ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNOLOGY TO SUPPORT ADEQUATE CHEMICAL FERTILIZER PREDICTION FOR TEF AND WHEAT PRODUCTION IN SOME SELECTED PARTS OF ETHIOPIA

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

ZEBIBA ALI ABEGAZ

JANUARY 2009

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNOLOGY TO SUPPORT ADEQUATE CHEMICAL FERTILIZER PREDICTION FOR TEF AND WHEAT PRODUCTION IN SOME SELECTED PARTS OF ETHIOPIA

By

ZEBIBA ALI ABEGAZ

JANUARY 2009

Name and Signature of Members of the Examining Board

_____ _____

_____ _____

_____ _____

Dedication

This thesis is dedicated to my parents, Ali Abegaz and Abebech Yimer and to my brother, Mohammed Ali, the most caring and loving.

Acknowledgment

Most of all, my gratitude goes to the MOST HIGH, ALLAH, who guides me throughout my life. I am very much grateful to my advisor Dr. Manoj V.N.V for his countless revision, priceless suggestions and guidance to profile my thesis. Had he not provided his support, this work would not have been fruitful within this short time.

My thank goes to the Department of Information Science providing me budget to cover the cost of this thesis work. In addition, I found the Department staff and librarian very sociable and thank you all for giving me such family atmosphere during my study.

I am also very much indebted to Ato Yesuf Assen, expert in National soil Testing Center, who devoted his precious time and expertise and lent me his hand at all steps of my research work. His provision of actual research data, which have been accumulated from various locations, for this experiment, is highly acknowledged.

My gratitude goes to Ato Tibebe Beshah who at present teaches in Addis Abeba University and Kumneger Fikra. They provided me their unreserved technical help during this research work. I am also very much grateful to all my friends who were with me whenever I need them.

Last but not least, I would like to extend my honor and prize to the members of my family and, most importantly, my brother, father and mother, whom I love the most. They all have raised me to this height and depth. Especially, my dad always shows me the hard work to believe in myself. Mam and my brother encourage me to improve myself from my childhood to this level. I wish them enjoyable and long life. I am also very much grateful to Shitaye, my uncle (Ato Eshetu), his family and Meki my friend for their friendship, financial and moral support during my stay in Addis Ababa for this study.

Zebiba Ali

Table of Content

Page No.

List of Figures

Page No.

List of Tables

Page No.

List of Acronyms

ARFF:-Attribute-relation file Format

CART: - Classification and regression tree

CHAID: - Chi-squared Automatic interaction detector

CRISP-DM:-Cross industry standard process for data mining

CSV:-Comma-separated value

DNA: - Deoxyribonucleic acid

FAIS: - financial crimes enforcement Network (FINCEN) AI system

GDP: - Gross domestic product

HNC: - Hecht-Nielsen Neurocomputer corp

KDD: - Knowledge discovery in databases

NSTC: - National soil testing center

OECD: - Organization of economic cooperation and development

OLAP: - Online Analytical processing

OLTP: - online transaction processing

PRISM: - Proactive Risk management system

SKICAT: - Sky Image Cataloging and Analysis Tool

SQL: - Standard query language

TASA: - telecommunications alarm-sequence analyzer

WEKA: - Waikato environment for knowledge analysis

Abstract

Though agriculture is the mainstay of Ethiopian economy, it is suffering from many disasters. Among these disasters, nutrient depletion of the soil is the major one. Applying organic and/or inorganic (chemical) fertilizer in the soil can curb nutrient depletion. Scarcity of organic fertilizer, here in Ethiopia, brings about the need to use chemical fertilizer. But, still, there is a problem of using sufficient amount of chemical fertilizer based on initial fertility status of the soil and nutrient requirement of crops to bring high yield. Hence, this and others arouse interest of developing a guideline for fertilizer recommendation.

This thesis developed a decision support system that can help agricultural researcher in the process of building a guideline for fertilizer recommendation. In doing so, the research aimed to assess the potential applicability of data mining technology specifically decision tree technique to help in fertilizer-grain yield data analysis in decision-making process.

In this research, in the process of building a model, different steps were undertaken. Among the steps, data collection, data preprocessing and model building and validation were the major ones. Different tasks performed in each step are mentioned as follows. The data were collected from National Soil Testing Center. Under preprocessing, data cleaning, discritization and attribute selection were done. The final step was model building and validation and it was performed using the selected tools and techniques.

The data mining tool used in this research was Weka. In this software the decision tree J48 algorithm was selected since it is capable to analyze numeric data. After successive experiments were done using this software, a model that can classify crop yield as high, medium and low with better accuracy to the extent of 85% and sound rule was selected. Experimental results show that decision tree is a very helpful tool to depict the contribution of soil-pH, initial available soil phosphorus, organic-matter, total nitrogen and treatment to bring high tef and wheat yield. The reported findings are optimistic, making the proposed model a useful tool in the decision making process. Eventually, the whole research process can be a good input for further in-depth research.

