Addis Ababa University

School of Graduate Studies

College of Natural Science

Department of Computer Science

Teff Scarcity Prediction Model for Ethiopian Context Using Multiple Linear Regression

Taye Mohammed Kemal

A Thesis Submitted to Addis Ababa University, Department of Computer Science in Partial fulfillment of the Requirements for the Degree of Master of Science in Software Engineering

Addis Ababa, Ethiopia

July 2023

Addis Ababa University
School of Graduate Studies
College of Natural Sciences
Department of Computer Science

Taye Mohammed Kemal

Advisor: Ayalew Belay (PhD)

This is to certify that this thesis prepared by Taye Mohammed Kemal, titled: *Teff Scarcity Prediction Model for Ethiopian Context Using Multiple Linear Regression* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Software Engineering complies with the regulations of the University and meets the accepted standards with respect ot originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor:	_____	_____	_____
Examiner:	_____	_____	_____
Examiner:	_____	_____	_____

Abstract

Scarcity prediction is vital in avoiding economic and political disturbances from food resource scarcity. In addition, it allows for well-managing resources for maintaining and improving livelihood, population's quality of life, use of budget, reducing cost, boosting productivity, seeing potential resource conflicts in the early stage for more responsive mitigation, reducing wastage of resources, and contributing sustainable and reasonable growth. Hence, the purpose of this paper is to use multiple linear regression models to predict the scarcity of Teff resources in the context of Ethiopia. However, predicting Teff's scarcity based on factors like resource consumption, population growth, productivity, and other factors is a significant problem to address. Thus, in this thesis work, we present a system that predicts Teff scarcity.

We use a multilinear regression approach to design the system. The teff scarcity prediction model consists of three components: the preprocessing components, the train-validate-test component, and the prediction component. The preprocessing part consists of data cleansing and data transformation. The train-validate-test consists of the training, validating, and testing data partition. The prediction part removes insignificant attributes and trains, validates, and tests the designed models. The preprocess component receives the raw dataset and performs data cleansing and transformation. The train-validate-test component performs partitioning into a train, validate, and test dataset. The prediction component predicts teff scarcity with the partitioned data and evaluates the model with the result.

The experiment result shows scarcity prediction statistically based on mean absolute error, root mean squared error, and R squared values. Hence, the experiment produced a mean absolute error of 7%, a root mean squared error of 6.43%, and an R squared of 97.07%. The designed model can predict Teff scarcity with an accuracy of 97.07%.

Keywords: crop production, population growth, consumption, scarcity prediction, Multiple linear regression, Root mean square root error, R squared value

Acknowledgment

First, next to God, I would like to express my deepest gratitude to my advisor Ayalew Belay (PhD) for his continuous support and professional guidance throughout this thesis work. I would like to thank Yaregal Assabie (PhD), whose comment and valuable input early at the proposal stage have meaningfully shaped the course of this thesis detailed suggestions on my entire proposal content and valuable discussion after that have been a great input that helped me well to shape my proposal. I would like to extend my thanks to the staff of CSA for their cooperation. My deepest gratitude goes to my family who always motivated me to work hard. I would also like to thank my friends for their support and companionship.

Table of Contents

List of Figures

List of Algorithms

Acronyms and Abbreviations

ABR	Average Birth Rate
ADR	Average Death Rate
ANN	Artificial Neural Network
Consup	Consumption
CSA	Central Statistical Agency
DT	Decision Tree
KNN	K-nearest Neighbor
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
NN	Neural Network
OLS	Ordinary Linear Sequences
Prod	Production
Pop	Population
RF	Rain forest
RMSE	Root Mean Square Error
R2	R squared
YieldPH	Yield Per Hector

Chapter 1: Introduction

1.1 Background

Food is the most important aspect of human life. Without food, human will not survive. Securing food resources for a country is a challenging and important task. Storing enough food for future consumption to address scarcity is a vital goal for all countries, especially developing countries. The main compulsion of the government of developing countries is securing enough food resources for the long term, particularly at the time of natural as well as human-made disasters [1].

It is necessary to organize the country's statistical data of food resources to ensure planning and monitoring of social and economic development. Hence, having organized and complete statistical information about population, food resources, consumption, and others helps to make a good plan regarding resource requirements and to plan for food resource security [2].

The world population is rapidly growing and will grow over 2050. With population growth, people will have to face a shortage of resources like minerals, energy, food, and other basic resources. Shortage of resources will seriously affect the improvement of people's life. Hence, strict control and management of basic resource requirements is an essential and urgent task for any country [3]. Market demand for food and other basic resources continues growing along with the population size. The population needs different resources to survive like food, water, energy, and other basic resources. However, the resource growth rate should be grown parallel with the population growth rate and the resource consumption. Predicting essential resource requirements; provides a good plan and management strategy for utilizing the resource adequately. Hence, the government used the information to produce an excellent technique to reduce the chaos created by the population because of the primary resource shortage [4].

Farmers rely on estimation and experience when deciding which crop types should be grown to maximize their profit. They also guess which crop is consumed more and scarce from the market. Though the farmer plant the scared crop from the market based on thier experience to benefit themeselves. Guesswork is a big challenge for farmers to estimate future agricultural production to grow and maximize their profits. If there is a reasonable prediction about the

resource scarcity of a crop, the farmer will have better information to select the short crop. They will depend on scientific results rather than guesswork, estimation, and experience [4, 5].

A meaningful relationship between population growth and resource shortage brings scarcity. Rapid population growth in Ethiopia puts pressure on resources. Whenever the population size increases rapidly, the crop consumption rate also increases. In Ethiopia, to our knowledge population growth rate, food consumption growth rate, product growth, and other data have not been analyzed an automated manner. No structured model exists for predicting food resource scarcity. As a result, different actors like the government, farmers, and others take desperate action on their function based on something other than scientific data. For example, farmers plant crops based on their experience, and they may face minimum benefits; government takes desperate action because of a lack of scientific data to take appropriate action, which might affect the country's economy. The Ethiopian government has start working with international partners to enhance productivity by collaborating with private sectors. Hence, the enhancement depends on experience data to tackle the current and future food resource scarcity [6].

1.2 Motivation

Internet service has increased its importance as a medium for communication. Even though predicting resource scarcity is one of the complex and challenging tasks, it is the most important service in controlling and preventing basic resource scarcity in the market.

Therefore, the motive for doing this research evolves under the following compelling factors:

- Most of the time, food items are exposed to scarce in the market.
- Consumers are exposed to unfair prices to purchase basic resources due to scarcity.
- Lack of better-predicted resource scarcity information for policymakers, farmers, and government so it is difficult to make the right decision.
- Lack of information to store different basic resources for future use to tackle the shortage that will happen in the future.
- Existence of an unstable market

1.3 Statement of the Problem

Currently, resource scarcity prediction is conducted traditionally in Ethiopia; Thus, scientifically analyzed data needs to be on the table for predicting future resource scarcity. Rapid population growth creates unbalanced demand and supply of different crops, which require farmers to produce more from the same agricultural land to increase supply and reduce food security problems. Accordingly, preparing for the solution before it happens depends on opinion and experience. Consumer and governments do their role without appropriate information on current and future food resource scarcity. The Ethiopian population was estimated at 117,876,227 in 2021, a 2.53% increase from 2020. Thus, this indicates that the population growth rate is fast, and the government needs to adjust the required resources increment accordingly. To reduce resource scarcity, there must be analyzed data, which helps the government to have an efficient plan [7].

With population growth, people will face a shortage of resources like minerals, energy, food, and other resources. Thus, this will seriously affect improvements in population life. Thus, strictly controlling and managing resource requirements is an essential and urgent task for any country [3]. Understanding and predicting food consumption and consumer in developing countries like Ethiopia is difficult. Without accurately predicting the scarcity, it misles the food producers and governments for resource allocation. Resource scarcity prediction plays a vital role in managing resources. It helps to decide what plane needs to accommodate the scarce resource and reduce the problems. Predicting the scarceness of primary resources would help the policymakers and other responsible entities for taking appropriate measures for marketing and storing scarce resources [8]. Agriculture planning also needs information on identifying the deficit corp for planting to reduce scarcity. Hence, the quantity of the crop is also required to utilize limited resources like Land. Because crop selection is a critical issue for agriculture planning, the crop that will be scarce in the future must be predicted using parameters such as production rate, population growth rate, consumption rate, and government policies. With the development of the economy, people are paying more attention to predicting resources in the future to make the right decision [9].

However, how to predict Teff resource scarcity based on factors like resource consumption, population growth, productivity, and other factors is a critical problem to address. Predicting

resource scarcity effectively and accurately is critical to maintaining a human life because it helps to store enough Teff resources to serve the people when an absence happens due to artificial and natural disasters. Predicting the scarcity of resources provides scientific information for farmers to plan to crop the scarce resource to maximize their profit. Thus, it will address the scarcity problem and prepare how to secure deficit crops [10].

Most of the research focuses on predicting a certain crop production quantity to maximize the crop quantity and profit. Using analyzed information like climate, soil type by measuring ph, the amount of fertilizer used to maximize the crop quantity, image processing of the land, and other parameters to achieve their goal. And the research is done specifically to the country because the parameters considered to predict crop production in the future are different from country to country. The reviewed research focuses on maximizing profit or crop production which does not consider the amount of crop required by the community. Targeting only maximizing crop production could create a surplus leading to surplus and inefficient use of Land resources.

To our knowledge, all research focused on predicting a particular crop production by considering different parameters. Predicting Teff resource scarcity to prepare to tackle the absence of resources using different mechanisms based on analyzed data is one of the essential scenarios to balance the market, economy, and people's life. Due to the current political and economic situation, Ethiopia needs more food resources to feed the people in the future. However, the country needs to produce more crops for the future consumption. Hence, the government collaborates working with partners for the enhancement of productivity. The government imports different crops to reduce the scarcity of resources. Therefore, no model exists for predicting the scarcity of a particular crop to help the government prepare for making appropriate decisions. Knowing scientifically how much a particular crop will be scarce in the future will help the government to adjust and allocate the appropriate budget to reduce the scarcity. Hence, scarcity prediction is vital in avoiding economic and political disturbances from food resource scarcity. In addition, it contributes management and control of market balance, economic growth, price control, and other advantages to the country [6].

1.4 Objectives

General Objectives

The general objective of this study is to develop a model for predicting Teff food resource scarcity using multiple linear regression in the Ethiopian context

Specific Objectives

To meet the general objective, the specific activities include the following are going to carried out to achieve the general objective:

- Reviewing literature and related work to organize the required data for the study
- Collect the dataset for training the scarcity prediction model.
- Design a model for scarcity prediction
- Develop a prototype for the proposed model.
- Evaluate the performance of the prototype.

1.5 Methodology

Research Approach

A Design Science Research methodology followed to achieve the objective set in this research. According to, [11, 12] design science is a research methodology focused on outcome-based information technology that offers guidelines for evaluation and iteration which is fundamentally a problem-solving paradigm. Design science research methodology is most familiar in Engineering and Computer Science disciplines focusing on the development and performance of designed artifacts with the explicit intention of improving the fundamental performance of the artifact.

Literature Review

We intensively reviewed different literature to explore related works on algorithms for predictive analysis, processes, and methods applicable to predictive models. Based on the review, appropriate tools and algorithms would be selected, applied, and developed.

Data Collection

Data will be composed of a different document review and governmental sector. Population growth rate, Teff item consumption rate, Teff product quantity growth rate, and other relevant data collected through document review and other appropriate mechanisms.

Tools, Environment, and Programming Methodology

To develop a prototype, python programming languages, and other appropriate tools were applied to handle the evaluation method.

1.6 Scope and Limitation

This research focused on developing a prediction model for scarcity prediction using a unified dataset that combines population growth, resource consumption, quantity of product, and another relevant dataset. However, this prediction can only cover some resources, and we pay attention to a single crop for our work. Hence, we select the Teff crop as a food resource since its data is available more than other food products.

1.7 Significance

Different actors like farmers, consumers, the government, and other parties would use this thesis work. The research project provides a service for different actors like:

- The consumer uses the system to identify scarce resources and manages their budget with their needs.
- The government uses the research product work for controlling the market price by imposing different appropriate rules and creating the appropriate mechanisms for reducing resource scarcity.
- The community will be benefit from this research by getting information about product scarcity.
- The industry as well as farmers also uses research work to produce scarce resources so that provide the product in the market and produce an appropriate plan to produce the required resource by the market with an effective budget.

- In addition, different parties like students, researchers, policymakers, etc. could use the research project result for a different purpose for making the genuine result for their works.

1.8 Organization of the Thesis

The rest of the thesis report is organized as follows: Chapter Two reviews literature relevant to the proposed approach, and Chapter Three presents the related work. Chapter Four presents the proposed Model and its components. Chapter Five presents the Experimental results of the proposed system. In Chapter Six, the conclusion and future work.

Chapter 2: Literature Review

2.1 Background

When the resource is well-managed, it generates different benefits that provide the basis for maintaining and improving livelihood, improving the population's quality of life, improving the use of budget, reducing cost, boosting productivity, seeing potential resource conflicts early on for more responsive mitigation, reduce wastage of resources, and contribute to sustainable and reasonable growth. Well-organized information helps to make good decisions for managing and controlling the resources [13].

There will be wide-reaching policy implications as any response to the possible changes in different essential resources. Lack of peace and stability, population growth, and the amount of land used for agricultural production will impact food security for the world's population. Improvements in farm productivity and the sustainability of the farming sector can come about by matching agricultural production systems to the specific environment and growing conditions for the farming system. Hence, this will allow farmers to make the best decisions on the types of varieties they can grow for their farming system. Agricultural research must ensure food security for its population by providing agricultural systems remain viable and productive. Therefore, achieving better adaptive environments, new agronomic practices, managing resources properly, predicting scarcity, and improving the efficiency of the agricultural product value chain can ensure food security. Agricultural infrastructure and resources provided by governments are essential for assisting food production to feed the nation. Understanding Teff production processes and controlling to reduce scarcity value can be added as critical information to the government for resource allocation. Monitoring crop production scarcity and yield prediction is very important for the economic development of a nation. The world's economies are influenced directly and indirectly by the prediction of crops. It plays a significant role in monitoring resource utilization, managing crop deficits, and ensuring food security. The government may become involved with providing an appropriate incentive to Teff producers for producing particular crops to make deficit crops available to consumers if there is scientifically analyzed data. Understanding and predicting food consumption and consumer in developing countries like Ethiopia is complex. In recent

years, this range and the increased differentiation of food products in developing countries have resulted in misleading signals back to food producers and governments for resource allocation. The government's investment in building facilities may be less desirable for several years once built the facility [14].

Goods and services must be in a limited supply, which limits options and demands of choices according to economic resources. Because we cannot have all the resources we require for our consumption, we must decide what help we will have, what goods we must forgo, and what resource is exposed to be scarce. Hence, society turns its face to the opportunity of getting the next best resources which could have been affordable based on their budget and appropriate data for making the right decision whenever the first best option in a scarce situation. The data presented for making a decision should be balanced so that the decision made by the stakeholders is as correct as possible. This action helps to overcome the scarcity of resources by managing the existing resource with appropriate mechanisms and providing reliable information to the community for utilizing their budgets wisely [15, 16].

