



AFRICAN CENTER OF EXCELLENCE
IN
INTERNET OF THINGS



UNIVERSITY OF RWANDA
COLLEGE OF SCIENCE AND TECHNOLOGY

Research Thesis Title:

Employing Machine Learning and Internet of Things Based Real Time NPK
Fertilizer Prediction for Cassava crop in Rwanda

Submitted By:

MUNEZERO Alphonse (REF.NO: 219014102)

A dissertation Submitted in partial fulfilment of the requirements for the award of

MASTERS OF SCIENCE DEGREE IN INTERNET OF THINGS-EMBEDDED COMPUTING SYSTEMS

November, 2022

Declaration

I, **Mr. MUNEZERO Alphonse**, hereby declare that this research proposal report is my original work and has not been submitted before for any academic award either in this or other institutions of higher learning for academic publication or any other purpose. The references used here from other journals or materials are indicated in the references section.

Name: **MUNEZERO Alphonse**

REG.NO: **219014102**

Signature:

Date: **28th February, 2023**

Bonafide Certificate

This is to certify that this submitted Research Thesis work report is a record of the original work done by **Mr. MUNEZERO Alphonse (REF.NO: 219014102)**, MSc. IoT-ECS Student at the University of Rwanda / College of Science and Technology / African Center of Excellence in Internet of Things. Certified further, that according to the best of our knowledge; the work reported here doesn't form a part of any other research work.

Main Supervisor:

Dr. UWITONZE Alfred

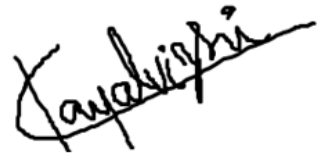
Signature:

Date:/...../.....

Co-Supervisor:

Prof. KAYALVIZHI Jayavel

Signature:



Date:/...../.....

THE HEAD OF MASTERS AND TRAINING

Dr. James RWIGEMA

Signature:

Date:

Acknowledgement

First, I want to thank the Almighty God for bringing me this far and keeping me sound and safe with good life during the entire academic period.

The development of this report took the effort, support and guidance of a number of people whom I wish to thank.

I acknowledge the continuous encouragement, supervision, timely suggestions and inspiration guidance offered by Dr.UWITONZE Alfred and Prof. KAYALVIZHI Jayavel who brought this research thesis at a successful completion.

I am grateful thanks to the management of Ministry of Agriculture, Rwanda Agriculture Board (RAB) and Ministry of Local Government for their guidance to make this research successful.

Finally, I express my sincere thanks to all of management of African Center of Excellence in Internet of Things for their helpful guidance and unlimited support. I'm grateful to all the lectures and all my beloved classmates who have patiently extended all kinds of help for accomplishing this undertaking.

Abstract

In the Agriculture sector the pressure is increased immensely due to the rise in population. In this present year we mainly witness a move from traditional methods were used in the agriculture to the advanced technology. Machine learning and IoT technologies have transformed the quality and the yield production. In fact IoT helps to collect real time data from the field via sensing technique and machine learning analyzes those data from sensors for generating information that will help in business growth. Knowing a suitable crop and soil fertilizer option is one of the things that can make a farmer more productive and help him to avoid losses.

Soil fertilization activities contribute a lot in crops production volume. However, if the quantity of soil composition (fertilizer) is not controlled and maintained consistent, this may lead to less crop production volume. Choosing the appropriate crop type and the corresponding quantity of soil fertilizer is one of the measures to be taken prior for preventing the inferior quality and less quantity of the crop production. Thus, the measurement of soil nutrients is greatly required for better plant growth and fertilization. Therefore, temperature, soil moisture, Nitrogen, phosphorus pentoxide, potassium oxide, pH level are among the parameters commonly measured to monitor the cassava crop as they are the ones mostly important and informative soil parameters to determine the soil fertility.

This research will focus on “Employing Machine Learning and Internet of Things Based Real Time Fertilizer Prediction for Cassava crop in Rwanda” by using a machine learning (ML) algorithm to build a model which may help the farmers to predict the cassava fertilizer components. Through this research, different parameters are respectively controlled by a network of sensors such as; temperature sensors, soil moisture sensors, soil nutrients sensors, and PH sensors then the data corresponding to these parameters will be feed to the different machine learning algorithms such as Linear Regression, Random Forest, Gradient Boosting, Random Forest, K-Nearest Neighbors and Decision Tree will be tested for optimizing the prediction accuracy using python programming packages. These algorithms have been selected because they are mostly used in classification and regression problems.

Keywords: Cassava crop, Machine Learning (ML), Soil fertility, Internet of Things (IoT), SQLite server.

List of symbols and Abbreviations

API: Application Programming Interface

Fig: Figure

HTTP: Hypertext Transfer Protocol

ICT: Information Communication Technology

IoT: Internet of Things

KNN: K-Nearest Neighbors

MCU: Microcontroller Unit

ML: Machine Learning

MLP: Multi-Layer Perceptron

MoA: Ministry of Agriculture

MSE: Mean Root Square Error

NPK: Nitrogen Phosphorus Potassium

RAB: Rwanda Agriculture Board

SDGs: Sustainable Development Goals

SVR: Support Vector Machine

TCP: Transmission Control Protocol

UR: University of Rwanda

WIFI: Wireless Fidelity

List of Figures

| | | |
|--|---|----|
| FIGURE 1: RESEARCH APPROACH..... | 9 | |
| FIGURE 2: DS18B20 WATERPROOF DIGITAL TEMPERATURE SENSOR..... | 11 | |
| FIGURE 3: NPK SENSOR | 11 | |
| FIGURE 4: PH SENSOR..... | 12 | |
| FIGURE 5: ESP8266 NODEMCU | 13 | |
| FIGURE 6: DECISION TREE CLASSIFIER | 15 | |
| FIGURE 7: BUILDING RANDOM FOREST ALGORITHM | 16 | |
| FIGURE 8: MODEL TRAINING AND EVALUATION | 18 | |
| FIGURE 9: SENSING SUBSYSTEM | 19 | |
| FIGURE 10: HTTP COMMUNICATION | 20 | |
| FIGURE 11: WIRELESS COMMUNICATION SYSTEM..... | 21 | |
| FIGURE 12: DETAILED BLOCK DIAGRAM OF SOIL DATA COLLECTION SYSTEM | 23 | |
| FIGURE 13: SYSTEM FLOWCHART FOR ESP8266 NODEMCU | 24 | |
| FIGURE 14: DATASET DESCRIPTION..... | 26 | |
| FIGURE 15: NITROGEN SAMPLE | 27 | |
| FIGURE 16: PHOSPHORUS SAMPLES | 28 | |
| FIGURE 17: POTASSIUM SAMPLES..... | 28 | |
| FIGURE 18: CORRELATION FOR NITROGEN | FIGURE 19: CORRELATION FOR PHOSPHORUS | 29 |
| FIGURE 20: CORRELATION FOR POTASSIUM | 29 | |
| FIGURE 21: NITROGEN FEATURES IMPORTANCE | FIGURE 22: PHOSPHORUS FEATURES IMPORTANCE | 30 |
| FIGURE 23: POTASSIUM FEATURES IMPORTANCE | 30 | |
| FIGURE 24: HARDWARE IMPLEMENTATION | 33 | |

List of Table

| | |
|------------------------------|--------------------------------------|
| TABLE 1: ON TRAINING DATASET | TABLE 2: ON TESTING DATASET 31 |
| TABLE 3: ON TRAINING DATASET | TABLE 4: ON TESTING DATASET 32 |
| TABLE 5: ON TRAINING DATASET | TABLE 6: ON TESTING DATA 32 |

Table of Contents

| | |
|---|-----|
| Declaration | ii |
| Bonafide Certificate | iii |
| Acknowledgement | iv |
| Abstract | v |
| List of symbols and Abbreviations | vi |
| List of Figures | vii |
| Chapter 1: General Introduction | 1 |
| 1.1 Background and Motivation..... | 1 |
| 1.2 Problem Statement | 2 |
| 1.3 Study Objectives | 4 |
| 1.3.1 General Objective | 4 |
| 1.3.2 Specific Objectives | 4 |
| 1.4 Hypotheses..... | 5 |
| 1.5 Study Scope | 5 |
| 1.6 Significance of the Study | 5 |
| 1.7 Organization of the Study | 5 |
| Chapter 2: Literature Review | 6 |
| Chapter 3: Research Methodology..... | 8 |

| | |
|---|----|
| 3.1 Overview | 8 |
| 3.1.1 Data Collection and Visualization | 8 |
| 3.1.2 Research Approach and Design of the System | 8 |
| 3.2 Data collection instruments..... | 9 |
| 3.2.1 Hardware and Software requirements..... | 10 |
| 3.2.1.1 Sensors | 10 |
| 3.2.1.2 Gateway and Communication Technology Requirements | 12 |
| 3.2.1.3 Cloud and IoT Analytics Platform requirements | 13 |
| 3.3 Data Preparation and Wrangling..... | 13 |
| 3.4 Predictive Models Modelling..... | 14 |
| 3.4.1 Logistic Regression..... | 14 |
| 3.4.2 Decision Tree Classifier..... | 15 |
| 3.4.3 Random Forest Classifier..... | 15 |
| 3.4.4 Gradient Boosting Classifier | 16 |
| 3.4.5 K-Nearest Neighbours Classifier (KNN)..... | 16 |
| 3.5 Model Training and Evaluation | 17 |
| Chapter 4: System Analysis and Design | 19 |
| 4.1 Introduction..... | 19 |
| 4.2 Sensing Subsystem..... | 19 |
| 4.3 Wireless Communication system with ESP8266 NodeMcu..... | 20 |

| | |
|---|----|
| 4.4 List of components and materials used | 22 |
| 4.5 System Block Diagram | 22 |
| 4.6 System flowchart diagram | 23 |
| 4.7 API design and communication protocol..... | 25 |
| Chapter 5: Results and Analysis | 26 |
| 5.1 Training and evaluation implementation | 26 |
| 5.2 Dataset Description..... | 26 |
| 5.3 Correlation computation | 29 |
| 5.4 Features Extraction | 30 |
| 5.5 Training and testing dataset | 31 |
| 5.6 Results of model training..... | 31 |
| 5.7 Real time soil data collection Hardware implementation..... | 33 |
| Chapter 6: Conclusion and Recommendation..... | 34 |
| LIST OF REFERENCES | 35 |
| APPENDICES | 39 |
| Appendix 1: Soil Dataset | 39 |
| Appendix 2: Python code for decision tree ML algorithm | 40 |
| Appendix 3: Python code of random forest algorithm..... | 40 |
| Appendix 4: Python codes of KNN | 40 |

Chapter 1: General Introduction

1.1 Background and Motivation

Nowadays, the usage of ICT in agriculture sector is increasing tremendously in the different countries of Africa Continent country. An agriculture sector is one among the factors that play a vital role in the development of the African continent. These are justified by a high percentage of citizens who are participating in this sector. The rate of increasing of the population in Africa is high and the demand for increasing the food production must be proportional. So, it is very important for taking measures on how to increase the crop production. The Crop production is mainly depending on the soil proper- ties of plant interaction. It is very important for the farmers to determine the crop type and soil fertility requirement for better and economical crop production. Soil pH is the most important parameter because it gives more information about many aspects of the soil fertility [1]. Major soil nutrients present in soil and that contribute in yield production include Phosphorus, Potassium, Nitrogen, Calcium and pH [1, 4]. Due to the insufficient rate of nutrients or excess fertilization may lead to the lower yield in the crop production [4]. However, the appropriate crop type and quantity of fertilizer is required for better plant growth.

