

Smart Agriculture Solutions using Machine Learning on Sensor Network Data

B.Tech Major Project - I Project report

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF
BACHELOR OF TECHNOLOGY
IN
COMPUTER ENGINEERING
by

Adarsh Jha

2K17/CO/022

Aditya Choudhary

2K17/CO/024

Under the supervision of

Dr. R K Yadav

Assistant Professor

COE Department



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of engineering)

Bawana Road, Delhi-110042

NOVEMBER, 2020

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

CANDIDATE'S DECLARATION

We Adarsh Jha, Aditya Choudhary, Roll no – 2K17/CO/022, 2K17/CO/024 , hereby declare that the project Dissertation titled “**Smart Agriculture Solutions using Machine Learning on Sensor Network Data**” which is submitted by us to the Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology (Computer Engineering), is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

Date: 28 November 2020



Adarsh Jha



Aditya Choudhary

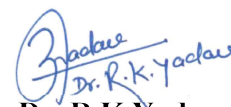
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

We hereby certify that the Project Dissertation titled “**Smart Agriculture Solutions using Machine Learning on Sensor Network Data**” which is submitted by Adarsh Jha, Aditya Choudhary, Roll Nos. – 2K17/CO/022, 2K17/CO/024 Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 28 November 2020



Dr. R K Yadav

SUPERVISOR

ACKNOWLEDGEMENT

We would like to express great appreciation to our supervisor Dr R. K. Yadav for his valuable and constructive suggestions during the planning and development of this project. His willingness to give his time so generously has been very much appreciated. We owe a special debt to our parents and friends for showing their generous love and care throughout the entire period of this time.

Place: Delhi

Date: 28 November 2020



Adarsh Jha



Aditya Choudhary

ABSTRACT

Each step of agriculture, be it preparation of soil, adding fertilizer, irrigation, sowing and harvesting requires proper analysis of soil, water, sunlight, weather etc. for better crop yield which is time consuming and costly. This project uses machine learning on data obtained from IOT sensor networks to analyse and predict various parameters for a better crop yield. This project also takes care of any missing data which may occur due to some sensors being offline.

The missing data is predicted using various linear and non-linear regression models such as Polynomial Regression, Random Forest, Gradient Boost taking various combinations of parameters. Water quality Index is used as a measure of quality of water and a model consisting of various linear and non- linear regression models is used for prediction of water quality index based on parameters measured using sensor data. Model for classification of water samples based in WQI is also developed which uses various Classification techniques like SVM, Decision tree etc. The results obtained are satisfactory and can be incorporated in for saving time and cost of manual lab tests in agriculture.

CONTENTS

Candidate's Declaration	ii
Certificate	iii
Acknowledgement	iv
Abstract	v
Contents	vi
List of Figures	viii
List of Tables	ix
List of Abbreviations	x
CHAPTER 1 Introduction	1
1.1 Overview	1
1.2 Motivation	1
1.3 Objectives	2
CHAPTER 2 Related Works	3
CHAPTER 3 Literature Review	4
3.1 Water Quality Index	4
3.2 Regression Models	5
3.3 Classification Models	7
CHAPTER 4 Proposed Work	10
4.1 Datasets	10
4.2 Implementation	11
4.2.1 Reconstruction of missing data	11
4.2.2 Calculation and Prediction of WQI	13
4.2.3 Classification based on WQI	15
4.3 Overall Architecture	17
CHAPTER 5 Experiments and Results	18

5.1	Experimental Setup	18
5.2	Evaluation Metrics	18
5.2.1	Metrics for Regression	18
5.2.2	Metrics for Classification	19
5.2.3	Cross Validation	19
5.3	Result of Experiments	20
5.3.1	Reconstruction of missing data	20
5.3.2	Prediction of WQI	21
5.3.3	Classification based on WQI	22
CHAPTER 6 Conclusion and Future Work		24
APPENDIX A		25
REFERENCES		29

LIST OF FIGURES

3.1	Working of random forest regressor	6
3.2	One hidden layer of MLP	7
3.3	Example of an ANN	7
3.4	Margins for an SVM trained with samples from two classes	8
3.5	Visualization of a Random Forest	9
3.6	Example of a decision tree	9
4.1	Workflow for Reconstruction of Missing Data	12
4.2	Workflow for Calculation and Prediction of WQI	14
4.3	Workflow for Classification Based On WQI	16
4.4	Overall Architecture	17
5.1	Comparison between actual class and predicted class for random forest classifier	23

LIST OF TABLES

3.1	Weights of various parameters according to WQI-NSF	4
3.2	Classification of water sample based on WQI	5
4.1	Water quality Dataset Parameters	10
4.2	Weather and Soil Dataset Parameters	10
5.1	Prediction scores for Data Reconstruction of Soil Temperature	20
5.2	Five fold cross validation scores for Data Reconstruction of Soil Temperature	20
5.3	Scores for prediction of Water Quality Index	21
5.4	Cross Val. Scores for prediction of Water Quality Index	21
5.5	Classification scores for Indian Water Quality Data	22
5.6	Cross Val. Accuracy for WQI Class	22

LIST OF ABBREVIATIONS

IOT	-	Internet of Things
ML	-	Machine Learning
WSN	-	Wireless Sensor Networks
SVR	-	Support Vector Regressor
MLP	-	Multilayer Perceptron
WQI	-	Water quality Index
ANN	-	Artificial Neural Network
SVM	-	Support Vector Machine
MSE	-	Mean Squared Error
MAE	-	Mean Absolute Error
TP	-	True Positives
FP	-	False Positives
TN	-	True Negatives
FN	-	False Negatives
BOD	-	Biological Oxygen Demand
DO	-	Dissolved Oxygen
PH	-	Potential Hydrogen

CHAPTER 1 INTRODUCTION

1.1 Overview

According to the economic survey of 2018[1] in India, around 50% of the workforce is involved in the agriculture sector. The contribution of the sector to the GDP is only 16% which has reduced significantly from 50% in 1950. The decline is not limited to India and is observed in the rest of the world. This low productivity depends on many factors, one of the major one is wastage of farming resources, money and time. Also, the majority of people involved in farming are from rural areas who are poor and have insufficient knowledge regarding farming practices. Each step in farming, be it preparation of soil, adding Fertilizers, irrigation or harvesting, requires proper analysis of soil nutrients, water quality, weather, sunlight etc for improving the productivity.

With improvement in technology, various farming equipment have been developed aiming to increase productivity and proper utilization of resources. The use of the Internet of Things, which consists of various devices connected through a network interacting with computers to transfer data, is becoming feasible with the internet and computers getting faster and cheaper. Various sensors can be placed in a farming environment to obtain large, real-time datasets such as irrigation water(pH, TDS, various chemical concentration), soil(moisture, pH, nutrients) , weather(temperature, humidity etc). These datasets can then be analyzed and used to train machine learning models that will help in proper utilization of resources, determining which crop could be suitable for a specific environment and increasing the overall efficiency of crop production.

The project aims to develop models for reconstruction of data, water quality prediction and classification. Chapter 1 focuses on the overview, motivation and objectives of the project. Chapter 2 focuses on related works that are being carried out independent to this project but some references are taken from those works with proper citations. Chapter 3 gives a literature review of various techniques used to develop the project. Chapter 4 gives the description of various datasets and architecture of various models developed. Chapter 5 gives the details of the experimental environment, evaluation criteria and results obtained by our models using various graphs and tables.

1.2 Motivation

Problem with agriculture sectors across the world is low productivity. This is because a huge percentage of people involved in agriculture are not properly skilled which leads to wastage of resources, time and money. Study of various parameters such as soil nutrition, water quality which require lab tests are expensive. Due to farmers being relatively poor, there is a need to develop methods which would find parameters cheaper than lab tests. With developments in cloud services and IOT, it is feasible to analyze data gathered from various sensors for helping in agriculture. Various AI techniques can be used on the gathered data to find out patterns that would help save time and cost of measuring parameters using expensive lab tests.

1.3 Objectives

This work is focused mainly on application of machine learning on various climate and agricultural datasets gathered from various sensors and other sources. The work is divided to fulfil the following objectives:

- Data gathered from various sensors can be missing due to them being offline or some failure. There needs to be a method to reconstruct those data using previously gathered time series data. Various linear and non-linear models have been used to predict the missing data in various datasets.
- Before irrigation, it is necessary to test and evaluate the quality of water according to a set of standards so as to prevent damaging the soil and crops. There needs to be a measure of water quality based on data gathered from various sensors. Water Quality Index, defined by the world health organization is used in this work as the quality measure for irrigation water. A model is developed consisting of various regression techniques to predict the water quality Index based on parameters measured by the sensors.
- Water Sample having a low Index is not suitable for farming and a method is needed for classifying water samples. Agricultural Water Quality Classes classification Table 3.2 has been used as the classification criteria. Classification Model is developed using various classification techniques such as SVM, decision tree etc. to classify water samples using sensor data like pH, Conductivity etc.

