Identification and Classification of Highly Productive Lentil Varieties Affected by Disease Using Machine Learning Techniques

Ebisa Bekele Fite

A Thesis Submitted to

Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of The Requirement for The Degree of Masters of Science in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

August 2021

Identification and Classification of Highly Productive Lentil Varieties Affected by Disease Using Machine Learning Techniques

Name: Ebisa Bekele Fite

Advisor: Dr. Biruk Yirga

A Thesis Submitted to Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of The Requirement for The Degree of Masters of Science in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

August 2021

DECLARATION

I hereby declare that this Master Thesis entitled "**Identification and Classification of Highly Productive Lentil Varieties Affected By Disease Using Machine Learning Techniques**" is my original work. That is, it has not been submitted for the award of any academic degree, diploma, or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation.

Ebisa Bekele _____ _____

Name of the Student Signature Date

RECOMMENDATION

I/we, the advisor(s) of this thesis, hereby certify that I/we have read the revised version of the thesis entitled **"Identification and Classification of Highly Productive Lentil Varieties Affected by Disease Using Machine Learning Techniques"** prepared under my/our guidance by **Ebisa Bekele Fite** submitted in partial fulfillment of the requirements for the degree of Master's of Science in Computer Science and Engineering. Therefore, I/we recommend the submission of a revised version of the thesis to the department following the applicable procedures.

Dr. Biruk Yirga _____ _____

Major Advisor Signature Date

APPROVAL SHEET

I/we, the advisors of the thesis entitled "**Identification and Classification of Highly Productive Lentil Varieties Affected by Disease Using Machine Learning Techniques**" and developed by Ebisa Bekele, hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

Major Advisor	Signature	Date
Co-advisor	Signature	Date

We, the undersigned, members of the Board of Examiners of the thesis by Ebisa Bekele Fite have read and evaluated the thesis entitled "**Identification and Classification of Highly Productive Lentil Varieties Affected by Disease Using Machine Learning Techniques"** and examined the candidate during the open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering.

Chairperson	Signature	Date
Internal Examiner	Signature	Date
External Examiner	Signature	Date

Finally, approval and acceptance of the thesis is contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

Department Head	Signature	Date
School Dean	Signature	Date
Office of Postgraduate Studies, Dean	Signature	Date

ACKNOWLEDGEMENT

First and Above all I would like to thank my almighty God, who paid his soul and gave me a life, for blessing me with a curious mind and good health that would help me to finalize this work.

I would like to express my deepest appreciation to my advisor Dr. Biruk Yirga for his valuable advice during this work by giving helpful advice, constructive criticism, and insightful suggestions.

Next, I really thank Mr. Asefa Funga one staff member of the Debre-Zeit agricultural research center for encouraging me during this work, by giving me some useful documents and data for implementation purposes. Mr. Asefa may God bless you for all your kindness.

I cannot leave my classmate students of the CSE department without mentioning them. I had a great pleasure to have friends like you. Thanks for sharing your knowledge and experiences with me. Finally, I would like to thank my beloved family. Mom and Dad, I am really glad to have you. If it wasn't for you, I couldn't be here, thanks for wishing me a bright future in every situation. May God bless you, wish you a long life.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATION

Acronyms

A

ANOVA · *Analysis of Variance*

C

CV · *Cross-Validation*

D

DT · *Decision Tree*
DZARC · *Debre Zeit Agricultural Research Center*

E

E.C · *Ethiopia Calender*

F

FN · *False Negative*
FP · *False Positive*

G

GUI · *Graphical User Interface*

I

ICARDA · *International Center for Agricultural Research in the Dry Areas*
ICRISAT · *International Crops Research Institute for the Semi-Arid Tropics*
IDE · *Integrated Development Environment*

M

MAE · *Mean Absolute Error*
ML · *Machine Learning*
MS · *Microsoft*
MSE · *Mean Squared Error*

N

NB · *Naïve Bayes*

P

PCA · *Principal Component Analysis*

R

RF · *Random Forest*
RFE · *Recursive Feature Elimination*
RQ · *Research Question*

S

SMOTE · *Synthetic Minority Oversampling Technique*
SVM · *Support Vector Machine*

T

TN · *True Negative*
TP · *True Positive*

U

UFS · *Univariate Feature Selection*

Abstract

Lentils are one of the most important pulse crops produced in Ethiopia. Our country ranks on the tenth level in the world and first in Africa for the volume of lentils produced. The most significant task in agriculture is finding improved crops to keep food security. The agricultural research centers focus on a breeding program to improve crops production. Genetic variety identification between genotypes of lentils is important for selecting genotypes for breeding programs. Accordingly, agricultural research centers doing different research on breeding programs to improve lentils production, but character analysis of lentil genotypes to find genetic diversity and lentils yield prediction without sowings are the active problems. To find the solution to these problems machine learning Techniques were proposed in this study. The data used in this research was collected from the Debre-Zeit agricultural research center. Using the K-means clustering algorithm, centroid initialized by K-means++ our dataset clustered into seven different groups of lentil genotypes. Each cluster was analyzed using traits of genotypes and compared to find highly productive lentil genotypes and cluster VII was highly productive. The clustering process was used for labeling datasets for the classification learning algorithms. 10-folds cross-validation was used to evaluate the regression learning and classification learning algorithms. The classification task was performed by Support Vector Machine(SVM), Random Forest(RF), Naïve Bayes(NB), and Decision Tree(DT) models to classify the newly inputted genotypes. Each algorithm was evaluated and compared based on their performance evaluation result, and the support vector machine has better accuracy of 99.43%. The yield prediction task of lentil genotypes affected by the disease was performed by Random Forest(RF), Gradient Boosting(GB), and Multiple Linear Regression(MLR) models. Those models were implemented without and with UFS and RFE feature selections. The random forest has better performance with MAE=17.51, MSE=1089.18, R-squared=0.96. The recommendation of lentil genotypes for the breeding process task was applied by a collaborative-based filtering algorithm. This study reveals that machine learning algorithms are important in the agricultural sector, to cluster, classify, yield prediction, and recommendation of lentil genotypes. We hope this study will be helpful for more findings concerning the identification and classification of lentil varieties using machine learning.

***Keywords**: Lentils, Clustering, K-means, genotypes identification, lentil genotypes recommendation, yield prediction, genotypes classification*

CHAPTER ONE

1. INTRODUCTION

1.1 Background of The Study

Ethiopia is gifted with many agricultural possessions as well as a broad range of natural zones. To increase the production of smallholder farms and expand large-scale commercial farms, the government has designated important priority intervention areas (Diriba 2020)(Administration 2020). In today's world, parallel to population growth, the food demand is also growing increasingly (Frona, Janos, and Harangi-Rakos 2019). Pulses play a vital role in the agriculture and diet of many developing countries and as a major source of dietary nutrients for numerous individuals. In Ethiopia, lentil is the eleventh most produced crop. Lentils are one of the most imperative pulses grown in our country. However, the low yields are mainly due to the use of low-yielding local varieties and soil fertility management practices; Our country ranks on the tenth level in the world and first in Africa for the volume of lentils produced (Dugassa, Legesse, and Geleta 2015).

Lentil is one of the highland pulses that are produced in cool and semi cooler climatic domains(Saikenova et al. 2021). Lent il contains protein, minerals, and carbohydrates. It is called a major source of protein 28 percent for human consumption and its straw is a valued animal feed consisted of minerals 2 percent and carbohydrates 59 percent (Kassa 2018). By occupying a unique position in cereal-based cropping systems, lentil plays a crucial role in health enhancement. It can fix nitrogen and capture and store CO_2 from the environment, which improves soil nutrient status and offers agricultural production systems with long-term sustainability (Kindie, Nigusie, and Yildiz 2018). Improving the production of lentils is very important because most of the world's poor people use lentils for their daily diet. In particular, Ethiopia benefited from lentils as their diet, source of income, and health benefits. Lentils contain good amino acids that contain all crucial amino acids enough for humans in proportions as suggested by the world health organization (Bogale, Mekbib, and Fikre 2017).

There are eleven varieties of lentils released between 1980-2010 in Ethiopia those are: ER-142, R-186, Chekol, Chalew, Adaa, Gudo, Alemaya, Assano, Alemtena, Teshale, Derso. Finding the lentil varieties started during 1970 to find the varieties that have improved productivity by the collaboration of the ICARDA (International Center for Agricultural Research in the Dry

Areas) and ICRISAT (International Crops Research Institute for the Semi-Arid Tropics). Except for ER-142 the others are derived from ICARDA(Regassa, Senait, Legesse Dadi, Demissie Mitiku, Asnake Fikre 2014). An efficacious breeding program is likely to produce a genetic improvement in yield overtimes that is one component of grain yield potential (Niazian and Niedbała 2020).

Genetic variability of lentils can be identified based on the agro-morphological traits of each genotype. To estimate the yield of the lentils there are yield components data, agronomic traits, and disease data of lentils those components are yield contribution traits like plant height, hundred-seed weight, seeds pod per plant, the number of pods per plant, grain yield, above-ground biomass, and other agronomic traits (Kassa 2018). Finding the genetic variability between genotypes, accurate yield prediction, and recommendation of genotypes for breeding purposes are important issues in supporting lentils production. To do this the machine learning techniques are proposed in this study.

Machine learning is the most evolving technology in today's world. Artificial Intelligence has given birth to this field. We can create better and smarter machines by using artificial intelligence (Chlingaryan, Sukkarieh, and Whelan 2018). Machine Learning is a strategy for learning from previous experiences and precedents without being explicitly programmed. Instead of writing code, the data is provided into a generic algorithm, which generates logic based on the information provided. Machine learning is used in the agriculture sector, health sector, spam filtering, ad placement, stock trading, and so on. Machine learning is valuable in agriculture to processing large datasets beyond the scope of human ability. It reliably converts analysis of that data into sector insights that aid agricultural workers in planning and estimating, ultimately leading to better outcomes (Chlingaryan, Sukkarieh, and Whelan 2018). Machine learning applications in Agriculture are yield prediction, quality assessment, species identification, crop management, crop disease, weed detection, the recommendation for the breeding program, and so on. Plants may have similar colors, leaf compositions, and shapes, it is hard to classify them using the human eye. We can now rely upon machine learning to accurately identify related plants, weed species, identifying plant genotypes, yield predictions, and recommendations to increase productivity (Computing 2019). This study was aimed to work on yield prediction and identification of genetic diversity between genotypes, characteristics analysis of each cluster, and recommend genotypes for the breeding program. The major objectives of this study were to find genetic variability between genotypes and cluster them based on agronomic traits and disease data to identify and classify genotypes that

produce good products and recommend them for the breeding program. In this study, regression algorithms of machine learning techniques used to predict yield of lentils, a clustering algorithm used for grouping genotypes of lentils depending on their traits, and classification algorithm to classify new genotypes to identify and classify highly productive lentil varieties that affected by the disease using a dataset collected from DZARC.

The data used in this research for prediction and clustering were collected from the Debre-Zeit agriculture research center. Debre-Zeit research center was established in 1953 under the Alemaya agricultural college, now a Haramaya university. In 1984 research center was then transferred to the Ethiopian agricultural research organization. Debre Zeit Agricultural Research Center (DZARC) is assigned for the enhancement of lentil, durum wheat, vegetables, chickpea, tef, forage crops, poultry, animal nutrition, dairy, fruits, small ruminants, and land and water[1]. Those data are yield data and data contains agronomic traits and disease data.

1.2 Motivation

In our country majority of our people are directly or indirectly engaged in agriculture. However, the agricultural sector is still traditional in our country. The agricultural sector needs to be supported by various techniques to be improved and improve production. Agriculture needs to be converted to modern farming to meet the needs of the traditional farming method. Finding the genetic variability to select genotypes of lentils for the breeding programme is an important task and this task is very difficult and effortful by human ability.

Lentils are often used as a daily diet in Ethiopian households. We are inspired to work on the identification and classification of highly productive lentil varieties affected by disease using machine learning because of its importance and usefulness for agricultural workers to save their time, accurate prediction. The other thing that motivates us to select this area is that the workers or agricultural experts can use the clustered genotypes of lentils to select good genotypes for breeding programs to get genetic improvement.

Lentil has a natural ability to nitrogen fixation and carbon dioxide capturing that is used in soil nutrient improvement and finding the good lentil varieties that give high product minimizes the use of inorganic fertilizer and is useful in managing soil by using naturally nitrogen fixer crop.

[1] http://www.eiar.gov.et/index.php/debre-zeit-agricultural-research-center

The main issue that initiates us to do this study is that the classical statistical analysis models are common methods for analyzing the characteristics of lentil genotypes and to find genetic variability. These methods may fail for the large dataset, machine learning is interesting to analyze the characteristics of lentil genotypes, and also it is an interesting area in prediction, recommendation, and clustering. To facilitate learning and build knowledge for researchers is another motivation of doing this study.

In this study, the proposed system is to identify and classify the highly productive lentil varieties by clustering the lentil genotypes, and predict the yield of lentils depends on their agronomic traits and disease data by using machine learning algorithms. The developed model is tested depends on training data using a test set. The system is significantly useful in identification and classification of lentil genotypes to increase the productivity of lentils that will lead to keeping food security and enhancing the breeding programs.

1.3 Statement of The Problem

Increasing the yield of pulses in a developing country is an important issue to improve food security and commercial problem. Problems raised in agriculture domains need different resources to increase yield. Though the majority of our people are farmers and they are surviving their life by engaging in farming, there is a lack of food security (Stellmacher and Kelboro 2019). They utilize a lot of effort and manpower to farm large areas but they do not get a product that is commensurate with their hard work. These problems may come from different perspectives like the way they choose the crop species for their agriculture, selecting genotype for breeding programs, shortage of great yielding varieties, and identification of disease-resistant genotypes. The lentils breeding program needs the identification of lentil genotypes with good quality (high yield, disease-resistant, and so on).

The problems raised in this study to be solved were difficulties in the analysis of lentil genotypes characteristics to find genetic variability; genetic variability is useful to know the divergence between genotypes. Inaccurate and difficulty in yield prediction to identify highly productive lentil genotypes is another problem. Difficulty to select genotypes for breeding purposes also problem that needs raised in this study. There were many genotypes of lentils produced in Ethiopia in different experimental areas. Identification and holding the genotypes which may have a good yield without analysis of their character is difficult.

The other problem was after some experiments done by the researchers; it may need additional techniques to cluster the lentil genotypes into different groups depending on their traits and products. To find out genotypes with a quality product machine learning algorithms are useful. Machine learning technology has a solution for such kinds of problems to learn from previous experience or sample of data and give appropriate solutions to the newly inputted data. In this study problem raised above tried to be solved by machine learning algorithms.

1.4 Research Questions

In this study, the following research questions are addressed and answered.

RQ1: What are a set of features that affect lentils yield to develop a machine learning model for the yield prediction?

RQ2: Which machine learning techniques are useful in clustering and classification of lentil genotypes to analyze their characteristics?

RQ3: How to build the recommender system to enhance the breeding programs?

RQ4: How to evaluate the performance of machine learning models for the identification and classification of highly productive lentil genotypes?

1.5 Objectives

1.5.1 General Objective

The main objective of this study is to design and implement machine learning models that can identify, classify and recommend highly productive lentil genotypes for breeding program.

1.5.2 Specific Objectives

To achieve the above general objective and to give the solution to the problem raised, the specific objectives of this study are:

- To collect data and pre-process it to build a quality dataset
- To train our dataset to identify the correlation between independent characteristics and yield of lentil genotypes to predict the yield using regression models.
- To implement the significant machine learning model to cluster lentil genotypes and analyze the characteristics of each cluster.

- To train different classification algorithms to classify new genotypes into different clusters.
- To Evaluate the machine learning model's performance to apply for the identification and classification of highly productive lentil varieties.
- To recommend existing genotypes of lentils for the breeding program.
- To build a GUI system and deploy with a better-performing machine learning model on the server.
- To report the finding of the study for the upcoming research area.

1.6 Significance of The Study

Agriculture plays an important role in the economy of Ethiopia. The need for nutrition and the population of our country are not equivalent. To improve food security there are different importance's of this research in our country. This work is important for the agricultural research centers. The following are the important issues that are used to study the identification and classification of highly productive lentils varieties affected by disease using the machine learning techniques.

- Accurate yield prediction.
- Easy analysis of lentils characteristics and clustering of different genotypes of lentils depends on agronomic traits and disease data.
- Easy to select genotypes for the breeding program.
- Easy to classify new genotypes to different groups of a cluster.
- The recommendation of genotypes for the breeding program

1.7 Scope and Limitation of The System

1.7.1 Scope of The Study

The focus area of our study was the Debre-Zeit agricultural research center based on different locations and different lentil varieties regarding geographical size. In this study, forty-three genotypes of lentils were collected from six different locations. Genotypes identification and classification is essential for holding a good lentils variety and way of enhancing breeding program for genetic improvement. Functionally, this work covers predicting lentils yield, clustering lentil genotypes based on their characteristics, characteristics analysis of genotypes, and recommending genotypes for breeding purposes using machine learning. In this study different machine learning models were used for prediction and clustering. The regression

learning model for predicting yield, a clustering algorithm for clustering our dataset, and a classification learning model to classify new genotypes into groups of genotypes.

1.7.2 Limitation of The Study

The limitation of this study was limited access to data in the research center. The other limitation of this study is that it only depends on agronomic traits and disease data; soil and weather conditions are not considered. The soil and weather data are important to understand the properties of the study area and their influence on the lentil's yield.

1.8 Application of Result

The main application of this study is the agricultural sector. This study focuses on developing ways to identify and classify highly productive lentil genotypes. Research center staffs or researchers and breeding experts are beneficiaries of the proposed system to save their time to cluster genotypes, character analysis within clusters, and selecting of genotypes for the breeding programs. The Debre-Zeit Agricultural Research center can be benefited from the investigation provided by this research. In addition to its main beneficiaries, farmers are beneficiaries of this research; by getting good products. Finally, we are also benefitted from facilitating learning studies for researchers.

1.9 Organization of The Thesis

The organizations of this study are discussed as follows:

Chapter One: describes the introduction part of the study which introduces the overall idea of the study, motivation, statement of the problem, objectives of the study, Scope and limitation of the study, and application area.

Chapter Two: This chapter describes the literature review, supportive ideas. Analyze the existing system and gain further study to the research area to identify trials, contributions, limitations, gaps to be fulfilled, and related works regarding the research that establish an understanding of the study area.

Chapter Three: describes the research methodology including methods, and techniques used in this study to design the solution. Data collection, analysis, processing methods, metrics used for performance evaluation are also discussed in this chapter.

Chapter Four: This chapter describes the design of the proposed solution is described. It includes the model architecture, pseudo-codes for implementation that we follow to solve the problem.

Chapter Five: This chapter describes the implementation and experiment of the proposed solution are described. Including specifying the working environment, the implementation of data preprocessing, model building, and model evaluation.

Chapter Six: describes the result and discussion. The result of data collection, data preprocessing, model building, performance evaluation of each model, and the overall discussion of this study are described.

Chapter seven: Concludes the overall research work and recommendation that provides possible future work.

CHAPTER TWO

2. LITERATURE REVIEW

2.1 Overview

This chapter is about reading and reviewing most related articles, journals, or books for more understanding. Analyze and compare the existing study and gain further study to the research area to identify trials, contributions, limitations, gaps to be fulfilled, and some related works. In this study, the tasks in the literature review are to summarize information, establish an understanding of the study area, and find the strength and weaknesses of the existing study related to lentil genotypes identification and classification.

2.2 Production of Lentils in Ethiopia

In our country, a small number of farmers grow different crops for their nutrition and economic profits. According to the report of the federal democratic republic of Ethiopia central statistical agency on Area and Production of Major Crops in 2010 E.C; pulse production on the second stage among the crops produced in Ethiopia by following cereals in terms of total production and area coverage (FDRECSA 2018).

Lentils mainly grow on thick black soil. It is one of the most important legumes in the Ethiopian highlands and grows in rotation with barley, tef, and wheat. However, there are some conditions for the production of lentil plants, such as soil type, altitude, and agro-ecological conditions; In Ethiopia, different varieties of lentils are found. In our country, its production is not automated and its production is handled by smallholder farmers with fragmented schemes of land mainly for household use. Ethiopia is one of the centers of diversity for lentils in the world (Kassa 2018). The production of lentils in our country has been low, mainly due to the planting of low-yield local varieties that are vulnerable to diseases. Ethiopian lentils are known for Low productivity per unit area and low grain quality. There are few improved lentil varieties established by Ethiopian breeders (Kassa 2018). There are eleven varieties of Lentils released in Ethiopia those are ER-142, R-186, Chekol, Chalew, Adaa, Gudo, Alemaya, Assano, Alemtena, Teshale, Derso.

2.3 Supportive Concepts of The Study

Most of our people are farmers and living by producing crops. Lentil is the most produced and used legumes by many Ethiopians. There are different genotypes of lentils produced in Ethiopia. There are problems with the identification of lentils genotypes that has high productivity using yield prediction and genotypes character analysis. Many types of research were done to find improved lentils varieties. Genotypes identification is the most important in finding the high productivity of crops. For this purpose, this study proposes the identification and classification of highly productive lentil varieties that are affected by disease and yield prediction using machine learning algorithms.

The proposed system works depending on the criteria and variables that are used by agriculture experts to identify the lentils varieties and yield prediction. Machine learning models are applied for different clustering and classification problems. In this study also machine learning algorithm was applied to cluster lentil genotypes and predict the yield of lentils. Identification of lentil genotypes is very useful in the breeding of plants; since breeding itself is finding the good genotypes from existing genotypes either in convention, nonconventional or hybrid breeding (Niazian and Niedbała 2020). The highly correlated traits with yields have a greater influence on the yield of lentils and breeders usually prefer to use those traits to analyze the yield. The evaluation of genetic diversity, analysis of yield components, clustering and classification of lentils genotypes, prediction of yield are the supportive concepts in this study.

2.4 Lentil Genotypes Identification Variables

Lentil has different properties for finding productive lentil genotypes. The main characteristics of lentil genotypes to identify the good genotypes crop can depend on agronomic traits and disease data that are categorized into two crop phonological data yield and yield components recorded at the maturity of yield (Kassa 2018). Phonological data are days to emergence, days of flowering, days to maturity. Yield components are Plant height, Above ground biomass yield, grain yield, Hundred Seeds weight, Harvest index. Disease data can be used to identify disease-resistant varieties features used in disease data can be Rust, RRot, ABL, Wilt, PMW

Days of Flowering: the days counted from the date of seedling up to a period of 50% flowering of the plant in a plot developed the first flower (Quartey et al. 2014).

Days of Maturity: number of days after seedling emergency up to the 90% maturity of plants in the plot and ready for harvesting. Maturity can be identified by a change in the pod color, foliage, and seed status in the pod.

Plant height (cm): The ruler is used to measure the height of sample plants at 90 percent maturity, from the ground to the tip. (Kassa 2018).

Above-ground biomass yield: Plants from the center row were manually harvested close to the ground surface when they reached physiological maturity. The weight was measured by sun-drying the above-ground biomass from the sample plants from rows of destructive sampling to a constant weight, then converting the weight to the number of plants per net plot and then kg/h.

Grain yield: The grains were been put in a material sack and been weighed to give yield per plot. After separating, Seeds were sampled after separation and brought to the laboratory to be measured for moisture content using a seed moisture meter. To maintain uniformity, each plot's final grain yields were adjusted to a seed moisture content of 10%. Finally, plot yields were converted to per hectare yields, and the average yield was reported in kilograms per hectare. (Kassa 2018).

Hundred Seeds weight: Hundred seed weights were recorded by weighing 100 randomly taken dry seeds from the harvested net plot using a sensitive balance and the weight would be adjusted to 10% seed moisture content(Kassa 2018).

Harvest index: is determined as a percentage of total above-ground biomass yield per plot divided by grain yield per plot (Pennington 2013).

In addition to agronomic traits, Disease data variables are listed below.

Rust: Around the flowering time brown to dark brown, round to oval, powdery pustules appear on older leaves. They are more often noticed on the lower surface of leaves but can be seen on the upper surface. In severe cases, pustules can be seen on petioles and stems.

Root Rot (RR): several fungi such as Fusarium solani, Rhhizoctonia bataticola, Rhizoctonia solani can cause RR. It can appear at any growth stage of the crop. It causes yellowing of the basal foliage, stunted growth, and reddening of the vascular tissue below the soil line.

Fusarium wilts (FW): is caused by Fusarium oxysporum f. sp. Ciceri. The affected plant can be seen within 10 days after emergence until the senescence stage.

Ascochyta blight (ABL): The disease is caused by Ascochyta rabiei and infection can be observed on leaves, petioles, stems, pods, and seeds of chickpea, producing dark lesions or

cankers of various sizes. Symptoms of the disease usually appear around flowering and podding time. The plants are destroyed and appear as patches of blighted plants in the field. Ordinary circular spots show up on takes off and cases, stretched injuries onstem, and profound cankerous injuries on seeds.

Powdery mildew (PWM): in which Erysiphe trifolii causes an essential foliar fungal disease that affects all above-ground sections of the plant, including leaves, stems, and pods. The infected leaves drop down, leaving just terminal leaves on the stems, significantly reducing photosynthate assimilation, resulting in a loss in crop output and seed quality.

2.5 Overview of Machine Learning

Machine learning is an AI (artificial intelligence) application that allows computers to learn and advance on their own without having to be programmed explicitly. Machine learning allows computer programs to access data and utilize it to learn for themselves, and it is concerned with the creation of computer programs. One of the research directions to tackle unpredictable circumstances at runtime is machine learning for modification of adaptation space (Judith Hurwitz 2018)(Sah 2020).

The learning methods begin with observations or data, corresponding to examples, direct experience, or instruction so that we will get patterns in data and make better decisions in the future based on the examples we provide. The fundamental goal of ML is for computers to learn on their own, without the requirement for human association, and to alter their behavior in a like manner. A more dynamic and cheaper solution is to use techniques such as those used in machine learning which can automatically update models so that the predictions of the model correspond to the situations that are observed at runtime(Eernisae and Snelling 2016). Machine learning algorithms are often divided into supervised learning, unsupervised machine learning, semi-supervised learning, and reinforcement learning.

2.5.1 Supervised Learning

Supervised learning is a type of machine learning in which train the machine using well-trained data. The labeled training data helps you to predict outcomes for unforeseen data(Sathya and Abraham 2013). The system can give class for any unseen data after enough training. By comparing the output of the learning algorithm with the correct result supervised learning finds errors to modify the model accordingly. Supervised learning is categorized into a classification for category output and regression for real value (Sathya and Abraham 2013).

2.5.2 Unsupervised Learning

Unsupervised learning is the type of machine learning technique, in which no need to supervise the model. It deals with unlabeled data. Unsupervised learning is more unpredictable than other learning techniques. This technique allows complex processing tasks when compared to supervised learning (Sathya and Abraham 2013). Instead, it needs to allow the model to work independently to discover information from the given model. It also mainly deals with unlabeled data to do tasks based on user requirements. Clustering and association are two types of unsupervised learning (Rajoub 2020).

2.5.3 Semi-Supervised Learning

Semi-supervised learning is the type of machine learning that contains a small amount of labeled and a large number of unlabeled datasets to train the model. It falls between supervised and unsupervised machine learning. Combination of supervised and unsupervised learning(Cholaquidis, Fraiman, and Sued 2020; Dietterich et al. 2005; Poczos n.d.). The advantages of both supervised and unsupervised learning are combined in this form of machine learning. You can utilize semi-supervised approaches to expand the amount of your training data if you don't have enough labeled data for training to build an accurate model and don't have the skills or resources to obtain more data. The semi-supervised learning method is valuable in improving learning accuracy (Cholaquidis, Fraiman, and Sued 2020). A text document classifier is an application of semi-supervised learning.

2.5.4 Reinforcement Learning

RL is one type of machine learning algorithm that enables agents to learn from the environment by trial and error using feedback from their actions and experiences that discover rewards(Morales and Zaragoza 2011). Markov decision process is one example of reinforcement learning. Here studying records offers feedback so that the machine adjusts to dynamic situations to acquire an aimed objective. The machine evaluates its overall performance primarily based totally on the feedback responses and replies in step with the feedback responses. The problem in reinforcement learning is mostly explained through games. Some terms describe reinforcement learning problems these are: Environment, State, Reward, Policy, Value. Its diagram is represented below.

