AFRICAN CENTRE OF EXCELLENCE

IN DATA SCIENCE

Applying Data mining to predict annual yield of major crops in districts of Rwanda during agriculture seasons 2017-2019

By

Jean Damascene HAGENIMANA

Registration Number: 219013782

A dissertation submitted in partial fulfillment of the requirements for the Degree of Master of Science in Data Science in Data Mining

University of Rwanda, College of Business and Economics

African Centre of Excellence in Data Science (ACE-DS)

Supervisor: Dr. Innocent NGARUYE

September, 2020

Declaration

I declare that this dissertation entitled "Applying Data mining to predict annual yield of major crops in districts of Rwanda during agriculture seasons 2017-2019" is my original work and that to the best of my knowledge, it has not been presented for the award of a degree in any other University and all sources of materials used for this thesis have been properly acknowledged.

Student Name: Jean Damascene HAGENIMANA

Signature: 

Approval sheet

This dissertation entitled “Applying Data mining to predict annual yield of major crops in districts of Rwanda during agriculture seasons 2017-2019“ written and submitted by Jean Damascene HAGENIMANA in partial fulfillment of the requirements for the degree of Master of Science in Data Science majoring in Data mining is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 15% which is less than 20% accepted by ACE-DS.

Dr. Innocent NGARUYE

Supervisor

12/8/2021

Dr. Ignace KABANO

Head of Training

Dedication

I dedicated this thesis to my wife, parents, siblings, classmates and friends for their support and encouragement.

Acknowledgement

I dedicate this piece of work to my wife, my parents, Mr. Jean Claude M, SAS Team, Mr.Donath. N, Mr. Moses RUHINDA, J. Jacque. K, to my brothers and Sister, classmates and friends for their encouragement, understanding, patience and all various support while pursuing the courses. It is with heartfelt gratitude that I first thank God the Almighty for his continued grace and who I believe has always been with me and has led me throughout this work. I express my warm gratitude to the University of Rwanda, African Centre of Excellence in Data Science (ACEDS) and all Lectures staffs of Data Science program that assisted me to reach this level. I would like to express my sincere gratitude to all those who helped me in one-way or the other in making this work successful. I am most grateful to my supervisors Dr. Innocent NGARUYE, for devoting his precious time to consistently guide and encourage me during the course of writing my research thesis. My special Thanks to my workmates, especially Seasonal Agriculture Survey staff, you did much more to me, I always recognize your encouragement, your motivation and your help in various ways, I remember and I will always do.

Many thanks to all my classmates and friends for the cordial relationship shared during the good and rigorous moments of the course. They will always be fond to me. Of course, it is a ritual for one to thank his family but mine deserve a special one especially my parents, my employee thanks for your usual love and encouragement. Finally, I beg the understanding of all those I have not mentioned in this short acknowledgement. To everyone who assisted me in this study in whatever way, am indeed indescribably grateful. May the Almighty God bless you abundantly.

Abstract

Crop yield estimate is crucial indicator computed by National Institute of Statistics of Rwanda (NISR) in cooperation with the Ministry of Agriculture and Animal Resources (MINAGRI) through seasonal Agricultural survey (SAS) for monitoring the agriculture programs and policies as well as addressing key agriculture issues and providing to policy makers and other stakeholders. After a comprehensive review on the prediction of crop yield, some re- cent work was focusing on estimating crop yield by considering the total reported quantity of harvested over the harvested area of that crop without considering the other factor which may affect the crop yield production in Rwanda. The cultivation period in Rwanda is divided into three cultivation seasons, Season A which is conducted from September to February of the following year, Season B which starts from March to June and Season C which take place from June to August. This study will analyze the crop yields of three consecutive years using data mining techniques. Specifically, the study will identify the effectiveness of Artificial Neural Network (ANN) model on crop yield prediction for typical environment factors, to compare the effectiveness of Polynomial Linear Regression Model, Multiple Linear Regression model with Artificial Neural Network model on yield prediction to assess the impact of various inputs on crop yield production of main crops in Rwanda. The focus of this study is the development of data mining techniques in agricultural field. Various descriptive methods will be used to summarize and to preprocess the attributes and to test statistical significant of the results obtained. Different regression models namely the, Linear Regression, Multiple Linear Regression, Polynomial Regression Model and Artificial Neural Network are proposed to accurately predict the yield of Maize, beans, Irish Potatoes, and paddy rice. Finally, regression models are proposed to accurately predict the yield of those crops. The Artificial Neural Network (ANN) model predict better results for maize and paddy rice rather than Polynomial Linear Regression (PLR) and Multiple Linear Regression (MLR) models with MAE (0.05), MSE (0.29), RMSE (0.54) and R^2 (0.99) for maize; and MAE (2.80), MSE (15.810), RMSE (12.57) and R^2 (0.99) for paddy rice while PLR predicts better results for Irish potatoes and bush beans with MAE (0.40), MSE (3.30), RMSE (1.82) and R^2 (0.95) for Irish potatoes; and MAE (0.11), MSE (3.47), RMSE (1.18) and R^2 (0.95) for bush beans respectively. The results approved that the best regression model selected based on R^2, MAE, MSE, and RMSE for predicting better solution for the farmers about how they may improve their yields, The findings proved that the ANN perform well in predicting the yields of major crops produced in Rwanda.

Keywords: *Artificial Neural Network, crop yield, data mining, food security, machine learning*

Table of Contents

List of Tables

List of Figures

Lists of Abbreviations

ANN : Artificial Neuron Network

BARI : Bangladesh Agriculture Research Institute

CAPI : Computer Assistant Personal Interviewing

CIP : Crop Intensification Program

CSPRO : Census Survey Processing System

DES : Department of Economics Statistics

GDP : Gross Domestic Product

GIS : Geographical Information System

GPS : Global Position System

GRNN : Generalized Regression Neural Networks

LSF : Large Scale Farmers

MAE : Mean Absolute Error

MFS : Multiple Frame Sampling

MINAGRI: Ministry of Agriculture and Animal Resources

MLR : Multiple Linear Regressions

NAEB : National Agricultural Export Development Board

NISR : National Institute of Statistics of Rwanda

PCA : Principal Component Analysis.

PLR : Polynomial Linear Regression,

RF : Random Forest

RMSE : Root Mean Square Error

SAS : Seasonal Agriculture Survey

1. General Introduction

1.1. Background of study

The agriculture sector is the backbone of economy of Rwanda where more than of 85% of active population is engaged in this economic activity. It represents 33% of Gross Domestic of Product (GDP) of Rwanda. Actually, Rwanda's GDP increased at the rate of 11.9% in 2019([Online], 2020c). The cultivation period in Rwanda is divided into three cultivation seasons, Season A which is conducted from September to February of the following year, Season B which starts from March to June and Season C which take place from June to August(NISR, 2019a)

Crop Yield prediction is considered as preeminent area of research which can contribute for ensuring food security and nutritional outcomes inside households all around the world(Ahamed, Mahmood, Hossain, Kabir, et al., 2015). Rwanda has developed the national Agriculture policy to outline how they will achieve the goal of agriculture pillar of vision 2020. It focus on increasing productivity by using modern inputs such as fertilizer, pests and also encouraged to switch to high value crops and consolidate farm plots through farmer cooperative. Therefore, in order to gain full whole advantage of soil and climate changes in Rwanda, farmers essentially need to know definitely the best crop to saw among major crops according to district they belong and also the whole economy depends on annual production harvested.

In order to take full advantage of the soil and Sub-Saharan climate change of Rwanda, farmers need to know exactly the best crop to plant among the main crops according to district they belong and also the entire economy depends on production from harvesting annually.

Nowadays the environment factors conditions are not like previous decades in all districts of Rwanda, day by day are changing because of the globalization, so farmers have faced difficulties to predict the environment factors such rainfall, humidity, temperature, and evaporation. It is relevant to take into consideration the climate factors for all districts separately and also by take into account the seasons according to Meteorology Stations as the climate change of Rwanda is one of the most meaningful factors influencing year to year crops

Production(Mikova,2015).

This can help the decision makers on advising farmers for the best crop to saw according to the season. The climate factors vary for a region to another and have strong impact on farming. For example, too little or too much rain can kill crops, the appropriate amount of precipitation and temperature lead to an increase of crop yield. While it rains, humidity also comes around and it increases the degree of water that is consumed by atmosphere and it may give rise to the crops by

remaining too dry or too wet which has effect to the yields (Ahamed, Mahmood, Hossain, & Kabir, 2015). The quantity of pesticides used can also affect the crop negatively or positively, that is why it should be taken into consideration during the sowing activity. In this research, I consider the effect of environment factors like rainfall, humidity, temperature, evaporation and sunshine, input parameters such as use of fertilizers, pesticides production area, types of seeds and quantity of seeds sown as factors to wards crop production in Rwanda. Taking these factors into consideration as datasets for various districts, we apply suitable data mining techniques to obtain crop yield predictions.

1.2. Statement of the problem

The agriculture sector is the main economic activity in Rwanda which occupies 80% of employee and 33% contribution of GDP. In recent years, this sector gained relevant growth of 4.5% to 6% per year and is progressively shifting from nourishment to a market-oriented model([Online], 2020d). Crop yield estimate is the most used impact indicator of agricultural productivity activities and tracks yield gaps for major commodities. It is computed by NISR in partnership of MINAGRI through Seasonal Agriculture Survey (SAS) for monitoring the agriculture programs and policies as well as addressing key agriculture issues relevant for policy makers and other stakeholders. Some of the most valuable recent works have focused on estimating crop yield by considering the total reported quantity of harvested over the harvested area of that crop without considering the other factor which may affect the crop yield production in Rwanda(NISR, 2014).

The prediction of yield for various crops by using different data mining techniques is an important study as it is a crucial indicator which may affect daily life of population. It could be done first and foremost, in order to ensure and understand food security, the ability to produce enough food to meet human needs in the foreseeable future(Raghuveer, 2014). Secondly, each crop has its potential yield which is an estimate of what its yield will be. The farmers could make comparison of their crop yields to predict yield and see how successful. Their agricultural activity and they could also compute the difference between their outputs and potential yield to check their yield gap. The stakeholders and policy makers often rely on accurate prediction to make timely decisions on export and import in order to strengthen national food security(Horie, 1992). For completion of this work, several data mining techniques can be chosen to process the data, improve the quality and the reliability of various datasets. The aim is to identify key factors that may affect crop yield, such as geographical location, environment factors, nutrition, type of seeds, harvested quantity, crop area, etc. In this study, the purpose is to use several data analysis methods with the focus on

mining techniques in order to help farmers find the combination of traits required to increase their crop yield as possible. Crop yield prediction is a key task that is important for planning and is used for taking various policy decisions. In Rwanda, we use the conventional techniques to compute annual yield prediction based on ground based visits and report. However, this method is subjective, unreliable, untrusted statistically as it does not consider various factors which could affect crop yield in general.

1.3.Research Objectives

1.3.1General objective

This study aims to contribute to the prediction of annual yield of major crops and elaborate a couple of recommendations for planting different crops in various districts of Rwanda.

1.3.2. Specific objectives

(i) To assess the effectiveness of Artificial Neural Network model on crop yield prediction for typical environment factors.

(ii) To compare the effectiveness of Multiple Linear Regression model with Polynomial Linear Regression model and Artificial Neural Network model on crop yield prediction

(iii) To assess the impact of various agriculture inputs that affect crop yield production of major crops in Rwanda through data mining techniques.

(iv)To propose some recommendations to decision makers about future crop yield based on finding from historical data.

1.4. Hypotheses of study

In my Research, I would like to set the following hypotheses based on the objective of my study.

H1: The Artificial Neural Network could effectively predict the crop yield production with respect to the environment factor and other agriculture inputs.

H2: There is no difference between yields predicted by Polynomial Linear Regression, Multiple Linear Regression and Artificial Neural Network models

H3: The use of improved agricultural inputs did not have greater impacts on the yield of major crops in Rwanda.

H4: There is effective impact of fertilizer and pests on crop yield production of main crops in Rwanda. The improved inputs

1.5. Scope of the study and limitation

The study will be carried out in sampled plots collected from SAS in all districts of Rwanda from 2017 to 2019. It will also take into consideration the data related to environment factor like soil temperature, sunshine, Rainfall, temperature, humidity and evaporation collected from Rwanda Meteorology Agency. In addition, the study aims to take into consideration the biotic data such as soil salinity and Hydrogen potential (PH) those are potential factors for yield prediction

1.6. Definition of key concepts

Data Mining: Data mining is defined as an extraction of useful information from existed dataset and transforms them into human understanding

Crop Yield: Is a highly complex trait determined by multiple factors such as genotype, environment and their interactions. Simply it can be defined as a measurement of the amount of Agricultural production harvested per unit of land area Agriculture: Is a basic source of food supply of all countries of the world whether under developed, developing or even developed. Due to heavy pressure of population in underdeveloped and developing countries and its rapid increase, the demand for food is increasing at a fast rate.

Food security: It is defined as the condition in which all people at all-time have social, physical and economic access to nutritious food that is sufficient, safe and meets their dietary needs and food preferences for an active and healthy life.

Machine learning: I is a subset of artificial intelligence which helps to use historical data to make better decisions. It is also a process where a machine takes data, analyze it to generate predictions, and use those predictions to make decisions. Those predictions generate results and those results are used to improve future predictions.

Artificial Neuron Network: It is a popular machine learning algorithm that attempts to mimic how the human brain processes information. It provides a flexible way to handle regression and classification problem without the need to explicitly specify any relations between the input and output variables.

1.7.Expected outputs

(i) At the end of this study, the following outputs are expected to be achieved:

(ii) Developed model which could be used to predict crop yield production in Rwanda

(iii) Recommendations related to future crop yield addressed to decision makers

(iv)The farmers will be advised about the type of crops by seasons that are appropriate to plant according to their district of residence.

(v) At least one scientific research paper will be published in a peer reviewed journal

2. Literature Review

2.1Overview of Agriculture Development

The Agricultural activity is key to food security as a source of food, source of income, nutrients and directly dictate the price of food (Satyal, 2010). Agriculture development is one of the effective tools to reduce and possibly end extreme poverty, boost prosperity sharing and feed a projected 9.7 billion people in the whole world by 2050.

Therefore agriculture driven growth, food security and poverty reduction are to be given high priority. Climate change has impact on crop yields and highly affects the world's most food insecure regions. In 2020, shocks associated to conflict, climate change, pests and emerging infectious diseases are hurting food production, disrupting supply chains and stressing the ability of people to have access to nutritious and affordable food.

Our beloved continent of Africa has enormous potential, not only in sustainability of feeding itself and eliminates hunger and food insecurity but also to become a major player in global food markets (Satyal, 2010). The big part of African continent has potential land, water and ocean that have been used by the Africans to make the agriculture as one of the pillars of African Development. The agriculture activity is a significant portion of all African countries as sector that contributes to major continent priorities to

- Eradicate poverty and hunger
- Boost investments and intra-Africa trade
- Enlarge rapid industrialization and economy
- Improve management of sustainable resources and environment
- Create Jobs
- Ensure human security and shared prosperity

According to the Comprehensive African Agriculture Development Programmed (CAADP), the African countries have increased investments in agriculture and they have seen reduction in hunger and poverty and increase in productivity. However, there has been no significant.Improvement in production factors such as land and labor but instead the agricultural growth in Africa is generally achieved by cultivating large area of land and larger agricultural labor force

The Sub-Saharan Africa itself represents more than 950 million peoples, approximately thirteen percent of the global population. The projection in 2050 is that this share may increase to almost 22%. Moreover, malnutrition has been a long- standing challenge, with inconsistency progress across the region. The fact that food security in African has progressed slowly in last decades has been attributed to low productivity of agricultural resources, political instability and high population growth rates. That is why role of agricultural sector in contributing to food security is considered as priority in the development agenda. The agricultural sector has high contribution to Gross domestic product (GDP), around 15% for African economies.

