KCA UNIVERSITY

Data Mining Approach for Building Predictive Model For Crop Yield Data

FACULTY OF COMPUTING AND INFORMATION MANAGEMENT

BY

BEATRICE MUTHONI KAGUCIA

15/03149

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MSC DATA ANALYTICS IN THE FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY AT KCA UNIVERSITY

NOVEMBER 2017

DECLARATION

I declare that this project is my original work and has not been previously published or submitted elsewhere for award of a degree. I also declare that this contains no material written or published by other people except where due reference is made and author duly acknowledged.

Student name: **Beatrice Muthoni Kagucia** Registration number: **15/03149**

Sign: ____________________ Date: __________________

I do hereby confirm that I have examined the Project Dissertation of

Beatrice Muthoni Kagucia

And have approved it for examination

Sign: ___________________ Date: ____________________

Dr. Simon Mwendia

Dissertation Supervisor

ABSTRACT

The application of computer science has led to advancements in various sectors of economies including agricultural production, manufacturing and marketing. Computer algorithms have been used for prediction. There has been immense interest and research on crop yield prediction aimed at addressing food security. This has been achieved through the development of various crop models. Some researchers have studied yield prediction by applying computer science solutions. However, critical issues related to agricultural output have not been well addressed. This study looked at issues related to crop production. By using a case study approach and knowledge discovery data mining process this study was preceded by literature review, followed by analysis of daily 1950-2016 meteorological and annual 1950-2016 crop produce data in Njoro Sub-County (Rift Valley, Kenya).

This study aimed to predict the agricultural outcome during pre-harvest season by building predictive models of Artificial Neural Networks (ANNs), Decision Tree (J48) and PART rule via WEKA and evaluating the effectiveness and the accuracy of the built models. Results showed that ANN that was built through MLP function has the highest accuracy followed by Decision tree. The obtained results could help decision makers for achieving food security and the country's productivity for the upcoming years continuously.

Keywords: Data Mining, ANN, WEKA, Pre-harvest, Yield, Yield Prediction

ACKNOWLEDGEMENT

Special thanks to God Almighty for the gift of life and His unending provision and protection.

I would like to recognize the people who assisted me throughout the project, because without them I wouldn't have come this far.

Special thanks go to Mr. and Mrs. Kagucia Peter, my father and mother, notably for their moral support and inspiration that has kept me going all the way.

I express my sincere gratitude to Dr. Simon Mwendia and Mr. Isaac Okola, my supervisors who have given me the guidance to ensure that this work comes to a successful completion.

Rose, James, David, Eunice, Wanini, Shanel, I love you all.

I am very thankful to my family and friends, who have encouraged me and ensured my progress till the very end. I cannot recognize all those people that supported me individually, but I would like them to know that I am really grateful and I love them all.

MAY THE LORD REWARD YOU ABUNDANTLY

TABLE OF CONTENTS

DEDICATION

I dedicate this work to Dad and Mum, and all those who have worked tirelessly to ensure that I achieve the best out of life.

LIST OF TABLES

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

ANN: Artificial Neural Network

CSM: Crop Simulation Model

DAP: Di-Ammonium Phosphate

FAO: Food and Agriculture Organization

FEWSNET: Famine Early Warning System Network

GDP: Gross Domestic Product

GHG: Green House Gases

IPCC: Intergovernmental Panel on Climate Change

KARI: Kenya Agriculture Research institute

KALRO: Kenya Agriculture and Livestock Research Organization

KDD: Knowledge Discovery and Data mining

KFSSG: Kenya Food Security Steering Group

KMD: Kenya Meteorological Department

MAM: March-April-May

MLN: Maize Lethal Neucrosis

MLP: Multi-Layer Perceptron

MoA: Ministry of Agriculture

SALUS: System Administration to Land Use Sustainability

WEKA: Waikato Environment Knowledge

OPERATIONAL DEFINITION OF TERMS

Algorithm: A process or set of rules to be followed in calculations of other problem solving operations especially by a computer

Climate: The long-term manifestations of weather and other atmospheric conditions in a given area or country.

Climate change: Change in average weather conditions or in the context of longer-term average conditions.

Crop: A plant, especially a cereal grown to be harvested as food, livestock fodder or fuel or for any other economic purpose

Data Mining: Analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Food security: A situation in which all people, at all times, have physical, social and economic access to sufficient safe and nuts food which meets their dietary needs and food preference for an active and healthy life

Model: A simplified representation used to explain the workings of a real world system or event

Prediction: A statement about the future. It’s a guess, sometimes based on facts or evidence.

Yield: Also known as agricultural output, refers to the measure of grains or seed generated from a unit of land

CHAPTER ONE: INTRODUCTION

1.0 Background of the study

The agriculture sector is the mainstay of the Kenya's economy. The sector directly contributes 24% of the GDP (Gross Domestic Product) and 27% of GDP indirectly through linkages with manufacturing, distribution and other service related sectors. Approximately 45% of government revenue is derived from agricultural sector and the sector also contributes ove75% of industrial raw materials and more than 50% of the export earnings. (KARI Report, 2008)

In Njoro Sub County, Agriculture is everything. The citizens living in this area depend on agriculture for their livelihood. The main crop in this area is Maize. Njoro is a town in Nakuru County. In the recent past, farmers and people in this area have suffered greatly due to food insecurity. This is due to poor crop harvest which is both the cash and food crop. The poor yield is caused by several factors e.g. climate change (Global warming), floods, drought, unreliable rainfall, soil type, soil acidity (this is caused by overusing DAP fertilizer) and diseases (MLN-Crop Lethal Virus) which adversely affects crop production. In the year 2015, the reduction in crop production was 40%. This was caused mainly by delayed rainfall, which causes delay in planting season hence affecting the whole process of crop production. (Ministry of Agriculture, Njoro Sub-County)

FAO defines Food security as a situation in which all people, at all times, have physical, social and economic access to sufficient safe and nuts food which meets their dietary needs and food preference for an active and healthy life (KARI report, 2008). Food insecurity in this area is an issue. This is because every year, farmers after harvesting the crop, they sell all of, and come April of the coming year, there's shortage of crop, which is the staple food in Njoro.

Data Mining comes in handy to solve the issue of food insecurity by predicting the crop yield using the available historical data. This is so because, previous studies have shown how Data Mining has been applied in various industries e.g. finance to detect fraud.

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. (Hand 2001)

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as *models* or *patterns*. Examples include linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series. (Hand 2001)

Data mining involves different algorithms to do diverse tasks and all these algorithms attempt to fit a model to the data. The algorithms examine the data and determine a model closest to the characteristics of the data being examined. The data mining algorithms can be characterized as consisting of three parts namely; **Model**: The purpose is to fit a model to the data, **Preference**: some criteria must be used to fit one model over another, **Search**: a technique to search the data.

1.1 Data Mining Techniques

The techniques in data mining can be categorized as either Predictive or Descriptive. A predictive Model predicts the future using the available historical data. The tasks in predictive modeling are Time series Analysis, classification, regression, and prediction. Descriptive model identifies patterns or relationship in data. It serves as a way of exploring the properties of the data examined. The tasks in descriptive modeling include; Clustering, Association rules, summarization and sequence discovery.

The goal of the data mining process is to extract knowledge from an existing data set and transform it into a human understandable formation for advance use. It is the process of analyzing data from different perspectives and summarizing it into useful information. There is no restriction to the type of data that can be analyzed by data mining. As far as data mining technique is concern; in the most of cases predictive data mining approach is used. Information mining is in countless demand in other applicative databases, including spatial databases, temporal databases, object-oriented databases, multimedia databases, etc. Spatial data mining can be used to forecast crop yield estimation besides other real world applications. Spatial data mining is the process of extracting interesting knowledge from spatial databases. The facts can be used for understanding spatial and non- spatial data and their associations. Information discovered from spatial data can be of several forms, like distinctive and discriminant rules, extraction and description of prominent structures or clusters, spatial associations, and others. The difficulty has been to get information from this raw facts; this has led to new methods and skills such as data mining that can bridge the information of the data to the crop yield estimation.

Crop yield estimation has been a matter of interest for farmers, professionals, and agricultural related organizations. Food and Agriculture of the United Nations, describes crop forecasting as the art of predicting crop yields and production before the harvest actually takes place, normally a couple of months in advance. Crop yield estimation is a major component of national food security assessment and food policy making. Crop development and yield data are critical for amending agricultural farming system, and farming operation and management. The crop production could be affected directly and indirectly under climate change. Crop models and statistical models are two main tools for reviewing effects of climate change on crop yields. Majority of study agriculture focus on biological mechanisms to identify crop growth and improve its yield. The result of crop yield primarily depends on factors such as crop variety, the type of seed and environmental factors such as sunlight, soil, water, rainfall and humidity. Crop yield estimation is necessary, particularly in countries which depends on agriculture as their main source of economy. Crop yield prediction helps decisions makers to lay down some strategies in respect to the excess or deficit production conditions and allow timely import and export resolutions. Crop identification and prediction of yield are the main concern of remote sensing application in agriculture.(“ Ch.N.Santhosh Kumar et al, 2012”)

1.2 Statement of the problem

According to a report prepared by Kenya Food Security Steering Group (KFSSG) joint assessment conducted in July estimated 1.2 million people are acutely food insecure and in need of immediate food assistance, up from about 700,000 in February 2016. (KFSSG, 2017).

A lot of data on crop production and food security is collected from different sources, though; collecting large amounts of data often is both a blessing and a curse. There is a lot of data available containing information which should be used to the farmer’s advantage. This is a common problem for which the term data mining has been coined. Data mining techniques aim at finding those patterns or information in the data that are both valuable and interesting to the farmer(Anon 2015). Early prediction of crop yield is important for planning and taking various policy decisions. According to (Manjula & Narsimha 2016) farmers use the conventional technique of data collection for crop monitoring and yield prediction based on ground based visits and reports. These methods are subjective, very costly and time consuming.

A common specific problem that occurs is yield prediction. As early into the growing season as possible, a farmer is interested in knowing how much yield he is about to expect. In the past, this yield prediction usually relied on farmers' long-term experience for specific fields, crops and climate conditions. Limited study has been done to evaluate the different data mining techniques with respect to crop yield prediction and food security data(Mwanjele, 2015). The study seeks to implement data mining techniques to build a predictive model that has the potential to predict crop yield to enable farmers and agricultural departments lay down strategies for improving food security issues.

1.3 Motivation of the Study

Agriculture plays a vital role in Kenya's economy. Farm crisis in Kenya is deepening, mainly due to climate change, and depreciating crop prices. One indicator of the growing agricultural distress is farmers are selling land to venture into other activities. Reason behind this is agricultural organizations failed in policy making. So solution to this issue we are making a data mining tool that help to predict annual yield of crop depending on their environmental as well as crop zone area. So as the demand of food is increasing, the researchers, farmers, agricultural scientists and government are trying to put extra effort and techniques for more production. And as a result, the agricultural data increases day by day.

As the volume of data increases, it requires involuntary way for these data to be extracted when needed. Still today, a very few farmers are actually using the new methods, tools and technique of farming for better production. Data mining can be used for predicting the future trends of agricultural processes.

1.4 Objectives

1.4.0 Main Objective/Aim of the study

The study proposes to implement data mining techniques to build a predictive model that has the potential of predicting crop yield.

1.4.1 Specific Objectives

1. To investigate the factors affecting crop production
2. To establish appropriate data mining methods used to build predictive model crop production
3. To develop the appropriate predictive model using the established data mining methods
4. To evaluate the effectiveness of the developed predictive model

1.5 Research Question

The study is expected to realize its stated objectives by solving the following questions:-

1. What are the factors affecting crop production?
2. What are the appropriate data mining methods used in building predictive model for crop production?
3. Which is the appropriate predictive model that can be used in crop production?
4. How effective is the developed predictive model?

1.6 Significance of the study

The findings of this study will be of great significance to several stakeholders:-

Farmers

The farmers will be able to know how much yield they are expecting during the harvesting season. This will enable them plan well in advance in harvesting mechanism, storage of the yield and marketing of the products.

The Government

The food security steering group and the ministry of agriculture, will be able to plan ahead to curb food insecurity if in a given year the crop yield expected is lower. They will know how many bags of crop to import. If in a given year the produce is high then, they are in a much better position to reduce importation and store the excess in their silos.

The National Cereal Produce Board

The NCPB will benefit from the findings of the study, since the will know how much money they will set aside to buy crop from farmers. They will also plan ahead to improve on their storage facilities during the harvesting season.

The findings will also enable them to know how many bags of fertilizer, and the type of fertilizer they will buy and supply it to the farmers for the next planting season.

Scholars

This study has the potential of adding knowledge to the body of scholars who are interested in the field of computing and agriculture. They will be able to know where to start and the path to follow.

The ICT sector will also benefit, because the study has the potential to create awareness of which data mining technique should be used in crop yield prediction.

Scope of the study

The purpose of the study is to predict crop yield production to enhance food security. The study is to be conducted in Njoro Sub County, Nakuru County.

CHAPTER TWO LITERATURE REVIEW

2.0 Introduction

This chapter focuses on the previous studies done in the area. Review of empirical Literature, summary and gaps as well as conceptual framework. The section gives light to the study from previous researchers. This chapter is subdivided into Factors affecting Crop Production, Existing methods for Crop Yield Forecast, Data Mining Techniques that can be used for yield prediction.

2.1 Factors affecting crop production

2.1.0 Impact of Climate Change on Crop Production

Over the 20th Century the Earth's surface temperature has increased by an average of 0.8 °C (IPCC, 2013). The last 50 years, in particular, have seen a more rapid and steady increase. The IPCC (2013) report that the warming trend is likely to continue as long as emission of greenhouse gases (GHGs) continues unchanged. The IPCC (2013) has further noted that anthropogenic activities have, over time, altered both the Earth's surface and the composition of its atmosphere. Not only are these activities driving climate change, but also resulting to unprecedented ecological damage that is indirectly exacerbating climate change. It has been stated that according to temperature records in the last 150 years, the last three decades (1980-2010) have been warmer than the previous ones, with the 2000s being the warmest (IPCC, 2013). In its recent report, the IPCC (2013) projected that by the end of the 21st Century the Earth will warm up by between approximately 0.3°C and 1.7°C and 2.6°C to 4.8°C for the more conservative and more aggressive scenarios respectively. The projections, which are based on 1986-2005 temperature levels, may produce both direct and indirect adverse impacts on the Earth's surface. For instance, the increase in temperature is likely to interfere with rainfall patterns, thus affecting major food crops (IPCC, 2014). Increased warming may also drive changes in rainfall patterns and temperature regimes in different regions leading to extreme natural occurrences such as droughts and floods (IPCC, 2013).