The world, especially developing countries, faces multiple challenges to food security. Ethiopia is an overpopulated country in Africa next to Nigeria and is highly vulnerable to severe food resource scarcity. This challenge includes malnutrition and overconsumption, rising food prices, population growth, inefficient production practices, and peace and stability, all leading to scarcity. One of the reasons that the challenge happens is the need for scientifically analyzed information for tackling food scarcity and making a better decision to provide a solution before the problem occurs. Managing the resources helps to put a good decision for resolving the problem. The analyzed information should be on the table for managing resources. The population growth, food price, and inefficient production might be handled by responsible organs if there would be an accurate prediction about the future as well as the current requirement of the resource using scientifically analyzed data which is a prediction of future events based on the present and past data occurrences [17, 18].

Efficient resource providing is one challenging problem since it needs accurate information about the requirement of the resources. Resource provisioning is complex and dynamic, which requires relevant information. Having a correct set of predicted information helps provide adequate resources which satisfy the expected results. We need a prediction model that can

accurately predict the required information so that we will have the correct information to make a decision [19].

Historical data, which is past data as well as current data used to create a mathematical tool to predict future, unseen cases to make a better decisions pretty much in everything in our life [20]. Predictive analytics is a paramount technique to anticipate events before happening. It is discovering exciting and meaningful patterns using current and past data occurrences. It uncovers the relationships and patterns within the data used to predict events in the future [21].

It is building a model for the dependent variable from one or more independent variables. The value of the dependent variable will predict from the known values of other variables. In predictive modeling, extracting the patterns for treating secret information from the available data is essential to build the model to predict future values [22].

2.2 Description of Scarcity Prediction

The objective of designing a scarcity prediction model is to acquire genuine information and inputs for scarcity prediction information; the country cannot afford to wait until the actual time of scarcity happens. Thus, in Ethiopia, assessing total food production, population size, and consumption and providing accurate and timely early warning signals to the emerging difficulties due to drought and other natural disasters are and remain to be the primary objectives of the efforts to be made by the government and the concerned stakeholders. Towards this end, many factors need to taken into consideration. Hence, compiling reliable, accurate, and timely quantitative crop production, population growth, consumption, and scarcity prediction estimates for communities should get prior consideration. Hence, the government and the concerned stakeholders could use the prediction information to plan and take all the necessary and appropriate measures in administering exports or imports, management of stocks and distribution of food to deficit areas, planning and controlling population growth, and planting deficit crop, regulation of price control at surplus or deficit production, managing financial capital accordingly, among others [23, 24].

The other objectives are to produce basic quantitative information on expected production, population growth, and scarcity consumption at a specified time. This information can be used

as an early indicator to warn policymakers and planners about the emerging difficulties that result from surplus or deficit crop production in the predicted years. Consequently, timely predicted information made on the particular resource of the expected quantity in a specified year is a primary input for policy preparation and implementation of timely measures, such as administering exports or imports, management of stocks and distribution of food to deficit areas, plan controlling population growth, planting deficit crop and regulation of prices at the time of surplus or deficits [25].

2.3 Scarcity Based on Demand and Supply

The macroeconomic analysis examines a nation's aggregate output and income, its competitive and comparative advantages, the productivity of its labor force, its price level and inflation rate, and the actions of its national government and central bank. The collective productivity of an economy is the produced values of all consumption in a specified period. To analyze to predict the aggregate output and revenue so that the economy will be stable, all products that the country should organize properly [26].

A theory must exist which can explain how the government made a decision. How does the government set prices, decide the level and style of output, the level of research and development expenditures, the financial policies, and the investment decisions for managing and addressing the deficit resources so that the right decision will make [27].

Consumers can decide what goods to buy based on their income and budget. Business firms can decide what products to plant and how to produce them to increase profit. Government entities decide what public services and goods to provide, how to finance them, and what rules and regulations should be applied to create a smooth environment for consumer and business firms. Analyzed and accurately predicted data should be available for all consumers, business firms, and governments to achieve what they want and reach a good decision [16]. With the rapid growth of economic society, a tremendous amount of resource demand leads to resource scarcity. The gap between resource supply and demand is growing, affecting the community's quality of life and livelihood. Consumption is rising continuously, which causes the serious problem of resource scarcity. To overcome resource scarcity, the consumption rate of a resource should be less than the resource productivity rate. The imbalance of productivity and

consumption rate becoming severe due to the increasing population, which creates resource scarcity, so the government should introduce a mechanism to control resource scarcity [28].

The demand for food crops escalates as the population increases, and the lifestyle and awareness of people affect the food demand as well. The population growth and lifestyle of people create a scarcity of food resources. Prediction of scarcity provides information for the government and community to facilitate decision-making to choose the best mechanisms for resolving the shortage of resources [29].

Regression: in contrast to classification, regression is concerned with predicting the numeric value for the class label. Regression models use input data features and continuous numeric output values to learn specific relationships between the inputs and consistent outputs. With this knowledge, the regression model can predict output responses for new unseen data instances similar to classification but with continuous numeric outputs. Linear regression and multilinear regression are some of the supervised regression algorithms [30].

Population, crop productivity, and consumption information are necessary for development planning. Analyzed datasets of population, crop productivity, and consumption help to predict the scarcity of a particular product which creates the opportunity to develop a plan for providing a solution to find out the way of overcoming the deficit of resources [31].

2.4 Predictive Modelling

With the development of the economy, people are paying more attention to predicting resources in the future to make the right decision. The growing variety of predicting methods are qualitative, quantitative, and mixed. [9]. The standard predictive methods for resources contain regression analysis, economic quality models, and energy consumption per capita [28].

Predictive Modeling is building a model by identifying the patterns for the dependent variable from one or more independent variables. The value of the dependent variable will predict from the known values of other variables. In predictive modeling, the pattern emanates from the data for predicting future values. Classification and regression are two forms of data analysis that can predict future data [22].

2.5 Machine Learning

Computer software, which has gained experience through learning from previous data using machine learning techniques, can predict new situations that can emerge in the future.

Machine Learning is a technique of training a machine with data to ensure that the designed machine learns by itself to make a decision. It is a subdivision of artificial intelligence area that assists machines in forecasting unknown events in the future based on experience and producing predictions. It may also be a predictive learning machine for making decisions based on prior information and experiences. Hence, the machine can learn and adapts through the experience without being explicitly commanded for a particular operation performed [32, 33]. It is an effective method, to recognize unknown objects through learning from prior information or experience. ML mainly studies extracting rules or patterns, which are rugged to be obtained by theoretical analysis from observed data, then how to use the extracted rules to predict future or unknown data [34].

Machine learning research focuses on automatically extracting information from data by computational and mathematical methods. According to [35] machine learning theory is related to statistical inference, wherein a model can learn to improve its performance based on prior experience. Machine Learning includes many advanced statistical methods for handling regression and classification tasks. ML models include artificial neural networks, linear regression, and support vector machines [36].

An ML system is trained rather than explicitly programmed and needs to learn from the historical data to gain experience, optimize for better computations, and generalize to provide accurate results. ML algorithm applied for classification, regression, clustering, or dimensionality reduction tasks of large datasets. Predicting resources is becoming very important to act proactively and implement preventive measures ahead of time [31]; thus, ML helps to recognize the complex patterns in the given large datasets by using a learning mechanism to make decisions or predictions when new data instances are coming. ML algorithms can be classified based on the algorithm's outcome and types of input fed during training as supervised, unsupervised, or semi-supervised learning [37].

ML allows computers to learn to decide data without being explicitly programmed. The machine is clever to use past or previously procured data to study and understand the patterns and analyze them further without being explicitly instructed to. This feature is beneficial in predicting future values like scarcity of resources [32].

2.5.1 Supervised Machine Learning

Supervised machine learning, also known as predictive learning, as it predicts the class of unknown matters based on previous information on similar items. The primary motivation behind this type of learning is to study the behavior or characteristics of a particular object by extracting information that has been gathered and provided in the past. A machine requires the primary data provided to it to accomplish the expected tasks or to understand the patterns or behavior of the object. The essential input is extracted from history or experience given to the machine in the form of training data which is the past information or data of a specific task [32].

Supervised learning is a process of driving function with the help of labeled data samples called training data. Hence, the algorithm analyzes the training data you fed to the machine, including the desired solution called label to extract function, for input to output mapping of a new instance. This newly learned information can then be used to predict an output, Y, for any further input data sample that says X, which was previously mysterious or invisible during the model training process. A supervised learning model performs two significant tasks to develop a model classification and regression [38].

Classification: It is a technique of categorizing a set of data into classes. It is concerned with predicting or determining which category the dependent belongs to one or more independent variables for what the model has learned in the training phase. Dependent variables, also known as classes or class labels, are categorical, which means they are unordered and discrete values. Standard algorithms for performing classification include LR, NNs, support vector machines, ensembles like RFs and gradient boosting, k-nearest neighbors, DTs, Naïve Bayes, and others [39].

Regression: Regression models use input data features and continuous numeric output values to predict constant values for class labels to learn specific relationships between the inputs

and corresponding output. It is a technique that helps find the correlation between variables and allows us to predict the continuous output variable based on one or more input variables. With this knowledge, the regression model can predict output responses for new unseen data instances similar to classification but with continuous numeric outputs. Linear regression and multilinear regression are some of the supervised regression algorithms [39].

2.5.2 Unsupervised Machine Learning

Unsupervised learning deals with how systems can learn to represent particular input vectors in such a way that it reflects the statistical structure or pattern of a collection of inputs [39]. In an unsupervised learning approach, there is no need for human intervention or labeled documents at any point in the whole process. Here the task is to group unsorted information according to patterns, similarities, and differences without prior training data. Clustering is one of the unsupervised learning methods that help cluster or group data points into different groups or categories without the availability of any output label in the input/training [40, 41].

2.5.3 Semi-supervised Machine Learning

A semi-supervised learning method is another type of ML that combines supervised and unsupervised learning. When a small number of labeled data is identified, a semi-supervised learning method is suitable for a particular application [37]. In semi-supervised learning, the clustering technique uses some knowledge of supervised data in the form of class labels is used along with the unlabeled data. It generates a function mapping from inputs of labeled and unlabeled data. This learning mechanism is to classify some of the unlabeled data using the tagged information set. The quantity of unlabeled data should be higher than the number of labeled data [42].

2.6 Popular Machine Learning Algorithms

This section presents some standard ML classification algorithms with their strengths and weaknesses.

2.6.1 K-Nearest Neighbor

The K-Nearest Neighbor algorithm is a method for classifying objects based on learning data closest to the thing. KNN algorithm works by determining the K-Nearest Neighbor based on learning data most proximate to the object to produce a better result [43, 44]. KNN algorithm is used for classification most of the time. This algorithm aims to classify new objects based on the attributes and training data. [45]. KNN is merely a lookup table to predict the target value of mysterious events not observed during training. The character K refers to the number of objects neighboring the data point that needs to make a better prediction. The mathematics behind the KNN algorithm concerns how computed the distance from an observed object to neighbors. The more neighbors included in the forecast, the smoother the predictions. KNN is a numerical algorithm that requires all the input data to be numeric [46].

2.6.2 Decision Trees

A DT is a supervised ML algorithm that can develop a learning model to predict a class or value of a target variable by learning the decision rule of training data. This algorithm can solve classification and regression problems [47]. Decision tree algorithms can handle nominal and continuous inputs and outputs [21]. One of the vital problems of DT is using records with unknown values from training and testing data [48].

DT is a reverse tree-like structure that allows for straightforward data interpretation and analysis in decision trees, which gives advantages over other classification methods. Decision trees and different classification algorithms of supervised machine learning often seek to predict the value of the unknown attribute based on the importance of known facts or prior evidence. Each interior tree node in a decision tree algorithm is associated with an input attribute, and the leaf node is an output attribute. Every leaf node in a tree is considered a value of the output attribute inferred from the input attributes and the path connecting the root and the leaf [49]. In the decision Tree, each node represents a test of an attribute, and the leaf node provides classification data [21]. DTs are nonlinear predictors where the decision boundary between output variable classes is nonlinear. The extent of the nonlinearities depends on the number of splits in the tree because each partition, on its own, is only a piecewise constant separation of the classes. The Root Node represents the entire data in which

the modeling process begins. The Root Node divides into two or more sub-nodes which is splitting. The node splitting continues until the criteria meet, and the sub-node where splitting happens is called a decision node. The node does not divide any further once the splitting criteria are satisfied. Once the requirements are satisfied, the splitting process stopThe last node, which is not splitting split after meeting the needs, is called a Leaf or Terminal node. The sub-nodes can remove by using the technique or process of Pruning [50].

2.6.3 Artificial Neural Network

Researchers from many scientific disciplines modeled artificial neural networks to resolve a variety of problems in pattern recognition, forecast, optimization, associative memory, and control by learning the relationship between input and output that are non-linear and complex. It is a machine-learning algorithm modeled by mimicking the human brain [51].

An artificial neural network is an information processing system where several processors are connected, inspired by how biological nervous systems such as the brain work. It offers a mathematical model that tries to mimic the human brain. An artificial neural network contains extensive parallel processing and self-learning machines, just like a brain, which is made possible by the neural network in the brain. It is a collection of processing elements interconnected with each other, which can produce some results after receiving input [52, 53].

Artificial neural networks are the essential processing elements of neural networks. A neuron is a unique biological cell that processes information from one neuron to another neuron with the help of some electrical and chemical change. It comprises a cell body and two types of out-reaching tree-like branches: the axon and the dendrites. The neurons in our nervous system can learn from historical data; similarly, the ANN can learn from the data and respond to predictions or classifications. A nonlinear statistical model which displays a complex relationship between the inputs and outputs to discover a new pattern [51].

2.6.4 Logistic Regression

The logistic regression algorithm is also called the logistic model or logit regression. It is considered a linear model that measures the relationship between the target variable and other variables by estimating the probability using a logistic function that calculates the likelihood

of an event occurring based on a given dataset of independent variables. Logistic regression is an algorithm that focuses mainly on classification and is used to predict a binary outcome based on independent variables. Since the result is the probability, only two possible scenarios occur (1 or actual), or it does not appear (0 or false). The logistic regression model can also find the probability of occurrence of categorical output by fitting the features in the logistic curve. The Logistic Regression model can be swapped by the simpler Linear Regression model when the output variable is continuous [54, 55, 56].

2.6.5 Linear Regression

Regression analysis is a statistical technique for exploring and modeling the relationship between dependent and independent variables [57]. The regression technique used for forecasting, time series, or continuous modeling and finding the causal effect relationship between the variables, which means predicting the outcome of the dependent variables from the independent variables [58].

Linear regression is a predictive model that uses regression analysis in mathematical statistics to establish a relationship between variables using a best-fit straight line known as a regression line. It creates a predictive model to make understandable the relationship between the dependent and independent variables, which is the most common predictive model. When there is only one independent variable, it is called Simple Linear Regression, and when the independent variable is more than one, it is called Multiple Linear Regression. Hence, linear regression divides into simple and multiple linear regression based on the independent variables in determining the dependent variable. Simple linear regression is a model that estimates the relationship between one independent variable and one dependent variable using a straight line. In contrast, a multiple linear regression model is an extension of a simple linear regression model that uses several independent variables to predict the outcome of a dependent variable [55].

