Application of Machine Learning Techniques in Predicting Productivity of Sorghum

By: Tigist Girma

A Thesis Submitted to the

Department of Computing

School of Electrical Engineering and Computing

Presented in Partial Fulfillment for the Degree of Masters of Science in Information System

Office of Graduate Studies

Adama Science and Technology University

Adama

July 2019

Application of Machine Learning Techniques in Predicting Productivity of Sorghum

By: Tigist Girma

Name of Advisor: Teklu Urgessa (PhD)

A Thesis Submitted to the

Department of Computing

School of Electrical Engineering and Computing

Presented in Partial Fulfillment for the Degree of Masters of Science in Information System

Office of Graduate Studies

Adama Science and Technology University

Adama

July 2019

DECLARATION

I declare that this thesis is my original work and has not been presented for a degree in any other Universities, and all sources of material used for the study have been accordingly acknowledged.

____

Tigist Girma Woldetsadik

July, 2019

This Thesis work has been submitted for examination with my approval as a University advisor.

____

Teklu Urgessa (PhD)

July, 2019

Application of Machine Learning Techniques in Predicting Productivity of Sorghum

By:

Tigist Girma

Name and Signature of Members of the Examining Board

Name	Title	Signature	Date
Dr. Tilahun Melak	Internal Examiner	______________	______________
Dr. Kula Kekeba	External Examiner	______________	______________
Dr. Mesfin Abebe	Chairperson	______________	______________

DEDICATION

I would like to dedicate this thesis to my beloved father Dr. Girma Woldetsadik who passed away during my study. I Love You Glory! May God put your soul in heaven.

ACKNOWLEDGMENT

First of all, I would like to thank God for allowing and helping me to do this thesis. Next, I am gratefully acknowledging Adama Science and Technology University for providing basic necessities for my study. I also would like to express my gratitude to my advisor Dr. Teklu Urgessa. I also express my gratitude for Melkassa Agricultural Research Center director, Dr. Mohammed Yesuf for accepting me to do my study in sorghum crop. Next, the deepest appreciation goes to Melkassa Agricultural Research Center Sorghum and Millet department head Dr. Taye Tadesse for allowing me to use the data for my study. I gratefully appreciate Sorghum and Millet department experts for supporting me in materials during my study.

My special thanks go to my mother W/ro. Wegayehu Tilahun and my younger brother Biruh Girma, for loving, caring and supporting me. Both of you are my reason behind my success. Finally, I would like to thank others who participated in this study.

LIST OF ABBREVISTIONS

AEZs	Agro-ecological zones
ANN	Artificial Neural Network
CSM	Crop Selection Method
DZARC	Debrezeit Agricultural Research Center
EBI	Ethiopian Biodiversity Institute
EIAR	Ethiopian Institute of Agricultural Research
ha	hectare
HARC	Holetta Agricultural Research Center
IAR	Institute of Agricultural Research
KB	Kilobytes
LASSO	Least Absolute Shrinkage and Selection Operator
LGP	Length of Growing Period
m.a.s.l	meter above sea level
MARC	Melkassa Agricultural Research Center
masl	meter above sea level
ML	Machine Learning
mm	millimeter
MoU	Memorandum of Understanding
MSE	Mean Square Error
NABI	National Agricultural Biotechnology Laboratory
OLS	Ordinary Least Squares
PAS	Plant Aspect Ratio

pH	potential of Hydrogen
RMSE	Root Mean Square Deviation
SIDA	Swedish International Development Agency
SVM	Support Vector Machine
TNSRC	Tepi Species Research Center
USSR	Union of Soviet Socialist Republics

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In agriculture sector, machine learning techniques play an important role for predicting productivity of agricultural products. Some of applications of machine learning in agriculture is mentioned in Appendix B. Existence of yield constraints in sorghum production decreases productivity. Drought, insect, bird, disease, inadequate improved sorghum varieties are some of constraints found in sorghum production. To address these problems after trail, correlation and regression models are used in the sample data of sorghum. To answer research question of identifying factors that increase and decrease productivity of sorghum from sorghum sample data, Pearson correlation is used. Pearson correlation shows linear relationship between dependent (yield per a plot) and independent variables (replication, varieties, heads per a plot, drought score, emergency day, flowering day, maturity day, bird score, disease score, insect score and plant height). Positive correlation means if independent variable increases the dependent variable also increases while negative correlation means if independent variable increase then dependent variable will decrease. Correlation of yield per a plot with replication, varieties, heads per a plot and plant height shows positive correlation. And correlation of yield per a plot with drought score, emergency day, flowering day, maturity day, bird score, disease score and insect score tell negative correlation in RStudio.

To answer how machine learning can be effectively used to predict productivity of sorghum, ridge regression, LASSO and elastic net regression are used in RStudio. Ridge regression describe variables with sum of squares of coefficient while least absolute shrinkage and selection operator define variables with sum of absolute values of coefficient and elastic net regression combines ridge regression and least absolute shrinkage and selection operator. Ridge regression, LASSO and Elastic net regression of root mean square deviation (RMSE) and mean square error (MSE) in RStudio indicates importance of variable as percentage value. Ridge regression RMSE variable are maturity day (100%), flowering day (79.5%) and drought score (32.9%) and MSE variables are maturity day (100%), flowering day (79.5%) and drought score (32.9%). For LASSO RMSE variables are maturity day (100%), flowering day (76%) and drought score (35.2%) and MSE variables are maturity day (100%), flowering day (76%) and drought score (35.2%). Elastic net regression RMSE variables are drought score (100%), maturity day (91.3%)

and flowering day (69.8%) and MSE variables drought score (100%), maturity day (91.5%) and flowering day (69.9%). Elastic net regression is selected than ridge and LASSO as it has better features than others. Drought score, maturity day and flowering day are selected as most important factors used in machine learning for predicting productivity of sorghum.

Therefore, machine learning is used to address a problem by applying different techniques in problem related data. In agriculture research centers, sorghum production constraints can be identified from historical data by applying different techniques including machine learning to problem related data.

Keywords: Sorghum Production, Machine Learning, Correlation, Ridge regression, LASSO, Elastic net regression

CHAPTER ONE

INTRODUCTION

1.1 Background

1.1.1 Sorghum Production in Ethiopia

Agriculture is the backbone of the Ethiopian economy. This particular sector determines the growth of all other sectors and consequently, the whole national economy. On average, crop production makes up 60 percent of the sector's outputs, whereas livestock accounts for 27 percent, and other areas contribute 13 percent of the total agricultural value added. The sector is dominated by small-scale farmers who practice rain-fed mixed farming by employing traditional technology, adopting a low input and low output production system. The land tilled by the Ethiopian small-scale farmer accounts for 95 percent of the total area under agricultural use and these farmers are responsible for more than 90 percent of the total agricultural output [4].

Cereals like Sorghum, Millets, Wheat, Maize and Rice are major staple foods of the most population [5].

Sorghum is one of the major staple crops grown in the poorest and most food insecure regions of Ethiopia. The crop is typically produced under adverse conditions such as low input use and marginal lands. It is well adapted to a wide range of precipitation and temperature levels and is produced from sea level to above 2000 m.a.s.l. Its drought tolerance and adaptation attributes have made it the favorite crop in drier and marginal areas. Ethiopian is often regarded as the center of domestication of sorghum because of the greatest genetic diversity in the country for both cultivated and wild forms. The large improvement in sorghum production is driven by both land expansion and yield improvement: yield increased from an average of 1.4 tonne/ha in 2004/05 to an average of 2.1 tonne/ha in 2010/11, increasing by 50 percent, while area under sorghum production increased by 51 percent (from 1.2 million ha in 2004/05 to 1.9 million ha in 2011. It should be noted that FAOSTAT yield and production figures during the period 2007 to 2011 are lower than the government. The major sorghum production regions of the country are Oromia at 38.5%, Amhara (32.9%), Tigray (14.1%), and Southern Nations and Nationalities People (S.N.N.P.) region (7.6%) [8].

Pearl millet is a climate hardy crop which is grown in harsh conditions, but as a subsistence crop. Harvested from an area of 20 m ha in the semi-arid regions of Africa pearl millet contributes 19% area to cereal production [5].

Maize is a major staple food crop grown in diverse agro-ecological zones and farming systems, and consumed by people with varying food preferences and socio-economic backgrounds in sub-Saharan Africa (SSA). An estimated 208 million people in SSA depend on maize as a source of food security and economic wellbeing. Maize occupies more than 33 million ha of SSA's estimated 200 million ha of cultivated land.The crop covers nearly 17% of the estimated 200 million ha cultivated land in SSA, and is produced in diverse production environments and consumed by people with varying food preferences and socio-economic backgrounds. The top 20 countries, namely South Africa, Nigeria, Ethiopia, Tanzania, Malawi, Kenya, Zambia, Uganda, Ghana, Mozambique, Cameroon, Mali, Burkina Faso, Benin, DRC, Angola, Zimbabwe, Togo, and Cote d'Ivoire, account for 96% of the total maize production in SSA. The planted land of maize and grain production have increased significantly across regions in SSA since 1961 [5].

Rice has become a highly strategic and priority commodity for food security in Africa. Consumption is growing faster than that of any other major staple on the continent because of high population growth, rapid urbanization and changes in eating habits. It is the single most important source of dietary energy in West Africa and the third most important for Africa as a whole. Rice consumption is increasing faster than that of any other food staple in Africa at about 5.5% per year (2000–2010 average). This increase is driven by urbanization and related changes in eating habits, and population growth. Rice consumption was approximately 24 million tonnes (Mt) per year in SSA in 2012. With only about 60% of rice consumption being satisfied by domestic production, rice imports stand at 10–12 Mt. This is equivalent to one-third of the rice traded on the world market. The percentage of Africa's population living in cities is expected to grow to 48% by 2030. Faced with the threat of shortages in the supply of rice, the need to increase domestic production has become a top priority to contribute to food security in Africa. Demand for milled rice in SSA is expected to increase by 30 Mt by 2035, equivalent to an increase of 130% in rice consumption [5].

Wheat is grown on around 10 million ha in Africa. It is a major staple crop for several countries and an imported commodity in all of Africa. In all African countries, wheat consumption has been steadily increasing during the past 20 years as a result of growing population, changing food preferences and a strong urbanization trend which has led to a growing ‘food gap’ in all regions, largely met by imports. In 2013 alone, African countries spent over $12 billion dollars to import more than 40 million metrics tons of wheat, equating to about a third of the continent’s food imports. During 2010-2013, the average quantity of wheat import in SSA was about 17.5 Million metric tons per annum, which is close to 80% of the total domestic wheat consumption in these countries. Each year, less than 30% of wheat consumption in the region is covered from domestic production. In addition to the increasing trend in volume of wheat import in SSA, wheat prices (both producers’ and world market prices) have increased substantially over the last half-decade. Domestic price volatility is very high. The average wheat productivity in SSA is 1.7 tons/ha, nearly 50% below the world average. Yield data from experimental stations and crop models indicate a very high yield potential, among the highest reported for spring wheats. Therefore, the yield gap between yield potential and average farm yields is significant, often greater than 5-fold. This yield gap can be filled through use of improved technologies (improved varieties/seeds, agronomic practices, fertilizer and pesticides), and better institutional and market arrangements creating incentives to wheat producers and other actors involved in wheat marketing and processing [5].

1.1.2 Overview of Machine Learning

Typically, ML methodologies involves a learning process with the objective to learn from “experience” (training data) to perform a task. Data in ML consists of a set of examples. Usually, an individual example is described by a set of attributes, also known as features or variables. A feature can be nominal (enumeration), binary (i.e., 0 or 1), ordinal (e.g., A+ or B-), or numeric (integer, real number, etc.). The performance of the ML model in a specific task is measured by a performance metric that is improved with experience over time. To calculate the performance of ML models and algorithms, various statistical and mathematical models are used. After the end of the learning process, the trained model can be used to classify, predict, or cluster new examples (testing data) using the experience obtained during the training process. ML tasks are classified into two main categories, that is, supervised and unsupervised learning, depending on the learning signal of the learning system. In supervised learning, data are presented with example inputs and the corresponding outputs, and the objective is to construct a general rule that maps inputs to outputs. In some cases, inputs can be only partially available with some of the target outputs missing or given only as feedback to the actions in a dynamic environment (reinforcement learning). In the supervised setting, the acquired expertise (trained model) is used to predict the missing outputs (labels) for the test data. In unsupervised learning, however, there is no distinction between training and test sets with data being unlabeled. The learner processes input data with the goal of discovering hidden patterns [29].

Machine learning lies at the intersection of computer science, engineering, and statistics and often appears in other disciplines. Machine learning uses statistics. [34] A long-standing problem in statistics is how best to use samples drawn from unknown probability distributions to help decide from which distribution some new sample is drawn. A related problem is how to estimate the value of an unknown function at a new point given the values of this function at a set of sample points. Statistical methods for dealing with these problems can be considered instances of machine learning because the decision and estimation rules depend on a corpus of samples drawn from the problem environment [33].

Nowadays, precision agriculture aims at increasing the productivity and maximizing the yields of a crop. It can benefit the entire crop cycle through an application of the correct amount of inputs (such as water, fertilizers, pesticides or fungicides) at the exact time and the right place, or by detecting diseases in plants [26]. Innovative technologies can be useful to face problems such as environmental sustainability, waste reduction, and soil optimization; the gathering and the analysis of agricultural data, which include numerous and heterogeneous variables, are of considerable interest for the possibility of developing production techniques respectful of the ecosystem and its resources (optimization of irrigation and sowing in relation to soil history and seasonal cycles), the identification of influential and non-influential factors, the possibility of carrying out market analysis in relation to the forecast of future hard-predictive information, the possibility of adapting crops to specific environments, and finally the ability to maximize technological investments by limiting and predicting hardware failures and replacements.[6]

Machine learning is selected as it uses predictive model to solve a problem and make a decision based on results obtained from experimentation. This study shows the effectiveness of machine learning in predicting productivity of sorghum by using different models and relationships of sorghum production with different factors that relates with productivity using correlation.