Chapter One: Introduction

1.1 Background of the Problem

Ethiopia, which lies between 20^0 and 45^0 latitude in mountainous region, is the origin of several crop plants. Ethiopia's major staple crops include a variety of cereals, pulses, oilseeds, and coffee. Grains are the most important field crops and the chief element in the diet of most Ethiopians. The principal grains are tef, wheat, barley, corn, sorghum and millet. The first three are primarily cool-weather crops cultivated at altitudes generally l,500 meters above see level [1]. Tef, indigenous to Ethiopia, furnishes the flour for injera, unleavened bread that is the principal form in which grain is consumed in the highlands and in urban centers throughout the country [1].

Agriculture becomes the mainstay of the Ethiopian economy. It accounts for about 50% of the GDP, provides employment for about 85% of the total working labor force and accounts for 90% of the total foreign exchange earnings [15]. Besides, many other economic activities depend on agriculture, including marketing, processing and export of agricultural products. Agricultural production is overwhelmingly of a subsistence nature and a large part of commodity exports are provided by the small agricultural cash-crop sector. Exports in Ethiopia are almost entirely agricultural commodities.

Ethiopian agricultural, however, faces various obstacles that hinder its development. The major obstacles are fragmentation of land holdings, limited infrastructure, shortage of skilled manpower and technological backwardness. Low agricultural production in Ethiopia has also been attributed to the lack of security of tenure, land fragmentation, soil erosion, and draught power, limited supplies of chemical inputs, adverse weather conditions, improper sectorial macro economic policies and inefficient marketing practices [2]**.** Despite much effort to increase productivity through introduction of improved inputs over the past three decades, average output per hectare has not shown significant increase to warrant optimism about the sector in the foreseeable future [40].

As it is mentioned, one of the factors that affect the development of agriculture negatively is limited supply of chemical input. It is further strengthened by the following idea: [30] pointed out that in Ethiopia, wheat and Tef are grown as a mono-crop on a given plot of land. This causes depletion and imbalance of plant nutrients in the soil and this is aggravated due to absence of organic fertilizer and low or unbalanced amount of chemical fertilizer application.

In Ethiopia, organic sources as fertilizer are scarce since they are highly used as a fuel for source of energy and feed for animals. Therefore, as soil fertility amendment, the use of chemical fertilizer is becoming indispensable. Recognizant to this fact, the Ethiopian Government is importing huge amount of chemical fertilizer from the scarce hard currency the country has in hand [30]. This importation is indicating an increasing trend from time to time to achieve the country's attempt for food self-sufficiency [5].

1.2 Gap in Chemical Fertilizer Use in Ethiopia

As far as the use of chemical fertilizer is concerned, because of different reasons, majority of Ethiopian farmers were illiterate in terms of fertilizer application based on soil testing. Now, agricultural researchers are working hard to reach the knowledge to different parts of Ethiopia and to recommend adequate chemical fertilizer to a given plot of land based on initial soil fertility status and crop nutrient requirements. This is very crucial, because, on the one hand, if the farmer applies low amount of chemical fertilizer, s/he will not be able to exploit the genetic potential of the crop to attain the possible maximum yield [39]. On the other hand, if s/he applies more than enough, the yield will reduces because of various reasons for instances due to lodging and will also cause environmental pollution [39]. Studies suggest that variation in soil fertility results in differences on nutrient requirements of crops growing on that specific soil. In this connection, if the farmer do not know the fertility status of the soil to determine the type and amount of fertilizer to be applied, then, the probability for the above problem to occur will be high, becoming a disaster in maximizing his/her benefit.

1.3 Fertilizer Management Data Analysis by National Soil Testing Center

National soil testing center in Ethiopia is one of the federal organizations in the country and established with the aim of doing experiments in order to develop a guideline for fertilizer recommendation on the basis of soil test results. To make this aim come true, a lot of field experimental data has to be generated by conducting sufficient numbers of soil test based field correlation experiments [30]. In this connection, the National soil testing center in Ethiopia collected, recorded and stored tremendous data in Microsoft Excel sheet and hard copies. The data is about where the selected sites are located (in which wereda and region of Ethiopia), to whom they belong to and what kind of crop has been planted (sown). Besides, data regarding the soil pH, the initial available soil phosphorus, total nitrogen, organic matter of the soil, condition of crops after they are planted, the amount of yield and the likes are all recorded. Agricultural researchers analyze this recorded data in different region and in different crop type by the help of statistical tool to know only the contribution of the added fertilizer on the amount of yield gained without taking into account the different soil properties mentioned above and others.

However, analyzing this huge dataset using the statistical tool, which isn't capable to show the relationship among more than two variables, is a tiresome job and prone to errors. Therefore, supporting the analysis with an automated system that is more accurate, less subjective and requiring less expertise is worthwhile.

Thus, through this research work an attempt has been made to apply data mining tools and techniques in analyzing and determining interesting patterns with respect to fertilizer application to get high yield.

Data mining incorporates steps of selection, exploration, and modeling of large quantities of data to discover previously unknown regularities or relations with the aim of obtaining clear and useful results for the owner of data in any organization Giudici (2003) as cited by [13].