It is represented by a formula:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \varepsilon}$$

Where **β0** intercept, β_1 is the slope of the line, and **ε** is the error term. The formula can use to predict the value of the dependent variable, which in this case is **Y,** based on the given independent variableX_1 [59, 60].

Multiple linear regression is one of the predictive models to search the relationship between a dependent variable and several independent variables. It allows estimating the relationship between the dependent variable and a set of explanatory variables and attempts to model the relationship between the input and output variables by fitting a linear equation to observed data. It is also a process to predict the direction of the effect or to understand which parameters have the most significant impact. In multiple linear regression, there are **p** explanatory variables, the independent variables, and the relationship between the dependent variable. The ordinary least square (OLS) technique uses to predict the future event in multiple linear regression.

The multiple regression model form is represented by the following formula:

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

Where β_0 is the constant term and β_1 to β_p are the coefficients relating the **p** independent and ε_i is the error term [61, 62].

2.7 Evaluation Metrics

The OLS (the least squares method) for regression will be used to estimate the unknown parameter in a model to find the best-fit line for data by minimizing the sum of squared error or residuals between the actual and predicted values. It relies on the difference between the actual and expected values, which helps to measure the model's performance. Among different performance measures of the model, some of the evaluation methods are R-squared (R^2), also known as the coefficient of determination, root mean squared error (RMSE), and mean absolute error (MAE). The mean fundamental error is the average amount of error in the measurement. MAE gives the mean of the absolute difference between the prediction and actual value. RMSE is a metric for measuring the performance of a regression-based machine learning model by showing the average of how far apart the predicted values are from the observed values in a dataset. Note that for the validation of forecast results, the smaller the

RMSE value, the better the model fits a dataset, and the greater the RMSE value, the further the forecast value of the actual data. R-squared is one of the performance evaluation measures for regression-based machine learning models. It is a statistical measure that represents the proportion of the variance in the response variable of the regression model that the predictor variables. Simply, it is the difference between the samples in the dataset and the predictions made by the model. The higher the R squared value, the better a model fits a dataset, which means the predicted value is close to the observed value. We will evaluate the performance of the designed model with metrics that are R squared and RMSE and MAE [63, 64].

Equations 3, 4, and 5 show how to calculate R squared value, root mean squared error, and mean absolute error, respectively.

$$R^2 = 1 - \frac{Sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares\ (SST)} = 1 - \frac{\sum(\boldsymbol{y_i} - \hat{y}_i)^2}{\sum(\boldsymbol{y_i} - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} ||\boldsymbol{Y(i)} - \hat{y}(\boldsymbol{i})||^2}{\boldsymbol{n}}} \quad (2)$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} |\boldsymbol{Y(i)} - \hat{y}(\boldsymbol{i})|}{\boldsymbol{n}} \quad (3)$$

Where;$Y(i)$is the actual Value, ŷ(i) is the Predicted Value, and n is the Total number of data points. For data preprocessing, data visualization, and data cleansing, python matplotlib, seaborn, Pandas, and other libraries are used for experimental and evaluation purposes [63, 65]

Chapter 3: Related Work

3.1 Introduction

This chapter discusses several works which are related to our research. Though no specific work covers teff scarcity prediction, we will discuss the pieces we believe are closely related to the challenges in our topic. Furthermore, we review papers that are mainly associated with our work. Hence, the techniques conveyed to the prediction model are discussed and reviewed.

3.2 Prediction of Using Linear Regression

Nwanze et al. [66] develop a machine learning linear regression model and compare it with the average projection and nature fund growth model by presenting the error percentage produced by the new model and some already existing models. Hence, the researcher predicted the population of Nigeria using a machine learning linear regression model and compared the result of the percentage of error with the other two models different scholars developed. According to their mark, the machine learning linear regression model is more accurate than the average projection model and the nature fund growth model researched by other scholars. They produced the result of the new proposed model compared to the old model. The linear regression model has lower percentage error margins, between 0.76% and 1.09%, than the average projection model and the nature fund growth model, with a percentage error margin between 4.73%-1.43% and 0.9%-1.89%, respectively. The research result partly agrees with the previous research, which they tried to compare with and gave qualitative support to a linear regression model. After justifying the motivation for the investigation, the authors modeled the population of Nigeria using a machine learning linear regression model and predicted the people to be 365 million in 2050.

Hiyam et al. [67] assumed that, recently, machine learning methods are accurate techniques and widely used for weather prediction. The aim was to develop a multiple linear regression model to predict the quantity of water falling to earth at a specific place within a specified time (precipitation) for Khartoum state. The prediction model has been developed based on some parameters taken as independent variables mean temperature, maximum temperature, minimum temperature, Dew point, sea level pressure, station pressure, mean visibility, and wind speed. The efficiency of the proposed model was measured by comparing the average

value of the mean square error of the training data with the test data. Hence, the researchers calculated the average mean squared error during the training and testing phase. The research result shows that the mean square error between actual and predicted values of the rainfall precipitation rate has been significantly decreasing during testing time. The researchers used an equal amount of data for testing and training in the first phase, and they found that the efficiency was 85%. But when the test data increased and was more significant than the training data, the model's efficiency was 59%; the researcher recommended that the model may need more data in the training phase and supplementary research should conduct.

Mustafa et al. [68] design a multilinear regression method to predict household natural gas consumption. Household natural gas consumption has a strong relationship with the season, which increases consumption amount during winter and decreases in the summer season. The time series data was divided into six datasets with seasonal behavior and called the model. The multiple linear regression method was applied and checked the probability. Using all data, a new Model, Model 7, has been created by removing multi-collinearity. The mean absolute percent error compares the model's result, and the predictive values create the merged model. The individual model and model seven compare using mean absolute percent errors. Model seven predicted consumption with a MAPE of 14.38%, and the individual model predicted a MAPE of 73.9%. Model 7 improved the error by 55.10%. Apart from MAPE, the research did not use other performance measurements like R squared and root means squared error.

Sreehari et al. [69] predict the climate variable using multiple linear regression by comparing the linear regression method. Using the expected result, the farmers can make appropriate decisions on crop yielding and, simultaneously, the scope of the occurrence of floods or droughts analyzed. The researcher collects a dataset of six years. Then they compared multiple linear and linear regression using the prediction results. The researchers calculated the errors in both regression methodologies by comparing the expected and actual values. The research did not use error metrics to measure the model's performance designed by both MLR and SLR. The result of multiple linear regression error rate is less than simple linear regression based on comparison. The researcher concludes that multiple linear regression methodology exploits better prediction for the climate variable than simple linear regression methodology.

MukhaTmmad et al. [70] evaluate the regression model with seven methods to select which method yields the best coefficient value for multiple linear regression for designing the required Model. The primary purpose was to see which Model predicts forest fires prediction better with minimum error values and which method is better for finding optimal coefficient values for multiple linear regression. The author evaluates Residual Standard Error (RSE), Coefficient of Determination (R Square), Correlation Coefficient (R), Adjusted R Square (Radj2), Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Percentage Error (MAPE) methods for finding the optimal coefficient values of the Model for predicting forest fires. The experiential work produces the result with no dominant method in the coefficient calculation for multiple linear regression methods. But by observing the three methods, which are the coefficient value of the Gauss Jordan, Gauss-Seidel, and Least Squares methods, the Least Square Method tends to be better than the other two methods because according to the calculation result of those evaluation methods, the Least Square Method have less error than the other two methods. From the calculation result of multiple linear regression coefficients, the research shows that the best value of residual standard error (RSE) 62.726442, mean square error (MSE) 3934.60648, root mean square error (RMSE) 64.6974 and mean absolute percentage error (MAPE) 310.44549 that proved the least square method is better.

According to Kavita et al. [71], population growth creates unbalanced demand and supply of different crops, requiring farmers to produce more from the same agricultural land to increase supply to reduce food security problems. Crop yield prediction using attributes of area, yield, production, and area under irrigation attained using machine learning techniques. The primary purpose of the researchers was to estimate crop yield using four different machine learning techniques, which are decision tree, linear regression, lasso regression, and ridge regression, on six significant crops grown in India, namely rice, jowar, maize, bajra, tobacco, and wheat by helping farmers for reducing the loss of production under unsuitable condition and increase the output under suitable and favorable condition. The prediction result shows that the decision tree 98.62, linear regression 89.62, lasso regression 86.33, and ridge regression 89.53 of accuracy. Cross validations methods for validation, mean absolute error, mean squared error, and root mean squared error, were used to validate the result. The decision tree performs better than the other three regression methods.

3.3 Prediction of Using Neural Network

According to Zhang et. al. [72] the economic development implies the contradiction of supply and demand of water consumption is increasing. Hence, optimizing water resource scheduling by predicting water consumption was the focus area of the researcher. The researcher used the neural network method since it is self-organization, self-learning ability, and good fault tolerance characteristics. The developed model achieved high prediction accuracy in short, medium, and long-term urban water demand forecasting. The influenced factors and water consumption were analyzed and selected as the NARX input based on the correlation coefficient. Finally, the researcher applied NARX to predict water consumption, and the experimental result shows that the developed model has higher prediction accuracy.

Chaudhari et. al. [73] developed a model that improves the existing routing schemes, with inherent limitations like low reliability due to resource scarcity, low scalability, limited robustness in a dynamic topology, limited quality of service, and interference of signal on a shared channel. This existing system MANET traffic is nonlinear and lacks quality service provisioning. Quality of service in Mobile Ad hoc NETwork(MANET) is required to meet relevant time service. The researcher tried to determine the future status of the wireless network by considering different external factors which affect the variation of traffic generated by users and interference based on the future availability of resources like buffer space, energy, and bandwidth. The designed model can predict the future availability of resources using wavelet neural networks-based traffic. the researcher evaluated the model performance using packet delivery ratio, computation overhead, memory overhead, and packet delay. The researcher recommended further analysis of the proposed model by comparing it with other well-established quality-of-service routing systems to understand the model behavior.

Wang et.al. [74] tried to predict the annual rainfall using generalized regression neural network. Predicting annual rainfall is vital to control floods, water resources, and other resources. The researcher used the GRNN model to predict the rainfall time series and trained on the rainfall past data and predictions made for some other period. However, the researcher compared the designed model with BP neural network and stepwise regression analysis methods to check whether the developed model predicts better than the existing two models.

The researcher showed that the simulation result of the developed model performs better than the BP neural network and stepwise regression analysis methods with great accuracy. Finally, the researcher concluded that the designed model produced a minor prediction error.

3.4 Prediction of Using Non-linear Machine Learning

Turgay et al. [75] evaluate different statistical model performance to determine which model predict wheat yield trend in Turkey with small error for contributing an essential role in forecasting crop yield for food security and food management. The research used five statistical models for predicting wheat yield trends in Turkey simple linear regression, quadratic regression, cubic regression, single exponential smoothing, and double exponential smoothing. Instead of using all the past yield values for the prediction, the researchers proposed a one-year ahead yield value called window size (number of past year's data used for forecast). Five statistical methods were selected and calculated based on the proposed windowed one-year-ahead wheat yield forecast. Among the five selected prediction methods, the cubic regression method performs and predicts better than the other methods. The second best method was quadratic regression; the other was double exponential smoothing. The researcher's problem was finding the best window size for a year-ahead prediction was impossible if the actual value of the predicted year's yield was unknown. And they proposed dynamic linear regression method can yield prediction using a window size approach in future work.

Bhanumathi et al. [76] proposed a model that can suggest to the farmer which crop is best among different crops from other climatic regions based on the agro-climatic part, which can grow some specific crops. The proposed model predicts how much crop yield is using the efficient algorithm and suggests how much fertilizer data uses to get a good result for the harvest. Random Forest algorithm and Back propagation algorithm implemented for crop production analysis. The backpropagation algorithm uses to train fertilizer data. It evaluates the result of how much nitrogen and phosphorus requires for the area of land, and the Random Forest algorithm is implemented to predict the crop yield rate. The main focus of parameters used in this model is state, district, crop year, season, crop, area, production, and how much phosphorous, potassium, and nitrogen in the soil should use to increase soil fertility. The

model predicts crop production yield and how much fertilizer is required to maximize crop production.

Kumar et al. [77] describe agriculture concerning the environment to predict crop yields by considering different factors. The research focus on the carried out is how to predict crop yields with past historical data on crop production, which includes factors such as temperature, humidity, ph, rainfall, and crop name. The research focused was predicting crops using a machine learning approach and Random Forest algorithm based on past data.

Neha et al. [78] seek to increase and improve the crop yield and quality of the crop to sustain human life by designing a model for predicting crop cultivation. The author develops a crop yield prediction model using machine learning techniques by comparing different linear and non-linear regression models based on 5-fold cross-validation utilizing the prediction performance. The researcher aims to propose and implement a model-based system to predict crop yield production from the collation of past data. The various machine learning algorithm and other techniques on agriculture data from 2013 and 2014 apply to achieve the objectives. Based on the researcher found that mostly default setting, the random forest regression performed the best, followed by nearest-neighbor regression, L2 linear regression with polynomial features, and finally, support-vector regression using a radial-based function kernel.

Kanaga et al. [79] focused on developing a system that advises the farmer on which crops to grow for a better profit by predicting crop yield and price that a farmer can obtain from his land by analyzing patterns in the past dataset. The researcher proposed a demand-based recommender system for the farmers so that the farmer will have adequate crop planning for growing. The past data on agriculture used by the researcher contains parameters of crop areas, types of crops cultivated, nature of the soil, yields, and the overall crops consumed for predicting the future demand of the crops. The future demand was predicted by classifying the collected past dataset based on the change in the market price of the crops. The researcher used the sliding window non-linear regression technique to predict crop yield and price based on different aspects which affect agronomic production, such as rainfall, temperature, market prices, area of land, and past yield of a crop. The researchers claimed that the proposed system reduces the loss and improves the farmers' output.

According to Tahmid et al. [80], agricultural technology plays a vital role in the success of agricultural growth. The goal of the research aimed to provide a learning agent that can aid the farmer in deciding to make farming more efficient and profitable. Using a decision-making algorithm, the researcher identified the list of good crops in a particular area. K-Nearest Neighbor and Decision Tree Leering algorithm used for designing the prediction model. By developing the prediction model, the researcher aimed to help farmers maximize profit margins by predicting crops that provide maximum output in a particular area. The twelve years of data were used for six significant crops predicting the accuracy. The researcher predicts yields in a specific area using the two supervised machine learning algorithms to help the farmers cultivate undesirable crops using previous experience and maximize their profits. The result of the research indicates that the decision tree algorithm gave less value of percentage error compared to the k-nearest neighbor algorithm without omitting the outliers of the dataset.