1.1.3 Food Security in Ethiopia

Ethiopia is the second most populous country in Africa with an estimated population of 94.3 million people in 2013. As indicated by Africa Food Security and Hunger/ Undernourishment Multiple Indicator Scorecard, Ethiopia ranked as first in having the highest number of people in state of undernourishment/ hunger which is 32.1 million people. This makes it, the fourth African country scoring (37.1%) of the population being undernourished/ in hunger. The livelihoods of rural Ethiopian people are highly sensitive to climate. Food insecurity patterns are seasonal and linked to rainfall patterns, with hunger trends declining significantly after the rainy seasons. Climate related shocks affect productivity, hamper economic progress and exacerbate existing social and economic problems. Food insecurity situation in Ethiopia is highly linked up to severe, recurring food shortage and famine, which are associated to recurrent drought. Currently there is a growing consensus that food insecurity and poverty problems are closely related in the Ethiopian context. Droughts and other related disasters (such as crop failure, water shortage, and livestock disease, land degradation, limited household assets, low income) are significant triggers, more important factors which increase vulnerability to food security and undermined livelihoods [1]. The relevance of machine learning is to solve the problem by finding a solution by applying techniques to data that includes information used for solving a problem.

1.2 Motivation

The motivation for this study is existence of crop production experience within family. As the researcher helps the family with crop production experience by inserting paper based manually written trail data to excel format in a computer. With this reason the researcher develops a feeling to choose a trial data. By the time the researcher went to nearby agricultural research center, sorghum department are the only one that uses server-based data storage system. Due to data storage system, sorghum crop is selected for this study. Before the study is accepted in computing department formerly it is proposed in datamining. During the proposal examination, the researcher was informed to use latest technology other than datamining. Machine learning is selected beside statistics used in agricultural research center, the researcher wants to show latest predictive technology in crop production.

1.3 Statement of the Problem and Research Gap

Several production constraints were identified as hindrance for sorghum production and productivity enhancement. The major constraints include drought, striga, insect pests (stalk borer, midge and shoot fly), disease (grain mold, anthracnose and smut), soil fertility decline, inadequate adoption of the existing improved varieties, lack of high yielding and good quality sorghum varieties and post-harvest management practices (grain storage management), storage pest control strategy and identifying alternative uses of sorghum grain [7].

Drought is one of the major yielding limiting factors. It is manifested by delay in onset, dry spell after sowing, drought during critical crop stage ad too early stop. Over 80% of sorghum in Ethiopia is produced under severe to moderate drought stress conditions. Most farmers grow long maturing local landraces some of which take 7 – 8 months to mature further complicating the drought problem. Although extent of yield loss due to drought was no studied in Ethiopia, complete yield loss was observed in some parts of the country (e.g. Mehoi area). The biotic constraint Striga is becoming the major pandemic in most of sorghum growing areas, where soil fertility and moisture stress are limiting factors. Nationally, Striga causes annual yield loss as high as 65 – 70 % and at times leaves plots uncultivable [7].

Other constraint in the sorghum sub sector include soil fertility ad number of other biotic stresses such as birds (Quelea quelea), disease and weeds. Low purchasing power and limited use of fertilizer and other agro inputs such as agro chemicals and seed, ineffective seed systems, lack of markets and poor market information, limited availability of storage, threshing and processing equipment and other value addition technologies are also another major constraint [7]. Therefore, machine learning is selected to address productivity problems as it depends on data stored historically where the technique takes inputs from the data outputs a solution for a problem.

Research Gaps

- Lack of improved varieties with high yield and factor tolerance varieties
- Inadequate of supportive technologies

Research Questions:

1. What are the factors that increase and decrease productivity of sorghum from sorghum trial data?
2. How machine learning can be effectively used to predict productivity of sorghum?

1.4 Objectives

1.4.1 General Objectives

The general objective of this study is to apply machine learning techniques in predicting productivity of sorghum.

1.4.2 Specific Objectives

The specific objectives are: -

- Review and summarized literatures related with machine learning techniques and sorghum crop production for better understanding.
- Discuss about sorghum production and production factors with sorghum experts.
- Develop architecture used for proposed system
- Prepare data for applying machine learning techniques.
- Apply machine learning techniques in prepared data.
- Test and evaluate findings and forward recommendation.

1.5 Scope and Limitations

The scope and limitation of the research work is: -

Scope: -

- The study focuses on factors of sorghum production using machine learning techniques.
- It consists of factors that results in increase and decrease of sorghum production.
- The study shows how machine learning can be effectively used to predict productivity of sorghum.
- Includes applying machine learning techniques for the sorghum data.

Limitation: -

- It is limited to sample sorghum trial data collected in the year of 2011 - 2014 from Melkassa Agricultural Research Center.
- Another limitation for this study is budget.

1.6 Significance of the study

The significance of the study: -

- Benefits agricultural research centers to give more effort for factors that increase and decrease sorghum production after completion of trials.
- Is used to show sorghum experts how effectively machine learning is used in predicting productivity of sorghum.
- Benefits farmers for better production.
- Provides recommendation future works in trial data for machine learning techniques.

1.7 Organization of the Thesis

This thesis consists of five chapters.

First chapter discusses introduction on background with sorghum production in Ethiopia, overview of machine learning and food security in Ethiopia, motivation, statement of the problem and research gap, objectives with general and specific objectives, scope and limitation and significance of study.

Second chapter provides literature review on introduction, origin and domestication of sorghum, advantages of sorghum, characteristics of sorghum races, application of machine learning in agriculture, application of machine learning in predicting crop productivity and related work. Third chapter discusses research methodology on introduction, data collection techniques, data preparation with raw data source and description, data preprocessing with data cleaning, model building on correlation, ridge regression, LASSO and elastic net regression.

Forth chapter provides implementation, analysis and evaluation as introduction, data description, experimentation with results for research questions.

Fifth chapter provides conclusion, recommendation and future work

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Sorghum [Sorghum bicolor (L.) Moench] is classified under the family Poaceae (grass family), tribe andropogoneae, genus bicolor, species bicolor. It the fifth most important staple food crop after wheat, rice, maize and barley [1]. In Ethiopia, sorghum grows well in 12 of 18 major agro-ecological zones with an altitude ranging between 400 – 2500 masl. The major agro-ecological zones are listed in table 1. Based on moisture condition agro ecological zones are classified into arid, semi-arid, sub-moist, moist, sub-humid, humid and per-humid. And depending on thermal, zones are grouped into hot, warm, tepid, cool, cold and very cold. Though capable of tolerating recurrent drought, for best production sorghum requires an average minimum rainfall ranging from 400-500mm; below this rainfall level, it requires the use of moisture conservation methods. Optimum temperature during growing season is $27^{o}c$ – $32^{o}c$. Sorghum can grow well on all soil types provided moisture is available and soil pH is within an acceptable level, usually from slight alkaline (6.0 to 7.5 respectively) [3].

For new sorghum variety seed that comes to agricultural research center for trial, there is a description of sorghum seed with the package. For agricultural research center, the description informs experts with what kind of agro ecological zones it needs to be produced during the trial. After field experimentation (trial) is complete experts examines the sorghum plant with different measurements such as the ability of sorghum plant in staying green in drought area, whether the sorghum plant yield is good amount of production, how much is the plant attacked by insect, disease and bird and so on. There are points that experts take into consideration which is used for predicting productivity of sorghum after it is released for farmers based on measurements taken from the trail. After complete trials are gathered mainly statistical methods are applied on the trial data that helps experts in releasing variety of sorghum to be produced across the country. For a sorghum plant that fulfills statistical criteria on statistics methods, the variety will be released and for plant that fails to fulfill the criteria that variety will not be accepted among experts.

Major Agro-ecological Zones of Ethiopia						
		Sub- zone	**Altitude**	**Rainfall**	**Temperature**	**Constraints**
1	Hot to warm arid lowland plains	Afar and Somalia Regional State	Below 1200 masl	1700 to 3000 mm	> 27 °C	Low rainfall, high temperature and infrastructure
2	Tepid to cool arid mid highlands	Somalia National Regional State	1200 – 1600 masl	500 – 600 mm	16 – 21 °C	Poor fertility status and moisture problem
3	Hot to warm semi-arid lowlands	Humera area (Western Tigray)	500 – 1600 masl	300 – 800 mm	21 to > 28 °C	Low rainfall and infrastructure
4	Tepid to cool semi-arid mid highlands	Oromia National Regional State (Bulbula)	1600 – 2200 masl	600 – 800 mm	16 – 27. °C	Moisture deficiency
5	Hot to warm sub-moist lowlands	North Western part (Metema)	400 – 1400 masl	200 – 1000 mm	> 21 °C	Low rainfall, deforestation and infrastructure
6	Tepid to cool sub-moist highlands	Assela zone (From Dera to Sire)	Between 1000 and 2000 masl	500 – 900 mm and 1400 – 1900 mm	16 – 21 °C	Surface stoniness, land degradation and moisture stress
7	Cold to very cold sub-	Mountains of Amhara	2800 – 4000	1000 – 1800	< 7 .5 - 16 °C	Low temperature, soil

	moist sub-afroalpine to afroalpine	Regional State	masl	mm		erosion and deforestation
8	Hot to warm moist lowlands	Southern People Nations and Nationalities Regional State (Bedeso)	1000 – 1600 masl	700 – 1300 mm	16 – 27 °C	Soil depth, moisture, topology and population density
9	Tepid to cool moist mid highlands	Around lake Tana	1000 – 2000 masl	1400 – 2200 mm	16 – 21 °C	Workability of soil type
10	Cold to very cold moist sub afroalpine to afroalpine	Mountains of Amhara Regional State	2800 – 3200 masl	1000 – 1800 mm	<7.5 – 16 °C	Steep slopes, low temperature, soil erosion and deforestation
11	Hot to warm sub humid lowlands	Gambella and Benishangul gumuz National Regional State	400 – 1000 masl	700 – 1000 mm	16 – 28 °C	Workability of Vertisols soil type
12	Tepid to cool sub humid mid highlands	Western part (Welega)	1000 – 2000 masl	700 – 1500 mm	16 – 21 °C	Soil erosion and deforestation
13	Cold to very cold sub	Southern People	2600 – 4300	700 – 1500	< 7.5 – 15 °C	Topology, very

	humid sub afroalpine to afroalpine	Nations and Nationalities Regional State and Oromia National Regional State	masl	mm		shallow soil, stoniness and drainage problem
14	Hot to warm humid lowlands	Southwestern part	600 – 1400 masl	1500 – 2000 mm	22 – 26 °C	Stoniness, shallow soil depth and topology
15	Tepid to cool humid mid highlands	Southern Rift valley part	1600 – 2000 masl	1200 – 1400 mm	17.5 – 21 °C	Stoniness, shallow soil depth and topology
16	Cold to very cold humid sub afroalpine to afroalpine	Bale mountains of Oromia National Regional State and Southern People Nations and Nationalities Regional State	3000 – 4200 masl	900 – 1800 mm	< 7.5 – 12 °C	Topology
17	Hot to warm per humid lowlands	Southern People Nations and Nationalities	500 – 1200 masl	1100 – 1500 mm	23.5 – 25.5 °C	Topology, shallow soil depth, stoniness, drainage and scarcity

		Regional State (Benji Maji zone)				of fire wood
18	Tepid to cool per humid mid highlands	Benji Maji zone and Kefacho Shekacho	1000 – 2800 masl	1100 – 2200 mm	13.5 – 25 °C	Topology

Table 1 Agro – ecological zones of Ethiopia (Source: Natural Resources Management & Regulatory Department MoA)

2.2 Origin and Domestication of Sorghum

It is believed that sorghum originated in Africa, more precisely in Ethiopia, between 5000 and 7000 years ago. From there, it was distributed along the trade and shipping routes around the African continent, and through the Middle East to India at least 3000 years ago. It then journeyed along the Silk Route into China. Sorghum was first taken to North America in the 1700-1800's through the slave trade from West Africa. It was re-introduced in Africa in the late 19th century for commercial cultivation and spread to South America and Australia. Sorghum is now widely found in the dry areas of Africa, Asia (India and China), the Americas and Australia [30].

2.3 Advantages of Sorghum

In Ethiopia, sorghum is used for making injera, kitta, kollo and locally made beverages (such as Tela and Areke). Being an indigenous crop, tremendous amount of variability exists in the country [8]. From different related literatures taken other importance of sorghum is summarized in Table 2.

Sorghum Importance Summary
Sorghum contains the following vitamins and minerals: vitamins B1, B2 and B3, calcium (Ca), potassium (K), iron (Fe), phosphorous (P), and sodium (Na).
It is a gluten-free cereal, which bears significance in the present-day scenario where the occurrence of Celiac Disease (CD), an immunological response to gluten intolerance is on the rise. Grain sorghum contains phenolic compounds like flavonoids, which have been found to inhibit tumour development.
Sorghum is consumed in various forms around the world like baked bread, porridge, tortillas, couscous, gruel, steam-cooked products, alcoholic, and non-alcoholic beverages, and so on.