Climate change in Africa, as argued by Hulme et al. (2001), is driven by both human and natural factors. Land cover changes together with the El Niño Southern Oscillation (ENSO) have been mentioned to be the major drivers of natural climate variability in Africa. The effects of climate change will be greatly experienced in the African continent due to the fact that the continent lacks sufficient technical and financial capacity to adapt to climate change (IPCC, 2007). For instance, the agricultural sector in the continent has been recognized as one of the sectors that will suffer most from the detrimental effects of climate change due to lack of adequate irrigation technology (Oseni and Masarirambi, 2011; Cairns et al., 2012; Cairns et al., 2013). Consistent with various climate change models, the Kenyan climate will experience several changes both in precipitation and temperature. A study carried out by the Famines Early Warning Systems Network (FEWS NET) in 2010 with the aim of establishing the trend of climate change in

Kenya, observed that a large part of the country will experience more than 100 mm decline in precipitation by 2025. However, findings from other computer models are conflicting, showing certain areas will get wetter (Anyah et al., 2006; DFID, 2009; Washington et al., 2012). It was also observed that there will be a significant increase in average air temperature of up to 1°C during the same period (FEWS NET, 2010). With spatial and temporal shift in rainfall patterns observed and predicted, coupled with a general reduction in recorded precipitation and increase in temperatures, agricultural production (especially maize) is likely to be affected in the country since it is mainly dependent on favorable climatic conditions for yield maximization.

2.1.1 Rainfall

One of the consequences of climate change is the alteration of rainfall patterns. The current study is based on the main growing season (March, April and May) and the associated effects on subsistence agriculture in Njoro District of Kenya. Due to global warming precipitation amount, type and timing are changing or are expected to change because of increased evaporation, especially in the tropics (Ritter, 2006). Although studies have shown a 2% overall increase in global land precipitation (IPCC, 2001), rainfall characteristics have shown considerable variations from region to region with some areas experiencing decline and in others increase in precipitation due to increased extreme weather patterns. Decrease in precipitation has been experienced in the Sahel, Mediterranean, southern African and parts of southern Asia where much of the rains fall as intense storms particularly in the dry areas (Ritter, 2006). A study by Xuebin Zhang and Francis Zwiers, of environment Canada in Toronto revealed that rainfall in Mexico and northern Africa decreased by nearly 70 mm per year in recent years compared to the 1925 rainfall as quoted in Boswell (2007). Wright and Jones (2003) note that rainfall in southwest Australia has declined by 15 to 20% from the late 1960s. On the other hand, significant increase in rainfall of 2.4% per decade was witnessed over the tropical lands during the 20th Century (IPCC, 2001). Similarly, rainfall has increased by up to 50% in parts of northern Europe (Ritter, 2006). Compared to the 1925 rainfall, Canada, Russia and Europe's annual rainfall has increased by 45 mm and in Peru and Madagascar, rainfall has increased by 60 mm annually (Boswel, 2007).

The seasonality, amount, distribution and the timing of the rainfall is of particular importance to the population that depends on rain-fed agriculture for their subsistence. But due climate change, the reliability of the rain for agricultural purposes has reduced in the recent years. For instance, despite the Indonesian annual rainfall decreasing by 2 to 3%, marked changes in precipitation patterns have been observed. Case et al. (2008) notes that the wet and dry seasons have changed with an increase in wet season rainfall in the southern region and a decreased dry season rainfall in the northern region of Indonesia. In Iberian Peninsula, Spain, the number of rain days increased over the 20th Century except in the Gulf of Cádiz and in western Portugal. The amount and intensity of rainfall declined in the Peninsula except for the two regions where the intensity increased (PhysOrg.com, 2011). IPCC (2001) indicates that in some parts of the equatorial East Africa December, January and February (DJF) rainfall may increase by 5 to 20% while June, July and August (JJA) rainfall may decrease by 5 to 10% under the lowest warming scenario. Climate change predictions indicates that North Africa will experience a significant decline in rainfall of 10 to 20% between March and November while South Africa rainfall will decrease by 5 to 15% between November and May by 2050.

Agricultural activities follow rainfall patterns especially in tropical regions. In Sub-Saharan Africa rain-fed agriculture, which provides food for the populace and represent a major share of the countries' economy follow precipitation pattern closely (United Nations Environmental Programme, 2008). Therefore, short-term as well as long-term variations in rainfall patterns have important effects on crop and livestock farming (IPCC, 2007). Seasonal rainfall has been marked by delayed onsets, declining number of rain days and increased intensities altering farming calendars with negative effects on the yields. Awuor and Ogola (1997) observed that global warming will likely affect the agricultural productivity due to changes in the length of growing periods. They observed that the length of growing period would increase by about 10 days/°C increase in mean annual temperature in Canadian Praire. In Kenya, they noted that increase in temperature by 4°C will result in a dramatic shortening of the length of the growing period. In Myanmar, Phyu (2010) notes that the monsoon period reduced by about 10 days in 2009 from an average of 144 days. The main growing rainfall seasons which occur in March, April and May (MAM) and December, January and February (DJF) have been declining since the 1980s in Ethiopia, Kenya, Tanzania, Zambia, Malawi, and Zimbabwe (UCSB, 2008). UCSB attributes the declining rainfall to anomalies in moisture transport between the ocean and land due to rising sea surface temperatures over the Indian Ocean.

According to IPCC (2007), changes in rainfall patterns have negatively affected mixed rain-fed and highland perennial systems in the Great Lakes region and in other parts of East Africa. Agriculture has progressively become more marginal in the arid and semi-arid areas. A significant decrease in suitable rain-fed agricultural land and production potential for cereals has been predicted under climate change by 2080. For instance, under increased El Niño Southern Oscillation (ENSO), South Africa would experience significant reductions in maize production. On the other hand, a combination of increased temperature and rainfall may result in longer growing seasons in some parts of the Ethiopian highlands and southern Africa such as Mozambique improving agricultural performance. Based on these observations, it therefore follows that rainfall patterns have been changing as climate changes and so has rain-fed agriculture. Thus, for agricultural purposes changes in growing period are very important and must be viewed against possible changes in seasonality of rainfall, onset of rain days and intensity of rainfall. This study is based on the long rainfall season (March, April and May (MAM)) which is the main growing rainfall season in Njoro District of Kenya since farmers largely depend on this season for subsistence food production. Rainfall during the short rains, which occur in October November and December (OND), is unreliable and inadequate particularly for maize growing, which forms the staple food for the inhabitants.

2.1.2 Temperature

The correlation between maize yield and annual temperature increases were determined to be insignificant, implying that the current change in temperature does not influence maize yields in the country. This situation can be attributed to several physiological reasons regarding maize crop. Elevated temperatures that exceed the threshold (above 32 °C for atmospheric and 26 °C or soil temperature) have detrimental effects on maize growth (Sundquist and Joanne, 1984, and Walker, 1969, respectively). Extreme temperatures affect the fundamental cell functions (such as cell division, sugar metabolism and starch synthesis) in maize crops (Monjardino et al., 2005).

Ultimately, the effects of extreme temperatures will lead to the reduction of both maize grain number and kernel weight (Schoper et al., 1987) leading to significant yield loss. The mean temperature in maize growing areas in Kenya was 18.68 °C (in 2010) compared to 18.01°C three decades ago. This falls under the threshold of optimum atmospheric temperature for maize which is 15-23°C (FAO, 2013a). However, there are other factors such as pests and diseases and maize storage that are directly affected by temperature change consequently affecting maize production.

2.1.3 Impacts of Soil PH

Soil acidity is a noteworthy yield constraining element for trim generation around the world. Land region influenced by acridity is assessed at 4 billion hectares, speaking to around 30% of the aggregate without ice arrive region of the world (Sumner and Noble, 2003). In the tropics, considerable weathering of soils over centuries has brought about the draining of yield supplement bases (chiefly K, Mg and Ca) trailed by their substitution by H, Al, Mn cations which have added to corrosive related weights on edit creation (Okalebo et al., 2009). Corrosive fruitlessness factors confine trim development and yield and also soil efficiency in exceedingly weathered soils of muggy and sub-damp locales of the world because of lack of fundamental supplement components (Akinrinade et al., 2006).

In Kenya, corrosive soils cover around 13% of aggregate land region and are conveyed broadly in the croplands of central and western Kenya locales, covering more than one million hectares under maize, vegetable, tea and coffee crops, developed by more than 5 million smallholder ranchers (Gudu et al., 2007). Crop production is low and declining on such corrosive soils and especially where corrosive shaping composts, for example, di-ammonium phosphate (DAP) and other alkali manures have been connected consistently to as of now fermented soils over years (Nekesa, 2007). As these soils suffer in multi-supplement insufficiencies, utilization of mineral manures has turned out to be obligatory to increase crop yields.

In any case, mineral manures are ordinarily rare, exorbitant; having imbalanced sustenance and their utilization could fuel the issue of soil sharpness (Oguike et al., 2006; Nottidge et al., 2006). The act of liming corrosive soils isn't normal in Sub-Saharan Africa (SSA), presumably as a result of restricted information onlime use and its adequacy, accessibility and high pulling expenses of liming materials (Okalebo et al., 2009).

Consistent farming utilizing off base manure sorts has heightened soil substance debasement of arable terrains bringing about diminished limit of soils to create trims reasonably (Nandwa, 2003; Ayuke et al., 2007; Mugendi et al., 2007). As per Kisinyo et al. (2005), persistent trimming has prompted advancement of soil acridity which is a noteworthy requirement to maize generation on tropical soils because of harmful levels of aluminum (Al) and the attending phosphorus (P) inadequacy that upset plant development.

2.2 Data mining methods that can be used for yield prediction

2.2.0 Overview

The vision of taking care of world requests for the expanding the product yield for population all through the world is turning out to be more vital in the current years. Predictive model and choice devices are progressively utilized as a part of farming fields to enhance creation effectiveness. The blend of cutting edge innovation and agribusiness used to enhance the generation of product yield. The data mining procedures like grouping, neural systems and relapse are required to apply on the sensible informational indexes for investigations and make the forecast on the farming product yield. The data mining tasks could be characterized in two classes.

- Descriptive data mining.
- Predictive data mining.

Descriptive Data Mining: Descriptive demonstrating is a scientific procedure that depicts genuine occasions and the connections between elements dependable for them. The procedure is utilized by buyer driven associations to help them focus on their promoting and publicizing endeavors. The primary parts of graphic demonstrating include:

- Customer Segmentation: Partitions a client base into gatherings with different effects on promoting and administration.
- Value-based Segmentation: Recognizes and measures the estimation of a client to the association.
- Behavior-based Segmentation: Analyzes client item use and acquiring designs.
- Needs-based division: Identifies approaches to profit by thought processes that drive client conduct.

Predictive Data mining: Predictive examination is the way toward separating data from huge sets so as to make forecast and gauges future results. The rundown of yield forecast models that a greater amount of them have been for the most part characterized in two gatherings strategies.

- Traditional Approach
- Artificial Intelligence

Conventional Approach: Most of the farmers depended on their long-terms encounters in the field on specific harvests to expect a higher yield in the next gathering period. Manmade brainpower: The advancement of PC frameworks ready to perform errands regularly requiring human knowledge, such as visual discernment, discourse acknowledgment, basic leadership, and interpretation between dialects.

2.2.1 Applications of data mining techniques in agriculture

(Khedr et al. 2015) in their paper, they proposed the use of ANN (Artificial Neural Network) via WEKA using Multilayer Perceptron(MLP) function as a data mining technique for prediction. The results showed that the data mining framework succeeded in predicting the annual needed amount of main strategic crops up to the year 2020.

In his research about the efficient management of feed resources using data mining techniques (Rani, 2010) mentioned that clustering of feed resource into different groups based on the composition can help in better feed management. The major findings of the study are for clustering techniques that could be used for classifying the feed resources into different clusters with reasonably good accuracy. Further, the K-means found to be the best technique for classification. He found that the k means clustering was not affected by sample size and the distribution of different feed resources in different data sets as the outputs were almost consistent across groups.

His study demonstrates that clustering could be effectively used for grouping of different feed resources without the aid of experts to an extent of 70% and thus can form a sound basis for efficient feed management. The application of the study was that it can be used for clustering new feed resource into a particular category which will be useful in determining the extent of usage of the new feed. Clustering the feeds into different groups provides multiple options to the end user (farmer or feed industry) to choose from a wide range of feed resources that is best suited in terms of local availability or price advantage.

To determine the vanished agricultural land in Tamilnadu India, (Megala and Hemalatha, 2011) used data mining approach. The study aimed to establish if data mining techniques can be used to assist in the clustering methods by determining whether meaningful patterns exist across various land profiles at various research sites across Tamil Nadu in India. The overall aim of the research was to determine the land utilization for agriculture and non-agriculture areas for the past ten year. The research used Software to conduct qualitative analyses and to create a benchmark for the analysis of the dataset. The dataset is analyzed using a clustering process within the data mining software.

(Salame, 2011) applied Data mining techniques to evaluate applications for agricultural loans. The study used Logistics regression, neural network and decision tree to identify the financial and non-financial variables that signal the capacity of borrowers to pay back the loan, and determine the best model(s) to evaluate credit risk. Financial institutions that serve agriculture need to continuously evaluate their models and methods to assess the probability of default on loans, especially when assessing the probability of default of a new borrower by examining the performance of three different methods.