3.5 Summary

The linear regression method predicts that Nigeria's population will be 365 million in 2050. However, multiple linear regression predicts rainfall. Using time series data divided into six groups and called a model for predicting household natural gas consumption, finally merging all data and called model seven. Hence, the reviewed research revealed that model seven predicted better accuracy using multiple linear regression. Multiple linear regression predicts climate with better accuracy than simple linear regression. Wheat yield trend in Turkey with minor errors contributing an essential role in forecasting crop yield for food security and food management predicted using different regression models. The researcher compared different regression models to find out which regression model performs better than the other. Random forest algorithm predicts the soil's fertilizer amount to maximize the crop. And the future status of a wireless network with different external factors anticipated affects the variation of traffic generated by users and interference based on future available resources like buffer space, energy, and bandwidth predicted using the neural network method. Finally, annual rainfall was predicted using generalized regression neural network model to control flood, water resources, and other related resources. All reviewed predictions targeted benefiting a wide range of communities.

Most of the reviewed research focuses on a single target, like a farmer, to see which regression model performs better in predicting a specific crop for maximizing a profit or production amount using MLR, non-linear regression, neural networks, and comparing different regression models. The reviewed work targeted a particular area for maximizing specified resources. We will take advantage of scarcity prediction using machine learning to benefit large target groups like farmers, government, researchers, consumers, and a whole community by predicting the scarcity of a crop.

Chapter 4: Design of Teff Scaricty Prediction

4.1. Introduction

This section is all about the proposed scarcity prediction model. We will discuss each component of the proposed solution, including preprocessing, train-validation-test, model construction, the model algorithm used, experiment result, and findings discussed in detail. Finally, model performance and prototype evaluation describe in detail.

4.2. Architecture of the Proposed System

The architecture of the proposed solution represents the design decisions related to overall architecture and behavior. It is the fundamental organization of a system embodied in its components, their relationships to each other, and the principles guiding its design. The proposed architecture aims to address the problem and meet the functional and quality requirements of the design [81, 82].

The main factor considered in designing the proposed architecture is whether the design delivers the required outcomes that will satisfy the statements of the problem, which is the goal of the proposed architecture. Figure 4.1 shows the proposed architecture. The architecture indicates a prediction component: the preprocessing, the train-validate-test, and the prediction. The preprocessing component performs preprocessing, data cleansing, and organizing activities. The train-validate-test part splits the prepared data into train, validate, and test activity, and the prediction component performs predicting activity. To predict scarcity in the future, we need to impute the missing data of all attributes. Treating the missing attribute makes the dataset suitable for machine learning, and the result uses for scarcity prediction [31, 83, 25, 84].

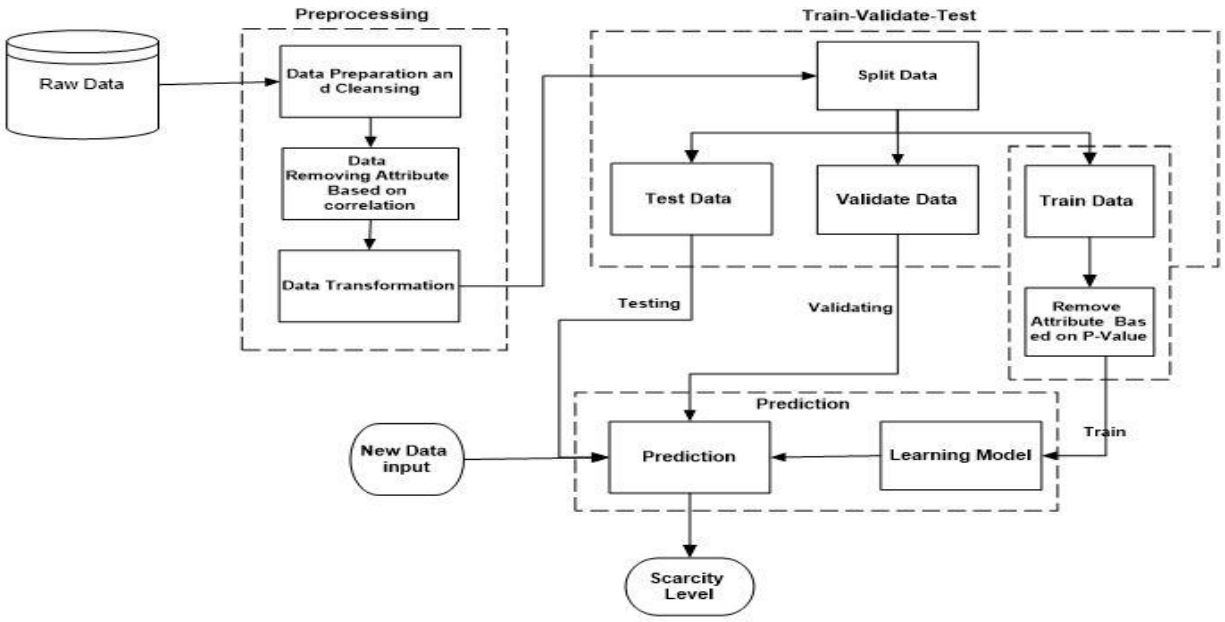

Figure 0-1 Scarcity Prediction Model

The teff scarcity prediction model consists of three components: the preprocessing components, the train-validate-test component, and the prediction component. The preprocessing part consists of data cleansing and data transformation. The train-validate-test component consists of the training, validating, and testing data partition. The prediction component includes creating model instances, removing insignificant variables, and testing the trained models. The raw dataset is fed to the preprocess component to perform data cleansing and transformation, preparing the raw data to be suitable for machine learning. The train-validate-test partitions the dataset into the train, validate, and test dataset. The training dataset is given to the prediction component of the model to train the model. The model will train with the training dataset to understand the behavior of the dataset. The trained model takes validation and a test dataset and validates and evaluates the model's performance. Finally, the model will present the scarcity prediction result. Our proposed model is shown in Figure 4.1.

4.2.1 Preprocessing Component

The data preprocessing component involves transforming raw data into an understandable format. Real-world data is often incomplete and not suitable for machine learning. The preprocessing part prepares the raw datasets for training, validating, and testing our scarcity

model. The preprocess component has actions that clean and transform the dataset suitable for machine learning. The library is imported for the collection of implementations of behavior implementation with a well-defined interface by the preprocess component to invoke the required behavior. Data cleansing activity is responsible for fixing or removing incorrect, corrupted, or incomplete data within the dataset. The data is filtered and modified to make the dataset easier to explore and understand by the algorithm. The dataset transformation activity is responsible for converting the dataset into a suitable format or structure that best represents model fitting. We used log transformation techniques to transform the skewed distribution to a normal distribution/less skewed distribution.

Preprocess Algorithm

```
Input: Raw_dataset
Output: Preprocess_dataset
IMPORT raw_dataset in CSV format
For each attribute in N do:
     if value == NaN
          if value == String
               remove comma
          else
               replace NaN with mean
end for
for each i in N do:
     transform←np.log(1+ mported_dataset_name.attribute_name)
end for
for each i in N do:
     value← correlate i with N-i attribute
     ifcorr_value greater than or equal to 0.8
          delete index[i]
Return preprocessed dataset
```

Algorithm 0.1 Preprocessing

4.2.2 Train Validates Test Component

The train-validate-test component prepares the required partitioned dataset for fitting the model. However, training, validating, and testing the model with a partitioned dataset are required. The partitioned dataset is organized and divided without crossing each other so that

the capability of the model to predict scarcity is efficient. Hence, the test dataset extracted from the total dataset is 15%, and the remaining dataset is 85%. The remaining 85% of the dataset split into validated datasets is 15%. The remaining dataset, 70% of the total dataset, is for the training dataset.

Train validates and test splitting algorithm

```
_____________________________________________________________________
Input: Preprocessed_dataset
Output: train, validate, and test dataset
_____________________________________________________________________
Read preprocessed dataset
     Split dataset into train and remaing_dataset (15%, 85%)
     x_train,x_test,y_train,y_test←train_test_split(x_multi_s
     ,y_multi_s,test_size=0.15)
Read remaining_dataset
Split remaining_dataset into validation and test(15%, 85%)
     x_train,x_val,y_train,y_val←train_test_split(x_train,y_t
     rain,test_size=0.15)
Return split dataset
_____________________________________________________________________
```

Algorithm 0.2 Train test and validate

4.2.3 Prediction Component

The prediction component addresses our thesis objectives and evaluates how well the model performs using the partitioned dataset. The model learns from the training dataset and gains experience from the sample dataset so that the model acquires the behavior of the dataset for extracting a pattern. However, the model is validated and evaluated using the experience gained during the training and validation activities. The prediction component learns the behavior of the dataset during the training as well as validating actions of the train-validate-test part. Hence, based on the acquired knowledge from the training activity, the prediction component uses the test dataset to evaluate the performance. The model used a multiple linear regression algorithm for performing all training, validating, and testing activities. The algorithm uses the test dataset to conclude whether the model is fit. Finally, after testing the model, the R squared, root mean squared error and mean absolute error results uses to decide whether the designed model predicted well as expected.

Prediction algorithm

```
Input:  train, validate, and test the dataset
Output: Scarcity prediction
Read train dataset
Store the independent and dependent variables on a separate
variable
     independent_variable←dataset_name.drop("dependent_variab
     le", axis=1)
     dependent_variabl←dataset_name[' dependent_variable ']
     Construct the model
     model←sn.OLS(dependent_varible,independent_variable).fit
     ()
for each i in N do:
     if p-value >=0.05
          delete index[i]
     independent_variable←dataset_name.drop
     ("dependent_variable", axis=1)
     dependent_variabl←dataset_name[' dependent_variable ']
     modified_model←sn.OLS(dependent_varible,independent_vari
     able).fit()
end for
Train_modelmodel_name.fit(x_train, y_train)
Predict ← lm_multi_sc.predict(x_train)
Return:
     R_squared← r2_score(y_train, y_pred)
     MAE ←mean_squared_error(y_val, y_val_s_pred)
     RMSE ←math.sqrt(MSE)
Repeat read validate and test dataset
```

Algorithm 0.3 Prediction

Chapter 5: Experiment

5.1 Introduction

This chapter will discuss the dataset preparation and analysis, tools, and experimental setup used to design and develop the prototype to implement the model. We will discuss the results acquired from the experiment of the model and give an interpretation of the experimental findings.

5.2 Data Preparation and Analysis

We live in a world where data is collected and analyzed in every aspect of our life for problem-solving, which leads to the success of design machine learning. Data collection and analysis play a prominent role in the world to help people to improve their quality of life by allowing them to measure and take action to establish baselines and goals to keep moving forward. It helps to see the solution clearly, which depends on a clear understanding of the problem using data collection and analysis [85]. Data analysis is the process of systematically applying statistical and logical techniques to describe and illustrate large volumes of data to extract meaningful information, which helps uncover hidden details and patterns and discover knowledge from large volumes of data [86]. The critical part of data analysis is data preprocessing which determines the applicability of machine learning [15]. In this study, relevant documents were reviewed and discussed thoroughly to get a deeper understanding of the problem.

Most of the dataset for this work relays on the Central Statistical Agency (CSA) with document review. However, in collecting appropriate data for the scarcity prediction model, the population and product datasets are paramount for completing the required missing attributes. However, this dataset's primary limitation is the organization of the data. The organization of population data must be in yearly series. However, some of the data organization is in a ten years pattern. CSA has very little time series data on population, production, consumption, scarcity, birthrate, mortality, and other relevant variables. The collected data covered the period from 1984 up to January 2037. The data collected from the document review called Area and Production of major crops and population projection for Ethiopia prepared by the federal republic of Ethiopian central statistics agency. On the other

hand, the missing data is calculated from the condition factor of the collected data directly from the reviewed documents and the existing population and product data [31, 83, 25]

The historical population, product, and other relevant data from 1/1/1984 to 30/12/2037, about 54 years dataset were computed and gathered from CSA and published paper. Hence, the multilinear regression method utilized the identified preprocessed attributes. The transformed data divides into train, validate, and test datasets.

Finally, the random data division method was applied. 70% of the dataset was partitioned and kept for training the model, and 30% of the dataset kept validating and testing models to evaluate the model's goodness or fitness [31, 83, 25, 84].

5.3 Data Preprocess

The data preprocessing goal was to organize the dataset in a yearly series.Usually, the data collected from different sources must be in a format suitable for machine learning and may contain invalid and missed values. Machine learning needs specific representations of the input data for training and testing, termed features, before feeding to an ML algorithm. Hence, using data preprocessing, the data is prepared to make it in a required format for machine learning. However, using the data preparation dataset, the population and other missing attributes from 1985-1993 and 1995-2006 were calculated in each year using the average annual growth rate equation indicated in equation 1 and the lost estimated population and other missing attribute values at a particular time calculated using the equation shown at 2. The summarized criteria considered to calculate the missing data from 1984-1993 and 1995-2006:

1. The missing population data is calculated from the existing dataset using equation1 and 2

$$r = \left(\frac{1}{T}\right) \ln\left(\frac{P_n}{P_o}\right) * 100 \quad (4)$$

Where r is the average annual growth rate, P_n is the population size at n year, $P_o$is the population size at the initial year, and T is the time. Using equation (1) which is exponential law, we manage to calculate the average annual growth rate of the population so that the missing data will be filled [31, 83, 24]

$$P_{n+1} = (P_n * r) + P_n \quad (5)$$

Where P_{n+1} is the population at n+1 year, P_n is the population at n year, and r is the average annual growth rate.

Using equation (2), the missing data of the population is calculated, and the result of population data, including the lost data, is indicated in Annex A.

The missing scarcity is calculated from the population and consumption of Teff at a household level. After calculating the shortages, the result was used to calculate the production from 1984-1993 and 1995-2006 to fill the missing data of the Teff crop, indicated in Annex B.

The other important data for predicting the model are consumption, household, and scarcity of a particular crop. Scarcity was calculated using equation 1 after consumption calculated from the existing data [87]. On average, sampled households consumed 24.2 kg of cereals, from which Teff covered 7.5 kilograms of the total consumption. The number of households ranges from 12 to 2, and the average household size is 5, which was used as a factor to calculate consumption. Estimating the population involves estimating and multiplying the average household size by the estimated number of households. However, the missing household values are calculated based on the information of the average number of households and the patterns of the household, collected from CSA population projection for Ethiopia 1984, 1991, 2007-2037, and 2001-2020 of area and production of significant crops reports which helps to calculate the consumption, scarcity values. After calculating missing values of consumption and scarcity, the total missing production is calculated by subtracting the total scarcity from the total consumption [31, 83, 25].

5.4 Data Quality

The datasets were checked for completeness and correctness of the required attributes and integrity before analysis and prediction. Details of data exploration discuss in the immediate sections below.

5.4.1 Checking the Dataset Contains the Required Attributes.

The study explored different studies related to scarcity prediction. Moreover, we have investigated the required attribute (column names) for our analysis of technical indicator

computation. Our data from CSA contained the needed features, Total population, Male, Female, Year, Area in Hectare, Production in quintals, yield per hectare, birth rate, death rate, migration, married, divorced, widow, disabled, and consumption. Our data contains enough attributes to compute the required indicators that can later be considered Scarcity features.

