WATER QUALITY PREDICTION USING ARTIFICIAL NEURAL NETWORK

A PROJECT REPORT

Submitted by

MANYA M (TKM21MCA-2027)

to

The APJ Abdul Kalam Technological University

In partial fulfillment of the requirements for the award of the degree of

MASTER OF COMPUTER APPLICATION

Thangal Kunju Musaliar College of Engineering
Kerala

DEPARTMENT OF COMPUTER APPLICATION

MAY 2023

DECLARATION

I undersigned hereby declare that the project report on **WATER QUALITY PREDICTION USING ARTIFICIAL NEURAL NETWORK**, submitted for partial fulfillment of the requirements for the award of degree of Master of Computer Application of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of **Prof.Jasmin M R**. This submission represents my ideas in my own words and where ideas or words of others have been included,I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in our submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not previously served as the basis for the award of any degree, diploma, or similar title by any other University.

Kollam

16-05-2023

MANYA M

DEPARTMENT OF COMPUTER APPLICATION

TKM COLLEGE OF ENGINEERING KOLLAM

2021 - 23

CERTIFICATE

This is to certify that the report entitled **WATER QUALITY PREDICTION USING ARTIFICIAL NEURAL NETWORK** submitted by **MANYA M** (TKM21MCA2027) to the APJ Abdul Kalam Technological University in partial fulfillment of the Masters degree in Computer Applications is a bonafide record of the project work carried out by her under our guidance and supervision. This report, in any form, has not been submitted to any other University or Institute for any reason.

Internal Supervisor Head of the Department External Examiner

ACKNOWLEDGEMENT

First and foremost, I thank GOD almighty and my parents for the success of this project. I owe sincere gratitude and heart full thanks to everyone who shared their precious time and knowledge for the successful completion of my project.I am extremely grateful to **Dr. Fousia M Shamsudeen**, Head of the Department, Department of Computer Application, for providing me with the best facilities.I would like to thank my coordinator **Prof. Vaheetha Salam** for her immense encouragement.I would like to thank my project guide **Prof.Jasmin M R**, Department of Computer Application, who motivated and guided me throughout my work.With a profound sense of gratitude, I would like to express my heartfelt thanks to my advisor **Prof. Natheera Beevi M** Department of Computer Application, for her expert guidance, co-operation.I profusely thank all other faculty members in the department and all other members of TKM College of Engineering, for their guidance and inspiration throughout my course of study.I owe my thanks to my friends and all others who have directly or indirectly helped me in the successful completion of this project.

MANYA M

ABSTRACT

WATER QUALITY PREDICTION USING ARTIFICIAL NEURAL NETWORK is a predition system for predicting the quality of water.This system predicts the water is useful for human conception or not.Water quality is a crucial aspect of ensuring public health and safety. The accurate prediction of water quality parameters can aid in the effective management of water resources.Water is used for various purpose like drinking,agriculture,business etc.Water quality prediction become an essential part in nowadays.

In recent years, machine learning and deep learning models have shown promising results in predicting water quality parameters.The proposed system use MultiLayer Perceptron Neural Network for train the model.Also the system make the use of machine learning classification algorithms such as Decision Tree,Random Forest,SVM and KNN to train the model.Machine learning classification algorithms are used for analyzing the performance best algorithm with MLP.The main objective of the system is to create an easy to use water quality prediction tool.

Contents

List of Figures

List of Abbreviations

ANN Artificial Neural Network

DT Decision Tree

KNN K Nearest Neighbors

MLP MultiLayer Perceptron

PCA Principal Component Analysis

SVM Support Vector Machine

Chapter 1

INTRODUCTION

WATER QUALITY PREDICTION USING ARTIFICIAL NEURAL NETWORK is a prediction system for predicting water quality using artificial neural network.Water is an essential resource for life and plays a crucial role in various aspects of human activities such as agriculture, industry, and domestic use. However, with increasing human activities and climate change, water quality has become a significant concern globally. Water quality prediction is an essential process that helps to prevent water pollution and protect human health.

During the last years, water quality has been threatened by various pollutants. Therefore, modeling and predicting water quality have become very important in controlling water pollution.Water quality can be predict using various methods. The automatic prediction of water quality has shown considerable promise when using artificial neural network.With the use of artificial neural network,intend to create a water quality predicting system in this project.Artificial Neural Networks (ANNs) are a powerful tool for predicting water quality parameters.There are several types of Artificial Neural Networks (ANNs), each with its own architecture and learning algorithm.The system will use MultiLayer Perceptron Model(MLP).A MultiLayer Perceptron (MLP) is a type of feedforward neural network and widely used model in the field of artificial neural networks.

In WATER QUALITY PREDICTION USING ARTIFICIAL NEURAL NETWORK build a MLP model which is able to handle large dataset and recognize pattern efficiently.Since there are multiple layers of neurons, MLP is a deep learning technique.Deep learning is a branch of machine learning that makes use of multiple-layered artificial neural networks to model and resolve complicated issues.Developing an MLP model for water quality prediction, the data is collected,preprocessed, and trained. The performance of the model is evaluated using various metrics.Various activation functions are used in this predicting system.In this system used machine learning classification algorithms for the comparison. The machine learning classification algorithms used in this system are Decision Tree Classifier,Random Forest,SVM,KNN.These algorithms are used for comparing the performance with MLP model.

This system aims to improve the accuracy of water quality prediction through the artificial neural network (ANN) technology. A comprehensive dataset consisting of various water quality parameters is utilized to train the model. The model is designed to take into account the intricate relationships between the different water quality factors and their impact on overall water safety.Collected real-time data from various water sources, which is then fed into the MLP model for prediction. The results from the model are analyzed and compared with traditional methods to determine the efficacy of the proposed approach. The project's outcome is expected to provide a robust, data-driven solution for water quality management, which will help to ensure the safety of water sources and improve overall public health.

1.1 Problem Statement

- The problem statement for this project is to predict the water quality by handling large dataset and recognize the patterns efficiently with the use of MultiLayer Perceptron(MLP) Model

1.2 Objectives

The goal is to accomplish the following:

- To create a reliable and accurate system for predicting quality of water.
- To compare the performance of the MLP and Machine learning classification algorithms.
- To create an easy-to-use water quality predicting tool

Chapter 2

LITERATURE SURVEY

A literature survey, also known as a literature review, is a comprehensive and critical analysis of the existing research and literature on a particular topic. It involves gathering and analyzing relevant academic and scholarly articles, books, and other sources of information to identify the current state of knowledge on a specific research question or topic.When conducting a literature review from an audit perspective, the main focus is on evaluating the relevant literature. This process covers information that has been published in a specific field of study and sometimes includes information published within a specific time frame. The purpose of a literature survey is to provide an overview of the research and literature that has been published on a particular topic, to identify gaps in the existing knowledge, and to provide a context for the proposed research. A well-crafted literature review can also enhance the credibility and authority of the author, as it demonstrates their familiarity with the current research and debates in the field. In certain cases, a literature review may include a meta-analysis, which involves analyzing the findings of numerous studies to uncover common patterns or trends. It also allows the researcher to identify key theories, methodologies, and findings that have been established in the field and to evaluate the quality and relevance of previous studies.Overall, a literature survey is an important step in the research process as it helps researchers to identify the existing knowledge and research gaps in their field of study, and to develop a research question that is grounded in the existing literature.

2.1 Purpose of Literature Review

1. It identify key theories, concepts, methodologies, and findings that have been established in the field.

2. It provide a comprehensive overview of the current state of knowledge on a particular research topic or question.

3. It evaluate the quality and relevance of previous studies and to identify areas where

further research is needed.

4. A literature review can be a standalone piece or part of a larger research project such as a thesis, dissertation, or research paper

5. It identify potential challenges and limitations in the proposed research and to develop strategies to address them.

6. It identify gaps, controversies, and inconsistencies in the existing literature.

2.2 Related Works

Various studies pertrained to predict the water quality are listed below:

Hamza Ahmad Madni et.al[1] propose an approach for predting water quality prediction based on H2O AutoML and explainable AI Techniques.H2O AutoML is a tool that automates the machine learning process, allowing for faster and more efficient modeling. The use of H2O AutoML in water quality prediction help identify important variables and their relationships to water quality. Explainable AI techniques, on the other hand, aim to make machine learning models more transparent and understandable by humans. This can be important for applications such as water quality prediction where the decisions made by the model can have significant impacts on human health and the environment.To handle the accuracy problem,makes use of the stacked ensemble H2O AutoML model.The study specifies that the author uses KNN imputer to handle the missing values.Overall, the paper highlights the potential benefits of using H2O AutoML and explainable AI techniques in water quality prediction. By automating the modeling process and making models more transparent, these techniques can help identify important variables and improve the accuracy of predictions, ultimately leading to better management of water resources.

Ali Najah Ahmed et.al[2] describes a study that explores the use of machine learning techniques to improve water quality prediction.Author discusses the significance of modelling water quality parameters in aquatic systems and the limitations of traditional modelling methodologies due to the large amount of unknown or unspecified input data.The system which proposed is based on The Johor River Basin.Adaptive Neuro-Fuzzy Inference System (ANFIS), Radial Basis Function Neural Networks (RBF-ANN), and Multi-Layer Perceptron

Neural Networks (MLP-ANN) are proposed in this paper.The Author proposed three evaluation techniques, including assessing the significance of input parameters, constructing models using single or combinations of input parameters, and developing prediction models for each station or based on the value of the same parameter at the previous station (upstream).Overall, the study highlights the potential of AI-based techniques for water quality prediction, which can serve as a powerful tool for better water resource management

Amir Hamzeh Haghiabi et.al[3] ,the author aims to predict water quality parameters using machine learning methods in this paper.The study focuses on the Karkheh River Basin in Iran, where water quality is affected by human activities and natural factors. The study uses three machine learning algorithms, including Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT), to predict the water quality.Also evaluates the performance of each algorithm using performance evaluation metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Correlation Coefficient (R). The study found that the SVM algorithm performed better in predicting the water quality parameters compared to ANN and DT. The study concludes that machine learning methods, particularly the SVM algorithm, can be useful in predicting water quality parameters and can assist in effective water resource management.

Yuanyuan Wang et.al[4] proposes a method for predicting water quality using a Long Short-Term Memory (LSTM) neural network.The paper describes the water quality prediction problem and the importance of accurate and timely predictions in preventing water pollution.Author presents the LSTM model and how it can be used to predict water quality based on historical data. The proposed method was tested on real-world water quality data, and the results showed that the LSTM model can effectively predict water quality with high accuracy. The paper concludes that the proposed method has potential for practical applications in water quality prediction and management.Author proposed a useful contribution to the field of water quality prediction, and the proposed LSTM-based approach may be beneficial for water management organizations and other stakeholders involved in maintaining water quality.

