 10(1):1–13
- Ana Elizabeth,Diego Cruz,Elena,et al,“Groundwater Quality Assessment: An Improved Approach to K-Means Clustering, Principal Component Analysis and Spatial Analysis: A Case Study , Water 2018, 10(4), 437; https://doi.org/10.3390/w10040437
- Athira K R , Varsha Murali , Aswin C R , Devaprabha B, Aneesh P C ,”Water Quality Modeling and Regression Analysis of Vembanad Lake and its Primary Inflow” Volume 10, Issue 06 (June 2021) DOI : 10.17577/IJERTV10IS060185
- B Warsito et al,”Evaluation of river water quality by using hierarchical clustering analysis”,2 021IOP Conf. Ser.: Earth Environ. Sci. 896 012072
- Bawa KV, Gaikawad VB (2013) Water quality assessment of Godavari River at Nashik, India: impact of sewage and industrial wastewa- ter. Univ J Environ Res Technol 3(4): 452–457. ISSN 2249 0256.
- Bhargava DS (1983) Use of a water quality index for river classifica- tion and zoning of the Ganga River. Environ Pollut Ser B 6:51–67 Bora M, Goswami DC (2017) Water quality assessment in terms of water quality index (WQI): case study of the Kolong River,
- Brown RM, McClelland NI, Deininger RA, Tozer RG (1970) A water quality Index- do we dare. Water Sew Works 117:339–343
- Chavan AD, Sharma MP, Bhargava R (2009) Water quality assess- ment of the Godavari River. Hydro Nepal J Water Energy Environ 5:31–34

- Consolata Gakii and Jeniffer Jepkoech ,"A Classification Model For Water Quality AnalysisUsing Decision Tree", European Journal of Computer Science and Information Technology Vol.7, No.3, pp.1-8, June 2019
- Dalkey NC (1968) DELPHI. The Rand Corporation, Santa Monica Das DN (2013) The impact of industrial and urban activities on the
- David, C.P. Heavy metal concentrations in growth bands of corals: A record of mine tailings input through time (Marinduque Island, Philippines). *Mar. Pollut. Bull.* **2003**, *46*, 187–196.
- Duan W, Meng F, Cui H, Lin Y, Wang G, Wu J (2018) Ecotoxicity of phenol and cresols to aquatic organisms: a review. Ecotoxicol Environ Safety 157:441–456
- Govorushko SM (2010) Effect of Human activity on Rivers Interna- tional Congress on River Basin Managemant, pp 464–476
- Horton RK (1965) An index number system for rating water quality, Journal of the Water Pollution Control Federation, 37, 300-306.
- Ichwana S, Nelly W (2016) Water quality Index by using National Sanitation Foundation—Water quality Index (NSF-WQI). Inter- national conference on technology innovation and society. 978-602-70570-4-3 © 2016 Published by ITP PRESS DOI 10.21063/ICTIS.2016.1019
- Isai KA, Shrivastava VS (2015) Detection and identification of organ- ics and metals from industrial wastewater by ICP-AES, FTIR and GC-MS. J Adv Chem Sci 1(04):164–166
- K.saikrishna, D.Purushotham, V.Sunitha," Data for the evaluation of groundwater quality using water quality index and regression analysis in parts of Nalgonda district, Telangana, Southern India", Elsevier, Volume 32, October 2020, 106235, https://doi.org/10.1016/j.dib.2020.106235
- Kharake A, Pathare J, Deshmukh P (2021) Spatio-temporal variability of intra-monsoonal rainfall in Pravara-Mula river basin, India. Arebian J Geosci 14(10):1–10