In their research, (Raorane and Kulkarini, 2010) discussed the role of data mining as an effective tool for yield estimation in the agricultural sector. As crop production depends on climatic, geographical, biological, political and economic factors, data mining can solve the challenge of extracting knowledge from this raw data and estimate the amount of crops production. Accurate and reliable information about historical crop yield is important for decisions relating to agricultural risk management. An accurate estimate of crop size and risk helps in planning supply chain decision like production scheduling.

In their paper, (Wang et al 2012) mentioned that evaluation and early warning prediction that Food safety risk early warning is one of the most important contents in food security management. An effective early warning system could improve the management of food security status. A great number of researchers are engaged in the research of food security early warning and they have got a lot of achievements, Data were collected via anonymous questionnaires, surveys, etc.

(Liu et al. 2010) proposed a data-mining technique to predict food quality using a back-propagation neural network. To reduce errors, data near the threshold values are selected to train their proposed system.

Table 1 Previous studies and findings

Author(s)	Case	Techniques	Findings
Raorane and Kulkarini, 2010	Data mining as an effective tool for yield estimation in the agricultural sector.	Classification, Clustering	Artificial Neural Network is a new technique used in flood forecast. Decision tree is one of the classification algorithms. Support Vector Machine is able to classify data samples in two disjoint clusters. SVM are a set of related supervised learning method used for classification and regression.

			K means method is used to forward the pollution in atmosphere. Different changes of weather are analyzed using SVM. K means approach is used to classify the soil and plants.
Rani, 2010	Efficient management of feed resources using data mining techniques	Clustering	Clustering can be effectively used for grouping of different feed resources without the aid of experts to an extent of 70% and thus can form a sound basis for efficient feed management
Megala and Hemalatha, 2011	Data mining approach to determine the vanished agricultural land in Tamilnadu, India	Classification/ Clustering/ Visualization	Reduced the time taken to undertake data analysis. Increased automation of the process. The integrity of the data is critical to ensure that results are not affected by outliers and null values in the data set, or other adverse factors. The establishment of clusters in the data

			required a large amount effort by the researchers when using current methods. Maps classification have been shown in order to understand the land lost with in the past seven years. Which helps in understanding of the biophysical and environmental management
Khedr, Kadry, & Walid, 2015	Proposed framework for implementing data mining techniques to enhance decisions in agriculture sector	Classification	Data_mining framework succeeded in predicting the annual needed amount of main strategic crops up to the year 2020.

2.3 Existing Methods for yield prediction

There are a few techniques for yield estimating. The traditional strategy for yield estimating is the assessment of product status by specialists. Perceptions and estimations are made all through the harvest developing season, for example, tiller number, spikelet number and their ripeness rate, rate of harm from nuisances and organisms, rate of weeds pervasion, et cetera. From the information acquired along these lines yield can be determined using relapse techniques, or by the learning from neighborhood expertizes. Other two strategies used to predict crop yield are the use of remote detecting and Crop Simulation models (CSM). The target of the yield conjecture is to give an exact, logical sound and autonomous figures of harvests' yield as ahead of schedule as conceivable amid the products' developing season by considering the impact of the climate and atmosphere.

The contrasts amongst gauges and last gauges are in the planning of the discharge. Conjectures are made before the whole yield has been collected while assessments are made after the product has been gathered. Signs are the after effect of applying a measurable estimator to the review information and the subsequent point evaluations are deciphered by ware analysts to make conjectures and gauges. Truly, agriculturists have been continually making "estimates" to arrange their agronomic practices. For instance, the planting window, the decision of a cultivar, the measure of manure to apply rely on upon the atmosphere. On the off chance that farmers realize that the resulting week there is a decent possibility for rain, then they will hurry into the field to sown their seeds. Gauging crop yield implies likewise knowing or determining other vital parameters. For instance, evaluating the range planted at the beginning of the developing season and measuring the region gathered.

Crop Yield Forecast using Statistical Models

Yield estimation using agro meteorological contributions to a factual relapse is somewhat normal and used as a part of many yield estimates research and projects (NASS, 2006; Lobell et al., 2009). As a rule, a straightforward measurable model is fabricated using a grid with memorable yield and a few agrometeorological parameters (e.g. temperature and precipitation). At that point, a relapse condition is inferred between yields as capacity of one or a few agronometeorological parameters. The NASS (2006) program utilizes a measurable model to estimate trim yield and generation. They utilize two techniques to conjecture yield, the previous allude to the utilization test determined models for guaging yields and their part; the last mentioned, is utilize models at state and local levels. For instance the NASS (2006) framework for corn estimate depends on the two levels depicted previously. A corn target yield study was initially led in the 10 noteworthy corn-delivering nations. They direct some point by point field studies for assurance of yield segments which can be utilized for the field-level yield gauge.

While to acquire state/territorial gauges an accumulation of the information is done first and after that a factual model is utilized. The upsides of a factual model is that the count is simple, less time is required to run the model and the information necessities are restricted. Be that as it may, they are restricted in the data they can give outside the scope of qualities for which the model is parameterized. Likewise the yield of such models won't not have any agronomic significance, while factually are still right. Moreover, they don't think about the dirt plant-air continuum, which is vital when managing areas having diverse soil sorts. For instance, the reaction of a harvest to a given measure of precipitation on a sandy soil is not quite the same as a product on a dirt soil.

The planning of the water stretch happening amid the developing season is likewise vital and frequently overlooked. For instance, a warmth stretch happening at blooming will lessen yield more than a warmth push occurring amid the vegetative stage. This is vital for effectively anticipating yield and for giving agriculturists essential agronomic advices (e.g. timing and measure of manure, time of sowing, water system, et cetera). There are endeavors of incorporate additionally importance into the factual models so as to stay away from a portion of the issues portrayed. For instance, the incorporation of harvest evapotranspiration, as well as the underlying soil dampness content (acquired through microwave detecting) as parameters of the model enhance the consistency control however leave the agronomic inquiries unanswered

2.3.1 Crop Yield Forecast using Process-Based Models

Agro-ecosystems are perplexing elements where crop yield is the resultant of numerous co-operations, for example, soil, air, water, and financial elements. CSMs are worked with the expect to consider the continuum soil-plant-climate and its every day changes on the day by day aggregation of biomass and nitrogen. There are numerous CSM around the globe, for instance, Asseng et al. (2013) utilized more 27 wheat models in their intercomparison. Not every one of the models are the same. Some are extremely basic without any than 7 parameters expected to depict a specific cultivar (Bondeau et al., 2007), to other which consider definite prepared like photosynthesis at leaf level and along these lines require numerous client indicated parameters. CSM have been broadly used to assess the agronomic outcomes from the between yearly atmosphere fluctuation (Paz et al., 2007; Semenov and Doblas-Reyes, 2007; Challinor and Wheeler, 2008). For instance, CSM can catch the impacts of and timing of wet/dry cycles on product development; which can essentially help famers in arranging their agronomic administration (Shin et al., 2009). Not all the CSM can be or will be utilized for agro technology exchange; fruitful utilizations of a CSM to yield gauge is capacity of numerous things yet the most imperative is the measure of parameters expected to depict trim soil-environment. A fruitful case of a CSM application is the yield prophet (http://www.yieldprophet.com.au/yp/wfLogin.aspx) which is an on-line CSM intended to advise agriculturists and advisors with continuous data about their products, giving danger appraisal data and observing choice bolster important to ranch administration. The framework is worked as a web interface for the Agricultural Production Systems Simulator (APSIM).

Another successful utilization of online straightforward interface crop reproduction model is the SALUS show as portrayed by Basso et al., (2012, 2010) (www.salusmodel.net). The model is intended to be utilized by agriculturists or augmentation experts to measure the effect of administration, soil and climate cooperation on yield and natural effect. The SALUS (System Approach to Land Use Sustainability) model is a progressing collaboration began by Joe Ritchie at Michigan State University in late nineties and at present carried on by Bruno (Basso et al.2006; Basso et al. 2010). SALUS is comparable in detail to the DSSAT group of models yet is intended to reproduce crop yield in pivot, soil, water and supplement rapid as capacity of administration systems for numerous years. SALUS represents the impacts of pivots, planting dates, plant populaces, water system and manure applications, and culturing hones. The models mimics every day plant development and soil flow forms on a day by day time venture amid the developing season and decrepit periods. SALUS contains: i) crop development modules; ii) soil natural matter and supplement cycling module and; iii) soil water adjust and temperature module. The model reenacted the impacts of soil-atmosphere and administration communication on the water adjust, soil natural matter, nitrogen and phosphorus (P) flow, warm adjust, plant development and plant improvement. i-Salus is additionally an electronic agronomic choice emotionally supportive network - named i-Salus - to help farmers enhance their water system and nitrogen administration hones over space and time. i-Salus permits clients to assess the best administration technique to enhance yield and nature of the harvests, to increment monetary net return and in the meantime to diminish nursery gas emanations and groundwater defilement from nitrate filtering. i-Salus is made by two interfaces: a straightforward interface and Web-GIS interface. SALUS with the basic interface is an easy to understand framework s focuses at farmers or augmentation pros who can reenact the effect of various administration techniques on yield, and ecological effect. SALUS-WebGIS is an online GIS coordinated with Google Earth and Salus model to reproduce in a spatial unequivocal way the impact of atmosphere soil-genotype-administration connection on product yield and natural effect. Both frameworks are accessible at www.salusmodel.net

The troubles of embracing CSM has as a rule been related with the serious information for models' parameterization. The requirement for adjustment can be very information broad and not material to some creating nations Actually, it has been contended that few factors are expected to adjust/assess the CSMs, finishing up the convenience of CSMs in some "genuine" circumstance on account of the difficulty of social affair sources of info and alignment datasets. Be that as it may, a more basic take a gander at the writing and the work done by other analyst brings up that the CSM can be run using "Least Data Set" (MDS) contributions (as examined prior). Models, similar to the i-Salus case detailed above, have appeared to be anything but difficult to use by anybody yet keep up its vigor in yield forecasts. It has been pointed that another constraint of the CSM is that they are "point-based" and lacking to keep running at provincial/national scale. In any case, Bondeau et al. (2007) and Challinor et al. (2004) built up a basic CSM that can be keep running at local and national scale with less requests on sources of info and adjustment dataset.

2.3.2 Yield estimate with remote sensing

The meteorological models used for forecasting yield are mainly based on two variable, temperature and precipitation because they are related to crop yields and can be easily obtained from meteorological stations or from satellite measurements such as the NASA Prediction Of Worldwide Energy Resource Project (POWER; White et al., 2011). These two inputs can be used singularly, or as a combination. They can be used as daily or monthly variables. Such models are generally a simple regression, and three main methods are commonly used: time-series; based on changes in space and time; and based on changes in space (Lobell and Burke, 2010). The application of agrometeorological models is more common nowadays because of the integration with RS. Doraiswamy et al. (2003) reported one of the first examples in which production is forecasted through satellite remote sensing and measured meteorological observations on the ground. The Large Area Crop Inventory Experiment (LACIE) project, which was launched in 1974, used satellite remote sensing to forecast wheat production in the major wheat-producing countries.

Kogan et al. (2012) estimated winter wheat, sorghum and corn yields 3/4 months before harvest using the AVHRR data. The errors of yield estimated were 8%, 6%, and 3% for wheat, sorghum and corn, respectively. While Meroni et al. (2013) using the SPOT platform (Satellite Pour l'Observation de la Terre) and a statistical model quantified wheat yield in north Tunisia and concluded that where crop conditions need to be quantified without ground measurements for calibration, the biomass proxies are preferred. The use of low resolution satellite images, along with the high temporal frequency, their wide geographical coverage, and their unitary low costs per area, means that these images are a good choice for yield estimation as showed by many reported findings. Another platform that is commonly used is the National Aeronautics and Space Administration's (NASA) Moderate Resolution Imaging Spectroradiometer (MODIS), which has been demonstrated to give better spectral and spatial resolution relative to the AVHRR (Doraiswamy et al., 2001; Ren et al., 2008; Funk and Budde, 2009; Becker-Reshef et al., 2010). Bolton and Friedl (2013) used satellite data from MODIS to develop empirical models for crop and soybean yield forecast in the Central United States. The EVI2 index showed better ability to predict crop yield than the NDVI, and the use of crop phenology information from MODIS improved the model predictability. Although MODIS has a low spatial resolution, the authors showed that MODIS was still able to identify the agricultural areas without affecting model's output compared to the higher-spatial resolution crop-type maps developed by the USDA. Kogan et al. (2013) used NDVI values from the MODIS, at 250 m spatial resolution for forecasting wheat yield in Ukraine. In Kenya the MODIS was used to derive images for six sugarcane management zones, over nine years, to estimate sugarcane yield based on each zone. Because of the zoning, the different management strategies were taken into account using the temporal series of NDVI which was normalized by a weighting method that includes sugarcane growth and the time-series of the NDVI. The challenge is to estimate the yield in small-holder farmers where the fields are often smaller that the spatial resolution of the MODIS used (250 m).

Recent studies have used both higher spatial resolution data from Landsat with higher temporal frequency data from MODIS or AVHRR (e.g., Mkhabela et al., 2011).

Yield forecast through RS and models has been made for several cropping systems. On vineyards, the final yield was evaluated using high spatial resolution RS for the estimation of canopy vigor and yield components (Hall and Wilson, 2013).

RS techniques have been extensively used in research for yield forecast but played a small role in understanding the cause of spatial yield variability. Also, it has been argued that while RS might not be suitable in developing countries because of their stratified agricultural systems and very small farm sizes. However, this problem is hard to overcome in the near-future because of the inability of RS to estimate yield in mixed agriculture. But, the increased availability of high-spatial resolution RS at a reasonable cost make this technique a possible interesting alternative for yield forecast.

2.4 Knowledge Gap

After scrutinizing the literature, limited study has been done to evaluate the different data mining techniques with respect to crop yield prediction and food security data in Kenya (Mwanjele,2015). There is a need for an objective methodology for pre-harvest crop forecasting. This involves building up suitable forecast model(s) which has certain merits over the traditional forecasting method. These merits include the objectivity of the forecast and its ability to provide a measure of reliability which a traditional forecast method cannot provide. This, as such, calls for the necessity of objective methods for pre-harvest forecast of crop yields.