Hongfang Lu et.al[5] proposes two novel hybrid decision tree-based machine learning models in this paper for obtain more accurate short-term water quality prediction results.The paper proposes two novel hybrid decision tree-based machine learning models, CEEMDAN-RF and CEEMDAN-XGBoost, for short-term water quality prediction.The results show that CEEMDAN-RF performs best in predicting water temperature, dissolved oxygen, and specific

conductance, while CEEMDAN-XGBoost performs best in predicting pH value, turbidity, and fluorescent dissolved organic matter.The author also discusses the stability of the prediction models and shows that CEEMDAN-RF and CEEMDAN-XGBoost have higher prediction stability than other benchmark models. The use of advanced data denoising technique CEEMDAN enhances the accuracy and stability of the hybrid models, making them a valuable tool for water quality prediction and management.The proposed hybrid decision tree-based machine learning models have the potential to improve the accuracy and stability of short-term water quality prediction

Jing Bi et.al[6] designs a work based on hybrid model by using an encoder-decoder neural network based on LSTM and a SG filter, named SE-LSTM, to predict the future water quality.Use of the Savitzky-Golay filter, which can remove any potential noise from time series of water quality, and the long short-term memory, which can look at nonlinear aspects in a complex water environment.The study highlights the potential of deep neural networks for large-scale water quality prediction and provides a framework for future research in this area.This work proposes an improved encoder-decoder network structure, with which the multi-step water quality time series data is better predicted. Thus, the proposed SE-LSTM can better handle long sequences in the time series data.Author innovatively combines and integrates the noise reduction ability of the SG filter, and the feature extraction ability of LSTM to significantly improve the multi-step prediction accuracy.

Tianan Deng et.al[7] describes a study that aims to predict marine water quality for coastal hydro-environment management using Machine learning.Two different ML methods – artificial neural networks (ANN) and support vector machine (SVM) are implemented and improved by introducing different hybrid learning algorithms.The system findings demonstrate the good applicability and accuracy of these two machine learning techniques for predicting the trend and amount of algal development.The results specifically show that ANN is better to provide satisfactory results with prompt response, while SVM is suited to precisely identify the ideal model while requiring more training time.Author describes that the applied ML techniques may provide robustness to learn complex relationships between algal dynamics and various coastal environmental variables and, as a result, to reliably identify significant variables. The study's findings, after analysis and discussion, also point to the advantages and potentials of the used ML models to shed light on the mechanisms underlying the spread and evolution of the HAB.The author highlights the potential of machine learning algorithms for environmental

monitoring and management, and provides insights into how these technologies can be applied to improve the quality

Elias Dritsas et.al[8] propose efficient data-driven machine learning models for Water Quality prediction.In this study, To create the desired model with the maximum accuracy, a comparative evaluation of a variety of machine learning (ML) classification models, including Naive Bayes, LR, kNN,Ensemble Learning (RF, AdaBoostM1, Stacking, Voting, and Bagging), was conducted.The study also investigates the use of feature selection techniques to identify the most important input variables for the models, in order to improve their accuracy and efficiency. The results of the study indicate that machine learning models can be effective in predicting water quality parameters, with the neural network model achieving the highest accuracy and efficiency. Author also discuss sbout the future work in this paper.

Ann Laverene Lopez et.al[9] describes a new approach for the prediction of water quality.Author combines Machine Learning and IOT.The objective of the research is to develop a water quality prediction system that utilizes real-time data collected from IoT sensors placed in water bodies. The study makes use of data collected from a lake and applies LSTM neural networks to develop the predictive model.The results of the study indicate that the LSTM neural network model is effective in predicting water quality parameters. The author highlights the potential of IoT and LSTM neural networks for water quality prediction and monitoring, and provides insights into how these technologies can be applied to improve the management of water resources.The key advantage of this method is that consumers can be alerted in advance if there is a risk of pollution, enabling them to sanitise the water before it gets contaminated.

Jin-Won Yu et.al[10] proposes water quality forecasting based on data decomposition, fuzzy clustering and deep learning neural network.The author proposes a novel hybrid model by combining data decomposition, fuzzy C-means clustering and bidirectional gated recurrent unit.The use of empirical wavelet transform, fuzzy C-means clustering, and bidirectional gated recurrent units allows for accurate and efficient prediction of water quality parameters.FCM is employed to reduce the unnecessary errors caused by real-time data decomposition.Results show that the proposed model provides highly accurate forecast results.

Di Wu et.al[11] proposes smart data driven quality prediction for urban water source management.Instead of data output from water treatment,collect the raw water data directly from water sources.Adaptive learning rate BP neural network (ALBP) and, 2-step isolation and random forest are proposed in this study.The results show that ALBP is theoretically simple and

easy to implement. 2sIRF considers the risk distribution and shows higher prediction accuracy. In addition,the system perform the correlation analysis of all the indicators and the importance analysis over different indicators.Author confirmed that this work is meaningful for future risk control and decision support in urban water supply systems.

Hong-Gui Hanet.al[12] introduce an efficient self-organizing RBF neural network for water quality prediction.The study addresses the challenge of predicting water quality, which is important for maintaining the safety and sustainability of water resources.The author proposed SO-RBF neural network is designed to learn patterns and relationships in water quality data without prior knowledge of the data's underlying structure. The network uses a self-organizing approach to adaptively adjust the number and location of radial basis functions based on the distribution of input data.The author also evaluates the performance of the SO-RBF network on a dataset of water quality measurements from a river in China. The results show that the SO-RBF network outperforms other traditional neural networks, such as the Back Propagation (BP) neural network, in terms of accuracy and efficiency.

Hieda Adriana Nascimento Silva et.al[13] describes artificial neural networks and remote sensing methods for predictiong water quality.These techniques can be used to estimate water quality variables.The proposed approach uses WNNs to learn patterns and relationships in water quality data, and remote sensing data to supplement and improve the accuracy of the predictions. Remote sensing data provides information on environmental factors such as water temperature, water color, and suspended solids, which can affect water quality. The WNNs are designed to model the complex relationships between these factors and water quality variables.The results show that the proposed approach outperforms other traditional neural networks, such as the Back Propagation (BP) neural network, in terms of accuracy and efficiency. The remote sensing data significantly improves the accuracy of the predictions.

Chapter 3

METHODOLOGY

WATER QUALITY PREDICTION USING ARTIFICIAL NEURAL NETWORK is a prediction system for predicting quality of water.The system is used to predict the water quality using MultiLayer Perceptron Neural Network model.

3.1 Algorithm

- Step 1:Import necessary libraries
  - Step 1.1:Load dataset
  - Step 1.2:Data pre-processing
  - Step 1.3:Build MLP model
  - Step 1.4:Training the model
  - Step 1.5:Testing the model
  - Step 1.6:Predicting water quality
- Step 2:
  - Step 2.1:Load dataset
  - Step 2.2:Data pre-processing
  - Step 2.3:Feature Extraction
  - Step 2.4:Build machine learning classification algorithms(SVM,KNN,Decision Tree,Random Forest)
  - Step 2.5:Compare each algorithm and find best one.
- Step 3:Comparison study of best machine learning algorithm and MLP
- Step 4:Result

3.2 System Architecture

Water quality has a direct impact on public health and the environment. Water is used for various practices, such as drinking, agriculture, and industry.During the last years, water quality has been threatened by various pollutants. Therefore, modeling and predicting water quality have become very important in controlling water pollution.This project predicts water quality Using Artificial Neural Network. The proposed system of water quality prediction uses MultiLayer Perceptron Model and other Machine Learning classification algorithms.

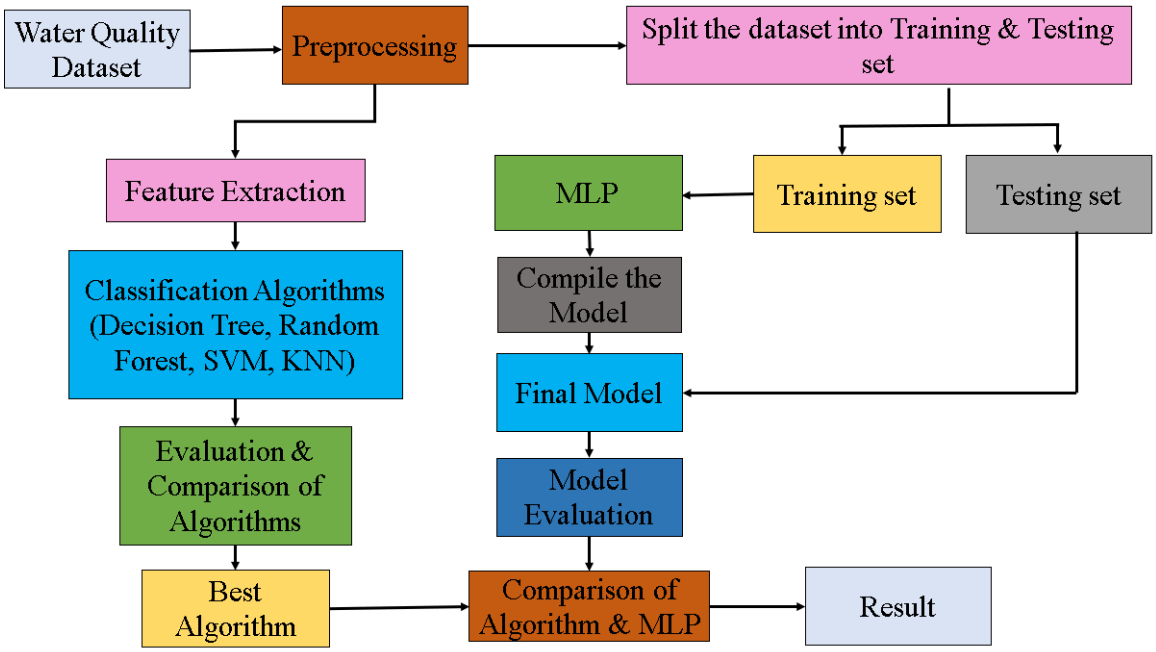

Figure 3.1: System Architecture

3.2.1 Dataset

WaterQuality dataset is used in this project.This dataset is used to train and evaluate models for the prediction of water qaulity and other water content based systems.This dataset consist of 20 features and 8000 corresponding records. The features which is contained in the dataset are Aluminium,Arsenic,Barium, Cadmium,Chloramine,Chromium,Copper,Flouride,Bacteria, Viruses,Lead,Nitrates,Nitrites,Mercury,Perchlorate,Radium,Selenium,Silver and Uranium.