In Rwanda, as the agriculture is the backbone of the economy, it is crucial for economy growth and reduction of poverty, it counts for 39% of the GDP([Online], 2020a). The agriculture sector in Rwanda faces the following challenges:

- Constraint of land due to population pressure
- Small average of land holdings
- Poor water management
- Poor access to output and financial markets that lead to limited commercialization

In partnership with World Bank, Rwanda has expertise in agriculture intensification and watershed management in hillside in the Easter-African region. This project has been established to increase the productivity and commercialization of hillside agriculture targeted areas. Rwanda continues to develop this sector by training different farmers, increasing the improved seeds to the farmers, by providing organic fertilizer to different farmers in various districts and by increasing the irrigation sites in different districts especially is Eastern province.

2.2. Crop Yield prediction

Crop yield is defined as the quantity of crop harvested per unit of land area (Kg per hectare) which is used to determine the efficiency of food production(NISR, 2019b). Crop yield prediction is an important subject of research in agriculture. Every farmer needs to know how much yield he/she should get from his/her expectations. The Agriculture yield primarily depends on environment factors such as weather condition and agricultural inputs. AccurateInformation about history of crop is also important thing for making decisions related to agriculture risk management(Manjula & Djodiltachoumy, 2017).

2.2.1. The role of crop yield prediction

The crop yield prediction may be important in the following ways:

- Crop yield is so important for decision making related to agriculture risk management and future prediction.
- Crop yield prediction is of great importance to predict global food production.
- Policy makers rely on accurate predictions to make timely import and export decision to provide nation food security.
- The farmers and growers can benefit from yield prediction to make financial decisions and informed management.

2.3 .The main influencing factors of crop yield prediction

2.3.1. The environment factors

2.3.1.1. Rainfall

Rainfall is defined as the amount of water in rain, snow within a given time and area, usually expressed as a hypothetical depth of coverage. Precipitation has effect on agriculture, all crops need at least some water to survive, and therefore rain is so important to agriculture while a regular rain is usually vital to healthy plants, too much or too little rainfall can be harmful even devastating to crops.

2.3.1.2.Temperature

Temperature is a weather parameter that directly influences the productivity of agricultural plants. All biological and chemical processes taking place in the soil are connected with air temperature. The heat supply of crops is characterized by a sum of average daily air temperatures that are higher than a biological minimum during a vegetable period.

The growth of plants is possible comparative broad temperature limits. There are three main types of temperature point of growth, Minimum temperature which is enough for growth to start. The optimal one which is most advantageous for growth process and maximum one where growth can immediately stop, for major part of the vegetative world, temperature rise up to 25-28 degree Celsius increases the activity of photosynthesis, and with a higher

than 30 degree Celsius, photo-respiration starts to prevail over photosynthesis significantly. Obviously, the temperature is so important on crop yield production as it is required for different several of plant species and it also varies for specific parts of plant.

2.3.1.3. Humidity

Humidity is the concentration of water vapor present in the air. The amount of water vapor needed to achieve saturation increases as the temperature increase([Online], 2020b). Humidity is very important to make photosynthesis possible. If the plant loses too much water, the stigmata will close the result that photosynthesis stop([Online], 2020). Reduction of photosynthesis leads to low crop yields. High air humidity are favorable for many plant diseases and insect pests. It increases the growth of shoots and leaves at the expense of crop yields.

2.4. Data Mining Concept in Agriculture

2.4.1. Introduction

In general, Data mining is the process of founding out previously potentially and unknown interesting patterns and insights in large datasets. It can also be defined as the process of discovering important and useful information from large sets of data(Abello J, 2002). Data Mining is mainly divided into descriptive and predictive data mining. However in the agriculture sector, predictive is mainly used for prediction(Jaganathan et al., 2014). Data mining software is an analytic tool that practitioners use to analyze data from many different dimension and angles, categorize and summarize the relationship identified(Manjula & Djodiltachoumy, 2017).

2.4.2. Data Mining Tasks

The data mining tasks is classified generally into two types based on what a specific task can accomplish. Those classifications are descriptive tasks and Predictive tasks. The descriptive data mining tasks can compute the characteristics of general properties of data while predictive data mining tasks compute the inference statistics on the available data set and forecast how the new data set will behave.

2.4.3. Different Data Mining Tasks

There are a various number of data mining tasks which are prediction, classification, time- series analysis, summarization, association and clustering. All these tasks are either descriptive or predictive data mining tasks. A data mining system can execute one or more of the above specified task as part of data mining activity.

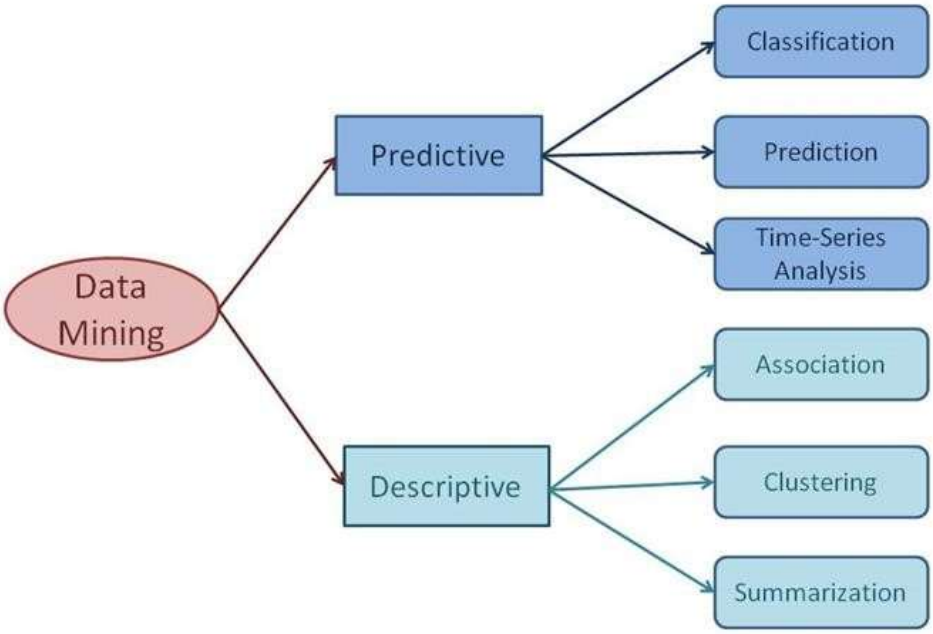

Figure 1: Types of data mining techniques

2.4.3.1. Predictive data mining tasks

Predictive data mining tasks are determined by created model from the available dataset that is helpful in predicting unknown or future values of another dataset of interest.

(i) Classification: Classification is driven by a model to determine the class of an object based on its attributes. A collection of records will be available, each record with a set of attributes. One the attributes will be class attribute and the goal of classification task is assigning a class attribute to new set of records as accurately as possible.

(ii) Prediction: Prediction task is a task that intends to predict the possible values of missing or future data. Prediction involves developing a model based on the the available data and this model is used in predicting future values of a new dataset of interest.

(iii) Time Series Analysis: Time series is a sequence of events where the next event is determined by one or more of the preceding events. Time series analysis is a method which can be used to analyze time on parametric and series data to extract the useful information and other characteristics of the data(Manjula & Djodiltachoumy, 2017).

2.4.3.2. Descriptive data mining tasks

Descriptive data mining tasks usually finds data description patterns and comes up with new significant information from the available data set. This task is categorized into the following types:

(i) Association: Association discovers the relationship among a set of items. Association analysis is used for advertisement, commodity management, catalog design, direct marketing etc.

(ii) Clustering: Clustering is used to discover data objects that are similar to one another. The similarity can be decided based on a number of factors like responsiveness to certain actions, purchase behavior, geographical location and so on.

(iii) Summarization

Summarization is the generalization of data where combination of relevant data is summarized which results in a smaller set that gives aggregated information on the data. Data can be summarized in different abstraction levels and from different angles.

2.4.4. The most Data Mining techniques used in Agriculture

Data mining techniques are most widely used in various sectors such as business and corporate sectors and can also be used in agriculture for data characterization, predictive, discrimination and forecasting purposes. Different techniques have been intended for mining data over the years, the most used data mining techniques are Linear Regression, Multiple Linear Regression, k-means clustering, bi-clustering, k nearest neighbor, Artificial Neural Network, support vector machine and Naive Bayes classifier in the Agriculture field(Raghuveer, 2014). The following classification or regression is most used in agriculture to predict crop yield.

2.4.5. Regression Models

2.4.5.1. Linear regression

Linear regression is a statistical approach to measure the relationship between on dependent variables and a series of other changing variables known as independent variables or explanatory variables. When the independent variable composed with multiple attributes like rainfall, temperature, humidity, solar radiation, sunshine etc., then it is termed as Multiple Linear Regression. Linear Regression provides a model for the relationship between a scalar variable and one or more explanatory variables(Ye, 2013). The linear relationship modeling may be done by fitting a linear equation to the observed data. The equation below may show how simple linear regression equation looks like

$$Y = \beta_0 + \beta_1 x_1 \quad (1)$$

Where:

- Y is dependent variable
- β_0, β_1 are the regression parameters
- x_1 is the independent variable

2.4.5.2. Multiple Linear Regressions

Multiple Linear Regression (MLR) is one of the important algorithms that attempts to model the linear relation between a single dependent variable with two or more independent variables by fitting a linear equation to observed data. Moreover, MLR is an extension of simple linear regression as it takes more than one predictor variable to predict the response variable. The Multiple Linear Regression equation is given by:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n \quad (2)$$

where

- Y is dependent variable
- β_0, β_1, β_2, $\beta_3 \ldots \beta_n$ are the regression parameters
- x_1, x_2, x_3,$\ldots x_n$ are the independent variables

Some key points about Multiple Linear Regression:

- For MLR, the target variable must be the continuous, but the predictor or independent variables may be of continuous or categorical form.
- Each Feature variable must model the linear relationship with the dependent variables
- Multiple linear regressions try to fit a regression line through a multidimensional space of data points. MLR is what you can use when you have a bunch of different independent variables, it has three main uses:
- You should have to look at the strength of the effect of independent variables on the dependent variable
- You should have to look at the strength of the effect of independent variables on the dependent variable

- You can use it to examine how much the dependent variable will change if the independent variables are changed
- You can also use to predict trends and future values

2.4.5.3. Machine Learning Polynomial Regression

Polynomial Regression is a regression algorithm that models the relationship between a dependent y and Independent variable x as n^{th} degree polynomial. The polynomial Regression equation is given below:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n \quad (3)$$

where:

- Y is a response variable
- x, x^2, x^3 and x^n are the explanatory variables
- β_0, β_1, β_2, β_3 and β_n are the regression parameter

Some key points about Polynomial Regression

- Polynomial Regression is also called the special case of Multiple Linear Regression in Machine Learning because we add some polynomial terms to the multiple linear regressions to transform it into polynomial regression

- It is a linear model with some modification in order to increase the accuracy
- The dataset used in polynomial regression for training is non-linear nature
- It makes use of linear regression model to fit the complicated and non-linear function and dataset. The need of Polynomial Regression in ML can be understood in the below points:
- If we apply a linear model on a linear dataset, then it provides us good results as we have seen in simple Linear Regression, but if we apply the same model without any modification on a non-linear dataset, then it will produce a drastic output. Due to which loss function will increase, the error rate will be high, and accuracy will be decrease.
- So for such cases, where data points are arranged in a non-linear fashion, we need the polynomial Regression model. We can understand it in a better way using the below comparison diagram of the linear dataset and non-linear dataset.
- In case where data points are arranged non-linearly. Then we should use the polynomial Regression model instead of Simple Linear Regression. The following figure can explain much more

about the polynomial Regression model and it shows how the data are fitted based on the complexity model, on the left the linear dataset and non-linear dataset on the right.

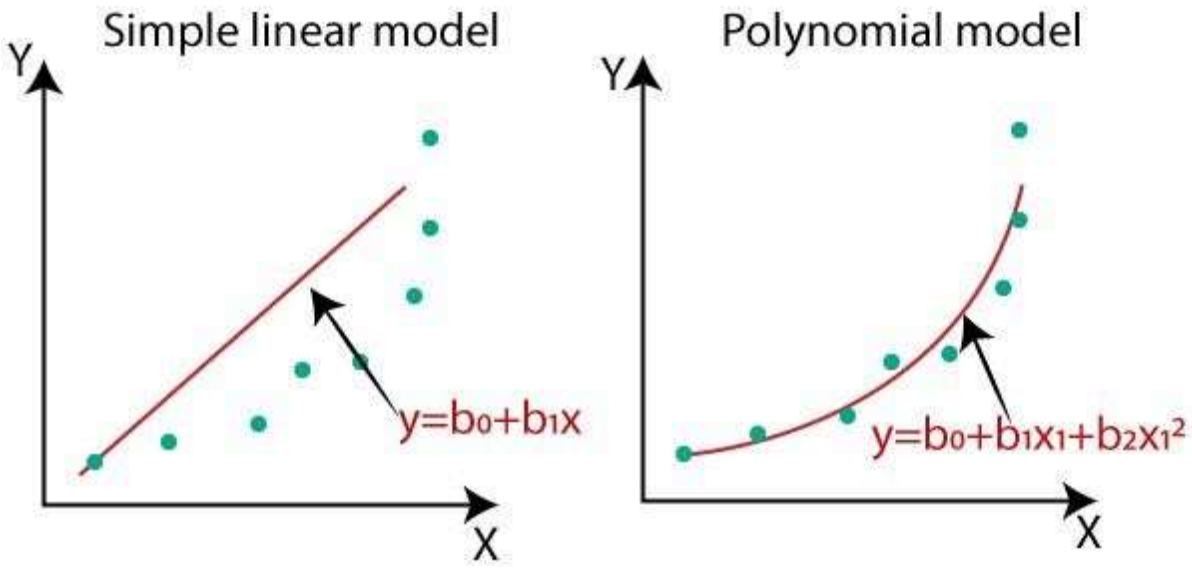

Figure 2: Comparison of linear and polynomial regressions model

2.4.6. Artificial Neural Network

Artificial Neural Network (ANN) is a mathematical model formulated based on the structure and functional aspects of biological neural networks for instance in our brains. In most cases an ANN is an adaptive system that transforms its structure based on external or internal information that flows through the network during the learning phase(Ye, 2013). An ANN is typically defined by three types of parameters:

(i) The interconnection pattern between different layers of neurons

(ii) The learning process for updating the weight of interconnections

(iii) The activation function that convert a neuron's weighted input to its output activation

In artificial Neuron network, the network receives inputs by neurons in the input layer, and the amount of the network is given by the neurons on an output layer. There can be one or more intermediate hidden layers(Lavanya & Parameswari, 2020).

.

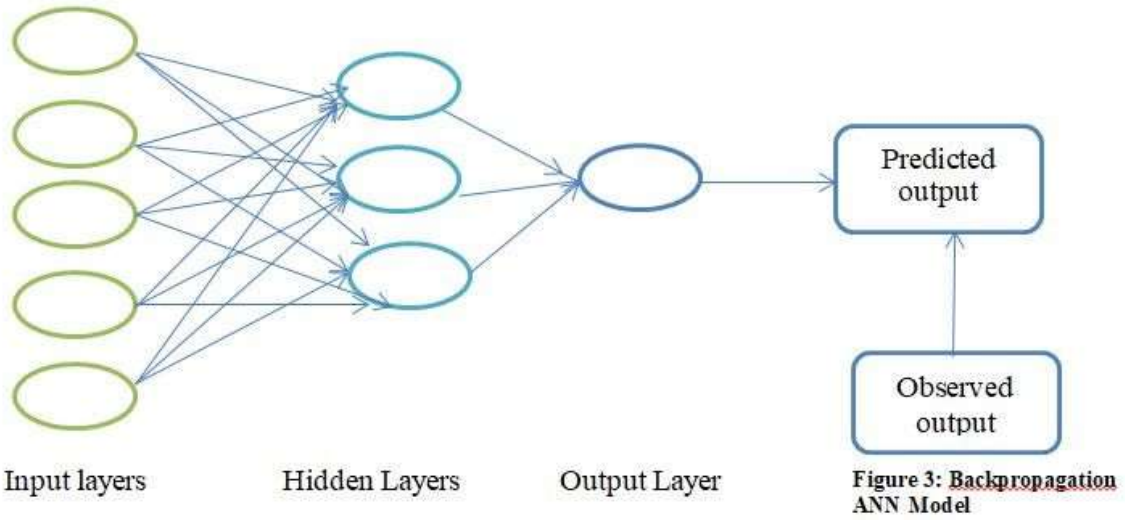

Figure 3: Backpropagation ANN Model

Figure 3: Back propagation of ANN model

The first term feed forward describes how this neural network processes and recalls pat- terns. In a feed forward neural network, neurons are only connected foreword. Each layer of the neural network contains connections to the next layer (for example, from the input to the hidden layer), but there are no connection back. The term back propagation describes how this type of neural network is trained. Back propagation is a form of supervised training.