Conceptual Framework

Figure 1 Conceptual Framework

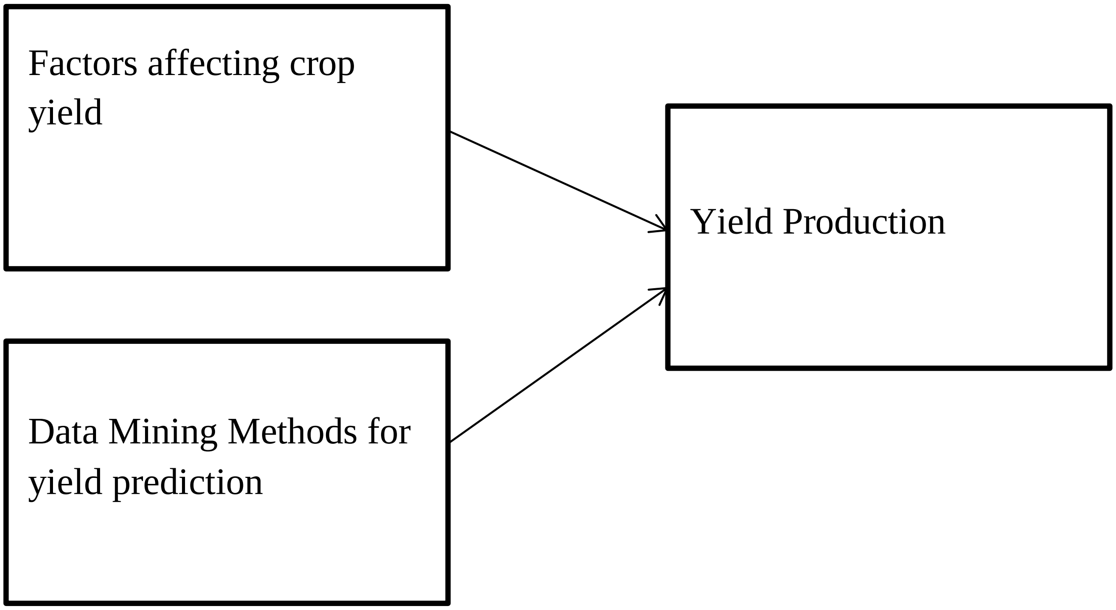

Factors affecting crop yield
Yield Production
Data Mining Methods for yield prediction

Table 2 Variable Operation

Variable	**Sub Variable**	**Indicators**	**Values**
Factors Affecting crop production	Climate Change	Amount of rainfall	High
			Average
			Low
		Temperature	High temperature
			Low Temperature
	Input factors	Labor	High cost
			Low cost
		Fertilizer	High Prices
			Low prices
		Cost of seed	High prices
			Low prices
	Soil PH	Acidic	Strongly acidic

			Slightly acidic
		Alkaline	Strongly alkaline
			Slightly alkaline
		Neutral	Yes
			No
Data Mining Methods for Prediction	Classification Methods	Trees	J48
			ID3
			Simple CART
		Functions	SVM
			Multi-Layer Perceptron
			Linear Regression
		Rules	PART
			Conjunctive Rule
Yield prediction	Prediction	Yield prediction accuracy	High
			Low

CHAPTER THREE: RESEARCH METHODOLOGY

3.0 Introduction

The chapter presents the research methodology used in this study. The chapter discusses the research design, target population, sampling design and procedures. It also includes the data collection methods.

3.1 Method for Achieving Objective 1 and 2

To achieve both objective 1and 2, survey of literature research design is implemented. Twenty five (25) Journal papers and 5 textbooks were downloaded.

3.1.0 Sampling Design

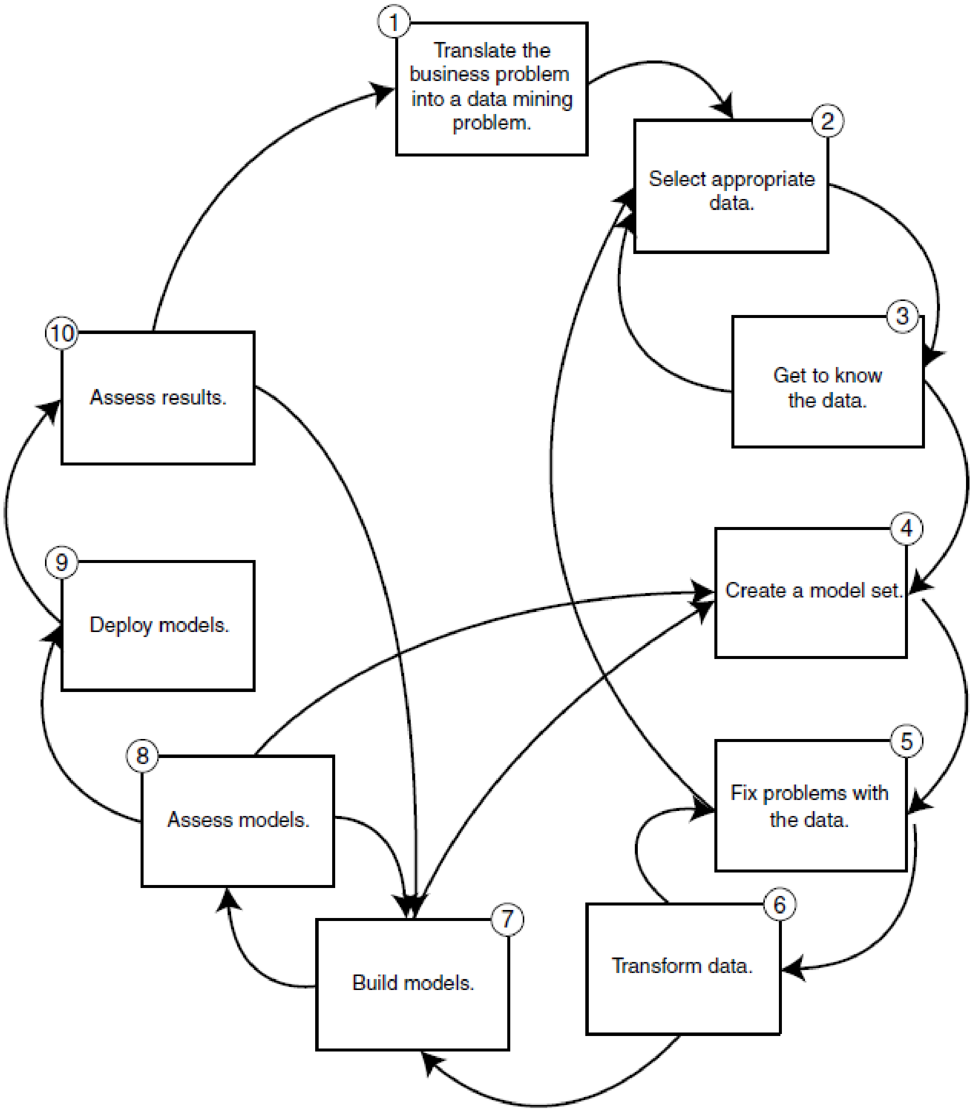
Purposive sampling technique was used. Purposive sampling is defined as sampling technique in which researcher relies on his or her/own judgment when choosing members of population to participate in the study. An inclusion and exclusion criterion was used to select the journal papers and textbooks to be used to conduct the study.

3.2 Method to Achieve Objective 3 and 4

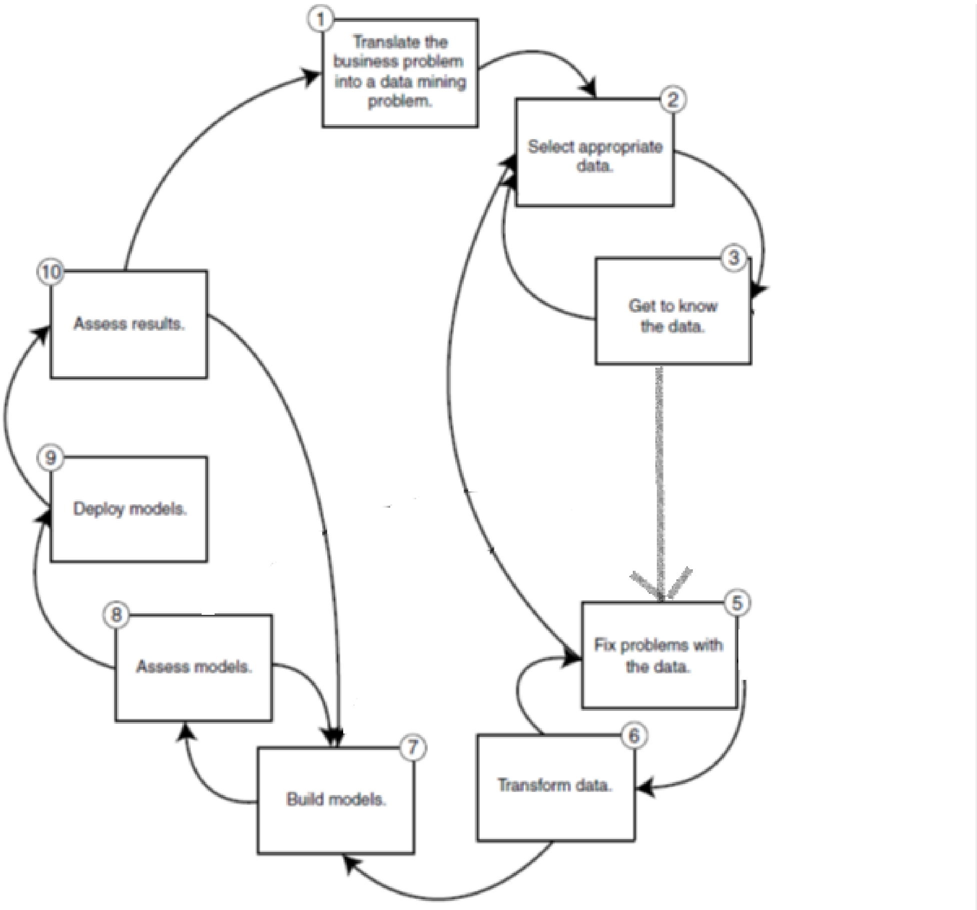
3.2.0 Research Design

(*Berry and Linoff*,*2009*) Define Data Mining methodology as a system of methods, rules, procedures and principles for regulating a data mining process. Methodology is designed to ensure that the data mining effort leads to unwavering model that positively addresses the business problem it is aimed to solve. As illustrated below, data mining is not a linear process but rather it is best thought as a set of nested loops.

Figure 2 Data Mining Methodology Source: (Berry & Linoff, 2009.)



The researcher adopted the berry and Linoff Data Mining Methodology for the for the purpose of this study. The Diagram was modified though there is no much difference as shown in the figure below



Step 1: Translating the business problem into a data mining problem

The proper destination for a data mining project is the solution of a well-defined business problem. Data mining goals for a particular project should not be stated in broad, but rather should be divided into more specific ones to make it easier to monitor progress in achieving them. This was done by identifying the problem statement and describing how the problem can be addressed using data mining technics, through the use of literature review.

Step 2: Select Appropriate Data

Data mining requires data. In the best of all possible worlds, the required data would already be resident in a corporate data warehouse, cleansed, available, historically accurate, and frequently updated. In fact, it is more often scattered in a variety of operational systems in incompatible formats on computers running different operating systems, accessed through incompatible desktop tools. The data sources that are useful and available vary, of course, from problem to problem and industry to industry. This was achieved through sampling as shown in the table below

Table 3 Sampling Method

Target/Sample size	Sampling Method	Subject
1	Convenience Sampling	Country (Kenya)
1	Convenience Sampling	County (Nakuru)
1	Convenience Sampling	Sub County (Njoro)
3	Convenience Sampling	Wards(Njoro, Mau,Lare)
100	Purposive Sampling	Farms

Step 3: Get to Know the Data

It is hard to overstate the importance of spending time exploring the data before rushing into building models. Good data miners seem to rely heavily on intuition—somehow being able to guess what a good derived variable to try might be, for instance. The only way to develop intuition for what is going on in an unfamiliar dataset is to immerse yourself in it. Along the way, you are likely to discover many data quality problems and be inspired to ask many questions that would not otherwise have come up. This was achieved through Human Computer Interaction.

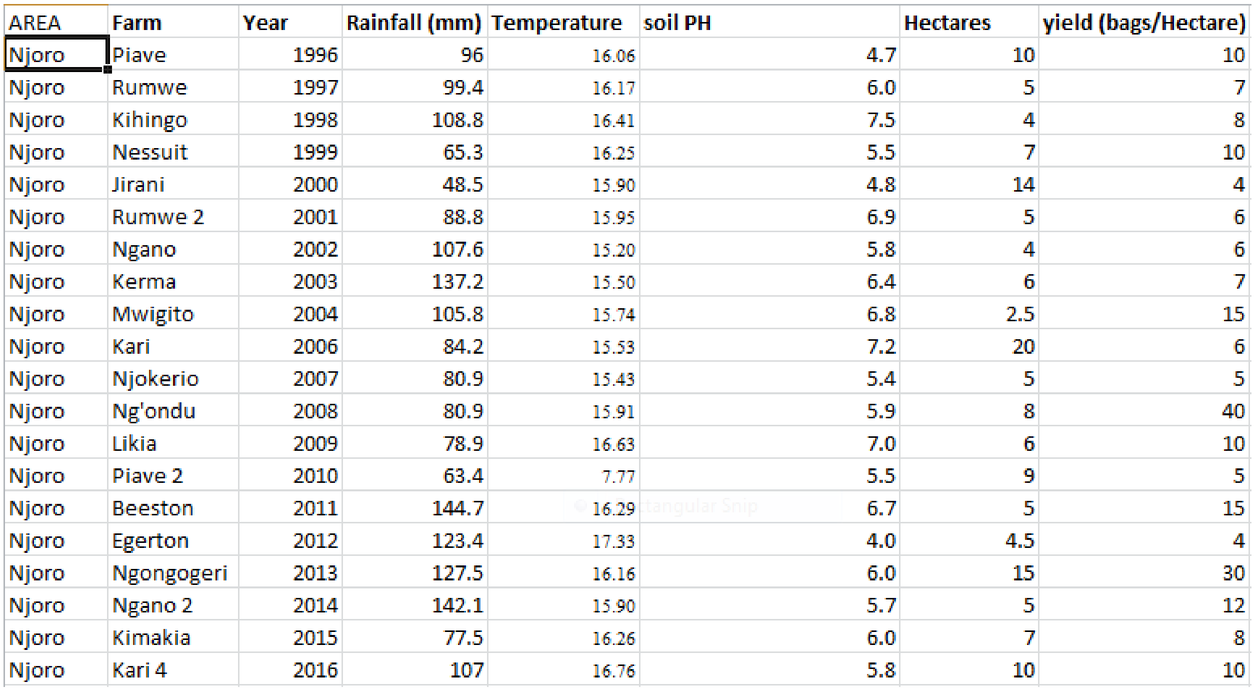
Step 4: Fix Problems with the Data

All data is dirty. All data has problems. What is or isn't a problem varies with the data mining technique. For some, such as decision trees, missing values, and outliers do not cause too much trouble. For others, such as neural networks, they cause all sorts of trouble. The researcher did data cleaning using a global constant "nil (0)" while cleaning the missing value and removing the columns or attributes with repetition and has no meaning or function.