1.4 Statement of the Problem

Ethiopia is the home of many agricultural products that determine the economic status of the country. Among these agricultural products wheat, tef and coffee can be mentioned in the first place. In many parts of Ethiopia, farmers usually grow wheat and tef on a given plot of land. This farming practice, through time, has caused depletion and imbalance of plant nutrients in the soils. Consequently, farmer's production is declined to the extent of not feeding himself. This again has negative impact in the economy of the country.

To alleviate this problem, considering the scarcity of organic fertilizer, scholars suggest the application of chemical fertilizers to boost crop production of the country.To achieve the objective of getting high yield using chemical fertilizer, the first step was to create awareness about chemical fertilizer. As a result, most farmers understand the role of chemical fertilizer and begin to use it.

Although chemical fertilizer is applicable in many parts of Ethiopia, the response of wheat and tef varies from place to place due to mainly variation in initial fertility status of the soil [30]. The researchers further investigate that putting any amount of fertilizer on the soil may bring yield reduction due to lodging, high production cost and less benefit. Hence, farmers should be in a position to supply adequate fertilizer by considering factors like initial fertility status of the soil.

To alleviate the problem related to inappropriate use of chemical fertilizer and to recommend adequate amount of fertilizer, Agricultural researchers begin to do field experiments. In the process, representative farmers' fields in different wereda are selected and soil sample from these selected farms are collected to test the initial soil fertility status. Then, different levels of nitrogen (N), phosphorus (P) and potassium (K) fertilizers as treatments together with the seed are broadcasted during planting. These specific treatments are shown in the following Table.

Table 1.1 Description of treatments

Treatment combinations N/ P/K (kg/ha)		
T1 = 0/0/0	T7 = 46/20/0	T13 = 138/0/0
T2 = 0/10/0	T8 = 46/40/0	T14 = 138/10/0
T3 = 0/20/0	T9 = 92/0/0	T15 = 138/20/0
T4 = 0/40/0	T10 = 92/10/0	T16 = 138/40/0
T5 = 46/0/0	T11 = 92/20/0	T17 = 46/20/25
T6 = 46/10/0	T12 = 92/40/0	T18 = 92/40/50

The above mentioned process is done on different farmers' fields in two replications (repetitions) to make the experiment reliable. All plots are equally treated as far as weeding and other agronomic parameters (Like weather) are concerned. After sometime, agricultural researchers begin to observe the status of the crop. All the above experimental information are recorded and kept in the form of table for final analysis. During the time of analysis, a statistical tool by the name MSTAT-C is used only to see the relationship between fertilizer and crop production and the analysis is specific to each wereda. Generally, the statistical tool is incapable of incorporating and showing the relationship of many determinant factors that affect the status of crop production. Moreover, according to the domain experts, this kind of experiment, which is uneconomical and tiresome, should be done through out the country unless and other wise there is a mechanism to study the historical data that they have at hand and to predict and extrapolate the reliable amount of fertilizer to be applied for a given fertility of the soil. Hence, to support this tiresome job, information technology particularly data mining will play a very important role.

The major problems to be addressed are the following:

- Identifying the effect of determinant variables like soil pH, initial phosphorous, organic matter and nitrogen on crop production
- Identifying the amount of chemical fertilizer that go together with the available initial fertility of the soil to get high or medium yield.
- Detecting the best combination of parameters that can generate high or medium yield.

This research is therefore, to avoid this gap by applying data mining techniques with the objective of analyzing the historical data by classifying amount of crop yield as high, medium, low and predicting adequate fertilizer level for a certain initial soil fertility. Eventually, using the information generated so far, it will be easy to improve production and productivity for similar agro-ecologies through extrapolation without repeating the experiment.

1.5 Objective of the Study

1.5.1 General Objective

The general objective of this study is to explore the potential applicability of data mining technology in developing a model that can support agricultural researchers in predicting the amount of adequate chemical fertilizer based on the initial soil fertility status of the soil to get high production of wheat and tef.

1.5.2 Specific Objectives

The specific objectives of this study are:

- To collect the valuable data needed to do the research from national soil testing center
- To prepare the data for analysis using different preprocessing techniques, this involves extracting the data, dealing with missing values and transforming into the format required for the data mining algorithm
- To select the data mining tools and algorithms to develop the model
- To design and develop the model
- To evaluate (test) the model
- To analyze the outcome of the research and make recommendations based on findings

1.6 Research Methodology

The research methods that were applied are adapted from CRISP-DM model. CRISP-DM is the industry standard methodology for data mining and predictive analytics [10]. The research follows the necessary data mining steps that are helpful to develop decision support system. These steps will be mentioned as follows:

1.6.1 Review of Related Literature

Books, research papers and different materials were reviewed to gain insight into the subject matters of data mining technology, principal crop of Ethiopia like wheat and tef production, chemical fertilizers and soil properties that indicate fertility of the soil.