When using a supervised training method, the network must be provided with both sample inputs and anticipated outputs. The anticipated outputs are compared against the actual output for given input. Using the anticipated outputs, the back propagation training algorithm then takes a calculated error and adjusts the weights of various layers backwards from the output layer to input layer. Know that the back propagation and feed forward algorithms are often together.

2.5.Empirical studies

This section has reviewed some of them; more attention has put in methodology used, data, the models, findings and limitations of those studies(Ahamed, Mahmood, Hossain, & Kabir, 2015) have conducted a research study of applying data mining to extract knowledge from the agriculture data to estimate crop yield for major cereal crops in major districts of Bangladesh. Primary data used in that research was collected from BARI (Bangladesh Agriculture Research Institute), those data were in pdf format which were converted to rtf format by using miscellaneous tools and tricks. In that study, a lot of pre-processing was required to handle missing value, noise and outliers. From the dataset, they have processed and selected only the important attribute like rainfall, temperature, humidity, irrigated area for all districts and cultivated area for every crop considered according to the districts. The other environment attributes sunshine and two further

biotic attributes which are soil salinity and soil pH were considered in that research, those data were collected from the Bangladesh Agricultural Research Council (BARC). After necessary pre-processing, the selected crops to work on that study were Rice-AMON, Rice-AUS, Rice-BORO, potato and Wheat. The average yearly of environment factors were considered by calculating the average from the monthly rainfall, temperature, etc of each district. Different Data mining techniques like linear regression, K-nearest Neighbor, Artificial Neural Network were used to predict crop yield in Bangladesh. The Root Mean Square Error (RMSE) is used to evaluate the performance of those models, different model provide better results for different crops but ANN provide better prediction for some of crops which have more missing values than others.

Manjula and Djodiltachoumy (2017) have done a study of crop yield prediction by using data mining techniques based on association rule of selected region in district of Tamil Nadu in India. Before starting data mining problem, they brought together the data, the proposed data were the data collected from 2000 to 2012 for district of Tamil Nadu in India. Each area in this collection has been identified by the respective longitude and latitude of the region.

The data were taken in nine input variable such as Year, district, Crop, Area, Tanks, Bore wells, Open Wells, production and Yield. The pre-processed data was clustered using k-means algorithm and the association rule mining process will apply on clustered data to find the rules.

Once the data has been collected and converted, association rule mining begun, in this step rules are created using frequent pattern mining. Association rules identify the relationships among a set of items or objects in a database.

In this study, different data mining techniques like ANN,K-nearest neighbours and decision Trees have been used for selection of appropriate crop that will be sown by considering different factors such type of soil and its composition, climate, geography of the region, crop yield, market prices. Crop selection based on the effect of natural calamities like famines has been done based on machine learning(Okori, W. and Obua, 2011).

Maize crop forecasting has been done by using multilayer feed forward network of ANN (Kaur & Singh, 2011). The Generalized Regression Neural Networks (GRNN) method has been used for fore- casting of agriculture crop production(Chaochong, 2008). They found that GRNN is good technique for predicting grain production in rural areas and it was reported GRNN Model is suitable for non-linear, multi-objectives and multivariate forecasting. The crop yield was predicted based on the generated rules. The district and crop is input to the prediction model, the overall accuracy was 89% and error rate was 11% of prediction results. The aim of this research was to propose and implement a rule based system to predict crop yield

production from the collection of past data and it has been achieved by applying association rule mining on agriculture data from 2000 to 2012.

2.6. Conceptual Framework

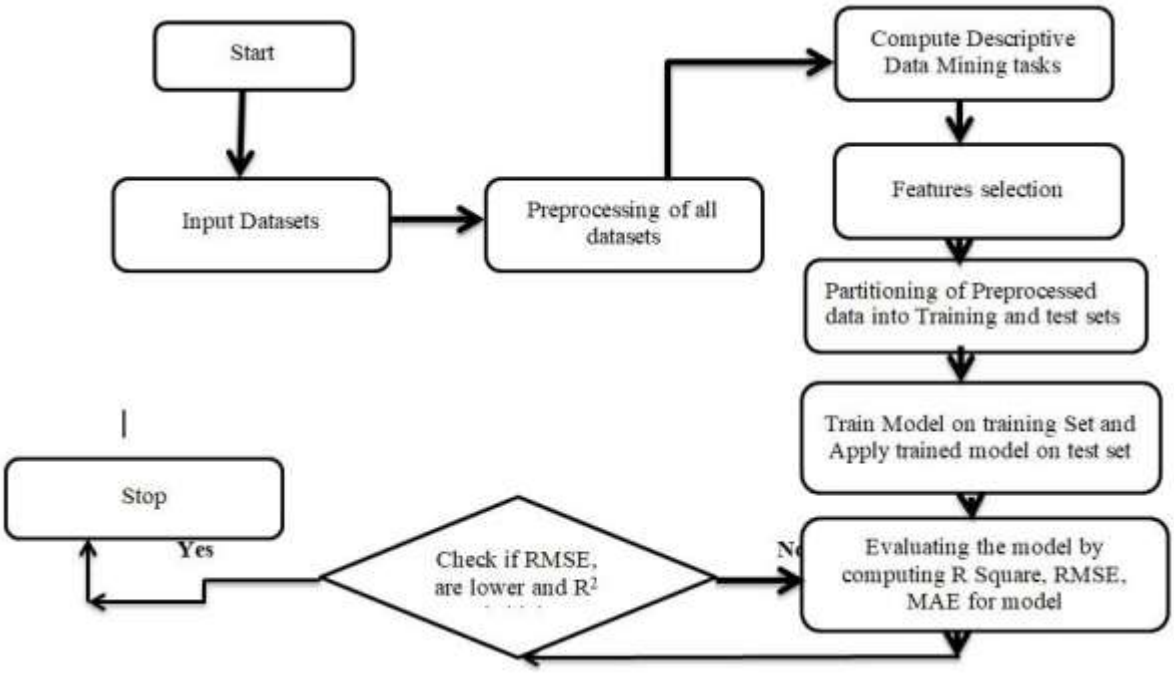

Figure 4: Conceptual Framework of models

3. **Methodology**

3.1 .Study area

The study was carried out in whole country of Rwanda. Rwanda is bordered by Tanzania to the East, Uganda to the North, to the West by The Democratic Republic of Congo (DRC) and by Burundi to the South. It is 26,338 square kilometers of Surface Area(Online, 2020c). In 2018/2019, the estimated agricultural land were of 1.4 million hectares (59% of total country land) from which 1.14 million of hectares was arable land. During this year the physical year crop cultivated was 1.1 million of hectare (79% of total agriculture land) and increased by 16.8% comparing to 2017/2018 agriculture year. Out of physical cultivated area, 1 million hectares was used for seasonal crops while 0.5 million hectares of land under permanent crops(NISR, 2019). The current population of Rwanda is 12891797 based on world meter of the latest United Nations data. The agriculture practiced by more than 80% of total population of Rwanda.

3.2. Sample Design

This survey is conducted based on complete coverage of the farm level and is better way of collecting the agriculture data and findings are much more precise. In sampling design SAS uses Multiple Frame Sampling (MFS) by which area frame and survey sample is drawn. A part from that, a List Frame is done for Large Scale Farmer, with at least ten hectares of agricultural holdings. List frame is done to complete area frame. This process of area frame is done by doing land Stratification and segment sampling. The purpose of stratification in this survey is to construct a frame by dividing the land of every district into homogenous land use groups, among which only strata with potential agriculture land are taken into consideration for allocating the sample to each stratum(NISR, 2019a).

3.2.1 Summary of sampling process

Among 10 strata, only 4 strata are considered to represent the country land potential for agriculture and they cover total area of 1787571.2 hectares. Those strata are 1.0 for tea plantation, stratum 1.1 for intensive agriculture land on hillsides, stratum 2.0 for intensive agriculture land on marshlands and stratum 3.0 for rangelands. SAS cannot consider the data from strata 1.0 as they are always monitored by NAEB. At sampling process, the agriculture strata 1.1.2.0 and 3.0 were divided into large sampling units called Primary Sampling Units (PSUs).

The strata 1.1 and 2.0 were divided into PSUs of around 100 hectares while strata 3.0 were divided into PSUs of around 500 hectares. After sample size determination, a sample of PSUs was done by systematic sampling method with probability proportion to size, then a given number of PSUs to be selected for each stratum, was assigned in every district. In 2019, the 2018 SAS sample of 780 segments has been kept the same for SAS 2019. At first stage, 780 PSUs sampled country wide were proportionally allocated in different levels of stratification such as hillside, marshland and rangeland for 30 districts of Rwanda, this done for allowing the publication results at district level. PSU sampling in each stratum is done systematically from frame. Secondary; 780 sampled PSUs were divided into secondary sampling units (SSUs) called segments. For stratum 1.1 and 2.0 the segment is estimated to be round 10 hectares and for stratum 3.0 is rounded 50 hectares. Thirdly, for each PSU, only one SSU is selected by random sampling method without replacement(NISR, 2019b).

3.3.Usage data

In my study, I manage to use secondary data; those data were collected in National Institute of Statistics of Rwanda under department of economics statistics (DES). The data to use are data

collected in three consecutive years of seasonal Agriculture Survey (SAS) such us 2017, 2018 and 2019. In study of crop yield prediction, the environment factors data like rainfall, min temperature, max temperature, humidity and solar radiation are mostly considered in order to increase the accuracy of our model and findings.

3.4. Data collection method

3.4.1. Data collection Design

Seasonal Agriculture survey Data collection is done in two distinct phases:

- The first phase know as screening activity which is consisting of visiting all sampled segments and delineating all plots in which the sampled grids points are fallen and thereafter recording the information related.
- The second phases consists of visiting the sub-sampled agricultural plots from screened plot in first phase as well as Large Scale Farmers having cultivated plots in the season the survey is being conducted.

3.4.2.Data collection tools and materials

Seasonal Agriculture survey uses two main questionnaires:

- Screening Questionnaire: Screening Questionnaire is a form used to collect information on the sampled plot mainly related to land use, plot area and crop cultivated.
- Plot Questionnaire: A plot questionnaire is a form used to collect information on the sampled plot mainly to crop production, inputs used (seeds, fertilizers, pests, labor,..) and agriculture practices. During data collection of SAS, the computer assisted personal Interview (CAPI) data collection methodology was based on three different applications:
- Collector which is GIS based and is used to help identify and navigate the enumerator team to the exact GPS point-sample location and for mapping the plot boundary for GIS area determination.
- Survey 123 is an electronic instrument for farmer interview data collection
- CS entry which is a CSPRO data entry application used on android tablets and suite of data processing tools.

3.4.3.Field data collection procedure and Methods

Before proceeding too field, the field worker have to check if they have required materials for the field work activities, every fieldwork is required to arrive early in field for both segment and LSF interview. The field worker are require to familiarize themselves with the segment boundaries by

using tablet GPS to guide them to each of point sampled and geographical coordinate, time and date are recorded automatically during enumeration process to know the actual time of field work.

- Screening Activity of the segment: After locating and identifying segment boundaries, the fieldwork team proceeds to let their tablet GPS guide them to locate accurately each of the sampled grid point inside the segment; arriving to grid point field work identify plot boundary with the guide of the farmer and delineate the plot by using collector app with GPS connected to tablet. For LSF complete enumeration of plots with at least 10 hectares is done.
- Farm Interview: After screening activity, only subsampled plots are given to enumerator to collect information provided in the plot questionnaire, CAPI data collection method allows the field workers in the field to collect and enter data with their tablets and then synchronize them to the server at headquarter where data are received by NISR staff, checked for consistency at NISR and thereafter cleaned data transmitted to analyst team for tabulation.

3.5. Data analysis

3.5.1Data mining tools

During the extraction of knowledge from data and building of predictive model, I have to manage to use python programming language, STATA and Microsoft Excel.

3.5.2.Major crops considered in this research

The main objective of data mining is to find the useful patterns from existing data. Among the main crops I should have to consider some crops which could be founded in each district of Rwanda and which have enough records for building a predictive model. The crops I have to focus on this study are beans, maize, paddy rice and Irish potatoes.

3.5.3.Used variables

The process of data wangling and crop yield prediction is very complex, since it deals with large data situation which comes from a number of factors. In our research of crop yield prediction in Rwanda, the following 16 factors are considered as predictors

3.5.3.1.Environment factors

(i) Rainfall: The average seasonal rainfall is considered by calculating average from monthly rainfall (mm) of each district. Usually, the year that contains the highest average rainfall should provide for maximum crop yield in that year.

(ii) Humidity: Similar to the way I compute the rainfall average, I also calculate and obtain the average monthly humidity for each district.

(iii) Max temperature: Variation in temperature through the year puts a great impact in that seasonal crop production, hence I have to consider the maximum as well as the minimum temperature in our research.

(iv)Min temperature: The average of Seasonal minimum temperature will be considered in our research in Celsius.

(v) Solar radiation: solar radiation is defined as light energy from the sun. It can furnish the light required for seed germination, leaf expansion, growth of stem and shoot, flowing, fruiting and thermal conditions necessary for the physiological functions of the plant. It always plays the role as regulator and controller of growth and development.

3.5.3.1. Agriculture inputs

(i) Organic fertilizer: This variable has the purpose of knowing if the farmer used organic fertilizer; it should be change to dummy variable

(ii) Inorganic fertilizer: This variable has the purpose of knowing if the farmer used inorganic fertilizer and it should be transformed into dummy variable.

(iii) Quantity of inorganic fertilizer: This variable should specify the quantity of inorganic fertilizer used in a plot in kilogram; the value is used as it.

(iv)Pesticide usage: This variable has the purpose of knowing if the farmer used pesticides; it should be transformed to dummy variable.

(v) Usage of irrigation activity: This variable has the purpose of knowing if the farmer irrigates the crop of that plot; it should be transformed into dummy variable

(vi)Type of seeds: In SAS, this variable specify the type seeds sown in a plot, it classify the seeds for improved seeds, tradition seeds and mixed. Those types of seeds will be used as a number 1, 2, 3 consecutively in order to be used in a model for yield.

(vii) Quantity of seeds sown: This is the quantity of seeds sown in each crop; it should be used as it is.

Only one variable is considered as targeted variable or predicted variable, this variable is called crop yield.

3.5.4.Descriptive Statistics

Descriptive statistics are employed to describe the general attributes and to summarize the pertinent information about them. Different crops characteristics are examined by using both frequency distribution and arithmetic mean. In this study, the interquartile used to check the outlier and data mining skills used to replace them on environment attributes. Test of Statistically significant difference such as t-test, p-value and chi-square test could be used to determine the difference between predicted yields from various used regression models.

3.5.4.1. Data Pre-processing

The data related to agriculture were collected by National Institute of Statistics of Rwanda and were in STATA formats. However, they contain the missing values, the outliers and unrelated values, those inconstancies were treated in order to increase data efficient. The outliers has been treated by computing the interquartile and the value which are higher than interquartile multiplied by one point five, plus third quartile and the lowest value which is calculated by quartile one subtract one point five multiply by interquartile are replaced by the mean of non-outliers. All plots with zero production were removed to our dataset in order to avoid the inconsistence to our dataset.

3.5.4.2. The Main statistical observation of Bush bean

In this study, I used different statistical method to summarize and getting more information from data such as mean, standard deviation and inferential statistical in order to draw the conclusion from data.