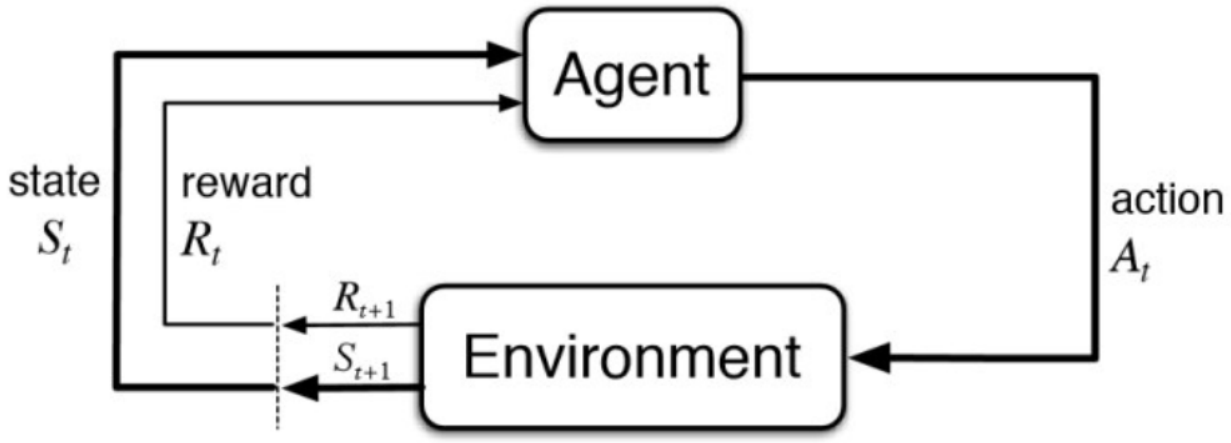


Figure 2. 1: Reinforcement learning process(Shweta Bhatt 2018)

2.6 Application of Machine Learning in Agriculture

By assisting farm management decisions with data-driven insights, digital agriculture plays a significant role in increasing crop production precision(Lutz n.d.). Crop cultivation is critical to the global food and biofuel economies. Machine learning is revolutionizing the way farmers contribute to the food and biofuel industries. Every year, farmers make a slew of complicated decisions that impact their risk, long-term viability, and financial rewards(Computing 2019). Applications of Machine learning in Agriculture are Species management, field condition management, crop management, livestock management(Sciforce 2019)(Benos et al. 2021).

2.6.1 Species Management

Under species management, there are subtasks like species identification, species breeding, species recognition. Species identification is most of the time there are difficulties in species identification using human eyes since many plants have similar leaf compositions, traits, colors, and shapes to label. Complex patterns may be assessed using machine learning, and related plant species can be properly identified. Plant species identification is beneficial since it saves farmers time and allows them to focus on other important tasks(Computing 2019).

Species breeding is a way of finding improved species that determine the efficiency of water and nutrient consumption, disease resistance, climate change adaptation, and nutritional content or flavor. In Machine learning, deep learning algorithms, for example, take decades of field data to examine crop performance in diverse climates and build new features in the process. They can create a probability model based on this data to forecast which genes are most likely to provide a beneficial characteristic to a plant. While the typical human approach to plant

classification involves comparing leaf color, traits, and shape, machine learning can deliver more accurate and faster answers by examining leaf vein architecture and traits analysis; which contains more information about the leaf attributes(Sciforce 2019).

2.6.2 Field Condition Management

In condition management, machine learning works tasks like soil management, water management. Soil management is the processes, practices done to protect and enhance the soil. To understand the dynamics of ecosystems and the impingement in agriculture, machine learning algorithms investigate evaporation processes, soil moisture, and temperature**.** Agriculture's water management has an impact on the hydrological, climatological, and agronomic balance. Machine learning is applicable in water management by estimating evapotranspiration allowing for more effective use of irrigation systems, which helps categorize expected climate phenomena and guess evapotranspiration and evaporation(Sciforce 2019).

2.6.3 Livestock Management

In livestock management, machine learning works tasks like livestock production, animal welfare. Livestock production is one application of Machine learning in which machine learning models accurately predict and estimate the farming parameters to optimize the livestock production system and economic efficiency. For example, weight prediction of cattle (Sciforce 2019). Machine Learning benefits farmers to sustain happy and healthy cattle(Computing 2019).

2.6.4 Crop Management

In crop management, machine learning is used in tasks like yield prediction, crop quality, disease prediction, weed detection. Machine learning is applicable in yield prediction by predicting the yield of the crop from available data like weather parameters, soil parameters. and historic crop yield (Priya, Muthaiah, and Balamurugan 2018). Machine learning in crop quality is using machine learning algorithm identification, detection, prediction, and classification of crop quality to increase product price and reduce waste (Sciforce 2019). In Disease detection machine learning plays a great role to save the high financial cost and significant cost for spray pesticides and other drugs for one or more diseases by detecting the right disease of the crops by accurately identifying or detecting and recommend optimal treatment (Computing 2019). The image-driven model is the best and most accurate for

detecting crop disease for image data (Mohanty, Hughes, and Salathé 2016). Weed detection using mechanical devices and machine learning to pull weeds from the arenas, protect pesticide impacts and save time and effort for the farmers(Computing 2019)(Benos et al. 2021). Machine learning models are important in the prediction of the yield of crops as important decision support for crop yield prediction(van Klompenburg, Kassahun, and Catal 2020).

2.6.5 Machine Learning for Clustering

Clustering algorithms are unsupervised machine learning techniques that are useful in crop management to cluster the genotypes into different homogenous clusters depending on the agronomic traits, disease data, and yield data, each genotype gives(Greenlaw and Kantabutra 2010). It is called because unsupervised since it uses unlabeled data. In clustering the items in the same cluster are similar and items in different clusters are dissimilar(Omran, Engelbrecht, and Salman 2007). Some common clustering algorithms are centroid-based clustering, distribution-based clustering, hierarchical clustering, density-based clustering.

Centroid-based clustering

Centroid-based algorithms are partitional methods to create a set of disjoint clusters fixed number of k data points are chosen to function as a centroid. After the centroid selection, the distance is calculated between each element and the centroid, then the items will be assigned to the closest cluster centroid based on the calculation. After the reassignment of items in all clusters, a new centroid is calculated and again the distance is calculated between each item and the new centroid in each cluster, and again the items are reassigned to the closest cluster centroid and this will continue until all items are in their closest cluster. This clustering method creates clusters by assigning items to closest clusters by fulfilling the following two requirements; the first requirement is each cluster contains at least one centroid point and the second requirement is each centroid point belongs to exactly one cluster. The centroid-based clustering algorithms are Kmeans, K-mode, K-medoids(Uppada 2014).

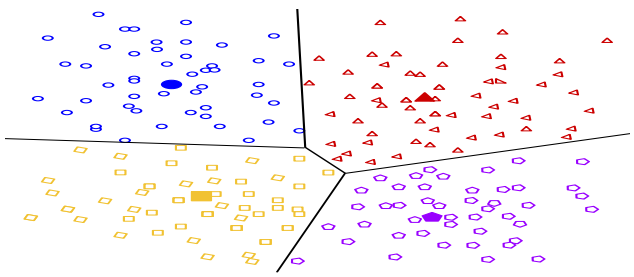


Figure 2. 2: Centroid-based clustering

Hierarchical clustering

Hierarchical clustering is a clustering algorithm that produces multiple hierarchies of clusters. This algorithm starts its clustering process with all the data points that are assigned to a group of their own, and the two nearest clusters are merged into the same cluster. In the end, the hierarchical algorithm terminates the process when only one cluster remains. After hierarchical clustering is done on the dataset, the result will be a tree-based representation of each data point, divided into clusters(Contreras and Murtagh 2015).

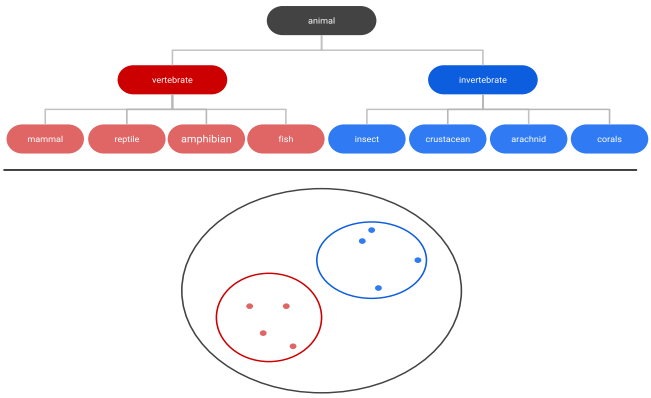


Figure 2. 3: Hierarchical clustering

Density-based clustering

This algorithm can better deal with large datasets with noise and also can select and cluster points with different sizes and shape In this clustering algorithm, clusters are selected based on the density of the points in the dataset, and clusters are considered high-density regions(Piiroinen 2018). The main demerits of this algorithm are in dealing with high-dimensional datasets and having high time complexity. The main algorithms that fall in this category are DBSCA, OPTICS, and SNN(Bhuyan and Borah 2013).

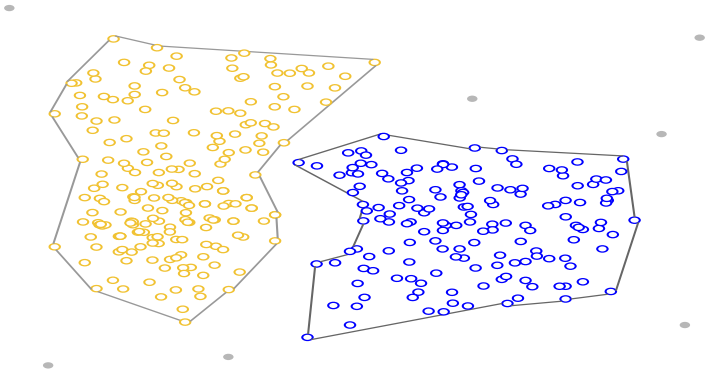


Figure 2. 4: Density-based clustering

Distribution-based clustering

The distribution-based clustering approach assumes data is composed of distributions. the distribution-based algorithm clusters data into different groups of distributions. As the distance from the distribution's center increases, the probability that a point belongs to the distribution decreases. The bands show a decrease in probability. When you do not know the type of distribution in your data, you should use a different algorithm. such as Gaussian distributions(X. Xu et al. 1998)(D. Xu and Tian 2015).

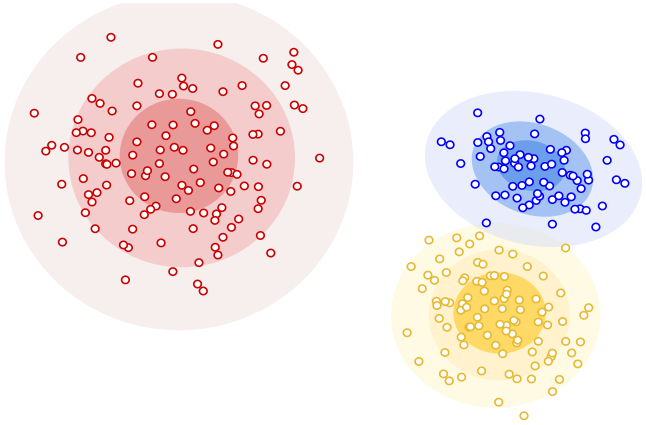


Figure 2. 5: Distribution-based clustering

2.6.6 Machine Learning for Lentils Yield Prediction

The regression learning algorithm is supervised machine learning that is important in the prediction of the labeled data. It works on continuous values prediction. It also important in the yield prediction of lentils.

Random Forest Regressor

The basic idea in a random forest is that the predictions will be closer to the true value on average; since the individual predictions made by the decision tree may not accurate. When combining predictions made by many decision trees into a single model it is called Forest (Cutler, Cutler, and Stevens 2012).

Gradient Boosting Regressor

is ensemble decision tree regressor model (Center 2019). It is boosting ensemble trees. It is important in heterogeneous data and can automatically detect non-linear feature interaction, but it is slow to train and requires careful tuning (Prettenhofer and Louppe n.d.).

Multiple Linear Regression

This is different from simple Linear regression by having multiple independent variables. It needs Multiple coefficients to determine and complex computation due to the added variables. Where Y is the dependent variable, β is the regression coefficient, and *x* is the independent variable. The equation of multiple linear regression is (Bremer 2012):

$$Y_i = \beta_0 + \beta_1\chi_i + \beta_2\chi_i \ \ldots\ldots \ \beta_n\chi_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{eq 2.1}$$

Equation 2. 1: The equation of multiple linear regression

2.6.7 Machine Learning for New Genotypes Classification

Classification refers to a classification label is predicted for a particular example of input data in an identification modeling issue. The results of classification predictive modeling algorithms will be evaluated based on their results. Classification accuracy is the most commonly used metric to evaluate the performance of the model based on the expected class labels(Granitzer et al. 2009). Although classification accuracy isn't ideal, it's a solid place to start for a lot of classification problems. As there is not exactly identified the best model for the classification purpose since different researchers have done using a different model and suggest the model based on their findings, in this study different machine learning algorithms, were used. Then, we compare and choose the model with the best accuracy. The following classification algorithms are used in the classification tasks in crop management.

Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm that is most commonly used to solve classification problems. Each data item is represented as a point in n-dimensional space in the SVM algorithm, with the value of each feature being the value of a certain coordinate. Then, the classification will be carried out by locating the hyper-plane that differentiates the classes very well (Awad and Khanna 2015).

(J. Brownlee 2016) The following is the equation for predicting a new input using the dot product of the input (x) and each support vector (xi):

$$f(x) = B0 + \sum_{i=1}^{n}(ai * (x * xi)) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{Eq 2.2}$$

Equation 2. 2:Equation of Support Vector Machine calculator

Random Forest (RF)

RF is one type of ensemble approach that generates forest. Overcome overfitting, deal with high dimensional data. It is the most classification algorithm that uses to create the forest with several decision trees that operate by constructing a multitude of the tree at training time. The algorithm that uses an ensemble algorithm will improve the accuracy and it is most powerful (J. (Machine L. M. Brownlee 2019). Random Forest uses both bagging and random feature selection to build the tree and creates an uncorrelated forest of trees whose prediction by the group is more accurate than that of any individual tree.

Decision Tree (DT)

Making predictions with DT is a rather simple process. The tree is traversed given a new input by evaluating the specific input starting at the root node of the tree. A learned binary tree is a partitioning of the input space. The dimension of the tree depends on the dimension of input variables. When p = 2 input variables, the decision tree divided this into rectangles, and when more input variables, DT split it into hyper-rectangles(J. Brownlee 2016). Because of their hierarchical structure, they are robust to outliers, scalable, and capable of modeling non-linear decision limits.

Naïve Bayes

NB is the machine learning algorithm that is used for binary class and multiclass domain problems. is a very simple classification method that makes some significant assumptions about each input variable's independence. It is, nonetheless, effective in a wide range of problem domains. Training from training data is fast; because no coefficients need to be fitted, only the probability of each class needs to be calculated (J. Brownlee 2016).

To calculate conditional probability:

$$P(\text{class} = 1) = \frac{\text{count}(\text{class} = 1)}{\text{count}(\text{class} = 0) + \text{count}(\text{class} = 1)} \quad \text{.....Eq 2.3}$$

Equation 2. 3: Equation of Naive Bayes calculator

2.6.8 Machine Learning for Genotypes Recommendation

Recommendation systems are used to recommend products, services, such as books, kinds of music, movies, and other products to assist users in finding good products or new items. This system plays a crucial role in decision-making for users to increase profits, reduce risks, and

minimize the waste of time to find an item. The Recommender System is applied in many computer science organizations mostly in social media like YouTube, Twitter, and LinkedIn (Portugal, Alencar, and Cowan 2018)(Sahu, Nautiyal, and Prasad 2017). Above mentioned machine learning algorithms are useful in this system to process items by classifying and clustering. The Recommender system contains different algorithms, such as collaborative-based filtering, content-based filtering, and hybrid filtering. In this study, a collaborative-based filtering algorithm was used.

Collaborative-based Filtering

Collaborative-based filtering is a recommendation technique that recommends an item to a given user that this given user didn't experience the item before based on the interests of users that have similar interests or liked the item the same in the past. This technique is good for recommending content like movies and music that are difficult to describe by metadata(Isinkaye, Folajimi, and Ojokoh 2015).

There are two types of collaborative-based filtering: Model-based filtering and Memory-based filtering. Collaborative-based filtering is the technique used in this system to recommend genotypes of lentils for breeding to enhance the yield of lentils. From collaborative filtering, the models of machine learning were used to improve recommendation accuracy. To improve the accuracy of this technique K-means clustering algorithm was used for clustering the genotypes that have similar characteristics into the same groups.

Content-based Filtering

Content-based filtering recommends based the user preferences using features extracted from the profile of the item that the user previously preferred. The main assumption in content-based filtering is that observing the past actions and histories of the user will allow us to predict the desired items in the future. This filtering technique unlike collaborative filtering techniques it doesn’t need the predictions of other users, it predicts its recommendation by using only a given user profile or preferences(Portugal, Alencar, and Cowan 2018).

2.7 Feature selection

It is known that most of the real-world datasets have many attributes especially in agriculture, However, not all the attributes have the same importance level in identification. Although there are many attributes to identify productive lentil varieties; not all attributes are equally

significant in identifying high products. So, there is a challenge to selecting a subset of features within the minimum risk or significantly essential to determine the dependent feature values. Features will be characterized with different behaviors like: Relevant, Irrelevant, and redundant. To overcome the challenges that come with feature characterization feature selection is important and selecting the relevant attribute only is very useful to increase performance and reduce test time. Feature selection helps in reducing dimensionality and making the model simpler and easy to use, thus leading to short training time and high accuracy(J. Brownlee 2016)(Deng 1999).

Feature selection is important for the following (Cunningham, Kathirgamanathan, and Delany 2021).

- It enables machine learning to train faster.
- It simplifies the model and makes it easy to understand.
- It decreases overfitting
- Feature selection is vital in improving the model's accuracy if the proper subset is selected.

In supervised machine learning models feature selection methods are classified into three Filter, wrapper, and embedded feature selection.

Filter method: This feature selection approach is referred to as preprocessing. The filter approach relies on the results of numerous statistical tests to determine the correlation between independent and dependent variables, as well as feature selection, without relying on machine learning. To select methods for filtering features we have to know our data. Depend on the dependent variable and independent variables we can categorize filtering methods into (Cunningham, Kathirgamanathan, and Delany 2021):

Numerical input, numerical output: use correlation coefficient such as Pearson's correlation coefficient for a linear correlation, Spearman's rank coefficient for non-linear correlation.

Numerical input, categorical output: most commonly used in classification problems such as ANOVA correlation coefficient for a linear, Kendall's rank coefficient for nonlinear.

Categorical input, categorical output: is common in classification predictive problems. Examples of Categorical input, the categorical output is the Chi-Squared test, mutual information.

Categorical input, numerical output: is most common in regression predictive modeling problems. ANOVA correlation coefficient and Kendall's rank coefficient can be used but in reverse.

Wrapper method: We try to use selected parameters and train the model using those selected parameters in wrappers techniques of feature selection. We then determine whether to add or remove parameters from these variables based on the trained models. This strategy is necessary for reducing search issues, although it is usually costly (Deng 1999).

Embedded method: it is the combination filter and wrappers method of feature selections. It will be implemented by algorithms that have their built-in feature selection method.

2.8 Related Works of The Study

The research centers are the center of research on agriculture in finding the best and improved genotypes of plants and animals. It is challenging to identify, predict, and operational optimization in the identification of productive lentils. The lentils breeding programs need identification of genotypes and analysis of characteristics to find the genetic variability. Classical statistics are the most common method in the analysis of characteristics in the identification of crops genotypes for breeding. The existing study for the identification and analysis of characters of lentil varieties was done using classical statistical analysis models. The lentils type identification researches were done to increase the productivity of lentils(Niazian and Niedbała 2020). Lentils genotype identification is useful to find the genotypes that have high products, good quality, and disease-resistant behavior. The groups of lentils were grouped by finding the variance using PCA and ANOVA between lentil genotypes agronomical traits and to cluster genotypes of lentils (Nath et al. 2014)(Mohammed et al. 2019). Some related works are listed down.

(Kindie, Nigusie, and Yildiz 2018) authors researched the Participatory evaluation of lentil varieties in Wag-lasta, Eastern Amhara to identify improved lentil varieties. Accordingly, seven newly released lentil varieties were assessed, including Alemaya, Checol, Danbi, Teshale, Alemtena, Derash, Gudo, and Local check, which were obtained from the Debrezeite Agricultural Research Center. From the first week to the middle of July in the 2016/17 cropping season, the trial was planted in a randomized complete block design with three replications. Days to flowering, days to maturity, plant height, pod per plant, seeds per pod, 100 seed weight, biomass, and grain yield were all reported. At harvest, biomass and grain yield were collected

from the four middle rows and recorded in gram per plot, but for analysis, it was converted to kg ha-1. Whereas the 100 seed weight was calculated by selecting 100 seeds at random. Finally, Derash and Danbi chose and recommended varieties for production in the designated regions and similar agro-ecologies based on the farmer's appraisal and the researcher's selection.

(Mohammed et al. 2019) In this study, 36 lentil genotypes and 24 agro-morphological traits of lentils were evaluated. The goal of this study is to analyze genetic diversity across lentil genotypes at the agro-morphological level to find the most capable genotypes for cultivation in dry environments. All genotypes tested and ANOVA showed highly important inter and intra block differences for all genotypes. PCA was used to analyze the cluster supported by hierarchical based on qualitative and quantitative traits. SAS 9.3 (statical Software) was used to analyze the standard deviation and mean values of morphological traits of lentils. The thirty-six genotypes were clustered into four main clusters.

(U. Singh, Waza, and Srivastava 2012) In this study, they work on eighty genotypes of lentils to evaluate. All genotypes were grouped into fourteen groups depending on their traits. The intra-cluster distance and inter-cluster distance between each genotype of lentils were used to analyze divergence. they conclude that the high inter-cluster distance between clusters is produced desirable isolation. The morphological traits of lentils are recorded during the experiment and used to cluster. The general objective of this study is to classify the available genotypes of lentils into trenchant groups depending on their genetic difference. They use multivariate analysis to identify the most desirable genotypes to use in hybridization programs for the evolution of high-yielding varieties with a qualitative and quantitative basis of improved plant traits.

(M. Singh et al. 2020) In this study 96 lentil accessions are comprised. The main aim of this research was to find economically feasible agro-morphological and resistance to biotic stresses by evaluating accessions of wild Lentils. PCA (principal component analysis) is used to analyze the traits of lentils. Depend on the agronomic traits the genetic variation of lentils produced genotypes were clustered into two major clusters. PCA exhibited that, number of pods per plant, number of seeds per plant, biological yield per plant, seed yield per plant, and harvest index contributed significantly to the total genetic difference assessed in wild lentil taxa. The 96 accessions of lentils were evaluated depending on the agronomic traits and disease data. ANOVA (analysis of variance) exposed that the wild lentil accessions differed significantly in terms of seed yield per plant.

(Nath et al. 2014). In this study, the experiment was done using 20 lentils genotypes were tested during 2010 and 2011 to determine genetic variability, diversity, characters association, and selection indices for better grain yield. The genotypes were clustered into four groups depending on the Euclidean distance. Statistical analysis was used for the analysis of morphological traits of lentils. To study genetic diversity between lentils genotypes Euclidean Ward's method was applied. PLABSTAT Version 2N software was used to calculate the analysis of variance. The major goal of this research was to identify genotypes that consistently perform well in ecological farming systems with no intercultural interactions.

(Bagheri et al. 2020). This study was conducted to predict the yield and biomass of lentils affected by weed interference using deep learning comparing with a multiple regression model. They measure the traits of lentils in two sampling stages. In the first stage sampling, the weed density and height, as well as canopy cover of the weeds and lentils, were measured. In the second stage, the yield and biomass of lentils were measured. the objectives of the authors are to increase awareness about the effects of weed on crop yield and to develop a system that predicts the yield and biomass of lentils using machine learning models. The result of this study shows that the Artificial neural network model is accurate in the prediction of lentils yield under weed interference than the multiple regression model by the result of R^2 =0.79.

Table 2. 1: Summary of related works

References	Contents	Descriptions
(Mohammed et al. 2019)	Title	Agro-Morphological Characterization of Lentil Genotypes in Dry Environments
	Objectives	The objective of this study is to assess genetic variability between lentil genotypes and identify the most capable genotypes for cultivation in dry environments.
	Methods	ANOVA was used for the analysis of variance and PCA for analysis of the clusters.
	Limitations	This study was limited to only analysis of characteristics. Methods used may fail for non-linear features of data.
(M. Singh et al. 2020)	Title	Evaluation and identification of wild lentil accessions for enhancing genetic gains of cultivated varieties
	Objectives	The main objective of this study was the identification of economically feasible agro-morphological and resistance to biotic stresses by evaluating accessions of wild Lentils.
	Methodology	PCA for analysis of lentils traits and ANOVA
	Limitations	This study was limited to find the accessions that are viable. It is not easy to select genotypes for breeding.
(Kindie, Nigusie, and Yildiz 2018)	Title	Participatory evaluation of lentil varieties in Wag-lasta, Eastern Amhara
	Objectives	The main objective of researchers is to identify improved lentils varieties that are recommended for Wag-lasta, Eastern Amhara.
	Methodology	Compare productivity or seed yield to select
	Limitations	Seed yield is not enough to find the genetic variability and as selection criteria for yield improvement
(U. Singh, Waza, and Srivastava 2012)	Title	Study the Evaluation of Genetic Diversity in Lentil (Lens culinaris Medik.)
	Objectives	The general objective of this study is to classify the available genotypes of lentils into trenchant groups depending on their genetic difference.

	Methodology	Intra-cluster distance and inter-cluster distance to find divergence.
	Limitation	Large intra-cluster distance values within a cluster, that needs additional cluster number
(Nath et al. 2014)	Title	Selection of Superior Lentil (Lens esculenta M.) Genotypes by Assessing Character Association and Genetic Diversity
	Objectives	The main objective of this study was to find the genotypes that perform constantly well even in ecological farming systems without any intercultural actions.
	Methodology	Ward's method was applied to find genetic divergence. ANOVA calculated by using PLABSTAT Version 2N software
	Limitations	Nothing stated about the recommendation of genotypes, methods used fail for large data size
(Bagheri et al. 2020)	Title	Artificial neural network potential in yield prediction of lentil (Lens culinaris L.) influenced by weed interference
	Objectives	objectives of the authors were yield prediction and environmental impact categories of lentil production using ANN models in Iran, and to investigate the sensitivity analysis of energy inputs on lentil yield and the environmental impact categories.
	Methodology	An artificial neural network and multiple regression
	Limitations	They got a low coefficient of determination

CHAPTER THREE

3. RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, techniques, methods, and approaches that take on to reach the objectives of the identification and classification of productive lentils varieties affected by disease using machine learning are described. This includes how to collect, process, tool selection (software and hardware) in data (data collection, analysis of data, and data preprocessing), model selection. Additionally, the overall research procedure, practices, systematic approaches are defined. Generally, to fulfill the research objectives and to answer the research questions the methodology used is described in this chapter. The research design used in this study is quantitative. Under quantitative research design, our work focuses on the experimental type. The research procedure or conceptual work of this study is as follows.