1.6.2 Business Understanding

Interviews, observations and document review were made to assess the needs of users, analyze the problems of the organization, and have good background knowledge in interpreting results of the data mining process.

1.6.3 Data Understanding

In this study, the data were collected from National Soil Testing Center. The data sources contain factors that affect the magnitude of crop yield produced. Some of the factors are initial fertility status of the soil such as soil pH, soil clay content, soil organic matter content, soil nitrogen content, kind and amount of chemical fertilizer applied and so on. The data also contain the amount of grain and straw yield of wheat and tef. More than 5600 data records for both wheat and tef were collected for this research. Almost all of the attributes of the records for both wheat and tef are numerical in nature. Many of the records were registered electronically in Ms-excel format though they were kept in separate worksheet of excel. But, there were also some that were kept in hard copies.

1.6.4 Preparing the Data for Analysis

The collected data was preprocessed before developing the model by removing attributes that are irrelevant for the objective of the research. The data kept in different hard copies and worksheet was integrated to form unified data. Moreover the data was edited, transformed and discretized before it was subjected to analysis.

1.6.5 Training and Building Model

After all the necessary steps in data preparation have been completed, the filtered data were used to build (train), test and validate classification and prediction models. Explanations (rules) of the dependent in terms of the independent (input) variables were also done.

1.6.6 Evaluation of the Models

The models were evaluated for their predictive performance. Measures of accuracy and confusion matrix were used to evaluate the models. Apart from this, the selected best model was evaluated whether it achieves the objective of the organization that it is supposed to benefit or not.

1.7 Application of Results

This research was conducted for the following applications: It will

- enable agricultural researcher to predict adequate amount of chemical fertilizer given the initial fertility status of the soil very easily
- curb the problem of scarcity of agricultural researchers. Once the model is developed, even junior staffs can use it
- contribute towards the successful achievement of one of the programs of the government i.e. food security by increasing grain yield production per unit area through appropriate use of inputs
- help researchers as a springboard for further study
- enable farmers to get improved services and their profitability will increase
- amplify the importance of chemical fertilizer to bring high production
- enable to minimize the potential risk of high amount of chemical fertilizer there by potentially saving economical risk of the farmer and the government

1.8 Scope and Limitation of the Study

The scope of this research is limited to assess the possible application of data mining technology in fertilizer data analysis at National Soil Testing Center; it is limited to investigate the impact of initial fertility status such as soil-pH, initial phosphate and nitrogen together with treatment upon the amount of grain yield. To analyze this, classification model particularly decision tree technique is selected.

Regarding the limitation that the researcher faced, there was time constraint to incorporate as many experiments as possible in the process of building models. Besides, there was absence of sufficient data for analysis. Above all, it wasn't easy to meet and discuss with agricultural experts because of frequent field travels. Any how, we tried to communicate through telephone, e-mails and using some office hour contact to handle the problem.

1.9 Thesis Organization

This thesis report is organized under six chapters. The first chapter deals with the general overview of the study including background, statement of the problem, objectives, methodology and the like of the research.

The second chapter is devoted to literature review of data mining technology. Available tools and techniques in the area are reviewed with the emphasis on the tools and techniques employed in this specific research. Assessment of the application areas of data mining in general and its application in agricultural data in particular is discussed in this chapter.

Under chapter three, an attempt has been made to review literatures and trends in status of Ethiopian agriculture, fertilizer management and data analysis system of the existing system

The next section, chapter four, deals with different pre-processing steps. Tasks like data collection and cleaning, attribute selection, data formatting and transformation are reported in detail.

Chapter five reports the experiment of the research. The experiment basically comprises training; building and validation of the models in addition to analysis and interpretation of the results. The last chapter presents conclusions and recommendations of the research. Next to this, reference as well as the appendix part that includes rules generated from different experiments and the results of the experiments in the form of screen shot are attached.

Chapter Two: Data Mining

In this chapter literatures related to concepts of data mining technology, tools and techniques available and the application area of data mining have been enlightened.

2.1 Overview

Exponential growth in the use of information technology by organizations has resulted in the availability of a tremendous volume of data to knowledge workers. The Internet, intranet, enterprise and groupware systems, and data warehouses have all been contributors to this increase in data volume, availability and importance [18]. This has generated the need to better explore and manage the knowledge hidden in mountains of data.

According to Thearling, 2003 as cited by [33] through evolutionary development of information technology, data mining comes into existence. In the early 1960s, the main concern was data collection for the purpose of retrospective and static data delivery. In 1980s, data was accessed with the aim of retrospective and dynamic data delivery at record level was practiced. Dynamic data delivery at multiple levels was the main feature of data warehousing and decision support in 1990s. Nowadays, data mining comes up with prospective and proactive information delivery.

Defining a scientific discipline, like data mining, is always a controversial task; researchers often disagree about the precise range and limits of their field of study. Hence, different scholars define data mining in various ways:

Data mining is defined as the process of discovering meaningful patterns and relationships through the automated analysis and classification of large stores of historical data [27]. The paper further states that any business or academic pursuit that collects and studies large quantities of data is a candidate for data mining.