CHAPTER 2 RELATED WORKS

Many equipment have been developed for helping in agriculture that includes modern farming tools such as mower, sprayer, seed drill etc. Also, with recent developments in IOT and cloud computing, many new and cheap sensors have come out that can help in agriculture. Also, many methods have been developed which make use of these equipment for efficient farming. Precision agriculture[2] is the process developed which involves gathering various time series, geolocation and individual data, processing and analyzing them, and making use of them for improving productivity in agriculture. Any step of agriculture where data can be gathered can come under precision agriculture. For example, in irrigation systems data can be gathered and analyzed to find the quality of water, determining and classifying soil and fertilizing it accordingly which will help in better fertilizing.

Irrigation involves watering the soil after sowing the seed for proper growth of crops. Vij. et al[3] proposes the use of wireless sensor networks for measuring various parameters such as temperature, moisture, water level, weather etc. , passes it through Support vector regression(SVR), Random Forest Regression algorithms to classify soil type and predict the amount of water required for irrigation. Janani and Jebakumar[4] measure soil, plant and water data and pass it through a MLP to estimate the irrigation amount.

Before Irrigation, it is essential to measure the quality of water so as to prevent crops from getting damaged by polluted water. Orozco et al.[5] studied various samples of water for irrigation and developed a water quality index(WQI) measure. Lerios et al.[6] developed a method for prediction and classification of water according to parameters such as pH, FC etc. using various classifying techniques like Naive Naves, Decision Tree, Random Forest, Gradient Boost and MLP.

Analysis of soil gives proper insights for fertilizing and irrigation of soil. Cai Y[7] uses a combination of meteorological and soil moisture data and uses a deep learning regressor network to determine weights for soil moisture prediction. Gholap[8] measures soil parameters like pH, Electrical Conductivity to classify soil using various algorithms and implementing automated soil sample classification. Suchitra et al.[11] classifies soil nutrients on the basis of mineral contents using classification techniques known as Extreme learning machines(feedforward neural networks) using various activation functions.

Sensors and wireless networks are susceptible to failure in a few cases, due to which some percentage of data may be missing. Gad I.[9] proposes the use of a deep learning imputation model using various optimizers (SGD, Adam, Rmsprop etc)) to compute missing parameters. Balducchi[10] uses a set of decision Tree and K nearest neighbours to predict the missing values.

CHAPTER 3 LITERATURE REVIEW

3.1 Water Quality Index

The Water Quality Index(WQI), formulated by Horton[13] is a linear function which gives the quality of water. The index is calculated using a weighted sum of various parameters such as pH, temperature etc. These parameters are selected such that they are available in all of the water sources that one wants to measure from. The WQI is calculated using four steps:

Parameter Selection and Weights

According to one of the most widely used index calculation, nine parameters are used to calculate WQI: Temperature, pH, turbidity, phosphate, Nitrate, total Solids, dissolved Oxygen (DO), biochemical Oxygen Demand (BOD) and fecal coliform. The weights for each of the parameters were obtained by the use of DELPHI technique [14] under WQI-NSF [15] as shown in table 3.1.

Table 3.1 Weights of various parameters according to WQI-NSF

Parameter	Weights
DO	0.17
Fecal coliforms	0.15
pH	0.12
BOD	0.10
Nitrate	0.10
Total phosphate	0.10
Temperature	0.10
Turbidity	0.08
Total solids	0.08

Q-value Normalization of Various Parameters

The values of various parameters are normalized to be in the range of 0-100 for easy index calculation. Table and graph for conversion of Parameter to Q-values are given in Appendix A.

Formula for Water Quality Index

After obtaining the weights and Q-values, WQI is calculated using the formula:

$$\text{Water Quality Index} = \sum (Rel(W_i) * Q_i) \quad (3.1)$$

Where, $Rel(W_i)$ is the relative weight of feature 'i' obtained by dividing Each weight by total weight,

Q_i is the Q-value of i-th parameter

Classification based in WQI

Based on calculated WQI, the water sample can be classified for various uses. Meirels et al [23] proposed a water quality index as shown in Table 3.2.

Table 3.2 Classification of water sample based on WQI

WQI	Restrictions	Soil	Plant
85–100	No restrictions (NR)	It can be used for most soils with low probability of solidification and salinization	Most plants won't be affected
70–85	Low restriction (LR)	Use for soil with fine texture or moderate permeability	Avoid use in plants with salt sensitivity
55–70	Moderate restriction (MR)	Can be used in soils with high or moderate permeability	Plants with moderate salt tolerance will be unaffected
40–55	High restriction (HR)	Can be used on soils with high permeability without layers of compaction.	It should be used to irrigate plants with moderate to high salt tolerance with special salinity control practices
0–40	Severe restriction (SR)	Use for irrigation under normal conditions should be avoided.	Avoided for all plants

3.2 REGRESSION MODELS

Linear Regression

Linear regression uses a linear approach to depict relationships between a scalar response and one or more variables [16]. This idea can be extended to predict multiple correlated dependent variables. The basic model of linear regression can be represented as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \epsilon_i \quad (3.2)$$

Mathematically it solves

$$\min_w ||Xw - y||_2^2 \quad (3.3)$$

This method is highly sensitive to the presence of outliers.

Random Forest Regression

Random forest regression uses the idea of random forests which is an ensemble learning method [17]. This method uses multiple decision trees to predict the output which in case of classification will be the output class and in case regression will be the output value. For regression the output is the mean value of the outputs. It relies on Bootstrap and Aggregation to compute results.

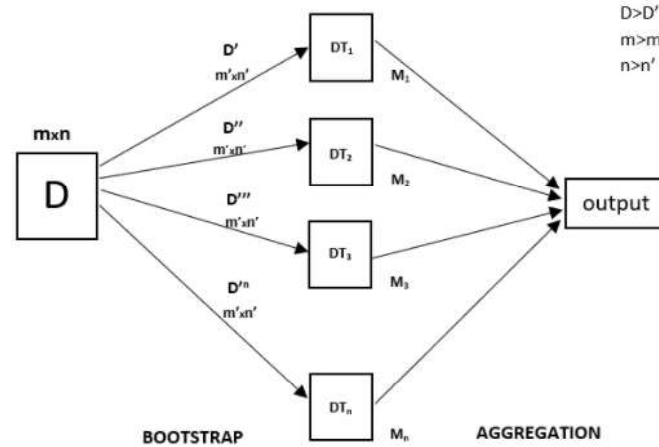


Figure 3.1 Working of random forest regressor

Gradient Boosting Regressor

Gradient Boosting Algorithm uses weak learners and makes changes in it to construct a strong learner [18]. Gradient boosting uses decision trees as their weak learners. It uses an additive model that allows for optimization of differentiable loss function. It is quite powerful but is prone to overfitting.

Polynomial Regression

This technique is used when the relationship between input variables and output is nonlinear [19]. The computation statistically is similar to the linear regression method, due to which it is also known as the special case of multiple linear regression.

It uses the given below equation:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_k x_i^k + \epsilon_i, \text{ for } i = 1, 2, \dots, n \quad (3.4)$$

MLP Regression

MLP (Multilayer perceptron) represents a feedforward artificial neural network where each perceptron has a linear function and an activation function [20]. These perceptrons are the building blocks of the neural network. Multi-layer perceptron (MLP) is a conventional model of neural net, which is mostly used for classification, but it can be used for regression as well by not using an activation function in the perceptron.