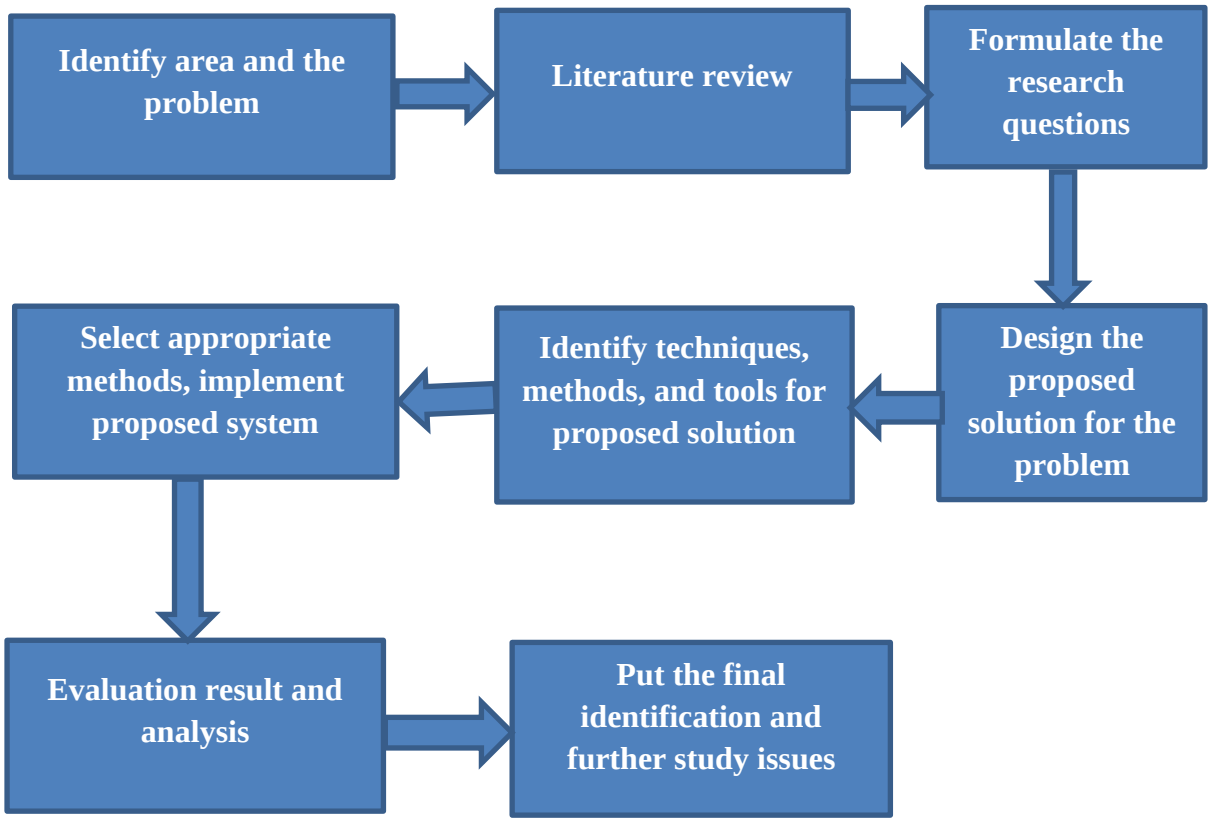


Figure 3. 1: The research procedure of this study

In this study the methodologies used to achieve the stated procedure are:

Review Literature: studying different resources like articles, journals, papers, books, magazines, websites, workshops, and so on. For more understanding, finding existing problems, gaps to be fulfilled related to identification and classification of productive lentil varieties and machine learning techniques for classification.

Data collection: Collecting useful data from available sources**.** The data collected from the Debre-Zeit Agriculture research center contains 43 genotypes collected from six different locations, from experiments done between 2013-2017.

Data preprocessing: to prepare a useful dataset, preprocessing process is useful to organize, clean, transform, and reduce the features.

Developing model: To identify and classify the highly productive lentil genotypes developing machine learning for clustering, analyzing characteristics, and predicting lentils yield. The model is developed to cluster and predict. To develop and implement the model environment set using python and anaconda libraries.

Evaluation: After developing the model that identifies and classifies lentil genotypes. we evaluate the model by using model evaluation metrics. The evaluation metrics used in this study are to find the accuracy and loss at validation. The assessment of model described by confusion matrix to represent TP, TN, FP, FN. Other evaluation metrics precision, recall, and F1-score were also used to evaluate the performance of the model. The error of prediction can be evaluated by MAE, MSE, and R-square. The clustering algorithms are evaluated depending on silhouette score and inter-cluster distance.

3.2 Data Collection

As data is the soul of machine learning and ML is a data-driven model data collection and model understating are important. Data collection is the process of collecting relevant data from different applicable sources. In this study, we gather lentil data that was available at the agricultural research center to train and test the model. These data were collected from the experiment researchers did in that organization. Having this train data set for the model, the study aims to identify and classify the highly productive lentil genotypes. In Ethiopia, it is difficult to find datasets because of limited access to data in the organization most of the time. Therefore, the dataset is prepared from the recorded data from the experiments done. To achieve a high-quality predictive model the data must be prepared in a suitable format to feed in the machine learning algorithms. The data used in this research is secondary data collected

by other people/ researchers gotten from the Debre-Zeit Agriculture research center. The data collected from agricultural organizations contains agronomic traits and disease data. The sample of data used in this study is shown below.

Year	Locn	GENOTYPE	DF	DM	Rust	RROT	WILT	ABL	PWM	Vigor	PLH	HSW	BMYP	YLD_GM
2016	AM	DZ-2012-Ln-0221	48	148	5	5	5	1	3	3	21.4	3.2	300	24
2016	AM	DZ-2012-Ln-0230	44	134	3	3	3	1	3	2	25.8	2	500	102
2016	AM	DZ-2012-Ln-0227	56	148	3	3	3	1	3	3	27.6	3.2	400	66
2016	AM	LOCAL CHECK	44	119	3	3	1	1	5	2	25.4	2	500	71
2016	AM	DZ-2012-Ln-0199	44	134	3	3	1	1	5	2	28.6	3.2	500	90
2016	AM	DZ-2012-Ln-0191	56	148	3	3	1	1	5	2	29	2.8	500	50
2016	AM	DERASH	44	125	1	1	3	1	5	1	31.4	2	700	211
2016	AM	DZ-2012-Ln-0226	58	148	3	3	3	1	5	2	27	2.4	500	58
2016	AM	DZ-2012-Ln-0223	44	134	3	3	3	1	5	3	25.6	4	400	57
2016	AM	DZ-2012-Ln-0225	58	151	3	3	3	1	5	3	26	2.8	400	70
2016	AM	DZ-2012-Ln-0222	51	148	1	1	1	1	5	2	27.6	3.6	500	61
2016	AM	DZ-2012-Ln-0056	46	148	1	1	3	1	5	3	29	2.8	400	36
2016	AM	DZ-2012-Ln-0045	42	127	1	1	1	1	7	2	27.4	2.8	500	107
2016	AM	DENBI	44	127	1	1	1	1	3	1	32.8	2	600	177
2016	AM	DZ-2012-Ln-0057	46	134	3	3	1	1	5	2	32.6	3.6	600	120
2016	AM	DZ-2012-Ln-0039	42	127	1	1	1	1	7	2	25	3	500	79
2016	AM	DERASH	44	127	1	1	1	1	3	1	34.8	2	700	216
2016	AM	LOCAL CHECK	42	120	3	3	1	1	7	2	29.8	2	600	96
2016	AM	DZ-2012-Ln-0226	44	151	3	3	1	1	5	3	31.6	2	500	63
2016	AM	DZ-2012-Ln-0191	54	148	1	1	1	1	7	2	32.8	2.8	500	18
2016	AM	DZ-2012-Ln-0045	42	134	1	1	1	1	7	2	30.8	2.4	500	112
2016	AM	DZ-2012-Ln-0222	46	148	1	1	1	1	5	2	28.4	3.2	400	61
2016	AM	DZ-2012-Ln-0223	46	148	1	1	3	1	3	3	29	3.2	400	31
2016	AM	DZ-2012-Ln-0221	51	148	5	5	3	1	5	3	26.2	3.2	400	35
2016	AM	DZ-2012-Ln-0230	42	127	3	3	1	1	3	2	34.6	2	200	143
2016	AM	DZ-2012-Ln-0057	46	134	3	3	3	1	5	2	35.2	3.6	600	149
2016	AM	DZ-2012-Ln-0056	46	148	1	1	1	1	5	2	34.4	2.4	500	65
2016	AM	DZ-2012-Ln-0039	42	127	1	1	1	1	7	1	34.4	3.6	600	192

Figure 3. 2: Sample of the dataset

3.2.1 Dataset Description

Machine learning needs data to feed to make a decision. Data used in machine learning has different formats either discrete or continuous. The data that is used in this research work is yield data for both supervised and unsupervised machine learning models. For predicting the yield using regression learning data that have independent and dependent parameters are used. The parameters used for prediction and their data type are stated below.

Table 3. 1: Dataset description

No.	Parameters	Datatypes	Number missing values
1	Year	Int	0
2	Location	Object	0
3	Genotype	Object	0
4	Days of Flowering (DF)	Int	3
5	Days of Maturity (DM)	Int	3
6	Rust	Int	0
7	Root Rot (RRot)	Int	0
8	Wilt	Int	0
9	ABL	Int	0
10	PWM	Int	0
11	Vigor	Int	0
12	Plant height	Float	2
13	100 seed weight	Float	0
14	Biomass yield per plot	Float	4
15	Yield per plot	Float	0

3.2 Data Preprocessing

After data was collected; The data preprocessing continued which is organizing, cleaning, normalizing, union and formatting into one file or data table. This process used the MS Excel preparation tool. Additionally, as the data available was manually recorded data, it contains inaccurate data, missing values, and unnecessary features that need to be processed before use it. For removing this unnecessary data from the dataset, this study has to be clean the raw data. After preparing on MS-Excel filling missed data and normalizing data are the most important tasks to increase the performance of the model. There are different techniques for normalizing data. Normalizing and standardize features by feature scaling techniques. In this study mean value was used for filling missing continuous values and Min-Max scaling methods for scaling our data. After being prepared and preprocessed dataset was used to train and test the model. Dataset collected from Debre-Zeit agricultural research center contains 22 features, 1587 instances. 8 features are dropped by the help domain expert. There are 8 incomplete rows of data including independent and dependent values. we remove the incomplete rows since those data don’t give full information.

Handling Missing Values: in this study, the data collected from Debre-Zeit agricultural research center have some missing values. Those data were missing due to mistakes while collecting data from the experimental areas in that center. Since the missing values influence the performance of our model we need to preprocess it(Kang 2013). To handle missing values we can use two techniques either dropping the missing values or filling missing values. In our study since we have limited access to data and the missing values are continuous values filling techniques are applied by using the mean imputation method(Gelman and Hill 2010). The mean imputation technique is better to fill the average value of the observed feature for continuous missing values. There are 12 missing values (3 in days of flowering, 2 in plant height, 4 in above-ground biomass, 3 in days of maturity) in our data and filled by using mean imputation techniques.

Handling Categorical Data: handling the categorical data is the technique in which the nominal data is converted into numerics to fed them into algorithms. The categorical data that contains string values need to be encoded. The encoding techniques in machine learning are useful in converting strings into integers(Hancock and Khoshgoftaar 2020). In this study location and genotypes name has string values. For encoding location name and genotypes name we used the label encoding technique.

Feature Scaling: To convert the required data to the required format which can fit data transformation is important in machine learning techniques. Feature scaling is the technique in which we can normalize or scale our data to important standards. Different machine learning algorithms need data scaled before fit into models. In some important classes of models feature scaling is mandatory. There are various techniques of feature scaling such as Min-Max scaler, decimal scaling, and Z-score(Patro and sahu 2015). In this study, since the K-means clustering algorithm is used and it needs feature scaling we scale our models using Min-Max scaler methods. Min-Max scaler keeps an existing relationship among data by working on the linear transformation of data. Min-Max scaler normalizes our data in between 0 and 1. Min-Max scaler is done as follows.

$$X_{scaled}=\frac{Xi-\ X\min}{Xmax-\ X\min)} \quad \text{.....Eq 3.1}$$

Equation 3. 1: Min-Max scaler formula

3.3Feature Selection

Feature selection can be either filter, wrappers, or embedded which are used for feature selection in different datasets. Filter and wrapper methods were used in this study. A filter method is a feature selection technique that is independent of algorithms and uses general characteristics of data for evaluating and selecting relevant features depending on feature importance results. The filter method works independently of the learning algorithm to remove irrelevant features and analyses the properties of a dataset to chooses the more important features. The filter method is popular; it is the most widely used approach due to its less complex nature. The wrapper method selects relevant feature selection while using the classification algorithm. The wrapper method is the best feature selection in terms of accuracy than filter feature selection. However, it requires high processing time. Univariate feature selection(UFS) method is used from filter methods because it is fast, efficient, scalable, and recursive feature elimination (RFE) is used from wrapper feature selection. RFE eliminates the redundant and weak feature whose deletion least affects the training error and keeps the independent and strong feature to improve the generalization performance of the model

3.4 Model Selection

To fulfill the proposed solution, we have to select the right methods and techniques. In machine learning, Regression learning is a type of supervised learning that involves predicting a labeled numerical value. It is a predictive modeling technique that evaluates the relationship between the independent variable and target variable in a dataset (Sah 2020). In this study, different regression algorithms are used for predicting the yield of the plant per plot. Those are random forest regressor, gradient boosting regressor, multiple linear regression.

3.4.1 Clustering Algorithms Selection

The clustering algorithms used in this study are the standard K-means and K-means++ (K-means initialized using k-means++) clustering algorithms.

K-means clustering: used in this study to cluster genotypes depends on their characteristics. It is centroid-based clustering. it is a simple, efficient, and effective clustering algorithm (D. Xu and Tian 2015)(Morissette and Chartier 2013). K value in K-means is pre-defined, distinct non-overlapping clusters where each data point belongs to only one group**.** Euclidean distance calculation is important in the k-means algorithm to find the correct centroids and reassign items

in their closest clusters to create clusters that contain homogenous items. Centroid initialization is the optimization process in K-means clustering in this study two methods of centroid initialization used. The two initialization methods are random data point and K-means++. In random data point initialization k random data points are selected from the dataset and used as the initial centroids. K-means++ initialization is smart centroid initialization. It is aimed to avoid converge at a local optimum caused by random initialization (Rukmi and Iqbal 2017)

3.5 Proposed System Evaluation

In this study, machine learning models are used and the model should give accurate predictions to create real value for a given organization. During training, a model developer splits data into training and test data to perform on it. The evaluation is important to analyze the work result very well. Using the test set in the sample data the models were evaluated. In this study, the researcher train, test, and evaluate the performance of the proposed system to evaluate how the model generalizes on unseen data. there are different techniques to evaluate machine learning models(Mutuvi 2019). The model evaluation aims to estimate the generalization accuracy of the model on unseen data. There are two categories for evaluating a model's performance these are:

Holdout: the holdout validation technique evaluates by dividing the data into training and test set. The trained models are based on the training set, and then the result of the prediction for the test set is compared with the actual result. It is a recognized technique to estimate the performance of models. It is very flexible, simple, and has speed. This technique is often associated with high variability(Raschka 2018).

Cross-Validation: is a mostly used data resampling method that is used to prevent overfitting and increase the accuracy of the model(Berrar 2018). There are different types of cross-validation techniques(Krstajic et al. 2014).

- **K-Fold cross-validation:** this evaluation techniques divide the dataset into k non-overlapped parts. The training set is k-1 data and the rest is used for testing. There are opportunities for each K-folds to be held back as a test set (Berrar 2018).

How does K-fold cross-validation works? The steps are shown below (Krstajic et al. 2014).

1. Split the dataset into K equal groups (folds).

2. Use the union of all folds except one testing fold as the training set.
3. Using the testing set compute the accuracy.
4. using a different fold as the testing set repeat steps 2 and 3 K times
5. To estimate out-of-sample accuracy use the average testing accuracy.

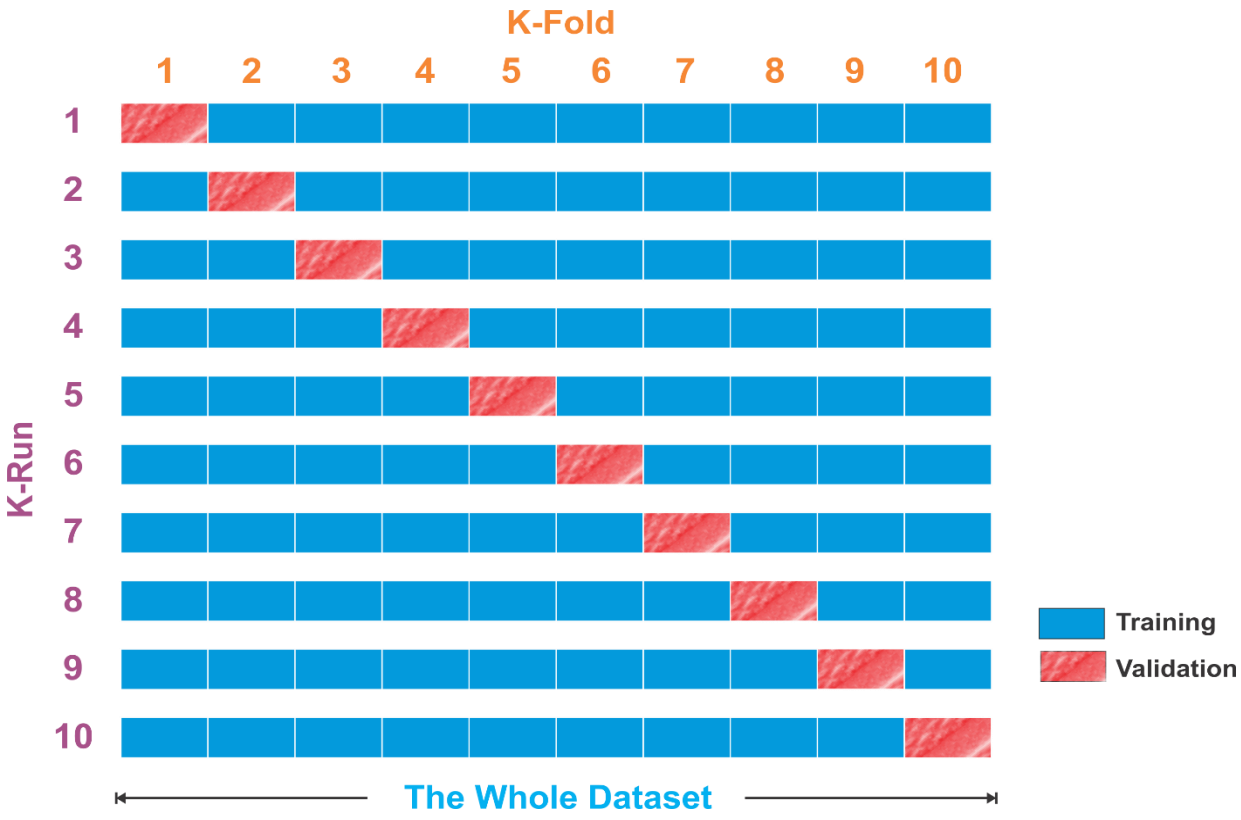


Figure 3. 3: Diagram of 10-fold cross-validation(Babatunde et al. 2015)

- **Leave-one-out cross-validation:** is a type of cross-validation in which the number of folds is equal to the number of instances in the dataset. A single item will be selected instance for the test set and all other instances used as the training set, then the learning algorithm is applied once for each instance. This method is advantageous in unbiased performance estimation. But there is a huge variance(Webb et al. 2011).

In this study, 10-Folds cross-validation was used to evaluate the performance of the models in regression and classification learning algorithms on unseen data.

3.6 Model Evaluation Metrics

Model evaluation metrics are a measure the performance of models and represent how well a model performs and how well it approximates the relationship (Yashwanth 2020). In this study,

different types of machine learning algorithms are used their evaluation metrics depend are different from one to another.

Regression Learning Model Evaluation Metrics

To evaluate the regression learning algorithms used in this study like gradient boosting regressor, random forest regressor, and multiple linear regression the following metrics are used.

- ✓ Mean Square Error
- ✓ Mean Absolute Error
- ✓ R-squared

Mean Squared Error: it is the most common metrics to measure the performance of regression. It is calculated by the average of the difference between squared values of predicted (pi) and actual values (ai) (Yashwanth 2020).

$$\text{MSE} = \frac{\sum_{i=1}^{k} (pi - ai)^2}{n} \quad \text{eq 3.4}$$

Equation 3. 2: Calculate Mean squared error

Mean Absolute Error: It is calculated by the average of the difference between the target value (yi) and the predicted value(xi) (Yashwanth 2020). Mathematically calculated as follows,

$$\text{MAE} = \frac{\sum_{i=1}^{n} |yi - xi|}{n} = \frac{\sum_{i=1}^{n} |ei|}{n} \quad \text{eq 3.5}$$

Equation 3. 3: Calculate Mean absolute error

R-squared: represents the proportion of how much variation of a dependent variable is explained by an independent variable/s in a regression model statically. Whereas correlation measures the strength of the relationship between your model and the dependent variable. R-squared explains to what degree the variance of one variable describes the variance of the other variable(Yashwanth 2020)(FERNANDO 2020).

$$R^2 = 1 - \frac{\text{UnexplainedVariation}}{Total\,Variation} \quad \text{eq 3.6}$$

Equation 3. 4: Calculate the R-squared

3.6.1 Clustering Model Evaluation Metrics

In clustering or unsupervised machine learning, there is no standard or well-accepted model evaluation unlike supervised; Because of that, it is challenging to evaluate clustering algorithms.

Elbow method: is helps data scientist in finding the optimal number clusters by fitting models with a range of values for K. as it is name indicates if the line of the chart looks like an arm the elbow it indicates that the model fits best at that point (Umargono, Suseno, and S. K. 2020).

3.6.2 Classification Learning Algorithms Evaluation Metrics

The classification algorithm must be evaluated to see if the model is performing well on unseen data and correctly predict. Model evaluation metrics are required to evaluate models since the classification algorithm does not give the same result. The classification metrics are Confusion matrix, Accuracy, precision, recall, and F1-score. The four listed metrics next to the confusion matrix just depend on the numbers inside the confusion matrix.

3.6.2.1 Confusion Matrix

The performance of the classification model can be described by using a confusion matrix. It is a table that has dimensions to describe the classification result. It is relatively simple to understand, but some dimensions or terminology can be confusing (Chauhan 2020).
The confusion matrix has four dimensions

True Positive is the condition at which actual value and predicted value are true.

True Negative is the condition at which actual value true and predicted value False.

False Positive is the condition at which the actual value is false and the predicted value is true.

False Negative is the condition at which both the actual value and the predicted value are false.

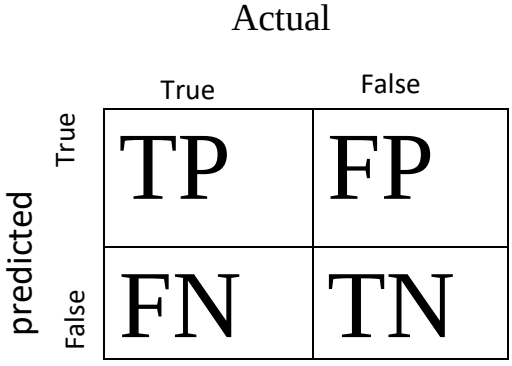


Figure 3. 4: Confusion Matrix

3.6.2.2 *Accuracy*

Accuracy is the metric that evaluates how often is the classifier correct? It's the number of correct predictions made as a ratio of overall predictions made(Zheng 2015). Accuracy can be calculated by using the equation:

$$Accuracy = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \ldots\ldots\ldots\ldots \text{Equation 3.8}$$

Equation 3. 5: Calculate the Accuracy

3.6.2.3 *Precision*

Precision is used to identify how many predicted are truly relevant out of items classified as relevant (Zheng 2015). It shows how many times the model predicts true.

Precision is calculated by the equation:

$$precision = \frac{\text{TP}}{\text{TP}+\text{FP}} \ldots\ldots\ldots\ldots\ldots\ldots \quad equation\ 3.9$$

Equation 3. 6: Calculate the precision of the model

3.6.2.4 *Recall*

The recall is used to answer the question from truly relevant items, how many are found by the classifiers(Zheng 2015). It is used to measure how our model identifies the true positive class. The recall is calculated by the equation:

$$Recall = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad \ldots\ldots\ldots \text{eq 3.10}$$

Equation 3. 7:Calculate the recall of the model

3.6.2.5 *F1-score*

F1-score is a weighted mean of the model's recall and precision. F1-score is calculated by the equation:

$$F1 - score = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{eq 3.11}$$

Equation 3. 8: Calculate the F1-score of the model

CHAPTER FOUR

4. DESIGNING PROPOSED SOLUTION

4.1 Overview

In this chapter, the overall model architecture of the proposed solution for the identification and classification of the highly productive Lentil varieties affected by the disease is described and discussed. In addition to the architecture, the process and steps that are used for the implementation of the models are also discussed. To answer the research questions defined; we defined the models used in this study in the previous chapter. The proposed solution architecture and process for the problem raised in the above chapters are discussed in this chapter.

4.2 The Proposed Solution for Identification of Productive Lentils

The proposed solution architecture for the identification and classification of highly productive lentil varieties is shown in **figure 4.1** below. The overall proposed architecture of this study is a guideline for the final work of this study. Activities that are important in this study are machine learning techniques for solving the problem. Data is the soul of machine learning algorithms since it is data-driven. After data collection the implementation process starts on collected data; it takes the dataset as input. After that data preprocessing tasks like cleaning, formatting, normalizing, and filling missing values of data followed. The next step is the model-building process. The model was built using machine learning algorithms, regression algorithms for predicting the yield of lentils like random forest regressor, gradient boosting regressor, and multiple linear regression were used. The clustering algorithms for grouping genotypes of lentils to apply characteristics analysis of clusters depends on their yield like a standard k-means clustering and K-means initialized by K-means++ algorithms were used. The random forest, decision tree, support vector machine, and naïve Bayes classification learning algorithms were used to classify new genotypes into clusters of genotypes. The recommendation of productive lentil genotypes is the final work of this study. To recommend the genotypes for the breeding program collaborative filtering was used. The model was evaluated using 10-fold cross-validation. The result of performance evaluation or testing accuracy was used to select the best prediction model. Finally, the models used in this study either prediction or classification algorithms are evaluated and selected based on the results obtained using the model evaluation metrics described in the chapter above.

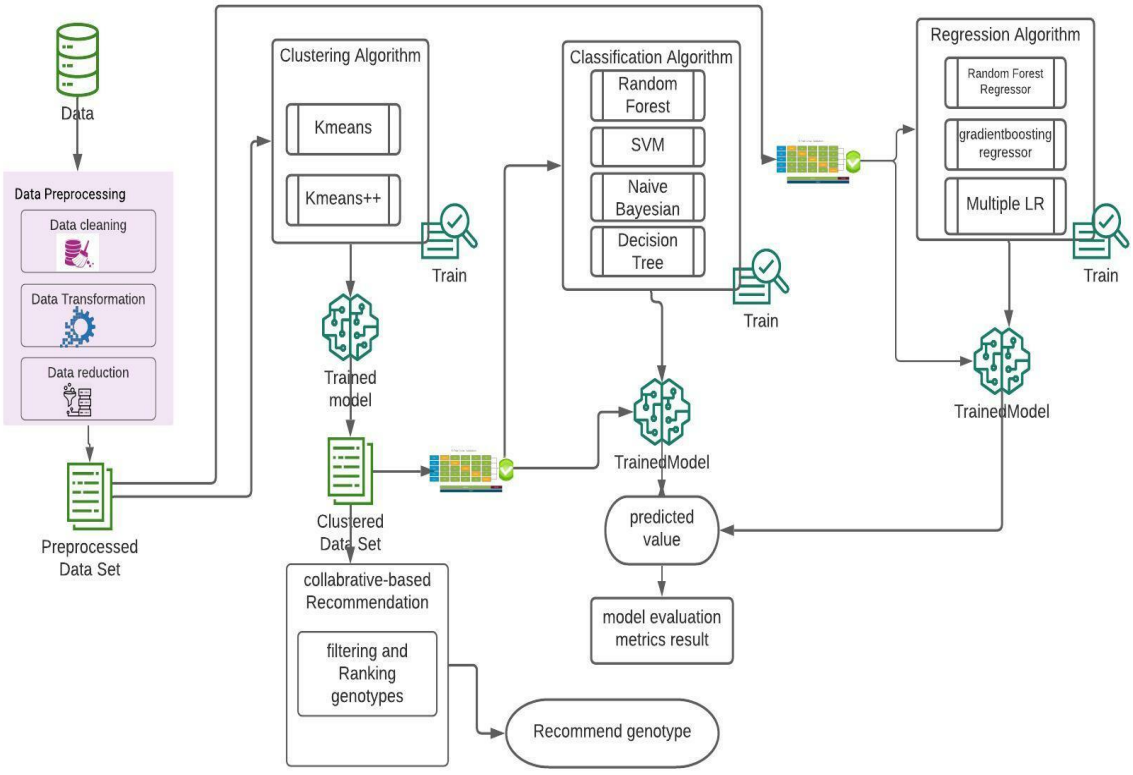


Figure 4. 1: The architecture of the proposed system

4.3 Data Preprocessing Algorithm

Our data may contain different things to be processed like missing values with nan, irrelevant features that may affect training time and prediction model, and unnormalized data that needs to scale. And also, machine learning models require integer data to train it, but since there are features that have string values it needs to label them. So the process and steps followed to process our data to get a cleaned dataset are shown below.