5.4.2 Conducting Data Integrity Test

In this study, we expected data to be missed in some rows because the data exists in a ten years sequence. We use the year value as a reference to find the lost attribute. The year column uses to locate the missing attributes in a row from the dataset. The missing values were observed in some rows of original datasets whose values were estimated using interpolation considering the neighbors [31, 83, 84]

5.4.3 Experiment Data Description

The data from CSA contains all the required information and is organized accordingly for the scarcity prediction model. The collected data from 1984 up to 2037 uses for the experiment. The data from CSA and computed contains Pop, Prod, Consup, Year, Male, Female, ABR, ADR, Migrant, Married, Divorced, Widow, Disable, Area, and YieldPH columns—the explored attributes for the study presented in Annex C [31, 83, 84].

5.5 Tools and Experimental Setups

We used Python programming language for the experiment. We chose Python programming language because there are Python libraries readily available for all our ML workflow. We have used an anaconda machine learning model for training, validating, and testing. Multilinear regression statistical methods use because the technique is used in many prediction problems and has proved that it is for function approximation problems and prediction. Three statistical metrics are used for evaluating the model: the root mean squared, the mean squared error, and the mean absolute error. A model with a considerable R2 value on the testing data and a small RMSE and MAE will be considered a fitted model [41, 63].

5.6 Model Performance Evaluation

The performance evaluation performed with three metrics variables. This section applies the multilinear regression technique as a metrics variable selection tool. Selection aims to reduce the set of dependent variables to necessary ones and account for nearly as much of the variance as is accounted for the entire group. It determines the level of importance for each variable. The P-value is the determinant point to determine the significance of the independent variables, and R squared value, RMSE, and MAE determine the model's accuracy [63]. The three major processes used to evaluate the model were preprocessing, train-validate-test, and prediction methods.

Based on [87], the weekly consumption of product of Teff at a household level increased from 7.27 kg in 2011 to 7.70 kg in 2014, which indicates that the population size increased from 2011 to 2017. However, as the consumption and population increase, the product size should be increased parallel to resolve the scarcity of the product. If the crop quantity decreases against the consumption and population, scarcity is introduced and requires management to reduce the problem. The scarcity is predicted using the pattern of the population and product size, which implies that the missing population and product data should fill in order make complete the dataset to predict the scarcity that will be happened in the future.

The scarcity model construction component has three major components to designing the multilinear regression prediction model: a preprocessed component, a train-validate-test component, and a prediction component [18, 87, 23]

OLS Regression Results

Dep. Variable:	Scarcity	**R-squared:**	0.930
Model:	OLS	**Adj. R-squared:**	0.916
Method:	Least Squares	**F-statistic:**	65.45
Date:	Thu, 09 Mar 2023	**Prob (F-statistic):**	1.58e-22
Time:	15:58:29	**Log-Likelihood:**	-920.62
No. Observations:	54	**AIC:**	1861.
Df Residuals:	44	**BIC:**	1881.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.024e+09	1.12e+08	9.168	0.000	7.99e+08	1.25e+09
Consup	-1.026e+08	4.77e+06	-21.529	0.000	-1.12e+08	-9.3e+07
ABR	2.234e+06	9.63e+06	0.232	0.818	-1.72e+07	2.16e+07
ADR	-8.503e+06	3.06e+06	-2.778	0.008	-1.47e+07	-2.34e+06
Migrant	8.309e+06	3.52e+06	2.357	0.023	1.21e+06	1.54e+07
Married	5.624e+06	3.14e+06	1.792	0.080	-7.01e+05	1.19e+07
Divorced	-5.136e+06	3.36e+06	-1.527	0.134	-1.19e+07	1.64e+06
Widow	1.148e+07	7.1e+06	1.617	0.113	-2.83e+06	2.58e+07
Disable	3.624e+05	1.23e+06	0.296	0.769	-2.11e+06	2.83e+06
Prod	2.793e+07	5.01e+06	5.581	0.000	1.78e+07	3.8e+07

Omnibus:	0.029	**Durbin-Watson:**	2.246
Prob(Omnibus):	0.986	**Jarque-Bera (JB):**	0.065
Skew:	-0.034	**Prob(JB):**	0.968
Kurtosis:	2.844	**Cond. No.**	5.24e+03

Figure 0-1 Scarcity P-value Result

Figure 5.1 shows the p-value of each independent variable, we can see that each predictor contributes to the model. However, Divorced, ABR, ADR, Migrant, Married, Widow, and Disable are insignificant for prediction. Since Divorced, ABR, ADR, Migrant, Married, Widow, and Disable p-values ≥ 0.05, they are insignificant for the model. We remove those attributes and identify the significant variables for the model.

The essential attributes for prediction are identified based on the p-value measure. Figure 5.2 shows the p-value of the selected attributes for the scarcity perdition. Figure 5.2 shows that the p-value is < 0.05 meaning the variable is significant in designing the model. The other influential value is the coefficient of the variable. Figure 5.2, there is an inversely proportional negative coefficient, which implies the independent variable increased and the dependent variable decreased [62].

OLS Regression Results

Dep. Variable:	Scarcity	**R-squared:**	0.914
Model:	OLS	**Adj. R-squared:**	0.907
Method:	Least Squares	**F-statistic:**	130.7
Date:	Thu, 09 Mar 2023	**Prob (F-statistic):**	1.69e-25
Time:	16:08:15	**Log-Likelihood:**	-926.27
No. Observations:	54	**AIC:**	1863.
Df Residuals:	49	**BIC:**	1872.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.114e+09	6.14e+07	18.135	0.000	9.91e+08	1.24e+09
Consup	-1.023e+08	4.98e+06	-20.546	0.000	-1.12e+08	-9.23e+07
ADR	-1.056e+07	3.01e+06	-3.512	0.001	-1.66e+07	-4.52e+06
Migrant	8.578e+06	3.59e+06	2.390	0.021	1.37e+06	1.58e+07
Prod	3.41e+07	4.37e+06	7.800	0.000	2.53e+07	4.29e+07

Omnibus:	0.385	**Durbin-Watson:**	2.045
Prob(Omnibus):	0.825	**Jarque-Bera (JB):**	0.534
Skew:	0.164	**Prob(JB):**	0.766
Kurtosis:	2.640	**Cond. No.**	1.93e+03

Figure 0-2 Scarcity P-value Modified Result

Looking at Figure 5.2, from all attributes prepared for the scarcity prediction model Consup, Prod, ADR, and Migrant attributes are identified as significant for the model based on the p-value.

Let’s check if the error terms are normally distributed or close to normal distribution for proceeding to the next phase. Figure 5.3 shows that the error terms resemble the scarcity of train stage actual and predicted values.

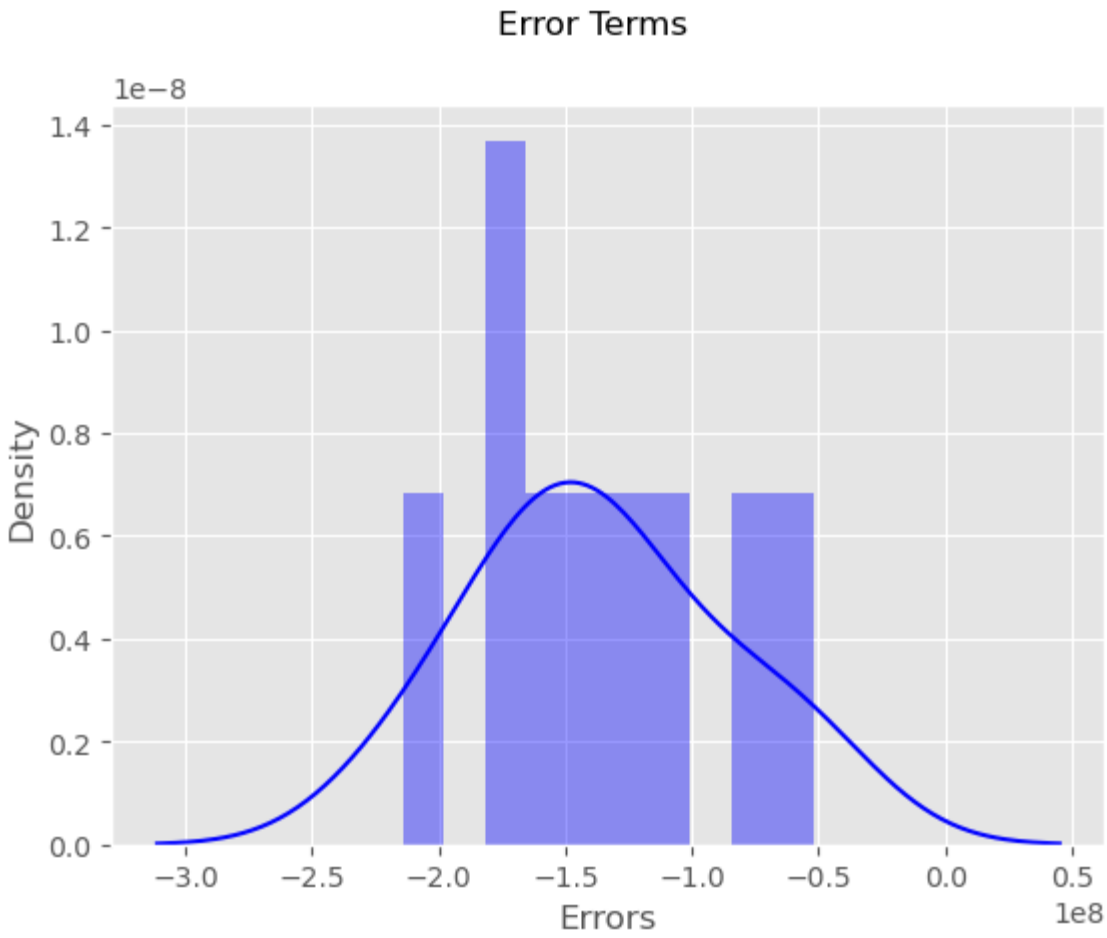

Figure 0-3 Scarcity Error Terms Distribution

As shown in Figure 5.3, the error terms resemble a normal distribution on the scarcity dataset. So, we can proceed and make predictions using validated and test datasets.

5.6.1 Scarcity Data Train-Validate and Test Split

The primary purpose of any model-building process is to develop a generalizable model on the available data, which will perform well on unseen data. But we need a separate dataset not previously involved in the training and validating stage to estimate a model's performance on unseen data. Creating a train, validation, and test split of the dataset is one method to evaluate the performance of an algorithm of the problem quickly. Hence, this is achieved by splitting our available data into training, validating, and testing datasets. The dataset is split to train, validate, and test the dataset using a train-validate-test component of scarcity. The model was

trained with the appropriate partitioned dataset and validated using the validated dataset to determine the model's capability. The testing dataset is used to test the finalized model to find the predictive capability of the final designed model.

It is important that the training, validating, and testing dataset are independent of each other and do not overlap. It is because crossing the dataset with each other will cause low generalizability. Therefore, to find high generalizability, the train, validate, and test datasets should be separated and not cross each other. The split size can depend on the size and specifics of the dataset. In our case, out of the total preprocessed dataset, the machine used 70% of the entire dataset for training, 15% of the whole dataset for validating, and the remaining 15% of it used for testing the model [88].

5.6.2 Scarcity Model Training and Validating

After the data is ready for machine learning, the prepared input is fed to the algorithm so that the model learns the behavior of the data to determine how the model works when applied for the final test for predicting scarcity with the new dataset. The model learns the data during the algorithm's training and produces the new predicted results. The predicted accuracy of the model based on the training dataset provided is 89.78%, which tells the model to predict well in the training stage. The model is validated using 15% of the dataset chosen out of 100% for validating the model to see if the model is predicted well utilizing the dataset, which is not part of the training dataset. Using the validation dataset, the model can predict with an accuracy value of 91.85%, indicating the model is predicted well in the validating stage [88].

5.6.3 Scarcity Model Performance Evaluation

During the training and validating process, the designed model predicts pretty well. The model was trained and validated with 85% of the entire dataset at this stage. The rest of the 15% of the entire dataset, which the model was unaware of during the training and validating process used for testing the model's performance. The testing phase tells how good the model's performance is when used in the real world. The testing phase shows how well the model can predict when the algorithm receives the new dataset. The model was tested with 15% of the entire dataset to see if the designed model predicts well, as it predicts well in the training and validation stage. The testing result shows that the model can predict with an accuracy of

97.07%, which is the model's performance. Thus, this concludes that 97.07% of the independent variable explained the prediction model [60].

5.6.4 Model Evaluation Result

We can predict the scarcity of teff crop using multilinear regression, and we will discuss the result in detail.

The scarcity model has repeatedly experimented with the prepared dataset to confirm the correct result. The experiment process has evaluated the required designed model. During model training with 70% of the total dataset, the **MAE** is 3517585.17, where the converted percentage value is 7%, the **RMSE** value is 17.80%, and the **R2** value is 89.56%. During model validation using 15% of the entire dataset, the experiment result produced by **MAE** is the same value during the training phase since it calculates the average error values, **RMSE** is 13.01%, and the **R2** result is 92.06%. The final phase of the process is testing the designed model using 15% of the total dataset for evaluating the performance of the developed model. During the testing phase, the **MAE** is unchanged and similar to the value during the train and validate step, the **RMSE** value is 6.43%, and the **R2** value is 97.07%. From the experiment results, the designed model produced an outstanding prediction value of **R2** of 97.07% of accuracy, a very small **RMSE** of 6.43%, and an **MAE** of 7%.

5.6.5 Prototype Development

The developed prototype was deployed locally on ‘localhost:8888/notebooks/Deskto/ProtoType/Scarcity_TVT.ipynb’ to test the efficiency of the prototype. The local machine the developed prototype deployed is cori5, which has 4GB RAM and 1TB of HD and Windows OS Environment (Win10, 64bit). The prototype displays the input window for a user to import the collected dataset in a CSV file extension. Figure 5.4 Shows the interface's screen shoot, allowing a user to import the stored data.