- Kharke AC (2008) Analysis of Geo-environmental effects due to human interventions along Pravara River at Sangamner city, Ahmednagar, Maharashtra. Multidiscip Approaches Remote Sens GIS 1(1): 89–93, ISSN: 2319–8648
- Khound NJ, Bhattacharyya KG (2018) Assessment of water quality in and around Jia-Bharali river basin, North Brahmaputra Plain, India, using multivariate statistical technique. Appl Water Sci 8(8):221
- Koichi F, Hong-Ying HU (2010) Effects of human activities on water quality. In: Water quality and standard, WATER QUALITY AND STANDARDS – Vol. II - Effects of Human Activities on Water Quality - Koichi Fujie, Hong-Ying Hu, ©Encyclopedia of Life Support Systems (EOLSS)
- Koli PB, Kailas HK, Uday GD (2018) Study of physico-chemical prop- erties, detection and toxicity study of organic compounds from effluent of MIDC Thane and GIDC Ankleshwar industrial zone. Appl Water Sci 8:196.
- Kumar A, Dua A (2009) Water quality index for assessment of water quality of river Ravi at Madhopur (India). Glob J Environ Sci 8:1 Marle SM (2011) Assessment of pilgrimage impact on river water qual- ity and health along river Indrayani, District Pune. Unpublished
- Lei Zhang ,Big Data, Knowledge Mapping for Sustainable Development: A Water Quality Index Case Study,2019
- Liberti, A.; Bertocci, I.; Pollet, A.; Musco, L.; Locascio, A.; Ristoratore, F.; Spagnuolo, A.; Sordino, P. An indoor study of the combined effect of industrial pollution and turbulence events on the gut environment in a marine invertebrate. Mar. Environ. Res. 2020, 158, 104950
- Md.Galal Uddin, stephenNash, Agnieziska, "A review of Water quality Index models and their use for assessing surface Water quality" , Volume 122, March 2021, 107218,Elsevier
- Muhammad AHR, Malik MM, Sana M (2018) Urbanisation and its effects on water recourses: an exploratory Analysis. Asian J Water Environ Pollut 15(1):67–74

- Mustakim,Rosdina, Dian Ramad," The Classification Status of River Water Quality in Riau Province Using Modified K-Nearest Neighbor Algorithm with STORET Modeling and Water Pollution Index", Journal of Physics, DOI 10.1088/1742-6596/1783/1/012020
- Nadikatla SK, Mushini VS, Mudumba PSMK (2020) Water quality index method in assessing groundwater quality of Palakonda man- dal in Srikakulam district, Andhra Pradesh, India. Appl Water Sci 10(1):30
- Nag SK, Kundu A (2018) Application of remote sensing, GIS and MCA techniques for delineating groundwater prospect zones in Kashipur block, Purulia district, West Bengal. Appl Water Sci 8(1):38
- Neetu Arrora, Amarpreet Singh Arora, et al" Use of Cluster Analysis-A Data mining tool for improved water quality monitoring of River Satluj" ISSN No. : 0975-0290,2015
- Obiora DN, Ibuot JC (2020) Geophysical assessment of aquifer vulner- ability and management: a case study of University of Nigeria, Nsukka, Enugu State. Appl Water Sci 10(1):29
- Patel SV, Chavda P, Tyagi S (2019) 'carrying out assessment of groundwater quality of villages of Bhavnagar district of Gujarat, India. Environ Claims J 31(1):79–92
- Patil S, Ghorade IB (2013) Assessment of physico-chemical character- istics of Godavari river water at Trimbakeshwar and Kopargaon, Maharashtra. Indian J Appl Res 3(3):149–152
- Phadatare SS, Gawande S (2016) Review paper on development of water quality index.
- Popovic NZ, Duknic JA, Atlagic JZ, Rakovic MJ, Marinkovic NS, Tubic BP, Paunovic MM (2016) Application of the water pollution index in the assessment of the ecological status of rivers: a case study of the Sava River, Serbia. Acta Zoologica Bulgarica 68(1):97–102