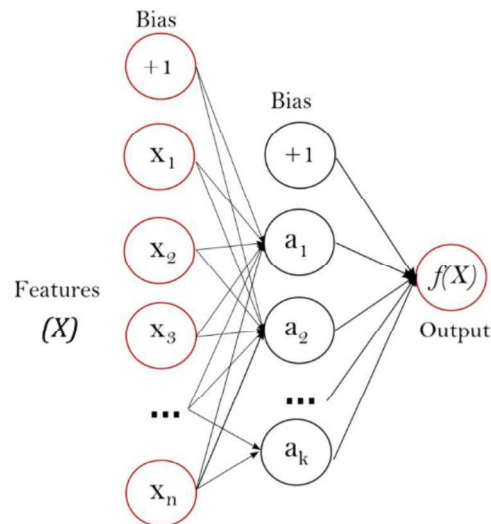


Figure 3.2 One hidden layer of MLP

ExtraTrees Regression

Also known as extremely randomized trees, ExtraTrees is quite similar to the random forests but differ in how the decision trees are constructed [29]. Here each tree has a random sample of features which lead to de-correlated trees. ExtraTrees is computationally faster than random forest and has low variance than the random forest algorithm.

3.3 CLASSIFICATION MODELS

Artificial Neural Networks

Artificial Neural Networks are inspired by the structure of our brain where dendrites receive the message which is passed through axon [21]. In ANN, neurons are responsible for receiving the input and producing the output after applying the activation function. Each neuron has a weight which increases or decreases as the learning proceeds. Typically, neurons are aggregated into layers. These layers transform the given input, finally producing and output which gets tested with the actual output and the error gets back propagated.

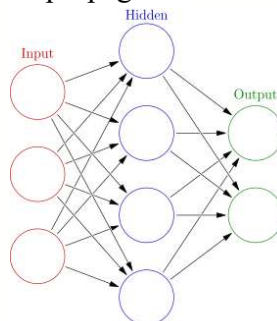


Figure 3.3 Example of an ANN

SVM

SVM presents one of the most robust prediction methods. Its objective is to find a hyperplane with maximum margin separation which distinctly classifies data points [22]. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. It can be used for classification, regression and also for outlier detection.

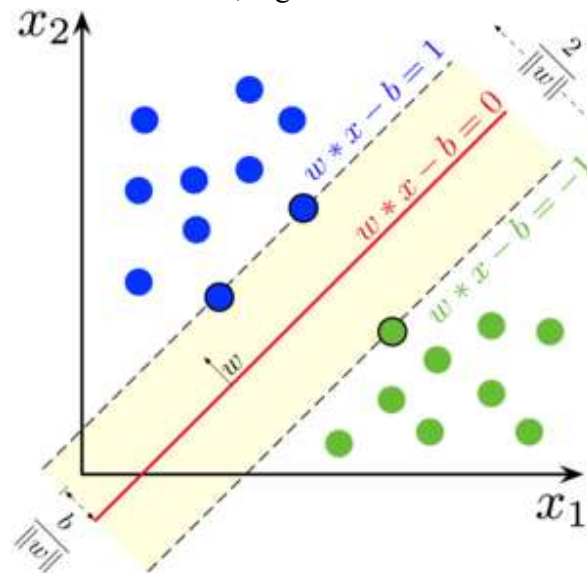


Figure 3.4 Margins for an SVM trained with samples from two classes

GRADIENT BOOST CLASSIFIER

Gradient Boosting Algorithm uses weak learners and makes changes in it to construct a strong learner [18]. Gradient boosting uses decision trees as their weak learners. It uses an additive model that allows for optimization of differentiable loss function. It is quite powerful but is prone to overfitting.

RANDOM FOREST CLASSIFIER

The random forest classifier uses multiple decision trees as an ensemble [17]. Each tree gives an output which contributes to the final result. The classifier is based on the idea that a group of classifiers outperforms a single classifier. The reason for this is that trees protect each other from their individual errors.

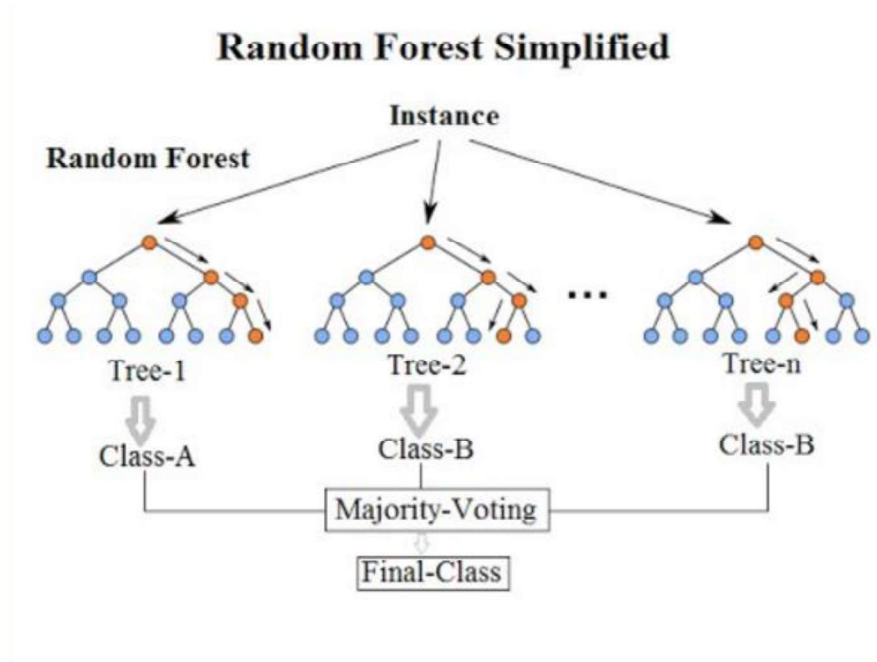


Figure 3.5 Visualization of a Random Forest Model

DECISION TREE

A decision tree is a flowchart like structure where nodes represent an if else condition, each branch represents the outcome and leaf nodes represent the actual class label [24]. Here, the path from root to leaf gives us the classification rule for that class. Commonly used algorithms for splitting are: Gini impurity, Chi-Square and Information Gain.

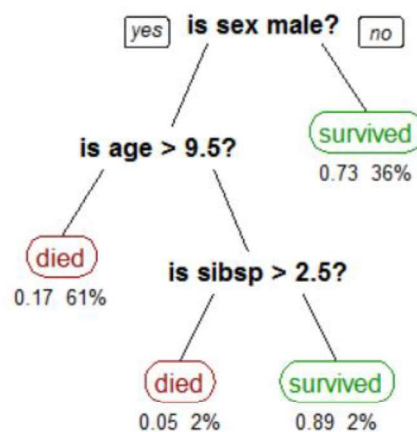


Figure 3.6 Example of a decision tree

CHAPTER 4 PROPOSED WORK

4.1 DATASETS

Indian Water Quality

Source: National Water Quality Monitoring Programme (NWMP) [30]

Data gathered by various underwater sensors and lab tests by Indian Government under NWQMP from various water bodies. This dataset contains the following parameters:

Table 4.1 Water quality Datasets' Parameters

Parameter	Description	Units
Temp	Temperature	Degree Celsius
D.O.	Total Dissolved Oxygen	mg/l
PH	Potential Hydrogen	None
Conductivity	Electrical Conductivity	μmho/ Cm
B.O.D	Biological Oxygen Demand	mg/l
Nitrate	Total Nitrate (NO ₃ ⁻)	mg/l
FC	Fecal Coliform	MPN/100ml
TC	Total Coliform	MPN/100ml

Weather and Soil Sensor Data

Source: the Johnston Draw catchment, Reynolds Creek Experimental Watershed and Critical Zone Observatory [31]

Hydrometeorological data gathered from a snow to rainfall terrain containing various Weather and soil parameters as mentioned below:

There is a total of 27093 rows after preprocessing and cleaning of data.

Table 4.2 Weather and Soil Dataset Parameters

Parameter	Unit
Air Temperature	Degree Celsius
Relative Humidity	None
Water Vapour Pressure	Pascal

Dew Point Temperature	Degree Celsius
Wind Speed	ms^{-1}
Wind Direction	Degree from North
Incoming Solar Radiation	W m^{-2}
Relative Soil Moisture	None
Soil Temperature	Degree Celsius

4.2 Implementation

The following sections describe three models prepared for missing data reconstruction, water quality prediction and classification respectively.

4.2.1 Reconstruction of Missing Data

Large amount of data can be gathered from sensors which can be used to train various machine learning algorithms to find out relationships among different parameters measured. The relationships can be used to predict missing values if some of the parameters are missing. Soil and Weather quality dataset containing more than 25k rows is used to develop this model. Various regression techniques are used in our model to find relationship between different parameters, details of which are mentioned in chapter 3.2 are trained and the model with the best score is selected for determination of parameters with missing values.