aluminium	ammonia	arsenic	barium	cadmium	chloramin	chromium	copper	flouride	bacteria	viruses	lead	nitrates	nitrites	mercury	perchlorat	radium	selenium	silver	uranium	is_safe
1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17	0.05	0.2	0	0.054	16.08	1.13	0.007	37.75	6.78	0.08	0.34	0.02	1
2.32	21.16	0.01	3.31	0.002	5.28	0.68	0.66	0.9	0.65	0.65	0.1	2.01	1.93	0.003	32.26	3.21	0.08	0.27	0.05	1
1.01	14.02	0.04	0.58	0.008	4.24	0.53	0.02	0.99	0.05	0.003	0.078	14.16	1.11	0.006	50.28	7.07	0.07	0.44	0.01	0
1.36	11.33	0.04	2.96	0.001	7.23	0.03	1.66	1.08	0.71	0.71	0.016	1.41	1.29	0.004	9.12	1.72	0.02	0.45	0.05	1
0.92	24.33	0.03	0.2	0.006	2.67	0.69	0.57	0.61	0.13	0.001	0.117	6.74	1.11	0.003	16.9	2.41	0.02	0.06	0.02	1
0.94	14.47	0.03	2.88	0.003	0.8	0.43	1.38	0.11	0.67	0.67	0.135	9.75	1.89	0.006	27.17	5.42	0.08	0.19	0.02	1
2.36	5.6	0.01	1.35	0.004	1.28	0.62	1.88	0.33	0.13	0.007	0.021	18.6	1.78	0.007	45.34	2.84	0.1	0.24	0.08	0
3.93	19.87	0.04	0.66	0.001	6.22	0.1	1.86	0.86	0.16	0.005	0.197	13.65	1.81	0.001	53.35	7.24	0.08	0.08	0.07	0
0.6	24.58	0.01	0.71	0.005	3.14	0.77	1.45	0.98	0.35	0.002	0.167	14.66	1.84	0.004	23.43	4.99	0.08	0.25	0.08	1
0.22	16.76	0.02	1.37	0.007	6.4	0.49	0.82	1.24	0.83	0.83	0.109	4.79	1.46	0.01	30.42	0.08	0.03	0.31	0.01	1
3.27	3.6	0.001	2.69	0.005	5.75	0.15	0.6	1.29	0.04	0.008	0.145	8.47	1.25	0.006	55.4	7.8	0.05	0.33	0.06	0
1.35	21.96	0.04	0.84	0.002	0.1	0.76	0.17	0.58	0.52	0.52	0.011	18.4	1.49	0.009	21.52	1.3	0.08	0.48	0.08	1
1.88	19.26	0.02	2.78	0.008	0.05	0.42	1	0.09	0.91	0.91	0.103	4.37	1.95	0.006	22.12	1.97	0.03	0.06	0.05	1
4.93	23.98	0.04	3.05	0.008	0.7	0.51	1.35	1.07	0.7	0.7	0.101	1.16	1.11	0.008	26.8	5.58	0.09	0.38	0.03	1
2.89	18.82	0.05	3.77	0.008	5.99	0.54	0.79	0.54	0.2	0.009	0.126	17.56	1.82	0.009	17.54	4.33	0.1	0.05	0.02	1
0.61	2.41	0.03	0.59	0.002	1.94	0.77	1.54	0.62	0.23	0.001	0.017	1.99	1.08	0.007	11.16	0.98	0.01	0.47	0.03	1
3.47	15.84	0.02	0.06	0.001	5.29	0.47	1.08	1.43	0.89	0.89	0.08	1.91	1.2	0.008	0.18	6.89	0.06	0.12	0.08	1
2.11	17.03	0.02	0.88	0.009	7.78	0.88	1.15	0.34	0.85	0.85	0.065	17.86	1.53	0.003	19.4	1.14	0.1	0.4	0.01	1
4.88	26.94	0.02	0.36	0.001	1.21	0.68	0.71	0.99	0.75	0.75	0.071	0.31	1.22	0.002	56.7	1	0	0.41	0.05	0
4.12	17.99	0.02	3.43	0.006	0.01	0.41	1.82	0.22	0.99	0.99	0.108	8.06	1.76	0.005	24.29	0.88	0.1	0.1	0.07	1
0.68	18.99	0.001	0.04	0.006	4.57	0.2	1.18	1	0.92	0.92	0.086	9.46	1.41	0.007	21.79	3.05	0.03	0.13	0.08	1
1.15	8.12	0.02	0.97	0.007	3.47	0.65	1.51	1.46	0.58	0.58	0.061	8.96	1.5	0.004	14.6	1.74	0.03	0.01	0.06	1
0.27	10.67	0.02	0.55	0.001	3.74	0.12	1.77	0.43	0.8	0.8	0.114	12.69	1.18	0.008	34.64	0.9	0.02	0.16	0.06	1
4.32	20.64	0.03	2.6	0.008	7.24	0.61	1.23	1.44	0.56	0.56	0.012	9.42	1.74	0.004	36.23	3.22	0.07	0.18	0.08	0
2.36	27.05	0.01	0.68	0.003	4.07	0.13	1.34	0.29	0.96	0.96	0.167	15.05	1.92	0.002	56.32	7.99	0.06	0.5	0.06	0
3.31	22.07	0.03	0.46	0.001	7.22	0.73	1.05	1	0.25	0.007	0.109	1.92	1.07	0.001	39.4	0.49	0.04	0.47	0.05	1

Figure 3.2: Dataset

3.2.2 Data Pre-processing

Data pre-processing is a critical step in deep learning because the quality of the data used to train a model has a significant impact on its performance.Data pre-processing in deep learning refers to the process of preparing and transforming raw data into a format suitable for training a machine learning model. It involves cleaning, formatting, and organizing the data to make it more meaningful and relevant for the model.Poor quality data can lead to inaccurate predictions and flawed models.Data pre-processing takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

In this stage,the input data is pre-processed by doing things like null values checking,removing unwanted entries,changing data types and balancing dataset. Here removing unwanted entries and data type changing is done by using numpy technique.For balancing the dataset oversampling is performed by using Synthetic Minority Oversampling Technique (SMOTE) technique.Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of cases in the dataset in a balanced way. The component works by generating new instances from existing minority cases that the supply as input. SMOTE works by creating synthetic examples of the minority class by interpolating between existing minority class samples. It randomly selects a minority class sample and then finds its k nearest neighbors in the feature space. It then generates new synthetic samples along the line segments joining the minority sample and its neighbors. The number of new samples generated is determined by a user-defined parameter that specifies the desired ratio of minority to majority class samples.

3.2.3 Building and Training the MultiLayer Perceptron Model

A MultiLayer Perceptron (MLP) is a type of artificial neural network (ANN) that consists of multiple layers of interconnected perceptrons (neurons). Each perceptron receives inputs from the previous layer, applies a weighted sum of these inputs, and passes the result through an activation function to produce its output.MLPs are commonly used for supervised learning tasks such as classification and regression. During training, the model adjusts the weights of the connections between neurons in order to minimize the difference between its predictions and the true labels or values of the training data. This process is typically done using backpropagation, where the error is propagated backwards through the network and the weights are updated accordingly.

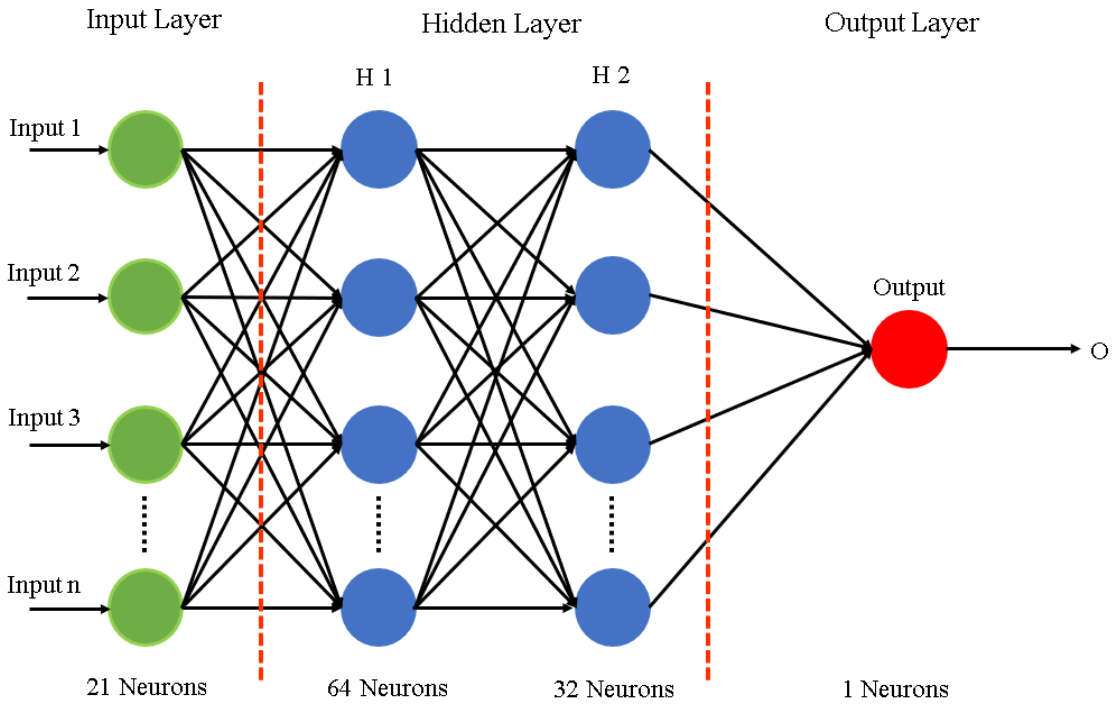

Figure 3.3: MultiLayer Perceptron Model

In this system creates a Sequential model for a binary classification problem using the Keras library.Keras is used for distributed training of deep learning models.The Sequential function from the tensorflow.keras module is used to initialize a new, empty sequential model. The list of layers that make up the model is passed as an argument to the Sequential function.

The model consists of six layers: three dense layers, two dropout layers, and one output layer. The dense layers are fully connected neural network layers. The first dense layer has

21 neurons and a ReLU(Rectified Linear Unit) activation function.The dropout layer has a rate of 0.2, which means that 20 percentage of the neurons will be dropped out during training.The second dense layer has 64 neurons and a ReLU activation function. The third dense layer has 32 neurons and a ReLU activation function. The dropout layer after the third dense layer has a rate of 0.1.The final layer is the output layer.ReLu is a non-linear activation function.The model trained with ReLU converged quickly and thus takes much less time when compared to models trained on other activation functions.It has 1 neuron and a sigmoid activation function, which is appropriate for binary classification problems. The sigmoid activation function outputs a probability between 0 and 1, which can be interpreted as the probability that the input data belongs to the positive class. This MLP model with a structure that can be fine-tuned for the specific binary classification problem at hand. The number of neurons, activation functions, and dropout rates can be adjusted to improve the performance of the model.

ADAM optimizer and Binary Cross-Entropy loss functions are used for the model compilation.An optimizer is a function or an algorithm that modifies the attributes of the neural network, such as weights and learning rates.It helps in reducing the overall loss and improving accuracy.The proposed system uses ADAM optimizer,it is a type of gradient-based optimization algorithm that adjusts the weights of the model in order to minimize the loss function.This optimization algorithm is a further extension of stochastic gradient descent to update network weights during training.Unlike maintaining a single learning rate through training in SGD, Adam optimizer updates the learning rate for each network weight individually.

The main advantage of using ADAM optimizer is Straightforward to implement,computationally efficient and little memory requirements.ADAM optimizer is well suited for problems that are large in terms of data or parameters.Loss function is a method of evaluating how well the algorithm models the dataset.Binary cross entropy is a commonly used loss function for binary classification problems.The binary cross-entropy loss measures the dissimilarity between the true label and the predicted probability of the positive class.

3.2.4 Testing the Model

The 20% of the dataset for testing. For training, the remaining 80% of the dataset will be used. After testing, the MLP model obtained 95.73 accuracy rating and 96 validation rating.

3.2.5 Feature Extraction for Machine Learning Classificaion

From the pre-processing stage takes the dataset for feature extraction technique. Before training the classification algorithm the the dataset should be perform feature extration process.Feature extraction refers to the process of extracting meaningful and relevant information or features from raw data.It is a crucial step in various fields, including signal processing,computer vision, natural language processing, and machine learning.In the context of machine learning, feature extraction involves transforming raw data into a representation that captures important characteristics of the data.The extracted features are usually more informative and suitable for the learning algorithms to make accurate predictions or classifications.By extracting relevant features, the dimensionality of the data can be reduced, removing redundant or irrelevant information and improving computational efficiency.

PCA-Pricipal Component Analysis

PCA (Principal Component Analysis) is used for feature extraction process in classification algorithms.PCA (Principal Component Analysis) is a widely used feature extraction technique that aims to reduce the dimensionality of data while preserving the most important information.It achieves this by transforming the original features into a new set of orthogonal variables called principal components.By selecting a subset of the principal components, PCA effectively selects the corresponding original features that contribute the most to the variance in the dataset. These selected features are the ones that carry the most information and are typically used for further analysis or modeling.