Besides, [19] define data mining as the analysis of often large observational data sets, which have already been collected for some other purpose, to find unsuspected relationships and to

summarize the data in novel ways that are both understandable and useful to the data owner. As it is inferred from the definition, data mining has no role in the data collection process. And this causes data mining to be referred as “secondary” data analysis.

According to [19] data mining is an interdisciplinary exercise. It is a discipline lying at the intersection of statistics, database technology, machine learning, pattern recognition, artificial intelligence and other areas. It is difficult to define sharp boundaries between these discipline and data mining. At the boundaries, one person’s data mining is another’s statistics, database or machine learning problem.

[6] say that no data mining algorithms were first invented with the intention of commercial application. The commercial data miner employs a grab bag of techniques borrowed from statistics, computer science and machine learning research.

2.2 Data Mining and KDD (Knowledge Discovery in Databases) Process

Many people consider data mining as a synonym of knowledge discovery in databases (KDD) though others treat it as an essential step in the process of KDD [18].

Historically, data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing are different names given to the notion of finding useful patterns in large data sets. The term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. The phrase knowledge discovery in databases was coined at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the Artificial intelligence (AI) and machine-learning fields [11].[11] pointed out that KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data. The additional steps in the KDD process, such as data preparation, data selection, data cleaning incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to

ensure that useful knowledge is derived from the data. Blind application of data-mining method that is considered as data dredging can be a dangerous activity and it is easily leading to the discovery of meaningless and invalid patterns.

In different literatures, different partitions of KDD process are indicated. According to [19], which is originated in the artificial intelligence (AI) research field, involves several stages: selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures. The researchers, however, state that it is still difficult to define the precise boundary of data mining. To many people data transformation is an intrinsic part of data mining.

In a more similar but detailed fashion [18] states that KDD incorporates iterative sequence of steps. These are:

- Data cleaning: It is useful to remove noise and inconsistent data.
- Data integration: In this step data from different sources will be merged and their consistency will be kept
- Data selection: The whole data in the database are not necessarily relevant to the analysis. The real data in which data mining is performed should be selected.
- Data transformation: If it is necessary, data are transformed or consolidated into forms appropriate for mining by performing operations like summary or aggregation.
- Data mining: It is an essential step that applies intelligent methods to extract data patterns.
- Pattern evaluation: This step is useful to identify truly interesting pattern based on some interestingness measures.
- Knowledge presentation: It is a way of presenting the mined interesting knowledge to the user.

[11] summarize all about KDD field as it is concerned with the development of methods and techniques for making sense of data. The KDD process addressed basic problems. The one is

mapping low-level, voluminous data into more compact and digestive. The other is developing a descriptive approximation or model of the process that generated the data. The last but not the least problem addressed by KDD is developing a predictive model for estimating the value of future cases. At the core of the process is the application of specific data-mining methods for pattern discovery and extraction.

2.3 Data Mining and Data Warehouse

Data mining can be performed in different kinds of data stores like relational databases, data warehouse, transactional databases, advanced databases, flat files and the World Wide Web. However, the challenges and techniques of mining may differ for each of the storage systems [18]. As it is explained, data warehouse is a data storage system in which data mining is performed. [27] also states that a data warehouse provides a central storage facility for an organization's data from which users can access data whenever they need it. And also most major institutions and corporations use data warehouse model to process huge amount of data.

Data warehouse is a storage system but data warehousing is a process of gathering data from different sources- billing records, scanner, registration forms, applications, call records, coupon redemptions and surveys and organizing it in a consistent and useful way for learning to take place [6]. Data warehousing has another virtue that it separates the reporting data on which data mining is performed from the online transaction processing (OLTP) data

Regarding the relationship between data warehouse and data mining, [27] pointed out that data warehouse do a good job of organizing and structuring data but we need data mining to extract useful information from the underlying unmanageable and too general data. Moreover, [6] depict that data warehouse provides the enterprise with a memory. But, memory is of little use without intelligence that allows us to comb through our memories, noticing patterns, devising rules, coming up with new ideas, figuring out the right questions, and making predictions about the future.

2.4 Data Mining Vs OLAP (On Line Analytical Processing)

Both data mining and OLAP serve as decision support tools, but each is designed for a different use. OLAP is primarily designed to store data in a summarized table to facilitate retrieval and navigation of this data by end users. In OLAP, in most cases the user is navigating through dimensions that contain meaning and relationships that are already well known. OLAP could also be used to try to discover new data, but since the data discovery is really being done by the end user, with the assistance of an OLAP tool, the data discovery is bound to be haphazard and incomplete. Data mining, however, is less concerned with allowing an end user to easily browse summary data. It is rather concerned with automatic discovery of new patterns and rules that can be applied to get future results. As a result of this difference, OLAP is an efficient storage and retrieval mechanism and data mining is a knowledge discovery tool [27].