- S. Brintha Rajakumari, C. Nalini ,"Analysis of lead in river water quality data using clustering"2016, ISSN: 0974-2115 www.jchps.com Journal of Chemical and Pharmaceutical Sciences
- Sattari, Mohmmd, Feizi , "Surface water quality classification using data mining approaches: Irrigation along the Aladag River,2021, DOI:10.1002/ird.2594
- Shah KA, Joshi GS (2017) Evaluation of water quality index for River Sabarmati, Gujarat, India. Appl Water Sci 7(3):1349–1358
- Shinde S, Shrivastava VS (2018) Physico-chemical assessment and analysis of hazardous organic substance from textile industrial effluents from Sachin and Kadodara, GIDC, Surat, India. Rev Res 8(02):1–18
- Tovar-Sánchez, A.; Rodríguez-Romero, A.; Engel, A.; Zäncker, B.; Fu, F.; Marañón, E.; Pérez-Lorenzo, M.; Bressac, M.; Wagener, T.; Triquet, S.; et al. Characterizing the surface microlayer in the Mediterranean Sea: Trace metal concentrations and microbial plankton abundance. *Biogeosciences* **2020**, *17*, 2349–2364.
- Tyagi S, Sharma B, Singh P, Dobhal R (2013) Water quality assessment in terms of water quality index. Water Res 1(3):34–43
- Unde M, Turkunde K (2008) Geo-environmental effects of urbanization in the river channel: a case study of River Sina around Ahmedna- gar city. Multidisciplinary Int Res J 1(03)
- UNESCO (1998) World Water Resources: a New Appraisal and Assessment for 21st century (Report) using water quality index (WQI) method and GIS in Aksu River (SW-Turkey). Sci Total Environ 584:131–144
- Vaishali AC, Ashali K (2019) Water quality assessment of Pra- vara river at Sangamner Tehsil, Ahmednagar District, India: an impact of anthropogenic activities. Rev Res 08(05): 1–5, ISSN- 2249–894X
- Venkaraman,Nikitha,Rishika" Water quality index estimation using linear regression model"2022, AIP Conference Proceedings 2418, 040033 (2022); https://doi.org/10.1063/5.0081794

- Verma RK, Murthy S, Tiwary RK, Verma S (2019) Development of simplified WQIs for assessment of spatial and temporal variations of surface water quality in upper Damodar river basin, Eastern India. Appl Water Sci 9(1):21
- Water quality of Tunia River, Assam. Unpublished Ph.D. Thesis submitted to Gauhati University, Assam.
- Wei Cong Leong, Aleriza,Jie,Ahmed," Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM) ", International Journal of River Basin Management Volume 19, 2021 - Issue 2
- Wu Z, Zhang D, Cai Y, Wang X, Zhang L, Chen Y (2017) Water qual- ity assessment based on the water quality index method in Lake Poyang: the largest freshwater lake in China. Sci Rep 7(1):1–10
- Yogendra K, Puttaiah ET (2008) Determination of water quality index and suitability of an urban waterbody in Shimoga Town, Karna- taka. In: Proceedings of Taal2007: The 12th world lake confer- ence, vol 342, p 346
- Zhaoshi W, Zhang D, Cai Y, Xiaolong W, Lu Z, Yuwei C (2017) Water quality assessment based on the water quality index method in Lake Poyang: the largest freshwater lake in China. Sci Rep 7:17999

**Reports**


- MCGM ESR English 2020 – 2021 Report (Environment Status Report 2020-21)
- Water Pollute Control Fed 37:300–306
- Maps_Maharashtra Government Maps
- Assam. India Appl Water Sci 7(6):3125–3135
- Water Quality of Rivers -2021(CPCB)
- Water Quality of Lakes, Ponds, Tanks, and Wetlands 2021 (State-wise)(CPCB)
- MPCB Annual Report 2020-21
- Market Research solution
- IS 10500:2012 Drinking Water Specification
- Guidelines for Water Quality Monitoring by MPCB


**Website**


- https://www.who.int/teams/environment-climate-change-and-health/water-sanitation-and-health/water-safety-and-quality/drinking-water-quality-guidelines
- https://cpcb.nic.in/nwmp-data/
- https://mpcb.gov.in/water-quality
- https://www.usgs.gov/mission-areas/water-resources
- https://portal.mcgm.gov.in/irj/go/km/docs/documents/MCGM%20Department%20List/Environment/Docs/MCGM%20ESR%20English%202020%20-%202021.pdf
- http://cgwb.gov.in/documents/wq-standards.pdf
- https://www.indiawaterportal.org/articles/indian-standard-drinking-water-bis-specifications-10500-2012-second-revision
- https://www.irctc.com/rail-neer-introduction.html

# Questionnaire

# Questionnaire

**Section I. Demographic Questions**

1. Name

……..