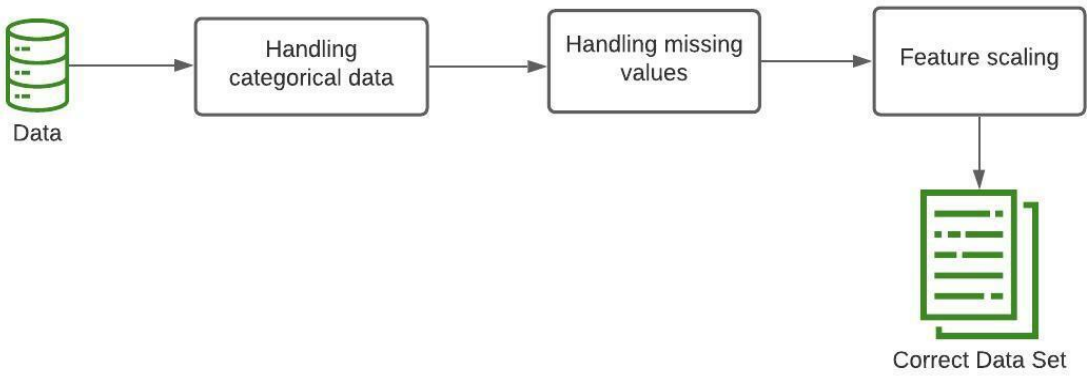


Figure 4. 2: Data preprocessing processes

4.4 Hyperparameter Optimization for Machine Learning Models

Hyperparameters are points of configuration that let a model of machine learning be customized for a specific dataset. It is used to guide the learning process for a specific dataset; it is a model configuration parameter definite by the developer (Yang and Shami 2020). In this study, the optimization method of hyperparameters is used to optimize machine learning algorithms. The most common optimization algorithms are grid search and random search.

Grid search: .specify the volume to be searched as the grid of hyperparameter values and evaluate every position in the grid.

Random Search: specify the volume to be searched as a bounded domain of hyperparameter values and random sample points in that domain

Hyperparameter steps are shown in the figure below.

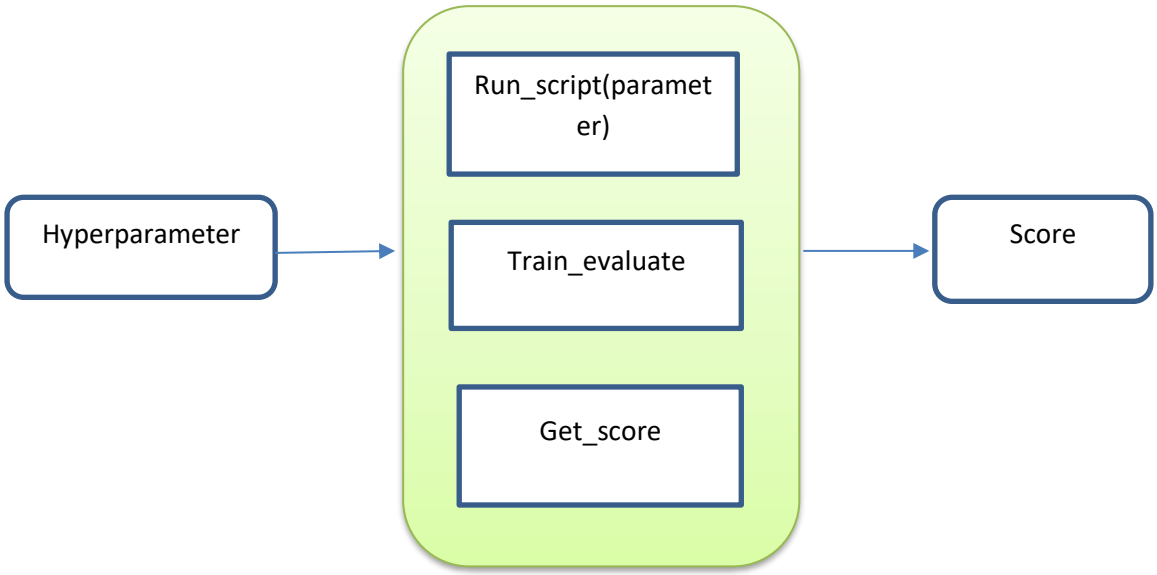


Figure 4. 3: Hyperparameter process

As shown in the figure above the process of hyperparameter optimization of models starts from hyper-parameter, Then run the scripts or parameters, after that evaluate all parameters, and find the best parameters depending on their score. In this study, we used random search methods for hyperparameter optimization of the machine learning algorithms.

4.5 Regression Algorithms Process in Lentils Yield Prediction

Yield prediction is a useful task in the identification and classification of highly productive lentil varieties. To predict the yield of lentils regression learning algorithms of supervised learning is used. The overall processes or design of architecture to predict the yield of lentils

in supervised learning are described in **figure 4.4** below. Input features to predict the yield of lentils are year, location, days of flowering, days of maturity, plant height, hundredweights of seed, Biomass yield per plot, rust, root rot, wilt, vigor, PMW. Output feature or dependent variable is the yield per plot.

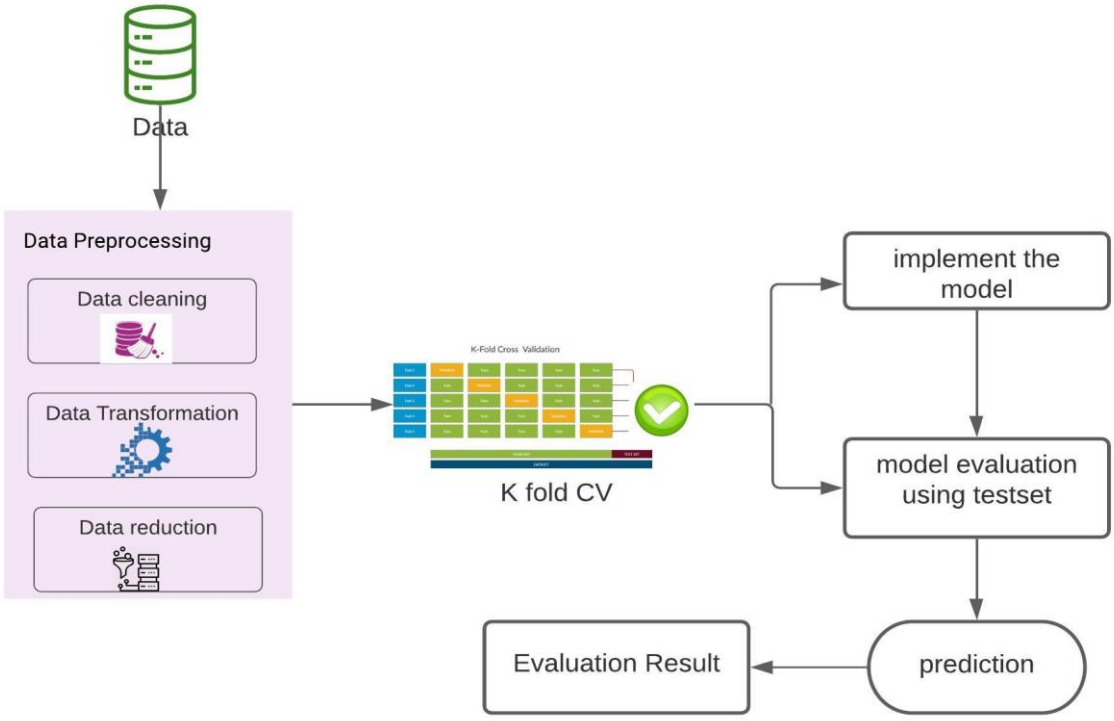


Figure 4. 4: Yield prediction algorithm steps

To predict the yield of lentils random forest, gradient boosting, and multiple linear regression are used and the performance of each algorithm is evaluated by using 10folds cross-validation and their result is compared by using evaluation metrics.

4.5.1 Random Forest Algorithm

In this study, a random forest regressor was used in the prediction of the lentil's yield per plot. It is a supervised machine learning algorithm used for regression and classification problems(Chakure 2019). Random forest algorithm has the following advantages than other supervised machine learning algorithms. It generates an internal unbiased estimate of the generalization error, As the forest grows. It offers a method for guessing missing data that works well and retains accuracy even when a high percentage of the data is missing. It is exceptionally precise, can manage a large number of input variables without deleting any of them, it is very manageable. In the diagram below we showed the steps and structure of the random forest.

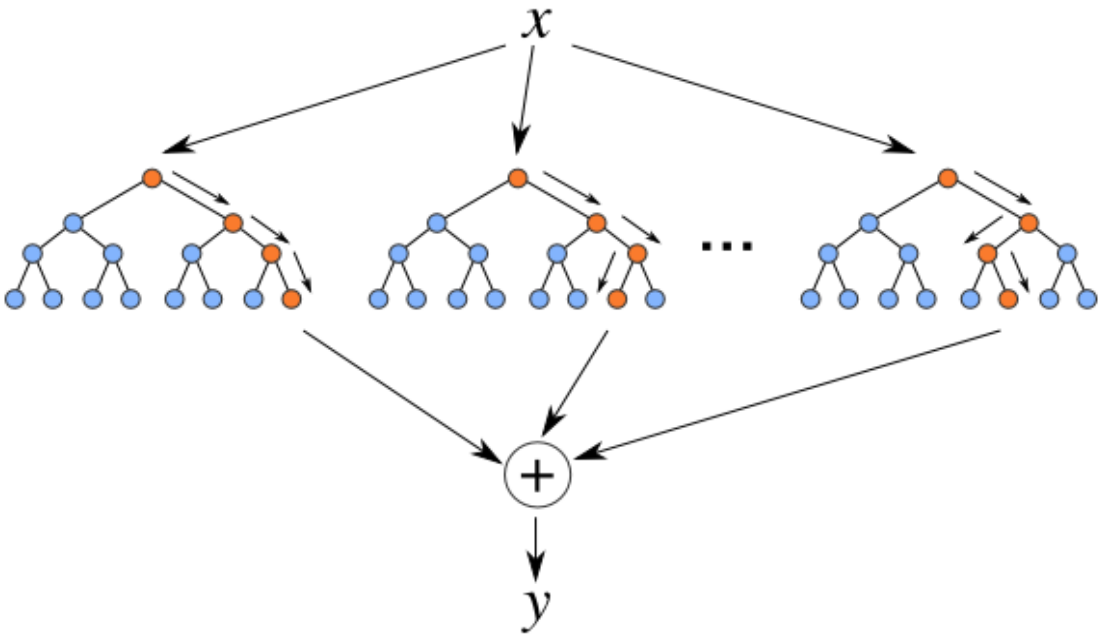


Figure 4. 5: Structure of Random Forest model(Chakure 2019)

Random Forest algorithm steps

Input: lentils agronomic traits and disease data

Output: classification label

Begin:

1. The algorithm selects randomly *k* data samples from our dataset.
2. The algorithm builds a decision tree associated with these *k* data points.
3. Choose the number *N* of trees you want to build and repeat steps 1 and 2.
4. Make each of the N-tree trees forecast the value of Y for the data point in question for a new data point, and then assign the new data point to the average of all of the predicted Y values. Predicting will be performed by using the mean for the regression problem.

End:

4.6 Genotype Clustering Algorithms

Grouping lentil genotypes using genotype features is one major task in this study. To analysis the characteristics of each cluster we need to cluster using clustering algorithms. We cluster and label the data into groups of genotypes. Depend on their characteristics and yield status groups of genotypes were analyzed. The process to implement clustering algorithms to cluster data is described in **figure 4.6** below.

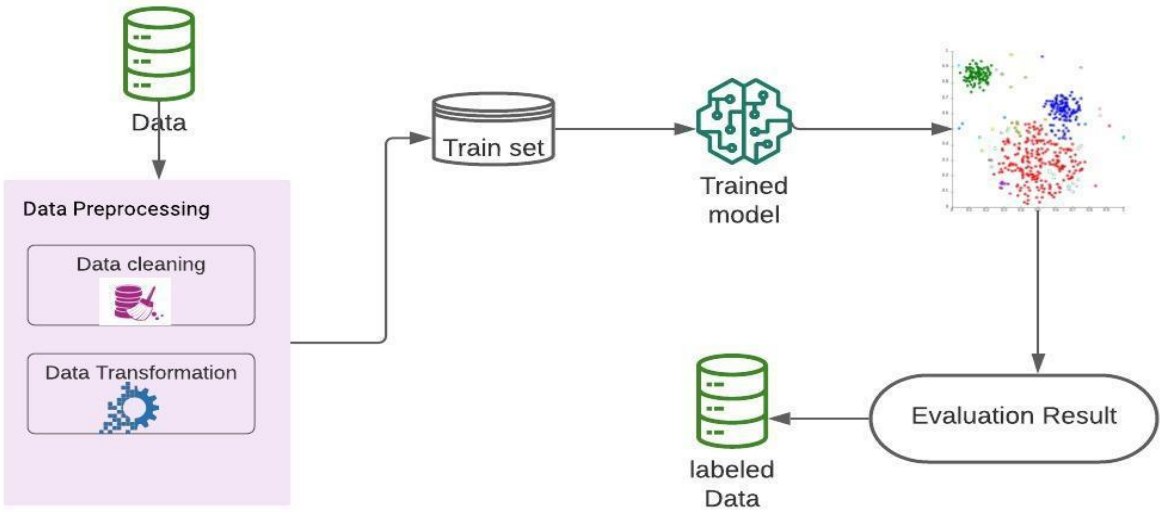


Figure 4. 6: Clustering algorithm steps

4.6.1 K-means Clustering Algorithms

K-means clustering is an unsupervised learning algorithm that is a partitional clustering. K-means clustering is a centroid-based algorithm. The centroid of each cluster in the K-means clustering algorithm can be specified depending on the K value. The algorithm of K-means clustering in this study is shown below(Ullman et al. 2014).

The steps how standard K-means Clustering works are shown below

Input: Lentils agronomic traits and disease data

Output: a set of clusters

1. Specify the number of k of clusters depend on the result from the elbow method.

2. Randomly initialize k centroids and calculate the distance (Euclidian distance) between the cluster center and each object.
3. Assign each object to the closest cluster center or centroid based on distance.
4. Compute the new centroid of each cluster by calculating the mean of objects in each cluster.
5. Repeat steps 3 and 4 until centroids' positions do not change.
6. Return cluster.

The steps how K-means Clustering initialized by Kmeans++ works are shown below

Input: Lentils agronomic traits and disease data

Output: The set of clusters

1. Select the first centroid at random from the data points.
2. Calculate the distance between each data point and the nearest, previously determined centroid.
3. Then, using the data points, choose the next centroid so that the likelihood of selecting a point as a centroid is proportional to its distance from the nearest, the previous centroid. (i.e. the location furthest from the nearest centroid is more likely to be chosen as the next centroid).
4. Steps 2 and 3 should be repeated until k centroids have been sampled.

4.7 Genotype Recommendation

4.7.1 Recommendation Algorithm

In this research as mentioned in the previous chapter, a recommender system was used to recommend genotypes of lentils for breeding programs. Recommendation of lentil genotypes for breeding is one work of this investigation. The collaborative-based filtering recommendation algorithm is applied to enhance the recommendation of lentil genotypes. Since K-means clustering models were used to improve the collaborative-based filtering model-based filtering types of collaborative-based was applied. In this study, the recommendation system is done in two ways at cluster level and genotype level.

1. Cluster level recommendation steps.

After applying the K-means clustering algorithm to cluster genotypes depending on their characteristics, those clustered data is used in the recommendation system. Since there are different genotypes in every cluster, need to find the relation between each cluster to find their similarity and recommend the genotypes that can produce good products when breeding with specified genotypes. This can be recommended at the cluster level.

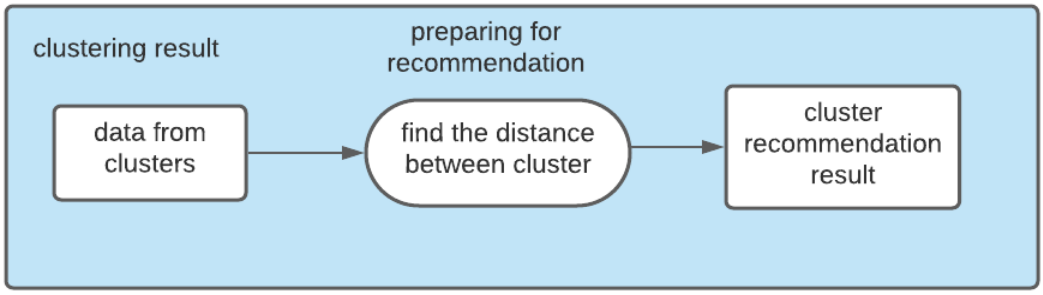


Figure 4. 7: Cluster-level recommendation

2. Genotype level recommendation

At this level also there is a similar process with the recommendation at the cluster level but, at this stage, the similarity between genotypes is calculated depending on every genotype's characteristic. The more diverged genotypes have the scope to create varied genotypes if hybridized. The process to recommend genotypes are shown figure below.

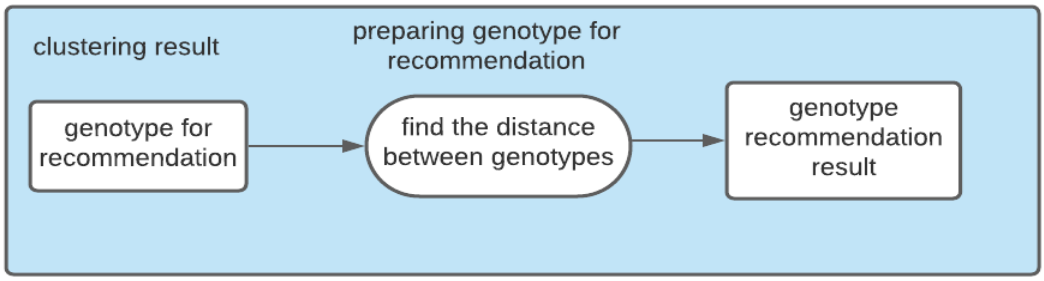


Figure 4. 8: Genotype level recommendation

4.8 Model Evaluation Process

The model evaluation is almost the final process in building the proposed solution. In this study, the model evaluation is performed on the dataset to check whether the model is properly trained depending on our dataset. As explained in chapter three elbow method has been used to find the proper number of clusters for clustering algorithms and cross-validation model evaluation has been used in this study to evaluate the performance of the model. After evaluate and validate using cross-validation the evaluation metrics like accuracy, precision, and recall are used to describe the result of classification algorithms. In yield prediction after cross-validation is applied and the evaluation metrics in regression like mean absolute error, mean square error, and R-square is used to check the performance of our model on the test set.

4.9 Integrating Machine Learning Models with Server

In this study, we aimed to predict yield and recommend productive genotypes of lentils that are affected by the disease. To classify the new genotype, we need to deploy it on the server after we train our model on our dataset. To implement the GUI interfaces, a HTML form was used to insert new input that connected to the flask server and the flask server connected to a trained model that is used to classify the new input to its group.

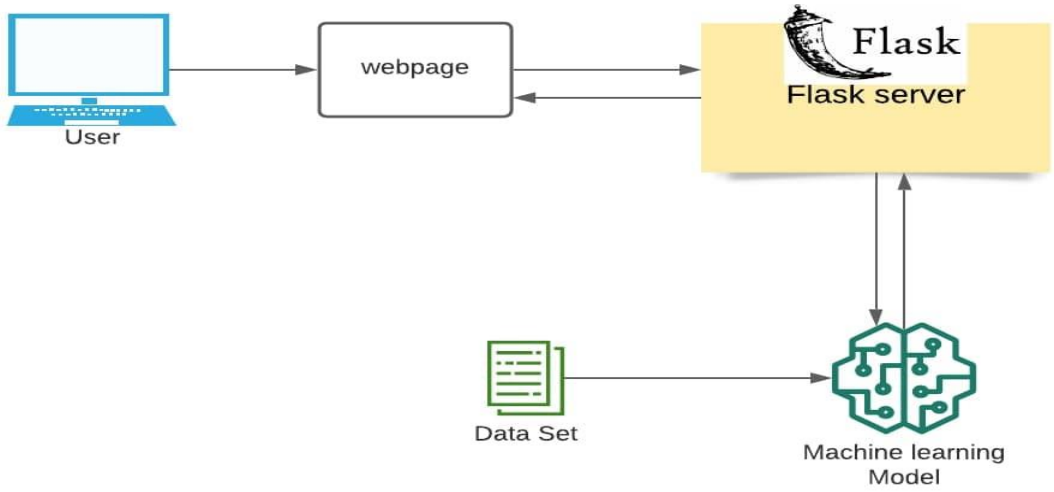


Figure 4. 9: Integrating with web server architecture

CHAPTER FIVE

5. EXPRIEMENT AND IMPLEMENTATION

5.1 Overview

The architecture of the proposed system is designed in chapter four using methods and techniques stated in the methodology part. Additionally, in this chapter, the experiments done on the datasets are described. The implementation environment and experimental setup are expressed in this section.

5.2 Working Environment

To accomplish the objectives of this study we have to use different techniques, tools, and programming languages.

- **Hardware Tools**

Hardware tools used in this study are:

- Laptop- with the following specification:
  RAM:8GB, HDD:1TB, core i5 processor

- **Software Tools**

Software tools used in this study in writing machine learning algorithms and editing documents are:

- Anaconda IDE
- MS office for writing documents and edit data.
- MS Visio and draw.io for diagram and figure drawing.

5.2.1 Programming Language

We used Python for creating and evaluating the model. For writing codes of yield prediction, clustering, classification, and recommendation.

Why Python?

Because it has better for numerical and categorical prediction and there are many frameworks and libraries are available in the python language.

5.2.2 Implementation Environments

To finalize this study different tools, packages and libraries are used in implementing the proposed solution. As stated above programming language used in this study is Python programming language. Python contains many packages and libraries. Starting from data preprocessing to model evaluation the tools, packages, and libraries used are listed below.

Table 5. 1: Description of Tools, and python packages used in the implementation.

Tools and packages	Version	Description
Package managers and Environments		
AnacondaNavigator[2]		is a desktop graphical user interface included in the Anaconda distribution that allows you to conveniently manage conda packages, environments, and channels without having to use command-line commands. Packages can be found on Anaconda.org or in a local Anaconda Repository using Navigator.
Jupyter Notebook[3]	6.1.4	Is an open-source web tool for creating and sharing documents with live code, equations, graphics, and narrative text. Data cleansing and transformation, numerical simulation, statistical modeling, data visualization, and machine learning are all examples of applications.
Python[4]	3.7.0	It is a powerful programming language to develop machine learning. It is easy to learn.
Data preparation tools and packages		
Microsoft Excel	2019	Is one product of Microsoft that was used in this study to prepare data for implementation.
Pandas[5]	1.0.5	is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. Easy to use data structures and analysis packages. Used for data reading, writing, and handling the data frame.

[2] https://docs.anaconda.com/anaconda/navigator/
[3] https://jupyter.org/
[4] https://www.python.org/
[5] https://pandas.pydata.org/

Numpy[6]	1.18.5	which provides support for multi-dimensional arrays.
Scikit-learn[7]	0.23.1	Is an open-source, commercially usable - BSD license Simple and efficient tool for predictive data analysis. Accessible to everybody, and reusable in various contexts.
Scipy[8]	1.5.0	This algorithm is used with NumPy.
Result and data visualization		
Matplotlib[9]	3.2.2	Used for data visualization, result display.
Seaborn	0.10.1	Used for statistical data visualization based on matplot. This package is used for drawing attractive statistical graphics.
Deployment tools		
Flask[10]	1.1.2	It is a web application framework written in python. it is easier for a web developer to write applications without having to worry about the details like protocol management, thread management.

[6] https://numpy.org/
[7] https://scikit-learn.org/stable/
[8] https://www.scipy.org/
[9] https://matplotlib.org/
[10] https://flask.palletsprojects.com/en/2.0.x/

5.3 Identification and Classification of Lentil Genotypes Implementation Steps

The technical and experimental part of this study is the implementation of the proposed system. To implement the proposed system, we followed the following steps starting from loading data from the file to the final evaluation of the model using cross-validation and evaluation metrics.