3.5.4.2.1.The distribution of Bush bean yield by agriculture year in Districts

In order to accomplish this project, I managed to summarize the data by visualizing them crop by crop so that I can retrieve the needed information. On the following image, I would like to show how the yield of bush bean is looking like in different district based on agriculture year.

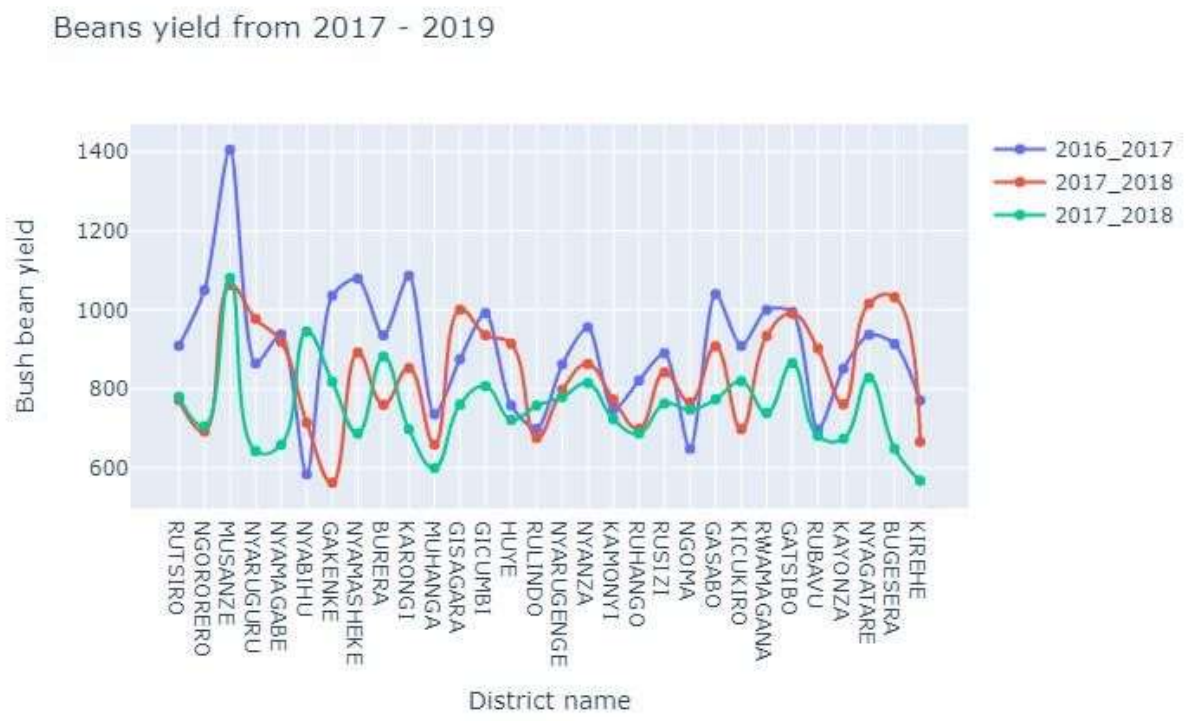

Figure 5: Bush bean yield per Districts in Agriculture year 2017 to 2019

According to the above image, the bush bean yield was extremely efficient in agriculture year of 2016 to 2017 when I compare with 2017 to 2018 and 2018 to 2019 agriculture year. Here we should also observe that Musanze district had high bush yield than the other district in all agricultural year.

3.5.4.2.2. The distribution of Bush bean yield by cropping system in each agriculture year

Generally the cropping system is one the factor which affects the crop yield production, sometime when you have the mixed crops, the production decrease but this statement is different to our case according to upcoming figure. The next figure shows that in our dataset, we had a high yield in case of mixed cropping system; this means that the crops planted with the beans in Rwanda had not the same characteristic to bush beans.

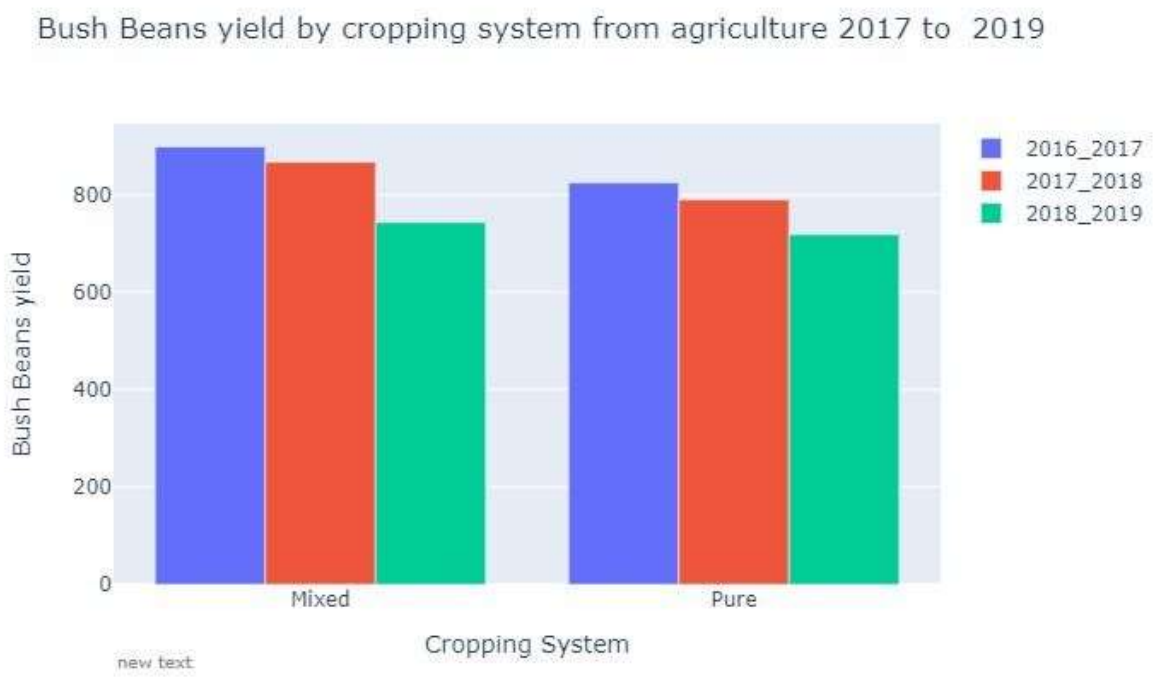

Figure 6: Comparison of Bush bean Yield agriculture year 2017 to 2019 by cropping system

3.5.4.2.3. The distribution of Bush bean yield by seasons in each agriculture year

Different data mining techniques are used to visualize the bush bean data, in the following schema, the crop yield is distributed based on different seasons. The following stacked bar showed that Season A 2018 in Rubavu district had a good yield compare to other seasons when you are considering districts.

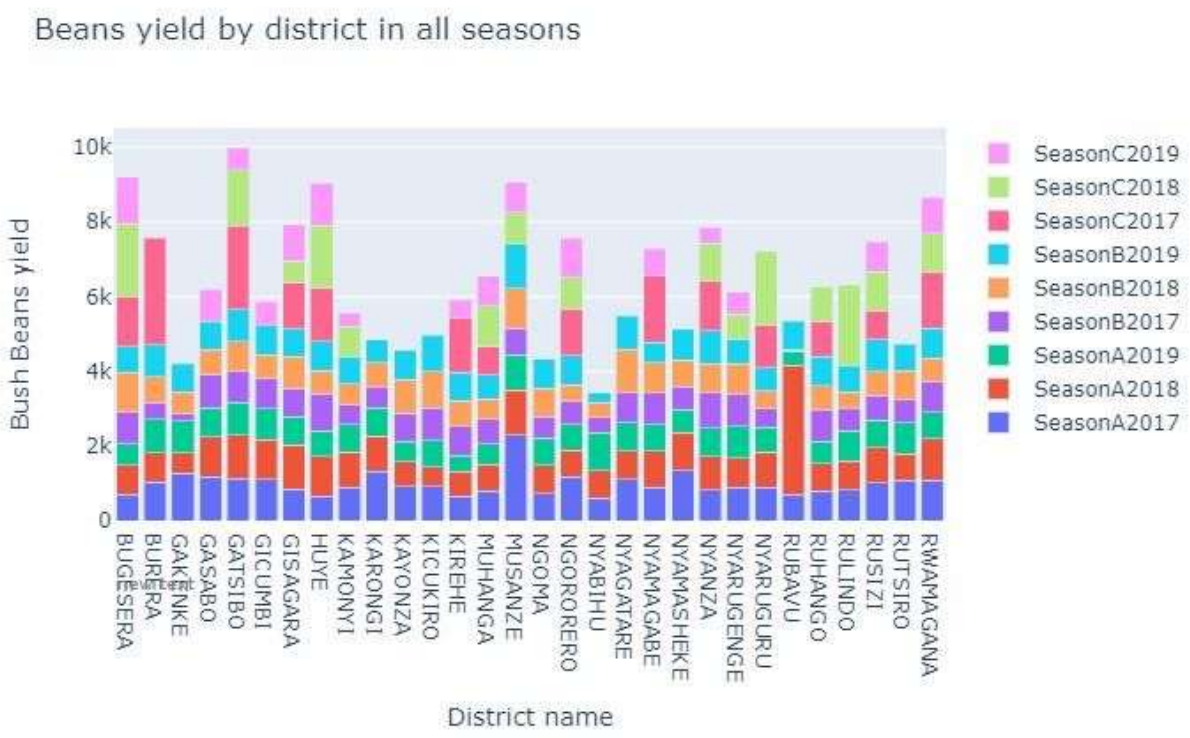

Figure 7: The stacked plot to compare Bush bean yield by seasons

3.5.4.3. The Main statistical observation of Paddy rice

3.5.4.3.1 The distribution of paddy rice yield by agriculture year in all districts

The following image, I want to make a show the value of paddy rice yield is in different district by considering the agriculture year. The following figure demonstrates that the paddy rice yield is high in in agriculture year of 2017 to 2018 especially in KAYONZA and RUSIZI Districts. It also show that NYAMASHEKE District have the lowest yield in Agriculture year in 2018 to 2019.

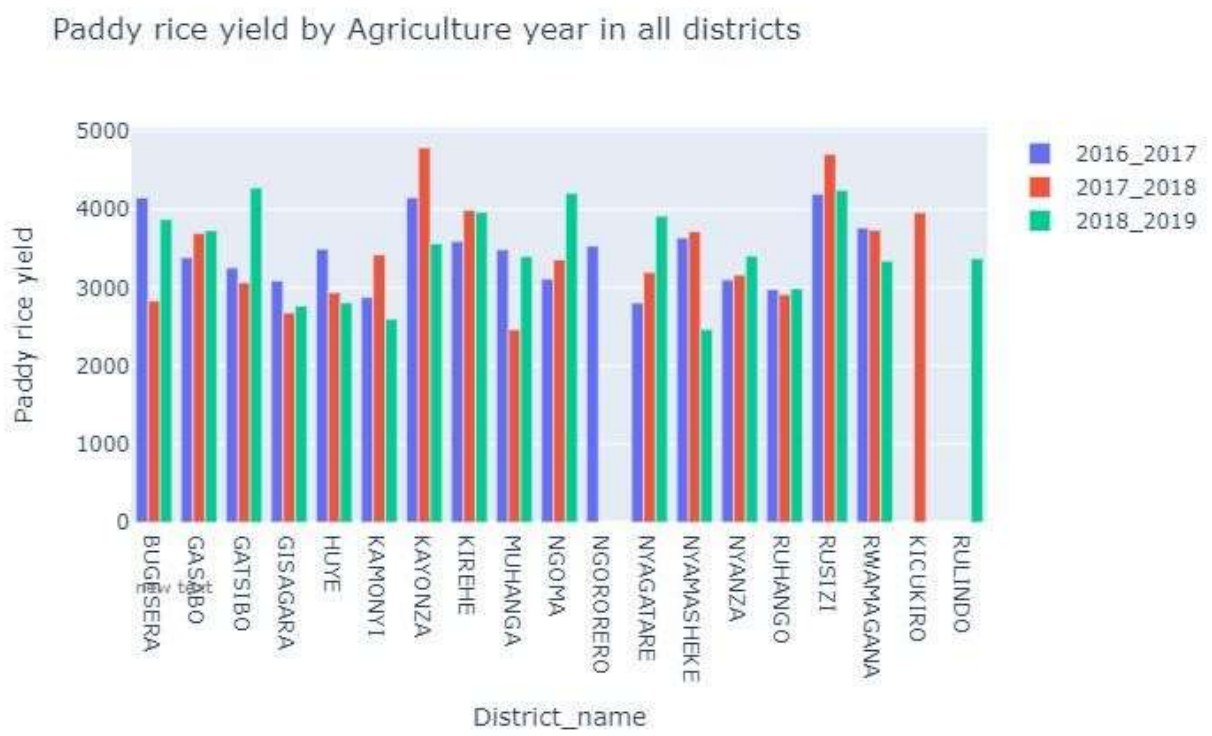

Figure 8: Distribution of Paddy rice yield in all Districts by agriculture years

The above image also indicates that, RULINDO do not have the paddy rice in agriculture year of 2016 to 2017 and 2017 to 2018.It also indicates that KICUKIRO did not have the paddy rice in agriculture year 2016 to 2017 and 2018 to 2019. It also demonstrates that NGORORERO and KICUKIRO do not have the paddy rice in 2018 to 2019 agriculture year. Zero value has been imputed in our data to replace the missing where the yield value does not belong accordingly.

3.5.4.3.2 The Distribution of Paddy rice yield by Season in all districts

The paddy rice crop is one of the main crops which is cultivated in two agriculture season such as Season A and Season B in Rwanda. Is one of the crops which cannot be mixed with the other while you are sowing it. The upcoming image exhibits the yield value in each district by season agricultural in each year.

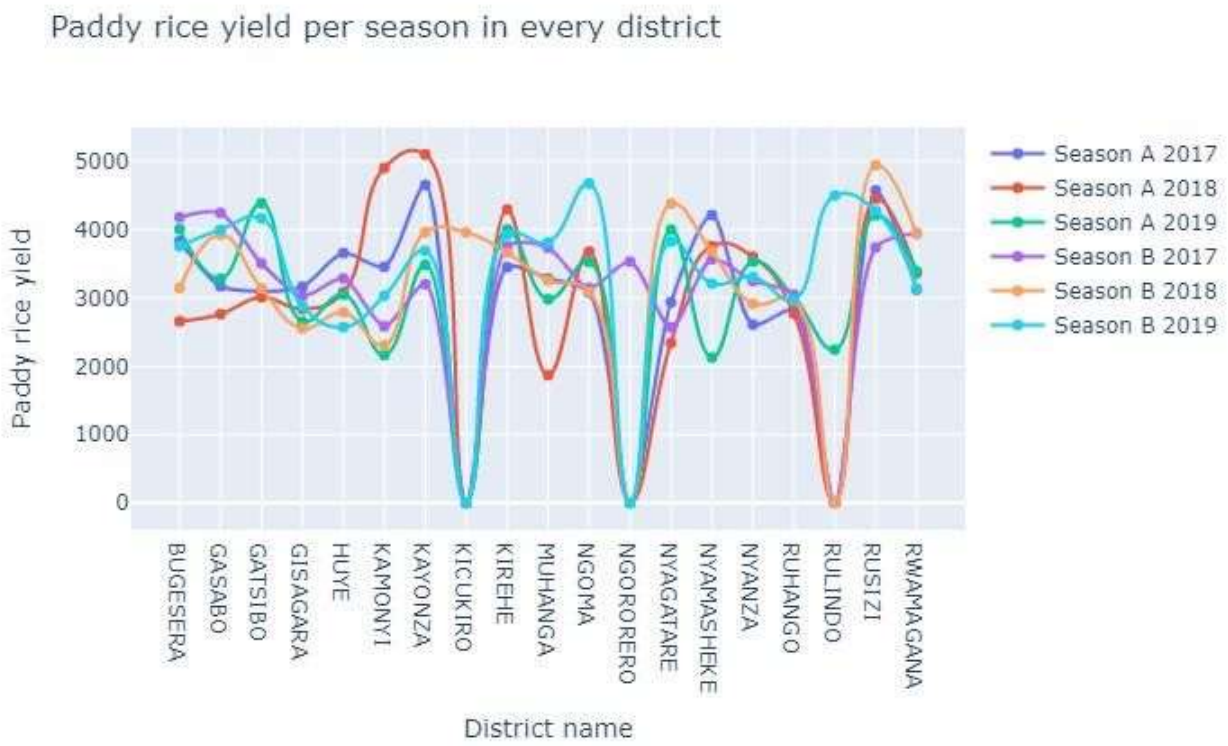

Figure 9: Paddy rice yield in all districts by Seasons

Based on previous image, we should observe that Season A 2018, paddy rice yield exceed five tons per hectare in KAYONZA District and followed by Season B 2018 with 4.9 tons per hectare RUSIZI District. It also demonstrate that some district like KICUKIRO, NGORORERO and RULINDO had not cultivated the paddy rice in every season, for example KICUKIRO contain the yield in season B 2018 but with zero value in other seasons which means that there was not paddy rice according to the season.