In Rwanda and even in the other countries of African continent, most farmers use to imagine the crop type and quantity fertilizer to be used during planting where the problem of bad selection of crop type and using much or less fertilizers is possible. Measuring the nutrients concentration present in the soil can help to get the soil nutrients to be provided and select the suitable fertilizer for the specific soil sample identified. Thus, the use of ICT in agriculture activities could play a vital role in sustainability of optimum agriculture and reducing the environmental impacts and economic losses.

It is in line the researcher in this current research, used the Machine Learning Algorithms concept to generate a predictive model that might help the farmers to know the quantity of soil composition and identification of the crops to be planted in the environment based on the soil composition predicted or identification of the quantity of nutrients to be added in the soil for keeping the soil fertility consistent. The dataset used by the machine learning algorithm includes different data such as nitrogen, phosphorus, potassium, temperature, soil moisture and PH data have been collected by the sensors through the soil samples for determining the soil nutrients.

Therefore, there is a need of an adequate technology to provide the adequate spatial data. The targeted values were nitrogen (N), phosphorous (P) and potassium (K). For selecting the best predictive model different machine learning algorithms such Linear Regression, K-Nearest Neighbors (KNN), Lasso, Ridge, Decision Tree, Random Forest and Gradient Boosting were implemented and investigated using python programming language packages to confirm the best algorithm for this particularly dataset. Python has seem to be a stable, flexible and popular language and makes many tools available for the researchers from development to deployment and maintenance of an AI project [7, 9].

1.2 Problem Statement

The crop production mainly depends on the rate of soil nutrients. It is important for the farmers to determine the soil fertility requirement for increasing the crop production yield. Due to having insufficient information needed about soil composition, this might cause a serious problem of planting without nutrients present in the soil knowledge. However, this may lead to low agriculture production quality and less crop production yield because using inappropriate rate of soil nutrients which in return lead to the crops degradation.

Improper usage of soil fertilizers may result into the poor quality of production [1]. For instance in China inappropriate usage of fertilizer caused the low product quality and even critical environmental problems [10]. Thus this current research evaluated various Machine Learning algorithms [11, 12] to confirm a good predictive model which allows the farmer to predict the appropriate fertilizer for the cassava crop based on the nutrients present in the soil.

Although these previous researches have been undertaken, the soil data used has been collected by the independent institutions and the researchers have not addressed the methodology approach in which the generated predictive models would be operated with IoT based real time data from field sensors for helping the predictive models to generate the predictions in the future. Thus, we found it important to investigate further research that includes also an IoT based real time data collection and continue reviewing different machine learning algorithms because the prediction accuracy depends on the quality of data collected, methodology approach used for analyzing the data, political situation and the social economic activities of any country. There is a need also to contextualize research broaden in different countries because the factors that contribute to the growth of cassava crop through the nutrients present in the soil might vary from one country with another based on the geographic location. Therefore, we need to develop a specific predictive model to be applied in the Rwandan context based on the real situation.

1.3 Study Objectives

1.3.1 General Objective

The main objective of this current project is to implement different machine learning algorithms using python libraries for generating a best predictive model to be used for helping the famers to know the type of fertilizer contents for cassava crop based on the type of the soil samples collected by using different sensors.

1.3.2 Specific Objectives

To achieve the main goals the research thesis has the following specific objectives

- To analyze the relationship between cassava crop production and the soil data factors that impact development of its production such as temperature, soil moisture, nitrogen (N), phosphorus pentoxide(P), potassium oxide (K), pH level in Rwanda
- Implement IoT based real time soil data collection system that uses the predictive model that will be generated to make prediction in the future.
- To implement and evaluate different machine learning classification models for optimizing soil fertilizer prediction accuracy.

1.4 Hypotheses

With the help of machine learning and internet of things based technologies, it is possible to predict the real fertilizer components for the cassava crop in the future and also helping in its general crop production.

1.5 Study Scope

This research study was carried out to design and implement a system that should predict the real fertilizer components for the cassava crop in the future based on soil data change such as temperature, soil moisture, nitrogen (N), phosphorus pentoxide (P), potassium oxide (K), and pH level in Rwanda specifically in Ruhango district.

1.6 Significance of the Study

The output from this research will help the ministry of agriculture here in Rwanda to monitor cassava crop production where farmers will be able to get timely information on selecting a particular fertilizer type based on soil conditions. This will also contribute to the crop quality and productivity and financial statement of farmers will also be improved.

1.7 Organization of the Study

This thesis report is organized into six chapters as follow: The first chapter deals with general induction about the project. The second chapter gives a brief description about the previous related research and the gaps identified. The third chapter shows different methodology approaches used by the research to carry out the research. The fourth chapter gives details about the system design and simulation models used. The fifth chapter explains the machine learning evaluation metrics used and results found throughout the research. Finally, the last chapter gives conclusion about the research and recommendation for the country and the future researchers.

Chapter 2: Literature Review

This part gives the brief description and analysis of existing related research projects. These include the problem investigated by the previous researchers, proposed technical solution, methodology used, and results found. Finally, the gaps identified through those existing project have been explained as the motivation of the current proposed research.

Through [1], Shylaja S.N. and Dr.Veena M.B.have developed a wireless sensor network architecture for collecting the soil nutrients information. Through this project the major soil nutrients collected were Potassium K, Phosphorous P and Nitrogen N. These data were collected by using the sensor technology and in return the collected information were sent to IBM cloud platform database to be stored through IoT backbone architecture. This system provided the people interface for accessing to the data through their mobile phone.

In [13], R. Ajith kumar et. al, have done a deep research on statistical soil nutrients analysis on three thousand and eight hundred soil samples of Thrissur district by using R software. The main nutrients analyzed were soil pH value, organic carbon, and electrical conductivity, phosphorus, potassium, calcium, Magnesium, Sulfur, Zinc, Boron, Iron, copper and Manganese. Their results of the study shows that there is a strong correlation between those soil nutrients in Thrissur district.

Amrutha A et.al [3] developed a system for collecting the soil nutrients information from the soil by using sensors technology. The interested soil properties were Phosphorous, Potassium and Calcium. At the end they have developed an automated system for identifying the amount of nutrients to be added based on the measured nutrients from the soil for avoiding using excess or insufficient fertilizers which may lead to plant degradation.

The researcher in [4], identified a soil fertility analyzer system with a Soil Test Kit. His study aimed to predict the soil nutrients present in the soil using image processing and artificial neural network implemented by using Matlab. Those nutrients include soil pH, Zinc, Phosphorus, Potassium, Nitrogen and Calcium.

Viraj. A. Gulhane et.al [14] preliminary have investigated the soil properties which include electrical conductivity, carbon, pH level, phosphorous and potassium contents of soil samples. Secondary they have collected soil spatial data by using the remote sensing data source which is satellite from improving the predictive model. Then the data have been analyzed via Matlab and the results showed the strong correlation between soil nutrients and wavelet transformation.

After analyzing different studies that have been undertaken by different researchers worldwide on soil data prediction by using different machine learning algorithms as clarified in the above research studies, we find that those different studies generated different values of prediction accuracy and even the best model candidate varies from country to country.

Sometimes, machine learning projects may require a huge amount and good quality of data for allowing the generated model to make the best prediction with high accuracy [12]. This indicates that with different geographic locations and even the vegetation, habitation, and political situation might impact the predictive accuracy [2].

An IoT is an interconnection of physical devices embedded with electronics and software, applications and virtual world for allowing different components of the system to communicate and exchange information via sensors connection and internet access, and finally providing data exchange with manufactures, operators and other connected devices [15, 16]. Internet of Things allows the physical objects embedded with sensors to be monitored and controlled remotely at any time and at any place by using existing network connectivity.

The research gaps identified in the existing related works are as follows:

- The previous researchers have done only the work of collecting soil data in order to find out information about the status of the soil and some of them are using prediction accuracy as evaluation metrics, however the prediction accuracy alone tends to hide useful information about model performance
- Lack of information about model performance on both training and testing data for some researchers.
- There is a lack of an automated tool that employs the model to make the real time prediction

Chapter 3: Research Methodology

3.1 Overview

For conducting this research activity, we have used different approaches. Thus, this section gives a brief explanation about research methodology approaches used by the researcher throughout in this study. These include data collection, data visualization, data preparation and wrangling, machine learning models training and evaluation, designing and implementation of real-time IoT based system, and finally integration of IoT system with machine learning predictive model for doing prediction on real time data from field sensors.

3.1.1 Data Collection and Visualization

For implementing this research study, the researcher has collected different types of sample from different soil data. These data include temperature, soil moisture, nitrogen (N), phosphorus pentoxide (P), potassium oxide (K), and pH level in Rwanda especially in Ruhango district. The next paragraphs explain in details how these data were collected.

3.1.2 Research Approach and Design of the System

This part describes the overview of the research approaches and the steps involved in system development from the step of gathering the ideas to the final step of prototype and getting result.

The development approach of this research thesis complies with two stages:

- The algorithm and flowcharts design approach
- Prototyping approach

In this research thesis, the existing systems are analysed and a new system used machine learning technology was developed.

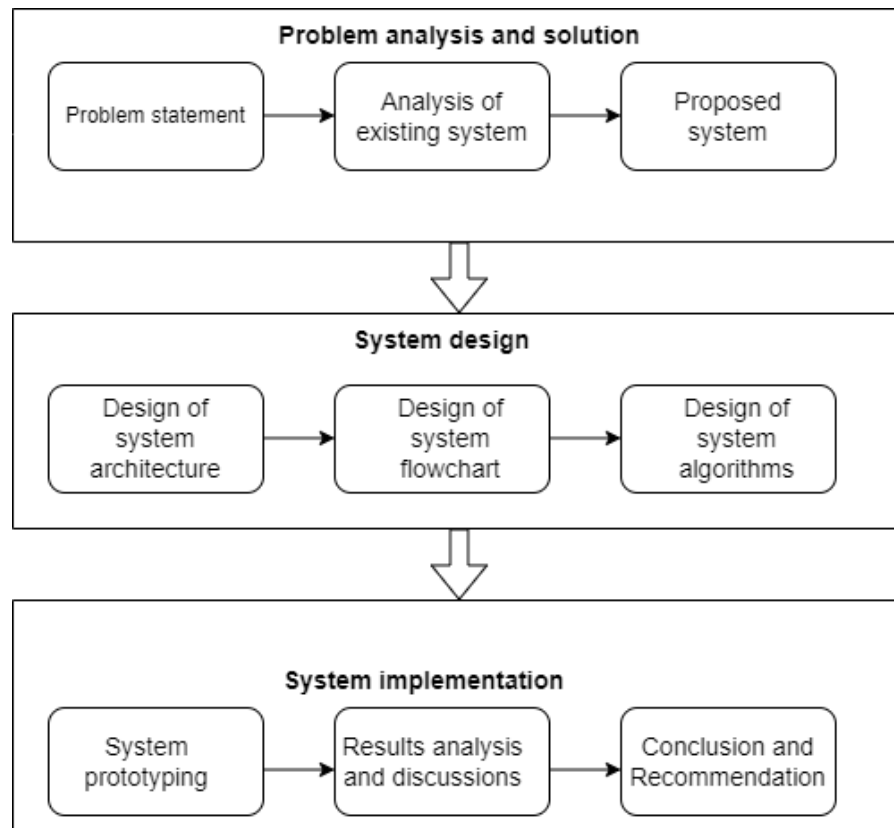


Figure 1: Research Approach

3.2 Data collection instruments

This study used primary and secondary information. The information provided the basis for the development of the various components of the application. The techniques included are:

- Literature review from many sources such as; thesis, government report, conference papers, and journals covered a large portion of the published information. The main sources of information on soil data and the way of collecting them were properly provided.
- Qualitative interview: this is a type of interview where the interviewer has no specific pre-set question that was to be asked in a particular order. The respondent does most of the talking. These interviews were used during site visits and were used to gather in-depth insights into how the farmers get information about soil status and how they know the amount of fertilizer to use when they are going to plant cassava.