Moreover, [6] pointed out that OLAP is a significant improvement over ad hoc query systems, because OLAP systems design the data structure with users in mind. This powerful and efficient representation is called a cube, which is ideally suited for slicing and dicing data. The cube itself is stored either in a relational database, typically using a star schema or in a special multidimensional database that optimizes OLAP operations. In addition, OLAP tools provide handy analysis functions that are difficult or impossible to express in SQL. If OLAP tools have one downside, it is that business users start to focus only on the dimensions of data represented by the tool. Data mining, on the other hand, is particularly valuable for creative thinking. They continue saying, OLAP and data mining complement each other. Data mining can help build better cubes by defining appropriate dimensions, and further by determining how to break up continuous values on dimensions. OLAP provides a powerful visualization capability to help users better understand the results of data mining, such as clustering neural networks. Used together, OLAP and data mining reinforce each other's strengths and provide more opportunities for exploiting data.

2.5 Data Mining and Statistical Applications

Data mining is a data-driven process that discovers meaningful patterns previously unseen or otherwise prone to be overlooked, while statistical inference begins with a hypothesis conceived

by a person, who then applies statistical methods to prove or disprove the thesis. In case of data mining it is the machine, not the operator, does the complex mathematics used to build the predictive model. Computers do the high level inductive reasoning required to analyze large quantities of raw data and can then output the results in a format that the owner understand only by analyzing large sets of cases can accurate predictions be made. Pure statistical analysis requires the statistician (the operator) to perform a high degree of directed interaction with the data sets, which interferes with the potential for making new discoveries. Discovery, in the world of data mining, is the process of looking at a database to find hidden patterns in the data. The process does not take preconceived ideas or even a hypothesis to the data. In other words, the program uses its own computational ability to find patterns, without any user direction. The computer is also able to find many more patterns than a human could imagine. What distinguishes data mining from the science of statistics is the machine [27].

Besides, [19] state that data mining is considered as a secondary analysis since the data on which mining is performed were originally collected for some other purpose. In contrast, much statistical work is concerned with primary analysis: the data are collected with particular questions in mind, and then are analyzed to answer those questions. In statistical analysis, the entire domain expertises are concerned with the best ways to collect data in order to answer specific questions.

Above all, the most fundamental difference between classical statistical applications and data mining is the size of the data set. To a conventional statistician, a “large” data set may contain a few hundred or a thousand data points. To someone concerned with data mining, however, many millions or even billions of data points is not unexpected [19].

[19] on the other hand, show how statistical techniques complement data mining. Statistical techniques alone may not be sufficient to address some of the more challenging issues in data mining, especially those arising from massive data sets. Nonetheless, statistics plays a very important role in data mining: it is a necessary component in any data mining enterprise. To undertake large data analysis projects, researchers have adapted established algorithms from statistics and others.

In summary, while data mining does overlap considerably with the standard exploratory data analysis techniques of statistics, it also runs into new problems, many of which are consequences of size and the non traditional nature of the data sets involved [19].

2.6 Functionalities of Data Mining

Many problems of intellectual, economic and business interest can be phrased in terms of the following six tasks: Classification, estimation, prediction, affinity grouping, clustering and description and profiling. These again can be clustered as directed data mining or undirected data mining. Directed data mining attempts to explain or categorize some particular target field. Hence, classification, estimation and prediction are categorized under this data-mining flavor. Undirected data mining, on the other hand, attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes. Affinity grouping and clustering are examples of undirected data mining. Description and profiling tasks may be either directed or undirected [6]. [37] also identified the most widely used data mining tasks as classification, association, clustering, dependency analysis, prediction, segmentation, description.

On the other hand, [18] with many scholars in the field of information science agree that data mining tasks are broadly categorized as predictive and descriptive.

2.6.1 Predictive Data Mining

It performs inference on the current data in order to made predictions [18]. [11] on their part say that predictive data mining involves using some variables or fields in the database to predict unknown or future values of other variables of interest.

Similarly, [16] states that predictive methods aim to describe one or more of the variables in relation to all the others; they are also called asymmetrical, supervised or direct methods. This is done by looking for rules of classification or prediction based on the data. These rules help us to predict or classify the future result of one or more response or target variables in relation to what happens to the explanatory or input variables. The main methods of this type are those

developed in the field of machine learning such as the neural networks (multilayer perceptions) and decision trees but also classic statistical models such as linear and logistic regression models.

2.6.2 Descriptive Data Mining

It characterizes the general properties of the data in the database [18]. [11] clarify by saying descriptive data mining focuses on finding human-interpretable patterns describing the data.

Similarly, [16] states that descriptive methods aim to describe groups of data more briefly; they are also called symmetrical, unsupervised or indirect methods. Observations may be classified into groups not known beforehand (cluster analysis, Kohonen maps); variables may be connected among themselves according to links unknown beforehand (association methods, log-linear models, graphical models). In this way all the variables available are treated at the same level and there are no hypotheses of causality.

TASKS IN PREDICTIVE MODELING

The goals of prediction and description can be achieved using a variety of particular data mining methods. Classification and Regression are the two most common tasks in predictive modeling. If the label is discrete (containing a fixed set of values), the task is called classification. If the label is a continuous value, the task is called regression [33].