Step 5: Transform Data to Bring Information to the Surface

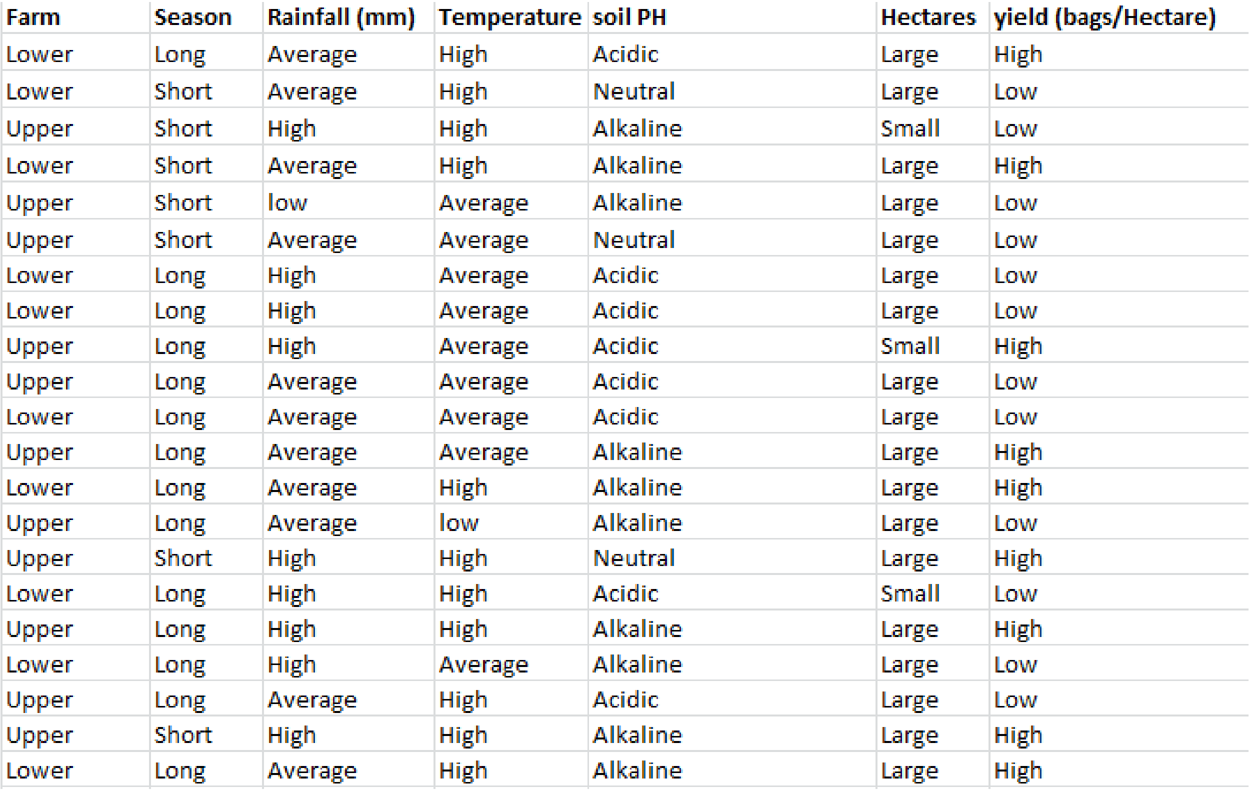
Once the data has been assembled and major data problems fixed, the data must still be prepared for analysis. This involves adding derived fields to bring information to the surface. It may also involve removing outliers, binning numeric variables, grouping classes for categorical variables, applying transformations such as logarithms, turning counts into proportions, and the like. This step was achieved by grouping data categorical variables

Figure 3 Data Set before Transformation



AREA	Farm	Year	Rainfall (mm)	Temperature	soil PH	Hectares	yield (bags/Hectare)
Njoro	Piave	1996	96	16.06	4.7	10	10
Njoro	Rumwe	1997	99.4	16.17	6.0	5	7
Njoro	Kihingo	1998	108.8	16.41	7.5	4	8
Njoro	Nessuit	1999	65.3	16.25	5.5	7	10
Njoro	Jirani	2000	48.5	15.90	4.8	14	4
Njoro	Rumwe 2	2001	88.8	15.95	6.9	5	6
Njoro	Ngano	2002	107.6	15.20	5.8	4	6
Njoro	Kerma	2003	137.2	15.50	6.4	6	7
Njoro	Mwigito	2004	105.8	15.74	6.8	2.5	15
Njoro	Kari	2006	84.2	15.53	7.2	20	6
Njoro	Njokerio	2007	80.9	15.43	5.4	5	5
Njoro	Ng'ondu	2008	80.9	15.91	5.9	8	40
Njoro	Likia	2009	78.9	16.63	7.0	6	10
Njoro	Piave 2	2010	63.4	7.77	5.5	9	5
Njoro	Beeston	2011	144.7	16.29	6.7	5	15
Njoro	Egerton	2012	123.4	17.33	4.0	4.5	4
Njoro	Ngongogeri	2013	127.5	16.16	6.0	15	30
Njoro	Ngano 2	2014	142.1	15.90	5.7	5	12
Njoro	Kimakia	2015	77.5	16.26	6.0	7	8
Njoro	Kari 4	2016	107	16.76	5.8	10	10

Figure 4 Transformed Data set



Farm	Season	Rainfall (mm)	Temperature	soil PH	Hectares	yield (bags/Hectare)
Lower	Long	Average	High	Acidic	Large	High
Lower	Short	Average	High	Neutral	Large	Low
Upper	Short	High	High	Alkaline	Small	Low
Lower	Short	Average	High	Alkaline	Large	High
Upper	Short	low	Average	Alkaline	Large	Low
Upper	Short	Average	Average	Neutral	Large	Low
Lower	Long	High	Average	Acidic	Large	Low
Lower	Long	High	Average	Acidic	Large	Low
Upper	Long	High	Average	Acidic	Small	High
Upper	Long	Average	Average	Acidic	Large	Low
Lower	Long	Average	Average	Acidic	Large	Low
Upper	Long	Average	Average	Alkaline	Large	High
Lower	Long	Average	High	Alkaline	Large	High
Upper	Long	Average	low	Alkaline	Large	Low
Upper	Short	High	High	Neutral	Large	High
Lower	Long	High	High	Acidic	Small	Low
Upper	Long	High	High	Alkaline	Large	High
Lower	Long	High	Average	Alkaline	Large	Low
Upper	Long	Average	High	Acidic	Large	Low
Upper	Short	High	High	Alkaline	Large	High
Lower	Long	Average	High	Alkaline	Large	High

Step 6: Build Models

The details of this step vary from technique to technique. In general terms, this is the step where most of the work of creating a model occurs. In directed data mining, the training set is used to generate an explanation of the independent or target variable in terms of the independent or input variables. This explanation may take the form of a neural network, a decision tree, a linkage graph, or some other representation of the relationship between the target and the other fields in the database. In undirected data mining, there is no target variable.

The model finds relationships between records and expresses them as association rules or by assigning them to common clusters. Building models is the one step of the data mining process that has been truly automated by modern data mining software. For that reason, it takes up relatively little of the time in a data mining project.

Step 7: Assess Models

This step determines whether or not the models are working. A model assessment should answer questions such as:

- How accurate is the model?
- How well does the model describe the observed data?
- How much confidence can be placed in the model's predictions?
- How comprehensible is the model?

Of course, the answer to these questions depends on the type of model that was built. Assessment here refers to the technical merits of the model, rather than the measurement phase of the virtuous cycle. To test the accuracy of the model, the N-fold cross validation test option was used. In this case N=10

Step 8: Deploy Models

Deploying a model means moving it from the data mining environment to the scoring environment. This process may be easy or hard. In the worst case, the model is developed in a special modeling environment using software that runs nowhere else. To deploy the model, a programmer takes a printed description of the model and recodes it in another programming language so it can be run on the scoring platform.

Step 9: Assess the results

The real test of data mining comes when you can measure the value of the actions you took as a result of the mining. Measuring lift on a test helps you to choose the right model.

Profitability models based on lift can help you decide how to apply the results of the model. However, it is very important to measure these things in the field as well.

Step 10: Begin Again

Every data mining project raises more questions than it answers. This is a good thing. It means that new relationships are now visible that were not visible before. The newly discovered relationships suggest new hypotheses to test and the data mining process begins all over again.

3.3 Target Population

Njoro district which is approximately 799.7 KM^2 was the case study area where this study applied the Knowledge Discovery and Data mining (KDD) process steps. There being no much study done on crop yield prediction in Kenya and further crop yield prediction problem being rather complicated to analytically explain, the case study approach was the best to yield a rich picture of the situation, which can well be further subjected to comparative analysis.

3.4 Sample Size and Sampling Procedures

The study employed simple random sampling technique to select farmers from each ward in Njoro Sub-County. The sample size is considered the major part of all statistical analyses. The computation of the appropriate sample size is generally considered the most important and the most difficult step in statistical study. The sample size plays a crucial role in those cases of statistical studies where the statistical studies like sample survey, experiments, observational studies, etc. are involved. The sample size to be employed for the identified target population will scientifically be computed through the Krejcie and Morgan's (1970) formulae.

$$S = \frac{X^2NP\ (1-P)}{d^2\ (N-1) + X^2P\ (1-P)}$$

Where:

S = *Required Sample size*

X = *Z value (e.g. 1.96 for 95% confidence level)*

N = *Population Size*

P = *Population proportion (expressed as decimal) (assumed to be 0.5 (50%)*

d = *Degree of accuracy (5%), expressed as a proportion (.05); It is margin of error*

Source: (Krejcie & Morgan, 1970)

3.5 Research Instruments

The researcher used the interview schedules.

3.5.0 Interviews Schedules

An interview is a conversation between two or more people where questions are asked by the interviewer to elicit facts and statements from the interviewee. This method is flexible, more explanatory in nature; firsthand information is collected to ensure the research achieves its objectivity (Weiss, 1994).

CHAPTER FOUR: DATA ANALYSIS AND PRESENTATION OF FINDINGS.

4.0 INTRODUCTION.

The chapter consist of data analysis, presentation and interpretation of the findings the study.

To understand the application domain daily Njoro KMD historical dataset on minimum/maximum temperatures and precipitations and annual Kenyan Ministry of Agriculture crop production dataset from Njoro district was obtained. The selection of samples of the datasets to use in analysis was done based on picking of data range without much of missing data. Data was analyzed, summarized and presented in form of tables.

4.1 Results for objective 1

The researcher was interested in investigating the factors that affect crop production directly in the area. The findings were grouped into Climate Change and Soil PH. There were other factors like farm input, market information and other activities which affect the production indirectly.

4.1.0 Climate

One of the factors affecting crop production is climate. The factor was divided into two sub categories namely rainfall (precipitation) and temperature (maximum and minimum). The average amount of rainfall was calculated for the long rain season in the area which is from the Month of March to July a period five months. The average maximum and minimum temperature was calculated for the whole year.

Table 4 Rainfall amount in mm between Mar-July Source (KALRO Njoro)

Year	March	April	May	June	July	Average
1996	102.0	28.1	57.1	126.0	166.0	96.0
2000	0.8	55.9	34.2	86.2	65.5	48.5
2006	68.0	122.1	98.3	65.7	50.3	80.9
2016	49.5	244.0	125.0	105.5	98.8	124.7

Table 5 Temperature in degree Celsius between Mar-July (Source KALRO Njoro)

Year	March		April		May		June		July		Average
Temperature	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	
1996	25.8	10.5	25.0	9.5	25.0	9.3	20.3	9.8	22.3	8.5	
2000	25.1	8.5	22.0	9.5	23.0	10.0	22.5	7.8	21.5	9.0	
2006	22.0	10.8	21.0	9.5	21.5	8.50	22.8	8.5	21.3	8.8	
2016	26.5	9.4	23.9	10.4	23.3	11.2	22.	10.5	20.9	10.1	

4.1.1 Soil PH

The researcher sought to find out the other factor affecting crop production which is soil PH. Soil pH is a measure of soil acidity or alkalinity. It is an important indicator of soil health. It affects crop yields, crop suitability, plant nutrient availability, and soil micro-organism activity which influence key soil processes.

Table 6 Soil PH range Source (Ministry of Agriculture Njoro)

Denomination	**PH Range**
<3	Ultra acidic
4.0-5.5	Strongly Acidic
5.6-6.0	Medium Acidic
6.1-6.5	Slightly acidic
6.6-7.3	Neutral
7.4-7.8	Slightly alkaline
7.9-8.4	Medium Alkaline
8.5-9.0	Strongly alkaline
>9	Very strongly alkaline

Other factors that affect crop production include the size of the land, Planting season and the geographical area.

Farm	**Season**	**Hectares**
Lower	Long	Large

Lower	Short	Large
Upper	Short	Small
Upper	Short	Large
Lower	Short	Large

4.2 Results for Objective 2

The researcher sought to find out the appropriate data mining methods used to build predictive model for crop production. This was achieved by the use of literature survey. The researcher found out that the appropriate data mining method(s) used in building predictive model for crop production is Classification.