3.2.1 Hardware and Software requirements

Sensors and actuators requirements Specifications:

Sensor is a device that when exposed to a physical phenomenon (temperature, displacement, force, etc) produces a proportional output signal (electrical, mechanical, magnetic, etc) while an Actuator is a component of a machine that is responsible for moving or controlling a mechanism or system. Sensors and actuators are two critical components of every closed loop control system. The controller accepts the information from the sensing unit, makes decisions based on the control algorithm, and outputs commands to the actuating unit. [17]

The following are specifications of sensors and actuators used in this research:

- pH Sensor module agricultural with power supply 5-10V, measuring range 4-10pH and resolution of 0.1pH.
- NPK sensor with Op range 0o C-50oC, accuracy of $\pm 3\%$, 5V-24V, measuring range: 0-1999mg/kg and response time less 10 seconds
- DS18B20 waterproof digital temperature sensor, 3V/5V, 2.5Ma and sampling period of 1second

3.2.1.1 Sensors

- DS18B20 Waterproof Digital Temperature sensor

This Maxim-made item is a digital thermo probe or sensor that employs DALLAS DS18B20. Its unique 1-wire interface makes it easy to communicate with devices. It can convert temperature to a 12-bit digital word in 750ms (max). Besides, it can measure temperatures from -55°C to $+125^{\circ}\text{C}$ (-67°F to $+257^{\circ}\text{F}$). In addition, this thermo probe doesn't require any external power supply since it draws power from the data line. Last but not least, like other common thermo probes, its stainless steel probe head makes it suitable for any wet or harsh environment. This sensor is mounted on the Node MCU and is used for sensing the water temperature from time to time.[18]



Figure 2: DS18B20 Waterproof Digital Temperature sensor

- NPK sensor

Principle of optical NPK sensors is based on the interaction between incident light and soil surface properties, such that the characteristics of the reflected light vary due to the soil physical and chemical properties [5].



Figure 3: NPK sensor

- pH sensor

This pH Sensor can be used for any lab or demonstration that can be done with a traditional pH meter, including: acid-base titrations, monitoring pH in an aquarium, and investigating the water quality of streams and lakes.[19] This sensor is mounted on Node MCU and is used for sensing the water pH from time to time.



Figure 4: pH sensor

3.2.1.2 Gateway and Communication Technology Requirements

An Internet of Things gateway is a physical device or software program that serves as the connection point between the cloud and controllers, sensors and intelligent devices. All data moving to the cloud or vice versa, goes through the gateway, which can be either a dedicated hardware appliance or software program. The IoT gateway provides a complex representation of the instances. The main tasks the IoT gateway perform are: Data forwarding, Gateway management, Device management, Data analysis and Diagnostics [20].

In this project, ESP8266 WiFi Module will be used as a Gateway and WiFi as Communication Technology.

The ESP8266 WiFi Module is a self-contained SOC with integrated TCP/IP protocol stack that can give any microcontroller access to your WiFi network. The ESP8266 is capable of either hosting an application or offloading all Wi-Fi networking functions from another application processor. Each ESP8266 module comes pre-programmed with an AT command set firmware, meaning, you can simply hook this up to your Arduino device and get about as much WiFi-ability as a WiFi Shield offers. [21].

This module has a powerful enough on-board processing and storage capability that allows it to be integrated with the sensors and other application specific devices through its GPIOs with minimal development up-front and minimal loading during runtime. The following figure illustrates the ESP8266 12-E chip pinout.

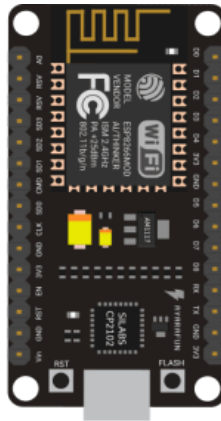


Figure 5: ESP8266 NodeMCU

3.2.1.3 Cloud and IoT Analytics Platform requirements

In the simplest terms, cloud computing means storing and accessing data and programs over the internet instead of your computer's hard drive. The cloud is also not about having a dedicated network attached storage hardware or server in residence, it's as a service. Cloud computing is intelligent. With all the various data stored on the computers in a cloud, data mining and analysis are necessary to access that information in an intelligent manner [22].

3.3 Data Preparation and Wrangling

Several categories of data used throughout this study were collected through the soil. The soil data used contains 498 samples recorded by the researcher from Ruhango district especially in Mbuye sector. However, nitrogen, phosphorus, potassium, temperature, soil moisture and pH data have been collected in samples from April 2022 to July 2022. Each sample has contained six data sources as mentioned above. Therefore, we have assumed that the soil data can be changed in a remarkable way after a certain period. Finally, the data have been combined all together to form a single meaningful dataset, of 498 records with six input variables type defined as dependent variables and independent variables.

3.4 Predictive Models Modelling

This part gives a detailed explanation and implementation procedures of different machine learning classification algorithm used by the researcher throughout this research work. These algorithms have been used for mapping the relationship between the dependent variables (model input) and independent variables (model target). They include Logistic Regression, Ridge, Lasso, Random Forest Classifier, K-Nearest Neighbors Classifier, Support Vector Regression Classifier, Decision Tree Classifier and Gradient Boosting Classifier. These algorithms were selected because they are popular in solving machine learning classification problems and availability of huge documentation because a large community is using them. The next paragraphs give a brief description about each one among these stated algorithms above.

3.4.1 Logistic Regression

A Logistic regression is a type of machine learning supervised algorithm used to make prediction when the target variables are discrete or categorical and commonly known in solving binary classification problems such as spam detection, cancer detection, anomaly detection. Unlike Linear regression which predicts unbound values, for logistic regression the range of predicted values is known [23]. The mathematical expression of logistic regression is given by $F(X)$:

$$F(X) = \text{sigmoid}(WX + b)$$

Here, \mathbf{X} is the input feature vector. \mathbf{W} , \mathbf{b} and **sigmoid** are the weight vector, bias and activation function respectively. Weight vector and bias are the model parameters to be identified during of model training. The sigmoid function or activation function is used for mapping the values between 1 and 0. If the output of the sigmoid function is above 0.5 we can classify this as 1 and 0 is the output is below 0.5[23].

$$\text{Sigmoid}(x) = 1 / (1 + e(-x))$$

3.4.2 Decision Tree Classifier

Decision tree classifier is a supervised machine learning algorithm used in machine learning and in statistics when the target variables are categorical. This predicting modelling approach uses a tree-like graph as a predictive model where observations are represented the branches and target values or the actual output or class represented in the leaves. The goal of this algorithm is to build a predictive model that can predicts the target value by learning decision rules identified from the features. These rules are implemented by using if-then-else statements. Decision trees generates predictions by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the observations [24].

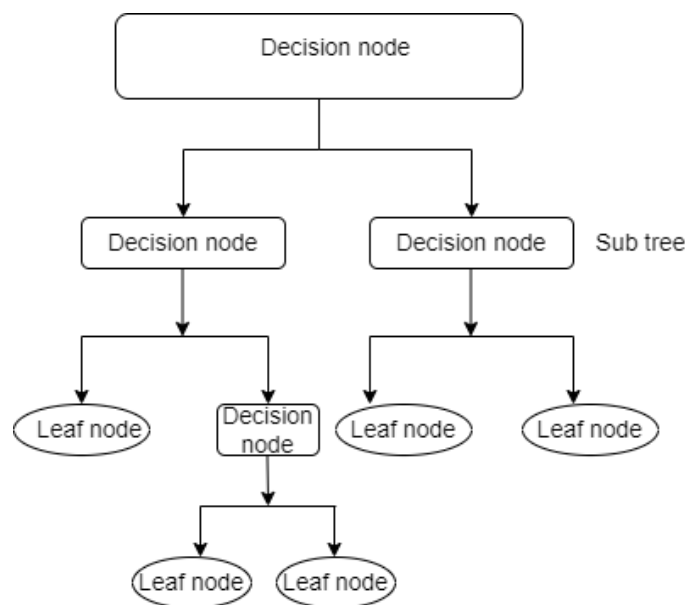


Figure 6: Decision tree classifier

3.4.3 Random Forest Classifier

The decision tree machine learning can sometimes suffer from high variance, these may impact their results negatively to the specific training data. This variance can be reduced by building multiple predictive models in parallel from multiple samples of your training data, however these trees might be highly correlated and this can make the predictions to be similar. Random Forest algorithm is a supervised machine learning algorithm that uses multiple trees identified from the

samples of your training data and forced them to be different by limiting the features that each model can evaluate for each sample. The final prediction is the class that comes many times in the output of the multipletrees used for the specific training data [25]. The figure 7 illustrates how the random forest algorithm is constructed.

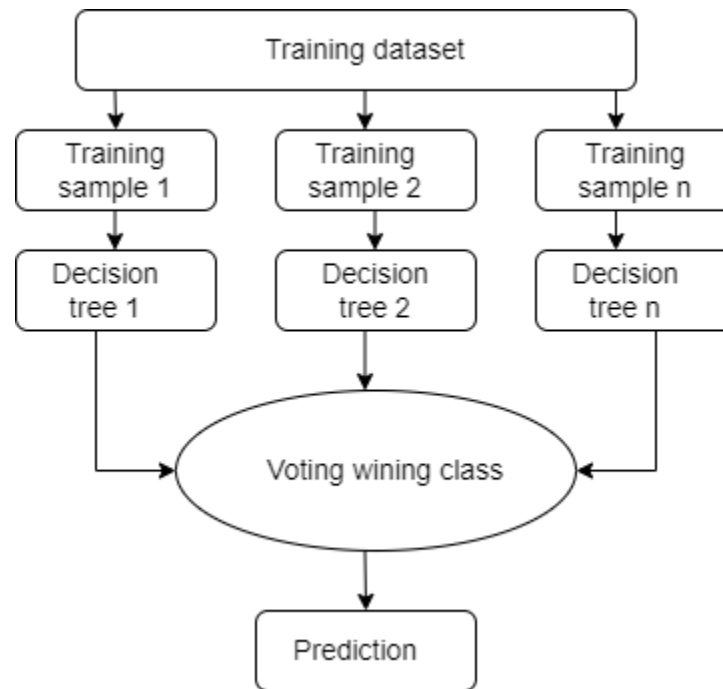


Figure 7: Building Random Forest Algorithm

3.4.4 Gradient Boosting Classifier

A Gradient boosting classifier is one type of ensemble techniques used in machine learning for increasing the prediction accuracy. It involves a collection of the weak models to build a strong predictive model. Decision tree algorithms are usually used to build a gradient boosting classifier. Gradient boosting classifier is used to make a prediction when the target variables are categorical [26, 27].

3.4.5 K-Nearest Neighbours Classifier (KNN)

The K-Nearest Neighbours is machine learning algorithm used in finding similarities between data. During the model training phase all of the data are used for learning the similarities between data. Then during of model prediction for unseen data, the model searches through theentire dataset the K-most similar training examples to new example and the data with K-most similar instance is returned as the prediction. The algorithm states that if you are similar to yourneighbours that means

that you are one of them [28]. In K-Nearest Neighbours, K means the number of neighbour points which contribute in voting.

In KNN the voting points are selected by using Euclidean distance between the new point and the existing points and then the points with least distances are selected. The general formula of Euclidean distance is given by the following mathematical expression. [29].