The workflow of the model is summarized below and shown in Figure 4.1:

- Data from various sensors are gathered with parameters mentioned in Table 3.2.
- Data Preprocessing: Rows containing missing values are removed, string values (such as Profession, Subject etc.) are mapped to integer values and all columns are normalized.
- Feature Selection: Data is split into features and labels for supervised learning. Here, Soil Temperature was selected as output feature (model can have any number of output feature less than number of input feature).
- Data is split into train-test and validation with a ratio of 80% data in training and 20% in testing.
- Regression: Linear, Polynomial, Random Forest, Extra Trees, Gradient Boost and MLP regression are trained on the training set and result is evaluated on the test set.
- Cross-Validation and model selection: Data is split into five parts for a five fold cross validation. Algorithm with the highest R2 score and cross-validation score is saved to the cloud and will be used in future to predict missing values.

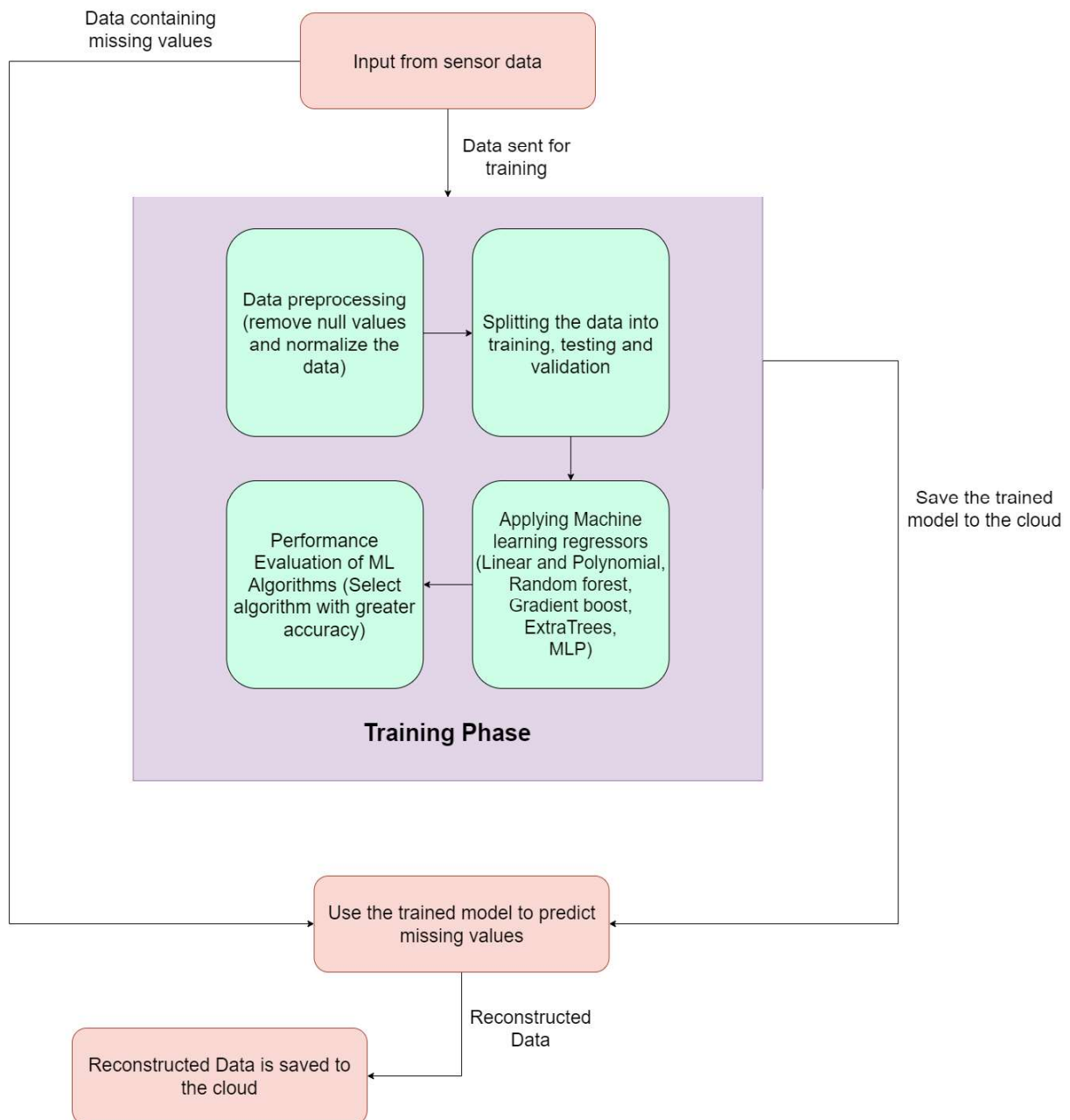


Figure 4.1 Workflow for Reconstruction of Missing Data

4.2.2 Calculation and Prediction of WQI

Water Quality Index is a standard for measuring the Quality of water. Further information about WQI is available in section 3.1. Values of various parameters are first Q-value normalized using and WQI is calculated. Then, WQI is taken as output and other parameters such as Temperature, PH, Conductivity etc as input features. Various prediction algorithms are used to find relationships between WQI and other parameters. Model with best accuracy is saved and can be used to predict WQI.

The workflow of the model is summarized below and shown in Figure 4.2:

- Data from various sensors are gathered with parameters mentioned in Table 4.1.
- Data Preprocessing: Rows containing missing values are reconstructed using the model mentioned in section 4.2.1.
- Q-Value and WQI calculation: Value of parameters calculated are normalized using Q-value normalization from Figure A.1. Then, Q-values are used to calculate the Water Quality Index using Equation 3.1.
- Feature Selection: WQI is chosen as the output feature and Temperature, D.O, BOD, PH, Conductivity and Fecal Coliform as input features.
- Data is split into train-test and validation with a ratio of 80% data in training and 20% in testing.
- Regression: Linear, Polynomial, Random Forest, Extra Trees, Gradient Boost and MLP regression are trained on the training set and result is evaluated on the test set.
- Cross-Validation and model selection: Data is split into five parts for a five fold cross validation. Algorithms with the highest R2 score and cross-validation score are saved to the cloud and will be used in future to predict WQI.