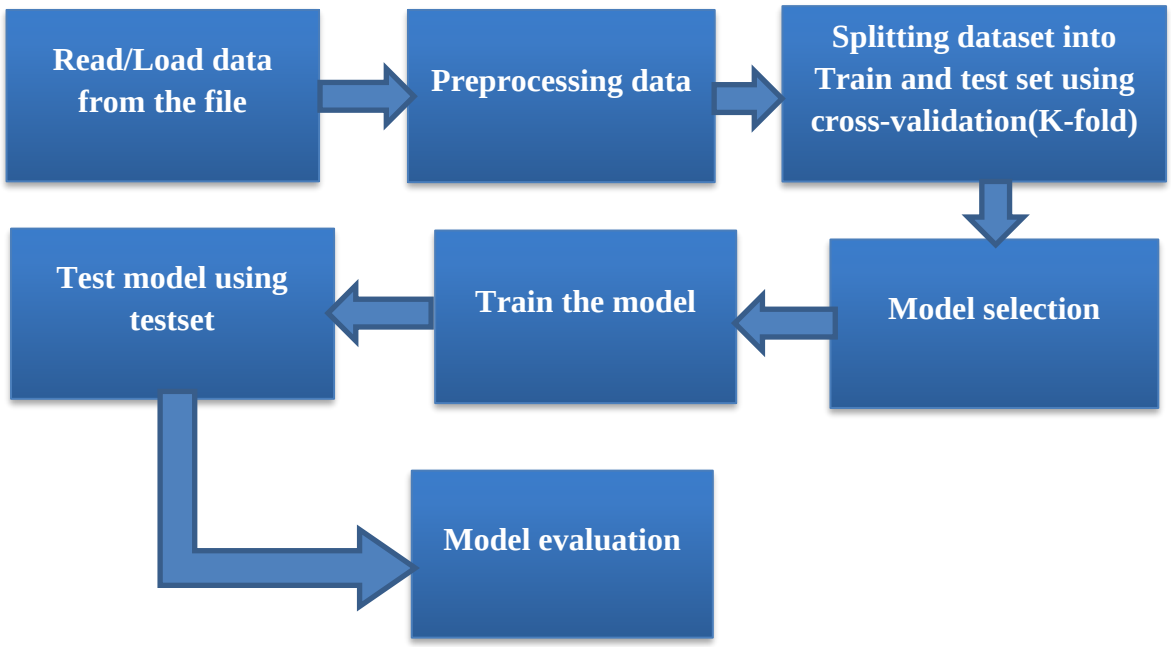


Figure 5. 1: Proposed system implementation steps

5.4 Data Preprocessing Implementation

Data needs preprocessing before being feed into our model. As stated in chapters three and four different preprocessing algorithms will be used to prepare the final dataset. The implementation of preprocessing activities in this study includes encoding categorical data, handling missing values, normalization, feature selection. To process data, we have to load data from the file using pandas and convert it into data frames. For this study, our data was saved in the CSV file extension. The code to load data using panda's library is as follows.

```
#load data
dataset = pd.read_csv("datakoo.csv")
```

Figure 5. 2: load dataset from the file using CSV extension

5.4.1 Implementation of Handling Missing Values

To handle missing values in data in machine learning there are different techniques. In this study, we used the mean value to replace missing values with fillna() method. The python code for filling missing data is written below.

```
dataset['BMYP'].fillna(dataset['BMYP'].mean(), inplace=True)
dataset['PLH'].fillna(dataset['PLH'].mean(), inplace=True)
dataset['DM'].fillna(dataset['DM'].mean(), inplace=True)
dataset['DF'].fillna(dataset['DF'].mean(), inplace=True)
dataset['HSW'].fillna(dataset['HSW'].mean(), inplace=True)
```

Figure 5. 3: Code to fill missing values

5.4.2 Implementation of Encoding Categorical Values

Non-numerical values in the dataset should be encoded to numeric to build an appropriate dataset. In this study, we use python programming and Labelencoder() method that is used to replace categorical string values with numerical values.

```
labelencoder_x=LabelEncoder()
dataset['Locn']=labelencoder_x.fit_transform(dataset['Locn'])
dataset['GENOTYPE']=labelencoder_x.fit_transform(dataset['GENOTYPE'])
```

Figure 5. 4: Code to implement label categorical data

5.4.3 Implementation of Feature Scaling

Normalization is important in a dataset to scale/transform the data that are not normalized. Since the data we have, has different range features; to make all the elements lie between 0 and 1 thus bringing all the values of numeric columns in the dataset to a common scale. So, we used a Min-Max scaler technique in this study for the K-means clustering algorithm. To scale our data MinMaxScaler() method was used.

```
from sklearn import preprocessing
re_minmax=preprocessing.MinMaxScaler(feature_range=(0,1))
dataset_norm = re_minmax.fit_transform(dataset)
```

Figure 5. 5: Code of the Min-Max scaler to normalize data

5.4.4 Feature Selection Implementation

The feature selection is important in machine learning to increase accuracy and reduce the time to train the model. In this study, as mentioned in chapter three Univariate Feature selection and Recursive feature eliminations were used and implemented to select features. To implement UFS SelectKBest() method was used. To implement recursive feature elimination RFE() method was used. The code below shows UFS and RFE feature selection code for random forest regressor.

```
def select_features(features, target):
    # configure to select a subset of features
    fs = SelectKBest(score_func=f_regression, k=10)
    # learn relationship from training data
    fs.fit(features, target)
    # transform train input input data
    X_train_fs = fs.transform(features)

    return X_train_fs, fs
X_train_fs, fs = select_features(features, target)
```

Figure 5. 6: Code of UFS feature selection

```
regressor = RandomForestRegressor(n_estimators=1600, min_samples_split= 2, min_samples_leaf=1,
                                  max_features='sqrt', bootstrap=False, max_depth=100, random_state=0)
rfe = RFE(estimator=regressor)
rfe = rfe.fit(features, target)
print("number of features selected %d" % rfe.n_features_)
print("selected fetures: %s" % rfe.support_)
print("feature ranking: %s" % rfe.ranking_)
```

Figure 5. 7: Sample code of RFE feature selection

5.5 Model Implementation

During the implementation of the model in this study, we follow some steps after the prepared dataset set is loaded. To build our model we used sci-kit learning library packages. The first thing we have done here is importing the important packages for preprocessing datasets, training models, and model evaluations.

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_regression
from sklearn.preprocessing import LabelEncoder
import seaborn as sns
from sklearn.model_selection import cross_val_score
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn import metrics
from sklearn import linear_model
from sklearn .model_selection import learning_curve
from scipy.stats import skew
from sklearn.feature_selection import RFE
from sklearn.model_selection import RandomizedSearchCV
from sklearn.model_selection import KFold
```

Figure 5. 8:Code to load important libraries for regression algorithms

After loading packages, to train the model we split the dataset into two features, the independent features, and dependent features. In this study, we split our dataset into a single target feature and thirteen independent features. The code to split the dataset features for yield prediction is written below.

```
features=dataset.drop(['YLD_GM'],axis=1)
target=dataset["YLD_GM"]
```

Figure 5. 9: Code to create independent and dependent features in yield prediction

5.5.1 Hyperparameter Optimization

Hyperparameter tuning is the way of setting the optimal parameter to optimize the learning algorithm. It is the value of parameters that are used to control the learning process. The model parameters values will be set before the learning process starts (Blogathon 2021). Searching algorithms like grid search and random search are most commonly used for hyper-parameter. Hyper-parameter can be tedious, because of searching algorithms. Optimal hyperparameter is used in control overfitting and underfitting of the model. The searching algorithm for hyperparameter used in this study is a random search. Random search is more effective than grid search. Random search randomly selects the values from random sample points on the grid. The code to implement the random search of hyperparameter tunning for the random forest model is implemented by the RandomizedSearchCV() method shown below.

```
from sklearn.model_selection import RandomizedSearchCV
# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
print(random_grid)
```

```
# First create the base model to tune
regressor = RandomForestRegressor()
# Random search of parameters, using 10 fold cross validation,
# search across 100 different combinations, and use all available cores
rf_random = RandomizedSearchCV(estimator = regressor, param_distributions = random_grid,
                               n_iter = 100, cv = 10, verbose=2, random_state=0, n_jobs = -1)
# Fit the random search model
rf_random.fit(features, target)
```

```
rf_random.best_params_
```

Figure 5. 10: Code to implement the hyperparameter tunning

The models are important to be created to tune the parameters. Different models have different parameters to create and train. The implementation code to create the base model of the random forest algorithm and fit the dataset and select the best parameters for the model is shown above.

5.5.2 Yield Prediction Algorithms Implementation

To train our model with the processed dataset to predict the yield of Lentils. Random forest regressor, gradient boosting regressor, and multiple linear regression proposed in this study to implement those algorithms we instantiate and fit our model. To implement random forest regression, RandomForestRegressor() method was used to train.

```
regressor = RandomForestRegressor(n_estimators=1600, min_samples_split= 2,
                                  min_samples_leaf=1, max_features='sqrt',
                                  bootstrap=False, max_depth=100, random_state=0)
regressor.fit(features, target)
```

Figure 5. 11: Code to instantiate random forest regressor and fit method

```
reg = GradientBoostingRegressor(n_estimators=400,learning_rate=0.1,subsample=0.5,max_depth=110,random_state=0)
reg.fit(features, target)
```

Figure 5. 12: Code to instantiate gradient boosting and fit method

```
regr = linear_model.LinearRegression()
regr=regr.fit(features, target)
```

Figure 5. 13: Code to instantiate Multiple Linear Regression and fit method

5.5.3 Clustering Algorithms Implementation

To cluster or group data depending on their traits and yield data some clustering models were trained on our dataset. Standard Kmeans and Kmeans++ algorithms are proposed in this study to implement those algorithms, we instantiate and fit our model. To implement K-means initialized by K-means++ clustering Kmeans() and fit method was used as shown below.

```
model=KMeans(n_clusters=7,init='k-means++',algorithm='auto',n_jobs=1,max_iter=1000, n_init=10, random_state=0)
y_predicted = model.fit_predict(dataset_norm)
y_predicted
```

Figure 5. 14: Code to implement K-means clustering

5.5.4 Classification Algorithms Implementation

To train our model with the processed dataset to classify the newly inputted genotype traits classification algorithm used in this study. Random forest, support vector machine, naive Bayes, and decision tree proposed in this study to implement those algorithms, we instantiate and fit our model. To implement algorithms different methods for models instantiation and fit() method to fit our dataset used in this study.

```
svclassifier = SVC(C=10, gamma=1, kernel='linear')
svclassifier.fit(X,y)
```

Figure 5. 15: Code to instantiate Support vector machine and fit method

```
randomforest = RandomForestClassifier(
    n_estimators=2000, min_samples_split= 5, min_samples_leaf= 1, max_features='sqrt',
    max_depth=30, bootstrap=False, random_state=0, n_jobs=-1)
randomforest.fit(X,y)
```

Figure 5. 16: Code to instantiate Random forest and fit method

```
gnb = GaussianNB()
gnb.fit(X,y)
```

Figure 5. 17:Naive Bayes instantiate and fit method

```
clf = DecisionTreeClassifier(max_depth=20, min_samples_leaf=4, min_samples_split=8)
clf.fit(X,y)
```

Figure 5. 18: Code to instantiate Decision Tree and fit method

5.6 Model Testing and Evaluation Implementation

In this study, machine learning model evaluation is an important part of our work. To evaluate the performance of the trained model we used k-fold cross-validation, 10-fold cross-validation in classification, and regression learning. In 10-fold cross-validation, the dataset is divided into ten equal parties. Where 9-folds were used for training and 1-fold for the test set. To instantiate the cross-validation cross_val_score() method and to predict cross_val_predict() method was used. The code of 10-fold cross-validation is shown below.

```
kfold = KFold(n_splits=10, random_state=0, shuffle=True)
result = cross_val_score(randomforest, X, y, cv=kfold, scoring='accuracy')
y_predif = cross_val_predict(randomforest, X, y, cv=kfold)
print(f'Accuracy: {result}')
print("Accuracy: %.3f%% (%.3f%%)" % (result.mean()*100.0, result.std()*100.0))
print(classification_report(y, y_predif))
model.append("RFC")
acc.append(result.mean())
```

Figure 5. 19: Code for 10-folds cross-validation for random forest regression

To evaluate the K-means clustering algorithm of unsupervised learning, we used silhouette score and inter-cluster distance to evaluate how the group of clusters are distant from each other. To select an appropriate number of clusters in the clustering algorithm elbow method is used in this study. Within cluster sum of squares (WCSS) to measure the variability between

data points. If the line chart looks like an arm, then the “elbow” is a good sign that the underlying model is more suitable at that point.