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

Where,

\mathbf{p}, \mathbf{q} = two points in Euclidean n-space

q_i, p_i = Euclidean vectors, starting from the origin of the space (initial point)

n = n-space

3.5 Model Training and Evaluation

During model training and evaluation, we have three training inputs such as temperature, soil moisture and pH data. The training dataset comprises 498 sample records that have the three features and three target also called class. As shown by the figure below, the data have been divided into two small subsets. One is used as testing subset while another is used as training subset. The training subset contains 90% of original dataset while the testing dataset contains only 10% of the original dataset. The training subset has been applied to the machine learning algorithms to generate the prediction model. However, to test if the predictive model does not have the over-fitting the researchers have used the testing during training and evaluation process. So, the training and evaluation are conducted by using the software packages of anaconda distributed software to process the training.

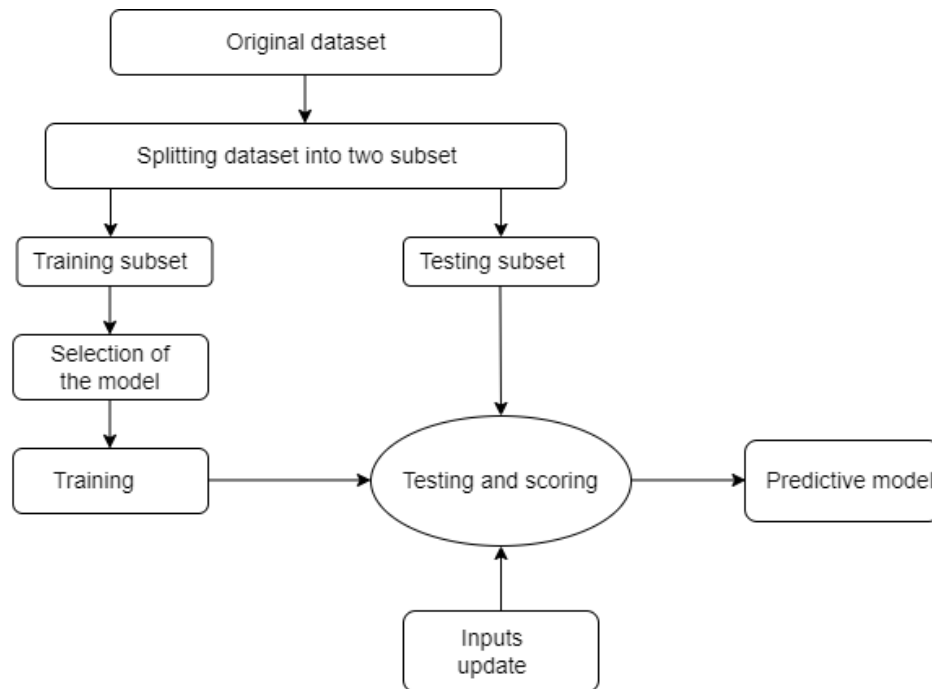


Figure 8: Model Training and Evaluation

After completing dataset preparation, different ML models have been used to choose which one can give the best training accuracy compared to others. However, the following machine learning models such as logistic regression, decision tree, KNN and Random Forest have been applied for modeling the relationship between fertilizer content and nutrients present in the soil. Due to the stability, popularity and flexibility of python codes and packages available, the python programming has been chosen for training and evaluation of dataset.

Finally, to select the best ML model among others, the researchers have considered some metrics that serve best to generate good predictive model such as prediction accuracy and precision. The prediction accuracy means the how best the model is performing well or how much the prediction is correct and it must be greater than 70%. In addition, the accuracy can serves best only if the testing and training accuracies are close one to another [30].

Chapter 4: System Analysis and Design

4.1 Introduction

During of the implementation of this research study, different hardware and software components have been used. This chapter gives a brief description on the architecture of the system implemented throughout this project. The system architecture used contains 3 subsystems. These include sensing subsystem, wireless communication subsystem and database subsystem. The following paragraphs describe separately each subsystem in details.

4.2 Sensing Subsystem

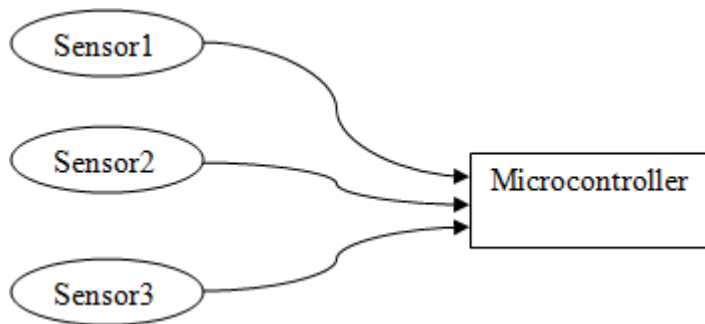


Figure 9: Sensing Subsystem

As illustrated by the figure 8 the sensing subsystem includes sensors and microcontroller platform used for collecting soil data such as temperature, soil moisture, nitrogen (N), phosphorus (P) pentoxide, potassium oxide (P) and pH level.

A sensor is an electronic device which converting any environmental physical change into a corresponded electrical signal. During of system prototyping digital temperature sensor, soil moisture sensor, pH sensor and NPK sensor have been used for measuring soil nutrients such as temperature, soil moisture, pH level as well as nitrogen, phosphorus and potassium nutrients present in the soil. The different sensors used during of prototyping are shown in the figure below.

For collecting the sensing parameters, the sensors were directly interfaced with microcontroller (Arduino Nano). Therefore NodeMCU ESP8266 Wi-Fi Module is a self-contained SOC with integrated TCP/IP protocol stack that allows any microcontroller access to your Wi-Fi network. The ESP8266 is capable of either hosting an application or offloading all Wi-Fi networking functions from another application processor [31, 32]. This board was programmed by using C programming via Arduino Integrated Development Environment (IDE) software platform.

4.3 Wireless Communication system with ESP8266 NodeMcu

The ESP8266 NodeMcu microcontroller operates in three operation modes such as access point mode, Wi-Fi station mode or both [33]. Consequently during prototyping I have used the board as Wi-Fi station mode to connect the soil data collected to SQLite server over the internet that handles all database logs. I have created an application programming interface (API) using TCP/IP protocol stack to enable communication between Node MCU and the remote server.

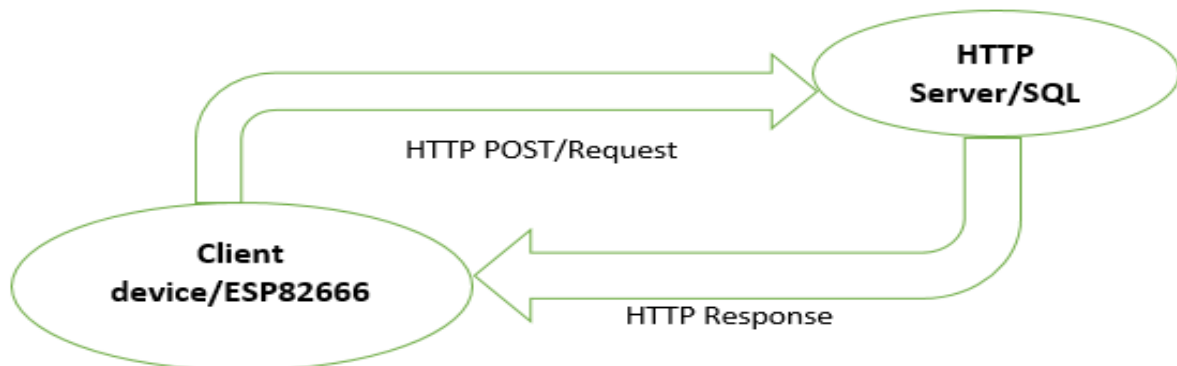


Figure 10: Http communication

Hyper Text Transfer Protocol is the best example of IoT network protocol. This protocol has formed the foundation of data communication over the web. It is the most common protocol that is used for IoT devices when there is a lot of data to be published. This scenario is illustrated in figure 10.

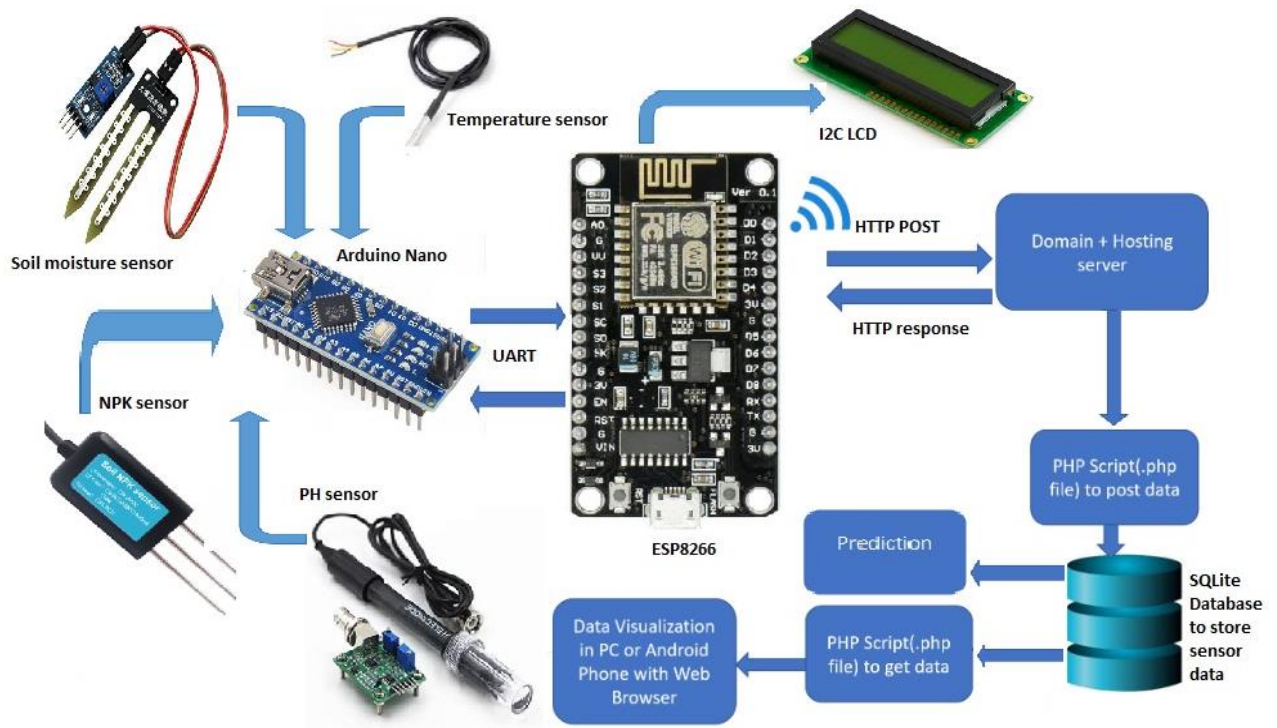


Figure 11: Wireless Communication system

The figure 11 shows how the data are being collected by sensors, processed and transferred by the microcontroller Arduino nano to the server through the NodeMCU over the internet.

4.4 List of components and materials used

- ESP8266-12E Board
- Arduino nano
- NPK Sensor
- Temperature sensor
- Soil moisture sensor
- PH sensor
- Liquid Crystal Display
- breadboards
- SQLite server
- Jumper wires

4.5 System Block Diagram

The system block diagram illustrates the important part of research solution. It is composed with three important subsystems. The first part deals with sensing unit where the sensors are collecting data from the soil and then sent to the microcontroller for processing. The second part is gateway responsible to connect the soil data collected subsystem to the server. Lastly the third is server and monitoring part, is responsible to communicate with sensor node by receiving the soil data which are coming from the sensor devices.