Table 7 Appropriate Classification Algorithms and Application

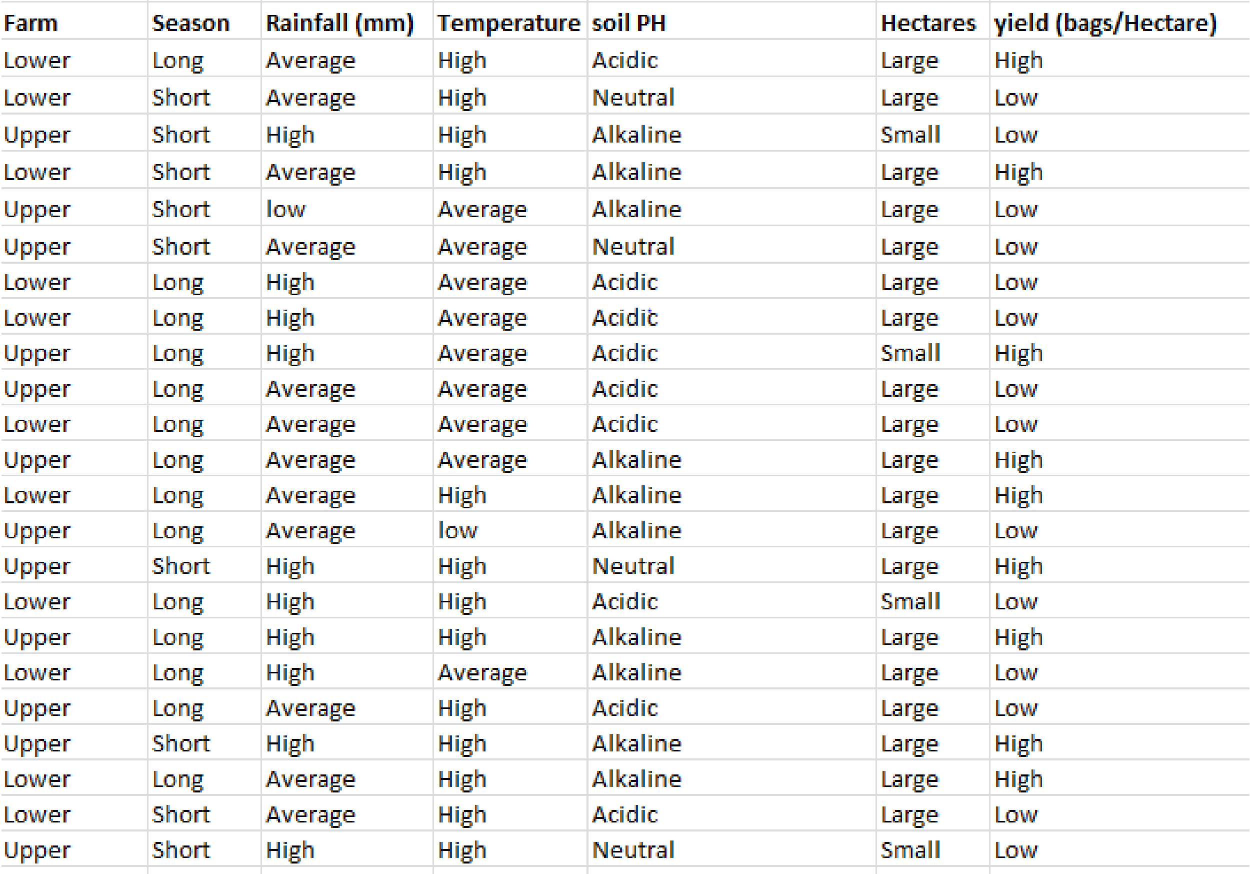
Name	Appropriateness	Application
Artificial Neural Network (MLP)	Data mining framework succeeded in predicting the annual needed amount of main strategic crops up to the year 2020.	data mining techniques to enhance decisions in agriculture sector applied in the MoA in Egypt (Khedr, Kadry, & Walid, 2015)
Rules (PART)	Improvement from previous researchers as the solution contributes to agricultural drought prediction with an emphasis specifically on agricultural drought.	Prediction of agricultural drought in Taita Taveta (S.Mwanjele,2015)
Isotonic Regression	Improvement from previous researchers as the solution contributes to agricultural drought prediction with an emphasis specifically on agricultural drought.	Prediction of agricultural drought in Taita Taveta (S.Mwanjele,2015)

Lbk	Improvement from previous researchers as the solution contributes to agricultural drought prediction with an emphasis specifically on agricultural drought.	Prediction of agricultural drought in Taita Taveta (S.Mwanjele,2015)

4.3 Results for Objective 3

Figure 5 depicts the data set that is used to create the predictive model. The data is already cleaned and transformed for model creation.

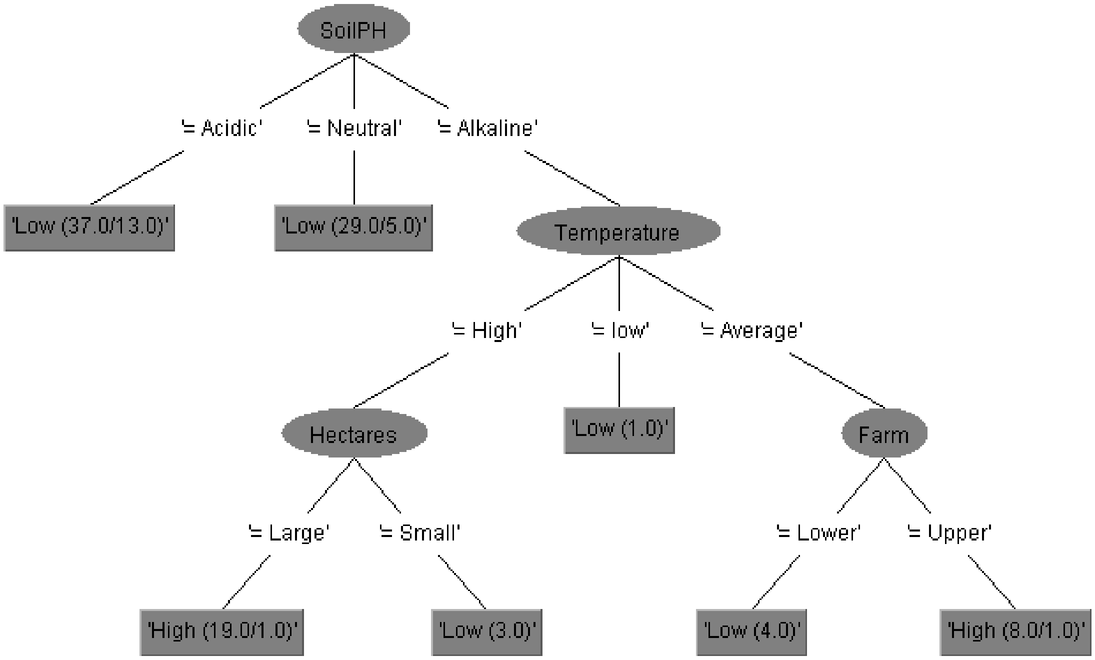
Figure 5 the created Data Set to be mined using WEKA software



Farm	Season	Rainfall (mm)	Temperature	soil PH	Hectares	yield (bags/Hectare)
Lower	Long	Average	High	Acidic	Large	High
Lower	Short	Average	High	Neutral	Large	Low
Upper	Short	High	High	Alkaline	Small	Low
Lower	Short	Average	High	Alkaline	Large	High
Upper	Short	low	Average	Alkaline	Large	Low
Upper	Short	Average	Average	Neutral	Large	Low
Lower	Long	High	Average	Acidic	Large	Low
Lower	Long	High	Average	Acidic	Large	Low
Upper	Long	High	Average	Acidic	Small	High
Upper	Long	Average	Average	Acidic	Large	Low
Lower	Long	Average	Average	Acidic	Large	Low
Upper	Long	Average	Average	Alkaline	Large	High
Lower	Long	Average	High	Alkaline	Large	High
Upper	Long	Average	low	Alkaline	Large	Low
Upper	Short	High	High	Neutral	Large	High
Lower	Long	High	High	Acidic	Small	Low
Upper	Long	High	High	Alkaline	Large	High
Lower	Long	High	Average	Alkaline	Large	Low
Upper	Long	Average	High	Acidic	Large	Low
Upper	Short	High	High	Alkaline	Large	High
Lower	Long	Average	High	Alkaline	Large	High
Lower	Short	Average	High	Acidic	Large	Low
Upper	Short	High	High	Neutral	Small	Low

Figure 6 shows the decision tree created using WEKA software, using J48 algorithm. The results as depicted on the tree indicate that for a farmer to have high yields. The soil has to be alkaline, the temperature could either be high or average and the size of the farm should be large and be on the upper part of Njoro.

Figure 6: Decision tree



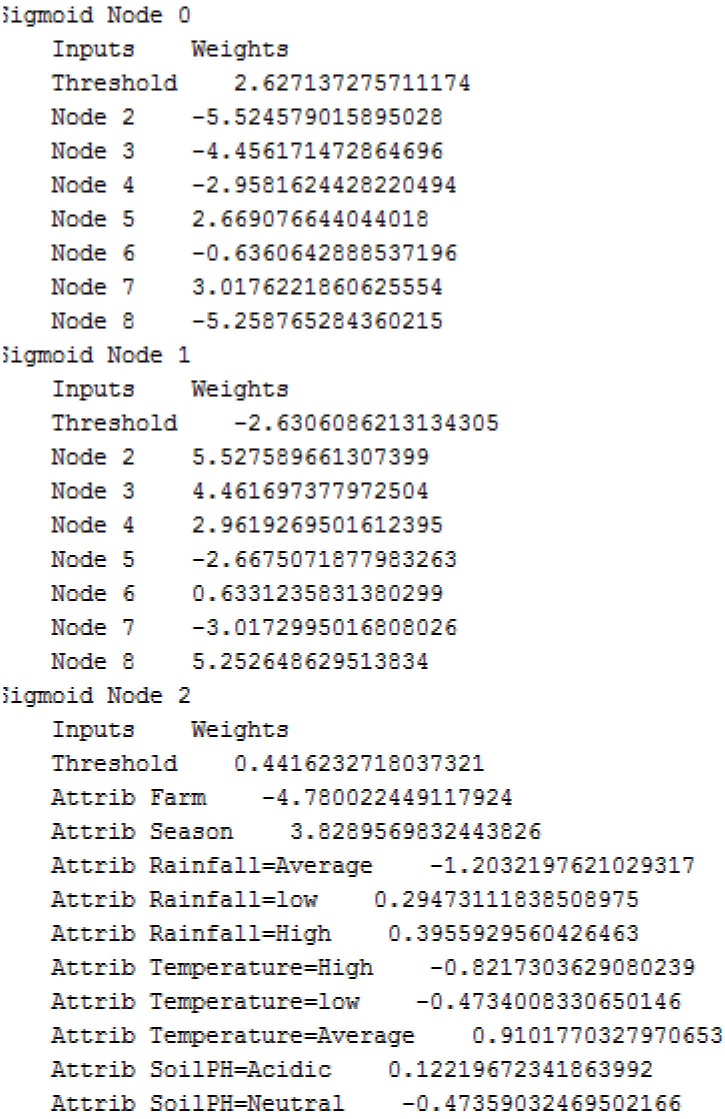
From the above tree, the researcher found out that;

If the soil PH =Alkaline, temperature=Average, Farm=Upper or soil PH=Alkaline, temperature=High, hectares= Large, then the farmer will get high yields.

If the soil PH=Alkaline, temperature=Average, Farm=Lower or soil PH=Alkaline, temperature=High, Hectares=Low, or if soil PH=Alkaline, Temperature=Low, then the yields will be low.

If the soil PH=Acidic or Soil PH= Neutral, then yields will be Low.[1]

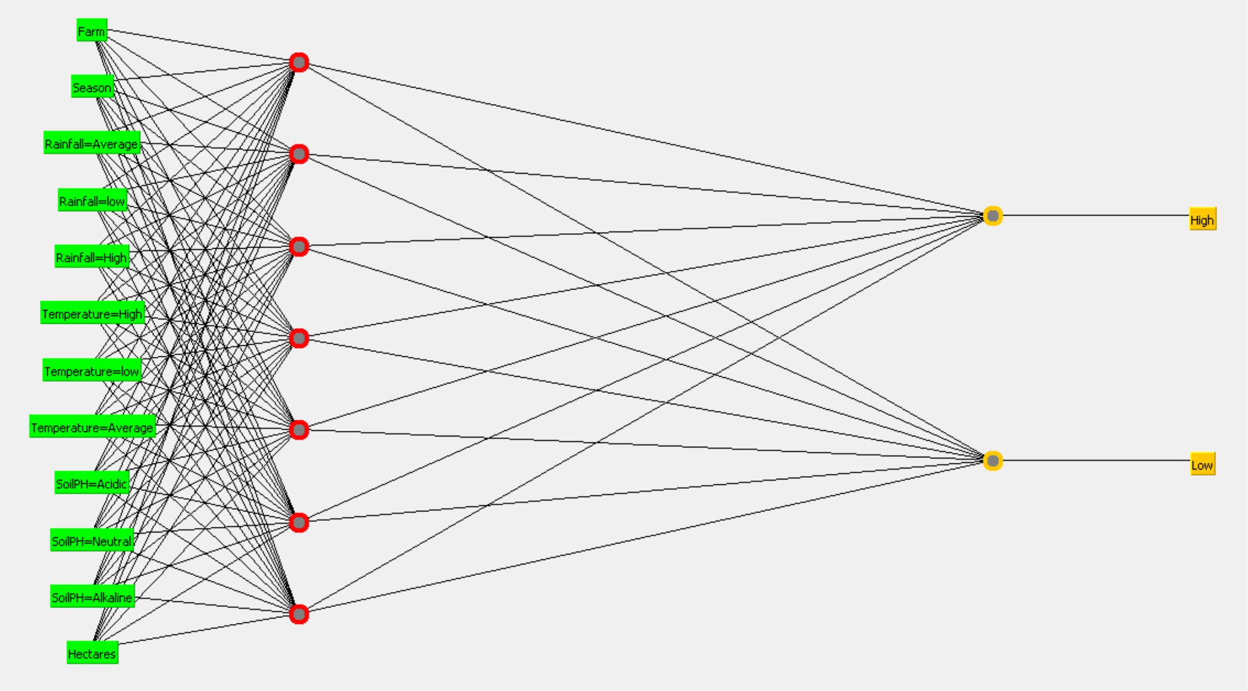
Figure 7: Multi-Layer Perceptron



```
Sigmoid Node 0
    Inputs    Weights
    Threshold    2.627137275711174
    Node 2    -5.524579015895028
    Node 3    -4.456171472864696
    Node 4    -2.9581624428220494
    Node 5    2.669076644044018
    Node 6    -0.6360642888537196
    Node 7    3.0176221860625554
    Node 8    -5.258765284360215
Sigmoid Node 1
    Inputs    Weights
    Threshold    -2.6306086213134305
    Node 2    5.527589661307399
    Node 3    4.461697377972504
    Node 4    2.9619269501612395
    Node 5    -2.6675071877983263
    Node 6    0.6331235831380299
    Node 7    -3.0172995016808026
    Node 8    5.252648629513834
Sigmoid Node 2
    Inputs    Weights
    Threshold    0.4416232718037321
    Attrib Farm    -4.780022449117924
    Attrib Season    3.8289569832443826
    Attrib Rainfall=Average    -1.2032197621029317
    Attrib Rainfall=low    0.29473111838508975
    Attrib Rainfall=High    0.3955929560426463
    Attrib Temperature=High    -0.8217303629080239
    Attrib Temperature=low    -0.4734008330650146
    Attrib Temperature=Average    0.9101770327970653
    Attrib SoilPH=Acidic    0.12219672341863992
    Attrib SoilPH=Neutral    -0.47359032469502166
```