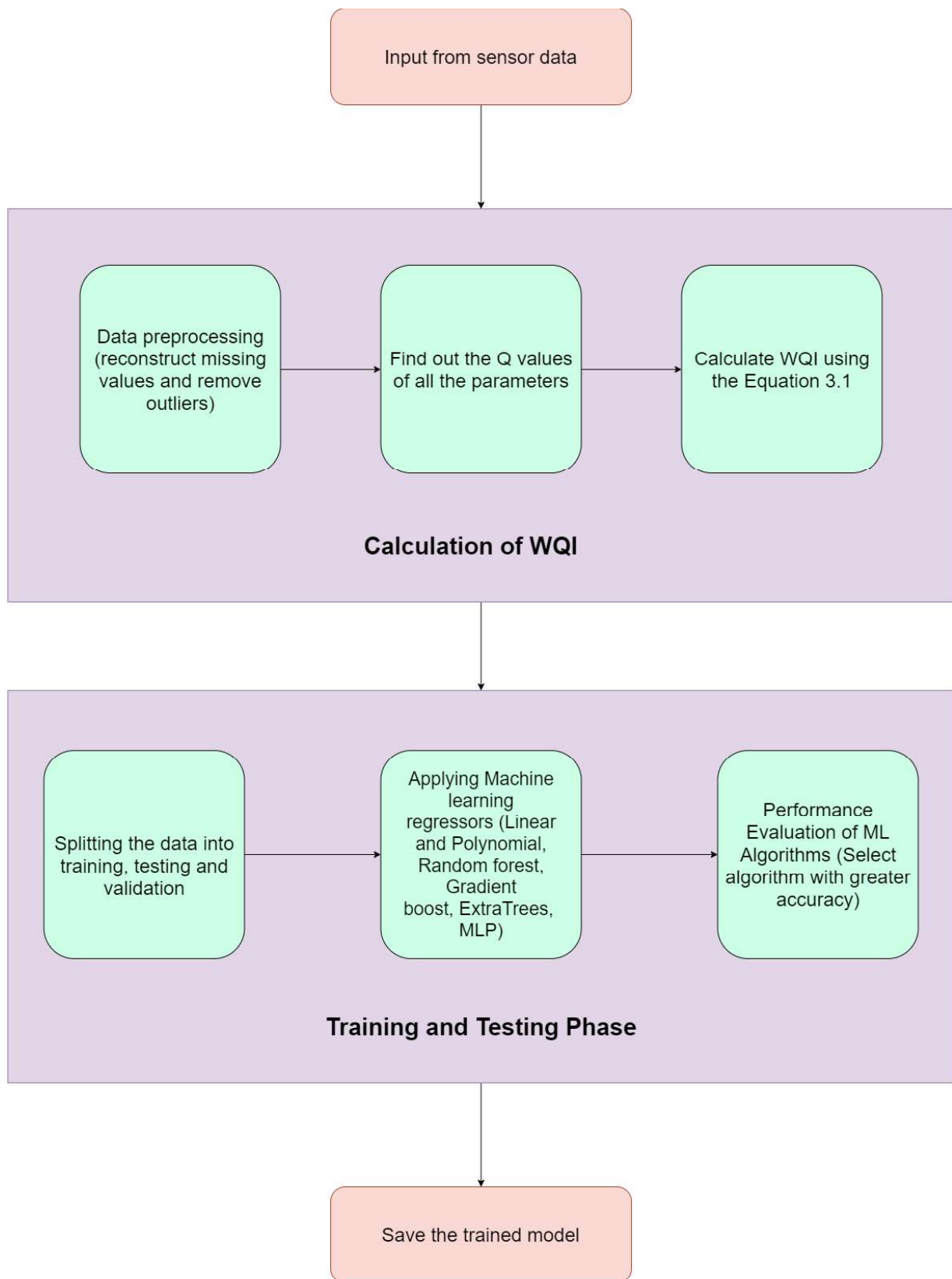


Figure 4.2 Workflow for Calculation and Prediction of WQI

4.2.3 Classification based on WQI

Different soils and crops need different qualities of water for ideal farming. Various research has been done and water samples can be classified based on WQI values. WQI class is taken as output and other parameters such as Temperature, PH, Conductivity etc as input features. Various classification algorithms are used to find relationships between WQI class and other parameters. Model with best accuracy is saved and can be used to predict WQI class.

The workflow of the model is summarized below and shown in Figure 4.3:

- Data from various sensors are gathered with parameters mentioned in Table 4.1.
- Data Preprocessing: Rows containing missing values are reconstructed using the model mentioned in section 4.2.1.
- Q-Value and WQI calculation: Value of parameters calculated are normalized using Q-value normalization from [Figure 1]. Then, Q-values are used to calculate the Water Quality Index using Eq(1).
- WQI class calculation: WQI class is calculated using the ranges given in Table 3.2 and actual WQI values.
- Feature Selection: WQI class is chosen as the output feature and Temperature, D.O , BOD, PH, Conductivity and Fecal Coliform as input features.
- Data is split into train-test and validation with a ratio of 80% data in training and 20% in testing.
- Classification: Decision Tree, ANN, SVM, Random forest and Gradient Boost classifiers are trained on the training set and result is evaluated on the test set.
- Cross-Validation and model selection: Data is split into five parts for a five fold cross validation. Algorithms with the highest accuracy and f measure are saved to the cloud and can be used in future to classify water samples.

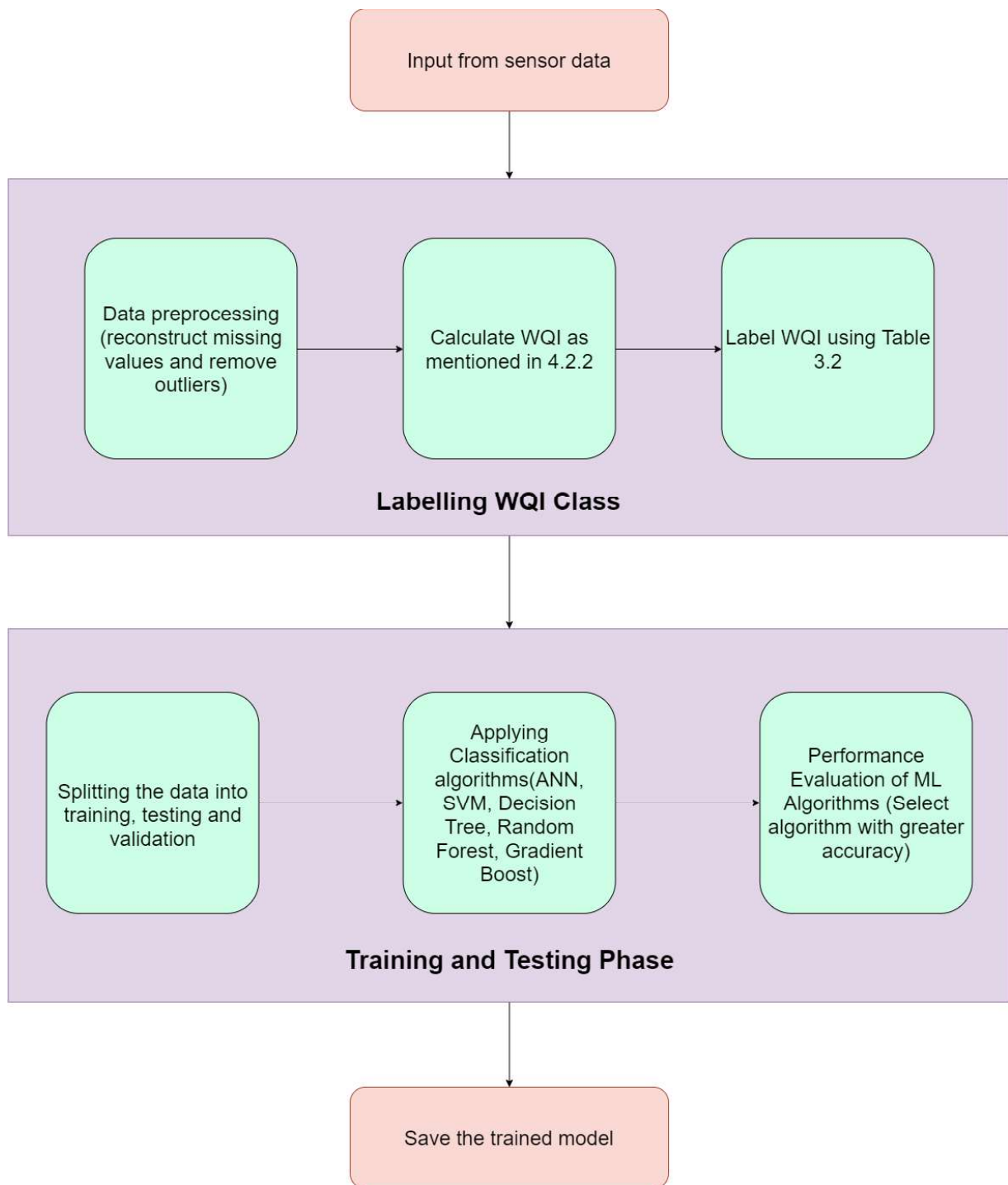


Figure 4.3 Workflow for Classification based on WQI

4.3 Overall Architecture

The model developed in section 4.2 for reconstruction of data, prediction and classification based on WQI can be incorporated into smart farming systems in the cloud. The overall architecture from gathering of data to providing results to users is shown in Figure 4.4.