The starches and sugars in sorghum are released more slowly than in other cereals and hence it could be beneficial to diabetics.
Sorghum strengthens the immune system, helps in the elimination of toxic waste from the body, increases endurance, assists in blood cell building, boost appetite, relieves diarrhea, aids rapid recovery, stimulates cardio-vascular system, stimulates free flow of blood, and lowers cholesterol levels.

Table 2 Summary in Importance of Sorghum

2.4 Characteristics of Sorghum Races

Of the five morphological races of sorghum (bicolor, guinea, caudatum, durra and kafir), all, except kafir, are grown in Ethiopia [8]. Sorghum races with their characteristics is described below.

Race Guinea – It is characterized by long and gaping glumes that reveal the obliquely twisted grain at maturity. Inflorescences are usually large and open with the branches often pendulous at maturity. The race is primarily a sorghum of West Africa, but is also grown along the East African rift from Malawi south to Swaziland. Guinea sorghums are grown extensively in northern Nigeria where hybridization with drought-tolerant kinds of race durra is evident [27].

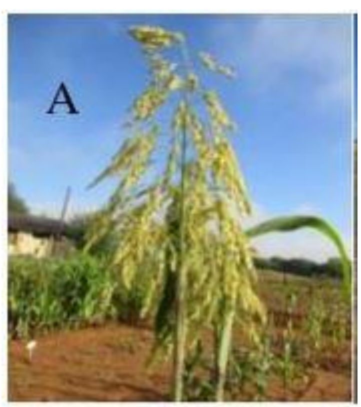

Figure 1 Race Guinea (Source: Google, Photo: T.Motlhaodi)

Race Caudatum –Caudatum sorghums have characteristics turtule-backed grains that are flat on one side and distinctly curved on the opposite side. The grains are usually exposed between the shorter glumes at maturity. The inflorescences range in shape from compact to open. It is widely grown in Chad, Sudan, northeastern Nigeria and Uganda. It is an old race of grain sorghum [27].

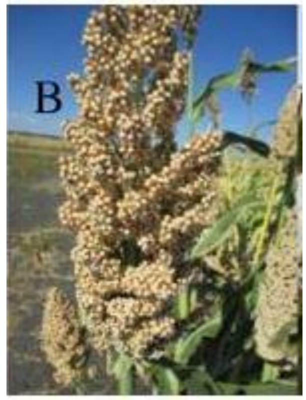

Figure 2 Race Caudatum (Source: Google, Photo: T.Motlhaodi)

Race Durra –The name is derived from the Arabic name for sorghum, and the distribution of race durra is closely associated with Islamic people in Africa. Inflorescences are usually compact. Sessile spikelet's are characteristically flattened and ovate in outline, and the lower glume is ceased near the middle or with the tip a distinctly different texture than the lower two-thirds. Durra sorghums are widely grown along the fringes of the southern Sahara, across arid West Africa, the Near East and parts of India [27].

Figure 3 Race Durra (Source: Google, Photo: T.Motlhaodi)

Race Kafir –It is characterized by more or less compact inflorescences that are often cylindrical in outline. Sessile spikelet's are typically elliptic with the glumes at maturity, tightly clasping the usually much longer grain. The name is derived from 'kafir' the Arabic for unbeliever, referring to the Bantu who grow this race. Kafir sorghums are important staples across the eastern and southern savanna from Tanzania to South Africa [27].

Figure 4 Race Kafir (Source: Google, Photo: T.Motlhaodi)

Race Bicolor – This race is based on subseries Bicoloria of Snowden. It is characterized by open inflorescences and long clasping glumes that usually enclose the elliptic grain at maturity. Bicolor sorghums, but they lack the ability of natural seed dispersal. Members of race bicolor, particularly those far removed from their African center of origin, probably represent relics of ancient cultivated kinds. Others may have originated as selections out of derivatives from hybridization between modern cultivated races and wild members of S.bicolor. bicolor sorghums are low yielding as cereals. They are rarely an important crop, but grown across the range of sorghum cultivation in Africa and Asia. Some selections are grown strictly for their sweet stems, while in Africa other kinds are grown for their bitter grains that are used to flavor sorghum beer [27].

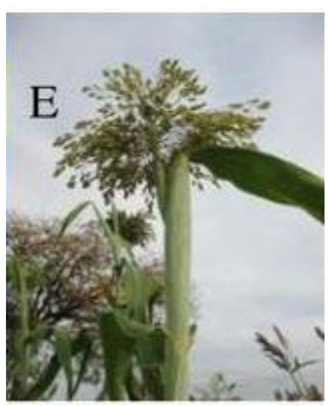

Figure 5 Race Bicolor (Source: Google, Photo: T.Motlhaodi)

2.5 Application of Machine Learning in Agriculture

An often-quoted example of the application of machine learning in agriculture is the use of the AQ11 program to identify rules for diagnosis of soybean diseases. In this early application the similarity-based learning program AQ11 was used to analyze data from over 600 questionnaires describing diseased plants. Each plant was assigned to one of 17 disease categories by an expert collaborator, who used a variety of measurements describing the condition of the plant. At the same time, the plant pathologist who had produced the diagnoses was interviewed and his expertise was translated into diagnosis rules using the standard knowledge engineering approach. Surprisingly, the computer-generated rules outperformed the expert-derived rules on the remaining 340 test instances- they gave the correct disease top ranking 97.6% of the time, compared to only 71.8% for the expert-derived rules. Furthermore, according to Quinlan , not only did AQ11 find rules that outperformed those of the expert collaborator, but the same expert was so impressed that he adopted the discovered rules in place of his own [35].

The most significant project so far carried out using the WEKA workbench has been the analysis of dairy herd data for the purposes of isolating rules that describe factors that farmers might use for culling decisions. This involved working with a large data set of 19,103 records containing 705 attributes spread across 10 herds and 6 years. About 40 new attributes were derived, including attributes like age and production index relative to herd, and these were added to the original data set which was then processed by various machine learning schemes. The high level of missing values in the data adversely affected the results, though experts consulted concurred that the rules produced appeared plausible. Other project in progress currently is the analysis of data pertaining to apple bruising and the prediction of when a cow will be in heat. In the case of the apples a variety of factors affect the size of an apple's bruise during the storage and packing stages of processing. Using these factors, which include apple temperature and the height it may fall, a set of rules that describe what factors results in larger bruises is being sought. A set of rules is also being sought with respect to predicting when a cow will be in heat based upon information contained in a herd milking database. This involves using information such as the volume of milk produced, its conductivity and the order the cow came into the shed, as well as recognizing temporal patterns in the data as the event in question is cyclical [36].

2.6 Application of Machine Learning in Predicting Crop Productivity

Machine Learning involves problems in which the input and output relationship is not known. Learning specifies the automatic acquirement of structural descriptions. In contrast to traditional statistical methods, machine learning does not make assumptions about the exact construct of the data model, which describes the data. This feature is very helpful to describe complex non-linear behaviors such as a crop yield prediction. Machine learning is a part of artificial intelligence employed to build an intelligent system. By utilizing the training samples, the test samples can be identified. The accuracy of the system can be measured using metrics such as mean square error, root mean square error, precision, recall, sensitivity specificity etc. Further, machine learning can be employed to address a variety of applications including crop yield prediction through supervised, unsupervised and reinforcement learning methods. Classification, clustering, regression, prediction is some of the techniques involved to attain the intelligent system [28].

Crop yield prediction (CYP) is a major problem in agriculture. Starting each growing season, agricultural planners require estimating the yield for all the involved crops. Regrettably, CYP is difficult because it depends on many interrelated factors. Moreover, yield is also affected by farmer decisions (such as applied irrigations, pest and fertilizers applications, crop rotation, and land preparation) and incontrollable factors (such as weather, subsidies and market) [2].

2.7 Related Work

Related works are as follows:

Article Name	Author	Year	Methodology	Results
Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique	Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh	2015	Crop Selection Method (CSM)	The result from this paper is only achieving maximum net yield rate of crops.
Agricultural Crop Yield Prediction Using Artificial Neural Network Approach	Miss. SnehalS.Dahikar, Dr. SandeepV.Rode	2014	Artificial Neural Network (ANN)	Predict crop using soil and atmosphere parameters.
Machine Learning in Soil Classification and Crop Detection	Ashwini Rao Janhavi U Abhishek Gowda N S Manjunatha Mrs.Rafega Beham A	2016	Support Vector Machine (SVM) Image Processing	Based on classification and grading of soil sample using image processing.

Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in India	Rajshekhar Borate, Rahul Ombale, Sagar Ahire, Manoj Dhawade, Mrs. Prof. R. P. Karande	2016	Clustering Classification K-Means Clustering Linear Regression: K-nearest neighbor Artificial neural network	Clustering used to divide regions and classification for crop yield prediction. No results have been shown.
This study focuses on correlation and Regression models to address the problem. It uses Pearson type of correlation to determine factors that increase and decrease productivity sorghum by identifying relationships between factors and sorghum production. It tells different factors that increase and decrease sorghum production. And it also uses ridge, LASSO and elastic net regression to show effectiveness of using machine learning for predicting productivity of sorghum. It shows the most important variables used for predicting productivity of sorghum				

Table 3 Summary of Literatures

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

In this study different methods were used to address problems found in sorghum production. Machine learning books is used to understand basics and techniques of machine learning. Sorghum research literatures were also used to gain knowledge from different works done by the researchers. Google is used for overall understanding of sorghum production supported by technologies in existing system. Paper and pen also used during discussion made between researcher and experts in agricultural research center. RStudio is for implementation of the system. And yEd Graph Editor also used for drawing architecture of proposed system.

3.2 Data Collection Techniques

For the purpose of this study both primary and secondary data collection methods were used. Some of data collection techniques include discussion, observation and documents. With the help of fieldscorer software that is installed on tablet, we observe sorghum plant information is stored is starting from field experiment up to server-based storage system. Formerly information from field experiment is stored spreadsheet layout in the fieldscorer and then copied to the server using combine software installed on computer. Discussion with sorghum experts done in different times in detail about sorghum crop and its production in Ethiopia. For further study secondary data collection methods mainly documents such as books were used. A book of Ethiopian Strategy for Sorghum prepared by Ethiopian Institute of Agricultural Research (EIAR) is used for additional understanding of sorghum production and common varieties of sorghum depending on yield, bird tolerance and staying green in drought area. Agro ecological zones of Ethiopia book is used for understanding of agro ecological zones based on altitude, annual rainfall, temperature and area.

3.3 Data Preparation

3.3.1 Raw Data Source and Description

The main source of sorghum data is Melkassa Agricultural Research Center. For this reason, by the time sorghum departments were the only one that uses server-based data storage system along the other departments. Due to this sorghum department trial data is selected for machine learning techniques. The raw data contains sorghum field trials collected in the year of 2011 – 2014 collected from different area. The raw data contains total of 1500 observations and 12 attributes and stored in Microsoft Excel Worksheet (xlsx) file format. The total size of trial data is 88.2 KB. The description raw data using full name, name, description and data type is shown below.

Raw Data Source Description			
Full Name	**Name**	**Description**	**Data Type**
Replicattion	Replicate	Number of repetitions for plot treatment.	Integer
Emergency Day	Emergence.Day	Number of days for sorghum plant to emerge from the soil.	Numeric
Flowering Day	Flowering.Day	Number of days for sorghum plant to display a flower.	Numeric
Maturity Day	Maturity.Day	Number of days for sorghum plant to be mature.	Numeric
Plant Height	Plant.Height	A height of sorghum plant.	Numeric

Drought Score	Drought.Score	A score given for sorghum plant depending on drought. Score 0 = No, Sorghum plant is not damaged by drought. Score 1 = Yes, Sorghum plant is damaged by drought	Integer
Bird Score	Bird.Score	A score given for sorghum plant depending on bird attack. Score 0 = No, Sorghum plant is not attacked by bird. Score 1 = Yes, Sorghum plant is attacked by bird.	Integer
Insect Score	Insect.Score	A score given for sorghum plant depending on insect. Score 0 = No, Sorghum plant is not eaten by insect. Score 1 = Yes, Sorghum plant is eaten by insect.	Integer

Disease Score	Disease.Score	A score given for sorghum plant depending on disease. Score 0 = No, Sorghum plant is not damaged by disease. Score 1 = Yes, Sorghum plant is damaged by disease.	Integer
Heads per a plot	Heads.Plot	Number of sorghum plant heads per a plot of land.	Numeric
Yield per a plot	Yield.Plot	Total yield obtained from sorghum plant per a plot of land.	Numeric
Varieties	Varieties	Name of varieties harvested during the trial	Factor

Table 4 Sorghum Data with Description and Data Types

3.4 Data Preprocessing

3.4.1 Data Cleaning

Sorghum trial data obtained from Melkassa Agricultural Research Center contain a lot of irrelevant variables for machine learning such as barcode source. The first step for data preprocessing is selecting relevant variables for machine learning techniques. After selecting relevant variables and complete trial data. The next step is checking for outliers in the data. Outliers in the sorghum data were handled by using boxplots in RStudio.