3.5.4.3.3. Distribution of Paddy rice yield by country wide based on Seasons countrywide

The paddy rice production can vary according to the season they are planted; the following figure is explaining the value of paddy yield in different season from 2017 to 2019. By computing the following descriptive statistics, we got that season A 2017 produce the highest paddy rice yield and also season A 2019 is one of the season which produced the lowest paddy rice yield. Broadly speaking, the Season B is the one of the season which can produce the highest value of paddy rice yield in Rwanda.

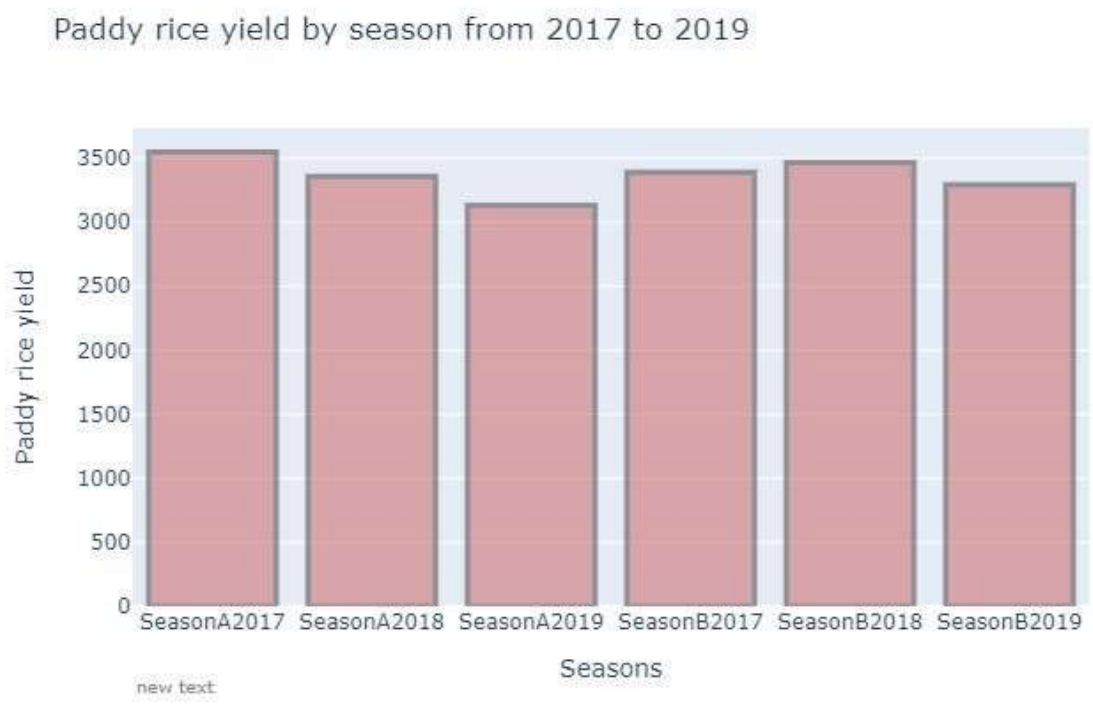

Figure 10: Paddy rice yield per seasons on country wide level

3.5.4.4. The Main statistical observation of Irish potatoes

3.5.4.4.1. The distribution of Irish potatoes' yield by Season in all districts

The Irish potato is the seasonal crop which can be cultivated in different districts of Rwanda, but because of behavior and conditions of this crop, you could not find its production in all districts in every season. For improving the data visualization, the missing value of yield in district by season has been replaced by zero, this means that the district which has not the Irish in specific season it is replaced by zero.

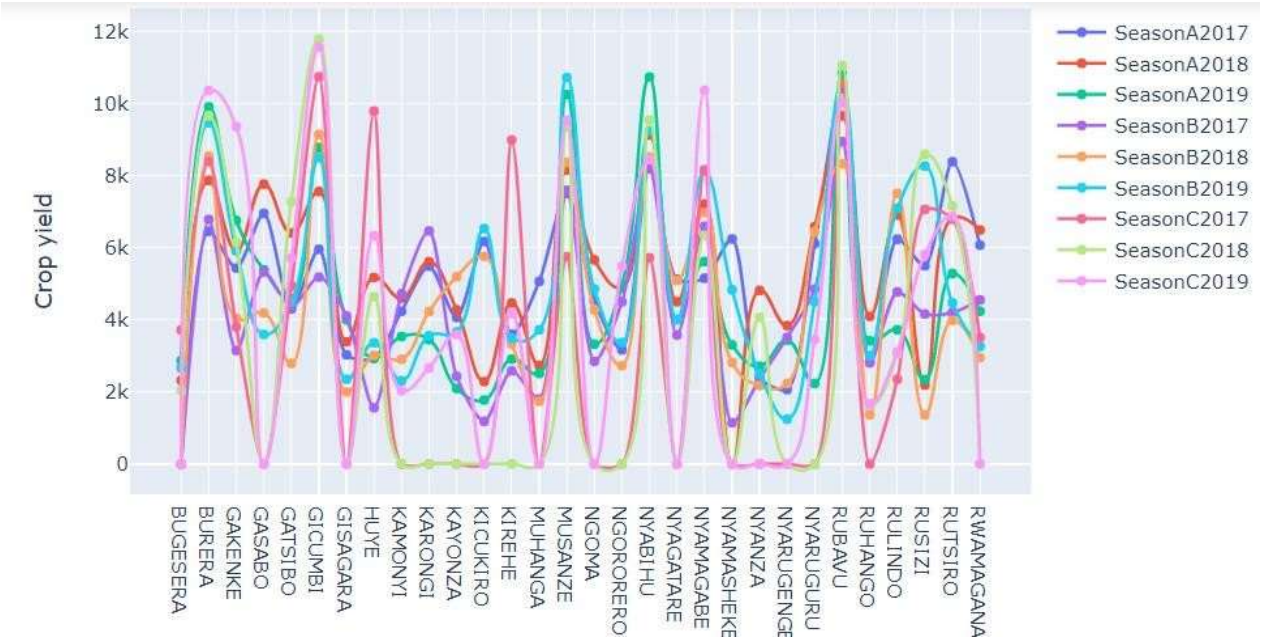

Figure 11: Irish potatoes yield in districts by seasons

By considering the Irish yield in every district by season, we can find that many of districts have zero value in season C that means that the season C is not appropriate on Irish in those districts. Generally, North districts and some of the west district are the one which are always produce the highest production on this crop in all seasons. Here GICUMBI has the highest yield in Season C 2017, Season C 2018 and Season C 2019.

3.5.4.4.2. The comparison of Irish potatoes' yield in all districts by considering the respective agricultural year

The Irish potatoes should not behave in the same way in different years, even the productions cannot be the same in the same district because of different environment conditions, the used inputs and so on. As results of the Irish yield computed in three different consecutive agriculture years, the production per hectares was different as it is shown in the following image.

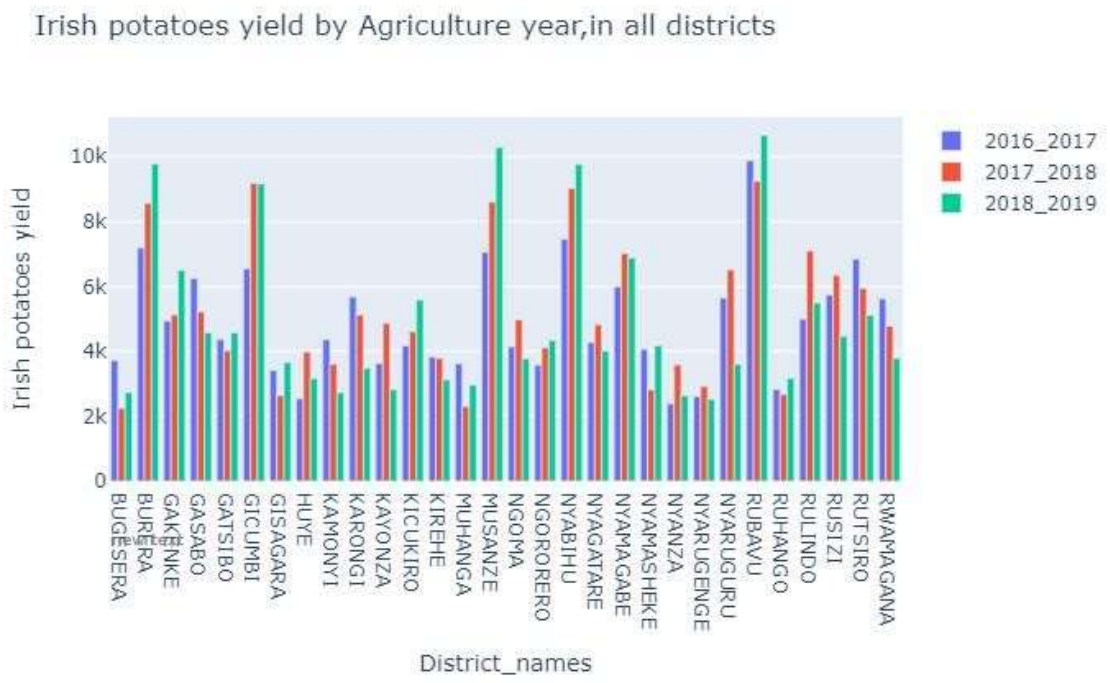

Figure 12: The comparison of Irish potatoes yield in all districts by agriculture years

Fig. 3.13: The comparison of Irish potatoes yield in all districts by agriculture years

By taking into account the results plotted on above image, the agriculture year 2018 to 2019 generated the highest yield in different districts like RUBAVU, MUSANZE, NYABIHU, BURERA and GICUMBI, this can implies that the volcanic regions are the one which can produce the highest value of Irish yield in Rwanda. This Agriculture year has been followed by the agriculture of 2017 to 2018 in productivity.

3.5.4.4.3. Distribution of Irish potatoes' yield by cropping system

The production of crops can depend on different factors; one of them is cropping system. In Rwanda the way the crops are planted may decrease or increase the productions. The average yield per cropping system was shown below on the following image.

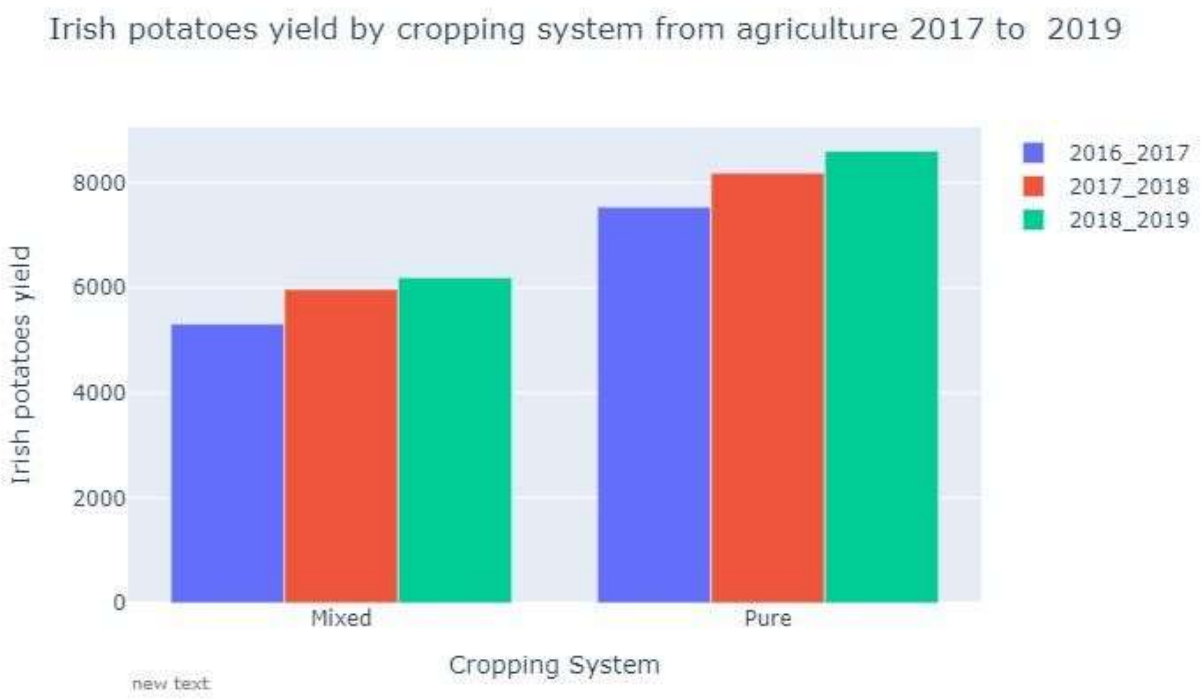

Figure 13: Irish potatoes yield in last three Agriculture years by cropping system

The preceding figure is demonstrating how the Irish yield has been produced based on cropping system in last three consecutive years. Pure cropping system is the one which could produce the highest yield in all Agriculture based on cropping system.

3.5.4.4.4. Comparison of Irish potatoes' yield by Agriculture Seasons countrywide

The agriculture year in Rwanda is comprised by three seasons, with the behavior and the type of crops you are planting, season can influence the production. The upcoming image shows the average yield by considering the seasons.

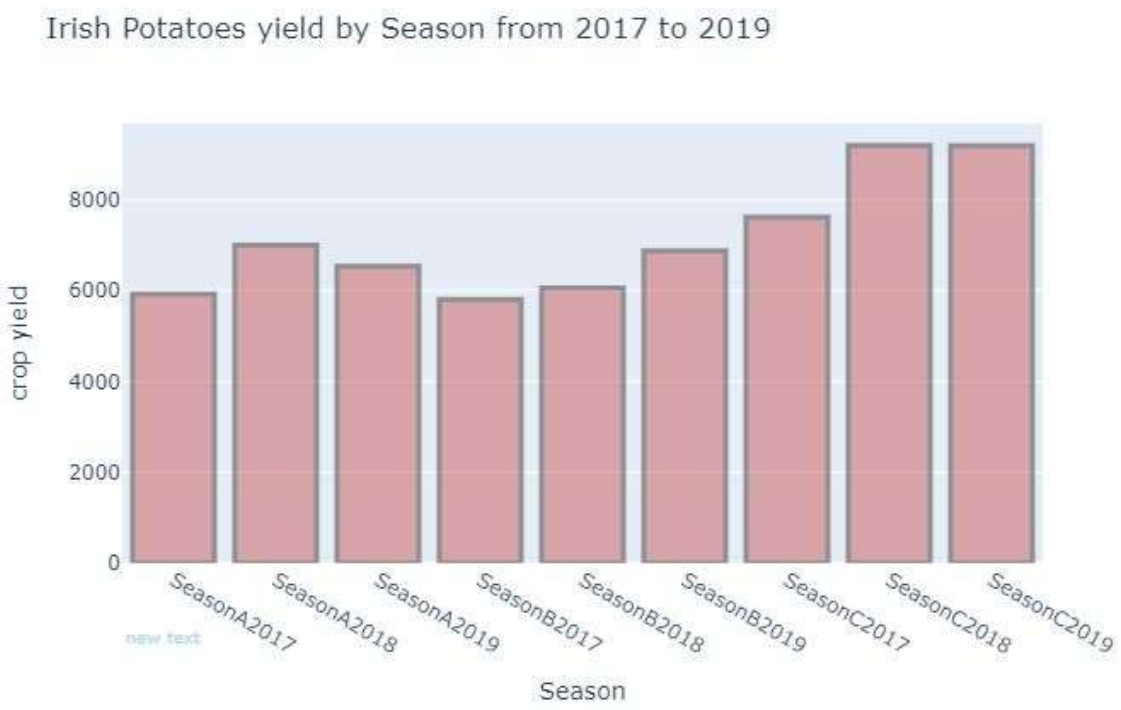

Figure 14: Comparison of Irish potatoes' yield per seasons

In accordance with the above image, different seasons produced several average yields. The comparison of Irish yield by season in last three consecutive agricultural years demonstrates that the Season C 2018 and Season C 2019 generated the highest production per hectares.