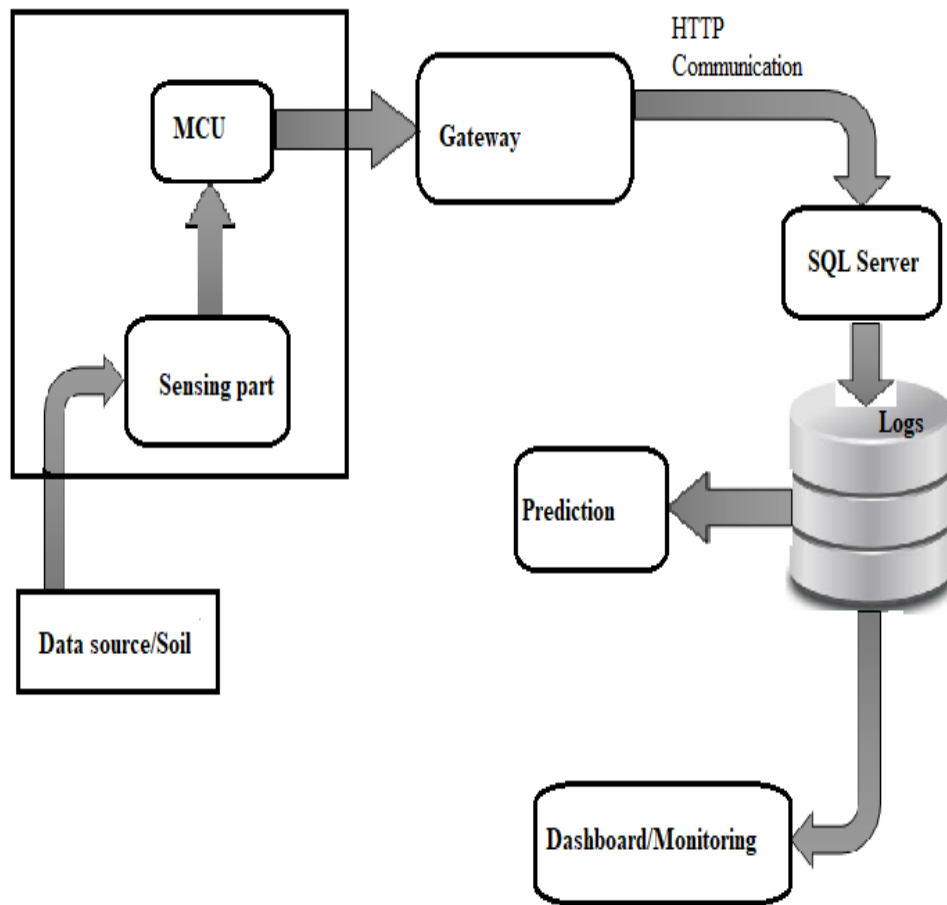


Figure 12: Detailed block diagram of soil data collection system

4.6 System flowchart diagram

A flowchart is a pictorial representation of an algorithm in which steps are drawn in the form of different shapes of boxes and the logical flow is indicated by interconnecting arrows. Flowcharts are also called block diagrams. The flow chart diagram illustrated below demonstrates how the current system works in general.

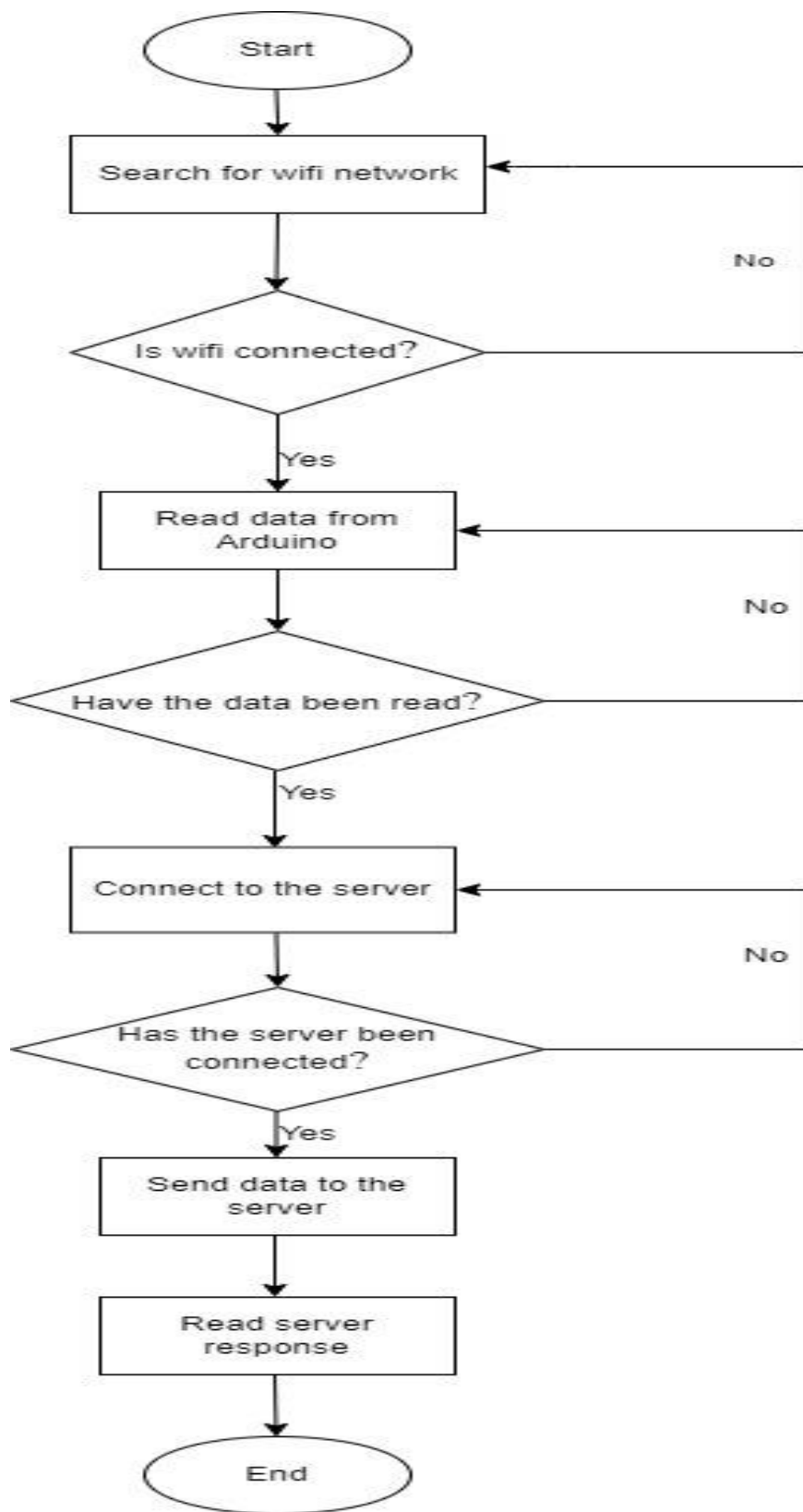


Figure 13: System flowchart for ESP8266 NodeMCU

4.7 API design and communication protocol

To design and implement system API, I have developed an API endpoint on the server side in php language that enables the ESP 8266 NodeMCU Client to connect to the server through a TCP port before publishing sensor data to the server. When data is successfully loaded into the database, the server responds by sending the ES8266 NodeMCU Client an acknowledgement. This technique is http post, which utilizes the TCP/IP Protocol.

Chapter 5: Results and Analysis

5.1 Training and evaluation implementation

The training and evaluation processes were implemented through python programming codes and python libraries. Different machine learning algorithms used were implemented using python libraries and these libraries are found in sklearn python package [33]. The training process and evaluation was started by importing the original dataset from CSV file by using data frame implemented with panda's python module.

5.2 Dataset Description

The dataset used during model training and evaluation contained 498 observations. As illustrated by the figure 14 each sample contains soil moisture, temperature and pH parameters as predictors and nitrogen, phosphorus and potassium as the response variables that are expressed in mg/kg.

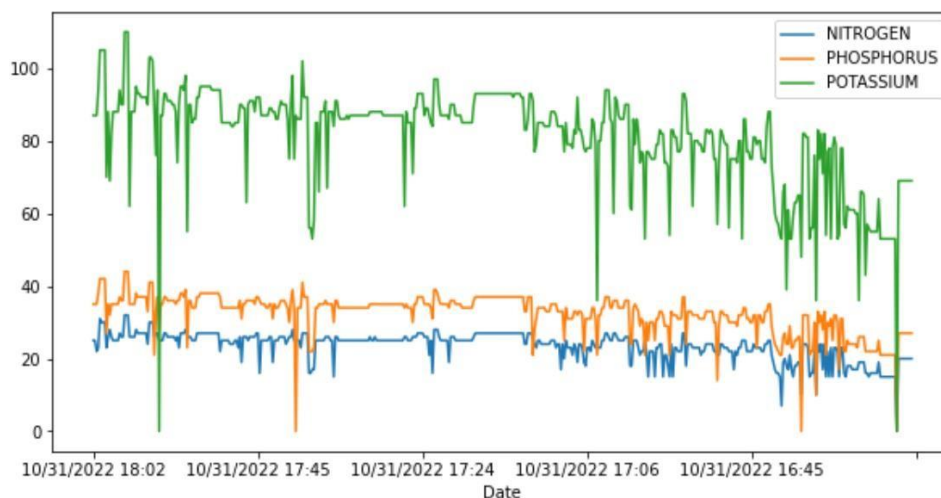


Figure 14: Dataset description

For optimizing the predictive accuracy each response variable was studied separately and three different predictive models for predicting nitrogen, phosphorus and potassium respectively have been generated. The details are provided in the following sections. Here the response variables values for nitrogen, phosphorus and potassium are changing due to the fact that the samples taken are also different depending on the type of soil. As figure 14 illustrates, potassium contains the highest content and nitrogen with the lowest content in mg/kg

The dataset used for building nitrogen predictive model include soil moisture, temperature and pH parameters as predictors and nitrogen as the target variable. As illustrated by the figure 15 the nitrogen samples range between zero and 35 levels.

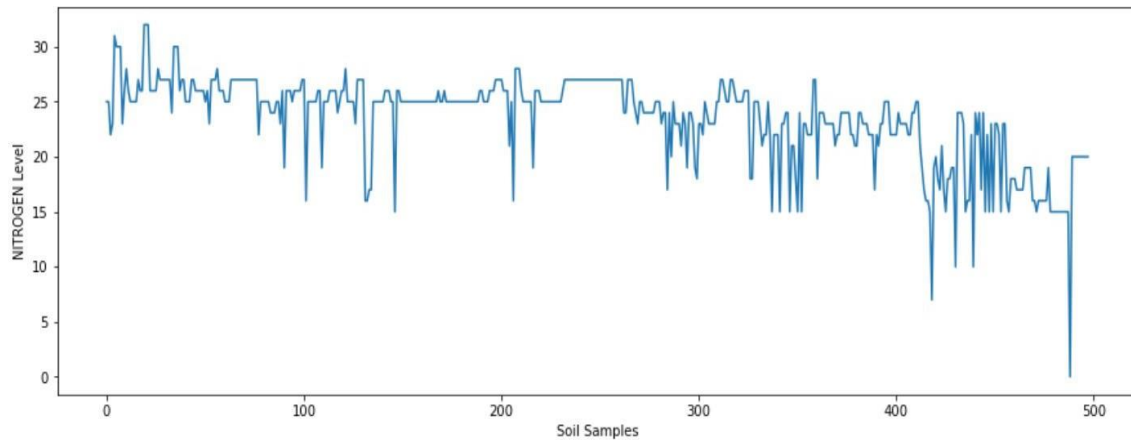


Figure 15: Nitrogen sample

The dataset used for building phosphorus predictive model include soil moisture, temperature and pH parameters as predictors and phosphorus as the target variable. As illustrated by the figure 16 the phosphorus samples range between zero and 50 levels.