3.2.6 Building different Classification Algorithms

Classification algorithms used in machine learning utilize input training data for the purpose of predicting the likelihood or probability that the data that follows will fall into one of the predetermined categories.

Decision Tree Classifier

A decision tree classifier is a type of machine learning algorithm that is particularly well-suited for classification tasks, where the goal is to predict a categorical target variable based on one or more input variables.In the context of water quality prediction, a decision tree classifier can

be used to predict whether a particular sample of water meets certain quality standards or not, based on the input data.

Gini used to evaluate the quality of a split in a Decision Tree.These measures are used to determine which feature or attribute of the data should be used to split the data at a given node in the decision tree.Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. In the context of a decision tree, the Gini impurity is used to evaluate how well a given split separates the data into classes.The Gini impurity of a node is defined as:Gini impurity = 1 - sum(p_i^2)

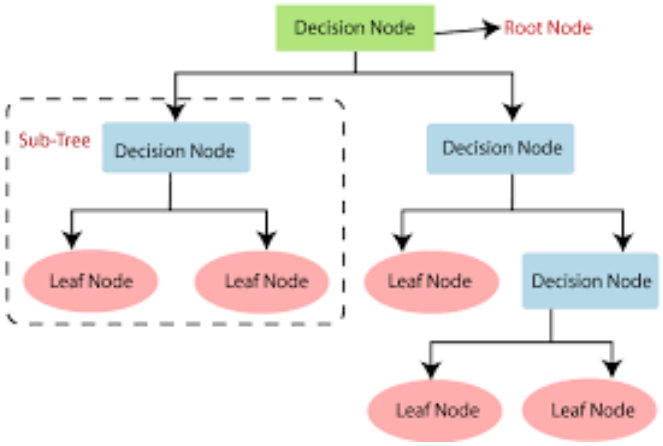

Figure 3.4: Decision Tree Classifier

For the Decision Tree classifier criterion used is gini,maximum depth of the tree is 5 and splitter is best.The water quality prediction model based on the decision tree classifier can be another effective tool for identifying potential contaminants and ensuring the safety of drinking water.The decision tree model is capable of handling both numerical and categorical data, making it a flexible and versatile.By using various water quality parameters as input features,the decision tree model can accurately classify water samples as safe or unsafe for consumption. By using Decision Tree claasification algorithm model predicts the water quality system with 90% accuracy.

Random Forest

Random Forest is a popular machine learning algorithm used for both classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to create a more accurate and robust model.In Random Forest, each tree is grown using a different

subset of the features, selected randomly.The model aggregates the predictions of the individual trees to arrive at a final prediction.This technique helps reduce overfitting and improve the model's generalization ability.

While training the Radom Forest Gini used as criterion and maximum depth is 5. Train the algorithm using the extracted data and evaluate the algorithm. Once the model is trained with the best parameters it predicts the test data using the trained model and calculates the accuracy score of the predicted values.Random forest classification algorithms gives 88% accuracy rate for the prediction system.

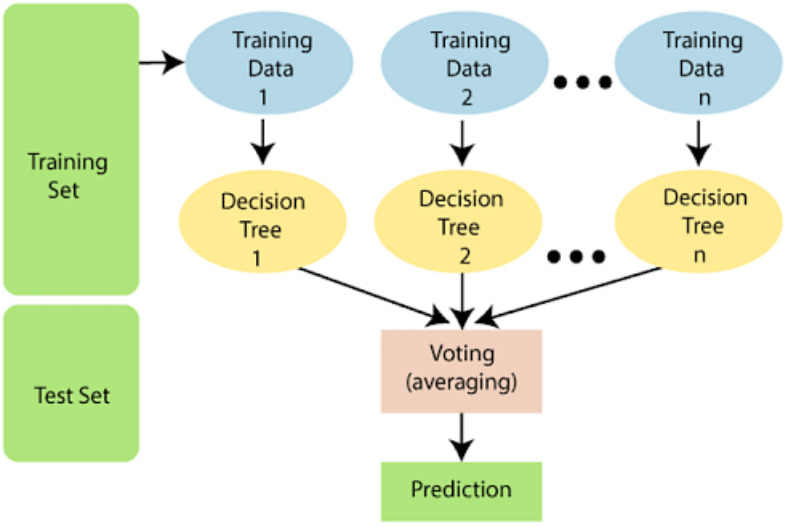

Figure 3.5: Random Forest

Support Vector Machine

Support Vector Machine (SVM) is a popular and powerful machine learning algorithm used for both classification and regression tasks.In SVM, the data points are mapped into a high-dimensional space using a kernel function. The hyperplane that best separates the data into different classes is then chosen by maximizing the margin, which is the distance between the hyperplane and the closest points from each class.SVM have several advantages like it can handle both linear and non-linear data by choosing the appropriate kernel function, effective in high-dimensional spaces and can handle datasets with many features.

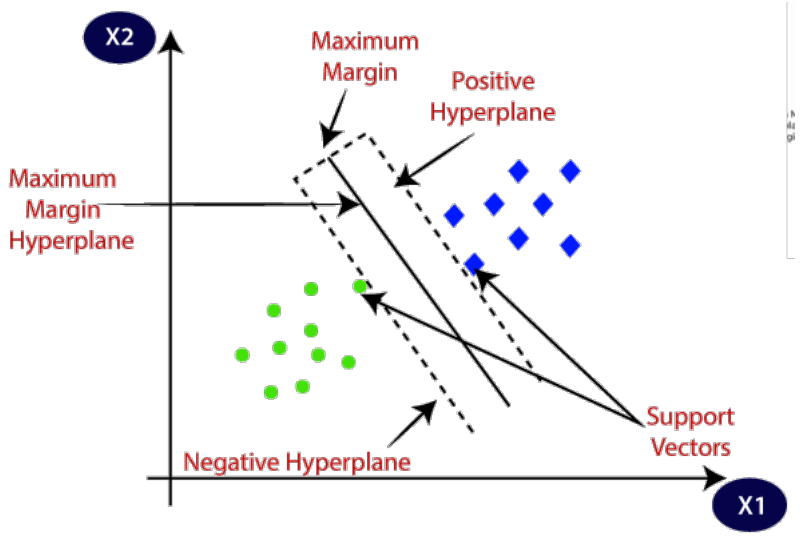

Figure 3.6: Support Vector Machine

For training the Support vector Machine algorithm regularization parameter C is used. Regularization parameter is controls the tradeoff between having a low training error and a low testing error.Linear, RBF and sigmoid are the most commonly used kernels,these are used for buliding the system.During the training of SVM Radial Basis Function (RBF) kernel is used.It is a popular kernel function used when the data is not linearly separable and degree of kernal assigned as 2.The usage of Shrinking Heuristic technique method it speed up the training process and improve its performance. 87% of accuracy rate obtained by using SVM classification algorithm.

K Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple and effective machine learning algorithm used for classification and regression tasks. KNN is a non-parametric algorithm that does not make any assumptions about the underlying distribution of the data.In KNN, the training data consists of labeled instances (i.e., data points with known classifications). The algorithm is called ”k-nearest neighbors” because it looks at the k closest labeled instances in the training data to the data point that needs to be classified, and then assigns the class label that is most common among those k neighbors.

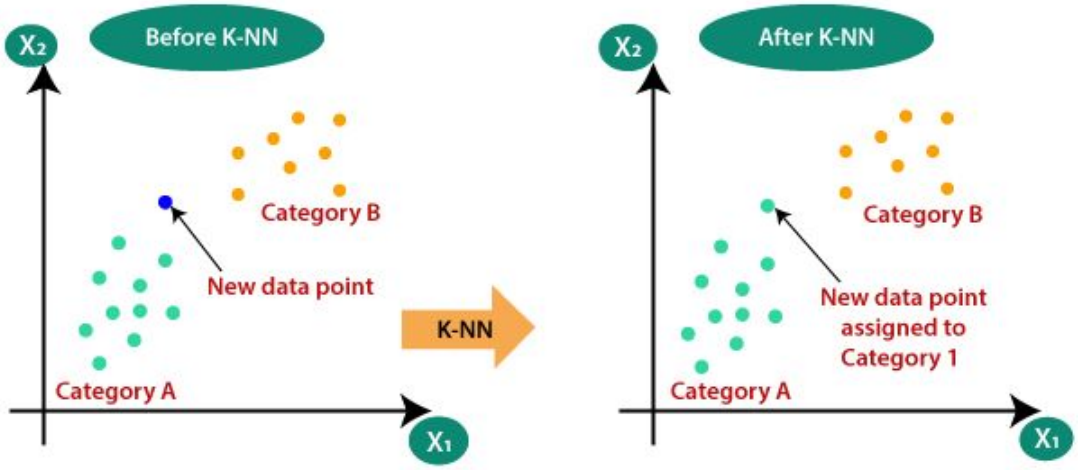

Figure 3.7: K Nearest Neighbors

KNN algorithm uses weight function ”Uniform”. ‘uniform’ means all points in each neighborhood are weighted equally.The algorithm used to compute nearest neighbors is ’ball tree’.Ball tree algorithm is efficient approach that uses a hierarchical structure of nested hyperspheres to partition the training set. The algorithm starts with a single hypersphere that encompasses all the data points, and then recursively subdivides it into smaller hyperspheres until each contains a fixed number of data points.For calculate the distance between two points using ’euclidean’ metric.Euclidean distance is a commonly used distance metric in k-Nearest Neighbors (k-NN) algorithm to measure the similarity or dissimilarity between data points. It calculates the straight-line distance between two points in a multi-dimensional space.

The Euclidean distance between two points, denoted as p and q, in an n-dimensional space can be calculated using the following formula:

$d(p, q) = \text{sqrt}((x1 - x2)^2 + (y1 - y2)^2 + ... + (zn - zm)^2)$

After the training and testing KNN algorithm predicts the water quality system with 85% accuracy.

3.3 Software Requirements and Specifications

3.3.1 Python

Python is a popular high-level programming language that is widely used in many different applications, including web development, scientific computing, data analysis, artificial intelligence, machine learning, and more. It was first released in 1991 by Guido van Rossum, and has since become one of the most widely used programming languages in the world.

One of the key features of Python is its simplicity and ease of use. Python code is typically easy to read and understand, making it a great language for both beginners and experienced developers. Python's syntax is concise and expressive, allowing developers to write code quickly and efficiently.Python is also known for its flexibility and versatility. It can be used for many different programming tasks, from simple scripts to complex applications. Python has a large and active community of developers who contribute to its libraries and frameworks, making it a powerful tool for many different applications.

Another important feature of Python is its support for multiple programming paradigms. Python can be used for procedural, object-oriented, and functional programming, giving developers the flexibility to choose the programming style that best suits their needs.Python is also cross-platform, meaning that code written in Python can run on many different operating systems, including Windows, macOS, Linux, and more. This makes it a great choice for developing applications that need to run on multiple platforms.Python has a rich ecosystem of libraries and frameworks that make it easy to develop a wide range of applications. Some of the most popular Python frameworks include Django, Flask, and Pyramid for web development, and NumPy, Pandas, and SciPy for scientific computing and data analysis.In addition to its many features and capabilities, Python is also an open-source language, which means that its source code is available for anyone to use, modify, and distribute. This has helped to build a large and active community of developers who contribute to the language's development and support.