3.5 Model Building

3.5.1 Correlation

Correlation and the correlation matrix are two terms that you will often come across in statistics, probability theory or Machine Learning. Correlation is basically a measure of the dependence between two variables. In fact, more than dependence, it is a measure of how two variables move together and how strongly they are related. Correlation is expressed in the form of a correlation coefficient. The correlation coefficient is measured on a scale from -1 to +1. A positive correlation coefficient means that there is a perfect positive relationship between the two variables, whereby they both move together in the same direction. An increase in one is accompanied by an increase in the other and a reduction in one is accompanied by a reduction in the other. A correlation coefficient of -1 represents a perfect negative correlation. This means when one increases, the other decreases and vice-versa. A value of 0, however, means that there is no correlation between the two and they are not related to each other at all. [10]

3.5.2 Ridge regression

Ridge regression is identical to least squares, unless the ridge coefficients are estimated by minimizing a slightly different quantity. It is hoped that the net effect will give estimates that are more reliable. If multicollinearity is present in the data, least squares estimates are unbiased, but the variances are large hence they are far away from the true value. The ridge coefficients minimize a penalized residual sum of squares, where $\lambda \geq 0$ is a tuning parameter that controls the amount of shrinkage. The coefficients are shrunk towards zero. The ridge estimates give an equation that predicts future observations better than least squares. While biased, the decreased variance of the ridge estimates frequently result in a minimum mean square error compared to the estimates of least squares method. The assumptions of ridge regression are same as the least squares regression except the normality is not assumed. It reduces the variability by shrinking the coefficients, resulting in more prediction accuracy at the cost of usually only a small increase of bias. It shrinks the coefficients nearly to zero but not exactly to zero, thus not good for feature selection. When the number of predictors is large, ridge regression will not provide an easily interpretable sparse model [32].

3.5.3 LASSO

The LASSO is a particular case of the penalized least squares regression with L1-penalty function. LASSO transforms each and every coefficient by a constant component λ, truncating at zero. Hence it is a forward-looking variable selection method for regression. It decreases the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. LASSO was originally defined in the context of least squares, but it can also be extended to a wide variety of models. LASSO improves both prediction accuracy and model interpretability by combining the good qualities of ridge regression and subset selection. If there is high correlation in the group of predictors, LASSO chooses only one among them and shrinks the others to zero. It reduces the variability of the estimates by shrinking the some of the coefficients exactly to zero producing easily interpretable models [32].

3.5.4 Elastic Net Regression

A third commonly used model of regression is the Elastic Net which incorporates penalties from both L1 and L2 regularization. In addition to setting and choosing a lambda value elastic net also allows us to tune the alpha parameter where $\boldsymbol{\alpha} = 0$ corresponds to ridge and $\boldsymbol{\alpha} = 1$ to lasso. Simply put, if you plug in 0 for alpha, the penalty function reduces to the L1 (ridge) term and if we set alpha to 1 we get the L2 (lasso) term. Therefore, we can choose an alpha value between 0 and 1 to optimize the elastic net. Effectively this will shrink some coefficients and set some to 0 for sparse selection. [9]

Architecture of the Proposed Solution

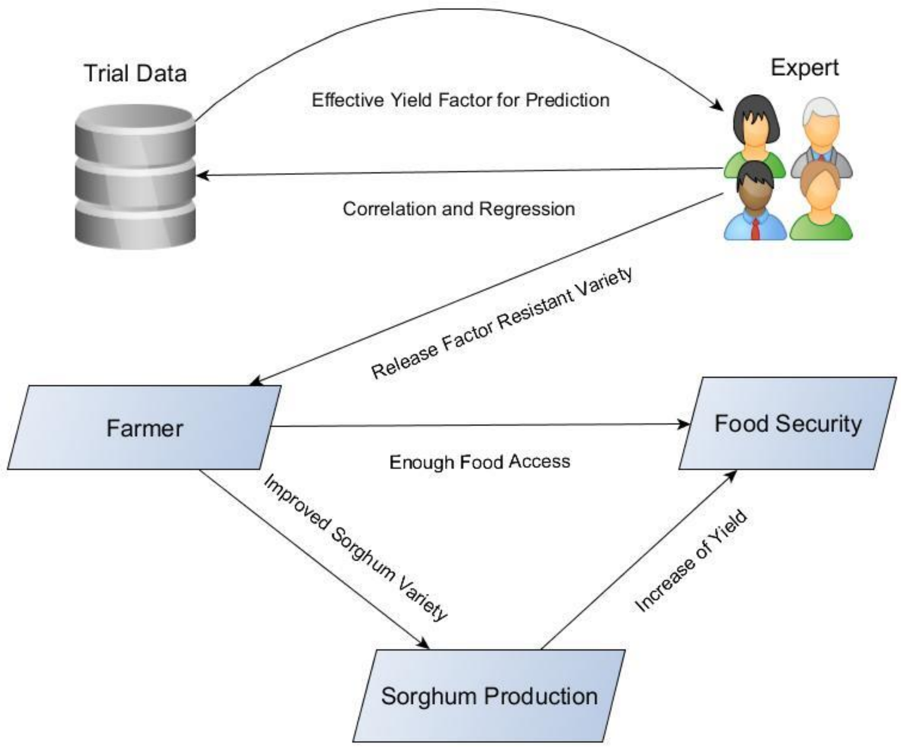

Figure 6 Architecture of the Proposed Solution

In above figure of architecture of proposed solution shows first experts in agriculture sector apply correlation to observe relationship between sorghum production factors and yield also regression models for effectiveness of machine learning for predicting productivity of sorghum on trial data gathered from field. In the proposed solution Pearson correlation is applied to identify factors that increase and decrease productivity of sorghum in RStudio. For effectiveness of machine learning used in predicting productivity of sorghum ridge regression, lasso and elastic net regression. The trial data then outputs factors relating to sorghum production that inform experts to give more focus on factor resistant varieties and important factors used for predicting sorghum production. After experts work on factor resistant varieties, then agricultural research centers release improved varieties for farmers. The farmers will be benefited from improved varieties and increase of sorghum production for people of country. The people of country including farmers will be benefit from increased amount of sorghum production. Increase amount of sorghum production in one country indicates that all individuals get sufficient amount of food access at all times which means food security.

CHAPTER FOUR

IMPLEMENTATION, ANALYSIS AND EVALUATION

4.1 Introduction

This chapter presents on implementation, analysis and evaluation of the study. The experiments were run in 1500 observations with 12 attributes. The models were implemented using RStudio 1.1.447 tool.

4.2 Data Description

The data obtained from MARC consists of 12 attributes with 1500 tuples.

> View (sorghum)

	Replicate	Emergence.Day	Flowering.Day	Maturity.Day	Bird.Score	Drought.Score	Disease.Score	Insect.Score	Plant.Height	Heads.Plot	Yield.Plot	Varieties
1	3	15	64	133	0	0	0	0	114	84	7408	ABON34X2002BK7020
2	3	10	66	115	0	0	0	0	190	71	6754	ICSA34XGambella1107
3	3	10	67	114	0	0	0	1	132	81	5518	ABON34XICSR24004
4	3	13	79	131	0	0	0	0	150	93	3011	ICSA21XPI308453
5	1	16	64	130	1	0	0	0	115	53	4429	ABON34XP89009
6	3	11	66	117	0	0	0	0	236	84	5666	ICSA34X2005MI5065
7	3	11	66	114	0	0	1	0	200	83	7127	ABON34X104GRD
8	3	8	67	116	0	0	1	0	210	87	7390	ABON34X2005MI5065
9	3	8	63	119	0	0	0	0	131	87	5375	ICSA10XP89009
10	1	15	66	131	0	0	0	0	122	61	4122	ABON34XPI308453
11	3	10	64	117	0	0	0	0	212	98	6825	ICSA10XGambella1107
12	3	9	65	133	0	0	0	0	235	71	7623	ABON34XIESV92031DL
13	1	9	67	119	1	0	0	0	221	68	4266	ABON34XGambella1107
14	1	9	63	117	0	0	0	0	130	48	4243	ABON34XPDL984953
15	3	16	61	117	0	0	0	0	154	88	7514	ICSA10X2002BK7020
16	3	11	61	115	0	0	0	1	184	104	7728	ESH-2
17	3	13	65	131	0	1	0	1	190	66	7972	ABON34XMisikir
18	1	18	66	131	0	0	0	0	150	31	4025	ICSA34XICSR24004
19	3	10	66	119	0	0	0	0	210	59	5961	ABON34XGambella1107
20	3	8	64	116	0	0	0	0	160	88	5751	ICSA10X2002BK7020
21	3	13	64	130	0	0	0	0	115	68	5169	ICSA10XP89009
22	3	16	67	134	0	0	0	0	241	71	8270	ICSA34X2005MI5065

Showing 1 to 23 of 1,500 entries

Table 5 Sample Sorghum Data

```
> summary(sorghum)
   Replicate      Emergence.Day   Flowering.Day    Maturity.Day     Bird.Score       Drought.Score
 Min.   :1.000   Min.   : 6.00   Min.   : 55.0   Min.   : 91.0   Min.   :0.0000   Min.   :0.0000
 1st Qu.:1.000   1st Qu.: 9.00   1st Qu.: 71.0   1st Qu.:119.0   1st Qu.:0.0000   1st Qu.:0.0000
 Median :1.000   Median :11.00   Median : 88.0   Median :146.0   Median :0.0000   Median :0.0000
 Mean   :1.835   Mean   :11.26   Mean   : 90.6   Mean   :142.5   Mean   :0.1753   Mean   :0.3247
 3rd Qu.:3.000   3rd Qu.:14.00   3rd Qu.:111.0   3rd Qu.:158.0   3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :3.000   Max.   :19.00   Max.   :171.0   Max.   :185.0   Max.   :1.0000   Max.   :1.0000

 Disease.Score     Insect.Score     Plant.Height     Heads.Plot       Yield.Plot        Varieties

 Min.   :0.0000   Min.   :0.0000   Min.   : 80.0   Min.   : 11.00   Min.   : 120   Tadesse   :  31

 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:150.0   1st Qu.: 59.00   1st Qu.:1838   ACC#203331:  12

 Median :0.0000   Median :0.0000   Median :184.0   Median : 75.00   Median :2860   ACC#203451:  12

 Mean   :0.1393   Mean   :0.3033   Mean   :181.3   Mean   : 80.17   Mean   :3176   ACC#203461:  12

 3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:210.5   3rd Qu.: 91.00   3rd Qu.:4142   ACC#203489:  12

 Max.   :1.0000   Max.   :1.0000   Max.   :283.0   Max.   :180.00   Max.   :9978   ACC#216054:  12

                                                                                   (Other)   :1409
```

Table 6 Sorghum Data Summary

Table 6 Summary function in RStudio summarize variables of sorghum data (numeric, integer and factor) as minimum, median, mean, maximum and quartiles.

4.3 Experimentation for Research Questions

Research Question 1 and Experimentation

1. What are the factors that increase and decrease productivity of sorghum from sorghum trial data?

Dependent (response) variable is yield per a plot.

Independent (explanatory) variables are replication, emergency day, flowering day, bird score, drought score, disease score, insect score, plant height, maturity day, heads per a plot and varieties.

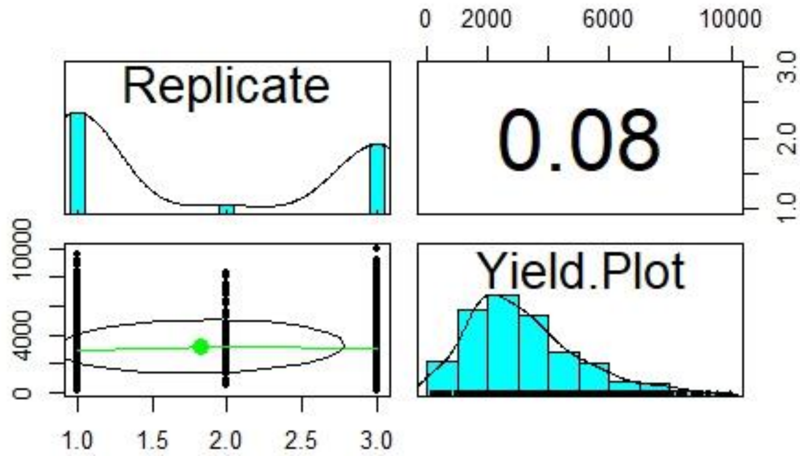

Figure 7 Correlation between yield per a plot and replication

Figure 7 shows the relationship between yield per a plot and replication is positive correlation with 0.08 = 8 %. Positive relationship between yield per a plot and replication indicate if number of repetitions for plot treatment increase yield per a plot also increase. And if there is no treatment applied for sorghum seed sown to the plot of land, there would be decreasing amount of yield during harvesting. Repeated treatment of sorghum plant in a plot of land during trial gives an advantage to protect the plant from different insects and diseases.

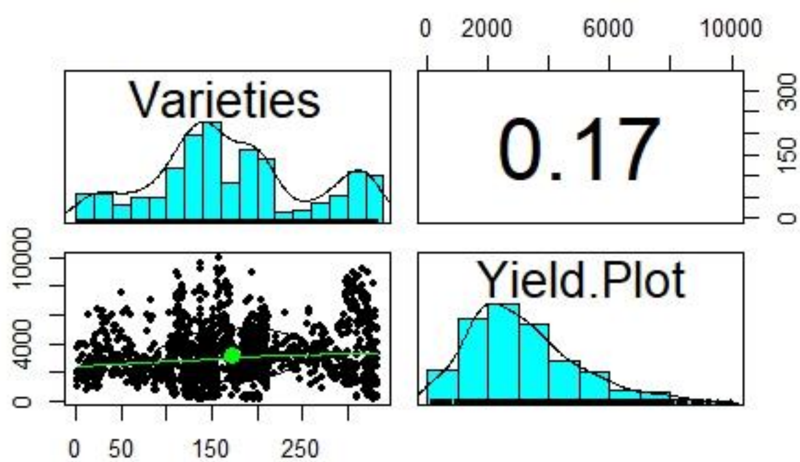

Figure 8 Correlation between yield per a plot and varieties

The relationship between yield per a plot and varieties also shows positive correlation with 0.17 = 17 % which means varieties play a great role in increasing production of sorghum per plot of land. For agricultural research center sorghum experts need to improve low yield of sorghum seed after the trail is complete for future to increase and achieve high yield. Sorghum variety that show low production during trail time should also experiment and give a second chance to be experiment in other agro ecological zones of agricultural research centers, than to conclude the variety as it is failed variety seed.