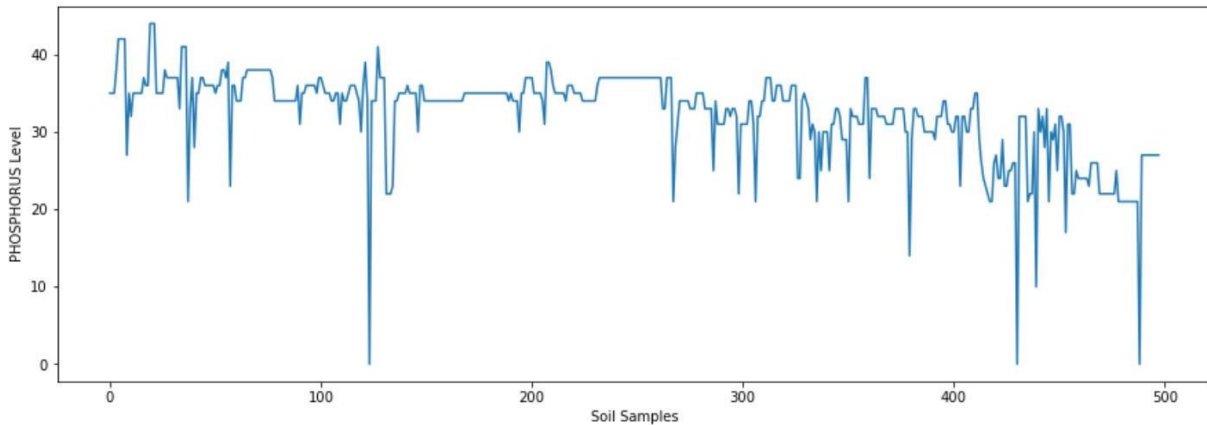


Figure 16: Phosphorus samples

The dataset used for building the potassium predictive model contained soil moisture, temperature and pH parameters as predictors and potassium. The values of potassium ranges between zero and 120. Details are available in the following figure 17.

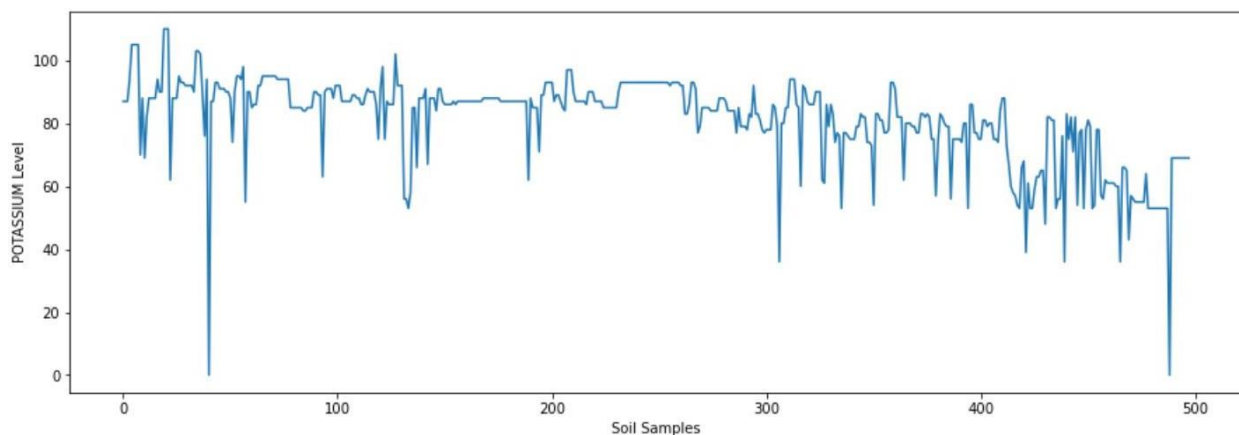


Figure 17: Potassium samples

5.3 Correlation computation

Before training the machine learning algorithms, the correlation between predictors and response variables have been computed for tracking the contribution of each predictor in predicting response variables.

The figure 18, 19 & 20 illustrate the correlation between nitrogen, phosphorus, potassium and their predictors respectively and pH shows to highly correlated comparing with others predictors.

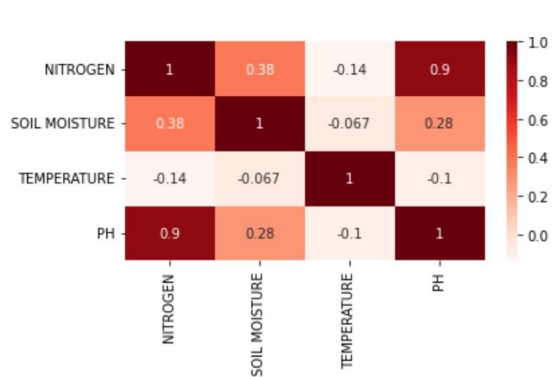


Figure 18: Correlation for nitrogen

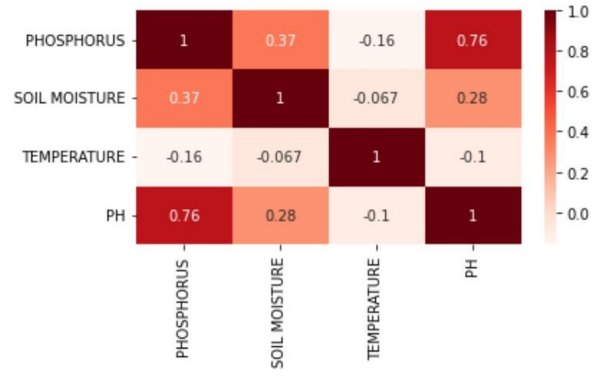


Figure 19: Correlation for phosphorus

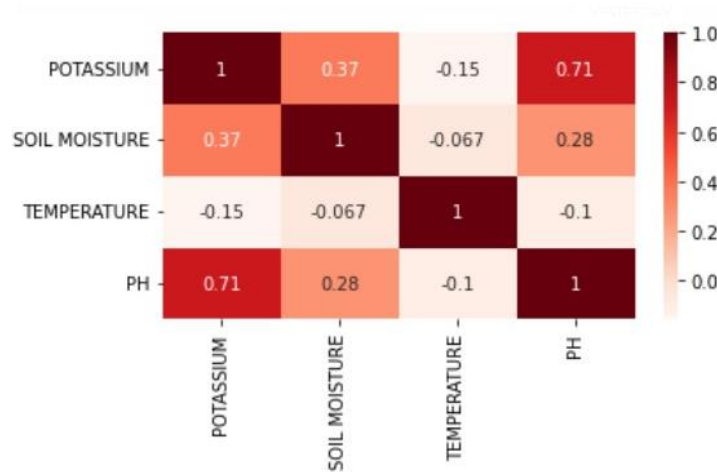


Figure 20: Correlation for potassium

The correlation between each predictor and the outcome has been computed by using packages provided by python programming and is ranging from 0 up to 1. As shown in the figures above, the pH predictor has a high correlation compared to other predictors in predicting response variables because with 0.9 for nitrogen, 0.76 for phosphorus and 0.71 for potassium respectively. This means that the pH variable has a high percentage in determining the value of the outcome compared with other features considered in this research study. The correlation helps us also in the identification of which features that are highly correlated for avoiding the problem of multicollinearity in model development.

5.4 Features Extraction

The figure 21, 22 & 23 show the features importance of predicting nitrogen, phosphorus and potassium respectively. The pH and temperature features demonstrated to be the features of high and low temperature respectively in all cases.

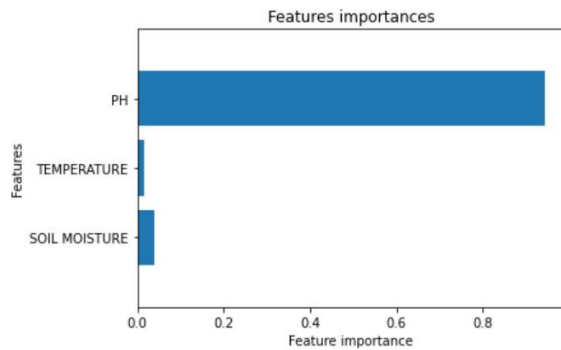


Figure 21: Nitrogen features importance

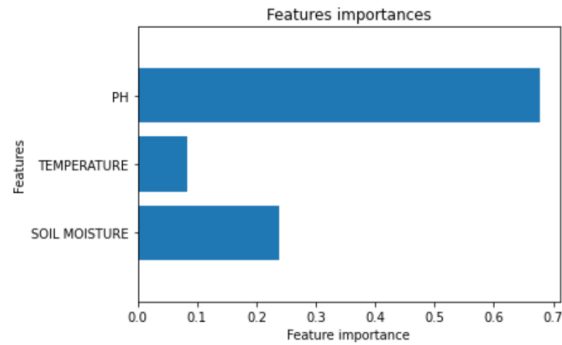


Figure 22: Phosphorus Features importance

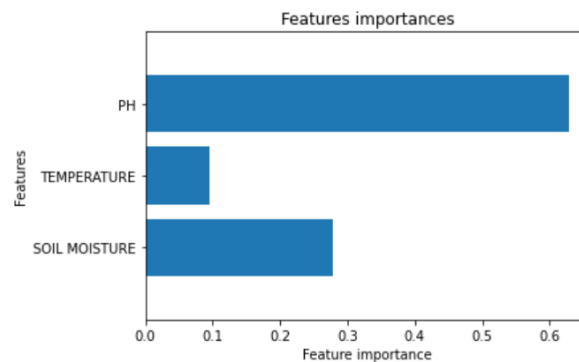


Figure 23: Potassium features importance

The term feature importance here refers to the weight of each among the predictors used in predicting the response variable and as the figures above show, pH has a highest contribution in predicting each of the response variables while temperature has the lowest contribution in predicting one of the response variables. The feature extraction helps us to identify which features that can be considered during the model training and evaluation for optimizing the prediction accuracy.

5.5 Training and testing dataset

During predictive modelling the original dataset was split into training and testing dataset. 70 % and 30% of the original dataset were used for training and testing the nitrogen predictive model respectively. However, 90 % and 10% of the dataset were used for training and testing respectively of phosphorus and potassium predictive models.

5.6 Results of model training

During the model training, different supervised machine learning algorithms such as linear regression, ridge, lasso, random forest, decision tree, support vector machine and gradient boosting were investigated.