```
wcss = []
for i in range(1,20):
 kmeans = KMeans(n_clusters=i,init='k-means++',max_iter=400,n_init=10,random_state=0)
 kmeans.fit(dataset_norm)
 wcss.append(kmeans.inertia_)
 print("Cluster", i, "Inertia", kmeans.inertia_)
plt.plot(range(1,20),wcss)
#plt.scatter(7, wcss[6], s = 400, c = 'red', marker='*')
plt.title('The Elbow Curve')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS') ##WCSS stands for total within-cluster sum of square
plt.show()
```

Figure 5. 20: Code to find K-value

The evaluation metrics are also used to evaluate the training performance of the model. The metrics used in this study were implemented by methods metrics, confusion_matrix(), classification_report() from sklearn metrics package. From the confusion matrix, we can calculate accuracy, precision, F1-score, and recall. we can show those metrics by using the classification_report() classification method. And also we can evaluate the performance of the regression by evaluation metrics by calculating errors and R-square by metrics method.

CHAPTER SIX

6. RESULT AND DISCUSSION

6.1 Overview

In this chapter, the discussion of this study and the result obtained from the implementation of the proposed identification and classification of highly productive lentil varieties affected by the disease are presented. The content we presented in this chapter consists of three parts. In the first part, the dataset preprocessing results are explained. In the second part, the result gained from the model building and the performance evaluation result is presented. At the last, a generalized discussion of the result is discussed.

6.2 Data Collection and Preprocessing Result

6.2.1 Feature Selection Result

To reduce overfitting and to train our model faster we need feature selection. Fifteen features were used in the clustering algorithm for characteristics analysis. In this study, feature selection methods were applied. The filter-based feature selection and wrapper feature selection were used, Univariate feature selection of filtering and recursive feature elimination of Wrapper was applied to select important features.

In UFS, the feature importance is used to find features to select. The level of importance of features in the prediction of yield is shown in order of importance in the model in **appendix B.1**. The feature importance provides a value that indicates how useful each feature is in the model. Depending on the value given the higher score features are more important in the model. The rank of the features will be given by comparing them with other features in the dataset. The result from this stage reduces features into features that are more important for prediction. The feature selection result is different depending on the models applied to train our dataset. The result from the feature selections is shown in the table below.

Table 6. 1: Feature selection result of regression models

Models	Selected Features	
	UFS	RFE
RFR	10	6
GBR	8	6
MLR	13	6

6.3 Dataset Correlation Analysis

Our dataset contains fifteen features, fourteen independent variables, and one dependent variable (grain yield). Correlation analysis measures the relationship between two variables. Correlation analysis is used to describe the linear relationship between two continuous variables. Correlation coefficients between two variables are expressed as integers between -1 and 1. As the value approaches zero it represents less of connection. Where the values closer to -1 or 1 represent the greater correlation between the variables.(James 2021). To find the relationship between two variables the formula we use to calculate is written below.

$$R_{xy} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\sum (xi - \bar{x})2 \sum (yi - \bar{y})2}} \ldots\ldots\ldots\ldots\ldots\ldots\text{Eq 6.1}$$

Equation 6. 1: Pearson correlation coefficient calculator

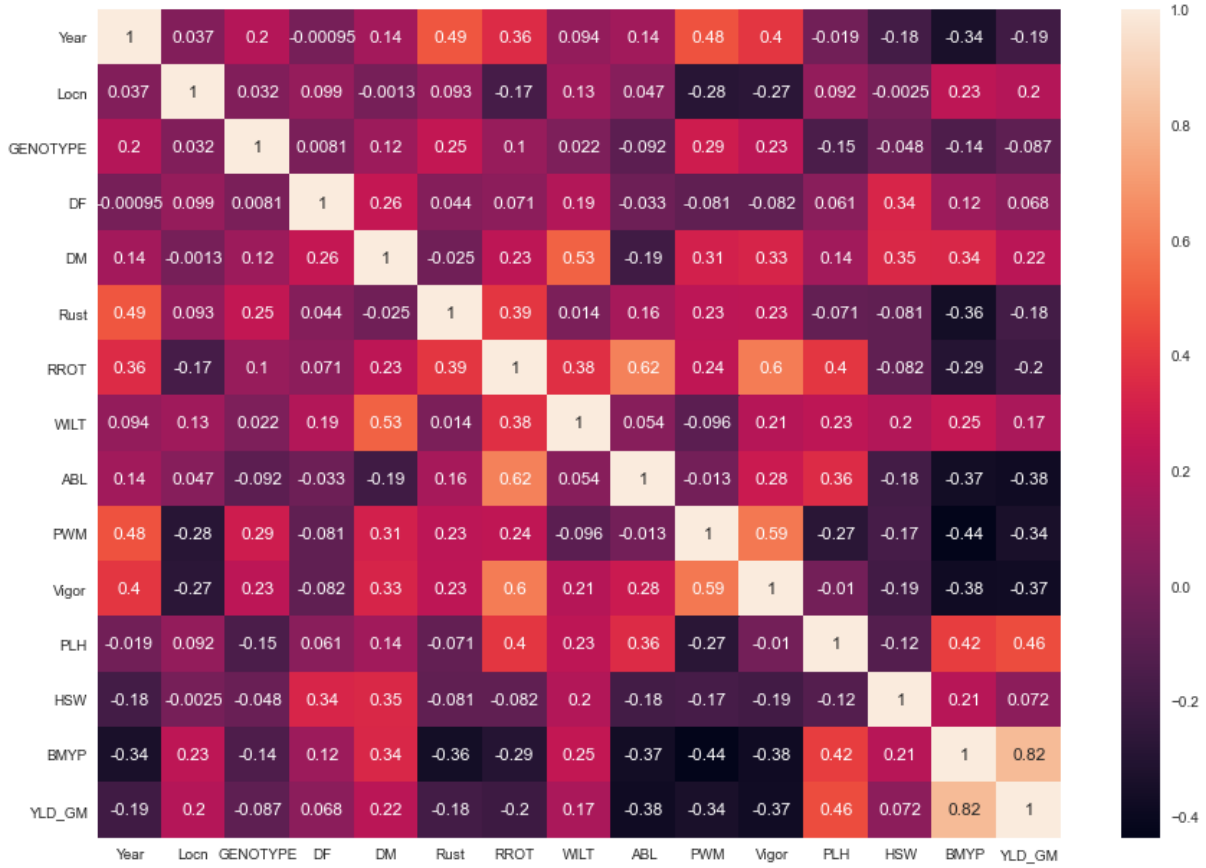


Figure 6. 1: Dataset correlation analysis result

The result of the correlation analysis of our dataset contains intervals between -1 and 1. Those numbers are positive, negative, and zero correlation.

A positive correlation indicates that as one variable increases the other variables will increase proportionally. The value between 0.5 to 1 indicates a strong positive correlation.

A negative correlation means that as one variable increases, other variables will decrease proportionally. The value between -0.5 to –1 indicates a strong negative correlation.

Zero correlation: sometimes called no correlation. It indicates that there is no correlation between the two variables. Affect in one variable doesn't affect the other variable.

In this study, the correlation between each feature is observed and some features have a negative and some have a positive correlation. In our dataset correlations between grain yield and other features are observed and significant positive and negative correlations are obtained. The features/traits that show a positive correlation with grain yield are days of flowering(DF), days of maturity(DM), Wilt, plant height(PHT), 100-seed weight(HSW), biomass yield per plot. A significant negative correlation was observed between lentils yield and Rust, Root rot, ABL, PWM, Vigor, and year. When we compare the correlation values between each feature and lentils yield, biomass per plot and yield per plot has the highest correlation relation value. 100-seed weight and days of flowering have a value near to zero relation with the yield of lentils.

6.4 Model Evaluation Results

6.4.1 Regression Model Result

To predict yields of lentils depending on the agronomic traits and disease data regression model was used. To evaluate the performance of the regression models K-fold cross-validation was used to evaluate. In this study, three regression models were used and compared. Those models are gradient boosting regression, multiple linear regression, and random forest regression. The evaluation metrics like MAE, MSE, and R- square are used to compare the model performance. As presented above the cross-validation was used to evaluate our dataset. 10-Fold cross-validation is used in the prediction of lentils yield depending on the agronomic traits and disease data. The results of evaluation metrics of regression models in each fold are shown below with and without feature selection methods.

- The result of regression models before feature selection or on the original features.

Table 6. 2: The result of random forest regression models before feature selection or on the original dataset

Metrics	10- Fold's cross-validation Random Forest regressor										Av. result
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
MAE	17.27	14.11	17.14	21.1	15.8	15.6	20.6	21.9	17.4	13.8	17.51
MSE	1245.3	996.9	994.4	1570	798.5	746.9	1347.65	1520.6	1016.39	654.8	1089.18
R2	0.95	0.96	0.96	0.95	0.97	0.97	0.95	0.95	0.96	0.97	0.96

Table 6. 3: The result of gradient boosting regression models before feature selection or on the original dataset

Metrics	10- Fold's cross-validation Gradient Boosting regressor										Av. result
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
MAE	18.6	15.4	17.4	20.6	16	16.22	19.9	20.7	17.4	14.7	17.72
MSE	1429.89	1083	913.63	1433.16	1118.7	794.5	1253.84	1316.43	1027.1	762.07	1113.24
R2	0.94	0.95	0.97	0.95	0.96	0.97	0.95	0.95	0.96	0.97	0.96

Table 6. 4: The result of multiple linear regression models before feature selection or on the original dataset

Metrics	10- Fold's cross-validation Multiple Linear regression										Av. result
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
MAE	66.5	60.4	61.3	68.4	69.6	63.6	69.7	76.8	71.1	64.8	67.2
MSE	6991.4	6085.5	5798.13	8212.8	7491.8	7483.27	7857.2	9120.9	8067.9	7023.76	7412.3
R2	0.73	0.75	0.82	0.75	0.79	0.72	0.74	0.72	0.71	0.74	0.75

In Table 6.2, Table6.3, and Table 6.4 above evaluation of three regression models results on each fold of 10-Folds cross-validations (CV) are described without applying the feature selection method. When models are compared the good result recorded by random forest and gradient boosting algorithms having almost close results. The random forest has the value of MAE= 17.51, MSE =1089.18, R2=0.96 than both algorithms.

- The result of regression models after feature selection UFS applied.

Table 6. 5: The result of random forest regressor models after feature selection UFS applied

Metrics	10- Fold's cross-validation Random Forest regressor with UFS.										Av. result
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
MAE	18.8	17.3	19.7	23.4	18.1	17.58	22.7	26.3	22.1	17.2	20.35
MSE	1569.6	1294.73	1364.45	2092.5	1055.3	890.7	1467.27	2267.29	1484.16	1047.9	1453.4
R2	0.94	0.94	0.95	0.93	0.97	0.96	0.95	0.93	0.94	0.96	0.95

The above table represents that the evaluation of the random forest regression model used in this study after applying Univariate Feature Selection (UFS). Ten features are selected depending on feature importance results for predicting the yield of lentils. 10-folds cross-validation was applied to evaluate the model. The performance random forest regressor was evaluated at each fold of cross-validation using evaluation metrics. The average values of evaluation metrics MAE=20.35, MSE=1453.4, R2=0.95.

Table 6. 6: The result of gradient boosting regressor models after feature selection UFS applied

Metrics	10- Fold's cross-validation Gradient Boosting with UFS										Av. result
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
MAE	19.8	17.79	18.2	24.7	17.58	16.4	22.2	26.5	23.35	18.4	20.71
MSE	1683	1308.09	1164	2261.15	1528.5	813.28	1473.36	2480.3	1696.18	1415.19	1582.31
R2	0.93	0.94	0.96	0.93	0.95	0.97	0.95	0.92	0.94	0.94	0.94

In the table above the evaluation of the gradient boosting regression model after applying Univariate Feature Selection (UFS) is described. Eight features are selected depending on feature importance results for predicting the yield of lentils. 10-folds cross-validation was applied to evaluate the model. The performance of the gradient boosting regressor was evaluated at each fold of cross-validation using evaluation metrics. The average values of evaluation metrics MAE=20.71, MSE=1582.31, R2=0.94.

Table 6. 7: The result of multiple linear regression models after feature selection UFS applied

Metrics	10- Fold's cross-validation Multiple Linear Regression with UFS										Av.result
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
MAE	66.5	60.48	61.37	68.4	69.67	63.6	69.77	76.8	71.16	64.8	67.2
MSE	6991.49	6085.54	5798.13	8212.84	7491.84	7473.27	7857.24	9120.9	8067.94	7023.7	7412.3
R2	0.73	0.75	0.82	0.75	0.79	0.72	0.74	0.72	0.71	0.74	0.75

The performance evaluation of the multiple linear regression model is shown in table 6.7. As shown above Univariate Feature Selection (UFS) is applied and thirteen features are selected depending on feature importance results for predicting the yield of lentils. 10-folds cross-validation was applied and The performance of the multiple linear regression was evaluated at each fold of cross-validation using evaluation metrics. The average values of evaluation metrics MAE=67.2, MSE=7412.3, R2=0.75.

When the result of each model is compared the random forest regressor is the better model for predicting the yield of lentils using UFS. When the results of evaluation metrics of algorithms are compared again random forest regressor is the best model with the values of MAE=20.35, MSE=1453.4, R2=0.95.

- ❖ The result of regression models after feature selection RFE applied.

The other feature selection method used in this study to predict yields of lentils using the regression model was Recursive Feature Elimination (RFE). RFE method of feature selection selects six features automatically for all models. The performance evaluation result of regression models used in this study is shown below; by applying 10-folds cross-validation and feature selection RFE.

Table 6. 8: The result of random forest regressor models after feature selection RFE applied

Metrics	10- Fold's cross-validation Random Forest regressor with RFE										Av. result
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
MAE	16.7	15.4	17.4	23.7	17.59	16.0	22.48	23.7	23.5	15.1	19.18
MSE	1147.9	1108.06	914.2	1819.46	1166.1	842.25	1556.78	2091.85	2257.29	859.25	1376.3
R2	0.95	0.95	0.97	0.94	0.96	0.96	0.94	0.93	0.92	0.96	0.95

Table 6. 9: The result of gradient boosting regressor models after feature selection RFE applied

Metrics	10- Fold's cross-validation Gradient Boosting Regressor with RFE										Av. result
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
MAE	19.8	17.79	18.28	24.7	19.58	16.4	22.2	26.5	23.35	18.4	20.7
MSE	1683.0	1308.09	1164.02	2261.15	1528.53	813.28	1473.36	2480.3	1696.18	1415.19	1582.3
R2	0.93	0.94	0.96	0.93	0.95	0.97	0.95	0.92	0.94	0.94	0.94

Table 6. 10: The result of multiple linear regression models after feature selection RFE applied

Metrics	10- Fold's cross-validation Linear Regression with RFE										Average
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
MAE	78.25	73.2	77.4	84.7	85.0	72.9	90.89	97.18	90.9	85.9	83.65
MSE	10870.33	8577.26	10646.94	10640.6	11663.98	9197.2	13680.0	15937.18	14364.45	13496.75	12029.8
R2	0.58	0.65	0.67	0.64	0.67	0.66	0.54	0.51	0.49	0.50	0.59

Using RFE feature selection also random forest is better than both algorithms with the values of MAE=19.18, MSE=1376.3, R2=0.95. In this study, when all models used in lentils yield predictions are compared based on their result of evaluation metrics random forest is better than gradient boosting and multiple regression in all experiments. From experiments done using random forest, it is better at the experiment done without feature selection with MAE=17.51, MSE=1089.18, R2=0.96 values.

6.4.2 Clustering Result

The result from clustering algorithms, when we analyze the result of the K-means clustering model; all genotypes are clustered using agronomic traits (days of flowering, days of maturity, Plant height, hundred seed weight, biomass yield per plot, vigor, yield of grain) and disease data (Rust, Root Rot, Wilt, ABL). As presented in chapter three, for clustering the K-means clustering algorithm was used in this study. K-means clustering uses the Euclidian distance to find the correct cluster centroids and reassign each item to its closest centroids. The steps in K-means clustering are presented in chapter four above. The result gained from each step is presented below.

6.4.2.1 K-means Clustering Algorithm Evaluation Result

Finding K values or number of clusters

In K-means clustering specifying the number of clusters (K) is the first step to build the model. To increase the quality of clustering and produce quality clusters selecting the K values can be done in two ways either manually repeatedly select a random number and check which cluster number is useful in K-means to produce quality clusters or using elbow methods that specify the cluster number depending on our dataset. In this research work, the appropriate number of clusters is selected from the result of the elbow method, starting from the place methods make elbow on the figure. As shown in the figure below according to our dataset the elbow is made at seven number of clusters.

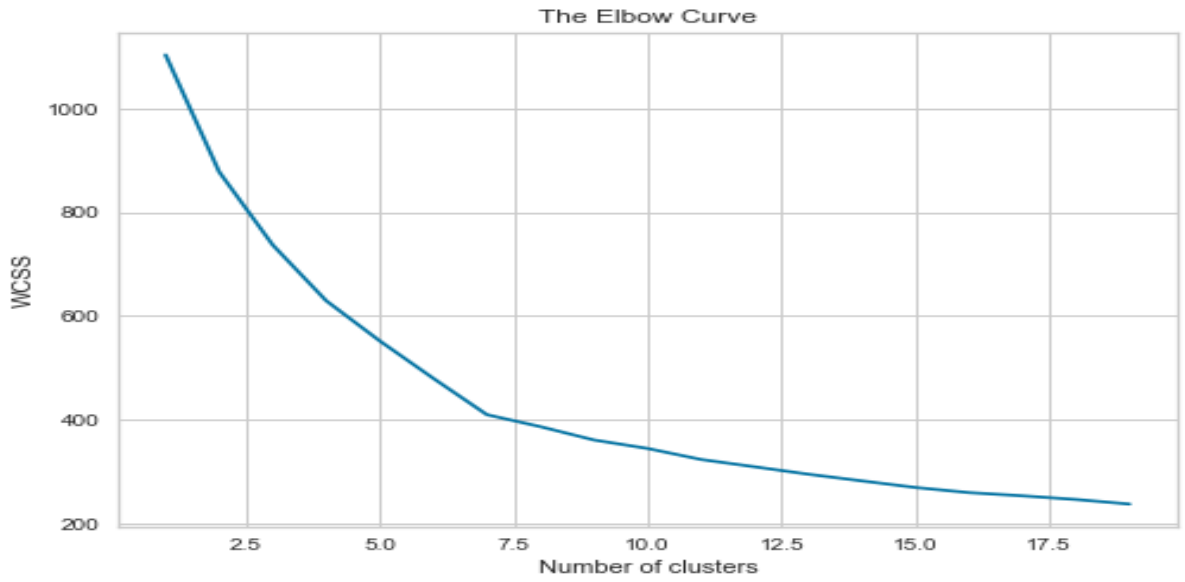


Figure 6. 2:Elbow method result for selecting K values

From the above figure, the elbow method assisted us to decide the number of clusters and select the K value. The place where the elbow method makes sharp elbow is an appropriate point that can be used as the number of clusters (K value). That was at seven clusters as shown in the figure below.

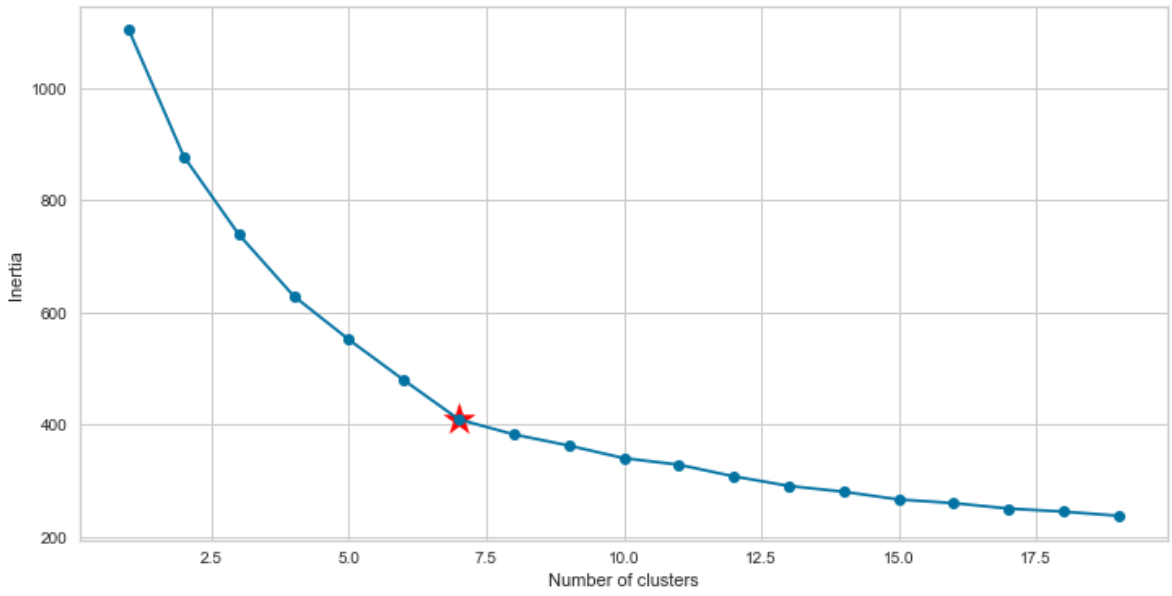


Figure 6. 3: Elbow method that represents the selected K value

After selecting clustering numbers(K) the model was built on the preprocessed dataset. In this study, seven clusters were produced to produce a group of genotypes depends on the number of clusters gained from the elbow method. K-means clustering algorithm uses Euclidean distance to cluster the seven groups of genotypes, seven centroids are specified, and depending on their closeness to the centroid's items are assigned to different clusters by calculating Euclidian distance. In this study, the analysis of characteristics of lentil genotypes was done depending on the agronomic characteristics and disease data. For this analysis and clustering, forty-three genotypes of lentils are collected from DZARC from experiments done between 2013-2017 produced in six locations Akaki, Ambo, Chafe Donso, Debre-Zeit, Arsi Robe, Eneweri. The result of the K-means clustering algorithm contains seven clusters.

6.4.2.2 Silhouette Score Result

The silhouette method was used to find the degree of separation between clusters(Dabbura 2018).

For each sample:

- Compute the distance between all data points in the same cluster (mean distance) (ai).
- Compute the distance between all data points in the closest cluster (mean distance) (bi).
- after that find the coefficient by the equation below.

$$\frac{\text{bi} - \text{ai}}{\max(ai, bi)}$$

The coefficient value should be between -1 and 1 intervals(Bouighoulouden 2020).

- If the coefficient value is 0, it represents the sample is close to neighboring clusters.
- If the coefficient value is 1, it represents the sample is far from neighboring clusters.
- If the coefficient value is -1, it represents the sample is allocated in the wrong clusters.

Depending on our dataset the analysis of genotypes character to identify the varieties between them for clustering. The high silhouette score result shows a good representation of the cluster. The result of the silhouette coefficient in this work has a good clustering result at the K values seven.

6.4.2.3 Inter-cluster distance

The distance measure within-cluster is intra-cluster. In all clusters, the value of inter-cluster distance within a cluster is zero. The distance between different clusters is measured by the inter-cluster distance. The inter-cluster distance between clusters centroids of lentil genotypes is shown below.

Table 6. 11: Inter-cluster distance between clusters of genotypes

	ClusterI	ClusterII	ClusterIII	ClusterIV	ClusterV	Cluster VI	ClusterVII
ClusterI	**0.00**	1.181404	0.897946	1.130955	1.126451	1.06770	1.0353466
ClusterII	1.18141	**0.00**	1.249713	1.191859	1.238072	1.15899	1.2272784
ClusterIII	0.89794	1.249713	**0.00**	0.925644	1.313416	1.02798	0.6311479
ClusterIV	1.13095	1.191859	0.925644	**0.00**	1.158289	0.95276	0.8380507
ClusterV	1.12645	1.23807	1.313416	1.158289	**0.00**	1.49470	1.4323985
ClusterVI	1.06770	1.158992	1.027981	0.952764	1.494705	**0.00**	0.9629769
ClusterVII	1.03534	1.227278	0.631147	0.83805	1.432398	0.96297	**0.00**

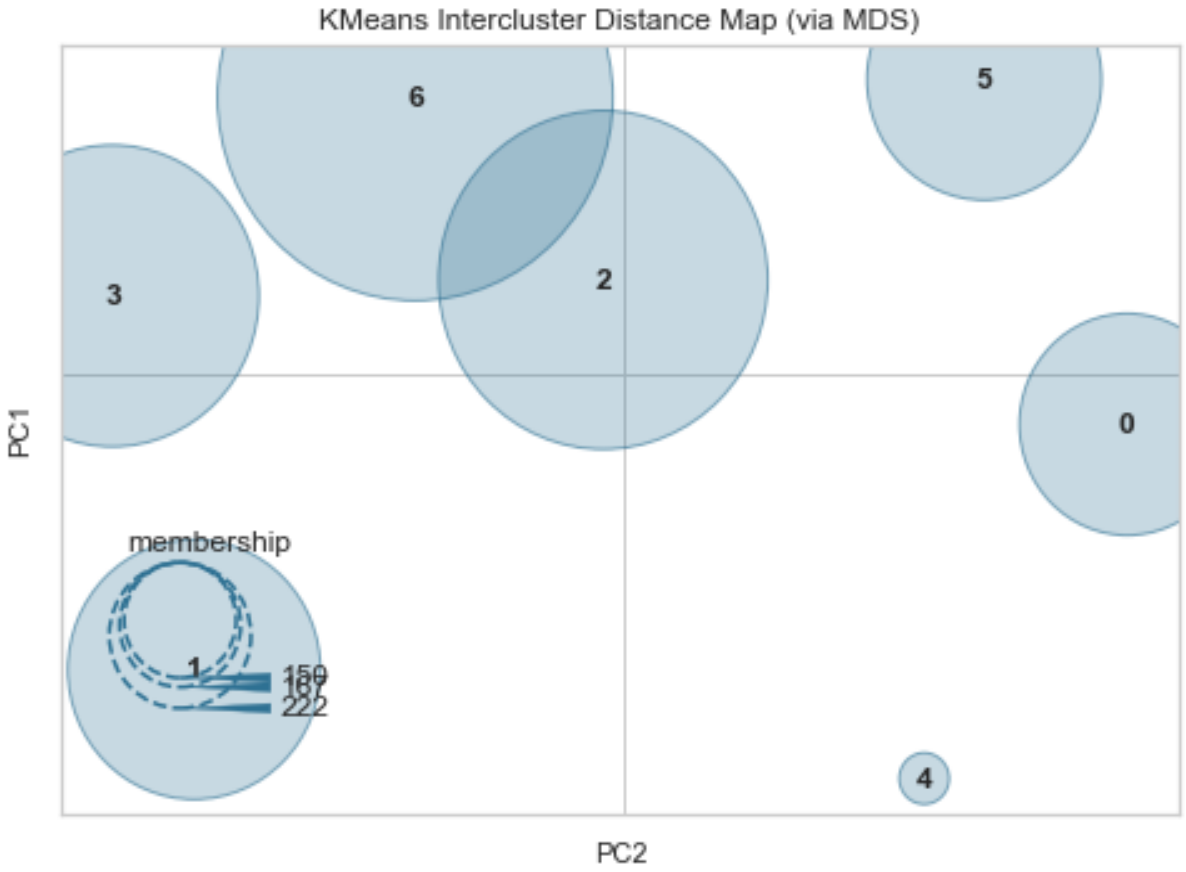


Figure 6. 4 Inter-cluster distance representation

Figure 6.4: Expresses that the inter-cluster distance between seven clusters of lentil genotypes. As represented on the table and figure based on the distance between clusters we can decide the probability to get a good pair of genotypes for the breeding program. It is better to breed dissimilar genotypes to get varied lentil genotypes. Cluster two and cluster six are more similar to each other. Cluster two, four, and six are very close to each other and the similarity of characteristics between them is very close. Clusters one, three, five, and seven are far from each other which indicates there is a wide diversity between them. Depending on their inter-cluster distance we can decide to select genotypes within-cluster with more diverged clusters to create varied genotypes.

6.4.2.4 *Descriptive Characteristics Analysis of Clusters Result*

To understand each cluster the analysis of lentil genotypes characteristics depends on the features used to cluster genotypes are important and each cluster description analysis is described as follows.

	Year	DF	DM	Rust	RROT	WILT	ABL	PWM	Vigor	PLH	HSW	BMYP	YLD_GM
count	136.000000	136.000000	136.000000	136.00000	136.000000	136.000000	136.000000	136.000000	136.0	136.000000	136.000000	136.000000	136.000000
mean	2016.352941	59.426471	112.661765	3.75000	1.066176	2.117647	1.058824	0.904412	1.0	34.161765	3.176471	835.459559	289.029033
std	0.479651	8.142829	14.764934	1.75436	0.326649	1.830033	0.292238	1.485130	0.0	3.845345	0.656730	295.756036	151.503850
min	2016.000000	46.000000	89.000000	1.00000	1.000000	0.000000	1.000000	0.000000	1.0	26.600000	1.800000	350.000000	26.000000
25%	2016.000000	53.000000	100.000000	3.00000	1.000000	1.000000	1.000000	0.000000	1.0	31.000000	2.600000	623.175000	206.799260
50%	2016.000000	58.000000	111.000000	3.00000	1.000000	2.000000	1.000000	0.000000	1.0	34.050000	3.100000	798.850000	277.766375
75%	2017.000000	66.000000	127.000000	5.00000	1.000000	3.000000	1.000000	2.000000	1.0	37.200000	3.700000	996.500000	354.358827
max	2017.000000	79.000000	140.000000	9.00000	4.000000	9.000000	3.000000	6.000000	1.0	46.700000	4.800000	1890.500000	776.600000

Figure 6. 5: Characteristics descriptive analysis of cluster I

In the above figure, cluster one is analyzed using descriptive analysis like count, mean, std, min, 25%, 50%, 75%, max for each feature value. To compare clusters this study focuses on the yield of lentils. 50% of the lentil genotype's yield lies around 277.76. 75% of lentils genotypes yield lies around 354.35. To generalize the value of each feature the mean value is used. The mean value of lentil genotypes yield in this cluster is 289.02 in cluster one.

	Year	DF	DM	Rust	RROT	WILT	ABL	PWM	Vigor	PLH	HSW	BMYP	YLD_GM
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	2015.874251	61.718563	133.814371	2.928144	6.329341	4.568862	6.149701	1.029940	1.832335	47.277468	3.162990	1007.131737	276.142831
std	0.603001	3.720484	8.931886	1.887080	1.918342	1.526785	1.748076	0.255662	0.691190	5.655814	0.640895	251.545238	93.522613
min	2013.000000	54.000000	112.000000	1.000000	3.000000	2.000000	3.000000	0.000000	1.000000	30.000000	1.000000	421.000000	38.000000
25%	2016.000000	58.000000	127.500000	1.000000	5.000000	3.000000	5.000000	1.000000	1.000000	44.000000	3.000000	883.500000	212.810902
50%	2016.000000	62.000000	131.000000	3.000000	7.000000	5.000000	6.000000	1.000000	2.000000	47.300000	3.000000	999.000000	282.340937
75%	2016.000000	65.000000	141.000000	4.000000	8.000000	6.000000	7.000000	1.000000	2.000000	51.100000	3.750000	1192.000000	355.098339
max	2016.000000	67.000000	150.000000	7.000000	9.000000	9.000000	9.000000	3.000000	4.000000	62.000000	4.300000	1738.000000	499.000000

Figure 6. 6: Characteristics descriptive analysis of cluster II

In the above figure, cluster two is analyzed using some of the descriptive analyses like count, mean, std, min, 25%, 50%, 75%, max for each feature value. To compare clusters this study focus on the yield of lentils. 50% of lentil genotypes yield lies around 282.34. 75% of lentils genotypes yield lies around 355.09. To generalize the value of each feature the mean value is used. The mean value of lentil genotypes yield in this cluster is 276.14.

	Year	DF	DM	Rust	RROT	WILT	ABL	PWM	Vigor	PLH	HSW	BMYP	YLD_GM
count	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000	294.000000
mean	2013.857143	59.744898	126.965986	1.047619	1.051020	2.982993	1.159864	0.013605	1.003401	39.711047	3.480272	1800.775510	422.643947
std	0.701047	5.151582	10.261373	0.256874	0.310426	1.237794	0.649196	0.116044	0.058321	2.968125	0.679303	455.322258	134.753379
min	2013.000000	47.000000	94.000000	1.000000	1.000000	0.000000	1.000000	0.000000	1.000000	28.300000	1.700000	700.000000	99.941680
25%	2013.000000	56.000000	121.000000	1.000000	1.000000	3.000000	1.000000	0.000000	1.000000	38.200000	3.000000	1500.000000	341.706737
50%	2014.000000	59.000000	130.000000	1.000000	1.000000	3.000000	1.000000	0.000000	1.000000	39.500000	3.300000	1734.000000	425.410712
75%	2014.000000	64.000000	134.000000	1.000000	1.000000	4.000000	1.000000	0.000000	1.000000	41.475000	4.000000	2200.000000	496.557296
max	2016.000000	71.000000	145.000000	3.000000	4.000000	7.000000	5.000000	1.000000	2.000000	50.700000	5.500000	2700.000000	711.659786

Figure 6. 7: Characteristics descriptive analysis of cluster III

In the above figure, cluster three is analyzed using some of the descriptive analyses like count, mean, std, min, 25%, 50%, 75%, max for traits. To compare clusters this study focuses on the yield of lentils. 50% of lentil genotypes yield lies around 425. 41% of lentils genotypes yield lies around 496.55. To generalize the value of each characteristic the mean value is used. The mean value of lentil genotypes yield in this cluster is 422.64.

	Year	DF	DM	Rust	RROT	WILT	ABL	PWM	Vigor	PLH	HSW	BMYP	YLD_GM
count	222.000000	222.000000	222.000000	222.000000	222.000000	222.000000	222.000000	222.000000	222.000000	222.000000	222.000000	222.000000	222.000000
mean	2014.297297	56.585586	113.833333	1.256757	1.414414	1.617117	1.900901	0.355856	1.031532	37.417866	3.430441	1029.959459	269.385428
std	1.114240	7.310153	10.956688	0.653445	1.006481	1.315510	1.371692	0.858514	0.175144	6.711440	0.801795	393.219837	130.450563
min	2013.000000	44.000000	90.000000	1.000000	1.000000	0.000000	1.000000	0.000000	1.000000	21.600000	1.700000	200.000000	13.546606
25%	2014.000000	51.000000	105.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	32.550000	2.800000	700.000000	172.500000
50%	2014.000000	56.000000	114.500000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	37.200000	3.200000	1022.000000	262.507159
75%	2016.000000	62.000000	120.000000	1.000000	1.000000	3.000000	3.000000	0.000000	1.000000	42.375000	3.966870	1303.000000	386.722661
max	2016.000000	72.000000	143.000000	5.000000	6.000000	5.000000	7.000000	5.000000	2.000000	56.900000	5.700000	1718.000000	576.000000

Figure 6. 8: Characteristics descriptive analysis of cluster IV

In the above figure, cluster four is analyzed using some of the descriptive analyses like count, mean, std, min, 25%, 50%, 75%, max for traits. To compare clusters this study focuses on the yield of lentils. 50% of lentil genotypes yield lies around 262.5. 75% of lentils genotypes yield lies around 386.72. To generalize the value of each characteristic the mean value is used. The mean value of lentil genotypes yield in this cluster is 269.38.

	Year	DF	DM	Rust	RROT	WILT	ABL	PWM	Vigor	PLH	HSW	BMYP	YLD_GM
count	103.0	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000	103.000000
mean	2016.0	51.252427	138.524272	1.796117	2.203883	2.077670	1.029126	5.106796	2.242718	31.007767	2.844660	530.844660	95.477002
std	0.0	37.381373	10.141010	1.239538	1.157746	0.996949	0.168983	1.187456	0.678320	3.354444	0.563207	120.240344	57.502728
min	2016.0	42.000000	119.000000	1.000000	1.000000	1.000000	1.000000	3.000000	1.000000	21.400000	2.000000	200.000000	10.000000
25%	2016.0	44.000000	127.000000	1.000000	1.000000	1.000000	1.000000	5.000000	2.000000	28.600000	2.400000	500.000000	56.500000
50%	2016.0	46.000000	138.000000	1.000000	3.000000	2.000000	1.000000	5.000000	2.000000	31.600000	2.800000	500.000000	79.851134
75%	2016.0	51.000000	148.000000	3.000000	3.000000	3.000000	1.000000	5.000000	3.000000	33.800000	3.200000	600.000000	129.500000
max	2016.0	424.000000	157.000000	7.000000	5.000000	5.000000	2.000000	7.000000	3.000000	37.000000	4.000000	1000.000000	282.765458

Figure 6. 9: Characteristics descriptive analysis of cluster V

In the above figure, cluster five is analyzed using some of the descriptive analyses like count, mean, std, min, 25%, 50%, 75%, max for traits. To compare clusters this study focuses on the yield of lentils. 50% of the lentil genotype's yield lies around 79.85. 75% of lentils genotypes yield lies around 129.5. To generalize the value of each characteristic the mean value is used. The mean value of lentil genotypes yield in this cluster is 95.47.

	Year	DF	DM	Rust	RROT	WILT	ABL	PWM	Vigor	PLH	HSW	BMYP	YLD_GM
count	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.0	150.000000	150.000000	150.000000	150.000000
mean	2013.906667	53.706667	106.006667	1.086667	1.080000	1.093333	5.706667	0.033333	1.0	38.534667	3.007333	881.866667	126.979110
std	0.354201	6.083917	10.492245	0.490674	0.456423	0.535287	2.041669	0.180107	0.0	6.064264	0.804526	291.619976	100.265890
min	2013.000000	42.000000	87.000000	1.000000	1.000000	0.000000	1.000000	0.000000	1.0	25.000000	1.600000	188.000000	14.000000
25%	2014.000000	49.000000	99.000000	1.000000	1.000000	1.000000	5.000000	0.000000	1.0	34.600000	2.400000	600.000000	48.500000
50%	2014.000000	54.000000	105.000000	1.000000	1.000000	1.000000	5.000000	0.000000	1.0	37.200000	2.800000	900.000000	102.000000
75%	2014.000000	59.000000	114.000000	1.000000	1.000000	1.000000	7.000000	0.000000	1.0	39.750000	3.400000	1032.250000	176.923996
max	2016.000000	69.000000	127.000000	5.000000	4.000000	5.000000	9.000000	1.000000	1.0	58.800000	5.100000	1700.000000	410.000000

Figure 6. 10: Characteristics descriptive analysis of cluster VI

In the above figure, cluster six is analyzed using some of the descriptive analyses like count, mean, std, min, 25%, 50%, 75%, max for traits. To compare clusters this study focuses on the yield of lentils. 50% of lentil genotypes yield lies around 102. 75% of lentils genotypes yield lies around 176.92. To generalize the value of each characteristic the mean value is used. The mean value of lentil genotypes yield in this cluster is 126.97.

	Year	DF	DM	Rust	RROT	WILT	ABL	PWM	Vigor	PLH	HSW	BMYP	YLD_GM
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.0	506.0	506.000000	506.000000	506.000000	506.000000
mean	2014.282609	60.195652	131.300395	1.005929	1.003953	3.600791	1.033597	0.0	1.0	40.039130	3.519763	1941.317906	448.908837
std	0.723833	4.672996	5.878698	0.099327	0.062807	1.016069	0.191031	0.0	0.0	3.139176	0.663746	433.533187	133.725096
min	2013.000000	51.000000	99.000000	1.000000	1.000000	0.000000	1.000000	0.0	1.0	21.800000	2.200000	100.000000	17.000000
25%	2014.000000	56.000000	129.000000	1.000000	1.000000	3.000000	1.000000	0.0	1.0	38.400000	3.000000	1700.000000	355.035052
50%	2014.000000	58.500000	132.000000	1.000000	1.000000	3.000000	1.000000	0.0	1.0	39.800000	3.400000	2036.500000	428.513495
75%	2014.000000	65.000000	135.000000	1.000000	1.000000	4.000000	1.000000	0.0	1.0	42.175000	3.975000	2285.250000	547.430779
max	2016.000000	68.000000	140.000000	3.000000	2.000000	10.000000	3.000000	0.0	1.0	50.000000	5.100000	2700.000000	712.000000

Figure 6. 11: Characteristics descriptive analysis of cluster VII

In the above figure, cluster seven is analyzed using some of the descriptive analyses like count, mean, std, min, 25%, 50%, 75%, max for each feature. To compare clusters this study focuses on the yield of lentils. 50% of lentil genotypes yield lies around 428.51. 75% of lentils genotypes yield lies around 547.43. To generalize the value of each characteristic the mean value is used. The mean value of lentil genotypes yield in this cluster is 448.90.

When all clusters are compared based on the yield genotypes within-cluster, cluster VII is highly productive than other clusters. More than 75% percent of lentil genotypes have 547.43 seed yield in cluster VII. The second highly productive lentil genotype is found in cluster III. More than 75% of genotypes in cluster III has 496.55 Yield product. Cluster V contains low-yielding genotypes around 50% of genotypes in this cluster have 79.85 yield products We summarize the results of each cluster from the highest to the lowest clusters VII, III, I, II, IV, VI, V in the productive sequence.

To find the disease-resistant genotypes there is a scale given by the agricultural experts (**shown on appendix D**) to analyze the ability of each genotype on the disease stress. When each cluster is analyzed, cluster IV is highly resistant than other clusters. In all disease data, the genotypes in cluster IV have highly resistant properties in the mean values of each feature of disease data. However, the response of each genotype to disease data is different more than 75% of genotypes in cluster II have around moderately susceptible status in rust, root rot, wilt, and ABL. Genotypes in clusterI are highly resistant to all diseases except resistant behavior to rust and wilt diseases. Genotypes in cluster III are highly resistant to all diseases except resistant behavior in wilt. Genotypes in cluster V are resistant to root rot and wilt and also susceptible to PWM. Genotypes in cluster VI are highly resistant to all diseases except susceptible to ABL. Genotypes in cluster VII are highly resistant to all diseases except resistant behavior to wilt.

6.5 Classification Models Evaluation Result

Imbalanced data may cause one class to dominate other classes. Dataset balancing is the way we can equalize the occurrence between classes. In this study, the data clustered or labeled by clustering algorithms are used to classify the newly inputted data. Labeled data are imbalanced and SMOTE method is used in this study to balance the dataset.

Data Before SMOTE

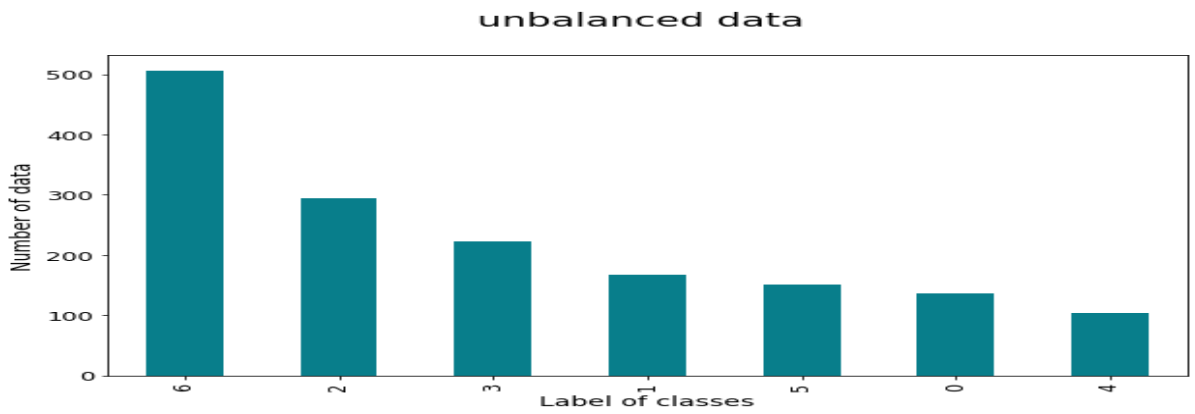


Figure 6. 12: Data before balanced/ unbalanced data

Data After SMOTE

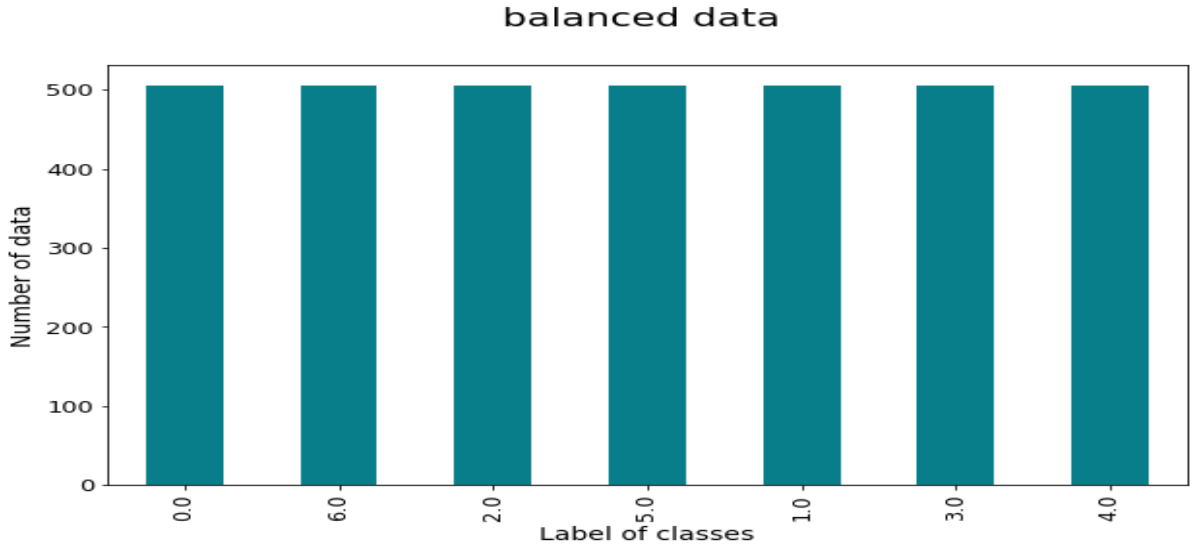


Figure 6. 13: Data after balanced using SMOTE method

Classification model trained on a seven-class dataset after balanced by SMOTE method. The classification report or evaluation metrics of each algorithm: SVM, Random Forest, Naive Bayesian, and Decision Tree models are calculated from the confusion matrix.

6.5.1 Confusion Matrix Result of Classification Algorithms