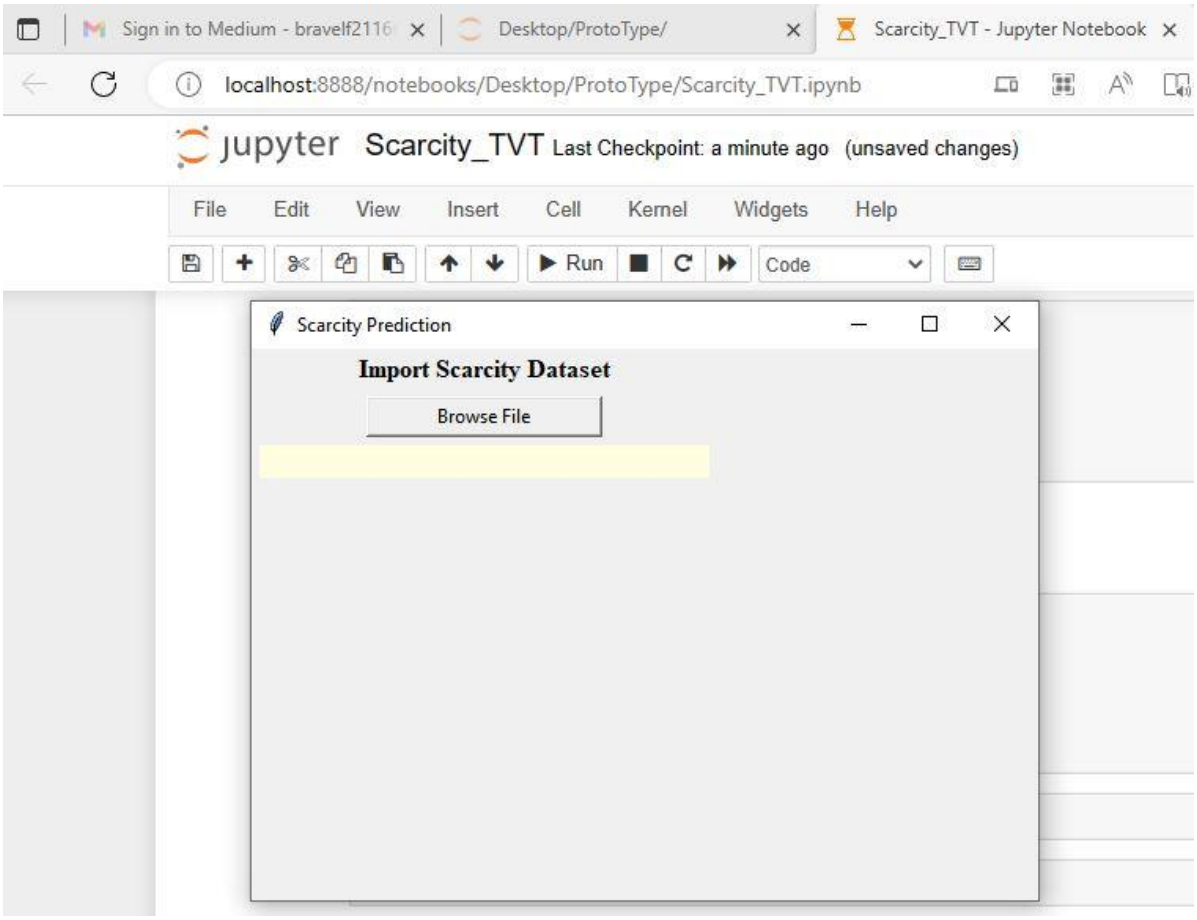

Figure 0-4 Importing dataset window for a user

After the dataset is imported using the interface indicated in Figure 5.4, the raw dataset is displayed. Figure 5.5 shows the list of rows ready for the preprocessing task.

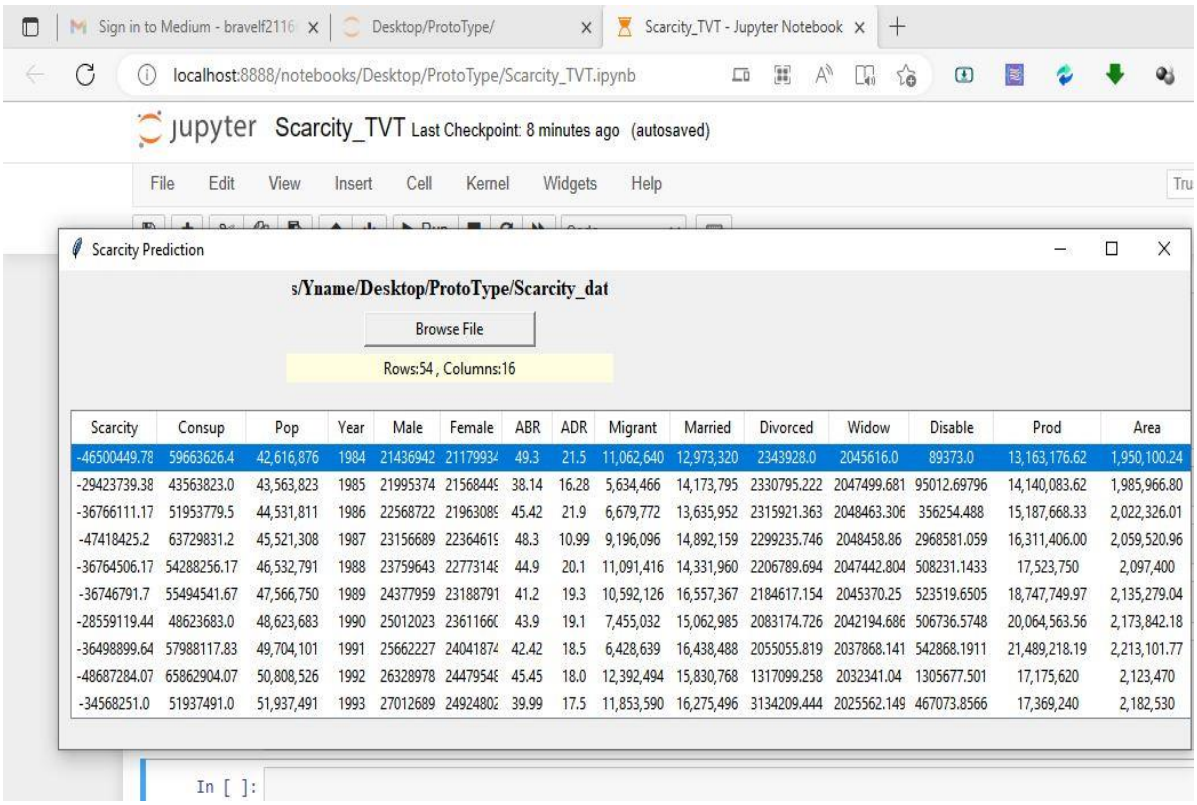

Figure 0-5 Imported dataset display window

After all preprocessed datasets are complete, the prototype will display the interface for a user to insert the significant attribute values to predict scarcity. Figure 5.6 shows the screen shoot interface allowing users to insert the input.

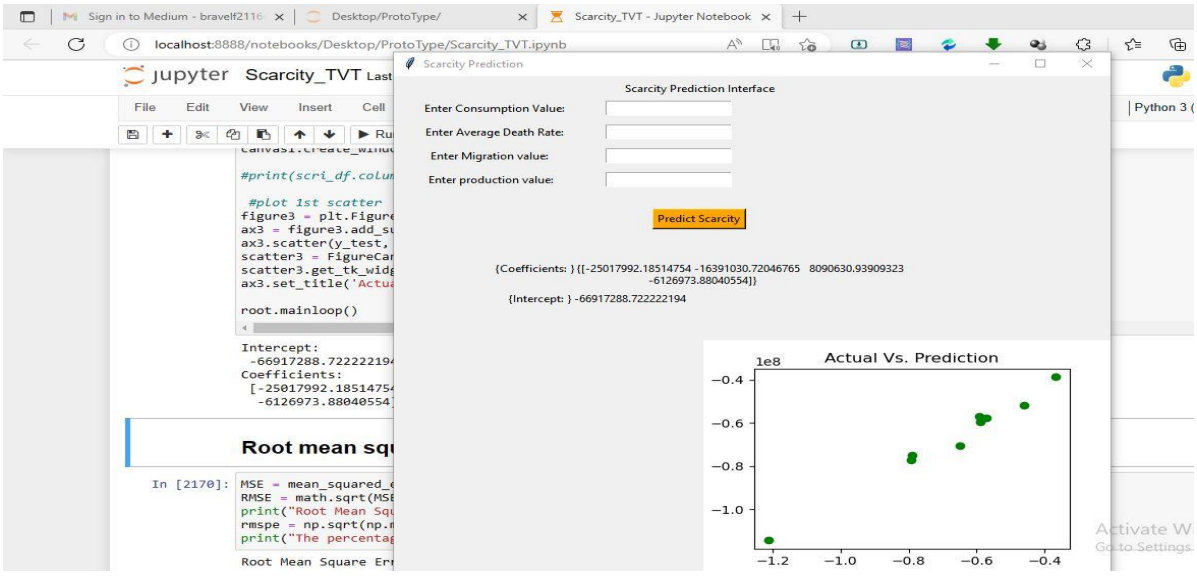

Figure 0-6 User input window for prediction

The prototype displays the result of the prediction for the user as depicted in Figure 5.7

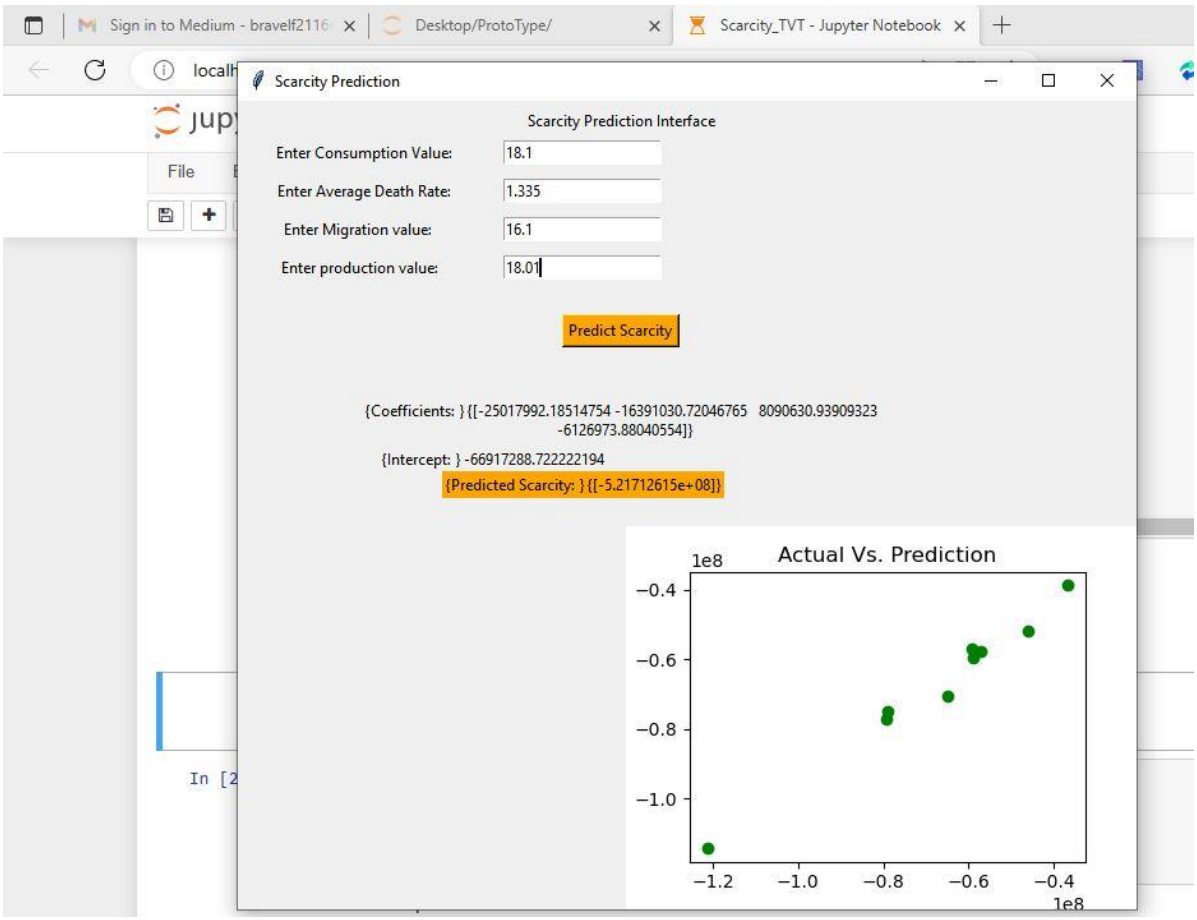

Figure 0-7 Prediction result display window

The prototype was evaluated and achieved what was expected to accomplish. It performs the prediction activity with the expected result and proved that efficient.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

Teff scarcity prediction remains a challenging issue since it needs a prediction result of other parameters used as independent variables, which generate the scarcity prediction in this study. The research implements machine learning techniques, a multilinear regression method for predicting Teff scarcity since it is a prominent staple grain in Ethiopia. As the population increases, food demand and supply chain have become challenging to maintain food security. We design a prediction model to predict the scarcity of teff for a specified time so that, using the result, stakeholders can minimize lavish spending of budget. Using a multilinear regression approach, we tried to propose a model that will predict teff scarcity at an early stage. The hypothetical metrics are tested individually against the independent variable. Then, the significant variables are entered into the regression model to construct the scarcity prediction model. After refining the dataset using reviewing different documents, the preprocessed activity makes the dataset suitable for prediction. The prototype has been developed and evaluated with the preprocessed dataset. The proposed model is evaluated statistically based on **MAE, RMSE,** and **R2** values. The final result of the experiment produced an **MAE** is 7%, **RMSE** is 6.43%, and **R2** is 97.07%, which tells us the value of **MAE** and **RMSE** produced in the experiment is very small, and the **R2** value is very great. The prediction revealed that the proposed model performs at the required level to predict scarcity, implying the model is fit. This is beneficial for different entities like farmers, the government, and the community as a whole. The developed model benefited the farmers, the government, and the community. Farmers can get reliable information about the crop deficit for the upcoming year from the prediction. They can grow the crop in compliance with the prediction. The government uses the prediction result to create sustainable solutions by introducing different rules and regulations based on the prediction result. This study result shows that it helps to improve crop management for better food security and resolve scarcity by alerting all stakeholders.

6.2 Contribution

The contributions of this thesis work are:

- Design a new model for scarcity prediction
- We tested our new model on teff prediction and have achieved a competitive result
- We generate an organized dataset that can be used for different research.

6.3 Future Work

The current work can be further extended and enhanced by adding the following feature:

- Development of scarcity prediction tools. Using the result of the research on dependent and independent variables, it will be worthy to develop standalone scarcity predictor applications based on this research result.
- The research can be scaled up for different areas for predicting the future situation.
- Applying the model on different datasets.

References

[1] Dey, U. Kumar, A. Masud and M. Uddin, "Rice Yield Prediction Model Using Data Mining," *In 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE),,* pp. 321-326, 2017.

[2] CSA, Central Statistics Agency, [Online]. Available: https://www.statsethiopia.gov.et/a-proclamation-to-establish-the-central-statistics-2/. [Accessed 1 2 2021].

[3] B. Li, T. Wang and L. Jia, "Application of Improved Logarithm Logistic Models in Population Prediction," *In 2012 Eighth International Conference on Computational Intelligence and Security,* pp. 99-102, 2012.

[4] S.Y.Chaganti, P.Ainapur, M.Singh, Sangamesh and S. R., "Prediction Based Smart Farming," *In 2019 2nd international conference of computer and informatics engineering (IC2IE),* pp. 204-209, 2019.

[5] N.Zannou and V.R.Houndji, "Sorghum Yield Prediction Using Machine Learning," *In 2019 3rd International Conference on Bio-engineering for Smart Technologies,* pp. 1-4, 2019.

[6] Institue, IFPRI, [Online]. Available: https://media.africaportal.org/documents/ESSP_II_Working_Paper_16.pdf. [Accessed 04 02 2021].