2. Email ID

……..

3. Phone Number

……..

4. Designation

……..

5. Organization Name

……..

6. Location

……..

**Section II. Drinking Water and its Quality**

7. How much water do you drink per day?

- 0.5 L
- 1 L
- 1.5 L
- 2 L
- 3 L
- More than 3L

8. Which is the source of water intake on daily basis?

- Tap water
- Filtered Water
- Packed Water
- Outsource
- Other

9. Percentage of sick leaves in your organization?

- 5% - 15%
- 15% - 25%
- 25% - 35%
- More than 35 %

10. What is the pH level of drinking water in your organization

- No idea
- <1
- 1 to 4
- 4 to 6.5
- 6.5 to 8.5
- 8.5 to 10
- 10 to 12
- 12 to 14

11. What is the TDS level of drinking water in your organization?

- Excellent (less than 300mg/L)
- Good (300mg/L-600mg/L)
- Fair (600mg/L-900mg/L)
- Poor (900mg/L-1200mg/L)
- Unacceptable (greater than 1200mg/L)
- No idea about TDS

12. Which Filtered water do you prefer in your organization?

- Bisleri
- Kinley
- Aquafina
- Bailley
- RO

- Other/Local Brands

13. According to you , where is the water of highest quality found?

- From wells
- From packed water
- From tap water
- RO
- No Idea

14. What percentage of overall budget is allocated in your company towards purified water for drinking water?

- 0 to 20%
- 20 to 40%
- 40 to 60%
- 60 to 80%
- 80 to 100%

15. Have you analyzed your drinking water quality in laboratory?

- Yes, Frequently
- Never
- Yes,Once upon a time
- No Idea

16. In your opinion, who should be responsible for drinking water quality

- Water Neutralization
- Filteration of water
- Chlorination
- Other water purification Techniques

17. In your opinion , who should be responsible for drinking water quality?

- Society for conversation /Pollution of water
- Municipalities
- Water suppliers
- In person
- Everyone as mentioned

18. How do you preserve drinking water in your home/Organization?

- Water tanks
- Earthen pots
- Plastic bottles
- Steel vessels
- Others

# Questionnaire for Interviewer

**General Questions:**

Name of the Employee:_____

Name of The Organization:_____

Type of the Company:_____

Designation:_____

**Water Quality Related Questions:**

1. Which filtered water do you prefer in your organization
    a. Water Purifier
    b. Outsourced /Packaged Drinking
    c. Tap Water
    d. Plastic Bottle
    e. Bring your own Water (BYOW)
2. Do you trust the Water Supply of your organization?
    a. Yes
    b. No
3. What are the pH and TDS levels of Drinking water in your organization?
    a. PH Level
    b. TDS Level
4. What are the reasons for sick leave in your Organization
   (Climate change, Water (water retention/dehydration), Food Infection, Stress)
5. How frequently do you test the Water Quality
    a. Quarterly
    b. Six Monthly
    c. Annually
    d. Ignore

6. Is there any overall budget is allocated in your company towards purified drinking water?
    a. Yes
    b. No
7. Are there effective complaint mechanisms at the level of the water service provider?
    a. Yes
    b. No
    c. In progress

# Data Analytical Model on Water Quality Assurance Based On the Hybrid Classification and Data Mining Techniques

A THESIS

SUBMITTED TO THE

**TILAK MAHARASHTRA VIDYAPEETH, PUNE**

FOR THE DEGREE OF

**DOCTOR OF PHILOSOPHY**

**IN COMPUTER SCIENCE**

**UNDER THE BOARD OF MODERN SCIENCES**

**& PROFESSIONAL SKILLS**

BY

**Swapnali Dattatray Mahadik**

(Registration No. 25315008500)