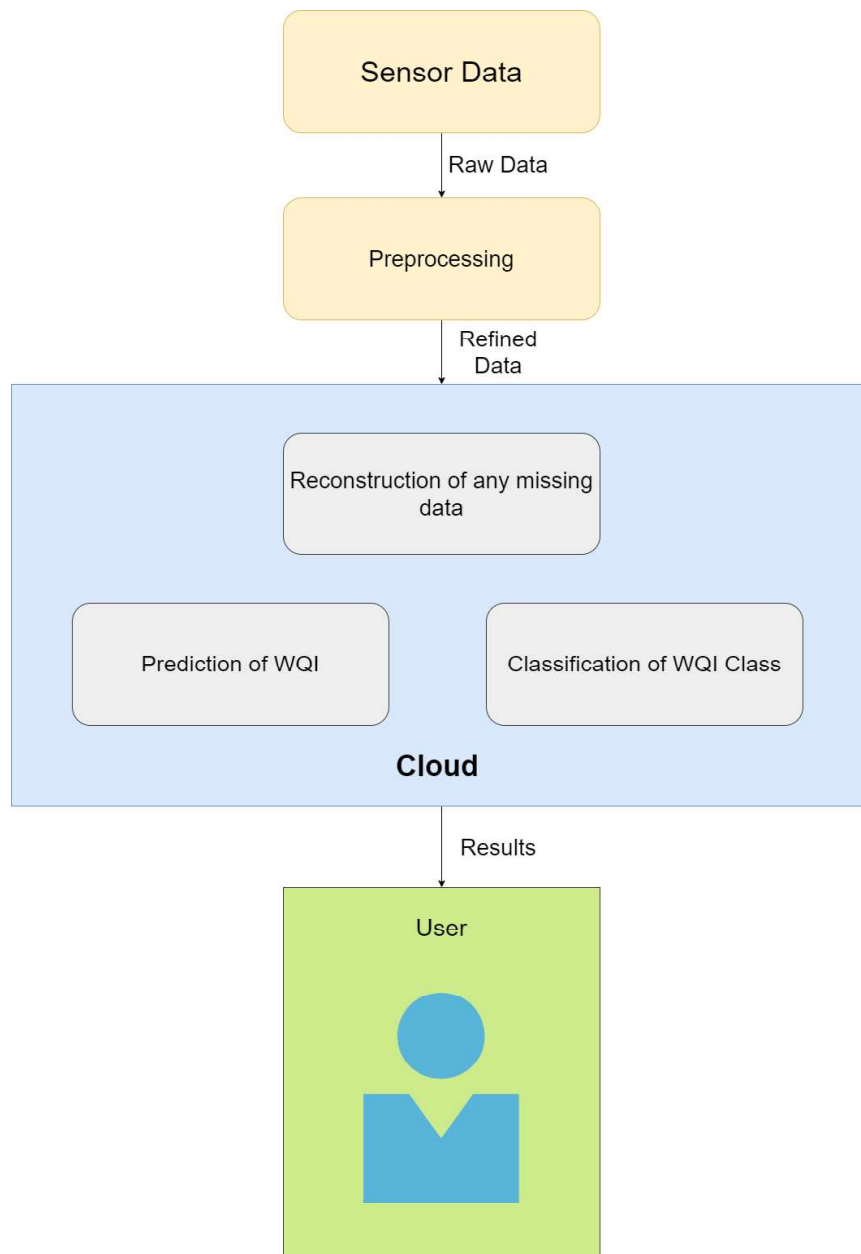


Figure 4.4 Overall Architecture of the System

CHAPTER 5 EXPERIMENTS AND RESULTS

This section describes the setup required to execute the models, evaluation metrics for regression and classification techniques and results and evaluation of various models.

5.1 Experimental Setup

The proposed models were implemented in Python3 and trained using Google Collaboratory. Following are the details of experimental setup for the project:

Operating System: Any Unix or Windows NT based

Programming Language: Python3

Python Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

5.2 Evaluation Metrics

This section describes the evaluation metrics that are used in this work for evaluation of regression and classification algorithms.

5.2.1 Metrics for Regression

Coefficient of determination

This metric is used to measure the relationship between two variables. It is also known as R squared and sometimes referred to as a goodness of fit. In simple terms, it is the ratio of explained variance to total variance. Its value is between 0 and 1 where 1 depicts perfect fit and 0 depicts that model fails to predict correct output at all.

$$R^2 = 1 - \frac{RSS}{TSS} \quad (5.1)$$

Mean square error

It is defined as the mean of the squared errors. It is used to measure the quality of the estimator and the closer its value to zero, the better the predictions of the model. Though it indicates the quality of the model, it doesn't indicate whether the model is overfitting the data or not.

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (5.2)$$

Mean absolute error

It is defined as the mean of the absolute errors. Just like the mean square error, it indicates the quality of the estimator and the closer its value to zero, the better the predictions of the model.

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (5.3)$$

For further information, refer [28]

5.2.2 Metrics for classification

Accuracy

It can be defined as the ratio of correct predictions to the total predictions. It helps us to find out how much data is classified correctly by the given model.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5.4)$$

Confusion Matrix

It is a square matrix in which how many values are classified to which class is depicted. This gives us the overall performance of the classification model and helps us to find what is wrong with the given model. It can be used to calculate various other metrics such as precision, recall, f-measure and accuracy.

Precision

It can be defined as the ratio of accurately classified values to the total positively classified values of that class. High precision indicates that a higher proportion of positively classified data was correct.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5.5)$$

Recall

It can be defined as the ratio of accurately classified values to the actual values belonging to that class. High precision indicates that a higher proportion of actually positive classes were classified as positive.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5.6)$$

F1 Score

It is a metric which captures both the relevance of precision and recall, captures features of both and puts it into a single score. It can be used to rate whether the given classifier performs better or not. It is high only when both precision and recall are high.

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5.7)$$

For further information, refer [28]

5.2.3 Cross validation

It is one of the techniques used for model validation which helps us estimate how our model would perform on the given dataset [26]. It helps us identify problems like overfitting and selection bias. Here the given dataset is partitioned into groups where some groups are used for training the model and remaining groups are used for testing the model.

5.3 Result of Experiments

This section describes the results we obtained from the models we created for reconstruction of missing data, prediction of WQI and classification based on WQI.

5.3.1 Reconstruction of Missing Data

This experiment aims at reconstruction of missing data of faulty sensors which needs a large amount of previous data available for the given sensor. We applied our data reconstruction model which contains various prediction algorithms on Soil and Weather data described in section 4.1. Soil temperature was chosen as the label and all other parameters (soil moisture, relative humidity, wind speed, wind direction and vapour pressure) as features. The results of the experiments are shown in Table 5.1 giving various evaluation metrics like MAE, MSE and coefficient of determination. We can see from Table 5.1 that ExtraTrees Regressor gives the best result with MSE 0.727, MSE 0.601 and R2 score as 0.9905.

Table 5.1 Prediction scores for Data Reconstruction of Soil Temperature

Method	MSE	MAE	R ² Score
Linear Regression	11.31585	2.70504	0.852188
Random Forest Regression	0.94134	0.6757	0.9877
ExtraTrees Regressor	0.727	0.60173	0.9905
Gradient Boosting Regressor	3.12163	1.3595	0.95922
Polynomial Regression (degree 3)	3.26492	1.42732	0.95735
MLP Regression	11.4655	2.71131	0.85023

Dataset is then divided into five parts and five-fold cross validation is applied on it whose results are shown in Table 5.2.

Table 5.2 Five fold cross validation scores for Data Reconstruction of Soil Temperature

Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Linear Regression	0.85524	0.85541	0.84908	0.84844	0.85592
Random Forest Regression	0.98837	0.98785	0.98751	0.98829	0.98799

ExtraTrees Regressor	0.99083	0.99046	0.99039	0.99068	0.99065
Gradient Boosting Regressor	0.96011	0.95885	0.96152	0.95931	0.95908
Polynomial Regression (3)	0.95844	0.9548	0.9542	0.95446	0.95588
MLP Regression	0.88623	0.85932	0.88844	0.89274	0.8263

5.3.2 Prediction of WQI

The experiment aims at prediction of Water Quality Index for Indian Water Quality Dataset mentioned in 4.1. Water Quality index is taken as the label and Temperature, D.O , BOD, PH, Conductivity and Fecal Coliform are taken as features. Various linear and nonlinear regression algorithms are used to predict WQI whose MAE, MSE and R2 scores are mentioned in Table 5.3. Gradient Boosting Regressor gave the best result with MSE 8.05242, MAE 2.06262 and R2 score of 0.93297.

Table 5.3 Scores for prediction of Water Quality Index

Method	MSE	MAE	R² Score
Linear Regression	42.56974	5.26115	0.56938
Random Forest Regression	7.00893	1.60631	0.9291
ExtraTrees Regressor	14.39321	2.40781	0.8544
Gradient Boosting Regressor	8.05242	2.06262	0.93297
Polynomial Regression (for degree 3)	29.84263	3.80581	0.69812
MLP Regression	36.46944	4.53955	0.69642

Dataset is then divided into five parts and five-fold cross validation is applied on it whose results are shown in Table 5.4.

Table 5.4 Cross Val. Scores for prediction of Water Quality Index

Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Linear Regression	0.58808	0.67945	0.64195	0.50281	0.66444
Random Forest Regression	0.90538	0.92066	0.91046	0.89434	0.92821

ExtraTrees Regressor	0.86189	0.90727	0.87003	0.83444	0.90454
Gradient Boosting Regressor	0.91411	0.94386	0.93138	0.90466	0.93309
Polynomial Regression (for degree 3)	0.71292	0.82328	0.73783	0.55293	0.74906
MLP Regression	0.69631	0.73555	0.72175	0.57071	0.72612

5.3.3 Classification based on WQI

This experiment aims at classification of WQI according to the WHO standards existing for water quality. We applied our model which contains various prediction algorithms on Indian Water Quality Data described in section 4.1. The WQI class was chosen as label and other parameters (Temperature, D.O, BOD, PH, Conductivity and Fecal Coliform) were used as features. The results of the experiments are shown in Table 5.5 giving various evaluation metrics like precision, recall, f measure and accuracy. We can see from Table 5.5 that Random Forest Classifier gives the best result with Precision 0.73373, Recall 0.7247, F measure 0.72445 and Accuracy as 0.92405.