Classification: It is learning a function that maps (classifies) a data item into one of several predefined classes [11]. Similarly, [6] state that classification is one of the most common data mining tasks, consists of examining the features of newly presented object and assigning it to one of a predefined set of classes. The task is to build a model of some kind that can be applied to unclassified data in order to classify it. The following are some of the real life application of classification tasks

- Classifying credit applicants as low, medium or high risk.
- Choosing content to be displayed on a web page.
- Determining which phone numbers correspond to fax machines.

Decision trees and nearest neighbor techniques are well suited for classification. Neural networks and link analysis are also useful for classification in certain circumstances.

[18] point out the differences and similarities of the terms prediction and classification. Prediction can be referred to as classification if it is used for predicting the class label of data objects. However, predicting some missing or unavailable values rather than class labels in numerical data is often specifically referred to as prediction, which is distinct from classification. Prediction also encompasses the identification of distribution trends based on available data.

[29] support the above idea by saying: classification and predictions are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Some of the basic techniques for data classification are decision tree induction and neural networks. These techniques find a set of models that describe the different classes of objects. These models can be used to predict the class of an object for which the class is unknown. The derived model can be represented as rules (If-then), decision tree or other formulae.

Regression: It is learning a function that maps a data item to a real-valued prediction variable [11]. [33] adds that regression is used to deal with non-discreet that means continuous variable. Regression is similar to classification, except that the label is not discrete. [11] mention some application of regression like predicting the amount of biomass present in a forest given remotely sensed microwave measurements and estimating the probability that a patient will survive given the results of a set of diagnostic tests.

TASKS IN DESCRIPTIVE MODELING

Clustering and Association are the two most common tasks in descriptive modeling and are explained below:

Clustering: It is a common descriptive task where one seeks to identify a finite set categories or clusters to describe the data. The categories can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or over lapping categories. Examples of

clustering applications include discovering homogeneous subpopulations for consumers in marketing databases [11].

Association: [18] point out that association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. The discovery of interesting association relationships among huge amounts of business transaction records can help in many business decision making processes, such as catalog design, cross marketing, and loss-leader analysis.

2.7 Data Mining Techniques

According to [6] data mining is largely concerned with building models. A model is simply an algorithm or set of rules that connects a collection of inputs (often in the form of fields in a corporate data base) to a particular target or out come. A model, under the right circumstances, can result in insight by providing an explanation of how out comes of particular interest are related to and predicted by the available facts. Models are created using data mining techniques like regression, neural networks, decision tree and most other. Predictive data mining techniques are briefly mentioned as follows:

2.7.1 Decision Tree

A decision tree is a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. In other words, a decision tree model consists of a set of rules for dividing large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable. The target variable is usually categorical and the decision tree model is used either to calculate the probability that a given record belongs to each of the categories, or to classify the record by assigning it to the most likely class. Decision tree can also be used to estimate the value of a continuous variable although there are other techniques more suitable to that task [6]

Besides, [27] pointed out that decision tree is a very well known algorithm used in one form or another by almost all commercially available data-mining tools. This umbrella term describes a number of specific algorithms, such as Chi-squared Automatic Interaction Detector (CHAID)

and C4.5, which results in models that look like trees. Decision tree algorithms are recommended for predictive tasks that require a classification-oriented model, and as such they are designed for problems best solved by segregating case into discrete groups, for example, decision tree are often used to predict those customers most likely to respond to direct mail marketing or those likely to be approved for loans. To strengthen this idea, [6], in their part, suggest that decision trees are powerful and popular for both classification and prediction. The attractiveness of tree-based methods is due largely to the fact that decision trees represent rules. Rules can readily be expressed in English so that we humans can understand them. They can also be expressed in a data base access language such as SQL to retrieve records in a particular category. Decision tree are also useful for exploring data to gain insight into the relationships of large number of candidate input variables to a target variable. Because decision tree combine both data exploring and modeling, they are a powerful, first step in the modeling process even when building the final model using some other technique.

A decision tree may be painstakingly constructed by hand or it may be grown automatically by applying any one of several decision tree algorithms to a model set comprised of pre-classified data [6].

[27] recommended to use decision trees under the following circumstances.

- When you want to reliably apply the segmentation scheme to a set of data that reflects a group of potential customers.
- When you want to identify possible interactive relationship between variables in a way that would lead you to understand how changing one variable can affect another.
- When you want to provide a visual representation of the relationship between variables in the form of a tree, which is a relatively easy way to understand the nature of the data residing in your database.
- When you want to simplify the mix of attributes and categories to stay with the essential ones needed to make predictions.
- When you want to explore data to identify important variables in a data set that can eventually be used as a target.