[7] Macrotrends. [Online]. Available: https://www.macrotrends.net/countries/ETH/ethiopia/population-growth-rate. [Accessed 04 02 2021].

[8] S. Veenadhari, B. Misra and C. Singh, "Machine Learning Approach for Forecasting Crop Yield Based on Climatic Parameters," *In 2014 International Conference on Computer Communication and Informatics ,* pp. 1-5, 2014.

[9] R. Kumar, M. Singh, P. Kumar and J. Singh, "Crop Selection Method to Maximize Crop Yield Rate Using Machine Learning Technique," *In 2015 international conference on smart*

technologies and management for computing, communication, controls, energy and materials, pp. 138-145, 2015.

[10] X.Wang, X.Wu and B.Sun, "Factor Selection and Regression for Forecasting Relief Food Demand," *In 2012 8th International Conference on Natural Computation,* pp. 226-228, 2012.

[11] Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Design_science_(methodology). [Accessed 12 3 2020].

[12] A. R. Hevner, S. T. March, J. Park and S. Ram, "Design Science in Information Systems Research," *Management Information Systems Research Center, University of Minnesota,* pp. 75-105, 2004.

[13] H. Freeman, B. Shiferaw and S. Swinton, "Assessing the Impacts of Natural Resource Management Interventions in Agriculture: Concepts, Issues and Challenges.," *In Natural resource management in agriculture: methods for assessing economic and environmental impacts,* pp. 4-16, 2005.

[14] L. Armstrong, D. A. Diepeveen and N. Gandhi, "Effective ICTs in Agricultural Value Chains to Improve Food Security: An International Perspective," *In 2011 World Congress on Information and Communication Technologies ,* pp. 1217-1222, 2011.

[15] V. R. Natalya and M. K. Dmitry, "Data Presentation and Application of Machine Learning Methods for Automating Retail Sales Management Processes," *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus),* no. 03 March 2019, pp. 1444-1448, 2019.

[16] F. S. McConnell Campbell, Microeconomics Principles, Problems, and Policies, New York: McGraw-Hill Education, 2018.

[17] S. Adeyeye, The Role of Food Processing and Appropriate Storage Technologies in Ensuring Food Security and Food Availability in Africa, United Kingdom: Emerald publication, 2017.

[18] C. B. Andrea, "National Food Security Assessment Through the Analysis of Food Consumption Data from Household Consumption and Expenditure Surveys: The case of Brazil's Pesquisa de Orçamento Familiares 2008/09.," *Food Policy,* pp. 20-26, 2017.

[19] M. R. Kanchana, "Utilization Based Prediction Model for Resource Provisioning," *In 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP),* pp. 1-6, 2017.

[20] K. Max, Applied Predictive Modeling, New York: Springer, 2013.

[21] D. Abbott, Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst, Indianapolis, Indiana, Published simultaneously in Canada: John Wiley & Sons, 2014.

[22] K. Micheline, Data Mining: Concepts and Techniques, San Francisco: Diane Cerra Elsevier Inc, 2006.

[23] Fikadu, Wedu, T.D. and Derseh, "Review on Economics of Teff in Ethiopia," *Open Access Biostatistics & Bioinformatics,* no. February 11, 2019;, pp. 1-8, 2019.

[24] Sintayehu and Matintu, "Mathematical Model of Ethiopia's Population Growth," *Journal of Natural Sciences Research ISSN 2224-3186 (Paper) ISSN 2225-0921 (Online),* vol. 6, 2016.

[25] C. Statistical, "Agricaltural Sample Survey Report on Area and Productionfor Major Crops," 1996 -2021.

[26] D. P. Christopher, Economics for Investement Decision Maker: Micro, Macro, and International Economics, Hoboken, New Jersey, Published simultaneously in Canada: John Wiley & Sons, 2013.

[27] A. Koutsoyiannis, Moderen Microeconomics, Waterloo, Ontario: The Macmillan Press ltd, 1975.

[28] L. X. Xingyuan, "Supply and Demand Prediction of Shandong Coal Resources Based on Mathematical Statistics Model," *2011 International Conference on Remote Sensing, Environment and Transportation Engineering,* 2011.

[29] R. A. D. D. M. Joe, "AGRITECHNO: A Development of a Revolutionized Farmer Assisted Agricultural Product Forecasting Mobile App System," *2019 2nd World Symposium on Communication Engineering (WSCE),* 2019.

[30] W. Jia, L. Chao, C. Wei and Z. Yuxiao, "Personalized Collaborative Filtering Recommendation Algorithm Based on Linear Regression," *2019 IEEE International Conference on Power Data Science (ICPDS),* no. 02 March 2020, pp. 139-142, 2019.

[31] C. S. A. Ethiopia, "Central Statistical Agency Population Projections for Ethiopia 2007-2037," Addis Abeba Ethiopia, 2013.

[32] A. K. Samidha, "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: a Comparison," *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) IEEE,* no. 09 April 2020, pp. 680-683, 2020.

[33] B.I. Halil, "Comparison of Classification Techniques Used in Machine Learning as Applied on Vocational Guidance Data.," *2011 10th International Conference on Machine Learning and Applications and Workshops,* pp. 298-301, 2011.

[34] C. Guo, "A New Machine Double-layer Learning Method and its Application in Non-linear Time Series Forecasting.," *In 2007 International Conference on Mechatronics and Automation,* pp. 795-799, 2007.

[35] T.M. Andres, "Forecasting Agricultural Commodity Prices Using Multivariate Bayesian Machine Learning Regression.," 2010.

[36] J. H. Trevor, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York: Springer Series in Statistics, 2009.

[37] Harimurti, R., Yamasari, Y. a. Asto and B.I.G.P, "Predicting Student's Psychomotor Domain on the Vocational Senior High School Using Linear Regression," *In 2018 International Conference on Information and Communications Technology (ICOIACT),* pp. 448-453, 2018.

[38] D. R. Kushal, "Analysing the Role of Supervised and Unsupervised Machine Learning in IoT," *In 2020 international conference on electronics and sustainable communication systems (ICESC),* pp. 75-79, 2020.

[39] B. Hoss and H. Alireza, "sciencedirect," Elsevier, 06 January 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128219294000044. [Accessed 06 January 2023].

[40] T. S. Dipanjan, Practical Machine Learning with Python, India: Berkely: Apress, 2018.

[41] Lee, C.H., Yang, Chen and S. T.C. Ma, "A Comparative Study on Supervised and Unsupervised Learning Approaches for Multilingual Text Categorization.," *In First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC'06),* pp. 511-514, 2006.

[42] R.K. Rajendra, "Semi-Supervised Clustering Using Seeded-kMeans in the Feature Space of ELM," *In 2016 IEEE Annual India Conference (INDICON),* pp. 1-6, 2016.

[43] F. Kholid, "Application of K-Nearest Neighbor Algorithm for Puzzle Game of Human Body's System Learning on Virtual Mannequin.," *2018 International Conference on Applied Science and Technology (iCAST),* pp. 530-535, 2018.

[44] S. A. Green, "Classification of Lower Back Pain Using K-Nearest Neighbor Algorithm," *2018 6th International Conference on Cyber and IT Service Management (CITSM),* pp. 1-5, 2018.

[45] G. N. Okfalisa, "Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification," *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE),* pp. 294-298, 2017.

[46] S. N. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician. 1992 Aug 1,* vol. 46, pp. 175-185, 1992.

[47] R. Saxena, "How Decision Tree Algorithm Works," Dataaspirant, 30 January 2017. [Online]. Available: https://dataaspirant.com/how-decision-tree-algorithm-works/. [Accessed 06 January 2023].

[48] S. Sachin, "Eager Decision Tree," *2017 2nd International Conference for Convergence in Technology (I2CT),* pp. 830-840, 2017.

[49] N. I. Nnamdi, "A Decision Trees Approach to Oil Price Prediction.," *2017 International Artificial Intelligence and Data Processing Symposium (IDAP),* pp. 1-5, 2017.

[50] StackExchange, "Are Decision Tree Algorithms Linear or Nonlinear," Data Science Stack Exchange, [Online]. Available: https://datascience.stackexchange.com/questions/6787/are-decision-tree-algorithms-linear-or-nonlinear. [Accessed 06 January 2023].

[51] K. M. Jain, "Artificial Neural Networks: A Tutorial," vol. 29, pp. 31-34, 12 January 1996.

[52] M. Z. M. Moradi, "A Neural Network Based System for Intrusion Detection and Classification of Attacks," *Proceedings of the IEEE international conference on advances in intelligent systems-theory and applications,* pp. 15-18, 2004.

[53] A. S. Christine, "Revenue Prediction Using Artificial Neural Network," *2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies IEEE,* pp. 97-99, 2010.

[54] V. Ashlesha, "Predictive and Probabilistic Approach Using Logistic Regression: Application to Prediction of Loan Approval," *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT),* pp. 1-6, 2017.

[55] K. Z. Xiaonan, "Logistic Regression Model Optimization and Case Analysis," *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT),* pp. 135-139, 2019.

[56] P. S. Ajay, "PPS—Placement Prediction System Using Logistic Regression," *2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE),* pp. 337-341, 2014.

[57] G. C. M. Douglas, "Introduction to Linear Regression Analysis," *International Statistical Review,* pp. 318-319, 2013.

[58] R. S. Kavitha, "A Comparative Analysis on Linear Regression and Support Vector Regression," *2016 Online International Conference on Green Engineering and Technologies (IC-GET),* pp. 1-5, 2016.

[59] W. M. Ching-Seh, "Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data," *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS),* pp. 16-20, 2018.

[60] A. M. Tamer, "A Comparison Study Between Soft Computing and Statistical Regression Techniques for Software Effort Estimation," *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE),* pp. 1-5, 2018.

[61] E. Mehtap, "Comparison of the Efficiency of Principal Component Analysis and Multiple Linear Regression to Determine Students' Academic Achievement," *2012 6th International Conference on Application of Information and Communication Technologies (AICT),* pp. 1-5, 2012.

[62] S. Areerachakul, "A Comparison Between the Multiple Linear Regression Model and Neural Networks for Biochemical Oxygen Demand Estimations," *2009 Eighth International Symposium on Natural Language Processing,* pp. 11-14, 2009.

[63] S. Ashish and B. Dinesh, "Survey of Stock Market Prediction Using Machine Learning Approach," *International Conference on Electronics, Communication and Aerospace Technology ICECA 2017,* pp. 506-509, 2017.

[64] S. Iwan, I. H. Dito, P. Ira, B. Tessy and S. Edi, "Corn Pests and Diseases Prediction Using Linear Regression and Natural Spline Method," *2018 International Conference on Applied Science and Technology (ICAST),* pp. 383-387, 2018.

[65] A. S. Mohan, A. Asfia and A. S. Aneeta, "A comparison of Regression Models for Prediction of Graduate Admissions," *2019 International Conference on Computational Intelligence in Data Science (ICCIDS),* pp. 1-5, 2019.

[66] A. D. Nwanze and D. N. Ndubuife, "Population Forecasting System Using Machine Learning Algorithm," *International Journal of Computer Trends and Technology,* vol. 68, no. 12, 40-43, December 2020, pp. 40-43, 2020.

[67] Y. A. Hiyam and W. A. M. Sondos, "Rainfall Prediction Using Multiple Linear Regressions Model," *2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE),* no. 17 May 2021, pp. 1-5, 2020.

[68] A. Mustafa and Y. Nejat, "Estimating Household Natural Gas Consumption with Multiple Regression: Effect of Cycle," *2013 International Conference on Electronics, Computer and Computation (ICECCO),* pp. 188-191, 2013.

[69] E. Sreehari and S. Satyajee, "Prediction of Climate Variable using Multiple Linear Regression," *2018 4th International Conference on Computing Communication and Automation (ICCCA),* no. 29 July 2019, pp. 1-4, 2018.

[70] A. W. Mukhammad, A. Mochammad, W. S. Ananto and U. Fitri, "Comparative Study based on Error Calculation in Multiple Linear Regression Coefficient for Forest Fires Prediction," *2018 International Conference on Sustainable Information Engineering and Technology (SIET),* no. 18 April 2019, pp. 115-120, 2018.

[71] M. Kavita and M. Pratistha, "Crop Yield Estimation in India Using Machine Learning," *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA) Galgotias University, Greater Noida, UP, India. Oct 30-31, 2020,* no. 10 November 2020, pp. 220-224, 2020.

[72] W. Zhang and Y. Bai, "Prediction of water consumption using NARX neural network based on grey relational analysis," *In 2018 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC) ,* pp. 471-475, 2018.

[73] S. Chaudhari and R. C. Biradar, "Resource prediction based routing using wavelet neural network in mobile ad-hoc networks.," *In International Conference on Circuits, Communication, Control and Computing,* pp. 273-276, 2014.

[74] Z. Wang and H. Sheng, "Rainfall prediction using generalized regression neural network: case study Zhengzhou," *In 2010 International conference on computational and information sciences,* pp. 1265-1268, 2010.

[75] A. D. Turgay and T. S. Anil, "Comparison of Statistical Methods for Predicting Wheat Yield Trends in Turkey," *2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics),* pp. 1-4, 2018.

[76] S.Bhanumathi, M.Vineeth and N.Rohit, "Crop Yield Prediction and Efficient use of Fertilizers," *In 2019 International Conference on Communication and Signal Processing (ICCSP),* pp. 769-773, 2019.

[77] Y.J.N.Kumar, V.Spandana, V.S.Vaishnavi, K.Neha and V.G.R.R.Devi, "Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector," *In 2020 5th International Conference on Communication and Electronics Systems (ICCES),* pp. 736-741, 2020.

[78] R. Neha, S. Raxitkumar, B. Doina, A. James and B. Wolfgang, "Prediction of Crop Cultivation," *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC),* pp. 227-232, 2019.

[79] Kumar, Y.J.N., Spandana, V., Vaishnavi, V.S., Neha, K. Devi and V.G.R.R, "Demand Based Crop Recommender System for Farmers," *2017 IEEE International Conference on Technological Innovations in ICT For Agriculture and Rural Development (TIAR 2017),* no. 01 February 2018, pp. 194-199, 2017.

[80] S. M. Tahmid, R. Karishma, R. N. Sumaiya and C. Amitabha, "Agricultural Production Output Prediction Using Supervised Machine Learning Techniques," *2017 1st International Conference on Next Generation Computing Applications (NextComp),* pp. 182-187, 2017.

[81] J. Knodel, M. Lindvall, D. Muthig and M. Naab, "Static Evaluation of Software Architectures," *In Conference on Software Maintenance and Reengineering (CSMR'06) ,* 2006.

[82] Osborne, W. Jason and W. Elaine, "Four assumptions of multiple regression that researchers should always test," *Practical assessment, research, and evaluation 8.1 (2002),* 2002.

[83] E. C. S. Agency, "The 1984 Population and Housing Census of Ethiopia analytical Report At National Level," 1991.

[84] E. C. S. Agency, "The 1994 Population and Housing Census of Ethiopia Results at Country level volume II Analytical Report," 1999.

[85] A. V. Pashentsev and V. V. Vedishchev, "Applying Big Data and Machine Learning Approach to Identify Noised Data," *Central Michigan University. Downloaded on May 14,2021 at 12:02:26 UTC from IEEE Xplore. Restrictions apply.,* 2020.