Table 5.5 Classification scores for Indian Water Quality Data

Method	Accuracy	Precision	Recall	F measure
ANN	0.63291	0.33629	0.40237	0.34677
SVM	0.83228	0.63661	0.60242	0.61597
GRADIENT BOOST CLASSIFIER	0.92089	0.73205	0.72357	0.72305
RANDOM FOREST CLASSIFIER	0.92405	0.73373	0.7247	0.72445
DECISION TREE	0.89241	0.69453	0.72275	0.70696

Dataset is then divided into five parts and five-fold cross validation is applied on it whose results are shown in Table 5.6

Table 5.6 Cross Val. Accuracy for WQI Class

Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
ANN	0.57312	0.50397	0.57937	0.52778	0.60714
SVM	0.81423	0.8373	0.79365	0.8373	0.83333

GRADIENT BOOST CLASSIFIER	0.86561	0.88889	0.88492	0.89286	0.94841
RANDOM FOREST CLASSIFIER	0.88142	0.90476	0.90079	0.90079	0.95238
DECISION TREE	0.86561	0.85317	0.88492	0.87302	0.8373

We obtained the best results from Random Forest Classifier for which the comparison between the actual class of the observation and predicted class of the observation is shown in Figure 4.3.

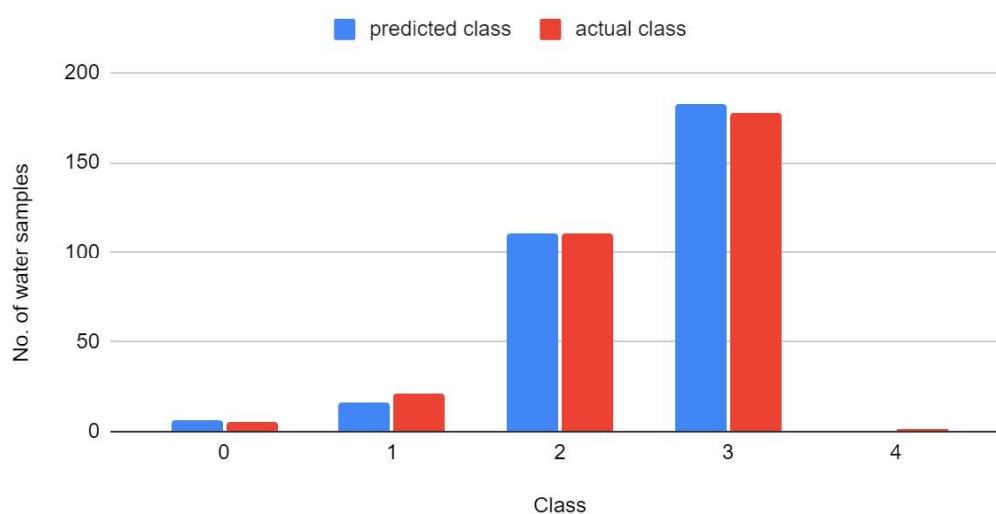


Figure 5.1 Comparison between actual class and predicted class for random forest classifier

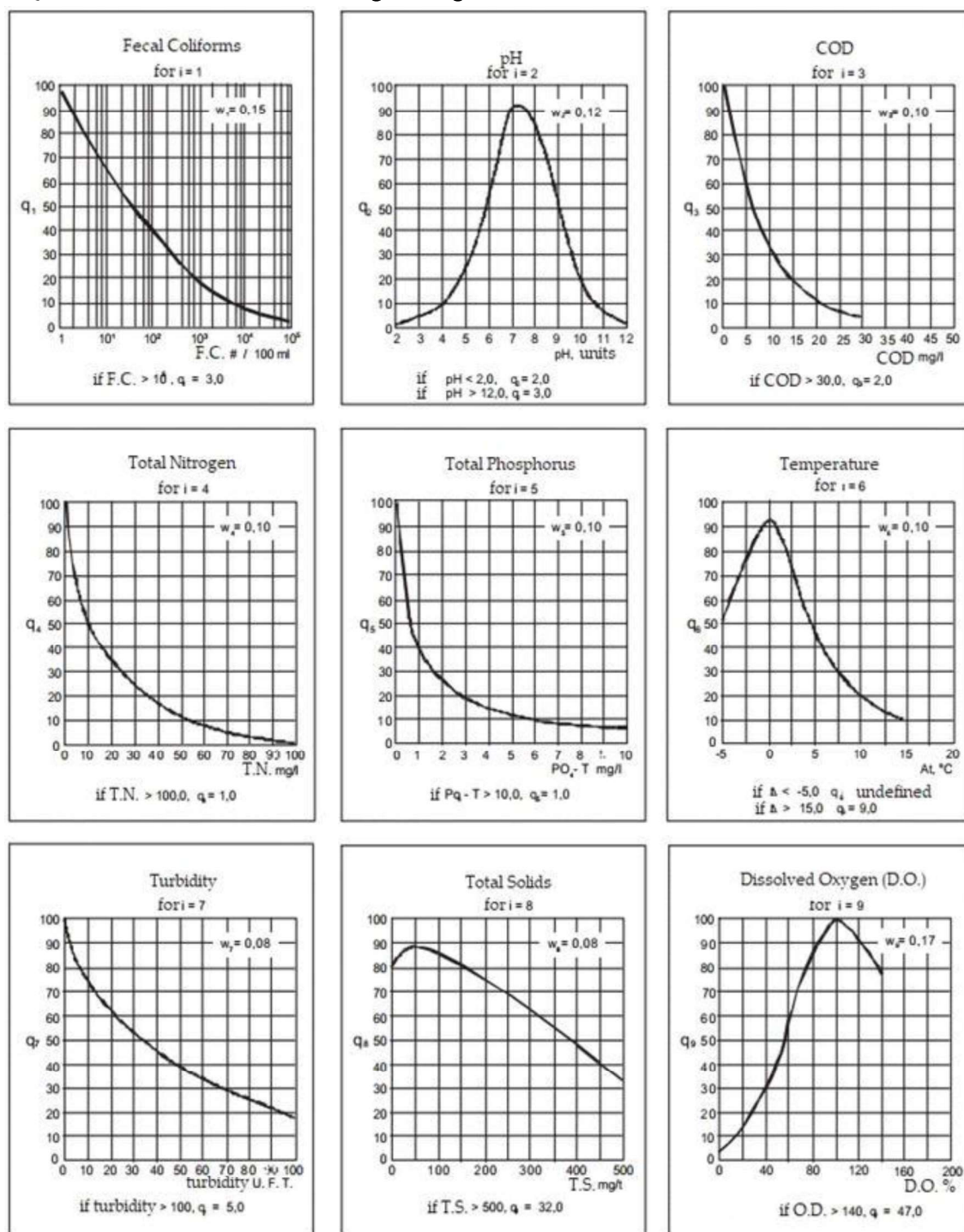
CHAPTER 6 CONCLUSION AND FUTURE WORK

The purpose of this study was to develop smart solutions for agriculture for reducing time, cost and resource utilization. This work mainly focuses on proper utilization of water for irrigation purposes. Data from various chemical sensors are gathered and WQI is obtained. A model is developed for prediction and classification of WQI using fewer number of parameters that can be obtained using cheap sensors. The classified Water Quality Data can be used in making decisions for selection of crops and soil preparation. Also, a model for prediction of missing parameter values is developed which can be used before data preprocessing for any of the sensors. Overall, satisfactory results were obtained for all of the models and these can be incorporated in any smart agricultural system that measures the parameters required by the models.

In future, we plan to work on other phases of agriculture such as soil preparation, crop selection, fertilizing and harvesting and developing models to analyze and assist in them. We plan to develop a mobile/web application that will provide an interface for using functionalities developed in this project. We plan to explore other possible domains of IOT such as networking, developing software for hardware like raspberry PI and Arduino and automation in Farming.