- The performance of Nitrogen predictive model on training dataset and testing dataset is illustrated by table 1 and 2 respectively. Decision tree, random forest, gradient boosting and KNN were performed better respectively.

| | Model | MAE | RMSE | R_squared |
|---|---------------------------|----------|----------|-----------|
| 0 | linear | 0.862980 | 1.754189 | 0.806293 |
| 1 | Lasso | 0.861929 | 1.754196 | 0.806291 |
| 2 | Ridge | 0.862980 | 1.754189 | 0.806293 |
| 3 | Rforest | 0.159895 | 0.566989 | 0.979763 |
| 4 | Gradient Boost | 0.209525 | 0.496554 | 0.984479 |
| 5 | K-Nearest Neighbors | 0.431034 | 1.255639 | 0.900752 |
| 6 | Decision Tree | 0.206897 | 0.736078 | 0.965893 |
| 7 | Support Vector Regression | 0.840464 | 2.116738 | 0.717950 |

Table 1: On training dataset

| | Model | MAE | RMSE | R_squared |
|---|---------------------------|----------|----------|-----------|
| 0 | linear | 0.681982 | 1.107899 | 0.909598 |
| 1 | Lasso | 0.680908 | 1.107273 | 0.909700 |
| 2 | Ridge | 0.681982 | 1.107899 | 0.909598 |
| 3 | Rforest | 0.280006 | 0.926380 | 0.936794 |
| 4 | Gradient Boost | 0.371790 | 0.934264 | 0.935714 |
| 5 | K-Nearest Neighbors | 0.488889 | 1.243651 | 0.886087 |
| 6 | Decision Tree | 0.248000 | 0.864780 | 0.944921 |
| 7 | Support Vector Regression | 0.788722 | 1.685327 | 0.790808 |

Table 2: On testing dataset

- The performance of Phosphorus predictive model was tested on training and testing dataset as illustrated by the following table 3 & 4 respectively. Here the random forest was the best performer 93% and 90% for training and testing accuracies.

| | Model | MAE | RMSE | R_squared |
|---|---------------------------|----------|----------|-----------|
| 0 | linear | 1.809203 | 3.585780 | 0.604734 |
| 1 | Lasso | 1.808579 | 3.585784 | 0.604733 |
| 2 | Ridge | 1.809203 | 3.585780 | 0.604734 |
| 3 | Rforest | 0.694620 | 1.491010 | 0.931659 |
| 4 | Gradient Boost | 1.125126 | 2.094833 | 0.865097 |
| 5 | K-Nearest Neighbors | 1.290923 | 2.846128 | 0.750982 |
| 6 | Decision Tree | 1.283743 | 2.813065 | 0.756734 |
| 7 | Support Vector Regression | 1.804528 | 3.967821 | 0.516021 |

Table 3: On training dataset

| | Model | MAE | RMSE | R_squared |
|---|---------------------------|----------|----------|-----------|
| 0 | linear | 1.410670 | 2.055141 | 0.800506 |
| 1 | Lasso | 1.410488 | 2.054779 | 0.800576 |
| 2 | Ridge | 1.410670 | 2.055141 | 0.800506 |
| 3 | Rforest | 0.853980 | 1.453185 | 0.900256 |
| 4 | Gradient Boost | 1.180616 | 1.921666 | 0.825578 |
| 5 | K-Nearest Neighbors | 1.433333 | 3.082207 | 0.551286 |
| 6 | Decision Tree | 1.069568 | 1.558403 | 0.885289 |
| 7 | Support Vector Regression | 1.363215 | 2.338767 | 0.741643 |

Table 4: On testing dataset

- During Potassium predictive model testing, random forest scored a high performance with 89% and 76% accuracies for training and testing dataset respectively as illustrate by the following tables.

| | Model | MAE | RMSE | R_squared |
|---|---------------------------|----------|-----------|-----------|
| 0 | linear | 5.327358 | 9.632187 | 0.526856 |
| 1 | Lasso | 5.327752 | 9.632188 | 0.526856 |
| 2 | Ridge | 5.327359 | 9.632187 | 0.526856 |
| 3 | Rforest | 2.125579 | 4.617073 | 0.891288 |
| 4 | Gradient Boost | 3.294305 | 6.377220 | 0.792601 |
| 5 | K-Nearest Neighbors | 3.838542 | 7.350106 | 0.724494 |
| 6 | Decision Tree | 3.590321 | 6.956107 | 0.753239 |
| 7 | Support Vector Regression | 5.348132 | 10.763424 | 0.409194 |

Table 5: On training dataset

| | Model | MAE | RMSE | R_squared |
|---|---------------------------|----------|----------|-----------|
| 0 | linear | 4.122869 | 5.831790 | 0.749634 |
| 1 | Lasso | 4.123107 | 5.832037 | 0.749612 |
| 2 | Ridge | 4.122869 | 5.831791 | 0.749634 |
| 3 | Rforest | 3.421243 | 5.694447 | 0.761287 |
| 4 | Gradient Boost | 3.933241 | 6.050152 | 0.730533 |
| 5 | K-Nearest Neighbors | 3.493333 | 5.744949 | 0.757034 |
| 6 | Decision Tree | 4.260681 | 7.516429 | 0.584094 |
| 7 | Support Vector Regression | 4.069993 | 6.968506 | 0.642520 |

Table 6: On testing data

According to the results shown in the above tables, it is seen that using Random forest, Decision tree and Gradient boosting machine learning algorithms could perform well respectively in both training and testing prediction accuracies comparing with the other algorithms for the current research with the specific dataset considered. Here models are performing in different manners, which means that they do not analyze the data in the same way depending on the dataset and their internal architecture. However before choosing a particular model that you will use in predictions, it is necessary to use all of them to know which one can perform well.

5.7 Real time soil data collection Hardware implementation

The hardware implementation is composed by four main sensors namely NPK sensor, temperature sensor, soil moisture sensor and pH sensor that are responsible for collecting soil data and connect directly to the microcontroller (arduino nano) for processing. Finally after computation, the data processed sent to the server using http protocol.

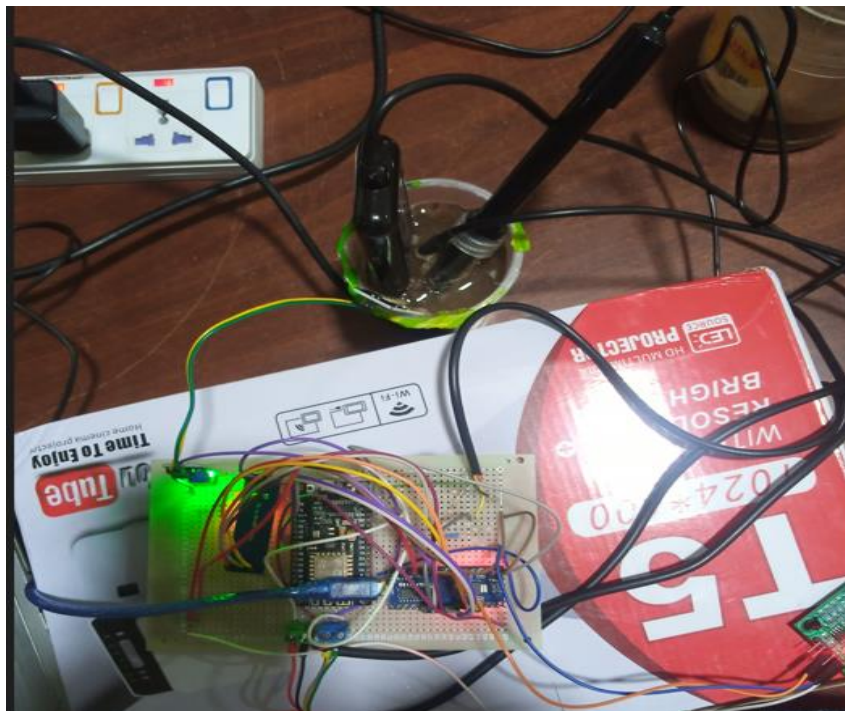


Figure 24: Hardware implementation

Chapter 6: Conclusion and Recommendation

This research aimed to map the dependency between fertilizer content for cassava crop and soil nutrients variables such as temperature, soil moisture, pH level and nitrogen, phosphorus as well as potassium here in Rwanda by using machine learning and internet of things. The outcome of the current research shows that the fertilizer of cassava crop could be predicted well by using decision tree and random forest machine learning algorithms respectively comparing with the other algorithms tested through this research. During the model performance evaluation, 93.1% and 90% of training and testing prediction accuracies respectively were achieved by using random forest for the soil samples taken in Ruhango district. However, 96.5% and 94.4% of training and testing prediction accuracies were generated by using the decision tree predictive model.

The models used in this current study, were trained and evaluated by using 498 soil data samples from Rwanda. However, the government agriculture policies like use of fertilizers and use of regular irrigation can impact the cassava crop productivity. For the future work, we expect to employ some advanced machine learning models like time series models (Recurrent Neural Networks) to predict the future behavior of cassava crop growth. Secondary, the number of observations will be increased and the government input policies to increase the productivity will be considered during the data analysis.

I recommend the ministry of agriculture (MINAGRI) in partnership of local government and national institute of statistics to extend this research in the rest districts for having the real image of the country. Secondly, I recommend the decision makers to help in developing an IoT based system for soil data information to the cassava farmers for taking measures accordingly in real time. Finally, I recommend the next researchers to investigate other research that includes more data, advanced machine learning algorithms such as recurrent neural networks to address the issue of soil data challenge and prediction for long term period as well as different government input policies to avoid agricultural losses due to the lack of information about the nutrients present in soil.

LIST OF REFERENCES

- [1] D. Vadalia, M. Vaity, K. Tawate, D. Kapse, S. V. Sem, and C. Engg, “Real Time soil fertility analyzer and crop prediction,” *Int. Res. J. Eng. Technol.*, vol. 4, no. 3, pp. 3–5, 2017.
- [2] S. N. Shylaja, “Real- Time Monitoring of Soil Nutrient Analysis u sing WSN,” 2017 *Int. Conf. Energy, Commun. Data Anal. Soft Comput*, pp. 3059–3062, 2017.
- [3] A. Amrutha, R. Lekha, and A. Sreedevi, “Automatic soil nutrient detection and fertilizer dispensary system,” *Proc. 2016 Int. Conf. Robot. Curr. Trends Futur. Challenges*, 2017, doi: 10.1109/RCTFC.2016.7893418.
- [4] J. C. Puno, E. Sybingco, E. Dadios, I. Valenzuela, and J. Cuello, “Determination of soil nutrients and pH level using image processing and artificial neural network,” *HNICEM 2017 - 9th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag.* vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/HNICEM.2017.8269472.
- [5] L. C. Gavade and A. Bhoi, “N , P , K Detection & Control for Agriculture Applications using PIC Controller : A Review,” *Int. J. Eng. Res. Technol.*, vol. 6, no. 04, pp. 638–641, 2017..
- [6] H. Zheng, J. Wu, and S. Zhang, “Study on the spatial variability of farmland soil nutrient based on the kriging interpolation,” 2009 *Int. Conf. Artif. Intell. Comput. Intell. AICI 2009*, vol. 4, pp. 550–555, 2009, doi: 10.1109/AICI.2009.137.
- [7] Prince Patel, “Why Python is the most popular language used for Machine Learning,” 2018. [Online]. Available: <https://medium.com/@UdacityINDIA/why-use-python-for-machine-learning-e4b0b4457a77>. [Accessed: 20-Mar-2022].
- [8] N. Gupta, “Why is Python Used for Machine Learning?” 2019. [Online]. Available: <https://hackernoon.com/why-python-used-for-machinelearning-u13f922ug>. [Accessed: 20-Mar-2022].
- [9] A. Beklemysheva, “Why Use Python for AI and Machine Learning.” [Online]. Available: <https://steelkiwi.com/blog/python-for-ai-andmachine-learning/>. [Accessed: 20-Mar-2022].