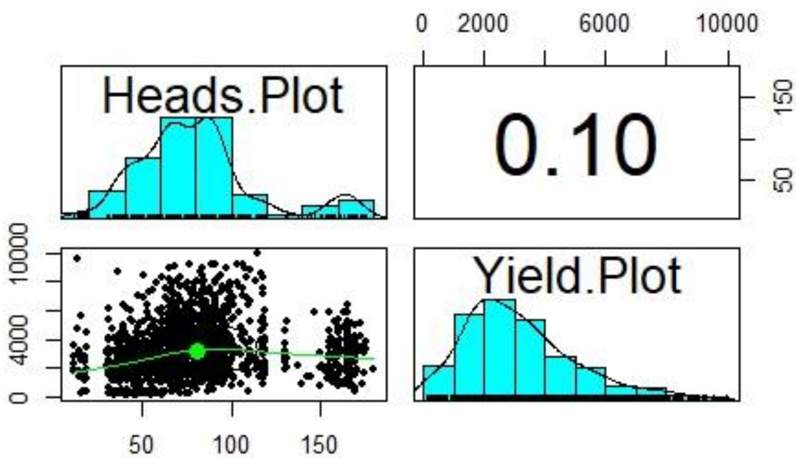

Figure 9 Correlation between yield per a plot and heads per a plot

Figure 9 shows the relationship between yield per a plot and heads per a plot is positive correlation with 0.10 = 10 %. This indicates if number of sorghum plant heads increase production of sorghum also increases and vice versa. Heads per a plot and number seeds found in sorghum plant determine and is used for predicting yield amount for a future

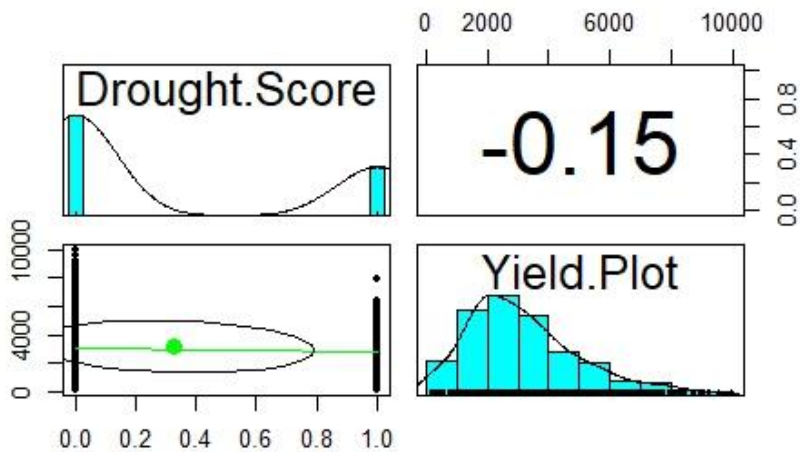

Figure 10 Correlation between yield per a plot and drought score

Figure 10 shows the relationship between yield per a plot and drought score is negative correlation with -0.15 = -15 %. Negative relationship tells if sorghum plant is damaged highly by drought then production of sorghum decreases. Highest percentage of drought score tells that the sorghum variety that is experimented depending on the agro ecological zone needs to have more moisture for a plant to resist drought and produce good amount of sorghum production.

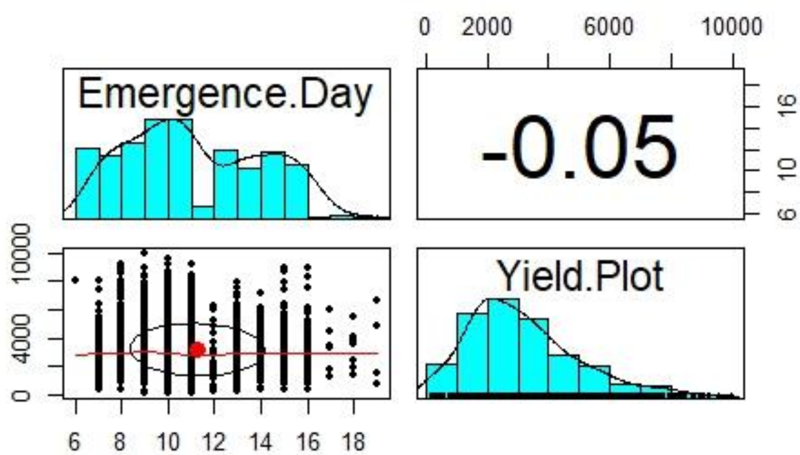

Figure 11 Correlation between yield per a plot and emergency day

Figure 11 shows the relationship between yield per a plot and emergency day is negative correlation with -0.05 = -5 %. The correlation tells the day for sorghum plant to emerge from soil increase sorghum production decrease. To define this, if the plant takes long time to emerge from soil in expected time achieving good amount of sorghum production within length of period is less. Because the plant may take long time to emerge a flower and be mature.

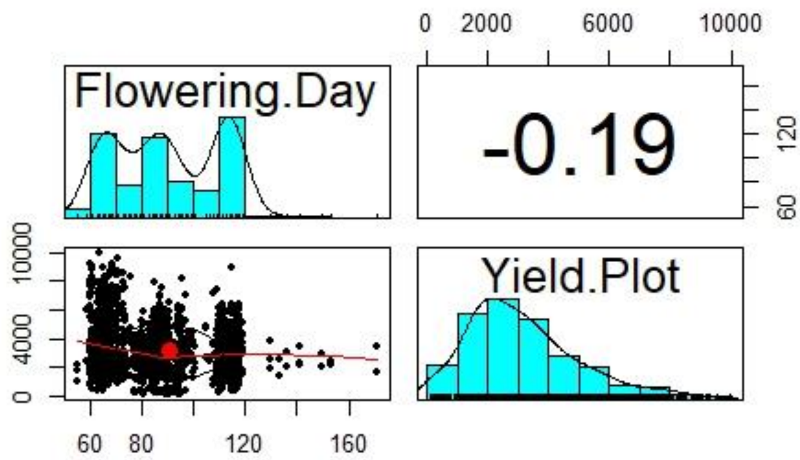

Figure 12 Correlation between yield per a plot and flowering day

In the above figure, the relationship between yield per a plot and flowering day is negative correlation with - 0.19 = -19 %. This indicates if the day for sorghum plant to display a flower increase then sorghum production decrease. When a plant flower it means that the flower is in reproductive stage in agriculture term. In field experiment some sorghum varieties takes long time for flowering and other might not take long time to display a flower. This plant reproductive stage also affected depending on agro ecological zones. Here also experts need to experiment the sorghum variety in other agro ecological zones for better production.

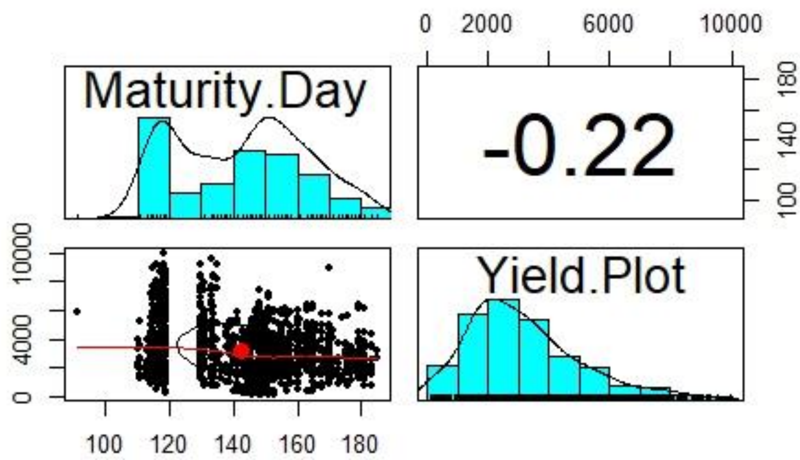

Figure 13 Correlation between yield per a plot and maturity day

Figure 13 shows the relationship between yield per a plot and maturity day is negative correlation with – 0.22 = -22 %. As the day for sorghum plant to be mature increase, this result in decreasing amount of sorghum production. Sorghum plant that has short maturity day produce sorghum in short time and vice versa. Sorghum maturity day also affected depending on sowing time of the seed as it is sown early or late time. It is also affected by agro ecological zones of the country.

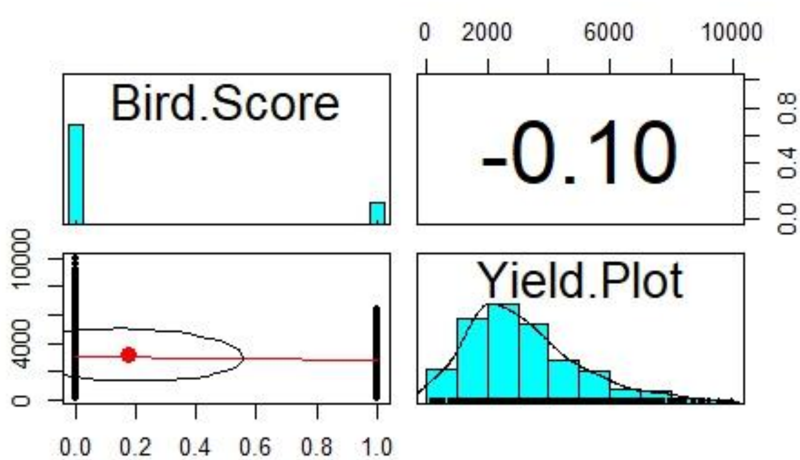

Figure 14 Correlation between yield per a plot and bird score

Figure 14 shows the relationship between yield per a plot and bird score is negative correlation with -0.10 = -10%. This tells if sorghum plant is not attacked by a bird, sorghum production increases and vice versa. If the trail shows whether high or low percentage of correlation between yield per a plat and bird score, the plant is exposed for bird to attack. In order to protect sorghum plant from birds using bird scare away methods needs to be used.

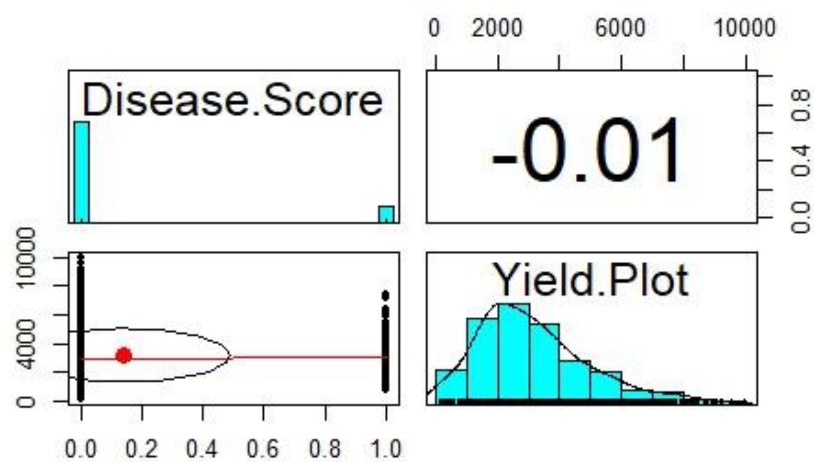

Figure 15 Correlation between yield per a plot and disease score

Figure 15 shows the relationship between yield per a plot and disease score is negative correlation with -0.01 = -1%. This tells if sorghum plant is not harmed by a disease, sorghum production increases and vice versa. Sorghum variety attacked by disease informs experts that the variety is not treated well during field experimentation or the treatment is not enough for preventing the type of disease found in plant. This experimentation also shows that the type of variety is not disease resistant in experiment taken agro ecological zone.

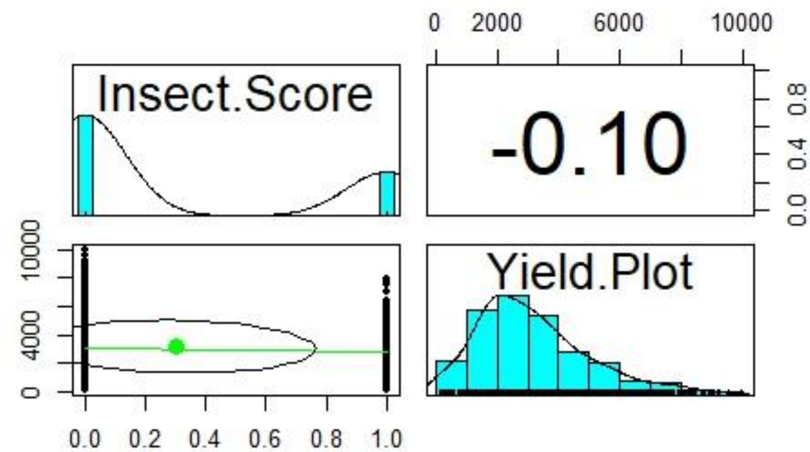

Figure 16 Correlation between yield per a plot and insect score

Figure 16 shows the relationship between yield per a plot and insect score is negative correlation with -0.10 = -10%. This tells if sorghum plant is not injured by insect, sorghum production increases. After field experiment complete, sorghum variety attacked by insects mainly affected by agro ecological zone. Agro ecological zones may result in increase or decrease of number insect reproduction in the area of field experiment taken.