[1] Benchmark values for yield per hectare: <15bags= Low, 15 bags and above=High
Bench mark values for the size of land:2-5 hectares=small, 6 and above=large
For the Soil PH the values were averaged

Figure 8 Multi-Layer Perceptron Network

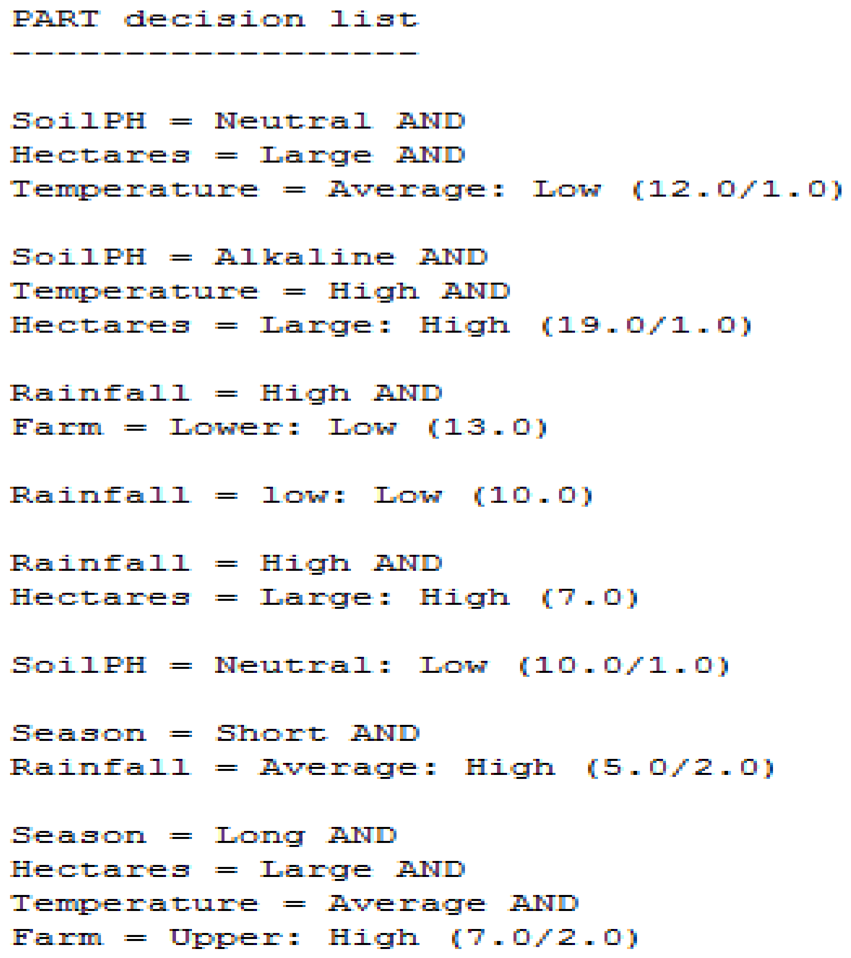


In the results above, there are 101 cases (instances) with 6 input variables namely; Farm, Season, Rainfall, Temperature, Soil PH and Hectares (size of the farm). Also there is one (1) output variable Yield. The Neural Network is using the given values of the 6 input variables to predict Yield.

There are 7 hidden nodes labeled Sigmoid Node 2..8. The weights are given for each variable that feeds into each Sigmoid Node, plus the Threshold weight, which is used to give some input to the output neuron in case of some problem with the other weights.

The fit of the network training to the actual values of Yield was about r = 0.61, which is not too bad of a correlation. One way to think of the weights is like a slope value. A high positive (+) weight means a high correlation between the variable and the outcome. A high negative (-) weight means a high negative correlation. A near zero weight means that variable has little or no effect on the outcome.

Figure 9: PART Rule



```
PART decision list
------------------

SoilPH = Neutral AND
Hectares = Large AND
Temperature = Average: Low (12.0/1.0)

SoilPH = Alkaline AND
Temperature = High AND
Hectares = Large: High (19.0/1.0)

Rainfall = High AND
Farm = Lower: Low (13.0)

Rainfall = low: Low (10.0)

Rainfall = High AND
Hectares = Large: High (7.0)

SoilPH = Neutral: Low (10.0/1.0)

Season = Short AND
Rainfall = Average: High (5.0/2.0)

Season = Long AND
Hectares = Large AND
Temperature = Average AND
Farm = Upper: High (7.0/2.0)
```

While using the PART rule, the researcher also found some rules that were interesting. In one of the scenario, the researcher found one rule stating that if the soil PH=Neutral, Hectares=Large and Temperature=Average, then the farmer will get low yields in that particular planting season.

In another case, if the planting season=Long and Hectares=Large, and Temperature=Average, and the Farm=Upper, then yields during that season will be high which is an advantage to the farmer.

4.4 Results for objective 4

To evaluate the models the following measures were used to test their accuracy.

Kappa Statistic: is used to measure how strongly data items in the same class resemble each other. That is, the level of intra-class correlation. It is similar to correlation coefficient.

Table 8 Kappa Statistic Measurements and Interpretation

Measurement	Interpretation
0.0	Complete disagreement (do not resemble each other).
0.40 to 0.59	Moderate agreement
0.60 to 0.79	Substantial agreement
above 0.80	Outstanding agreement
1.0	Complete agreement (items strongly resemble each other)

Mean Absolute Error (MAE): Can be defined as sum of absolute errors divided by number of predictions. MAE measures set of predicted value to actual value i.e. how close a predicted model to actual model.

Root Mean Squared Error (RMSE): Is defined as square root of sum of squares error divided number of predictions. RMSE measures the differences between values predicted by a model and the values actually observed. Small value of RMSE means better accuracy of model.

F-Measure: Measures both the Precision and recall. It is used to provide a single measurement of model accuracy combining both precision and recall.

ROC Area/Curve: Receiver Operating Characteristic (ROC) curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied.

A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system:

0.90-1 = excellent (A)

0.80-0.90 = good (B)

0.70-0.80 = fair (C)

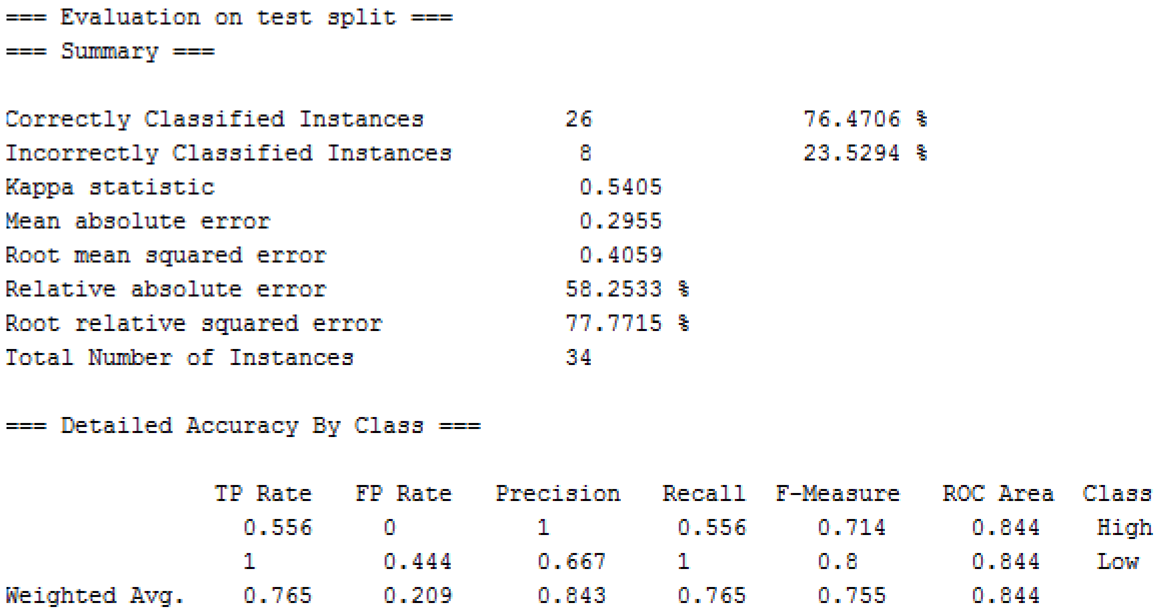
0.60-0.70 = poor (D)

0.50-0.60 = fail (F)

A value near 0.5 means the lack of any statistical dependence.

The evaluation results of the decision tree, PART rule and Multi-layer Perceptron are as shown below

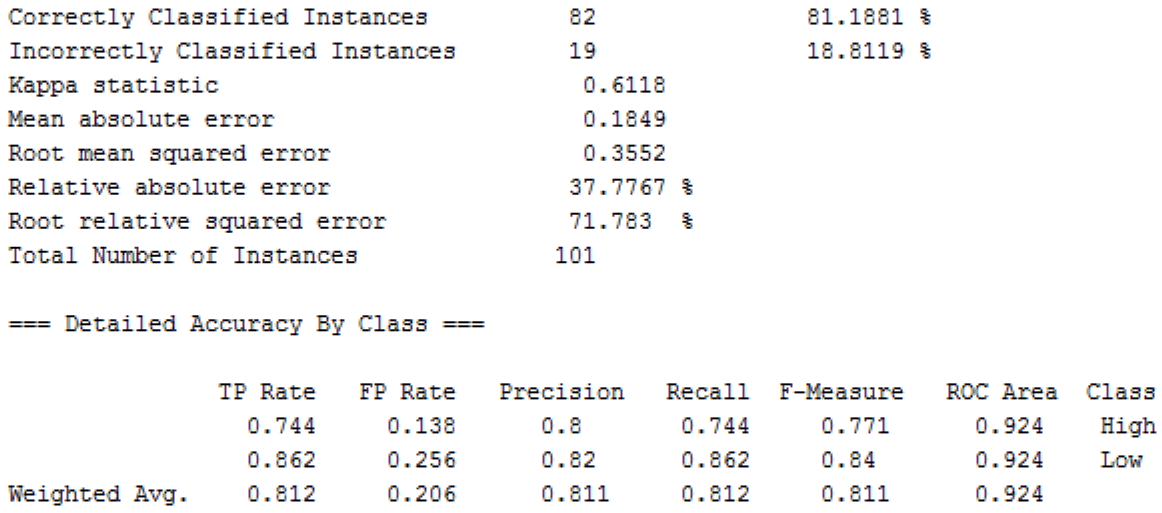
Figure 10 J48 Evaluation Results



```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          26               76.4706 %
Incorrectly Classified Instances         8               23.5294 %
Kappa statistic                          0.5405
Mean absolute error                      0.2955
Root mean squared error                  0.4059
Relative absolute error                 58.2533 %
Root relative squared error             77.7715 %
Total Number of Instances               34

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.556     0         1          0.556     0.714      0.844    High
                 1         0.444     0.667      1         0.8        0.844    Low
Weighted Avg.    0.765     0.209     0.843      0.765     0.755      0.844
```

From figure 10, there are 26 instances which are classified correctly and 8 cases incorrectly classified. The Kappa statistic (measures the agreement of prediction with the true class) of 0.54 indicates that there is a moderate relationship/agreement of the items in the class. The researcher also found that the ROC area is 0.844, which indicates that the model is good.