3.3.2 Google Colaboratory

Google Colaboratory, also known as Google Colab, is a cloud-based platform that allows users to run and execute code in a Jupyter notebook environment using Google's infrastructure. Colab

provides users with free access to computing resources, including GPU and TPU, which are used to accelerate machine learning and deep learning algorithms.Google Colab is widely used by researchers, developers, and data scientists for machine learning projects, data analysis, and educational purposes. It allows users to easily share their code with others and collaborate on projects in real-time. Additionally, Colab provides pre-installed libraries, such as TensorFlow and PyTorch, making it easy to get started with machine learning and deep learning.

Colab notebooks can be saved in Google Drive, making it easy to access and share them with others. Users can also import and export notebooks to and from GitHub, making it easy to collaborate with others on shared projects. Colab also provides integration with Google Sheets, allowing users to import and analyze data directly from a spreadsheet.Overall, Google Colaboratory is a powerful and user-friendly platform that provides a great environment for data analysis, machine learning, and collaborative work.

3.3.3 Django

Django is a free and open-source web framework for building web applications. It is written in Python and follows the Model-View-Controller (MVC) architectural pattern. Django was developed to help developers create web applications quickly and easily by providing a set of tools and features to handle common web development tasks.Object-Relational Mapping (ORM) system is one of the main feature of Django,which allows developers to interact with a database using Python code, rather than writing SQL queries directly.ORM makes it easier to manage data and reduces the likelihood of errors.Django also provides a robust admin interface that allows developers to easily manage and modify data in the database. Another key feature of Django is its built-in security features. Django provides protection against common web application vulnerabilities, such as cross-site scripting (XSS) and cross-site request forgery (CSRF). Additionally, Django has a built-in authentication system that can be used to secure access to different parts of the application.Django is also known for its scalability and versatility. It can be used to build small applications, as well as large-scale, complex applications. It is also flexible enough to be used for a variety of applications.

3.4 Hardware and Experimentation Environment

The hardware used for the experiments includes Windows 11 OS, 64-bit operating system, x64-based processor, Processor Intel(R) Core(TM) i3-10110U CPU @ 2.10GHz,, 8 GB RAM. The experimental environment was prepared by using Python 3.11 programming language. Framework used is Keras with TensorFlow.

Chapter 4

RESULT AND DISCUSSION

Water quality prediction can be done using a variety of approaches, including statistical models, physical models, and machine learning algorithms.The MultiLayer Perceptron model has been quite successful in predictong water quality. The efficiency of the MultiLayer Perceptron model is rated high. After the model is constructed, optimized using Adam optimizer and it's compiled using a categorical binary cross entropy loss function.The dropout method was used for reduce the overfitting of neural networks after fully connected layers.

4.1 Training and Validation Results

In terms of outcomes, research has demonstrated that artificial neural network model can predict water quality with high accuracy.In one case, for instance MultiLayer Perceptron model to predict water quality had 95.73% success rate.

For Decision Tree Classifier algorithm had 90.37% success rate,Random Forest had 88.43% success rate,Knn had 85.5% success rate and finally SVM had 87.68% success rate.The MLP model is trained for 100 epochs,in the training phase 337 steps per epoch for training was employed. Thus, training for 100 such epochs yielded optimized results, with high accuracy of 95.73% for the MultiLayer Perceptron model.While comparing all machine learning classification algorithms decision tree had the highest accuracy rate.

4.2 Performance Metrics for Validation Phase

A machine learning or deep learning project's validation phase involves evaluating the model's performance using a variety of performance indicators. The particular challenge and the kind of model being assessed determine the performance metrics to be used.In Water Quality Prediction system use Accuracy,Precision,Recall and F1-score as performance metrics.

The most used performance metric for categorization issues is accuracy. It calculates the proportion of accurate predictions the model made based on a specific set of data.Precision is a performance indicator that counts the proportion of accurate positive predictions among all the model's positive forecasts. It is used to assess how well the model can detect real positive cases. Recall is a performance indicator that counts the proportion of accurate positive predictions among all real positive instances in the data. It is used to gauge how effectively the model can locate each positive case.F1-score,The harmonic mean of recall and precision is known as the F1-score. It is used to assess how well the model's predictions strike a balance between precision and recall.

Recall= TruePositives / (TruePositives + FalseNegatives)

Precision=TruePositives / (TruePositives + FalsePositives)

F1 Score=2*((precision*recall)/(precision+recall))

Classification Report of MLP

	precision	recall	f1-score	support
0	0.99	0.92	0.96	1401
1	0.93	0.99	0.96	1433
accuracy			0.96	2834
macro avg	0.96	0.96	0.96	2834
weighted avg	0.96	0.96	0.96	2834

Classification Report of Decision Tree

	precision	recall	f1-score	support
0	0.93	0.97	0.95	1400
1	0.67	0.46	0.54	200
accuracy			0.90	1600
macro avg	0.80	0.71	0.74	1600
weighted avg	0.89	0.90	0.90	1600

Classification Report of Random Forest

	precision	recall	f1-score	support
0	0.88	1.00	0.94	1400
1	1.00	0.07	0.14	200
accuracy			0.88	1600
macro avg	0.94	0.54	0.54	1600
weighted avg	0.90	0.88	0.84	1600

Classification Report of SVM

	precision	recall	f1-score	support
0	0.88	1.00	0.93	1400
1	1.00	0.01	0.03	200
accuracy			0.88	1600
macro avg	0.94	0.51	0.48	1600
weighted avg	0.89	0.88	0.82	1600

Classification Report of KNN

	precision	recall	f1-score	support
0	0.89	0.95	0.92	1400
1	0.33	0.15	0.21	200
accuracy			0.85	1600
macro avg	0.61	0.55	0.57	1600
weighted avg	0.82	0.85	0.83	1600

4.2.1 Confusion Matrix

A confusion matrix is a performance evaluation tool that is commonly used to assess the accuracy of a machine learning model.It is a table that summarizes the number of correct and incorrect predictions made by a model on a set of data.For binary classification issues, when the model is trained to predict one of two possible outcomes, the confusion matrix is especially helpful.A confusion matrix is typically represented in a table format, where the rows correspond to the actual (true) labels, and the columns correspond to the predicted labels.

The four quadrants of the table represent the following outcomes:

- True Positive (TP): The model correctly predicted a positive (yes) outcome when the actual outcome was positive.
- False Positive (FP): The model incorrectly predicted a positive (yes) outcome when the actual outcome was negative.
- True Negative (TN): The model correctly predicted a negative (no) outcome when the actual outcome was negative.
- False Negative (FN): The model incorrectly predicted a negative (no) outcome when the actual outcome was positive.