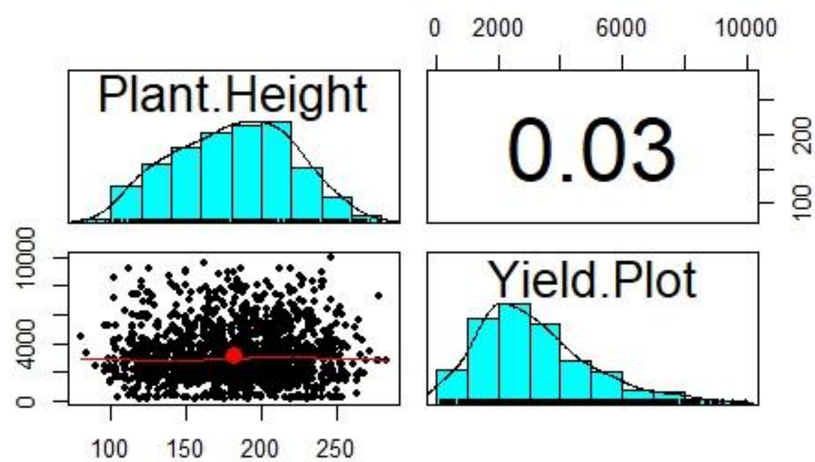

Figure 17 Correlation between yield per a plot and plant height

Figure 17 shows the relationship between yield per a plot and plant height is positive correlation with 0.03 = 3%. This tells if sorghum plant height increases production increases. For longest plant height the plant grows more branches. And the plant branches hold sorghum heads which means a number of seeds is produced in that sorghum plant heads. For sorghum plant that holds number of seeds indicates yield amount to be produced. Plant height is also dependent on type of sorghum variety.

Research Question 2 and Experimentation

2. How machine learning can be effectively used to predict productivity of sorghum?

Experimentation I: Ridge Regression

Dependent (response) variable is yield per a plot.

Independent (explanatory) variables are replication, emergency day, flowering day, bird score, drought score, disease score, insect score, plant height, maturity day and heads per a plot

In the independent variables, varieties are not included because it contains 1500 different variety names in the sample data. If it were limited varieties names with small number it would be included for prediction. In RStudio for ridge regression model, fscaret and caret library is used to select important variables for predicting yield per a plot. By using read function the sample data from excel is imported in to RStudio. Then the variables from dataset is ordered by making yield per a plot a last column. After this the sample data is split in to 1126 observations for training and 374 observations for testing. Then ridge regression is selected as model for the data set. Next the model trains a tests sorghum sample data. Next it displays variable importance using root mean square deviation (RMSE) which measure predicted and observed values and mean square error (MSE) measures average squares of error between estimated values in ridge model. This experiment result is shown in table 7 below. It shows the sum of each important variables in percentage with input number that the model thinks. Ridge regression RMSE variable are maturity day (100%), flowering day (79.5%) and drought score (32.9%) and MSE variables are maturity day (100%), flowering day (79.5%) and drought score (32.9%).

$matrixVarImp.RMSE

	ridge	SUM	SUM%	ImpGrad	Input_no
4	33.95608992	33.95608992	100.00000000	0.000000	4
3	27.00527705	27.00527705	79.52999627	20.470004	3
8	11.18914488	11.18914488	32.95180601	58.566821	8
1	7.62022766	7.62022766	22.44141678	31.896246	1
6	5.68360889	5.68360889	16.73811356	25.414185	6
5	5.51829220	5.51829220	16.25125923	2.908657	5
10	4.89061170	4.89061170	14.40275282	11.374543	10
2	3.62760001	3.62760001	10.68320886	25.825229	2
9	0.49701790	0.49701790	1.46370770	86.298988	9
7	0.01212981	0.01212981	0.03572203	97.559483	7

$matrixVarImp.MSE

	ridge	SUM	SUM%	ImpGrad	Input_no
4	33.95608992	33.95608992	100.00000000	0.000000	4
3	27.00527705	27.00527705	79.52999627	20.470004	3
8	11.18914488	11.18914488	32.95180601	58.566821	8
1	7.62022766	7.62022766	22.44141678	31.896246	1
6	5.68360889	5.68360889	16.73811356	25.414185	6
5	5.51829220	5.51829220	16.25125923	2.908657	5
10	4.89061170	4.89061170	14.40275282	11.374543	10
2	3.62760001	3.62760001	10.68320886	25.825229	2
9	0.49701790	0.49701790	1.46370770	86.298988	9
7	0.01212981	0.01212981	0.03572203	97.559483	7

```
$model
list()
```

Table 7 RMSE and MSE for Ridge Regression

Next the model outputs input numbers shown in table 7 by arranging the variables from number1 to 10 as labels shown in table 8. Labels with input number indicates the name of label is most important variables used to predict yield per a plot of land. In table 8, input number 4 is maturity day, 3 is flowering day and 8 is drought score. Therefore, the most important variables used to predict yield per a plot of land using ridge regression are maturity day, flowering day and drought score.

```
> FSResults$PPlabels
```

	Orig Input No	Labels
1	1	Replicate
2	2	Emergence.Day
3	3	Flowering.Day
4	4	Maturity.Day
5	5	Bird.Score
6	6	Insect.Score
7	7	Disease.Score
8	8	Drought.Score
9	9	Plant.Height
10	10	Heads.Plot

Table 8 Labels for Ridge Regression

Experimentation II: LASSO

In RStudio for lasso, the same step with dependent and independent variables is used as of ridge regression until sample data is split. Then lasso is selected as model. Next the model trains a tests sorghum sample data. Next it displays variable importance using root mean square deviation (RMSE) and mean square error (MSE) in lasso. This experiment result is shown in table 9 below. LASSO RMSE variables are maturity day (100%), flowering day (76%) and drought score (35.2%) and MSE variables are maturity day (100%), flowering day (76%) and drought score (35.2%)

```
$matrixVarImp.RMSE
        lasso        SUM       SUM%   ImpGrad Input_no
4  34.9197748 34.9197748 100.000000  0.000000        4
3  26.5390764 26.5390764  76.000136 23.999864        3
8  12.3164756 12.3164756  35.270776 53.591167        8
1   7.6413833  7.6413833  21.882682 37.958036        1
5   5.2389471  5.2389471  15.002809 31.439808        5
6   4.8074361  4.8074361  13.767088  8.236597        6
10  4.6975035  4.6975035  13.452273  2.286720       10
2   3.2695218  3.2695218   9.362952 30.398737        2
9   0.4561163  0.4561163   1.306183 86.049449        9
7   0.1137651  0.1137651   0.325790 75.057871        7

$matrixVarImp.MSE
        lasso        SUM       SUM%   ImpGrad Input_no
4  34.9197748 34.9197748 100.000000  0.000000        4
3  26.5390764 26.5390764  76.000136 23.999864        3
8  12.3164756 12.3164756  35.270776 53.591167        8
1   7.6413833  7.6413833  21.882682 37.958036        1
5   5.2389471  5.2389471  15.002809 31.439808        5
6   4.8074361  4.8074361  13.767088  8.236597        6
10  4.6975035  4.6975035  13.452273  2.286720       10
2   3.2695218  3.2695218   9.362952 30.398737        2
9   0.4561163  0.4561163   1.306183 86.049449        9
7   0.1137651  0.1137651   0.325790 75.057871        7

$model
list()
```

Table 9 RMSE and MSE for LASSO

LASSO outputs input numbers shown in table 9 by arranging the variables from number1 to 10 as labels shown in table 10. In table 10, input number 4 is maturity day, 3 is flowering day and 8 is drought score. Therefore, the most important variables used to predict yield per a plot of land using lasso are maturity day, flowering day and drought score.

```
> FSResults$PPlabels
   Orig Input No          Labels
1              1       Replicate
2              2  Emergence.Day
3              3  Flowering.Day
4              4   Maturity.Day
5              5      Bird.Score
6              6    Insect.Score
7              7   Disease.Score
8              8   Drought.Score
9              9    Plant.Height
10            10      Heads.Plot
```

Table 10 Labels for LASSO

Experimentation III: Elastic net Regression

Elastic net regression, fscaret, caret and glmnet library is used to select important variables for predicting yield per a plot. The same step with dependent and independent variables is used as of lasso until sample data is split. Then elastic net regression selects glmnet as model for the dataset Next the model trains and tests sorghum sample data. Next it displays variable importance using root mean square deviation (RMSE) and mean square error (MSE) in elastic net regression model. This experiment result is shown in table 11 below. Elastic net regression RMSE variables are drought score (100%), maturity day (91.3%) and flowering day (69.8%) and MSE variables drought score (100%), maturity day (91.5%) and flowering day (69.9%).

```
$matrixVarImp.RMSE
       glmnet      lasso       ridge        SUM        SUM%    ImpGrad Input_no
8  54.5746139 12.3065580 11.18914488 78.0703168 100.000000  0.000000        8
4   2.4697999 34.8916564 33.95608992 71.3175463  91.350399  8.649601        4
3   0.9936168 26.5177064 27.00527705 54.5166003  69.830126 23.557942        3
1  15.5095184  7.6352303  7.62022766 30.7649764  39.406752 43.567691        1
7  22.6266612  0.1136735  0.01212981 22.7524645  29.143553 26.044265        7
5   0.0000000  5.2347285  5.51829220 10.7530207  13.773507 52.739095        5
6   0.0000000  4.8035650  5.68360889 10.4871739  13.432985  2.472299        6
2   3.4636785  3.2668891  3.62760001 10.3581676  13.267741  1.230135        2
10  0.0000000  4.6937210  4.89061170  9.5843327  12.276539  7.470771       10
9   0.0000000  0.4557490  0.49701790  0.9527669   1.220396 90.059121        9

$matrixVarImp.MSE
       glmnet      lasso       ridge       SUM        SUM%    ImpGrad Input_no
8  54.3769931 12.2966484 11.18914488 77.862786 100.000000  0.000000        8
4   2.4608565 34.8635607 33.95608992 71.280507  91.546309  8.453691        4
3   0.9900188 26.4963537 27.00527705 54.491650  69.984202 23.553224        3
1  15.4533567  7.6290822  7.62022766 30.702667  39.431759 43.656199        1
7  22.5447275  0.1135820  0.01212981 22.670439  29.115885 26.161334        7
5   0.0000000  5.2305134  5.51829220 10.748806  13.804805 52.586690        5
6   0.0000000  4.7996971  5.68360889 10.483306  13.463821  2.470038        6
2   3.4511361  3.2642585  3.62760001 10.342995  13.283617  1.338426        2
10  0.0000000  4.6899414  4.89061170  9.580553  12.304406  7.371574       10
9   0.0000000  0.4553821  0.49701790  0.952400   1.223177 90.059029        9

$model
list()
```

Table 11 RMSE and MSE for Elastic net Regression

Elastic net regression (glmnet) outputs input numbers shown in table 11 by arranging the variables from number1 to 10 as labels shown in table 12. In table 12, input number 8 is drought score, 4 is maturity day and 3 is flowering day. Therefore, the most important variables used to predict yield per a plot of land using elastic net regression are drought score, maturity day and flowering day.

```
> FSResults$PPlabels
   Orig Input No          Labels
1              1       Replicate
2              2  Emergence.Day
3              3  Flowering.Day
4              4    Maturity.Day
5              5      Bird.Score
6              6    Insect.Score
7              7   Disease.Score
8              8   Drought.Score
9              9    Plant.Height
10            10      Heads.Plot
```

Table 12 Labels for Elastic net Regression

CHAPTER FIVE

DISCUSSION AND CONCLUSION

5.1 Conclusion

This chapter summarizes results and activities on the questions raised in the study. Research centers contribute to the production of new and improved agricultural products for farmers, thereby contributing to the benefit of people of the country. It also helps students to gain insights and access to information from experts. In Appendix A, some of the agricultural research centers in Ethiopia are mentioned.

The first research question asks what are the factors that increase and decrease productivity of sorghum from trial data. To answer this question, Pearson type of correlation is used. The result in experimentation describes positive and negative relationship between response(dependent) and explanatory (independent) variables of sample sorghum data. In correlation experimentation, replication of plot of land with agricultural process will increase productivity of sorghum. And improved status of variety and good quality of sorghum variety used in a plot of land tells better production of sorghum. Yield per a plot of land has positive correlation with heads per a plot. This means a sorghum plant heads that holds high amount of seeds tells and predicts good amount of productivity in sorghum production. The height of sorghum plant also specifies in increase amount of sorghum production. Highest sorghum plant can have branches with many heads that hold large amount of number of seeds. Negative correlation is seen in relationship of yield per a plot of land and drought score. The relationship tells productivity of sorghum decreases if sorghum plant is highly damaged by drought. Yield in a plot of land decreases for sorghum plant that takes long time to emerge from the soil, display a flower and mature within length of plant maturity period. For sorghum plant attacked by bird, injured with disease and eaten by insect also decrease amount of sorghum production. Even though correlation for research question experimentation indicate weak positive and negative relationship, these factors play a vital role in showing increase and decrease amount of productivity of sorghum.