[86] S. P. Priyanka and V. D. Nagaraj, "Analysis of Banking Data Using Machine Learning," *International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017),* 2017.

[87] K. A. Mottaleb and R. B. Dil, "Household Production and Consumption Patterns of Teff in Ethiopia," *International Maize and Wheat Improvement Center (CIMMYT), Socioeconomics Program, CarreteraMéxico-Veracruz Km. 45, El Batán, Texcoco, C.P., 56237, México,* no. 7 November 2017, 2017.

[88] M. Ramesh, S. R. Vijay and B. Rashmi, "Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning," *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA),* no. 13 September 2018, pp. 1-6, 2017.

Annex A

Year	1984	1985	1986	1987	1988	1989
Population	42,616,876	43,563,823	44,531,811	45,521,308	46,532,791	47,566,750
Year	1990	1991	1992	1993	1994	1995
Population	48,623,683	49,704,101	50,808,526	51,937,491	53,477,265	54,665,530
Year	1996	1997	1998	1999	2000	2001
Population	55,880,198	57,121,856	58,391,104	59,688,554	61,014,834	62,370,584
Year	2002	2003	2004	2005	2006	2007
Population	63,756,458	65,173,126	66,621,273	68,101,598	69,614,816	73,750,932
Year	2008	2009	2010	2011	2012	2013
Population	75,762,000	77,744,000	79,788,000	81,888,000	84,040,000	86,222,000
Year	2014	2015	2016	2017	2018	2019
Population	88,434,000	90,668,000	92,931,000	95,223,000	97,540,000	99,880,000
Year	2020	2021	2022	2023	2024	2025
Population	102,235,000	104,606,000	106,983,000	109,372,000	111,768,000	114,166,000
Year	2026	2027	2028	2029	2030	2031
Population	116,567,000	118,959,000	121,349,000	123,739,000	126,125,000	128,504,000
Year	2032	2033	2034	2035	2036	2037
Population	130,875,000	133,240,000	135,600,000	137,953,000	140,300,000	142,577,000

Annex B

Year	1984	1985	1986	1987	1988	1989
Teff_Production	13,163,176.62	14,140,083.62	15,187,668.33	16,311,406.00	17,523,750	18,747,749.97
Year	1990	1991	1992	1993	1994	1995
Teff_Production	20,064,563.56	21,489,218.19	17,175,620	17,369,240	16,273,155	17,419,933
Year	1996	1997	1998	1999	2000	2001
Teff_Production	16,773,480	20,255,214	21,755,977	24,377,495	29,935,361	30,272,263
Year	2002	2003	2004	2005	2006	2007
Teff_Production	31,788,757	34,834,826.26	34,976,895	37,652,411.66	44,186,421.95	47,506,572.79
Year	2008	2009	2010	2011	2012	2013
Teff_Production	44,713,786.91	50,204,400.47	52,834,011.56	54,034,790.51	57,357,101.87	61,407,241.25
Year	2014	2015	2016	2017	2018	2019
Teff_Production	65,730,459.33	70,353,415.85	72,085,719.62	73,838,158.52	75,607,431.48	77,391,789.65
Year	2020	2021	2022	2023	2024	2025
Teff_Production	79,184,779.60	80,986,972.48	82,790,323.05	84,599,282.79	86,409,684.77	88,217,344.59
Year	2026	2027	2028	2029	2030	2031
Teff_Production	90,022,751.18	91,816,207.51	93,602,853.25	95,383,895.33	97,155,824.35	98,915,886.97
Year	2032	2033	2034	2035	2036	2037
Teff_Production	100,662,866.12	102,397,861.13	104,121,157.70	105,830,661.74	107,526,589.10	109,158,287.08

Annex C

Scarcity	Consup	Pop	Year	Male	Female	ABR	ADR	Migrant	Married	Divorced	Widow	Disable	Prod	Area	YieldPH
-46500449.78	59663626.4	42,616,876	1984	21436942	21179934	49.3	21.5	11,062,640	12,973,320	2343928	2045616	89373	13,163,176.62	1,950,100.24	6.75
-29423739.38	43563823	43,563,823	1985	21995374	21568449	38.14	16.28	5,634,466	14,173,795	2330795.222	2047499.681	95012.69796	14,140,083.62	1,985,966.80	7.12
-36766111.17	51953779.5	44,531,811	1986	22568722	21963089	45.42	21.9	6,679,772	13,635,952	2315921.363	2048463.306	356254.488	15,187,668.33	2,022,326.01	7.51
-47418425.2	63729831.2	45,521,308	1987	23156689	22364619	48.3	10.99	9,196,096	14,892,159	2299235.746	2048458.86	2968581.059	16,311,406.00	2,059,520.96	7.92
-36764506.17	54288256.2	46,532,791	1988	23759643	22773148	44.9	20.1	11,091,416	14,331,960	2206789.694	2047442.804	508231.1433	17,523,750	2,097,400	8.35
-36746791.7	55494541.7	47,566,750	1989	24377959	23188791	41.2	19.3	10,592,126	16,557,367	2184617.154	2045370.25	523519.6505	18,747,749.97	2,135,279.04	8.78
-28559119.44	48623683	48,623,683	1990	25012023	23611660	43.9	19.1	7,455,032	15,062,985	2083174.726	2042194.686	506736.5748	20,064,563.56	2,173,842.18	9.23
-36498899.64	57988117.8	49,704,101	1991	25662227	24041874	42.42	18.5	6,428,639	16,438,488	2055055.819	2037868.141	542868.1911	21,489,218.19	2,213,101.77	9.71
-48687284.07	65862904.1	50,808,526	1992	26328978	24479548	45.45	18	12,392,494	15,830,768	1317099.258	2032341.04	1305677.501	17,175,620	2,123,470	8.09
-34568251	51937491	51,937,491	1993	27012689	24924802	39.99	17.5	11,853,590	16,275,496	3134209.444	2025562.149	467073.8566	17,369,240	2,182,530	7.96
-76454329.52	92727484.2	53,477,265	1994	26910698	26566567	46.42	16.28	6,916,653	17,877,824	2402348.48	2032136.07	512173.1578	16,273,155	1,818,375	8.95
-59111808.85	76531742	54,665,530	1995	28519198	26146332	41.65	21.3	9,056,870	18,372,921	2287107.377	2022624.61	597226.3818	17,419,933	1,851,215	9.41
-39106718	55880198	55,880,198	1996	29145502	26734696	38.46	20.1	5,196,780	19,951,298	1476086.606	2011687.128	572537.3327	16,773,480	1,989,068	8.43
-79708034	99963248	57,121,856	1997	29785558	27336298	44.6	18.3	6,207,798	20,496,864	627894.8655	1999264.96	119784.532	20,255,214	2,135,553	9.48
-59991568.6	81747545.6	58,391,104	1998	30439668	27951436	38.99	5.5	8,467,644	22,175,014	2651843.666	1985297.536	293064.951	21,755,977	2,246,017	9.69
-45259151.33	69636646.3	59,688,554	1999	31108140	28580414	38.7	17.2	9,591,581	22,560,901	1790191.049	1969722.282	671973.7409	24,377,495	2,404,674	10.14
-55485406.6	85420767.6	61,014,834	2000	31791291	29223543	40.1	13.9	7,248,892	22,952,987	2018175.461	1952474.688	347015.7669	29,935,361	2,565,155	11.67
-32098321	62370584	62,370,584	2001	32489442	29881142	40.9	12.9	8,387,796	22,839,297	2156388.097	1933488.104	223935.3448	30,272,263	2,481,333	12.2
-57470284.2	89259041.2	63,756,458	2002	33202922	30553536	39.1	6.5	9,573,733	22,709,222	2400966.198	1912693.74	415207.5571	31,788,757	2,588,661	12.28
-41200487.41	76035313.7	65,173,126	2003	33932069	31241057	39.99	15.4	18,274,962	19,955,164	2551879.818	1890020.654	521385.008	34,834,826.26	2,761,190	12.62
-81610332.75	116587228	66,621,273	2004	34677225	31944048	46.2	21.8	19,725,506	21,730,993	2403082.614	1865395.644	389334.7194	34,976,895	2,731,112	12.81
-57689825.54	95342237.2	68,101,598	2005	35438744	32662854	43.2	11.1	11,293,887	22,894,872	2246412.932	1838743.146	392265.2045	37,652,411.66	2,730,273	13.79
-25428394.05	69614816	69,614,816	2006	36216983	33397833	39.88	10.1	9,015,063	23,194,752	2192114.864	1809985.216	1015332.091	44,186,421.95	3,016,521.90	14.65
-62335240.83	109841814	73,750,932	2007	37217130	36533802	38.14	9.7	12,218,893	25,467,390	1517605	2084871	68102.34812	47,506,572.79	3,016,053.75	15.75
-31048213.09	75762000	75,762,000	2008	38215000	37547000	37.99	11.9	20,128,281	30,015,919	1792650.139	2369001.978	783197.2512	44,713,786.91	2,866,052.99	15.6
-60949194.3	111153595	77,744,000	2009	39155000	38589000	35.9	10.5	17,064,216	31,034,394	1955930.198	2664209.136	797155.8784	50,204,400.47	3,017,914.36	16.64

-58869188.44	111703200	79,788,000	2010	40123000	39665000	40.32	9.99	16,261,958	33,206,728	2253468.442	2973618.972	771486.1296	52,834,011.56	3,023,283.50	17.48
-27853209.49	81888000	81,888,000	2011	41119000	40769000	44.6	9.1	14,213,233	36,537,361	2435365.498	3297547.872	743936.1024	54,034,790.51	3,076,595.02	17.56
-86125824.96	143482927	84,040,000	2012	42135000	41905000	36.6	8.99	12,606,000	37,818,000	3025440	3781800	672320	57,357,101.87	3,101,177.38	18.5
-46370258.75	107777500	86,222,000	2013	44058000	42164000	33.8	15.2	24,085,098	41,386,560	2974917.666	3793768	740164.1368	61,407,241.25	3,157,184.64	19.45
-81659540.67	147390000	88,434,000	2014	45164000	43270000	41.6	12.3	26,089,445	43,332,660	2778454.786	3802662	1598709.852	65,730,459.33	3,214,203.39	20.45
-20314584.15	90668000	90,668,000	2015	45987000	44681000	35.6	14.9	23,453,672	45,442,802	2712913.495	3808056	838932.8704	70,353,415.85	3,272,251.90	21.5
-79197303.63	151283023	92,931,000	2016	47055000	45876000	31.6	10.9	22,582,103	47,385,517	5979589.436	3810171	956464.4382	72,085,719.62	3,331,348.77	22.61
-65028716.48	138866875	95,223,000	2017	48365000	46858000	30.89	9.9	21,646,169	49,506,438	5833341.935	3808920	1040025.606	73,838,158.52	3,391,512.93	23.77
-38189235.19	113796667	97,540,000	2018	49430000	48110000	29.5	15.1	18,628,306	44,380,700	5674409.008	3804060	1122334.256	75,607,431.48	3,452,763.65	24.99
-89074877.02	166466667	99,880,000	2019	50500000	49380000	30.4	8.1	25,927,210	54,934,000	5661018.616	3795440	1157649.152	77,391,789.65	3,515,120.56	26.28
-66865220.4	146050000	102,235,000	2020	51978000	50257000	29.4	9.2	15,809,784	59,296,300	3690806.182	3782695	298730.67	79,184,779.60	3,578,603.73	27.63
-23619027.52	104606000	104,606,000	2021	52989000	51617000	28.4	13.1	29,800,890	56,905,664	3619806.945	3765816	1282344.033	80,986,972.48	3,643,233.31	29.05
-104429927	187220250	106,983,000	2022	53990000	52993000	26.4	9.5	16,047,450	60,552,378	3372061.367	3744405	224343.351	82,790,323.05	3,709,030.10	30.54
-68521517.21	153120800	109,372,000	2023	55028000	54344000	29	7.9	27,184,126	53,373,536	3283631.807	3718648	320153.7184	84,599,282.79	3,776,015.18	32.11
-66997374.05	153407059	111,768,000	2024	56988000	54780000	27.2	8	23,718,041	66,949,032	3010806.384	3688344	1510164.509	86,409,684.77	3,844,210.01	33.76
-102059322.1	190276667	114,166,000	2025	58011000	56155000	28.9	7.2	20,078,169	59,309,237	3246310.21	3653312	913328	88,217,344.59	3,913,636.44	34.5
-45972082.15	135994833	116,567,000	2026	59400000	57167000	26.9	6.4	35,682,744	61,920,390	3674145.213	3613577	1273144.774	90,022,751.18	3,984,316.71	36.28
-59586156.13	151402364	118,959,000	2027	60074000	58885000	26.1	5.9	32,092,045	71,518,151	3927621.719	3568770	2349083.373	91,816,207.51	4,056,273.47	38.15
-79752861.03	173355714	121,349,000	2028	61133000	60216000	28.4	6.9	28,327,031	75,843,125	4380844.519	3519121	970792	93,602,853.25	4,129,529.77	40.11
-121159354.7	216543250	123,739,000	2029	62187000	61552000	28.9	19.1	24,388,313	74,119,661	4652363.67	3464692	3897036.066	95,383,895.33	4,204,109.08	42.17
-49990008.99	147145833	126,125,000	2030	63999000	62126000	25.6	5.8	28,673,283	73,657,000	5131118.15	3657625	4045888.2	97,155,824.35	4,280,035.29	44.34
-105522294.9	204438182	128,504,000	2031	65282000	63222000	26.9	4.99	19,275,600	80,700,512	5420273.019	3855120	4497691.402	98,915,886.97	4,357,332.73	46.62
-78969486.82	179632353	130,875,000	2032	66319000	64556000	25.1	4.8	16,927,137	78,551,175	5923978.35	4057125	4591670.85	100,662,866.1	4,436,026.16	49.02
-76963677.33	179361539	133,240,000	2033	67353000	65887000	25.99	9.3	17,265,719	81,673,455	6230488.936	4263680	5063972.736	102,397,861.1	4,516,140.79	51.54
-54078842.3	158200000	135,600,000	2034	68383000	67217000	40.3	3.89	22,465,910	79,732,800	3550604.64	4474800	1486718.4	104,121,157.7	4,597,702.29	54.19
-32122338.26	137953000	137,953,000	2035	69408000	68545000	25.1	3.5	20,692,950	799,989,45	4966308	4690402	6863630.79	105,830,661.7	4,680,736.79	56.96
-32773410.9	140300000	140,300,000	2036	70685000	69615000	24.9	3	36,419,579	84,194,030	2886953.1	4910500	8218549.52	107,526,589.1	4,765,270.90	59.89
-57181546.26	166339833	142,577,000	2037	71987000	70590000	24.28	2.8	18,475,641	85,546,200	7841735	5132772	9610174.562	109,158,287.1	4,851,331.69	62.97

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Declared By:

Name: Taye Mohammed Kemal_____

Signature: _____

Date: _____

Confirmed by advisor:

Name: _____

Signature: _____

Date: _____