The results of the performance evaluation in the classification algorithms are calculated from the confusion matrix. It is the best evaluation matrix that describes how our model is trained on our dataset. The result of the confusion matrix of classification algorithms used in this study is shown below.

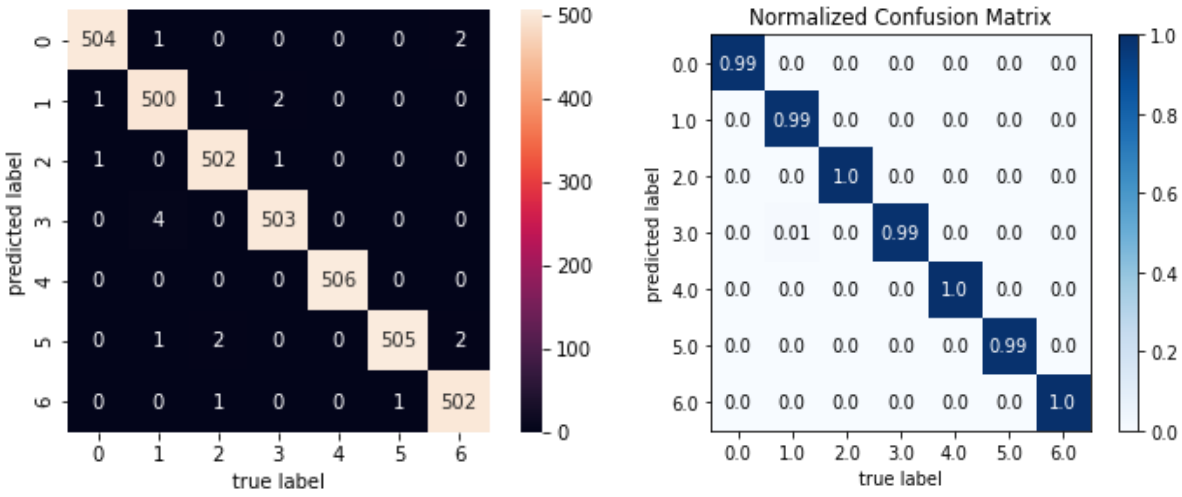


Figure 6. 14: Confusion- matrix result of the Support Vector Machine model

Figure 6.14: Represents the confusion matrix of seven classes dataset trained with 10-folds cross-validation using the SVM model. As shown above, in all classes the model trained well, since the value of true positive is greater than others. The normalized confusion matrix represents SVM classifies 99% ‘class 0 ‘, 99% ‘class 1 ‘, 100% ‘class 2 ‘, 99% ‘class 3 ‘, 100% ‘class 4 ‘, 99% ‘class 5 ‘, 100% ‘class 6 ‘.

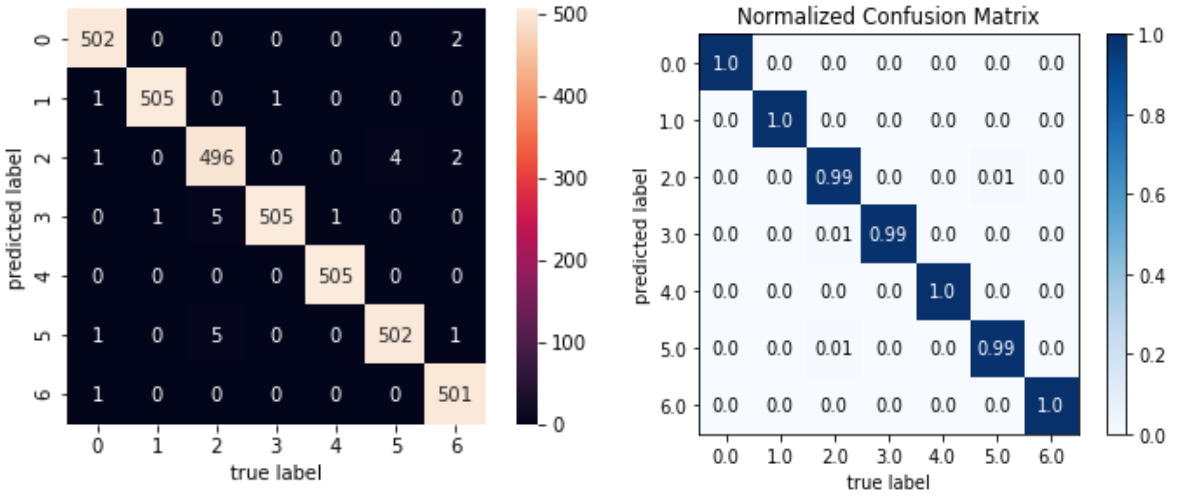


Figure 6. 15: The confusion-matrix result of the Random Forest model

Figure 6.15: represents the confusion matrix of seven classes dataset trained with 10-folds cross-validation using the RF model. As shown above, in all classes the model trained well, since the value of true positive is greater than others. The normalized confusion matrix represents RF classifies 100% ‘class 0 ‘, 100% ‘class 1 ‘, 99% ‘class 2 ‘, 99% ‘class 3 ‘, 100% ‘class 4 ‘, 99% ‘class 5 ‘, 100% ‘class 6 ‘.

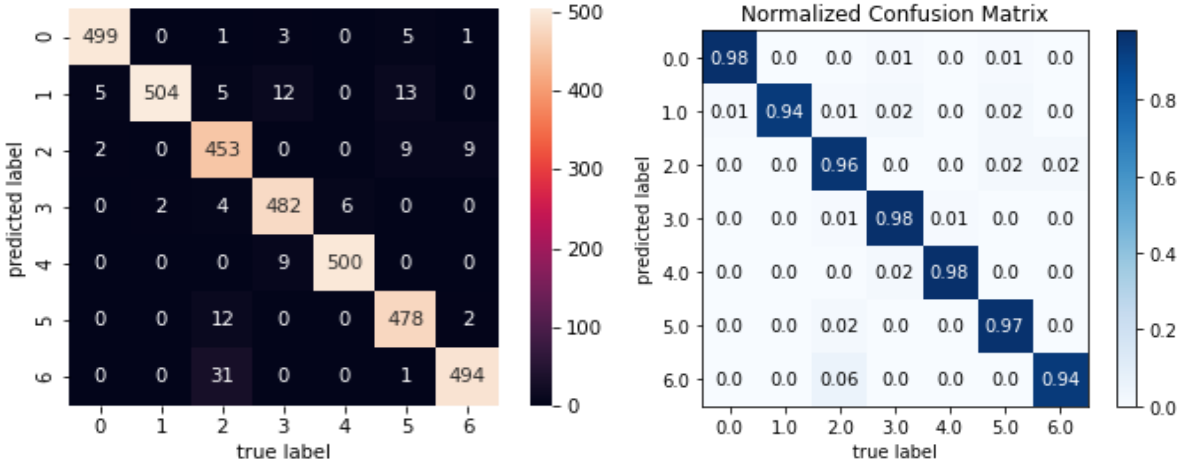


Figure 6. 16: The confusion-matrix result of the Naive Bayes model

Figure 6.16: represents the confusion matrix of seven classes dataset trained with 10-folds cross-validation using the NB model. As shown above, in all classes the model trained well, since the value of true positive is greater than others. The normalized confusion matrix represents NB classifies 98% ‘class 0 ‘, 94% ‘class 1 ‘, 96% ‘class 2 ‘, 98% ‘class 3 ‘, 98% ‘class 4 ‘, 97% ‘class 5 ‘, 94% ‘class 6 ‘.

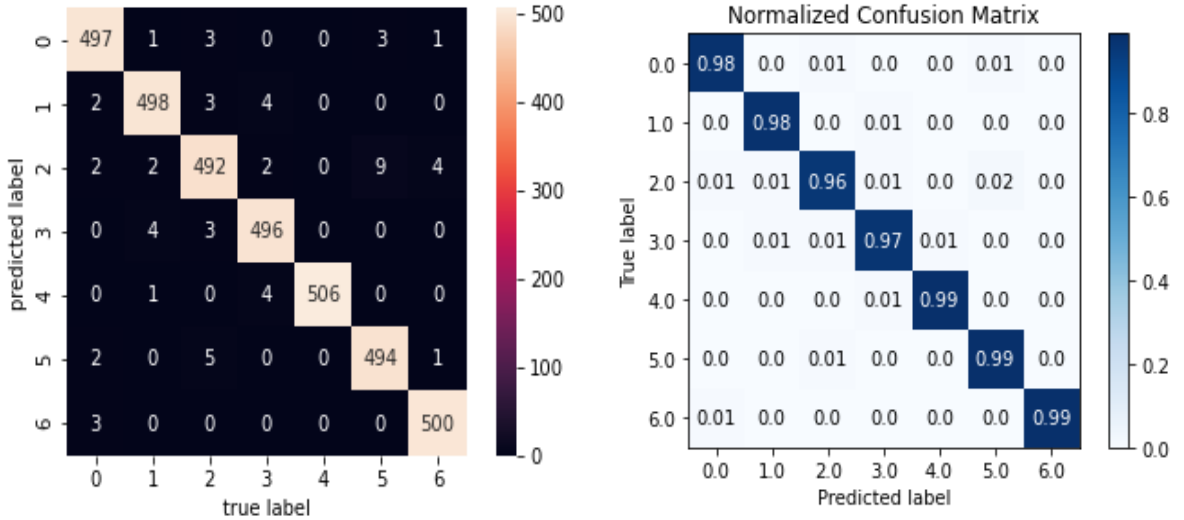


Figure 6. 17: The confusion-matrix result of the Decision Tree model

Figure 6.17: represents the confusion matrix of seven classes dataset trained with 10-folds cross-validation using the DT model. As shown above, in all classes the model trained well, since the value of true positive is greater than others. The normalized confusion matrix represents DT classifies 98% ‘class 0 ‘, 98% ‘class 1 ‘, 96% ‘class 2 ‘, 97% ‘class 3 ‘, 99% ‘class 4 ‘, 99% ‘class 5 ‘, 99% ‘class 6 ‘.

6.5.2 Classification Report

The dataset labeled and obtained from the clustering algorithm is used for training classification algorithms. The data obtained from clustering algorithms has seven classes. The classification evaluation metrics report a precision score, Re-call, and F1-score of SVM, RF, NB, DT models shown in the table below.

Table 6. 12: The result of classification evaluation metrics report

Models	Class	Precision	Re-call	F1-score
SVM	0	0.99	1.00	1.00
	1	0.99	0.99	0.99
	2	1.00	0.99	0.99
	3	0.99	0.99	0.99
	4	1.00	1.00	1.00
	5	0.99	1.00	0.99
	6	1.00	0.99	0.99
RF	0	1.00	0.99	0.99
	1	1.00	1.00	1.00
	2	0.99	0.98	0.98
	3	0.99	1.00	0.99
	4	1.00.	1.00	1.00
	5	0.99	0.99	0.99
	6	1.00	0.99	0.99
NB	0	0.98	0.99	0.98
	1	0.94	1.00	0.96
	2	0.96	0.90	0.93
	3	0.98	0.95	0.96
	4	0.98	0.99	0.99
	5	0.97	0.94	0.96

	6	0.94	0.98	0.96
DT	0	0.98	0.98	0.98
	1	0.98	0.98	0.98
	2	0.97	0.97	0.97
	3	0.98	0.98	0.98
	4	0.99	1.00	1.00
	5	0.98	0.98	0.98
	6	0.99	0.99	0.99

To evaluate the performance of classification models, evaluation metrics are useful to represent the performance of our model numerically. Accuracy is the metric that calculates the performance of the model. In this study, 10-Fold cross-validation is used to evaluate our model on each fold i.e., 90% of the dataset for training and 10% of the dataset for testing. There is an accuracy score on each iteration of 10-Fold cross-validation, and the average accuracy score is used to represent the overall accuracy of our model.

In this study, the classification algorithms used to classify new genotypes into the clustered class are SVM, Random Forest, Naive Bayesian, and Decision tree. Their accuracy was calculated and shown in the table below.

Table 6. 13: Accuracy result of classification algorithm

Models	10- folds Accuracy for SVM, RF, NB, DT										Mean accuracy
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
SVM	99.4	99.4	100	98.8	99.7	98.8	99.4	100	98.8	99.7	99.43
RF	99.1	98.8	99.4	99.4	99.1	99.1	99.1	100	98.8	99.4	99.26
NB	94.9	96.6	97.1	96.3	97.1	96.6	97.1	95.7	93.2	97.7	96.27
DT	98.5	98.5	97.7	97.7	98.5	98.0	97.7	99.7	97.4	98.3	98.24

When the classification algorithms used in this study for prediction groups of new genotypes are evaluated and compared using performance evaluation metrics; the SVM model is better than the RF, NB, and DT with an accuracy value of 99.43%. Naïve Bayesian is the least accurate model when compared with SVM, RF, and DT with 96.27% accuracy.

Table 6. 14: Summary of evaluation metrics of all classification algorithm to compare its performance

Models	Accuracy	F1-score	Precision	Re-call
SVM	99.43	0.99	0.99	0.99
NB	99.26	0.99	0.99	0.99
RF	96.27	0.96	0.96	0.96
DT	98.24	0.98	0.98	0.98

6.6 Recommendation System Result

The recommender system was applied for recommending genotypes for breeding. The experiment was done by using the dataset gained from the clustering algorithm. As stated in the chapters above to recommend genotypes for the breeding program two stages of recommendation were applied. At cluster level and single genotype level. For both recommendation processes to find the similarity between each cluster and each genotype the distance between them was calculated by Euclidean distance measure. When the similarity between each genotype was measured the more similar genotypes are not recommended for breeding; since the narrow genetic variability genotypes were not good in creating varied genotypes(Dugassa, Legesse, and Geleta 2015)(U. Singh, Waza, and Srivastava 2012).

The result of the recommendation system at the cluster for ‘cluster0’ is shown below depend on their distance.

	cluster0
cluster1	1.181404
cluster3	1.130956
cluster4	1.126451
cluster5	1.067701
cluster6	1.035347
cluster2	0.897946
cluster0	0.000000

Figure 6. 18: Cluster level recommendation system

The result of the recommendation system at the genotype level for the ‘DERASH’ genotype is shown below depending on the distance between each genotype. The best 20 genotypes are listed below.

GENOTYPE	
DZ-2012-LN-0018	697.409567
DZ-2012-Ln-0023	543.048185
DZ-2012-Ln-0222	531.250331
DENBI	517.288321
DZ-2012-LN-0024	516.370942
DZ-2012-Ln-0027	515.725800
ALEMAYA	515.193272
DZ-2012-LN-0026	513.782082
DZ-2012-LN-0017	511.340308
DZ-2012-Ln-0026	511.099457
DZ-2012-Ln-0226	507.697220
DZ-2012-Ln-0015	506.758178
DZ-2012-Ln-0056	505.970079
DZ-2012-Ln-0025	505.860988
DZ-2012-Ln-0017	505.616880
DZ-2012-LN-0025	504.444301
DZ-2012-LN-0015	503.777235
DZ-2012-Ln-0024	502.645705
DZ-2012-LN-0019	501.871860
DZ-2012-LN-0022	501.776386

Figure 6. 19: Genotypes recommended for single genotypes

6.7 Discussion

This study focuses on the identification and classification of highly productive lentil varieties affected by disease using machine learning models. There were problems of finding genetic variability, accurate yield prediction, classify new genotypes depending on their traits, and recommendation problems for the breeding program. To solve those problems machine learning Techniques were applied.

To understand the dependencies between features used in yield prediction; according to our knowledge, the Pearson correlation method was used. Using the correlation method, the relation between every two features was identified and calculated. In lentil yield prediction target feature is YLD_GM. When the dependencies between YLD_GM and other features calculated" DF, DM, Wilt, PHT, HSW, BMYP" gives positive dependencies and "Rust, RRot, ABL, PWM, Vigor" has negative dependencies. To predict the yield of lentils different regression machine learning algorithms such as random forest regression, gradient boosting regression, and multiple linear regression were applied. When we compare them using evaluation metrics the random forest gives a better performance than other models used with the value of evaluation metrics experiment done using all features MAE=17.51, MSE=1089.18, and R-squared=0.96.

Characteristics of genotypes of lentils analyzed, differentiated, and clustered in this study depend on their agronomic traits and disease data. To group lentil genotypes, since there is no labeled data or grouped data unsupervised learning algorithms were used in the experiment. The clustering algorithm used in this study was K-means clustering centroid initialized by kmeans++ to cluster lentil genotypes depending on their traits. The elbow method was applied

to find K-values. As shown in **figure 6.2,** The number of clusters seven was the better number to cluster our dataset. Depending on the result of the K-means clustering algorithm analysis of the characteristics of each cluster were described. This algorithm was evaluated using silhouette score and distance between clusters intra-cluster and inter-cluster distance. The intra-cluster and inter-cluster distances between each cluster are shown in **figure 6.4.** when all clusters are analyzed depending on their traits and yield, genotypes in cluster VII is highly proactive and genotypes in cluster IV is highly resistance to all disease than other clusters.

To classify new genotypes into the clustered groups of lentil genotypes, the labeled dataset gained from the K-means clustering algorithm was used. The clustering algorithm labels our data into seven classes. Those classes are not balanced. Since the classification algorithm needs balanced data to equalize each class the SMOTE method was used and applied to oversample our dataset. After the dataset was balanced, the dataset was oversampled to 3542 from 1579. For the classification process, four classification algorithms were applied and compared. Those algorithms were SVM, NB, RF, DT. The performance of each algorithm was evaluated using evaluation metrics as shown in **table 6.13** depending on our dataset SVM is the better algorithm to classify a given genotype with an accuracy of 99.43%.

The final result of this study was recommending the genotypes for the breeding program. The genotypes recommender system was built with a collaborative-based filtering algorithm. Since the recommended genotypes are analyzed and clustered using a K-means clustering algorithm model-based filtering algorithm was applied. The recommendation system in this study depends on the Euclidian distance between clusters and each genotype. Depending on the distance calculated between genotypes the more varied or far genotypes are good in cross-breeding programs. Finally, to find the best models for our study, since different algorithms were used and evaluated. Every learning algorithm has its metrics to evaluate. The clustering algorithm evaluated depends on silhouette score and inter-cluster distance. For regression learning, MAE, MSE, And R-square metrics were used. classification algorithms evaluated by Accuracy, precision, Re-call, and F1-score. To evaluate the performance of models 10-folds cross-validation was used. Depend on their values during the experiments the algorithm that perform than others were chosen for the deployment process. Therefore, for lentils yield prediction Random Forest (RF), for genotypes clustering K-means clustering centroid initialized with kmeans++ algorithm, for classification of new genotypes SVM was selected for the implementation.

CHAPTER SEVEN

7. CONCLUSION AND RECOMMENDATION

7.1 Conclusion

Identification and classification of lentil genotypes are important in producing better lentil varieties. We need to have genotypes that have high yield, disease-resistant behavior, and good quality for the breeding purpose. In this paper, we focus on the identification and classification of highly productive lentil genotypes affected by the disease. For identifying and grouping lentil genotypes, lentil genotypes characteristics were analyzed. Agronomic traits and disease data are the features used in this study. Based on those traits, lentil genotypes were clustered and analyzed to find genetic variability. The analysis process is useful in the breeding program to understand the characteristics of genotypes to improve lentil products.

The purpose of this study is to solve the problems raised like the problem of grouping genotypes to find genetic variability, inaccurate yield prediction, classify new genotypes depending on their traits, and recommendation problems for the breeding programs. To experiment with this study, the dataset was collected from the Debre-Zeit agricultural research center. Significantly those data were preprocessed to make it more understandable and used for building the machine learning models to find the solution.

Based on our dataset the model was developed to cluster into groups of genotypes. The clustering model used in this study was the K-means clustering algorithm to analyze the characteristics of each genotype within the cluster. Depending on the result from the elbow method K-means clustering algorithm groups our dataset into seven clusters. The analysis of each cluster depending on the features exposed and described. The yield prediction experiments were done using regression learning algorithms like the random forest, gradient boosting, multiple linear regression. The random forest model was better in the prediction when compared with gradient boosting and multiple linear regression models using evaluation metrics like MAE, MSE, and R-squared. To find the optimal parameters for algorithms randomly search hyperparameter tuning techniques used and optimized. The Results of random forest regression in all experiments were as follows, without feature selection MAE=17.51, MSE=1089.18, and R-squared=0.96, with UFS MAE=20.35, MSE=1453.4, and R-squared=0.95, with RFE MAE=19.18, MSE=1376.3, and R-squared=0.95 using 10-Folds cross-validation. To classify the new genotypes of lentils into clustered groups of lentil

genotypes the output of the K-means clustering algorithm was exposed to implement the classification algorithms like SVM, RF, NB, DT. The classification experiment was performed on the seven-classes labeled dataset. Since this dataset was not balanced the SMOTE method was used to balance. Each model was evaluated using 10-Folds cross-validation techniques, as we do have limited data and to reduce overfitting. SVM model better test accuracy on our dataset as compared with RF, NB, and DT algorithms with accuracy result 99.43. The other investigation of this study was the recommendation of genotypes for the breeding program to enhance the yield of lentils. The recommendation process was implemented using collaborative-based filtering to filter genotypes based on their similarity and recommend the more divergent genotypes for the breeding program.

This study reveals that machine learning algorithms are important in the agricultural sector, in the traits analysis, cluster, yield prediction, and recommendation of lentil genotypes.

7.2 Recommendation

Agriculture and technologies like machine learning integration are the most important area to analyze large data, to enhance and simplify the workload of workers. It is a huge burden for the agricultural research center on their work improvement since technology integration is not available. For the researchers whose interested in this area, it is better to focus on machine learning techniques rather than classical statistical analysis models. It is better to extend the study to deep learning by including more data.

7.3 Future Work

However, different tasks were done in this study to enhance the identification and classification of lentils varieties using different techniques of machine learning, this area needs more works to improve. We applied different machine learning techniques for clustering, classification, and recommendation. Here K-means clustering needs more optimization of centroids to improve the performance of clustering. The recommendation system in this study uses collaborative-based filtering techniques to recommend genotypes for the breeding program. Collaborative filtering cannot work on new items to recommend, it needs improvement using other recommendation algorithms. In this study, the agronomical traits and disease data were used for characteristics analysis and yield prediction the weather data and soil data are better to be considered to understand the study area and their effects on the yield for future work.

* * *

This research was funded by Adama Science and Technology University with a grant number **ASTU/SM-R/243/21**.

References

Administration, International Trade. 2020. “Agricultural Sector.” *international trade administration*. https://www.trade.gov/knowledge-product/ethiopia-agricultural-sector (January 22, 2021).

Awad, Mariette, and Rahul Khanna. 2015. “Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers.” *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers* (January): 1–248.

Babatunde, Oluleye, Leisa Armstrong, Jinsong Leng, and Dean Diepeveen. 2015. “Comparative Analysis of Genetic Algorithm and Particle Swam Optimization: An Application in Precision Agriculture.” *Asian Journal of Computer and Information Systems Asian Online Journals (www.ajouronline.com)* 03(1): 2321–5658.

Bagheri, Alireza, Negin Zargarian, Farzad Mondani, and Iraj Nosratti. 2020. “Artificial Neural Network Potential in Yield Prediction of Lentil (Lens Culinaris L.) Influenced by Weed Interference.” *Journal of Plant Protection Research* 60(3): 284–95.

Benos, Lefteris et al. 2021. “Machine Learning in Agriculture: A Comprehensive Updated Review.” *Sensors* 21(11): 1–55.

Berrar, Daniel. 2018. “Cross-Validation.” *International Journal of Trend in Research and Development* 1(January 2018): 542–45.

Bhuyan, Rupanka, and Samarjeet Borah. 2013. “A Survey of Some Density Based Clustering Techniques.” *Conf. Advancements in Information, Computer and Communication* (September 2014).

Blogathon, Data Science. 2021. “Evaluating Machine Learning Models Using Hyperparameter Tuning Hyperparameter Types : Hyperparameter Tuning.”

Bogale, Daniel Admasu, Firew Mekbib, and Asnake Fikre. 2017. “Genetic Improvement of Lentil [Lens Culinaris Medikus] Between 1980 and 2010 in Ethiopia.” *Malaysian Journal of Medical and Biological Research* 4(2): 129–38.

Bouighoulouden, Amine. 2020. “School of Science and Engineering Crop Yield Prediction Using K-Means Clustering.”

Bremer, Martina. 2012. “Math 261A -Spring 2012.” *Math 261A -Spring 2012* (1): 18–36. http://mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement 5 - multiple regression.pdf.

Brownlee, Jason. 2016. “Master Machine Learning Algorithms Discover How They Work and Implement Them From Scratch.” *Machine Learning Mastery With Python* 1(1): 11. http://machinelearningmastery.com.

Brownlee, Jason (Machine Learning Mastery). 2019. “MachineLearning Mastery with Python Mini-Course.” *Journal of Chemical Information and Modeling* 53(9): 1689–99.

Center, ModelOp. 2019. “Gradient Boosting Regressor | Open Data Group.” https://opendatagroup.github.io/Knowledge Center/Tutorials/Gradient Boosting Regressor/ (March 28, 2021).

Chakure, Afroz. 2019. “Random Forest Regression. In This Blog We’ll Try to Understand… | by Afroz Chakure | The Startup | Medium.” https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f (April 3, 2021).

Chauhan, Nagesh Singh. 2020. “Model Evaluation Metrics in Machine Learning - KDnuggets.” https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html (May 18, 2021).

Chlingaryan, Anna, Salah Sukkarieh, and Brett Whelan. 2018. “Machine Learning Approaches for Crop Yield Prediction and Nitrogen Status Estimation in Precision Agriculture: A Review.” *Computers and Electronics in Agriculture* 151(May): 61–69. https://doi.org/10.1016/j.compag.2018.05.012.

Cholaquidis, A., R. Fraiman, and M. Sued. 2020. 29 Test *On Semi-Supervised Learning*.

Computing, Object. 2019. “Machine Learning in Agriculture | Machine Learning and Artificial Intelligence Solutions | Object Computing, Inc.” https://objectcomputing.com/expertise/machine-learning/machine-learning-in-agriculture (March 11, 2021).

Contreras, Pedro, and Fionn Murtagh. 2015. “Hierarchical Clustering.” *Handbook of Cluster Analysis* (February): 103–24.

Cunningham, Padraig, Bahavathy Kathirgamanathan, and Sarah Jane Delany. 2021. “Feature

Selection Tutorial with Python Examples." http://arxiv.org/abs/2106.06437.

Cutler, Adele, D Richard Cutler, and John R Stevens. 2012. "Ensemble Machine Learning." *Ensemble Machine Learning* (February 2014).

Dabbura, Imad. 2018. "K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks | by Imad Dabbura | Towards Data Science." https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a (June 21, 2021).

Deng, Kan. 1999. "Chapter 7 Feature Selection." *Phd Thesis*.

Dietterich, Thomas et al. 2005. "Semi-Supervised Learning."

Diriba, Getachew. 2020. *Agricultural and Rural Transformation in Ethiopia: Obstacles, Triggers and Reform Considerations*. https://media.africaportal.org/documents/Agricultural_and_rural_transformation_in_Ethiopia.pdf.

Dugassa, A, H Legesse, and N Geleta. 2015. "Genetic Variability, Yield and Yield Associations of Lentil (*Lens Culinaris* Medic.) Genotypes Grown at Gitilo Najo, Western Ethiopia." *Science, Technology and Arts Research Journal* 3(4): 10.

Eernisae, E P, and J B Snelling. 2016. Hand, The *Machine Learning for Dynamic Software Analysis*.

FDRECSA. 2018. "The Federal Democratic Republic of Ethiopia Central Statistical Agency Report on Area and Production of Crops." *THE FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA CENTRAL STATISTICAL AGENCY* V. I(April 2018): 128.

FERNANDO, JASON. 2020. "R-Squared Definition." https://www.investopedia.com/terms/r/r-squared.asp (March 28, 2021).

Frona, Daniel, Szenderak Janos, and Monika Harangi-Rakos. 2019. "The Challenge of Feeding the Poor." *MDPI Journal for Sustainability* (2017): 3–4.

Gelman, Andrew, and Jennifer Hill. 2010. "Missing-Data Imputation." *Data Analysis Using Regression and Multilevel/Hierarchical Models*: 529–44.

Granitzer, Michael et al. 2009. "Machine Learning Based Work Task Classification." *Journal*

of Digital Information Management 7(5): 305–12.

Greenlaw, Raymond, and Sanpawat Kantabutra. 2010. “Introduction to Clustering.” *Dynamic and Advanced Data Mining for Progressing Technological Development* (June): 224–54.

Hancock, John T., and Taghi M. Khoshgoftaar. 2020. “Survey on Categorical Data for Neural Networks.” *Journal of Big Data* 7(1). https://doi.org/10.1186/s40537-020-00305-w.

Isinkaye, F. O., Y. O. Folajimi, and B. A. Ojokoh. 2015. “Recommendation Systems: Principles, Methods and Evaluation.” *Egyptian Informatics Journal* 16(3): 261–73.

James, Emily. 2021. “What Is Correlation Analysis? Definition and Exploration.” https://blog.flexmr.net/correlation-analysis-definition-exploration? (May 21, 2021).

Judith Hurwitz, Daniel Kirsch. 2018. 35 Journal of the American Society for Information Science *Approaches to Machine Learning*.

Kang, Hyun. 2013. “The Prevention and Handling of the Missing Data.” *Korean Journal of Anesthesiology* 64(5): 402–6.

Kassa, Zelalem. 2018. “RESPONSE OF LENTIL (Lens Culinaris Medik) VARIETIES TO RHIZOBIUM INOCULATION AND INORGANIC FERTILIZER APPLICATION IN SIYA DEBRINA WAYU DISTRICT, NORTH SHEWA ZONE, ETHIOPIA.” Debre Berhan University.

Kindie, Yirga, Zinabu Nigusie, and Fatih Yildiz. 2018. “Participatory Evaluation of Lentil Varieties in Wag-Lasta, Eastern Amhara.” *Cogent Food & Agriculture* 4(1): 1561171. https://doi.org/10.1080/23311932.2018.1561171.

van Klompenburg, Thomas, Ayalew Kassahun, and Cagatay Catal. 2020. “Crop Yield Prediction Using Machine Learning: A Systematic Literature Review.” *Computers and Electronics in Agriculture* 177.

Krstajic, Damjan, Ljubomir J. Buturovic, David E. Leahy, and Simon Thomas. 2014. “Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models.” *Journal of Cheminformatics* 6(1): 1–15.

Lutz, Brian. “Transforming Field Data Into Meaningful Insights | Cropscience.”

https://www.cropscience.bayer.com/innovations/data-science/a/transforming-field-data-meaningful-insights (March 11, 2021).

Mohammed, Nabil A. et al. 2019. “Agro-Morphological Characterization of Lentil Genotypes in Dry Environments.” *International Journal of Agriculture and Biology* 22(6): 1320–30.

Mohanty, Sharada P., David P. Hughes, and Marcel Salathé. 2016. “Using Deep Learning for Image-Based Plant Disease Detection.” *Frontiers in Plant Science* 7(September): 1–10.

Morales, Eduardo F., and Julio H. Zaragoza. 2011. “An Introduction to Reinforcement Learning.” *Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions*: 63–80.

Morissette, Laurence, and Sylvain Chartier. 2013. “The K-Means Clustering Technique: General Considerations and Implementation in Mathematica.” *Tutorials in Quantitative Methods for Psychology* 9(1): 15–24.

Mutuvi, Steve. 2019. “Introduction to Machine Learning Model Evaluation | by Steve Mutuvi | Heartbeat.” https://heartbeat.fritz.ai/introduction-to-machine-learning-model-evaluation-fa859e1b2d7f (April 22, 2021).

Nath, U K et al. 2014. “Selection of Superior Lentil (Lens Esculenta M .) Genotypes by Assessing Character Association and Genetic Diversity.” 2014: 1–7.

Niazian, Mohsen, and Gniewko Niedbała. 2020. “Machine Learning for Plant Breeding and Biotechnology.” *Agriculture (Switzerland)* 10(10): 1–23.

Omran, Mahamed G.H., Andries P. Engelbrecht, and Ayed Salman. 2007. “An Overview of Clustering Methods.” *Intelligent Data Analysis* 11(6): 583–605.

Patro, S.Gopal Krishna, and Kishore Kumar sahu. 2015. “Normalization: A Preprocessing Stage.” *Iarjset* (March): 20–22.

Pennington, Dennis. 2013. “Harvest Index: A Predictor of Corn Stover Yield - MSU Extension.” *Michigan State University Extension*. https://www.canr.msu.edu/news/harvest_index_a_predictor_of_corn_stover_yield (March 29, 2021).

Piiroinen, Jarkko. 2018. “Density-Based Clustering.” (May).

Poczos, Barnabas. “Semi-Supervised Learning.”

Portugal, Ivens, Paulo Alencar, and Donald Cowan. 2018. “The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review.” *Expert Systems with Applications* 97(November 2015): 205–27.

Prettenhofer, Peter, and Gilles Louppe. “Boosted Decision Trees Slides.”

Priya, Muthaiah, and Balamurugan. 2018. “Predicting Yield of the Crop Using Machine Learning Algorithms.” *International journal of engineering sciences and reseach technology* 7(4): 1–7.

Quartey, Emmanuel Kwatei et al. 2014. “Agronomic Evaluation of Eight Genotypes of Hot Pepper (Capsicum Spp L .) in a Coastal Savanna Zone of Ghana.” *Journal of Biology, Agriculture and Healthcare* 4(24): 16–28.

Rajoub, Bashar. 2020. “Supervised and Unsupervised Learning.” *Biomedical Signal Processing and Artificial Intelligence in Healthcare* (January): 51–89.

Raschka, Sebastian. 2018. “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning.” *arXiv*.

Regassa, Senait, Legesse Dadi, Demissie Mitiku, Asnake Fikre, Aden Aw-Hasan. 2014. *Impact of Research and Technologies in Selected Lentil Growing Areas of Ethiopia*. Ethiopian Institute of Agricultural Research.

Rukmi, Alvida Mustika, and Ikhwan Muhammad Iqbal. 2017. “Using K-Means++ Algorithm for Researchers Clustering.” *AIP Conference Proceedings* 1867(August).

Sah, Shagan. 2020. “Machine Learning: A Review of Learning Types.” *ResearchGate* (July). www.preprints.org.

Sahu, Satya Prakash, Anand Nautiyal, and Mahendra Prasad. 2017. “Machine Learning Algorithms for Recommender System - a Comparative Analysis.” *International Journal of Computer Applications Technology and Research* 6(2): 97–100.

Saikenova, Alma Zhumabaevna et al. 2021. “Crop Yield and Quality of Lentil Varieties in the Conditions of the Southeast of Kazakhstan.” *OnLine Journal of Biological Sciences*

21(1): 33–40.

Sathya, R., and Annamma Abraham. 2013. “Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification.” *International Journal of Advanced Research in Artificial Intelligence* 2(2).

Sciforce. 2019. “Machine Learning in Agriculture: Applications and Techniques | by Sciforce | Sciforce | Medium.” https://medium.com/sciforce/machine-learning-in-agriculture-applications-and-techniques-6ab501f4d1b5 (March 11, 2021).

Shweta Bhatt, Youplus. 2018. “Reinforcement Learning.” *KDnuggets*. https://www.kdnuggets.com/2018/03/5-things-reinforcement-learning.html (August 11, 2021).

Singh, Mohar et al. 2020. “Evaluation and Identification of Wild Lentil Accessions for Enhancing Genetic Gains of Cultivated Varieties.” *PLoS ONE* 15(3): 1–14. http://dx.doi.org/10.1371/journal.pone.0229554.

Singh, Umesh, S A Waza, and R K Srivastava. 2012. “Study the Evaluation of Genetic Diversity in Lentil (Lens Culinaris Medik .).” 3(6): 1199–1202.

Stellmacher, Till, and Girma Kelboro. 2019. “Family Farms, Agricultural Productivity, and the Terrain of Food (in)Security in Ethiopia.” *Sustainability (Switzerland)* 11(18).

Ullman, Shimon et al. 2014. “9.54 Class 13.” *Unsupervised learning Slides*: 54.

Umargono, Edy, Jadmiko Endro Suseno, and Vincensius Gunawan S. K. 2020. “K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based-on Mean and Median.” 474(Isstec 2019): 234–40.

Uppada, Santosh Kumar. 2014. “Centroid Based Clustering Algorithms- A Clarion Study.” *International Journal of Computer Science and Information Technologies* 5(6): 7309–13. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.658.3904.

Webb, Geoffrey I. et al. 2011. “Leave-One-Out Cross-Validation.” In *Encyclopedia of Machine Learning*, Springer US, 600–601. https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_469 (April 23, 2021).

Xu, Dongkuan, and Yingjie Tian. 2015. “A Comprehensive Survey of Clustering Algorithms.” *Annals of Data Science* 2(2): 165–93.

Xu, Xiaowei, Martin Ester, Hans-peter Kriegel, and Jörg Sander. 1998. “Published in the Proceedings of 14th International Conference on Data Engineering (ICDE ’ 98) A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases D-80538 München 3 . A Notion of Clusters Based on the Distance Distribution.” : 324–31.

Yang, Li, and Abdallah Shami. 2020. “On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice.” *Neurocomputing* 415(August): 295–316.

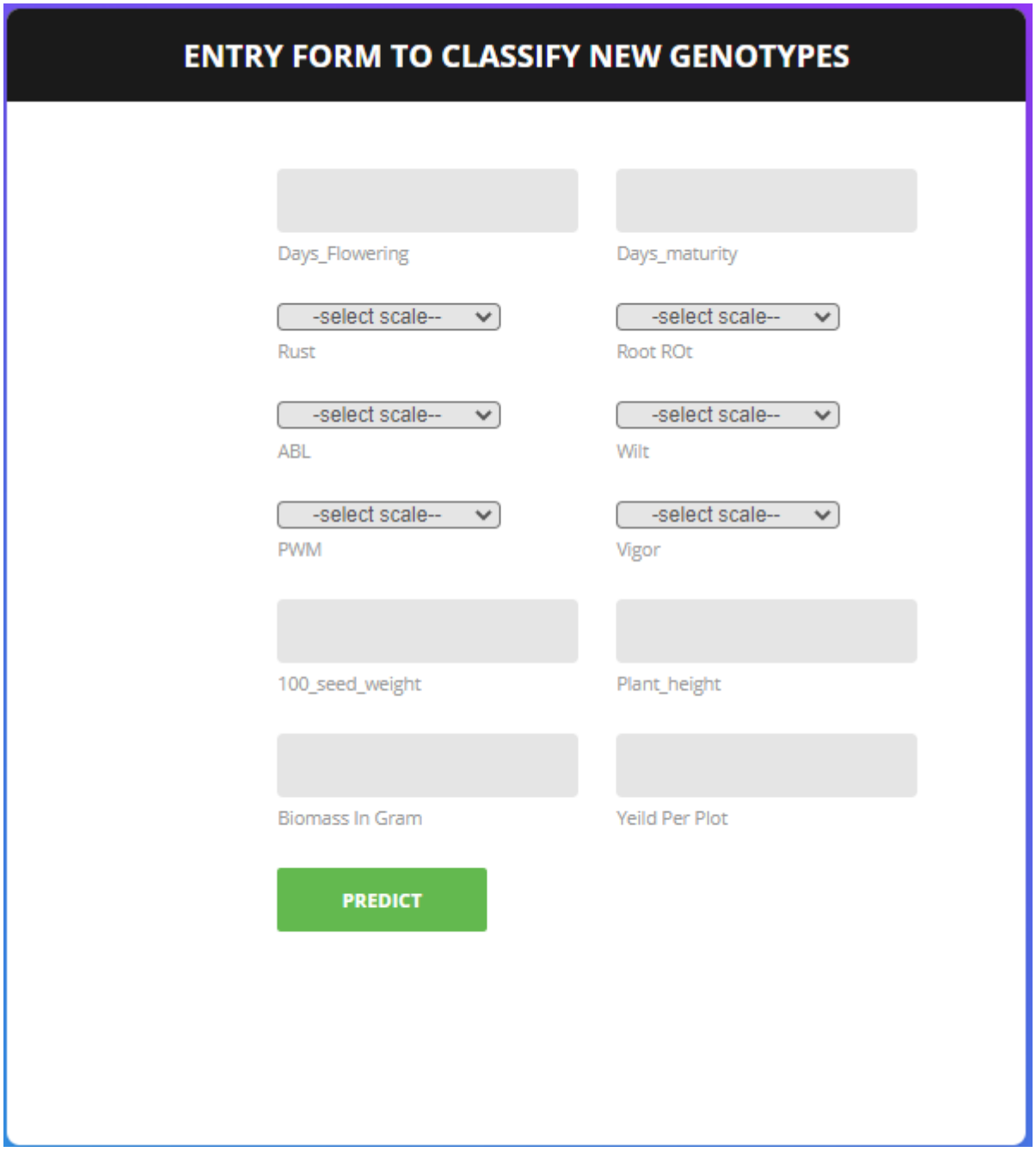
Yashwanth, NVS. 2020. “Evaluation Metrics & Model Selection in Linear Regression | by NVS Yashwanth | Towards Data Science.” https://towardsdatascience.com/evaluation-metrics-model-selection-in-linear-regression-73c7573208be (March 28, 2021).

Zheng, Alice. 2015. Springer *Evaluating Machine Learning Algorithms*.

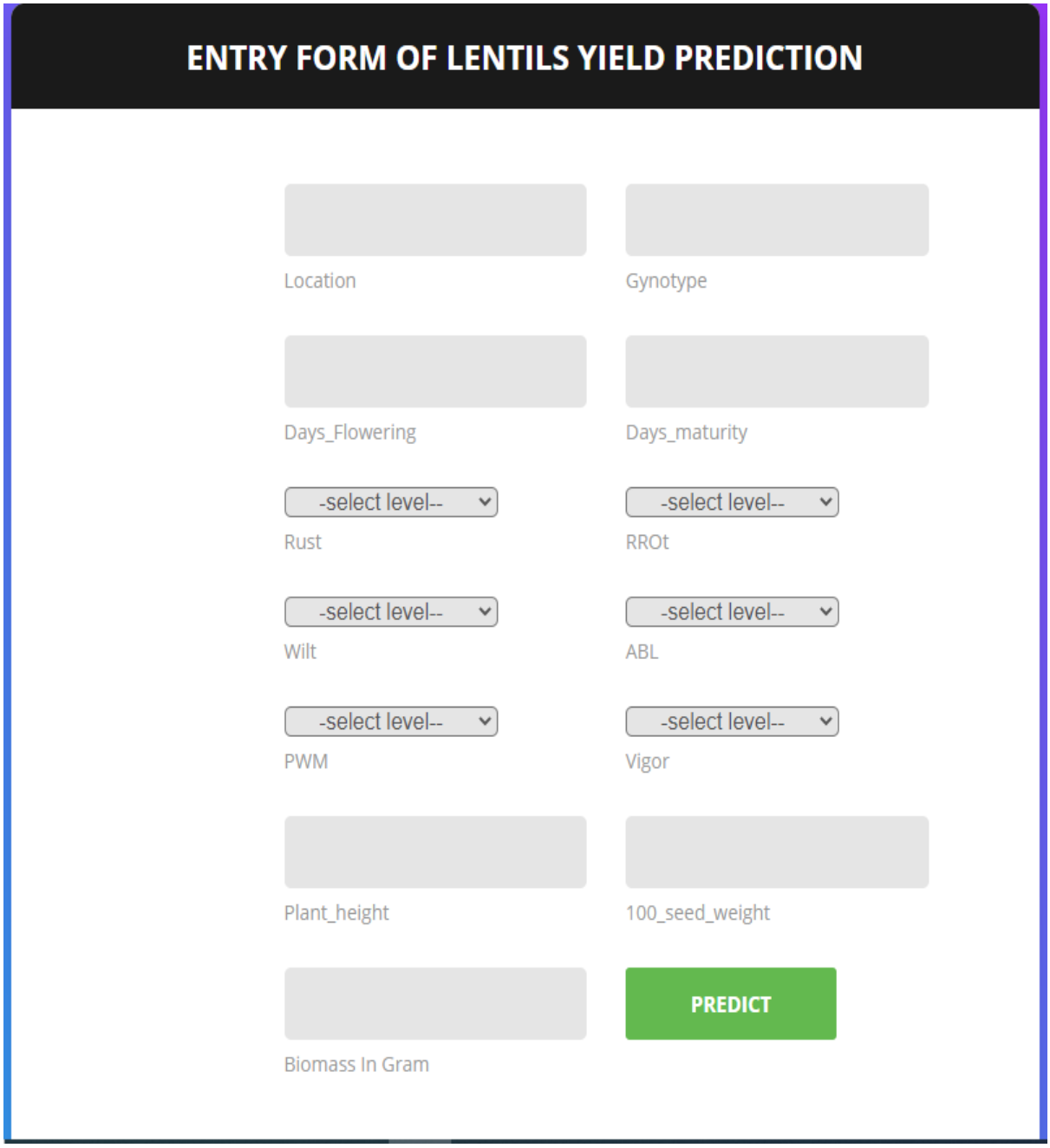
Appendices

Appendix A: GUI

Appendix A.1: GUI for inserting inputs from the keyboard for lentil yield prediction

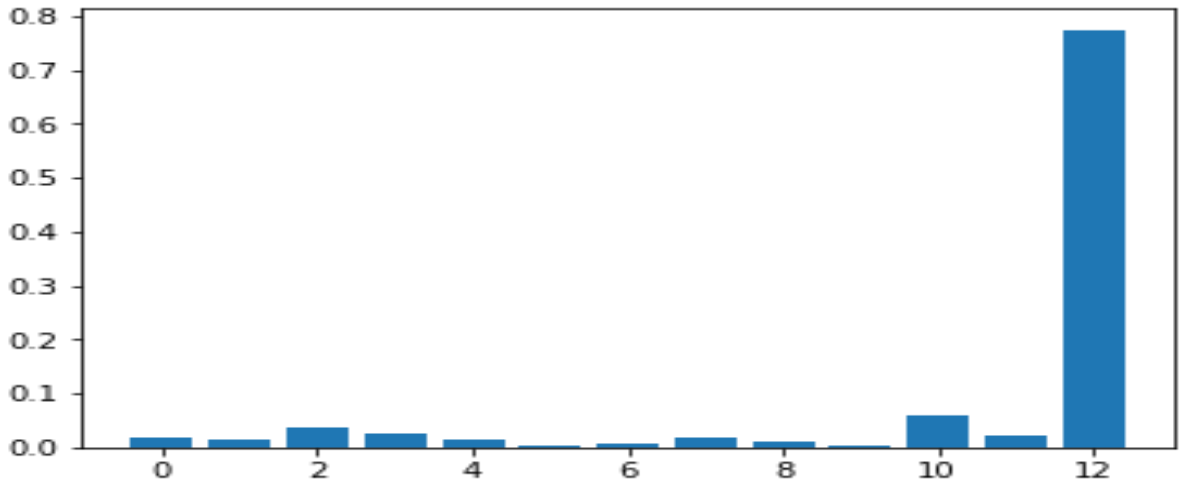


Appendix A.2: GUI to insert new input to classify to clusters

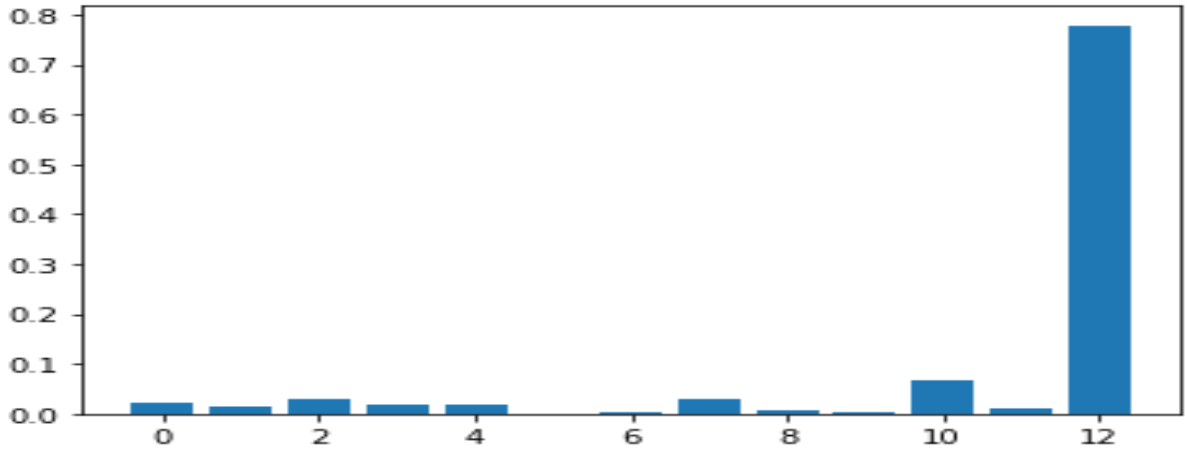


Appendix B: Selected features using feature selection

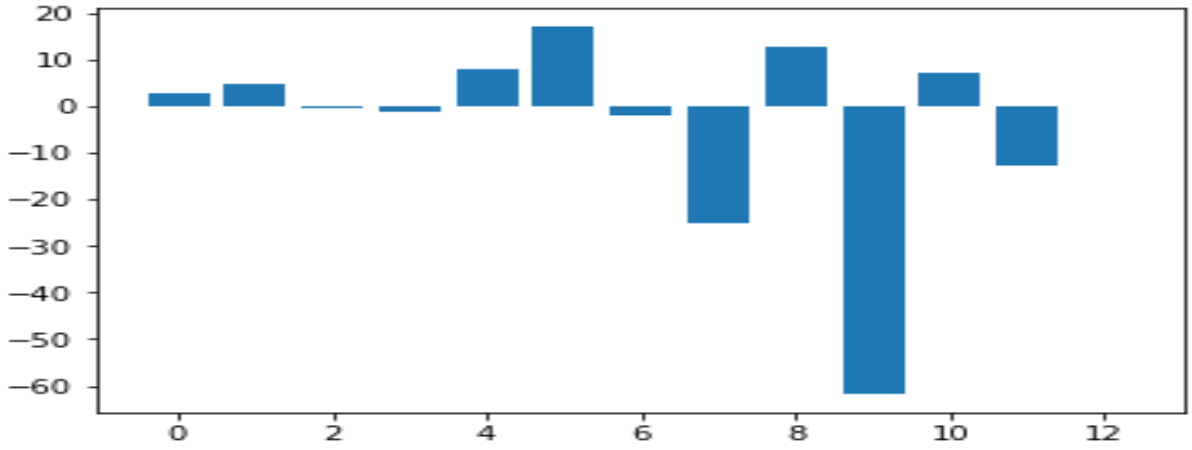
Appendix B.1: Feature importance with regression models



Random forest regression



Gradient boosting regression



Multiple linear regression

Appendix B.2: Features selected in the prediction of lentil yield

Models	UFS	RFE
Random forest regressor	DF, DM, RRot, Wilt, ABL, PWM, Vigor, PLH, HSW, BMYP	Locn, DF, DM, ABL, PLH, BMYP
Gradient boosting regressor	DF, DM, Rust, RRot, Wilt, Vigor, PLH, HSW,	Locn, DF, DM, PLH, HSW, BMYP
Multiple linear regression	Locn, year, DF, DM, Rust, RRot, Wilt, ABL, PWM, Vigor, PLH, HSW, BMYP	Locn, ABL, PWM, Vigor, PLH, HSW

Appendix C: Classification Reports

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	506
1.0	0.99	1.00	0.99	506
2.0	0.99	0.99	0.99	506
3.0	1.00	0.99	0.99	506
4.0	1.00	1.00	1.00	506
5.0	0.99	0.99	0.99	506
6.0	1.00	0.99	0.99	506
accuracy			0.99	3542
macro avg	0.99	0.99	0.99	3542
weighted avg	0.99	0.99	0.99	3542

Support Vector Machine

	precision	recall	f1-score	support
0.0	1.00	0.99	0.99	506
1.0	1.00	1.00	1.00	506
2.0	0.99	0.98	0.98	506
3.0	0.99	1.00	0.99	506
4.0	1.00	1.00	1.00	506
5.0	0.99	0.99	0.99	506
6.0	1.00	0.99	0.99	506
accuracy			0.99	3542
macro avg	0.99	0.99	0.99	3542
weighted avg	0.99	0.99	0.99	3542

Random Forest

	precision	recall	f1-score	support
0.0	0.98	0.99	0.98	506
1.0	0.94	1.00	0.96	506
2.0	0.96	0.90	0.93	506
3.0	0.98	0.95	0.96	506
4.0	0.98	0.99	0.99	506
5.0	0.97	0.94	0.96	506
6.0	0.94	0.98	0.96	506
accuracy			0.96	3542
macro avg	0.96	0.96	0.96	3542
weighted avg	0.96	0.96	0.96	3542

Naïve Bayes

	precision	recall	f1-score	support
0.0	0.98	0.98	0.98	506
1.0	0.98	0.98	0.98	506
2.0	0.97	0.97	0.97	506
3.0	0.98	0.98	0.98	506
4.0	0.99	1.00	1.00	506
5.0	0.98	0.98	0.98	506
6.0	0.99	0.99	0.99	506
accuracy			0.98	3542
macro avg	0.98	0.98	0.98	3542
weighted avg	0.98	0.98	0.98	3542

Decision Tree

Appendix D: Rating scales of disease status

The scales are tentative and have been developed for easy scoring. Nine points (1-9) scales are being adopted. The interpretation of the scales is given as follows:

1- Highly resistant

3– Resistant

5– Moderately resistant

7– Moderately susceptible

9 - Highly susceptible

Appendix E: Sample code

Appendix E.1: Sample code to integrate random forest regression with flask server