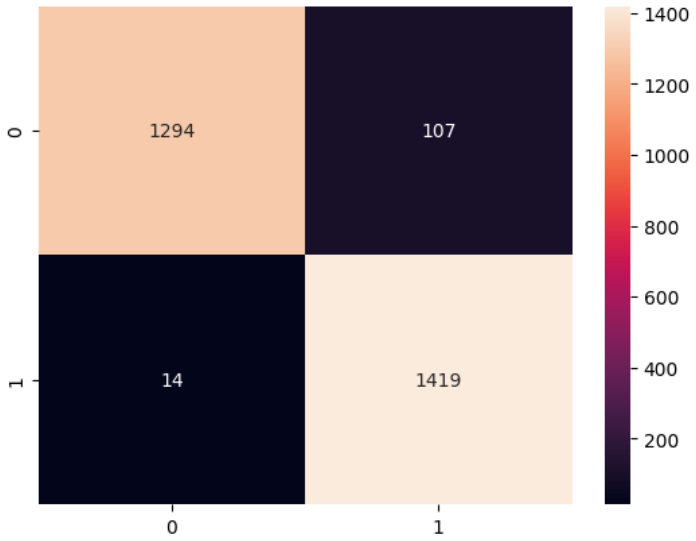

Figure 4.1: Confusion Matrix of MultiLayer Perceptron Model

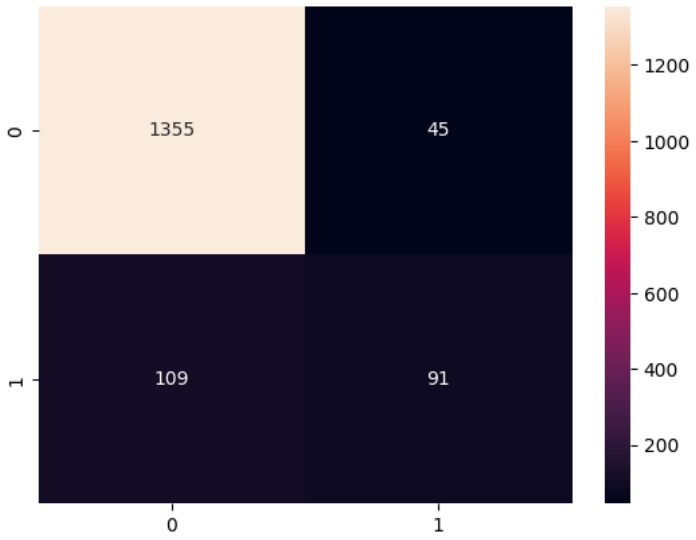

Figure 4.2: Confusion Matrix of Decision Tree Classifier

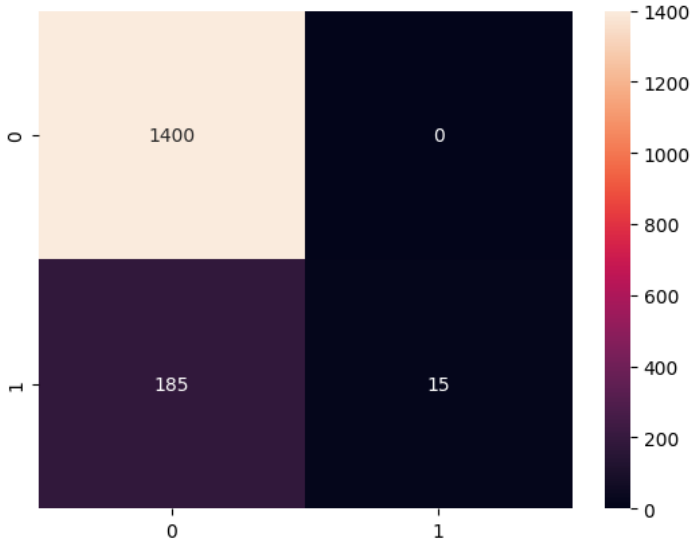

Figure 4.3: Confusion Matrix of Random Forest

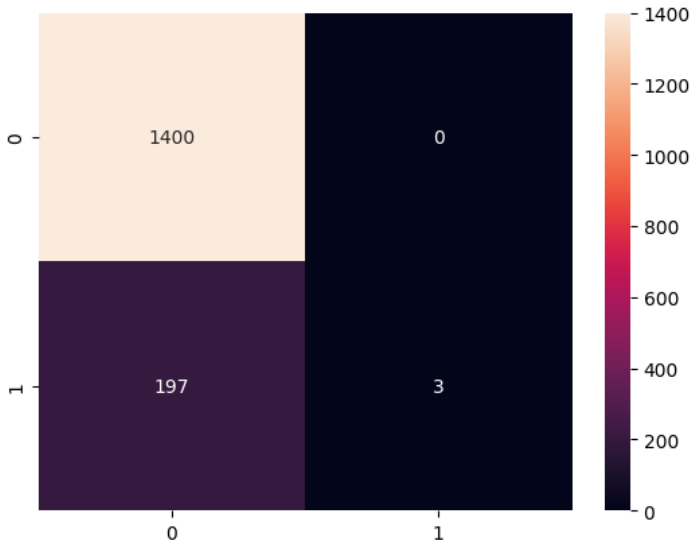

Figure 4.4: Confusion Matrix of SVM

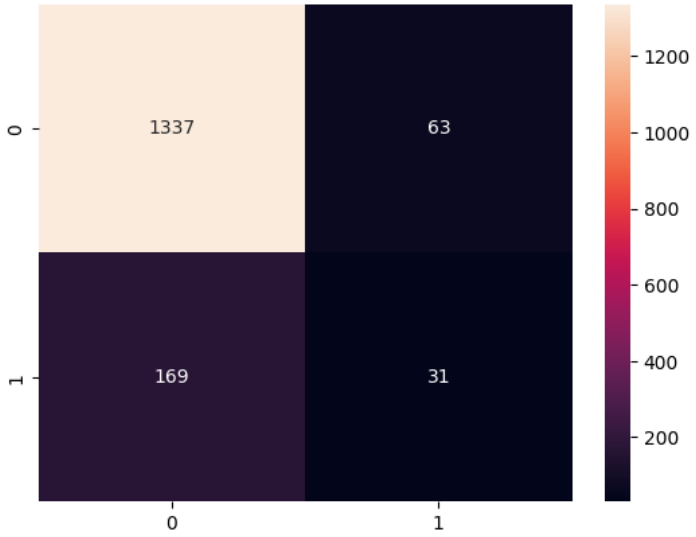

Figure 4.5: Confusion Matrix of KNN

4.2.2 Results

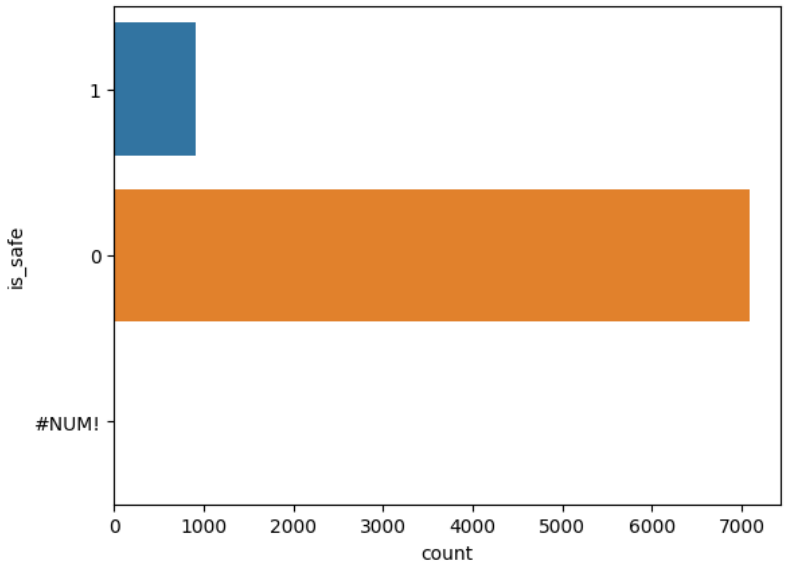

Figure 4.6: Inbalanced Dataset

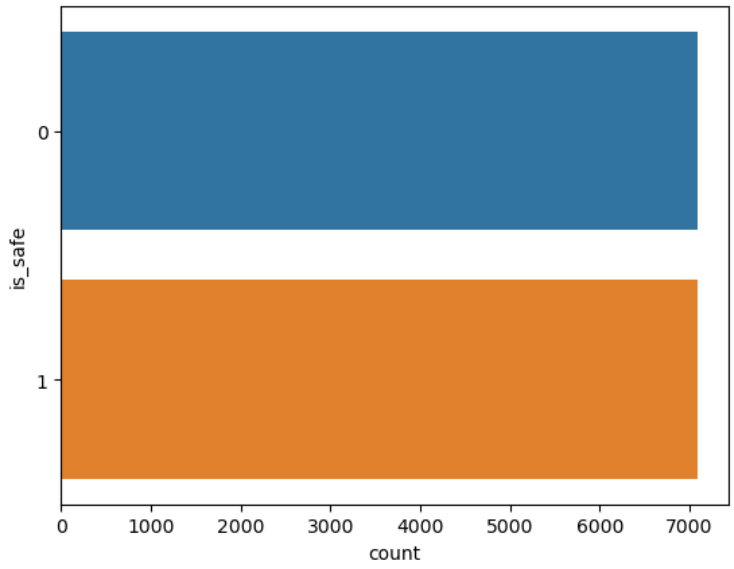

Figure 4.7: Balanced Dataset after SMOTE

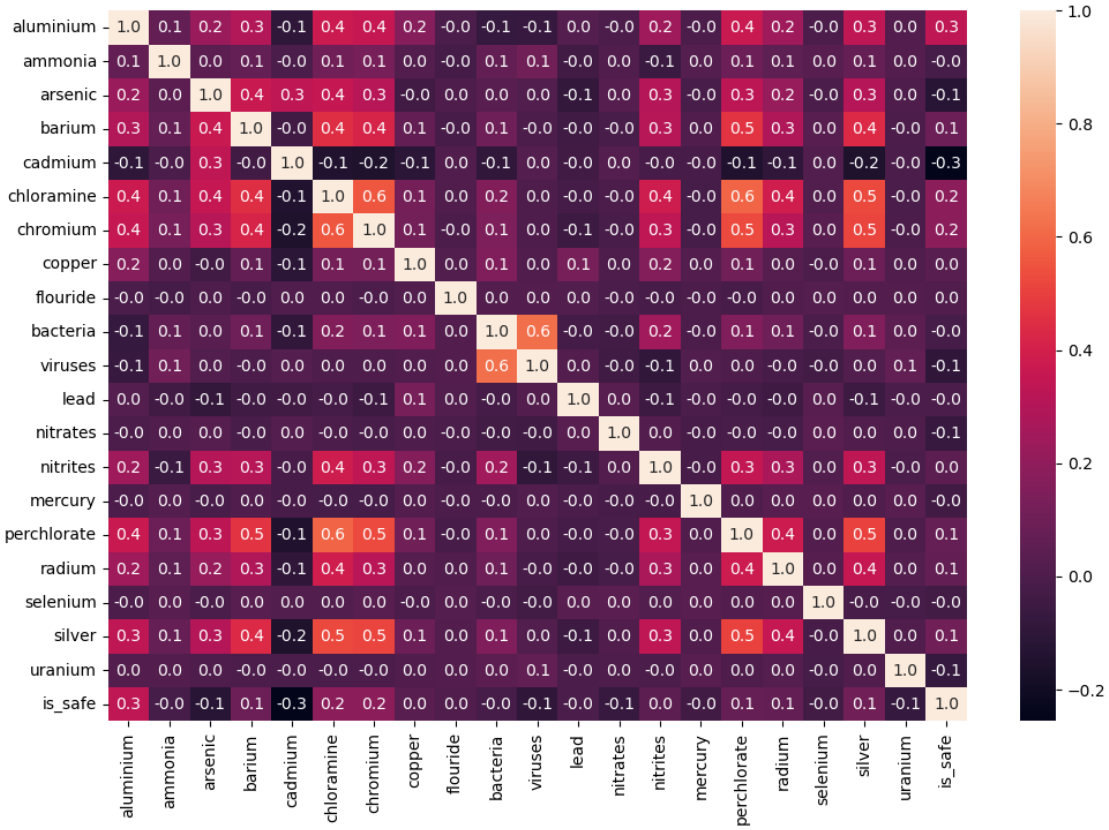

Figure 4.8: Correlation Matrix on a Heatmap

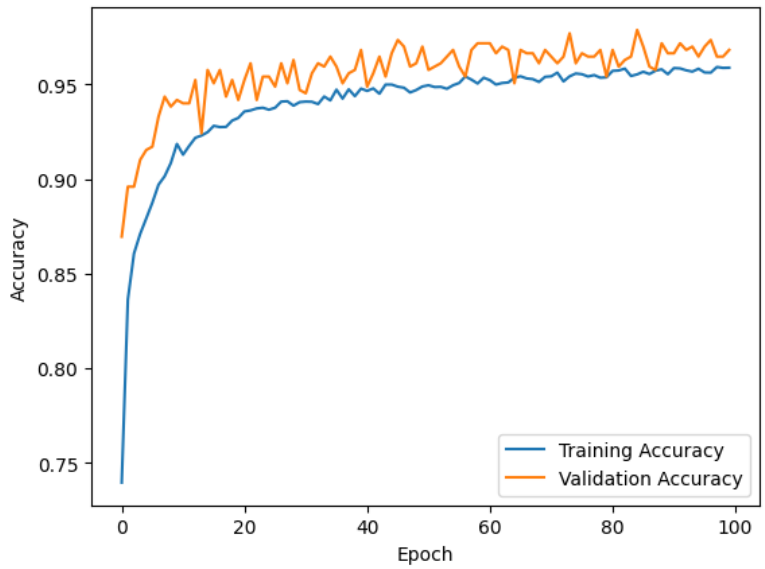

Figure 4.9: Training and Validation Accuracy Combiation of MLP Model

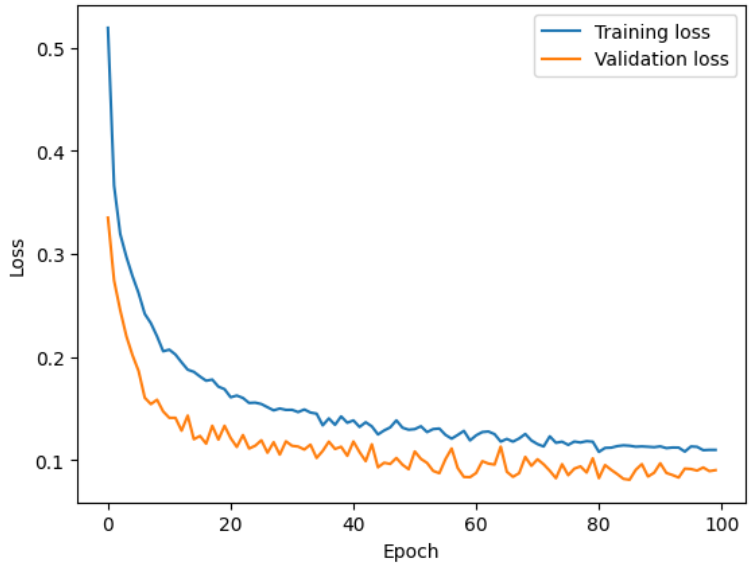

Figure 4.10: Training and Validation Loss Combiation of MLP Model

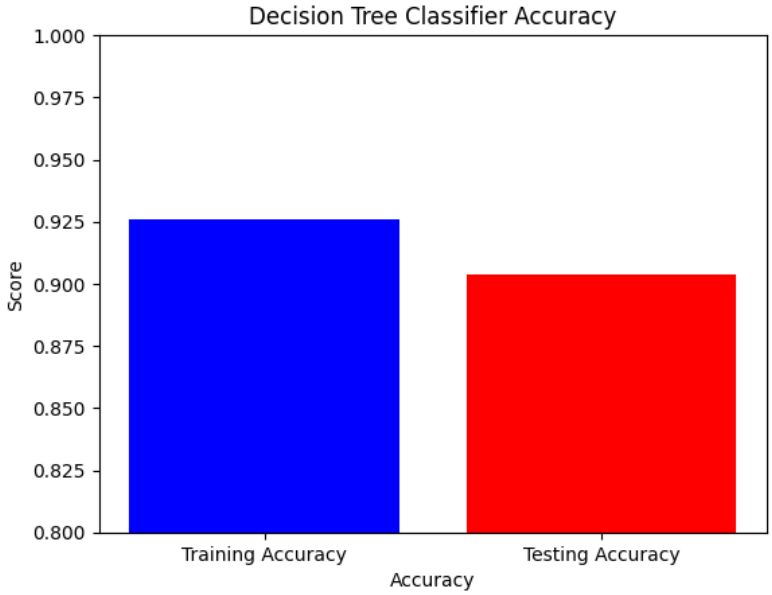

Figure 4.11: Training and Testing Accuracy Combiation of Decision Tree algorithm

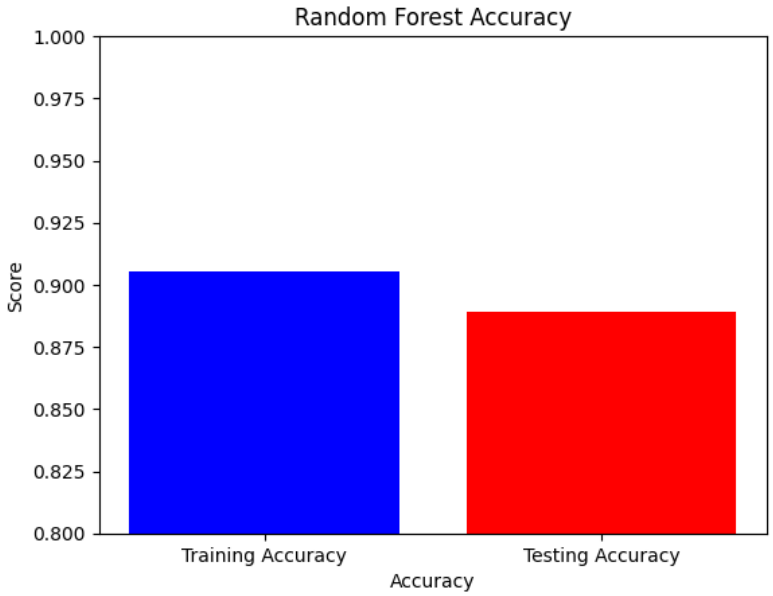

Figure 4.12: Training and Testing Accuracy Combiation of Random Forest algorithm

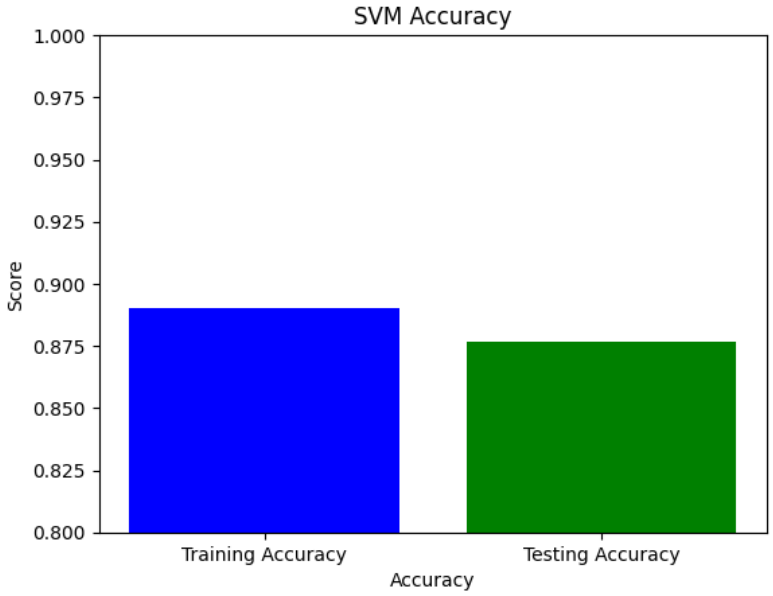

Figure 4.13: Training and Testing Accuracy Combiation of SVM algorithm

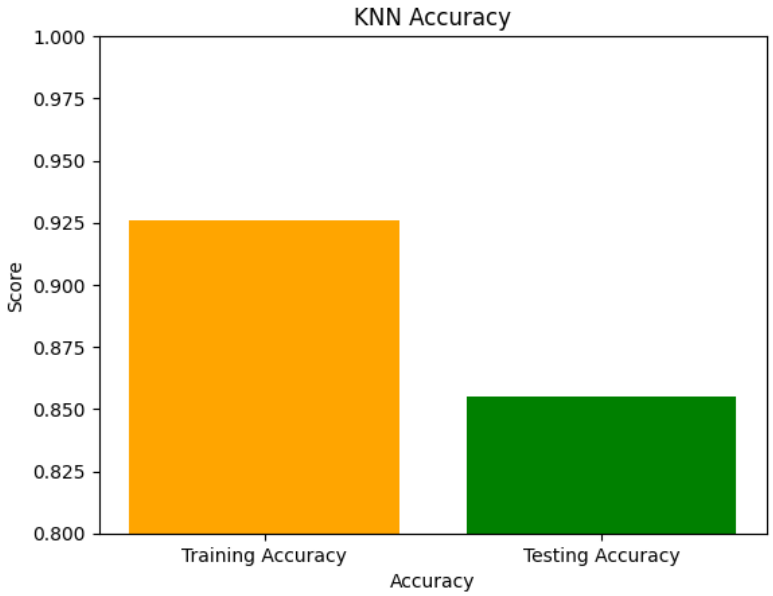

Figure 4.14: Training and Testing Accuracy Combiation of KNN algorithm

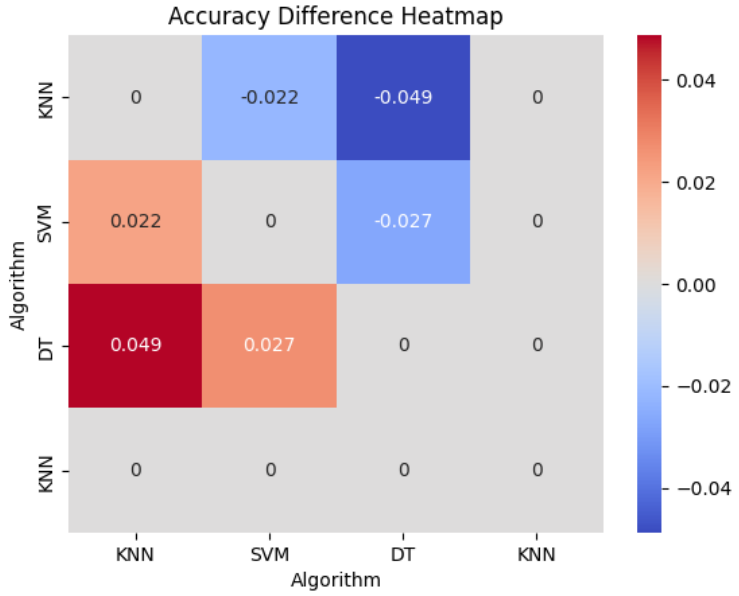

Figure 4.15: Accuracy difference on a Heatmap

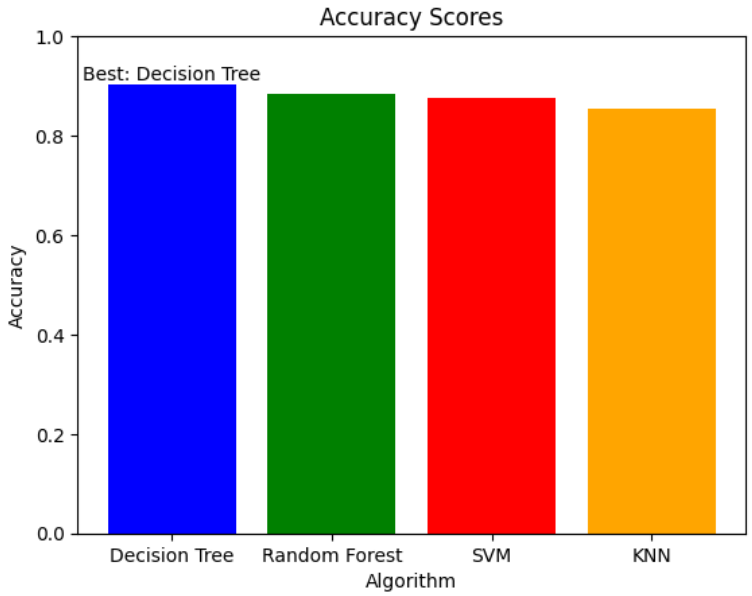

Figure 4.16: Comparison of Machine Learning Algorithms

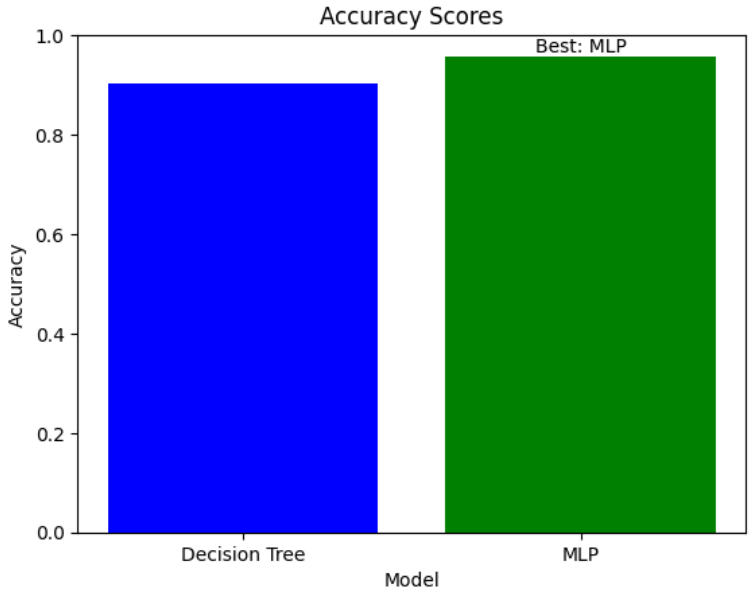

Figure 4.17: Comparison between Decision Tree and MLP

Chapter 5

CONCLUSION

In conclusion,Predicting water quality is crucial for a number of reasons. It first aids in ensuring safe drinking water, which is crucial for maintaining public health. A safe water supply can be ensured by taking the proper precautions and identifying potential dangers, such as the presence of dangerous contaminants, with the aid of water quality forecasting. Predicting water quality can also help to save the environment and ecosystems.Water quality prediction using ANNs is a powerful tool for predicting water quality parameters. ANNs can learn complex relationships between input and output data, and can adapt to changing conditions. However, ANNs require a large amount of data for training and can be computationally expensive. When used appropriately, ANNs can provide accurate predictions of water quality parameters and improve water resource management.In Water Quality Prediction system it predicts the quality of water using MultiLayer Perceptron model.The MultiLayer Perceptron model gives a high accuracy rate of 95.73% .The results of this proposed model verified that the best performance scores of MultiLayer Perceptron model rather than the Machine Learning Classification algorithms.

5.1 Future Enhancement

Deep Learning and machine learning methods are commonly employed in water quality predicting systems.In this proposed system the required data for predicting water quality is entered by the user.Real time water quality analysis could be included in the study will improve the performance.Using sensors for analysing or detecting the water contents and directly fed into the model make the system more simpler.