APPENDIX A

The Q-values are calculated according to “Fig. A.1”



[Figure A.1] Q values for various parameters

The following table can be used to obtain Q-values of Fecal coliform, BOD and Dissolved Oxygen

Table A.1 Q-val Conversion chart for FC, BOD and DO

Fecal Coliforms (per 100 ml)	Q-FC	BOD (mg/L)	Q-BOD	Dissolved Oxygen (saturation)	Q-DO
1	99	0	100	0	2
2	91	1	95	5	5
3	86	2	80	10	7
4	82	3	67	15	10
5	80	4	61	20	12
6	78	5	56	25	15
7	76	6	51	30	19
8	74	7	46	35	23
9	73	8	42	40	30
10	72	9	38	45	37
20	63	10	34	50	44
30	58	11	30	55	51
40	55	12	28	60	57
50	52	13	25	65	66
60	50	14	23	70	75
70	48	15	20	75	81
80	47	16	18	80	87
90	45	17	16	85	91
100	44	18	14	90	95

200	37	19	13	95	98
-----	----	----	----	----	----

The following table can be used to obtain Q-values of pH, Temperature and Nitrate.

Table A.2 Q-Val Conversion chart for pH, Temperature and Nitrate

pH	Q	Temp. change (deg C)	Q	Nitrate (mg/L)	Q
2.4	3	-7	66	3	90
2.6	3	-6	70	4	70
2.8	4	-5	74	5	65
3	4	-4	78	6	60
3.2	5	-3	82	7	58
3.4	6	-2	85	8	56
3.6	7	-1	89	9	53
3.8	8	0	93	10	51
4	9	1	89	12	48
5.2	33	7	61	24	33
5.4	38	8	56	26	31
5.6	44	9	50	28	29
5.8	49	10	45	30	27
6	55	11	40	32	25
6.2	60	12	36	34	23
6.4	68	13	34	36	21
6.6	75	14	33	38	19
6.8	83	15	31	40	18
7	88	16	29	42	16

7.2	92	17	27	44	15
7.4	92	18	26	46	13
7.6	92	19	24	48	12
7.8	90	20	22	50	10

REFERENCES

- [1] India economic survey 2018
<https://www.financialexpress.com/budget/india-economic-survey-2018-for-farmers-agriculture-gdp-msp/1034266/>
- [2] Precision Agriculture, International Society for Precision Agriculture
<https://www.ispag.org/about/definition>
- [3] Anneketh Vij, Singh Vijendra, Abhishek Jain, Shivam Bajaj, Aashima Bassi, Arushi Sharma, IoT and Machine Learning Approaches for Automation of Farm Irrigation System, *Procedia Computer Science*, Volume 167, 2020, Pages 1250-1257, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2020.03.440>.
- [4] Janani M and Jebakumar R. A Study on Smart Irrigation Using Machine Learning. *Cell & Cellular Life Sciences Journal* ISSN: 2578-4811, 2019, DOI:10.23880/cclsj-16000141
- [5] De La Mora-Orozco C, Flores-Lopez H, Rubio-Arias H, Chavez-Duran A, Ochoa-Rivero J. Developing a Water Quality Index (WQI) for an Irrigation Dam. *Int J Environ Res Public Health*. 2017;14(5):439. Published 2017 Apr 29. doi:10.3390/ijerph14050439
- [6] Jefferson L. Lerios, Mia V. Villarica, Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir, *International Journal of Mechanical Engineering and Robotics Research* Vol. 8, No. 6
- [7] Research on soil moisture prediction model based on deep learning
 Cai Y., Zheng W., Zhang X., Zhangzhong L., Xue X.
 (2019) *PLoS ONE*, 14 (4), art. no. e0214508
- [8] Gholap, Jay & Ingole, Anurag & Gohil, Jayesh & Gargade, Shailesh & Attar, Vahida. (2012). Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction. *International Journal of Computer Science Issues*. 9.
- [9] Gad, I., Hosahalli, D., Manjunatha, B.R. et al. A robust deep learning model for missing value imputation in big NCDC dataset. *Iran J Comput Sci* (2020). <https://doi.org/10.1007/s42044-020-00065-z>
- [10] Fabrizio Balducci, Donato Impedovo * ID and Giuseppe Pirlo Dipartimento di Informatica, Università degli studi di Bari Aldo Moro, 70125 Bari, Italy; fabrizio.balducci@uniba.it (F.B.); giuseppe.pirlo@uniba.it (G.P.)
- [11] M.S. Suchithra, Maya L. Pai, Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters, *Information Processing in Agriculture*, Volume 7, Issue 1, 2020, Pages 72-82, ISSN 2214-3173, <https://doi.org/10.1016/j.inpa.2019.05.003>.
- [12] Meeradevi and Monica R Mundada, Automated Control System for Crop Yield Prediction using Machine Learning Approach, *International Journal of Applied Engineering Research* ISSN 0973-4562 Volume 14, Number 2 (2019) pp. 480-484
- [13] Horton RK. An index number system for rating water quality. *Journal of Water Pollution Control Federation*. 1965;37(3):300-305
- [14] Grisham, Thomas. (2009). The Delphi technique: A method for testing complex and multifaceted topics. *International Journal of Managing Projects in Business*. 2. 10.1108/17538370910930545.
- [15] Brown RM, McClelland NI, Deininger RA, Tozer RG. A water quality index—Do we dare? *Water & Sewage Works*. 1970;117:339-343
- [16] Francis Galton. "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute*, 15:246-263 (1886)
- [17] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. *R News* 2(3), 18--22.

- [18] Friedman, J. H. (February 1999). "[Greedy Function Approximation: A Gradient Boosting Machine](#)"
- [19] [Gergonne, J. D.](#) (November 1974) [1815]. "The application of the method of least squares to the interpolation of sequences". *Historia Mathematica* (Translated by Ralph St. John and [S. M. Stigler](#) from the 1815 French ed.). 1 (4): 439–447. [doi:10.1016/0315-0860\(74\)90034-2](#)
- [20] Günther, F.; Fritsch, S. *Neuralnet: Training of neural networks*. R J. 2010, 2, 30–38.
- [21] Daniel G.G. (2013) Artificial Neural Network. In: Runehov A.L.C., Oviedo L. (eds) *Encyclopedia of Sciences and Religions*. Springer, Dordrecht. [https://doi.org/10.1007/978-1-4020-8265-8_200980](#)
- [22] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- [23] Meireles ACM, Andrade EM, Chaves LCG, Frischkorn H, Crisostomo LA. A new proposal of the classification of irrigation water. *Revista Ciência Agronômica*. 2010;41(3):349-357. DOI: 10.1590/S1806-66902010000300005
- [24] Fürnkranz J. (2011) Decision Tree. In: Sammut C., Webb G.I. (eds) *Encyclopedia of Machine Learning*. Springer, Boston, MA. [https://doi.org/10.1007/978-0-387-30164-8_204](#)
- [26] Refaeilzadeh P., Tang L., Liu H. (2009) Cross-Validation. In: LIU L., ÖZSU M.T. (eds) *Encyclopedia of Database Systems*. Springer, Boston, MA. [https://doi.org/10.1007/978-0-387-39940-9_565](#)
- [27] Divyanshu Mishra .(2019, Dec. 7) Regression: An Explanation of Regression Metrics And What Can Go Wrong [Online]. Available: [https://towardsdatascience.com/regression-an-explanation-of-regression-metrics-and-what-can-go-wrong-a39a9793d914#:~:text=Root%20Mean%20Squared%20Error%3A%20RMSE,value%20predicted%20by%20the%20model.&text=The%20MAE%20is%20more%20robust,errors%20as%20extremely%20as%20mse.](#)
- [28] Harikrishnan N B. (2019, Dec. 19). Confusion Matrix, Accuracy, Precision, Recall, F1 Score [Online]. Available: [https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd](#)
- [29] Simm J, de Abril I and Sugiyama M (2014). Tree-Based Ensemble Multi-Task Learning Method for Classification and Regression, volume 97 number 6. [http://CRAN.R-project.org/package=extraTrees](#).
- [30] Water Quality Database [Online]. Available: [http://www.cpcbenvi.nic.in/water_quality_data.html](#)
- [31] Enslin, Clarissa L., Godsey, Sarah, Marks, Danny G. (2017). Eleven years of mountain weather, snow, soil moisture and stream flow data from the rain-snow transition zone - the Johnston Draw catchment, Reynolds Creek Experimental Watershed and Critical Zone Observatory, USA. v1.1 [Online]. Available: [https://agris.fao.org/agris-search/search.do?recordID=US2019X00237](#)