For second research question experimentation, ridge regression, LASSO and elastic net regression is used in RStudio to show effectiveness of machine learning in predicting productivity of sorghum. Ridge regression reduces variability by shrinking the coefficients, resulting in more prediction accuracy at the cost of usually only a small increase of bias. When the number of predictors is large, ridge regression will not provide an easily interpretable spares model so it is not good for selecting important variables from the data. To overcome erroneous assumption, LASSO is used. LASSO regression model improves prediction accuracy and model interpretability by combining good qualities from ridge regression. For LASSO, if there is high correlation in the group of predictors, LASSO only chooses only one and leaves others by shrinking to zero. This is the limitation for LASSO. To overcome limitation of LASSO, Elastic net regression is used. Elastic net regression combines ridge regression and LASSO. Due to this elastic net is selected for indicating importance of variables from sorghum sample data. Generally, drought score, maturity day and flowering day are the most important variables for predicting yield per a plot of land for experiment taken in a year of 2011 - 2014.

The contribution of this study for experts is identifying constraints of sorghum production that increase and decrease after field experimentation data is gathered from different trials. It also benefits agricultural resaech centers to give more effort on factors that increase and decrease amount of sorghum production. It also helps agricultural research sectors to focus on release improved varieties and good quality of sorghum for achieving better production. Improved varieties benefit farmers to produce increased amount of sorghum which also benefits people of the country in accessing available sorghum foods at all times for individuals. It also helps sorghum experts to do further study on predicting productivity of sorghum on trials based on agro ecological zones.

5.2 Recommendation and Future Work Direction

As recommendation, for further study to produce good sorghum production, Ethiopian agricultural research centers should focus on national based improved sorghum varieties. And in the data of sorghum trail including additional factors such as type of bird, insect and disease attacked the plant during the trial of sorghum production helps to get further data for predicting in productivity of sorghum. In addition, for sorghum plant head number of seeds is better to be included in the dataset.

Results obtained from this thesis provides a future work. The first future work is combining machine learning techniques with image-based processing for identifying factors of sorghum production based on images. The second future work is applying machine learning techniques in genetic and genomic datasets.

REFERENCES

[1] Abduselam Abdulahi Mohamed, “Food Security Situation in Ethiopia: A Review Study,” International Journal of Health Economics and Policy, Vol. 2, No. 3, 2017, pp. 86-96

[Online] Available: http://www.sciencepublishinggroup.com/j/hep

[2]. Alberto Gonzalez-Sanchez, Juan Frausto-Solis and Waldo Ojeda-Bustamante, “Predictive ability of machine learning methods for massive crop yield prediction,” Spanish Journal of Agricultural Research, Vol.12, No.2, 2014, pp. 313-328 [Online] Available: http://revistas.inia.es/index.php/sjar/article/view/4439

[3] Alemu Tirfesa and Assefa Ayele, “Sorghum Prodcution – Technology Manual,” January 2014, pp. 5

[4] Atsbaha Gebre-Selassie and Tessema Bekele, “A Review of Ethiopian Agriculture: Roles, Policy and Small-scale Farming Systems,” pp. 1 [Online] Available: http://global-growing.org/en/content/review-ethiopian-agriculture-roles-policy-and-small-scale-farming-systems

[5] Dr. Harold Macauley, “Cereal Crops: Rice, Maize, Millet, Sorghum, Wheat,” October 2015, pp. 2-10

[6] Fabrizio Balducci, Donato Impedovo and Giuseppe Pirlo, “Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement,” September 2018, pp. 1

[7] Getachew Gashaw, “Response of Sorghum (Sorghum bicolor (L.) Moench) to Plant Population and Nitrogen rates at Fedis, Eastern Ethiopia,” May 2014, pp. 19

[8] Hailegebrial Kinfe and Adane Tesfaye, “Yield Performance and Adoption of Released Sorghum Varieties in Ethiopia,” Edelweiss Applied Science and Technology, Edelweiss Appli Sci Tech 2: 46-55

[9] Gurkamal Deol, “An Introduction to Ridge, Lasso, and Elastic Net Regression,” February 2019 [Online] Available:https://hackernoon.com/an-introduction-to-ridge-lasso-and-elastic-net-regression-cca60b4b934f

[10] Bingran Lu, “ Demystifying the Correlation Matrix,” April 2109 [Online] Available: https://www.datadriveninvestor.com/2019/04/24/demystifying-the-correlation-matrix/#

[11] Ethiopian Institute of Agriculture Research, “Ambo Research Center,” February 2011 [Online] Available:http://www.eiar.gov.et/ambo-plant-protection-research-center

[12] Ethiopian Institute of Agriculture Research, “Bako Maize Research Center,” February 2011 [Online] Available:http://www.eiar.gov.et/index.php/bako-national-maize-research-center

[13] Ethiopian Institute of Agriculture Research, “Chiro Sorghum Research and Training Center,” February 2011 [Online] Available: http://www.eiar.gov.et/index.php/chiro-sorghum-research-and-training-center

[14] Ethiopian Institute of Agriculture Research, “Debre Zeit Research Center,” February 2011 [Online] Available:http://www.eiar.gov.et/index.php/debre-zeit-agricultural-research-center

[15] Ethiopian Institute of Agriculture Research, “Holetta Research Center,” February 2011 [Online] Available: http://www.eiar.gov.et/index.php/en/holeta-agricultural-research-center

[16] Ethiopian Institute of Agriculture Research, “Jimma Research Center,” February 2011 [Online] Available:http://www.eiar.gov.et/index.php/en/jimma-agricultural-research-center

[17] Ethiopian Institute of Agriculture Research, “Kulumsa Research Center,” February 2011 [Online] Available: http://www.eiar.gov.et/index.php/kulumsa-agricultural-research-center

[18] Ethiopian Institute of Agriculture Research, “Mehoni Research Center,” February 2011 [Online] Available: http://www.eiar.gov.et/index.php/mehoni-agricultural-research-center

[19] Ethiopian Institute of Agriculture Research, “Melkassa Research Center,” February 2011 [Online] Available:http://www.eiar.gov.et/index.php/melkassa-agricultural-research-center

[20] Ethiopian Institute of Agriculture Research, “National Agricultural Biotechnology Research Center,” February 2011 [Online] Available:http://www.eiar.gov.et/index.php/national-agricultural-biotechnology-research-centers

[21] Ethiopian Institute of Agriculture Research, “Pawe Research Center,” February 2011 [Online] Available: http://www.eiar.gov.et/index.php/Pawe-Agricultural-Research-Center

[22] Ethiopian Institute of Agriculture Research, “Assosa Research Center,” February 2011 [Online] Available: http://www.eiar.gov.et/index.php/2017-01-26-13-34-49

[23] Ethiopian Institute of Agriculture Research, “Tepi Spices Research Center,” February 2011 [Online] Available:http://www.eiar.gov.et/tepi-national-spices-research-center

[24] Ethiopian Institute of Agriculture Research, “Werer Research Center,” February 2011 [Online] Available: http://www.eiar.gov.et/werer-agricultural-research-center

[25] Ethiopian Institute of Agriculture Research, “Wendo Genet Research Center,” February 2011 [Online] Available: http://www.eiar.gov.et/wondo-genet-agricultural-research-center

[26] Jérôme Treboux, “Improved Machine Learning Methodology for High Precision Agriculture,” pp. 1

[27] J.M.J. de Wet, “Systematics and Evolution of Sorghum Sect.Sorghum (Gramineae),” Americal Journal of Botany, Vol.65, No.4 (Apr., 1978), 477-484. [Online Available: http://links.jstor.org/sici?sici=0002-9122%28197804%2965%3A4%3C477%3ASAEOSS%3E2.0.CO%3B2-8.

[28] Kodimalar Palanivel, “An approach for Prediction of Crop Yield using Machine Learning and Big Data Techniques,” International Journal of Computer Engineering and Technology (IJCET), Volume 10, Issue 03, May-June 2019, pp. 110-118, [Online] Available: http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=10&IType=3

[29] Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson and Dionysis Bochtis, “Machine Learning in Agriculture: A Review,” August 2018, pp. 3 – 4

[30] Mamoudou H. Dicko, Harry Gruppen, Alfred S. Traoré1, Alphons G. J. Voragen and Willem J. H. van Berkel, “Sorghum grain as human food in Africa: relevance of content of starch and amylase activities,” African Journal of Biotechnology Vol. 5 (5), March 2006, pp. 384-395. [Online] Available: http://www.academicjournals.org/AJB.

[31] Sciforce, “Machine Learning in Agriculture: Applications and Techniques,” March 2019 [Online] Available: https://medium.com/sciorce/machine-learning-in-agriculture-applications-and-techniques-6ab5014d1b5

[32] Muthukrishnan R and Rohini R, “LASSO: A Feature Selection Technique in Predictive Modeling for Machine Learning,” IEEE International Conference on Advances in Computer Applications (ICACA) 2016, pp. 1-2

[33] Nils J. Nilsson, “Introduction to Machine Learning,” November 1998, pp. 12 – 13

[34] Peter Harrington, “Machine Learning in Action,” Manning Publications, 2012, Ch. 1 pp. 5

[35] Robert J. McQueen,Stephen R.Garner, Craig G.Nevill-Manning and Ian H.Witten “Applying Machine Learning to Agricultural Data,” Volume 12, Issue 4, June 1995, pp. 275 – 293 [Online] Available: https://www.sciencedirect.com/science/article/pii/0168169995986019

[36] Stephen R. Garner, “WEKA: The Waikato Environment for Knowledge Analysis,” 1995 pp. 6 [Online] Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.3371

APPENDICES

Appendix A

Agricultural Research Centers

Ambo Research Center

The center was established by a bilateral agreement among the Ethiopian and the then Union of Soviet Socialist Republics (USSR) government in 1977 by the name of pathological laboratory. During this time, the research focus was only on plant diseases. After some years of its establishment Agricultural Entomology and Weed Science research was launched in addition to the plant diseases research. In 1989, due to the termination of the bilateral agreement, the center was restructured under the then Ethiopian Science and Technology Commission, now called Ministry of Science and Technology. In 1994, over again, the Center was restructured under the then Institute of Agriculture Research (IAR), now EIAR. Currently, the Center is working as one of the federal research centers of EIAR [11].

Total area: 149.215 ha

Location (Latitude/longitude): 8° 57' N and 38° 07' E

Agro-ecological zones (AEZs): Temperate (Intermediate high land)

Altitutde:2175 m.a.s.l.

Temperature (10 years): 10.02°C min and 26.89°C Max (Mean: 18.44°C)

Rainfall: 1018.19 mm

Length of growing period (LGP): 4-5 months

Soil type and coverage in percent: Vertisols consisting of 67% clay, 18% silt, 15 % sand and 1.5% organic.

Coordination role: plant protection research [11]

Assosa Research Center

Major deriving force for the establishment of the Assosa Agricultural Research Center was the serious famine incident in 1986. The Center, originally known as Assosa Research Station carried out research activities under the close supervision of Bako Research Center. The station was put up on an area of 271.84 ha in village 12 with few research staff supporting the agricultural activities of the settlers while research was the primary objective. After some years of operation, the Center was totally devastated during the last civil unrest in 1990. After 15 years of closing from this year, the research center started its activity in Nov 5, 2005 under the Ethiopian Institute of Agricultural Research [22].

Establishment year:1986

Location (Latitude/longitude): 10° 03’ N and 34° 59’ E

Altitude: 1580 m.a.s.l.

Temperature (min/max):14°C and 39°C

Rainfall: 1,275 mm

Length of growing period (LGP): 4-7 months

Soil type: Dystric Nitosols [22].

Chiro Sorghum Research and Training Center

Chiro Sorghum Research and Training Center is a recently established federal research center in eastern part of Ethiopia, in Chiro town from where the Center name is derived from. The Center commenced its research and training activities on May 2014 with an ultimate aspiration to become a center of excellence mainly for sorghum research and training. By playing a leading role in sorghum research and training, the Center envisions bringing about market-led sorghum production system through the promotion of improved sorghum technologies and husbandry practices in the commodity growing areas of the country [13].

Total area: 7.5 ha

Establishment year: 2014

Location (Latitude/longitude):9° 03' 49" N, 40° 52' 26" E(Latitude/longitude)

Agro-ecological zones (AEZs): Intermediate altitude

Altitude: 1855m.a.s.l.

Rainfall: 980mm

Length of growing period (LGP):175-240 days

Coordination role: Sorghum research [13].

Bako Maize Research Center

Bako Maize Research Center is one of the oldest research centers in the country. It was established in 1964 with an agreement signed between the Federal Republic of Germany and the Government of Ethiopia. The Center is located at 8 km from the nearest District's town, Bako. Since 1991, the main Center (Bako Agricultural Research Center) has been governed by the Oromia Regional Government while Bako National Maize Research Center has been maintained under Federal Government of Ethiopia. Based on the MoU signed between the federal and regional governments, the two centers have shared common resident house and transportation service, which are under the administration of regional research center [12].

Total area: 40 ha

Establishment year: 1964

Location (Latitude/longitude): 9° 06' N and 37° 09' E

Agro-ecological zones (AEZs): Mid-altitude sub-humid

Altitude: 1590 m.a.s.l.

Temperature (min/max): 9°C min and 34.4°C (Mean: 19°C)

Rainfall: 12.45mm

Soil type: Nitosol

Excellence: Maize [12].

Debre Zeit Research Center

Debre Zeit Agricultural Research Center was established in 1953 under the then Alemaya Agricultural College, now Harmaya University. The Center was then transferred to the then Ethiopian Agricultural Research Organization in 1984.Debre Zeit Agricultural Research Center (DZARC) is mandated for the improvement of tef, durum wheat, chickpea, lentil, forage crops, vegetables (shallot and garlic), fruits (grapevine), poultry, dairy, small ruminants, animal nutrition, and land and water [14].