REFERENCES

[1] H. A. Madni et al.,"Water-Quality Prediction Based on H2O AutoML and Explainable AI Techniques",in *Water 2023* ,vol. 15, no. 3, pp. 475, Jan. 2023

[2] Ali N Ahmed,Faridah B O,Haitham Abdulmohsin, Rusul KI,Chow Ming F,Md Shabbir Hossain,M Ehteram and Ahmed Elshafie,"Machine learning methods for better water quality prediction", in *Journal of Hydrology*,vol. 578,Nov. 2019

[3] Amir H H,A H Nasrolahi,Abbas Parsaie,"Water quality prediction using machine learning methods", in *Water Quality Research Journal*,vol. 53, no. 1,pp. 3-13,Jan. 2018

[4] Y Wang,J Zhou,K Chen,Y Wang,Linfeng L,"Water quality prediction method based on LSTM neural network", in *2017 IEEE 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pp. 1-5,Nov. 2017

[5] Hongfang Lu,Xin Ma,"Hybrid decision tree-based machine learning models for short-term water quality prediction", in *Chemosphere* ,vol. 249,June. 2020

[6] Jing Bi,Yongze Lin,Quanxi Dong,Haitao Yuan,MengChu Zhou,"Large-scale water quality prediction with integrated deep neural network", in *Information Sciences*,vol. 571,pp. 191-205,September. 2021

[7] Tianan Deng, K W Chau, H F Duan, "Machine learning based marine water quality prediction for coastal hydro-environment management", in *Journal of Environmental Management*,vol. 284,April. 2021

[8] Elias Dritsas and Maria Trigka,"Efficient Data-Driven Machine Learning Models for Water Quality Prediction", in *Computation*,vol. 11, no. 2, pp. 16, Jan. 2023

[9] A. L. Lopez, N. A. Haripriya, K. Raveendran, S. Baby and C. V. Priya,"Water quality prediction system using LSTM NN and IoT", in *2021 IEEE International Power and Renewable Energy Conference (IPRECON)*,pp. 1-6,September. 2021

[10] Jin-Won Yu,Ju-Song Kim,Xia Li,Yun-Chol,Kwang-Hun,G Ryang,,"Water quality forecasting based on data decomposition fuzzy clustering and deep learning neural network", in *Environmental Pollution*,vol. 303,June. 2022

[11] Di Wu,Hao Wang,Razak Seidu,"Smart data driven quality prediction for urban water source management", in *Future Generation Computer Systems*,vol. 107,pp. 418-432,June. 2020

[12] Hong-Gui Han, Qi-li Chen,Jun-Fei Qiao,"An efficient self-organizing RBF neural network for water quality prediction", in *Neural Networks*,vol. 24,no. 7,pp. 717-725,September. 2011

[13] H. A. Nascimento Silva, A. Rosato, R. Altilio and M. Panella,"Water Quality Prediction Based on Wavelet Neural Networks and Remote Sensing", in *2018 International Joint Conference on Neural Networks (IJCNN)*,pp. 1-6,July. 2018

APPENDIX

Screenshots

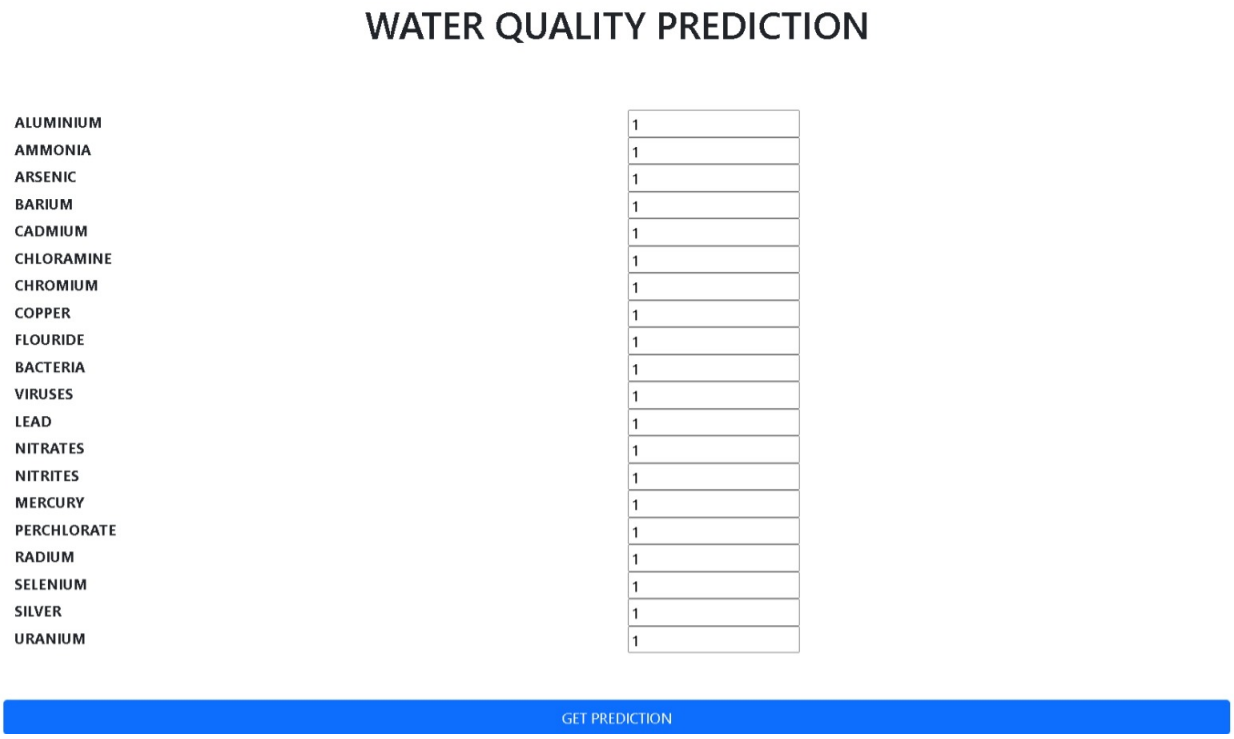

Figure A.1: Prediction 1

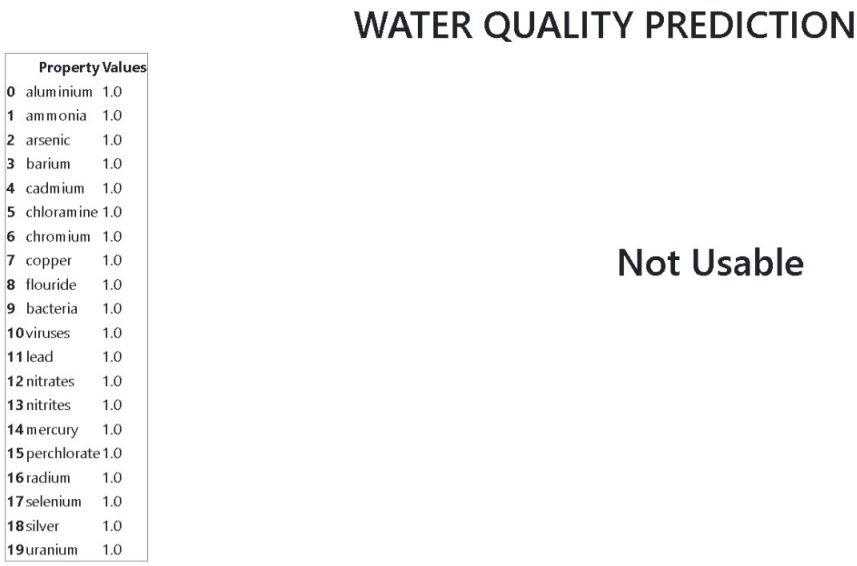

Figure A.2: Prediction 1 result

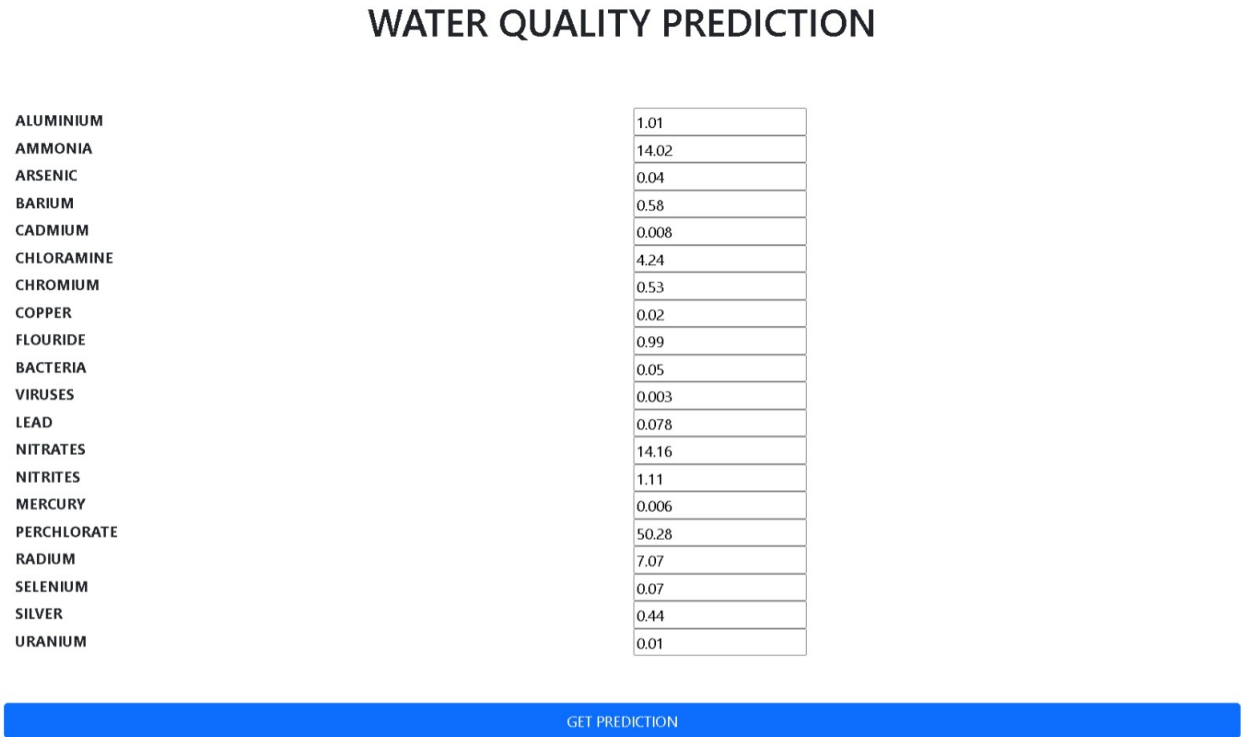

Figure A.3: Prediction 2

	Property	Values
0	aluminium	1.010
1	ammonia	14.020
2	arsenic	0.040
3	barium	0.580
4	cadmium	0.008
5	chloramine	4.240
6	chromium	0.530
7	copper	0.020
8	flouride	0.990
9	bacteria	0.050
10	viruses	0.003
11	lead	0.078
12	nitrates	14.160
13	nitrites	1.110
14	mercury	0.006
15	perchlorate	50.280
16	radium	7.070
17	selenium	0.070
18	silver	0.440
19	uranium	0.010

WATER QUALITY PREDICTION

Usable

Figure A.4: Prediction 2 result