```
from flask import Flask,render_template,request
import pickle
import numpy as np
app=Flask(__name__)
model = pickle.load(open('modelrf.pkl','rb'))
@app.route("/")
def home():
    return render_template("forprediction.html")
@app.route("/predict",methods=['POST', 'GET'])
def predict():
    int_features = [int(x) for x in request.form.values()]
    out=[np.array(int_features)]
    prediction=model.predict(out)
    we=format(prediction)
    return render_template('forprediction.html', prediction_text='The yield of lentils predicted is= {}'.format(we))
if __name__=="__main__":
    app.run(debug=true)
```

Appendix E.2: Sample code to integrate SVM with flask server for classification

```
from flask import Flask,render_template,request
import pickle
import numpy as np
app=Flask(__name__)
model = pickle.load(open('SVMmodel.pkl','rb'))
@app.route("/")
def home():
    return render_template("index.html")
@app.route("/predict",methods=['POST', 'GET'])
def predict():
    features = [float(x) for x in request.form.values()]
    out=[np.array(features)]
    prediction=model.predict(out)
    we=format(prediction)
      if(prediction==0):
       return render_template('index.html', prediction_text='clustered to moderatly good productive cluster of genotype is= {}'.format(we))
    elif(prediction==1):
       return render_template('index.html', prediction_text='clustered to moderatly productive cluster of genotype is= {}'.format(we))
    elif(prediction==2):
       return render_template('index.html', prediction_text='clustered to high productive cluster of genotype is= {}'.format(we))
```

```
    elif(prediction==3):
        return render_template('index.html', prediction_text='clustered to productive cluster of genotype is= {}'.format(we))
    elif(prediction==4):
        return render_template('index.html', prediction_text='clustered to very low productive cluster of genotype is= {}'.format(we))
    elif(prediction==5):
        return render_template('index.html', prediction_text='clustered to low productive cluster of genotype is= {}'.format(we))
    elif(prediction==6):
        return render_template('index.html', prediction_text='clustered to highly productive cluster of genotype is= {}'.format(we))
    #return render_template('index.html', prediction_text='The cluster of genotype is= {}'.format(we))
if __name__=="__main__":
    app.run(host="10.240.69.23",port='8080')
```