3.5.4.5. The Main statistical observation of Maize

3.5.4.5.1. Comparative Analysis of Maize yield in all districts by taking into account the respective agricultural years

Maize is one of the crops which is cultivated in all district of Rwanda, is one of the crop that the government of Rwanda is mobilizing to the farmer for ensuring the food security to his or her population. So all districts cannot produce the same production according to their soil fertility and environment factors of them, the following schemas explain the production of maize in all districts in consideration of agriculture years of 2017 to 2019 respectively

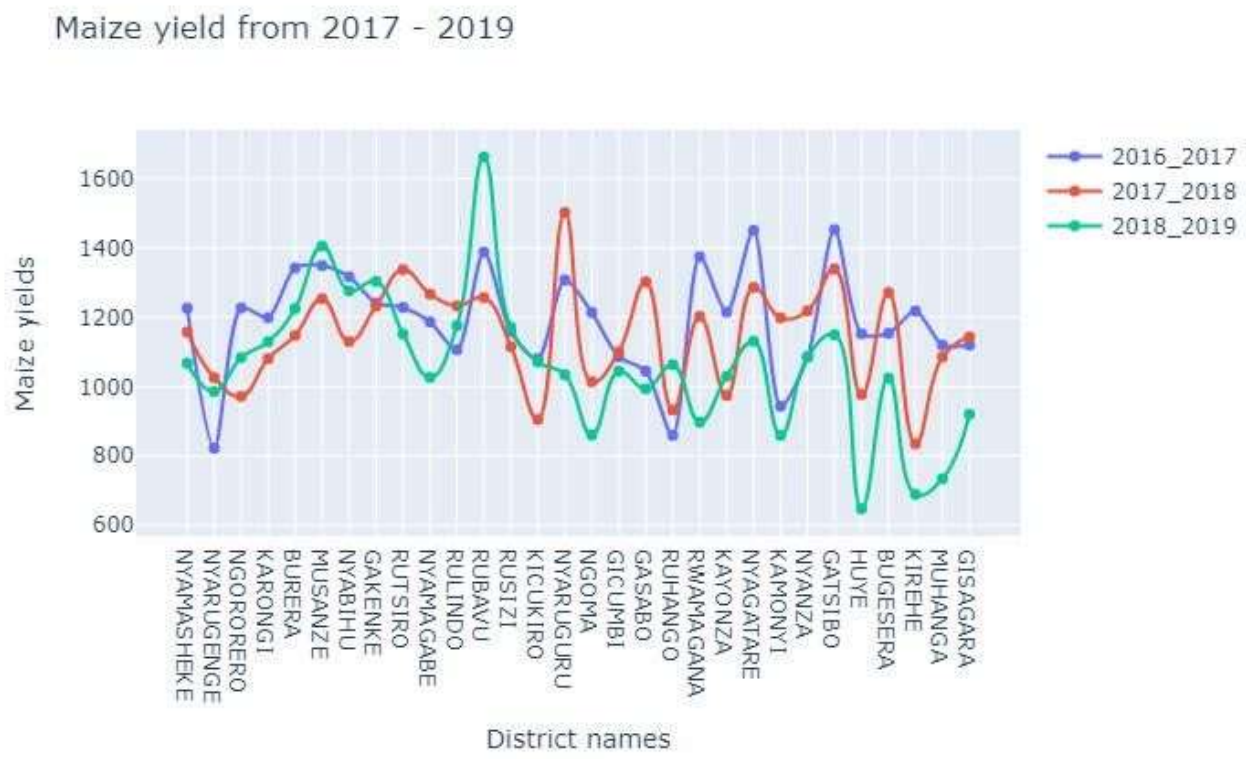

Figure 15: Maize yield in all districts by Agricultural year

The General observation on previous figure, RUBAVU district is the district which produced the highest production per hectares in 2018 to 2019 agricultural year. On the other hand, the agriculture year of 2016 to 2017 had the highest yield average and followed by 2018 to 2019 agriculture year.

3.5.4.5.2. The comparison of Maize production per seasons in different districts

The production of Maize crops can sometime depend on agriculture seasons; the maize is one the main crop which is not considered as season C crop in Rwanda because of environmental factors of our country. So the forthcoming figure show the productivity of maize per hectare in every district by taking into account the seasons.

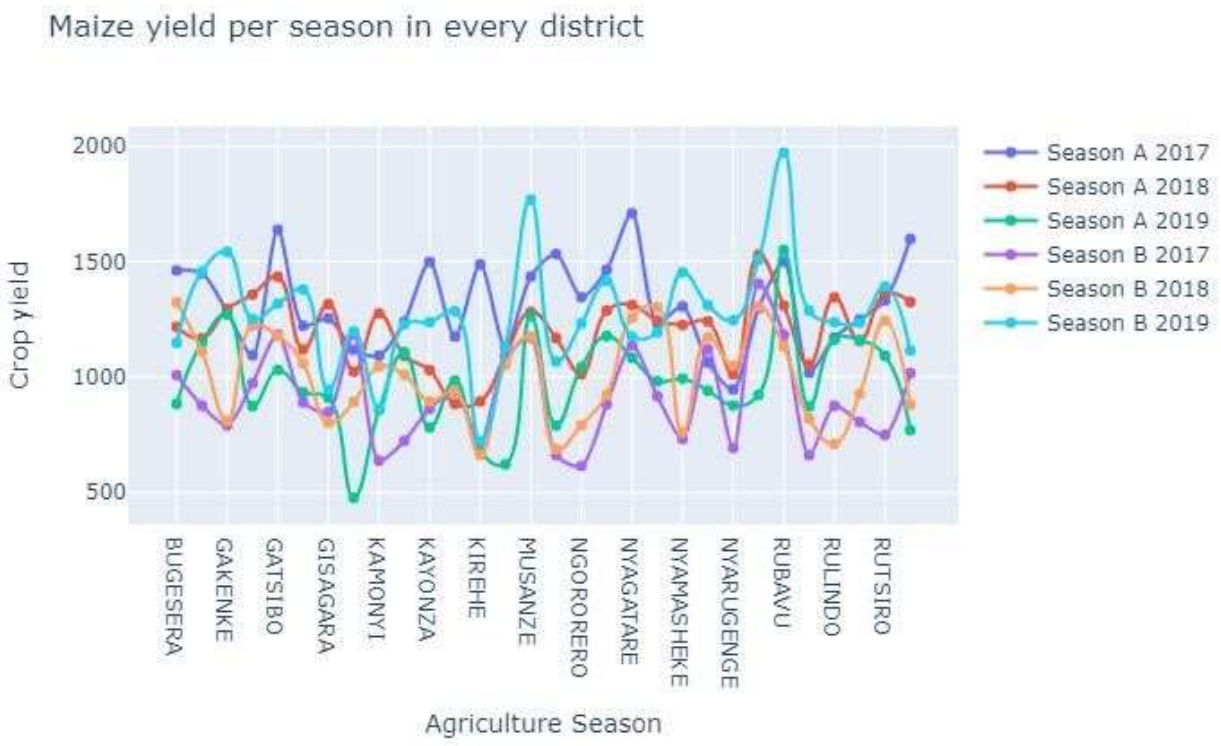

Figure 16: Maize yield in districts by seasons

By examining the above figure, we should observe that season B 2019 in RUBAVU district and in MUSANZE district have been produced the highest production of maize per hectares. We can also notice that the Season B 2017 production was low in many districts.

3.5.4.5.3. The comparison of Maize yield by cropping system in different districts

The cropping system is one of the farming factors which can increase or decrease the production of maize, in following figure; the three consecutive are plotted by considering cropping system.

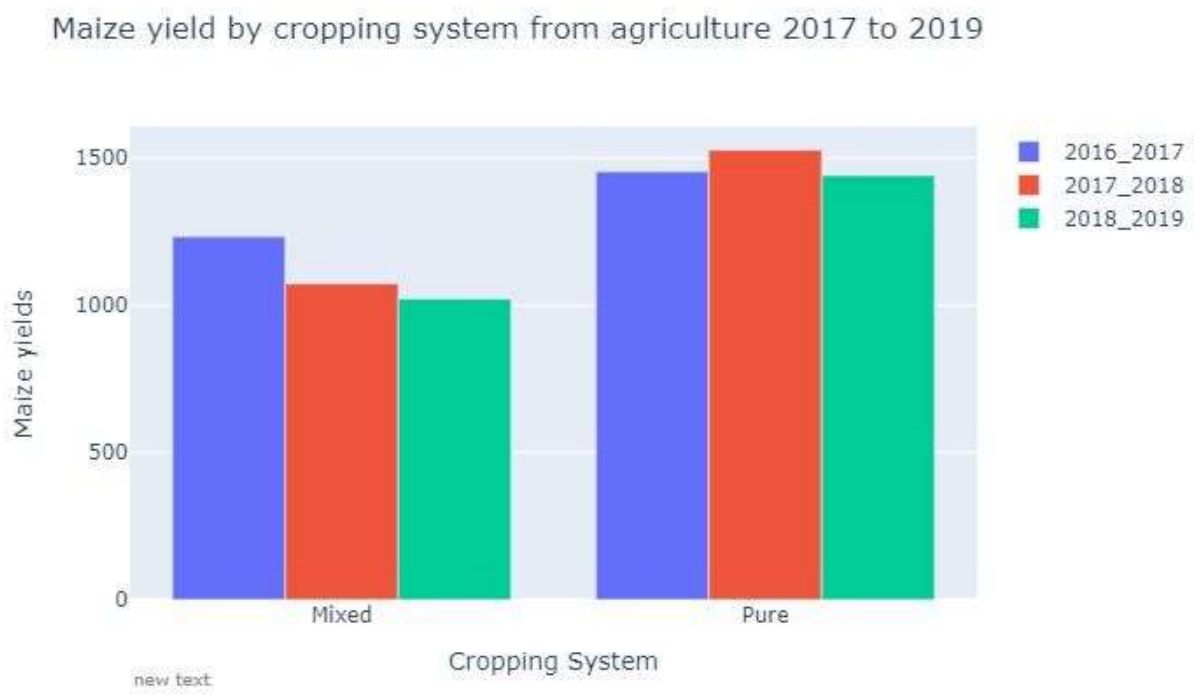

Figure 17: Maize yield by cropping system in last three agriculture years

Fig. 3.19: Maize yield by cropping system in last Three Agriculture years

The previous figure, show that when you plant the maize purely in plot, you should obtain or increase the production per hectares as it has been proved by the last there consecutive agriculture years. It also demonstrates that 2018 to 2019 agriculture year had the best yield of maize in case cropping system is pure.

3.6 Predictive models

The main objective of predictive model is to assist decision makers in forming right decision by making them efficient and effective. In case of building predictive model, we should have to use different data mining techniques; in this study, I manage to use the regression model such Linear, Multiple Linear Regression, Polynomial Linear regression and Artificial Neural Network models were employed for crop yield prediction

3.6.1. Artificial Neural Network for linear regression

3.6.1.1. Brief introduction on ANN

Artificial neural network are a family of models effective at solving problem of function approximation and pattern recognition. Neural network are composed of multiple simple

Computational of blocks called neurons. Simply, a neuron receives an input signal and then computes an output, every input should have a weight associated with it, and the larger of the weight the more impact on the corresponding input channel has on the output. Also a neuron has a bias which could be considered as an addition input to the neuron, x_0, that is equal to 1 and has a weight identical to the value of bias, $w_0 = b$. Additionally, a neurons has a transfer or activation function that can define the type of neurons

3.6.1.2. ANN description in our study

Artificial Neural Network (ANN) is referred to as nonlinear statistical data models that replicate the role of biological ANNs (A.K. Jain, J. Mao, 1996). Statistical pattern approach has been the most commonly studied utilizes in practices (C.Shang,et al., 2017; J.K. Basu, D. Bhattacharyya, 2010). ANN is increasingly attractive, effective, and successful in achieving pattern recognition (PR) in many problems (R. Bala, 2017; S. Knerr, L. Personnaz, 1992). Neural Networks are composed of multiple simple computational of blocks called neurons. Simply, a neuron receives an input signal and then computes an output, every input should have a weight associated with it, and the larger of the weight the more impact on the corresponding input channel has on the output. Also a neuron has a bias which could be considered as an addition input to the neuron, Xo, that is equal to 1 and has a weight identical to the value of bias, w_0=b (Abiodun et al., 2019).

$$h_1 = g_1(w_1 x - b_1)\cdot \quad (4)$$

$$h_2 = g_2(w_2 x - b_2)\cdot \quad (5)$$

$$h_3 = g_3(w_3 x - b_3)\cdot \quad (6)$$

$$h_{12} = g_{12}(w_{12} x - b_{12})\cdot \quad (7)$$

$$Y = g_{12}\, w_{12} g_{11}\, w_{11} g_{11}(\ldots..w_2 g_1(w_1 x)\ldots..+b_1 + b_2 + \cdots\ldots.+b_1 + b_0)\cdot \quad (8)$$