UNDER THE GUIDANCE OF

**Dr. Anup Girdhar**

**DEPARTMENT OF COMPUTER SCIENCE**

**July 2023**

# 80_Recommendation

## Conclusion

The following are the main key findings reached after analyzing the data gathered:

- The proposed NavJal Model is an effective method for water quality assurance which includes data collection, data preprocessing, classification, data mining.
- The classification and data mining techniques are important methods for monitoring water quality and identifying the factors that contribute to poor water quality.
- By accurately classifying water quality and identifying the factors that contribute to poor water quality, effective measures can be taken to mitigate their impact and improve the overall quality of water.
- Currently in the Market there is not a single system that provides water quality analysis based on 18 parameters. But here the Proposed NavJal Model helps to analyze the water with 18 parameters with heavy metals. i.e. not only ph, TDS, DO, and Turbidity is been considered but other parameters are also taken into consideration.
- The proposed model can be further improved by incorporating more advanced data mining techniques such as deep learning and natural language processing.
- Future work in water quality assurance includes the development of more sophisticated and accurate methods for water quality monitoring and the implementation of policies and regulations to protect water resources and promote sustainable water use.
- The limitations of the research include the availability and quality of data, the complexity of the water quality parameters, and the generalizability of the proposed framework.

## Suggestions

The proposed model is a comprehensive approach to water quality assurance that leverages the power of advanced statistical and Data Mining techniques to provide a more accurate and efficient method for monitoring water quality and identifying potential problems. The proposed model can be further improved by incorporating more advanced data mining

techniques such as deep learning and natural language processing.

So the suggestions are as below:

- Here in the proposed data model only drinking water is considered but the model can also be customized to meet the different classes of water with the different purposes of use.
- Also proposed model can be modified in a way with the specific needs of different water bodies.
- Additionally, the model can be extended to include different parameters which may especially treat for various factors.
- Make sure the parametric values should not affect any age group with any basic medical history

**Scope for Future Work**

Future water quality assurance research will cover a wide range of topics. Future study in this area may focus on a variety of topics, including:

- Creation of new IOT based Device: There is possibility to create a device which can measure maximum parameters of water quality and auto generate the reports which concludes he deficiencies of water and suggest the changes through NavJal Model,
- Creation of cutting-edge water quality monitoring methods: There is room to create more complex and precise systems for monitoring water quality given the speed at which technology is growing. These might include methods like real-time monitoring systems, machine learning, and remote sensing.
- Combining environmental data with information about water quality: The effects of land use, biodiversity, and climate change all interact with the effects of water quality on the environment. A more thorough understanding of the effects of water quality on the environment can be provided by integrating water quality data with other environmental data, which can aid in the development of efficient environmental management plans.
- Development of new and sustainable water treatment solutions: With the rising demand for clean water, innovative and sustainable water treatment solutions are required. This might entail the application of bioremediation methods,

nanotechnology, and natural treatment systems.

- Implementing policies and regulations to protect water resources: To prevent pollution and deterioration of water resources, it is necessary to implement policies and laws that are effective. This could involve actions like trading on water quality, managing watersheds, and providing incentives to businesses to switch to more environmentally friendly production methods.
- Cross-disciplinary cooperation in water quality research: Environmental science, public health, engineering, and the social sciences are just a few of the disciplines that are involved. The skills and knowledge from several sectors can be combined through cross-disciplinary collaboration to provide practical solutions for ensuring the quality of the water.
- Future research in ensuring water quality will include a wide range of topics and necessitate a multidisciplinary strategy. We can assure the sustainability of water supplies and safeguard the environment by continuing to develop new and sustainable solutions, laws, and regulations. The hybrid classification and data mining strategy combines two distinct methods for categorizing and analyzing data on water quality. The first strategy is classification, which entails grouping data on water quality into various classes in accordance with predetermined criteria or traits. The second strategy is known as data mining, and it entails the use of sophisticated analytical methods to find patterns, connections, and trends in the data.