- [10] D. V Ramane, S. S. Patil, and A. D. Shaligram, "Detection of NPK nutrients of soil using Fiber Optic Sensor," *Int. J. Res. Advent Technol. ACGT*, no. February, pp. 13–14, 2015.
- [11] P. Ongsulee, "Artificial intelligence, machine learning and deep learning," *Int. Conf. ICT Knowl. Eng.*, pp. 1–6, 2018, doi: 10.1109/ICTKE.2017.8259629.
- [12] A. Angra, Sheena; Sachin, "., /,7(5\$785(6859(i;" pp. 57–60, 2017.
- [13] R. A. Kumar, M. K. M. Aslam, V. P. J. Raj, T. Radhakrishnan, K. S. Kumar, and T. K. Manojkumar, "A statistical analysis of soil fertility of Thrissur district, Kerala," *Proc. 2016 Int. Conf. Data Sci. Eng. ICDSE 2016*, pp. 7–11, 2017, doi: 10.1109/ICDSE.2016.7823953.
- [14] V. A. Gulhane and S. V. Rode, "Correlation analysis on soil nutrients and wavelet decompositions of satellite imagery," *2015 Int. Conf. Ind. Instrum. Control. ICIC 2015*, no. Icic, pp. 1225–1230, 2015, doi: 10.1109/IIC.2015.7150934.
- [15] D. Nelson, "Gradient Boosting Classifiers in Python with Scikit-Learn," 2020. [Online]. Available: <https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/>. [Accessed: 04-Nov-2022].
- [16] V. KURAMA, "Gradient Boosting In Classification: Not a Black Box Anymore!," 2022. [Online]. Available: <https://blog.paperspace.com/gradient-boosting-for-classification/>. [Accessed: 04-Nov-2022].
- [17] J. F. Figueroa, H. R. B. Everett, S. S. Ipson, and C. Liu, "Sensors and Actuators 16."
- [18] P. Description, "DS18B20 Waterproof Temperature Sensor Cable," pp. 0–2.
- [19] O. C. Ph-bta, "pH Sensor," pp. 3–6.
- [20] I. Zolotová, M. Bundzel, and T. Lojka, "Industry IoT Gateway for Cloud Connectivity Industry IoT Gateway for Cloud Connectivity," no. September, 2015.

[21] Javier Bonilla, “ESP8266 NodeMCU pinout for Arduino IDE,” 2019. [Online]. Available: <https://mechatronicsblog.com/esp8266-nodemcu-pinout-for-arduino-ide/>[Accessed: 11-Oct-2022].

[22] S. Jagirdar, “Cloud computing basics,” no. August 2013, 2014.

[23] M. Stojiljković, “Logistic Regression in Python,” 2022. [Online]. Available: <https://realpython.com/logistic-regression-python/>. [Accessed: 04-Nov-2022].

[24] Scikit-learn, “Decision Tree regression,” 2022. [Online]. Available: <https://scikit-learn.org/stable/modules/tree.html#mathematical-formulation>. [Accessed: 08-Oct-2022].

[25] DataCamp.com, “Random Forest classifier,” 2022. [Online]. Available: <https://www.datacamp.com/community/tutorials/random-forests-classifier-python>. [Accessed: 10-Nov-2022].

[26] D. Nelson, “Gradient Boosting Classifiers in Python with Scikit-Learn,” 2022. [Online]. Available: <https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/>. [Accessed: 04-Nov-2022].

[27] V. KURAMA, “Gradient Boosting In Classification: Not a Black Box Anymore!,” 2022. [Online]. Available: <https://blog.paperspace.com/gradient-boosting-for-classification/>. [Accessed: 04-Nov-2022].

[28] J. Brownlee, “Develop k-Nearest Neighbors in Python From Scratch,” 2022. [Online]. Available: <https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/>. [Accessed: 04-Nov-2022].

[29] A. Robinson, “How to Calculate Euclidean Distance,” 2022. [Online]. Available: <https://sciencing.com/how-to-calculate-euclidean-distance-12751761.html>. [Accessed: 04-Nov-2022].

[30] D. S. E. Nagesh Singh Chauhan, “Model evaluation metrics in ML,” 2022. [Online]. Available: <https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html>. [Accessed: 18-Nov-2022].

[31] H. Chen, X. Jia, and H. Li, “A brief introduction to iot gateway,” IET Conf. Publ., vol. 2011, no. 586 CP, pp. 610–613, 2012, doi: 10.1049/cp.2011.0740.

[32] S. Guoqiang, C. Yanming, Z. Chao, and Z. Yanxu, “Design and implementation of a smart IoT gateway,” Proc. - 2013 IEEE Int. Conf. Green Comput. Commun. IEEE Internet Things IEEE Cyber, Phys. Soc. Comput. GreenCom-iThings-CPSCoM 2013, pp. 720–723, 2013, doi: 10.1109/GreenCom-iThings-CPSCoM.2013.130

[33] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, “Scikit-learn,” GetMobile Mob. Comput. Commun., vol. 19, no. 1, pp. 29–33, 2015, doi: 10.1145/2786984.2786995.

APPENDICES

Appendix 1: Soil Dataset

| 1 | Date | NITROGEN (mg/kg) | PHOSPHORUS(mg/kg) | POTASSIUM(mg/kg) | SOIL MOISTURE(%) | TEMPERATURE(°C) | PH |
|----|------------------|------------------|-------------------|------------------|------------------|-----------------|-------|
| 2 | 31/10/2022 18:02 | 25 | 35 | 87 | 71.16 | 24.19 | 10.78 |
| 3 | 31/10/2022 18:02 | 25 | 35 | 87 | 79.57 | 24.19 | 10.78 |
| 4 | 31/10/2022 18:01 | 22 | 35 | 87 | 86.51 | 24.19 | 7.78 |
| 5 | 31/10/2022 18:01 | 23 | 38 | 94 | 69.4 | 24.13 | 8.78 |
| 6 | 31/10/2022 18:01 | 31 | 42 | 105 | 69.7 | 23.75 | 11.78 |
| 7 | 31/10/2022 18:01 | 30 | 42 | 105 | 70.48 | 24.19 | 15.78 |
| 8 | 31/10/2022 18:01 | 30 | 42 | 105 | 81.92 | 24.19 | 15.78 |
| 9 | 31/10/2022 18:01 | 30 | 42 | 105 | 85.53 | 24.19 | 15.78 |
| 10 | 31/10/2022 18:00 | 23 | 27 | 70 | 68.72 | 24.06 | 8.78 |
| 11 | 31/10/2022 18:00 | 26 | 35 | 88 | 68.72 | 24.06 | 11.78 |
| 12 | 31/10/2022 18:00 | 28 | 32 | 69 | 68.91 | 24.06 | 13.78 |
| 13 | 31/10/2022 18:00 | 26 | 35 | 82 | 68.62 | 24.13 | 11.78 |
| 14 | 31/10/2022 18:00 | 25 | 35 | 88 | 68.91 | 24.19 | 10.78 |
| 15 | 31/10/2022 18:00 | 25 | 35 | 88 | 68.82 | 24.19 | 10.78 |
| 16 | 31/10/2022 17:59 | 25 | 35 | 88 | 68.82 | 23.63 | 10.78 |
| 17 | 31/10/2022 17:59 | 25 | 35 | 88 | 69.01 | 23.69 | 10.78 |
| 18 | 31/10/2022 17:59 | 27 | 37 | 94 | 69.01 | 24.13 | 12.78 |
| 19 | 31/10/2022 17:59 | 26 | 36 | 90 | 69.31 | 24.19 | 11.78 |
| 20 | 31/10/2022 17:59 | 26 | 36 | 90 | 69.31 | 24.13 | 11.78 |
| 21 | 31/10/2022 17:59 | 32 | 44 | 110 | 69.7 | 24.19 | 9.78 |
| 22 | 31/10/2022 17:58 | 32 | 44 | 110 | 70.09 | 24.19 | 9.78 |
| 23 | 31/10/2022 17:58 | 32 | 44 | 110 | 72.14 | 24.06 | 9.78 |
| 24 | 31/10/2022 17:58 | 26 | 35 | 62 | 81.43 | 23.5 | 11.78 |

| | | | | | | | |
|-----|------------------|----|----|----|-------|-------|------|
| 476 | 31/10/2022 16:32 | 16 | 22 | 55 | 54.84 | 24.38 | 5.01 |
| 477 | 31/10/2022 16:32 | 16 | 22 | 55 | 54.74 | 24.31 | 5.01 |
| 478 | 31/10/2022 16:31 | 16 | 22 | 55 | 57.58 | 23.63 | 5.01 |
| 479 | 31/10/2022 14:58 | 19 | 25 | 64 | 28.15 | 24.06 | 8.23 |
| 480 | 31/10/2022 14:58 | 15 | 21 | 53 | 49.56 | 24.81 | 8.21 |
| 481 | 31/10/2022 14:57 | 15 | 21 | 53 | 51.81 | 24.13 | 8.21 |
| 482 | 31/10/2022 14:57 | 15 | 21 | 53 | 51.61 | 24.06 | 8.2 |
| 483 | 31/10/2022 14:57 | 15 | 21 | 53 | 52.79 | 24.69 | 8.21 |
| 484 | 31/10/2022 14:57 | 15 | 21 | 53 | 52.79 | 24.69 | 8.23 |
| 485 | 31/10/2022 14:57 | 15 | 21 | 53 | 52.3 | 24.63 | 8.21 |
| 486 | 31/10/2022 14:57 | 15 | 21 | 53 | 47.21 | 24.88 | 8.21 |
| 487 | 31/10/2022 14:56 | 15 | 21 | 53 | 52.3 | 24.56 | 8.2 |
| 488 | 31/10/2022 14:56 | 15 | 21 | 53 | 46.33 | 24.06 | 8.22 |
| 489 | 31/10/2022 14:56 | 15 | 21 | 53 | 28.15 | 24.56 | 8.2 |
| 490 | 31/10/2022 14:56 | 0 | 0 | 0 | 28.05 | 24.56 | 8.18 |
| 491 | 31/10/2022 14:56 | 20 | 27 | 69 | 28.15 | 24.44 | 8.19 |
| 492 | 31/10/2022 14:55 | 20 | 27 | 69 | 28.15 | 24.56 | 8.2 |
| 493 | 31/10/2022 14:55 | 20 | 27 | 69 | 28.15 | 23.94 | 8.2 |
| 494 | 31/10/2022 14:55 | 20 | 27 | 69 | 28.25 | 24.56 | 8.19 |
| 495 | 31/10/2022 14:55 | 20 | 27 | 69 | 28.05 | 23.88 | 8.22 |
| 496 | 31/10/2022 14:55 | 20 | 27 | 69 | 28.05 | 24.5 | 8.19 |
| 497 | 31/10/2022 14:55 | 20 | 27 | 69 | 28.15 | 24.5 | 8.2 |
| 498 | 31/10/2022 14:54 | 20 | 27 | 69 | 27.47 | 24.5 | 5.58 |
| 499 | 31/10/2022 14:54 | 20 | 27 | 69 | 14.37 | 24.38 | 5.63 |

Appendix 2: Python code for decision tree ML algorithm

```
#import decision tree algorithm from the sklearn library  
from sklearn.tree import DecisionTreeClassifier  
#instantiate the model  
dec = DecisionTreeClassifier()  
#Train the model with input features (X_train) and targets (Y_train)  
dec.fit(X_train,Y_train)  
#Making Prediction on testing data (X_test)  
Y_pred=dec.predict(X_test)
```

Appendix 3: Python code of random forest algorithm

```
#import decision tree algorithm from the sklearn library  
from sklearn.ensemble import RandomForestClassifier  
#instantiate the model  
RandF = RandomForestClassifier()  
#Train the model with input features (X_train) and targets (Y_train)  
RandF.fit(X_train,Y_train)  
#Making Prediction on testing data (X_test)  
Y_pred=RandF.predict(X_test)
```

Appendix 4: Python codes of KNN

```
#import decision tree algorithm from the sklearn library  
from sklearn.neighbors import KNeighborsClassifier  
#instantiate the model  
Kn = KNeighborsClassifier()  
#Train the model with input features (X_train) and targets (Y_train)  
Kn.fit(X_train,Y_train)  
#Making Prediction on testing data (X_test)  
Y_pred=Kn.predict(X_test)
```