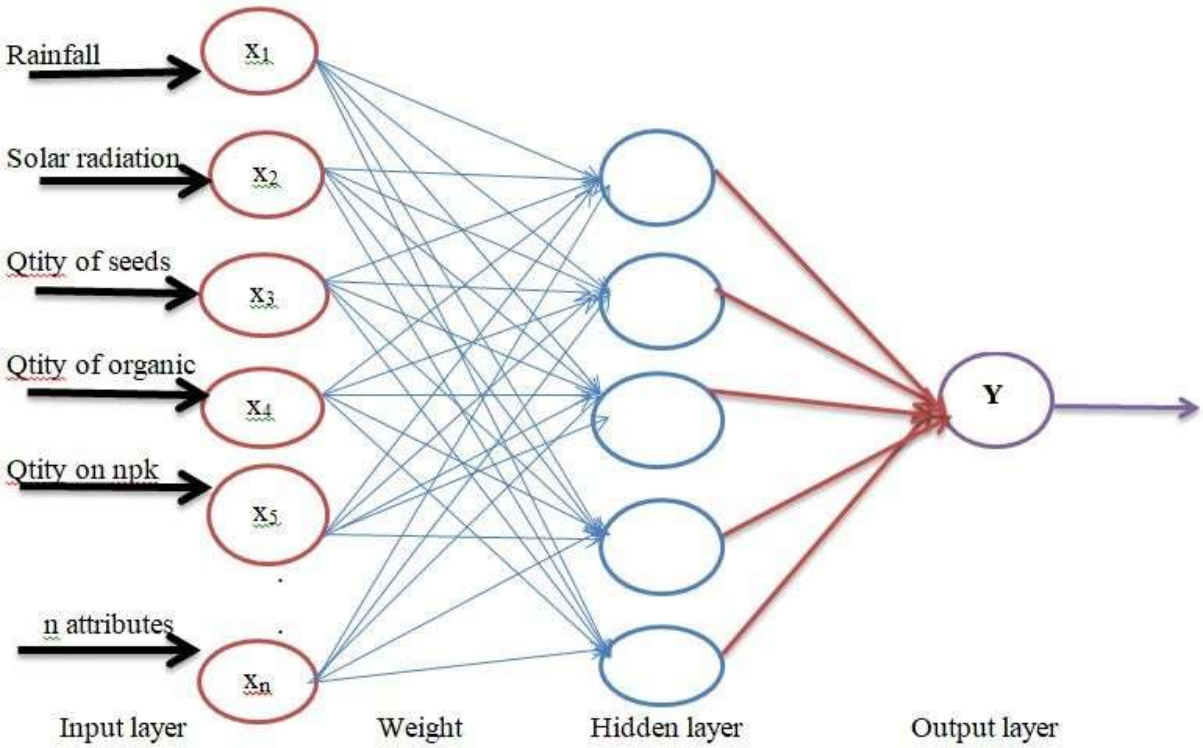

Figure 18: ANN Image base on our attributes

The following diagram shows the ANN Prediction process

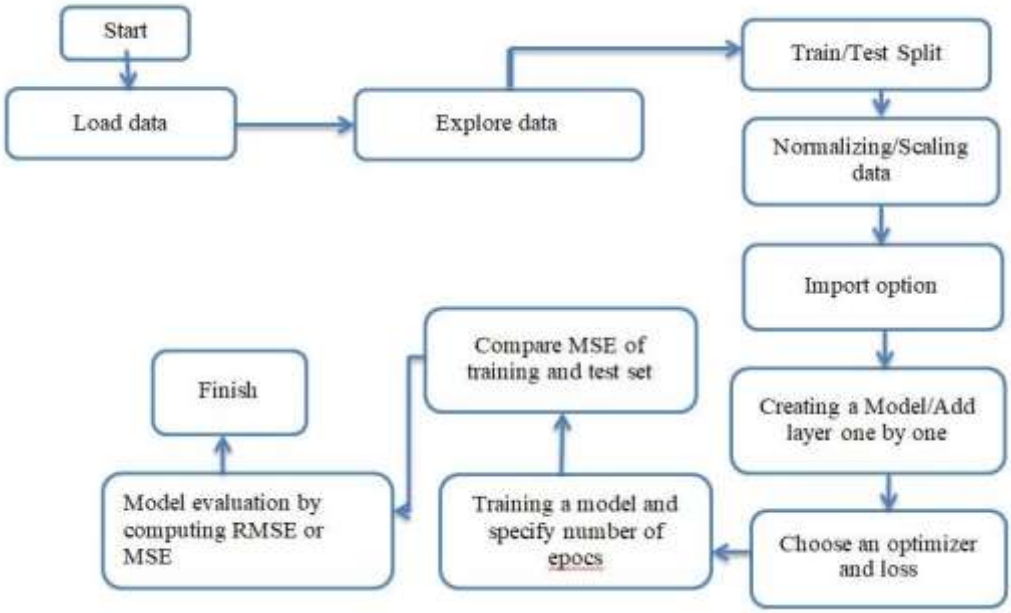

Figure 19: ANN model prediction process

3.6.1.3 Normalization or data scaling

Data Normalization is the method of rescaling the data from the original range so that all values are within 0 and 1. The value should be normalized by using the following formula:

$$Y = \frac{X - min}{max - min} \quad (9)$$

In our case, I normalized my data by using the scikit-learn object called MinMaxScaler in python programming language. This technique has been used in the following steps:

- Fitting the scaler by using the available training data: For data scaling this means that the training data are used to estimate the minimum and the maximum observable value. This is computed by calling fit() function in Python programming language
- Applying the scaler to training data: This signify that you can use the normalized data to train your model, this is done by calling transform() function in Python programming language
- Applying the scale data on going forward: This means that you prepare the new data in the future for making predictions.

Standardizing the input data imply rescaling the distribution of values so that the mean of observed value become zero and the standard deviation becomes 1.

3.6.2. Polynomial Linear Regressions

The polynomial linear regression is one of linear regression model in which the relationship between predictors x and predicted variable y is modeled as an nth degree polynomial. It fits a nonlinear relationship value x and corresponding conditional mean of y. The construction of that model could be done in the following steps:

(i) Importing the needed libraries
(ii) Separating the predictors and predicted variables
(iii) Fitting polynomial regression to data and choose the number of degree
(iv) Predicting the new results with both linear and polynomial regressions

3.6.3. Multiple Linear Regressions Model

Multiple Linear Regression model is used to model the linear relationship between several explanatory variables and a response variable. It consists of more than just fitting a line through data points. It comprises three stages that are:

- Analyzing the correlation and directionality of data
- Model estimation and fitting the lines

- Validity and usefulness evaluation of model

3.6.4. Features selection by using Principal Component Analysis

3.6.4.1. Overview of PCA

After data cleaning, data exploration and descriptive analysis; features selection stage was so important for choosing the best predictors to use to our model for ensuring accuracy and goodness of fi. So, the PCA is considered as an ordination and dimensionality reduction techniques which is used in ecological data analysis. It is the one which can transform the numerical predictors into a set of uncorrelated variables developed as a linear combination of predictors which are known as principal components for explaining the maximum variation in the data. It is also a technique which can be used for decreasing dimension, higher dimensional to lower dimension data, this can be taken as Normalization of linear combination of predictors.

In this method, principal component is firstly intended as the linear combination of original predictors which can apprehend the maximum variance in the data set. It can identify the direction of highest variability in the data. This option can minimize the sum of square distance between a data point and lines. The other principal components capture the remaining variance and uncorrelated to the first PC and they should be orthogonal.

3.6.4.2. The steps of PCA computations

The following steps are used to implement dimensionality reduction by using PCA (i) Standardization of data

(ii) Computing the covariance matrix

(iii) Computing the eigenvectors and eigenvalue

(iv)Computing the principal component

(v) Reducing dimension of dataset

3.7. The regression Metrics

After construction of every regression model, you should have to evaluate for examining its efficiency. In my project, I managed to use several scikit-learn metrics module to implement various scores, loss and utility functions to assess the regression performance.

3.7.1. Mean Absolute Error (MAE)

The Mean Absolute Error is a risk metric corresponding to expected value of the absolute error loss. If we denote by y_i, the predicted value of the i-th sample and y_i the corresponding true value, then the mean absolute value can be estimated from n samples by

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (10)$$

3.7.2. Mean Square Error (MSE) and Root Mean Square Error (RMSE)

The sciklearn library of Python programming language contains the mean square error function which can be used to compute the mean square error, this is the risk metrics which is corresponding to the expected value of squared (quadratics) error of loss. The metrics is correspondent to the following formula

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \ddot{y}_i)^2 \quad (11)$$

One the other hand, the RMSE is used to assess the performance of in model evaluation studies as does the MAE. However, the RMSE is more appropriate to represent model per- formance than the MAE when the error distribution is expected to be normally distributed.

It is defined as the square root of MSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \tilde{x}_i)^2}{n}} \quad (12)$$

3.7.3. Coefficient of Determination R^2 score

The scikit-learn environment with r2 score function computes the coefficient of determination, which is denoted as R^2, it optimize the proportion of variance of dependent variable thathas been explained by the independent variables in the model. It indicates how the model is good and how it fits the data, it is also a measure of how well unseen samples are likely to be forecasted by the model, through the proportion of explained variance.

Therefore, the variance is dataset dependent; R^2 cannot be meaningfully comparable across different datasets.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \text{ or } SST = SSR + SSE \quad (13)$$

Whereas, i=variable i, n: number of non-missing data points, yi: observed values, $\hat{y}_i$=predicted values, y_i=observed values and $\bar{y}_i$=estimated values.

Now, we define the Coefficient of determination by

$$R^2 = \frac{SSR}{SST} \quad (14)$$

Whereas R^2: Coefficient of determination, RSS: Sum of squares of residuals and TSS: Total Sum of Squares. Best possible score is 1.0 and it can be negative as the model can be arbitrarily worse. A constant model that always predicts the expected value of y, ignoring the input features, would get a R^2 score of 0.0. We should notice that the scikit-learn r2 score function calculate unadjusted R^2 without correcting for bias in sample variance of y.

4. Findings discussion and results

4.1. General observation of bush bean& paddy rice

The algorithm of predicting the crop yield was developed by using Python programing language; various python library like numpy, pandas, matplotlib, plotly, searborn, scikit-learn and so on were used to finalize this study. Applying descriptive statistics on agricultural dataset over the period of 2017-2019 resulted in different output printed out. By analyzing crop-by-crop, there were founds a higher diversification of national agricultural yield for different agricultural commodities focused on this study.

The figure3 shows that, the bush bean yield was extremely efficient in agriculture year of 2016 to 2017 compared with 2017 to 2018 and 2018 to 2019 agriculture years. Among all Districts of Rwanda, Musanze had higher bush bean yield (8.4t/ha in 2017, 10.2t/ha in 2018 and 10.2t/ha in 2019) while Muhanga had a low yield (2.3t/ha in 2017, 1.5t/ha in 2018 and 2.1t/ha in 2019) for the whole period of study. The results have been approved that the bush bean crop yield did not much more depend on areas size used for agricultural production. In average, the agriculture season C for all years presented the higher yields in all Districts of Rwanda. The paddy rice is one of the main crops planted in at least all region of Rwanda, except the in the Northern Province. It had the consistent yield as it is one and only one crop produced as pure crop in all plots of study. The paddy rice yield was higher in all agriculture year; especially for agricultural season A of 2018-2019. The yield of paddy rice estimated to be higher in Eastern province, as major Districts in that region cover the yield which was greater than 5 tones/ha followed by South-West region namely RUSIZI District with an average yield of 4.9tons/ha. Some districts, especially those located in central regions (KICUKIRO [2t/ha], NGORORERO [1.76t/ha] and RULINDO [3.35t/ha]) predicted low average seasonal yield, and majority of them produced the paddy rice in one season per year.

4.2. General observation of Irish potatoes& Maize

The Irish potatoes are produced in all districts of the country, but with the higher dominance in the Northern and other higher mountains zone across the country. The Irish potatoes and Maize are between the 2 important crops, which are more focused more than others in the program of the Ministry of Agriculture of Rwanda called Crop Intensification Program (CIP). In the past 10 years the areas used for the maize and Irish potatoes production have been extended in considerable manner, and yields have been increased considerably. The higher Irish potatoes yield was higher in higher mountains zone, especially in RUBAVU [30.2t/ha], BURERA [25.9t/ha], NYABIHU [26t/ha], MUSANZE [25.8t/ha], and GICUMBI [26.4t/ha],

especially in Agriculture year of 2017 to 2018 and 2018 to 2019 while seemed to be very low in Southern province, especially in MUHANGA [5.9t/ha], NYANZA [7t/ha], GISAGARA [6.3t/ha], and RUHANGO [7t/ha] for all the period.

Since 2017, the Rwandan farmers had been mobilized to make an effort in maize production through the program of land use consolidation and crop intensification for ensuring the food security among the Rwanda people. The maize yield was very higher in 2016-2017 and at least in all districts of Rwanda, but it intended to be higher diversified in 2018-2019 whereas the higher yield estimated to be 1.5t/ha in Rubavu District while the lower value counted to be 0.86t/ha in Huye district.

4.2.1. ANN regression through crop yield prediction by using Environmental factors

The model has been developed for assessing the effectiveness of Artificial Neural Network model on crop yield prediction for typical environment factors mainly the rainfall and solar radiation. The evaluation metrics depicting the performance of the ANN model to determine the relationships between climatic factors and crops yields of different commodities have been used in this study. The metrics like MAE, MSE and RMSE used to provide the results that may allow us to conclude that the ANN was the best model that could be used to predict the crops yields. The results clearly show that maize, Irish potatoes, paddy rice and bush beans production have the strong correlation with climatic factors. However, some of those factors were not statistically significant at 5% level, the combination of all agricultural seasons from 2016 up 2019 have produced the good results. As indicated in the Table 1, these results are very stable.

Table 1: The ANN regression metrics through the consideration of environmental factors

Artificial Neural Network Model				
Regression Metrics	**Maize**	**Irish potatoes**	**Paddy Rice**	**Bush Beans**
MAE	0.20	0.60	16.89	0.29
MSE	4.40	61.48	5884.03	57.78
RMSE	2.10	7.84	76.71	7.60
R^2	0.83	0.71	0.78	0.57

Source: Authors Analysis, 2020

As described in Table 1, the ANN model provides better prediction results for all crops especially the Maize as it computed lowest mean value of MAE and RMSE. These evaluation metrics were within acceptance range of R^2=0.83 for maize with MAE=0.20 and RMSE=2.1, R^2= 0.71 for Irish potatoes with MAE=0.60 and RMSE=7.84, R^2=0.78 for paddy rice with MAE=16.86 and RMSE=76.71 and R^2=0.57 for bush beans with MAE=0.29. In contrast the results provided by R^2 for bush bean was very small compared to the results values of other crops. Therefore, in terms of evaluation metrics, MAE has performed very well compared to other error measurements. However, other metrics have also performed better well with higher acceptable

coefficient of determination value. The same research as conducted by Amaratunga et al., 2020 in Sri Lanka have carried out the prediction models using neural networks to determine the relationships between climatic factors and rice production. Their prediction models for environmental factors, especially minimum and maximum temperature, average rainfall, humidity, climate, weather, types of land, types chemical fertilizer, types of soil, soil structure, soil composition, soil moisture, soil consistency, soil consistency, and soil texture on paddy rice yield were within lower mean square error values (Amaratunga et al., 2020).

4.3. The regression metrics values of Maize, Irish Potatoes, paddy rice, and Bush beans.

The dimensionality reduction has been performed using the Principal components Analysis in order to improve the model precision whereas the number of component has been considered according to the crop but mostly four number of component produced better results for all crops. Three types of regression models have been used to predict the crop yields. The various regression metrics have been performed for model evaluation on the same units for comparing model performance. The evaluation results of 3 types of regression models namely Artificial Neural Network Model (ANN), Polynomial Linear Regression (PLR), and Multiple Linear Regression (MLR) have been used for ensuring study efficiency and comparing model results in order to determine the best model that can be used to predict the crop yield in Rwanda.

Table 2: The regression metrics of various models for major crops

Maize crop			
Regression Metrics	**Artificial Neural Network Model**	**Polynomial Linear Regression**	**Multiple Linear Regression**
MAE	0.05	0.29	0.42
MSE	0.29	1.26	6.80
RMSE	0.54	1.12	2.61
R^2	0.99	0.96	0.74
Irish Potatoes crop			
MAE	0.54	0.40	0.92
MSE	32.84	3.30	35.39
RMSE	5.73	1.82	5.95
R^2	0.84	0.95	0.83
Paddy Rice			
MAE	2.80	3.45	3.62
MSE	158.10	234.18	143.24
RMSE	12.57	15.30	11.97
R^2	0.99	0.99	0.99
Bush Beans			
MAE	0.15	0.11	0.18
MSE	15.76	3.47	14.64
RMSE	3.97	1.86	3.83
R^2	0.88	0.95	0.89

Source: Authors Analysis, 2020

The variables, which were correlated and also which were statistically significant to crop yield were taken into consideration as predictors, the following variables were the one which were taken as predictors to our models precipitation, solar radiation, elevation, total quantity of inorganic fertilizers, total quantity of seeds, types of seeds sown, quantity of organic fertilizers used, quantity of pesticides used, number of crops produced, cropping system and use of irrigation. They have been used to predict the yield of maize, paddy rice, Irish potatoes and bush beans. After model construction, the results showed that some variables have the highly positive coefficients, which means that, while the value of predictors increase, the predicted value also tends to increase. The explained variance score (R^2) is regression metric, which is used to measure the discrepancy between model and actual data. So the higher percentage of it indicates stronger strength to make the better predictions. The use of regression techniques in agriculture sector has also been used in different region of the world whereas the regression techniques were used to predict the annual crop yields of several crops, mainly maize, beans, wheat, and potatoes, in conjunction with data mining techniques (Khaki, S., & Wang, 2019), this study has been increased suggestions on how to utilize machine learning techniques to predict the annual

crop yields in the Mexico. In Southern America (Mexico), the use of multiple linear regression, M5-Prime regression trees, the perceptron multilayer neural network, SVM and KNN methods (Gonzalez-Sanchez et al., 2014) was justified for that scenario since there were 10 different crops including maize, beans, wheat, and potatoes, Its focus was solely on multiple linear regression and the results prove that model was the best method that should be used to predict the maize, beans, wheat, and potatoes yields in South American countries.

The mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) have been evaluated for all crops. The results obtained after evaluating the regression models for maize crop by basing on MAE (0.05), MSE (0.29) and RMSE (0.54) and even also on R^2 (0.99), we found that the Artificial Neural Network regression model had the small value for MAE, MSE and RMSE with an adequate R^2 value comparatively to PLR and MLR. So ANN precise highly the maize yield prediction rather than any other models used in this study and is followed by Polynomial Linear Regression according to the obtained results. The research conducted by Shastry et al. of prediction of crop yield using regression techniques, the different regression techniques such as quadratic, pure-quadratic, interaction and polynomial were used for predicting of wheat, maize, and cotton crops. The pure quadratic model accurately predicts the maize yield better than the remained model (Shastry e`t al., 2017). In contrast, the study conducted by Fashoto et al., on the implementation of machine learning for predicting maize crop yields using multiple linear regression and backward elimination, the researchers have been found that the backward elimination showed the strong correlation between the predictors used in the model. The use of normalization helped to improve the Root Mean Squared Error score, however, the results obtained did not give a clear picture of how well normalization helped since they returned perfect values (Fashoto et al., 2021)

Even the used dataset was well processed and cleaned, and all outliers treated very well as required, we got higher mean value for every estimator used for Paddy rice. We really know that the MSE and RMSE are very higher biased value and we're very sure that all the outliers have been treated as simple possible, so here found all models as they have the highest value of coefficient of determinants with an average R^2=0.99, as the MEA is always more robust to outlier at least ANN has the lowest MAE (2.80), RMSE (12.57) and for MLR the metrics results were MAE (3.62), RMSE (11.97). Ahamed et al., 2015 conducted a research study on applying data mining techniques to extract knowledge from the agricultural data to estimate crop yield for major cereal crops in major districts of Bangladesh using linear regression, k-NN, and Artificial Neural Network techniques. Ahamed et al made the crop yields estimation on Rice-AMON, Rice-AUS, Rice-BORO, Potato, and Wheat. The Root Mean Square (RMSE) have been used to predict the yields of Rice-AMON, Rice-AUS, Rice-BORO, Potato, and Wheat, all models provide better results for different crops but ANN provide better prediction for Rice-AMON, Rice-AUS, Rice-BORO which have low Root Mean Square

(RMSE) than those of others (Ahamed, Mahmood, Hossain, & Kabir, 2015). However, Hossain et al. presented an Artificial Neural Network-based prediction model to analyze the paddy harvest in Bangladesh. They have highlighted the necessity of such study to Bangladesh with ongoing climate change. They were successful in predicting paddy harvest to Bangladesh with an error threshold (M. Hossain, M.Uddin, 2017).

In Rwanda, The Irish potatoes produced abundantly in higher mountains zone and its production differ to the regions, in the analysis of outliers treatment, some districts have been considered as they should produce the different yield, some value could be taken as an outliers for others districts or higher value due to worse estimation performed by the data collectors, this at back could affect our model performance. Therefore, taking account the characteristics of our data, the evaluation of our model on Irish potatoes is based on Mean Absolute Error (MAE) as it is sensitive to higher value and less biased to outliers. So by taking into consideration of all behaviors, the PLR is well predicting the Irish potatoes yield as it has the lowest MAE (0.4), MSE (3.30), and RMSE (1.82); and the highest value of coefficient of determination (R^2=0.95). Bush beans considered as basic food in Rwanda, and at least each farming household's in Rwanda produced once the year the bush beans, in fact it is cultivated in all most districts of Rwanda. All the independent variables used are statistically significant at p-value=0.00 and they were all correlated. As it shown in Table 2, the PLR model is a best model to predict the bush bean yield as it is the one which has the lowest MAE (0.11), MSE (3.47) and RMSE (1.86), and the highest value of coefficient of determination (R^2=0.95) comparatively to ANN and MLR.

The study conducted by Suvidha et al., on yield prediction of crops like wheat; maize and Irish potato was done with huge of data divided into training and testing sets. The algorithm like Random Forest(RF) was compared with Multiple Linear Regression (MLR), Author found that RF performed better than MLR(Lavanya & Parameswari, 2020). Machine learning algorithms like Multiple Linear Regression, Random forest Regression and Multivariate Adaptive Regress-ion splines were used for predicting the yield for chosen crops, MLR gave good prediction(Suvidha J. al, 2018). The same study conducted in Mexico by Gonzalez-Sanchez et al. (2014), several factor were chosen to be used as predictor attributes for the study (planting area, irrigation water depth, solar radiation, rainfall, maximum, average and Minimum temperatures), however, other factors which could prove to be significant in Eswatini were lacking. The authors further proceeded to use multiple linear regression, regression trees, and artificial neural network, support vector regression in comparison to finding the method that most accurately predict yield for all crops were considered.

5. Conclusion and Recommendations

5.1. Conclusion

The main purpose of this study is to assess the effectiveness of Artificial Neural Network (ANN) on crop yield prediction for typical environment factors, to compare the effectiveness of Polynomial Linear Regression (PLR), and Multiple Linear Regression (MLR) with ANN model in order to assess the impact of various agriculture inputs and environmental factors on crop yield in Rwanda. Regression analysis was used as a predictive modeling technique for predicting crop yield production. Various regression models have been developed for predicting the main crop yield in Rwanda. The result of correlation coefficients depicting the performance of the ANN trained to determine the relationships between climatic factors and crops yields of different commodities using the four different evaluation metrics (MAE, MSE, RMSE and R^2) providing the results that allow us to conclude that the ANN and PLR are the best model that could be used for predicting the crops yields.

Generally, through the evaluation of the regression between the agriculture inputs and different environmental factors within the crops yields, the results have been proven that the ANN model is good enough to predict the Irish potatoes and Bush beans yields; furthermore, the polynomial linear regression (PLR) model is more effective to predict the paddy rice and Maize yields. The correlation coefficients between the agriculture inputs (quantity of organic and inorganic fertilizers used, quantity of seeds sown, pesticides) and using of agriculture technologies providing the positive association for all crops which means that, the increase of one unit of agriculture input increase the yield; and the results showed that, their correlation was statistically significant as their p-value are less than 5% (p-value<0.05). So, it is clear that the agriculture inputs are effectively affect positively the crop yield in Rwanda.

The results of Artificial Neural Network Model (ANN), Polynomial Linear Regression (PLR), and Multiple Linear Regression (MLR) showed that the regression analysis could be used to predict the production of maize, Irish potatoes, paddy rice, and bush beans with precision. Accurate predictions of these parameters would result in accurate production predict in the future. Hence Data mining techniques will be used for decision-making in the agriculture data-mining sector. The National Institute of Statistics of Rwanda (NISR) should use the data mining techniques to predict the future crop yield production for decision making in the agriculture sector. This will help to reduce the seasonal money spent for agricultural data collection every year.

5.2. Recommendations

In our research we have found that accurate prediction of Maize, Paddy rice, Irish potatoes and Bush beans of crops yields across the country could help a lot of farmers, for researchers, policy makers and others alike. As the Rwandan farmers predict to produce for market oriented, the farmer could produce different crops in different districts of Rwanda based on simple predictions made by this study and if they take much effort on that, each and every farmer would get opportunity at increasing their production which will lead back the increasing of their profits, and increasing the country overall income. Eventually this study exhibits a combination of academic contributions, for researchers, farmers, and policy makers. Together, the policy makers and others stakeholders in agriculture should design an effective model that could allow them to predict the agricultural production based on the historical data. The government's institutions which are involved in agriculture must provide enough required agriculture inputs to the farmers in order to allow the farmers to increase the yields and ensuring the food security among the population of Rwanda.

References

[Online]. (2020). *https://openweather.co.uk/blog/post/influence-temperature-plant-productivity-agriculture-accumulated-temperature*. Assessed on 04/224/23. https://openweather.co.uk/blog/post/influence-temperature-plant-productivity-agriculture-accumulated-temperature

[Online]. (2020a). *https://en.wikipedia.org/wiki/Geography_of_Rwanda*. 4/27. https://en.wikipedia.org/wiki/Geography_of_Rwanda

[Online]. (2020b). *https://en.wikipedia.org/wiki/Humidity*. 4/23. https://en.wikipedia.org/wiki/Humidity

[Online]. (2020c). *https://rdb.rw/investment-opportunities/agriculture/*. 4/6. https://rdb.rw/investment-opportunities/agriculture/

[Online]. (2020d). *https://www.worldbank.org/en/results/2013/01/23/agricultural-development-in-rwanda*. Assessed on 04/22. https://www.worldbank.org/en/results/2013/01/23/agricultural-development-in-rwanda

A.K. Jain, J. Mao, & K. M. M. (1996). "Artificial Neural Networks: A tutorial,." *Computer*, *3*, 31–34.

Abello J, P. P. (2002). *Handbook of massive data sets*.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., Arshad, H., Kazaure, A. A., Gana, U., & Kiru, M. U. (2019). Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition. *IEEE Access*, *7*(February 2017), 158820–158846. https://doi.org/10.1109/ACCESS.2019.2945545

Ahamed, A. T. M. S., Mahmood, N. T., Hossain, N., & Kabir, M. T. (2015). *Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh*. *December 2017*. https://doi.org/10.1109/SNPD.2015.7176185

Ahamed, A. T. M. S., Mahmood, N. T., Hossain, N., Kabir, M. T., Das, K., Rahman, F., & Rahman, R. M. (2015). Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh. *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2015 - Proceedings, June*. https://doi.org/10.1109/SNPD.2015.7176185

Amaratunga, V., Wickramasinghe, L., Perera, A., Jayasinghe, J., Rathnayake, U., & Zhou, J. G. (2020). Artificial Neural Network to Estimate the Paddy Yield Prediction Using Climatic Data. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/8627824

C.Shang, A. Palmer, J. Sun, K.S. Chen, J.Lu, & J. B. (2017). "VIGAN: Missing view imputation with generative adversarial networks" In 2017 IEEE International Conference on Big Data. *Big Data*,

766–775.

Chaochong, J. (2008). *Forecasting Agricultural Production via Generalized Regression Neural Network.IEEE*.

Fashoto, S., Mbunge, E., & Opeyemi, O. G. (2021). IMPLEMENTATION OF MACHINE LEARNING FOR PREDICTING MAIZE CROP YIELDS USING MULTIPLE LINEAR REGRESSION AND BACKWARD IMPLEMENTATION OF MACHINE LEARNING FOR PREDICTING MAIZE CROP YIELDS USING MULTIPLE. *Journal of Computing, January*. https://doi.org/10.24191/mjoc.v6i1.8822

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). No TitPredictive ability of machine learning methods for massive crop yield predictionle. *Spanish Journal of Agricultural Research. Https://Doi.Org/10.5424/Sjar/2014122-4439*.

Horie, T. (1992). *Yield forecasting. Agricultural Systems*. *40*, 211–236.

J.K. Basu, D. Bhattacharyya, T. H. K. (2010). "Use of artificial neural network in pattern recognition,." *International Journal of Software Engineering and Its Applications.*, *4*(2).

Jaganathan, P., Vinothini, S., & Backialakshmi, P. (2014). A Study of Data Mining Techniques to Agriculture. *International Journal of Research in Information Technology*, *2*(4), 306–313.

Kaur, A., & Singh, H. (2011). *Artificial Neural Networks in Forecasting Minimum Temperature*. *7109*, 101–105.

Khaki, S., & Wang, L. (2019). No TitleCrop yield prediction using deep neural networks. *Plant Science. Https://Doi.Org/10.3389/Fpls.2019.00621*.

Lavanya, M., & Parameswari, R. (2020). A multiple linear regressions model for crop prediction with adam optimizer and neural network Mlraonn. *International Journal of Advanced Computer Science and Applications*, *11*(4), 253–257. https://doi.org/10.14569/IJACSA.2020.0110434

M. Hossain, M.Uddin, and Y. J. (2017). "Predicting rice yield for bangladesh by exploiting weather conditions." *Proceedings of International Conference on Information and Communication Technology Convergence (ICTC), IEEE, Jeju, South Korea.*, 589–594.

Manjula, E., & Djodiltachoumy, S. (2017). A Model for Prediction of Crop Yield. *International Journal of Computational Intelligence and Informatics*, *6*(4), 298–305.

Mikova, K. (2015). Effect of Climate Change on Crop Production in Rwanda. *Earth Sciences*, *4*(3), 120. https://doi.org/10.11648/j.earth.20150403.15

NISR. (2014). Seasonal Agricultural Survey. *National Institute of Statistics of Rwanda*.

NISR. (2019a). *SEASONAL AGRICULTURAL SURVEY, ANNUAL REPORT, RWANDA*.

NISR. (2019b). *Seasonal Agriculture Survey Annual Report,Department of Economics*

Statistics(National Institute of Statistics of Rwanda).

Okori, W. and Obua, J. (2011). Machine Learning classification techniques for famine prediction. In proceedings of the world Congress on Engineering. *In Proceedings of the World Congress on Engineering*, *2*, 6–8.

R. Bala, D. & K. (2017). "Classification Using ANN: A review, "International Journal Of Computational Itelligence Research. *International Journal Of Computational Itelligence Research.*, *13*(7), 1811–1820.

Raghuveer, K. (2014). Data Mining in Agriculture: A Review. *AE International Journal of Multidisciplinary Research*, *2*(9), 1682–1690. https://doi.org/10.1007/s13398-014-0173-7.2

S. Knerr, L. Personnaz, & G. D. (1992). "Handwritten digit recognition by neural networks with single-layer training.''." *IEEE Transactions on Neural Networks*, *3*(6), 962–968.

Satyal, V. R. (2010). Agriculture in. *Economic Journal of Development*, *11*(1), 144–157.

Shastry, A., Sanjay, H. A., & Bhanusree, E. (2017). Prediction of Crop Yield Using Regression Techniques. *International Journal of Soft Computing*, *12*(2), 96–102.

Suvidha Jambekar; Shikha Nema; Zia Saquib. (2018). Prediction of Crop Production in India Using Data Mining Techniques. *IEEE*.

Ye, N. (2013). *Data Mining:Theories, Algorithms,and Examples,CRC Press*.

Applying Data Mining Techniques to predict annual yield of major crops in districts of Rwanda during agricultural seasons 2017-2019

ORIGINALITY REPORT

15%	12%	7%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.periyaruniversity.ac.in Internet Source	3%
2	A. T. M. Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir et al. "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh", 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 Publication	2%
3	www.javatpoint.com Internet Source	2%
4	yeanrwanda.org Internet Source	1%
5	www.wideskills.com	

	Internet Source	1%
6	ijesc.org Internet Source	1%
7	www.fedoa.unina.it Internet Source	1%
8	ijarcce.com Internet Source	1%
9	openweather.co.uk Internet Source	1%
10	Submitted to Swinburne University of Technology Student Paper	1%
11	scikit-learn.org Internet Source	1%
12	Irina S. Zhelavskaya, Yuri Y. Shprits, Maria Spasojevic. "Reconstruction of Plasma Electron Density From Satellite Measurements Via Artificial Neural Networks", Elsevier BV, 2018 Publication	1%
13	Nabaz T Khayyat. "Energy Demand in Industry", Springer Science and Business Media LLC, 2015 Publication	1%
	computer-trading.com	

14 Internet Source

1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		

Applying Data Mining Techniques to predict annual yield of major crops in districts of Rwanda during agricultural seasons 2017-2019

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47

PAGE 48

PAGE 49

PAGE 50

PAGE 51

PAGE 52

PAGE 53

PAGE 54

PAGE 55

PAGE 56

PAGE 57

PAGE 58

PAGE 59

PAGE 60

PAGE 61

PAGE 62

PAGE 63

PAGE 64

PAGE 65

PAGE 66

PAGE 67

PAGE 68