Figure 11: Multi-Layer Perceptron evaluation results

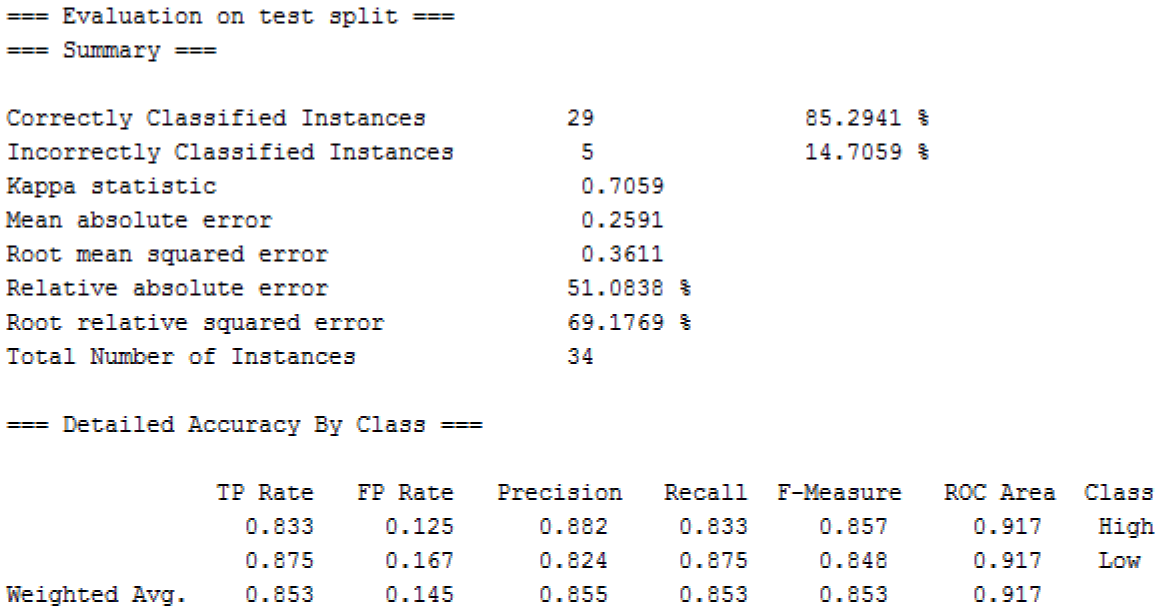


```
Correctly Classified Instances          82               81.1881 %
Incorrectly Classified Instances        19               18.8119 %
Kappa statistic                          0.6118
Mean absolute error                      0.1849
Root mean squared error                  0.3552
Relative absolute error                 37.7767 %
Root relative squared error             71.783  %
Total Number of Instances              101

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.744     0.138     0.8         0.744     0.771       0.924    High
                 0.862     0.256     0.82        0.862     0.84        0.924    Low
Weighted Avg.    0.812     0.206     0.811       0.812     0.811       0.924
```

From the above results, there are 82 instances which are classified correctly and 19 cases incorrectly classified. The Kappa statistic of 0.611 indicates that there is a substantial relationship of the items in the class. The Mean Absolute error is 0.184, which indicates that the model, is good when it comes to prediction.

From the evaluation results in figure 11 shows that the model is excellent, since the precision of the model is 0.811, the Recall is 0.812, the F-Measure of the model is 0.811 and the ROC area of the model is 0.924.

Figure 12: PART Rule Evaluation Results



```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          29               85.2941 %
Incorrectly Classified Instances         5               14.7059 %
Kappa statistic                          0.7059
Mean absolute error                      0.2591
Root mean squared error                  0.3611
Relative absolute error                 51.0838 %
Root relative squared error             69.1769 %
Total Number of Instances               34

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.833     0.125      0.882     0.833     0.857      0.917    High
                 0.875     0.167      0.824     0.875     0.848      0.917    Low
Weighted Avg.    0.853     0.145      0.855     0.853     0.853      0.917
```

From the above results, there are 29 instances which are classified correctly and 5 cases incorrectly classified. The Kappa statistic of 0.705 indicates that there is a relationship of the items in the class. The Mean Absolute error is 0.259, which indicates that the model is good when it comes to prediction.

From the evaluation results above it shows that the model is also excellent, since the precision of the model is 0.855, the Recall is 0.853, the F-Measure of the model is 0.853 and the ROC area of the model is 0.917.

Table 9: Evaluation results for different Algorithms used in the study

Evaluation Results	**Multi-Layer Perceptron**	**PART Rule**	**Decision Tree (J48)**
Kappa Statistics	0.6118	**0.7059**	0.5405
Mean Absolute Error	**0.1849**	0.2955	0.2955
Root Mean Square Error	**0.355**	0.361	0.405
Relative absolute Error	**37.77%**	51.08%	58.25%
F-Measure	0.811	**0.853**	0.755
ROC Area	**0.924**	0.917	0.844

From the results on table 7, the researcher found out that the ROC area of Multi-layer Perceptron, and that of PART rule was 0.924 and 0.917 respectively, meaning that the model was excellent. For J48 algorithm, the ROC area was 0.844, which shows that the model was good.

4.5 Discussion of Results

In this study, three data mining algorithms namely J48, ANN and PAT rule were used and the results compared for the researcher to know the best model that can be used in yield prediction to enhance food security in the study area. The three algorithms when tested using a 10-fold Cross-Validation, gave the researcher different results when evaluating the models. In this case ANN provided the best prediction accuracy with a Root Mean Square Error (RMSE) of 0.355 and an ROC of 0.924.

(Khedr et al. 2015) in their research, proposed the use of ANN (Artificial Neural Network) via WEKA using Multilayer Perceptron(MLP) function as a data mining technique for prediction. The RMSE of their model was given as 0.371 and the ROC of the same model was 0.917. The results showed that the data mining framework succeeded in predicting the annual needed amount of main strategic crops up to the year 2020.

The quality metrics were ranked, from the best to the worst, according to Kappa statistic, Mean Absolute Error, Root Mean Square Error, Relative Absolute Error, F-Measure and ROC Area, in the following order: ANN (Multi-Layer Perceptron), PART Rule and J48. In this study, ANN, give us the best mechanism to apply in order to predict the crop yield and regulate the production by controlling the independent factors.

CHAPTER FIVE: SUMMARY, CONCLUSIONS, RECOMMENDATIONS

5.0 Introduction

This chapter discusses the summary of the findings of the sturdy, conclusions made and the recommendations that are to be used to further this work

5.1 Summary

1. The findings in this study are important as we try to develop strategies and policies to address the food security issue.
2. Identifying the factors affecting crop and food production is very significant as it enables agricultural organizations and farmers plan to develop ways of ensuring that citizens are food secure.
3. The findings presented in this study, along with others, can inform the development and implementation of strategies that can be used to improve agriculture and encourage people to invest in farming to ensure the study area and the whole country of Kenya there is minimal or no shortage of food..

5.2 Conclusions

Agriculture is the most influencing and significant application area particularly in the developing countries like Kenya. Use of information technology in agriculture can change the situation of decision making and farmers can yield in better way and agricultural organization get idea about yield and they make better policies for famers

Data mining plays a crucial role for decision making on several issues related to agriculture field. This study integrates the work of various authors in one place so it is useful to get information of current scenario of data mining techniques and applications in context to agriculture field and predictive data mining.

The study of crop yield prediction consists of three stages namely, preprocessing, feature reduction and forecast. The study used input data as real world data. Real world data is often incomplete, inconsistent, and/or lacking in certain performances or trends, and is likely to contain many inaccuracies. Food production in the country can be improved with automatic prediction of crop yield based on the reliable variables. The system helps farmers to do right things at right time. The productivity gets improved in agriculture with sustained research in the field of spatial data mining to realize precision agriculture.

5.3 Recommendations

The researcher recommends further studies and more research to be conducted in the livestock and fisheries departments to enhance the production of meat, eggs and dairy products in order to ensure there is enough for the citizens and export.

Since the study used three algorithms (J48, PART Rule and MLP), the researcher recommends further studies using different algorithms and different mining and analytics tools that are available.

REFERENCES

Awuor VO, Ogola JS (1997). Effects of climate change on agriculture. In Ogola JS, Abira MA, Awuor VO (eds), Potential impacts of climate change in Kenya. Climate African Network, Nairobi. pp. 95-115.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P.J., Rotter, R.P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P.K., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R., Heng, L., Hooker, J., Hunt, L.A., et al., 2013. Uncertainty in simulating wheat yields under climate change. Nature Climate Change DOI: 10.1038/nclimate1916

Basso B., Cammarano D.,Troccoli A., Chen D., Ritchie J.T.. 2010. Long-term wheat response to nitrogen in a rainfed Mediterranean environment: Field data and simulation analysis. *European Journal of Agronomy 33 (2010) 132–138*

Bondeau, A., Smith, P. C., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., Gerten, D., Lotze-campen, H., Müller, C., Reichstein, M., and Smith, B. 2007. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. Global Change Biology 13, 679-706.

Bourdieu, Pierre. 1984. Distinction: A Social Critique of the Judgement of Taste. New York: Cambridge University Press

Ch.N.Santhosh Kumar, V. Sitha Ramulu, K.Sudheer Reddy, Suresh Kotha and Ch. Mohan Kumar, Aug 2012 "Spatial Data Mining using Cluster Analysis", International journal of computer science and information technology, Vol.4, No.4, pp. 71-77,.

Challinor, A.J., Wheeler, T.R., 2008. Crop yield reduction in the tropics under climate change: Processes and uncertainties. Agric. Forest Meteorol. 148, 343–356.

De Groote, H., G. Owuor, C. Doss, J. Ouma, L. Muhammad and K Danda (2012). The Crop Green Revolution in Kenya Revisited. e Journal of Agricultural and Development Economics , 32-49.

Doraiswamy, P.C., Moulin, S., Cook, P.W., Stern, A., 2003. Crop Yield Assessment from Remote Sensing. Photogrammetric Engineering & Remote Sensing 69, 665–674.

FAOSTAT, (2006). FAO Statistical Data [online]. FAO. Available at http://www.fao.org/faostat/foodsecurity

FAOSTAT. (2010) Food and Agricultural Organization of the United Nations (FAO), FAO Statistical Database, 2010, from http://faostat.fao.org

Gommes, R., 1998. List of variables used for crop modelling purposes by the Space Applications Institute/MARS, pp. 313-326. In: Rijks, D., Terres, J.M., Vossen, P. (Eds.).

Gudeman, Stephen. 1986. Economics as Culture. London: Routledge.

Hatfield, J.R., 1983. Remote Sensing estimators of potential and actual crop yield. Remote Sens. Environ. 13, 301–311.

Heisey, P.W., and Smale, M., (1995). Crop Technology in Malawi: A Green IPCC., (2007). New Assessment Methods and the Characterisation of Future Conditions: In Climate change 2007: Impacts, adaptation and vulnerability,pp: 976. Contribution of working group II to the fourth assessment report of the Intergovernmental panel on climate change. Cambridge university press, Cambridge, UK.

Hand, D., 2001. *Principles of Data Mining*,

IPCC (2001). Climate change 2001: The scientific basis. Cambridge University Press, UK. p.146.

IPCC (2007). Climate change 2007: Impacts, Adaptation and Vulnerability. Cambridge University Press, UK. pp. 127-131.

Khedr, A.E., Kadry, M. & Walid, G., 2015. Proposed framework for implementing data mining techniques to enhance decisions in agriculture sector Applied case on Food Security Information Center Ministry of Agriculture , Egypt. *Procedia - Procedia Computer Science*, 65(Iccmit), pp.633–642. Available at: http://dx.doi.org/10.1016/j.procs.2015.09.007.

Kogan, F., Salazar, L., Roytman, L., 2012. Forecasting crop production using satellite-based vegetation health indices in Kansas, USA. International Journal of Remote Sensing 33, 2798-2814.

Kogan, F., Kussul, N., Adamenko, T., Skakun, S., Kravchenko, O., Kryvobok, O., Shelestovb, O., Kolotii, A., Kussul, O., Lavrenyuk, A., 2013. Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models. International Journal of Applied Earth Observation and Geoinformation 23, 192–203.

Liu HB, Gou Y, Wang HY, Li HM, Wu W 2014 Temporal changes in climatic variables and their impact on crop yields in southwestern China. Int J Biometeorol 58:1021–1030

Manjula, A. & Narsimha, G., 2016. Crop Yield Prediction with Aid of Optimal Neural Network in Spatial Data Mining : New Approaches. , 6(1), pp.25–33.

Megala S., Hemalatha M. A Novel Datamining Approach to Determine the Vanished Agricultural Land in Tamilnadu, International Journal of Computer Applications (0975 – 8887) Volume 23– No.3, June 2011, pp. 23, 24, 26.

Messages, K.E.Y., 2017. KENYA Food Security Outlook Update Food security typically deteriorates as lean season continues. , pp.2016–2017.

Michael Beery and Gordon S. Linoff, 2009, *Data Mining Techniques : Theory and Practice Course Notes*

Mkhabela, M.S., Bullock, P., Raj, S., Wang, S., Yang, Y., 2011. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agricultural and Forest Meteorology 151, 385–393.

Mugendi DM, Mucheru-Muna M, Mugwe J, Kung'u JB, Bationo A (2007). Improving food production using 'best bet' soil fertility technologies in the Central Highlands of Kenya. In: Bationo et al. (eds). Advances in integrated soil fertility management in Sub-Saharan Africa: Challenges and Opportunities, pp. 345-351.

Mwanjele, S.M., 2015. INTELLIGENT PREDICTION OF AGRICULTURAL DROUGHT USING CLASSIFICATION ALGORITHMS.

Nekesa AO (2007). Effect of Minjingu phosphate rock and agricultural lime in relation to maize, groundnut and soybean yields on acid soils of western Kenya. M. Phil Thesis. Moi University Eldoret, Kenya. P. 79.

,Paz, J.O., Fraisse, C.W., Hatch, L.U., Garcia y Garcia, A., Guerra, L.C., Uryasev, O., Bellow, J.G., Jones, J.W., Hoogenboom, G., 2007. Development of an ENSO-based irrigation decision support tool for peanut production in the southeastern US. Comput. Electron. Agric. 55, 28–35.

Raorane A. and Kulkarini F. Data Mining: An effective tool for yield estimation in the agricultural sector, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 1, Issue 2, July – August 2012, pp.1, 2, 4.

Salame F. Applying data mining techniques to evaluate applications for agricultural loans, University of Nebraska-Lincoln, 2011, pp. 164,165,166.

Wang Y., Tang J. and Cao W. Grey prediction model-based food security early warning prediction, School of Business, Jiangnan University, Wuxi, China, Grey Systems: Theory and Application Vol. 2 No. 1, 2012, Emerald Group Publishing Limited, 2012, pp. 2.

White J, Hoogenboom G, Wilkens P, Stackhouse P, Hoel J (2011) Evaluation of satellite-based, modeled-derived daily solar radiation data for the continental United States. Agronomy Journal, 103, 1242-1251.