Total area: 147 ha

Establishment year:1953

Location (Latitude/longitude): 08° 44’ N and 38° 58’ E

Agro-ecological zones (AEZs): Tepid to cool sub-moist highlands (SM2)

Altitude: 1900 m.a.s.l.

Temperature (min/max): 8.9°C min and 28.3°C (Mean: 19°C)

Rainfall

Type (bimodal/unimodal): Bimodal (small rains occur between February and April)

Annual total (mm):851 mm

Length of growing period (LGP): 61-120 days

Soil type and coverage in percent: Light soils (Alfisols/Mollisols, 20 ha or 14% and mostly used for construction and other purposes. Black soils (Vertisols) also constitute 127 ha, 86% [14].

Holetta Research Center

Holetta Agricultural Research Center (HARC) was established in 1966 under the Institute of Agricultural Research (IAR), now EIAR. The Center is located in Holetta town and has two sub-centers and two testing sites. Ginchi and Ada Berga subcenters focus on Vertisol and dairy research, respectively, while Adadi and Jeldu testing sites are for highland crops research activities [15].

Total area: 396 ha

Establishment year:1966 GC

Location (Latitude/longitude): 9º 00’ N and 38º30’ E

Agro-ecological zones (AEZs): M2-5

Altitutde:2400 m.a.s.l

Temperature (min/max): 6ºC and 22ºC

Rainfall: 1144 mm

Length of growing period (LGP): 4-6 months

Soil type and coverage: Nitosols and Vertisols

Coordination Role: dairy and forage crops, highland (temperate) fruits, acidic soil management, Vertisol management, soil fertility and plant nutrient management, highland oilseeds, highland pulses [15].

Jimma Research Center

Jimma Agricultural Research Center was established in 1967. Currently, the Center has two subcenters: Gera and Haru and three testing sites, i.e., Mettu, Mugi and Agaro. The Center is mandated to coordinate coffee research nationally. Moreover, nationally, integrated soil fertility and crop production, and integrated watershed management research are coordinated by the Center [16].

Total area: 183 ha

Establishment year: 1967

Location (Latitude/longitude): 7º46' N and 36º 00'E

Agro-ecological zones (AEZs): Sub humid tepid to cool mid highlands

Altitutde:1753 m.a.s.l.

Temperature (min/max): 9ºC and 28ºC

Rainfall: 1561 mm

Length of growing period (LGP): 4-7 months

Soil type and coverage: Upland: Chromic Nitosol and Combisol;Bottom land: Fluvisol [16].

Kulumsa Research Center

Kulumsa Agricultural Research Center was established in 1966 by government of Ethiopia and the Swedish International Development Agency (SIDA). The research Center is mandated to wheat, malt barley and highland pulse crops research nationally and serves as Wheat Center of Excellence for East Africa (Ethiopia, Kenya, Uganda, Tanzania), regionally [17].

Total area: 442.7 ha

Establishment year: 1966

Location (Latitude/longitude): 8º2' N and 39º10'E

Agro-ecological zones (AEZs): From cool highland to semi-arid

Altitutde:2200 m.a.s.l.

Temperature (min/max): 10ºC and 22ºC

Rainfall:840mm

Length of growing period (LGP):120-135 days

Soil type: Clay soil

Coordination role: wheat, faba bean and field pea research [17]

Mehoni Research Center

Mehoni Agricultural Research Center was established on March 2004 to conduct research on lowland high-value tree crops and horticulture crops. The Center conducts its research activities on the Fachagama Research Site, located in Raya Azebo District of the Tigray Region. The Center has also engaged in source seed multiplication for those of the crop varieties that are suitable to nearby districts of the region [18].

Total area: 100 ha

Establishment year: March 2004

Agro-ecological zones (AEZs): lowland

Altitutde:1574 m.a.s.l.

Temperature (min/max):18ºC and25ºC

Rainfall: Raya Azebo woredas: 300-750mm

Length of growing period (LGP): 7-8 months

Soil type: Vertisols, Aluvisols

Coordination role: Lowland high value tree crops and horticultural crops research [18].

Melkassa Research Center

Melkassa Agricultural Research Center (MARC) was established in 1969 as a national horticulture research center with staff members of five personnel and land area not exceeding five hectares. Over the years, other important research programs and commodities have been included in the research program that contributed to improved rainfed and irrigated agricultural development with a main focus on dry and lowland areas of moisture stress and irrigable areas of the country [19].

Total Area: 275ha

Establishment year: 1969

Location (Latitude/longitude): 8º24'N and 39º21'E

Agro-ecological zones (AEZs): Arid to semi-arid

Altitude: 1550 m.a.s.l

Temperature (min/max): 14ºC and 28.4ºC

Rainfall: 763 mm

Length of growing period (LGP):3-6 months

Soil type: Andosol of volcanic origin with pH ranging from 7 to 8.2.

Coordination Role: Sorghum, Low land Pulses, Lowland Maize, Horticulture (Vegetable and Tropical Fruits), Sericulture, Climate and Geospatial, Agricultural Mechanization, Water shade Management [19].

Pawe Research Center

The Center came in to existence on 1986; a year after Ethiopia was smashed by the worst famine in the living memories that claimed breathes of millions of its citizens, succeeding the resettlement of the highly-affected people from the northern and southern part of the country to deliver improved technology options. At the time of establishment, the Center was intended to provide support for the settlers. As a result, putting trials on the ground was started with the aim to identify crops which best suit the agro-ecology[21].

Total Area: 294.18 ha

Establishment year:1985 GC

Location (Latitude/longitude):11°19’N and 36°24'E

Agro-ecological zones (AEZs): Hot to warm moist

Altitude: 1120 m.a.s.l.

Temperature (min/max):16.3°C and 32.6°C

Rainfall max: 1587mm

Length of growing period (LGP): 5-7 months

Major soil types: Nitosols, Vertisols and Luvisols

Coordination Role: Soybean [21].

Tepi Spices Research Center

Tepi Spices Research Center (TNSRC) is one of the centers of the Ethiopian Institute of Agricultural Research (EIAR). It used to be one of the testing sites under Jimma Research Center of the EIAR since 1973 for coffee and spices research, representing hot to warm humid lowland agro-ecologies of the country. Now the center coordinates the national spice research activities [23].

Total area:104 ha

Location (Latitude/longitude): 7º 3' N and 35º18’ E

Agro-ecological zones (AEZs): Hot to warm humid low, Lands (H1), hot to warm, sub-humid lowlands (SH1)

Altitutde:1200 m.a.s.l.

Temperature (min/max): 15ºC and 30ºC

Rainfall: 1522.1 mm

Length of growing period (LGP): 170-270 days

Soil type: Nitosols, red coffee soils, fertile loam soils

Coordination Role: Spices [23].

Wendo Genet Research Center

Wendo Genet Agricultural Research Center was established in 2009 to conduct agricultural research mainly on Aromatic, medicinal and bio-fuel plants. before this year, the Center was a perfume factory owned by Belgium and France citizens who were engaged in extracting and producing perfumes and aromatic oils up until. Subsequently, the Center was operational under the then Ethiopian Chemical Corporation and Ethiopian Biodiversity Institute (EBI) up to 2003. Since then the center had been served as sub-center for Aromatic Oils Research Center up until 2008. In 2009 the Center was upgraded to a level of research center under the Ethiopian Institute of Agricultural Research [25].

Land Holding (ha): 113.5 ha, that is, forest 54.3 ha, open woodland 19 ha, field site trial: 31.9 ha, field genetic bank 3.1 ha, built up area, 4.6 ha, road 4.6 ha

Location (Latitude/longitude): 7º 19' N and 38º 38' E

Agro-ecological zones (AEZs): H3 (Tepid humid highland)

Altitude: 1760-1920 m.a.s.l.

Temperature (min/max):11.5 ºC and 26.2 ºC

Rainfall: 1372 mm

Length of growing period (LGP):210-239 days

Soil type and coverage: Luvisols (sandy loam with pH 5.2-7.7)

Coordination Role: Aromatic, medicinal and bio-energy plants [25].

Werer Research Center

Werer Agricultural Research Center, the then called Melka Werer Agricultural Research Station, was established in 1964 by the Research Department found under the Ministry of Agriculture. It was established to conduct research on cotton crop with the objective of giving support to the Cotton Farms established in Awash Valley. After two years, the station was transferred to the newly established Institute of Agricultural Research (IAR) in 1966. Presently, the center is one of the federal agricultural research centers conducts research activities on Cotton, Oil Crops and Irrigated Agriculture [24].

Land holding:56 ha Research Farm: 200 ha Residence: 47 ha other purpose

Total: 301 ha

Location (Latitude/longitude): 9º16'N and 40º9'E

Agro-ecological zones (AEZs): AI (Hot to warm humid low land)

Altitude: 750m m.a.s.l.

Temperature (min/max): 26.7ºC and 40.8ºC

Rainfall:590 mm

Length of growing period (LGP): 3-5 Months

Soil type: Vertisols and flui-soil

Coordination Role: Cotton and irrigated agriculture crops [24].

National Agricultural Biotechnology Research Center

The National Agricultural Biotechnology Research Center, is one of the recently established research centers of the Ethiopian Institute of Agricultural Research. Before the center was upgraded to a full-fledged research center level, it was a National Agricultural Biotechnology Laboratory (NABL). The Laboratory was established as an independent biotechnology research facility under the administration of Holetta Agricultural Research Center (HARC) on September

2010.Currently, the Center conducts research activities at four of its laboratories in areas of (1) plant tissue culture biotechnology research, (2) central molecular biotechnology research, (3) animal biotechnology research, and (4) microbial biotechnology research [20].

Establishment year:2015

Location (Latitude/longitude): 9º4’ N and 38º30’ E

Agro-ecological zones (AEZs): M2-5

Altitutde:2400 m.a.s.l

Temperature (min/max): 6 ºC and 22 ºC

Rainfall: 1100 mm average

Length of growing period (LGP): 4-6 months

Soil type and coverage: Nitosols and Vertisols

Coordination Role: Agricultural biotechnology [20].

Appendix B

Machine Learning in Agriculture: Applications and Techniques

Machine learning is everywhere throughout the whole growing and harvesting cycle. It begins with a seed being planted in the soil—from the soil preparation, seeds breeding and water feed measurement—and it ends when robots pick up the harvest determining the ripeness with the help of computer vision.

Species management

Species Breeding

Our favorite, this application is so logical and yet so unexpected, because mostly you read about harvest prediction or ambient conditions management at later stages. Species selection is a tedious process of searching for specific genes that determine the effectiveness of water and nutrients use, adaptation to climate change, disease resistance, as well as nutrients content or a better taste. Machine learning, in particular, deep learning algorithms, take decades of field data to analyze crops performance in various climates and new characteristics developed in the process. Based on this data they can build a probability model that would predict which genes will most likely contribute a beneficial trait to a plant.

Species Recognition

While the traditional human approach for plant classification would be to compare color and shape of leaves, machine learning can provide more accurate and faster results analyzing the leaf vein morphology which carries more information about the leaf properties.

Field conditions management

Soil management

For specialists involved in agriculture, soil is a heterogeneous natural resource, with complex processes and vague mechanisms. Its temperature alone can give insights into the climate change effects on the regional yield. Machine learning algorithms study evaporation processes, soil moisture and temperature to understand the dynamics of ecosystems and the impingement in agriculture.

Water Management

Water management in agriculture impacts hydrological, climatological, and agronomical balance. So far, the most developed ML-based applications are connected with estimation of daily, weekly, or monthly evapotranspiration allowing for a more effective use of irrigation systems and prediction of daily dew point temperature, which helps identify expected weather phenomena and estimate evapotranspiration and evaporation.

Crop management

Yield Prediction

Yield prediction is one of the most important and popular topics in precision agriculture as it defines yield mapping and estimation, matching of crop supply with demand, and crop management. State-of the-art approaches have gone far beyond simple prediction based on the historical data, but incorporate computer vision technologies to provide data on the go and comprehensive multidimensional analysis of crops, weather, and economic conditions to make the most of the yield for farmers and population.

Crop Quality

The accurate detection and classification of crop quality characteristics can increase product price and reduce waste. In comparison with the human experts, machines can make use of seemingly meaningless data and interconnections to reveal new qualities playing role in the overall quality of the crops and to detect them.

Disease Detection

Both in open-air and greenhouse conditions, the most widely used practice in pest and disease control is to uniformly spray pesticides over the cropping area. To be effective, this approach requires significant amounts of pesticides which results in a high financial and significant environmental cost. ML is used as a part of the general precision agriculture management, where agro-chemicals input is targeted in terms of time, place and affected plants.

Weed Detection

Apart from diseases, weeds are the most important threats to crop production. The biggest problem in weeds fighting is that they are difficult to detect and discriminate from crops. Computer vision and ML algorithms can improve detection and discrimination of weeds at low cost and with no environmental issues and side effects. In future, these technologies will drive robots that will destroy weeds, minimizing the need for herbicides.

Livestock management

Livestock Production

Similar to crop management, machine learning provides accurate prediction and estimation of farming parameters to optimize the economic efficiency of livestock production systems, such as cattle and eggs production. For example, weight predicting systems can estimate the future weights 150 days prior to the slaughter day, allowing farmers to modify diets and conditions respectively.

Animal Welfare

In present-day setting, the livestock is increasingly treated not just as food containers, but as animals who can be unhappy and exhausted of their life at a farm. Animals behavior classifiers can connect their chewing signals to the need in diet changes and by their movement patterns, including standing, moving, feeding, and drinking, they can tell the amount of stress the animal is exposed to and predict its susceptibility to diseases, weight gain and production [31].