2.7.2 Artificial Neural Network

Neural network is the most widely known and the least understood of the major data mining techniques. The way neural network process information is similar to the way our brain process information. The most basic components of the neural networks are modeled after the structure of the brain. Neural network attempts to simulate within specialized hardware and sophisticated software, the multiple layers of simple processing elements called neurons. Each neuron is linked to certain of its neighbors with varying coefficients of its connectivity that represent the strengths of these connections. Learning is accomplished by adjusting these strengths to cause the overall network to output appropriate results [21].

Since neural networks have similarity to the biological brain, its terminologies are borrowed from neuroscience.

The basic element of the human brain is a specific type of cells, which are very large in number and provide us with the abilities to remember, think, and apply previous experiences to our action. These cells are known as neurons; each of these neurons can connect with many other neurons. The power of the brain comes from the numbers of these basic components and the multiple connections between them. All natural neurons have four basic components: dendrites, soma, axon, and synapses. Basically biological neuron receives input from other sources, combines them in some way, performs generally non-linear operation on the result, and then the final result [21].

2.7.3 Linear Regression

2.7.3.1 BIVARIATE LINEAR REGRESSION

It is applicable to evaluate whether one variable, called the dependent variable or the response, can be caused, explained and therefore predicted as a function of another, called the independent variable, the explanatory variable, the covariate or the feature. Y is used as the dependent (or response) variable and X is used as independent (or explanatory) variable. The simplest statistical model that can describe Y as a function of X is linear regression. The linear regression model specifies a noisy linear relationship between variables Y and X, and for each paired observation (xi, yi) this can be expressed by the so-called regression function: yi = a +

bxi + ei (i = 1, 2, . . . , n) where *a* is the intercept of the regression function, *b* is the slope coefficient of the regression function, also called the regression coefficient, and ei is the random error of the regression function, relative to the ith observation. The regression function has two main parts: the regression line and the error term. The regression line can be built empirically, starting from the matrix of available data. The error term describes how well the regression line approximates the observed response variable. From an exploratory viewpoint, determination of the regression line can be described as a problem of fitting a straight line to the observed dispersion diagram. The regression line is the linear function yi = a + bxi (i = 1, 2, . . . , n) where yi indicates the fitted ith value of the dependent variable, calculated on the basis of the ith value of the explanatory variable *xi* . Having defined the regression line, it follows that the error term ei in the expression of the regression function represents, for each observation yi , the residual, namely the difference between the observed response values yi , and the corresponding values fitted with the regression line, yi :ei = yi . .yi. Each residual can be interpreted as the part of the corresponding value that is not explained by the linear relationship with the explanatory variable. To obtain the analytic expression of the regression line, it is sufficient to calculate the parameters a and b on the basis of the available data. The method of least squares is often used for this. It chooses the straight line that minimizes the sum of the squares [16].

Regression is a simple and powerful predictive tool. To use it in real situations, it is only necessary to calculate the parameters of the regression line, according to the formulae above, on the basis of the available data. Then a value for *Y* is predicted simply by substituting a value for *X* into the equation of the regression line. The predictive ability of the regression line is a function of the goodness of fit of the regression line, which is very seldom perfect.

However, although at a simplest senses regression uses standard statistical techniques such as linear regression, because of the complex nature of the real world problems, more complex techniques like decision tree and neural networks may be necessary to forecast future values [35].

2.7.3.2 MULTIPLE LINEAR REGRESSIONS

It is applicable if there is more than one explanatory variable. Suppose that all variables contained in the data matrix are explanatory, except for the variable chosen as response variable. Let *k* be the number of such explanatory variables. The multiple linear regression is defined by the following relationship, for i = 1, 2, . . . , n: yi = a + b1xi1 + b2xi2 +…+bkxik + ei [16].

2.7.3.3 LOGISTIC REGRESSION

The above two considered a predictive model for a quantitative response variable; where as logistic regression considers a predictive model for a qualitative response variable. A qualitative response problem can often be decomposed into binary response problems (e.g. Agresti, 1990). The building block of most qualitative response models is the logistic regression model, one of the most important predictive data mining methods. Let yi (i = 1, 2, . . . , n) be the observed values of a binary response variable, which can take only the values 0 or 1. The level 1 usually represents the occurrence of an event of interest, often called a 'success'. A logistic regression model is defined in terms of fitted values to be interpreted as probabilities that the event occurs in different subpopulations:

ði = P(Yi = 1), for i = 1, 2, . . . , n. [16]

2.8 General Application of Data Mining

It is expected that data mining have broad applications. It can help business managers find and reach suitable customers as well as develop special intelligence to improve market share and profits. Besides, data mining has other applications like DNA data analysis in biomedical researches. Recent research in DNA data analysis has enabled the discovery of genetic causes of many diseases as well as discovery of new medicines. Moreover, financial data analyses, data analysis for retail and telecom industry, intrusion detection and network security are some other applications of data mining [29].

[11] again state that KDD (data mining) have been deployed in science and in business areas. In science, astronomy is one of the application areas of data mining. Here, a notable success was achieved by SKICAT, a system used by astronomers to perform image analysis, classification,

and cataloging of sky objects from sky survey images. In its first application, the system was used to process the 3 terabytes (10^{12}bytes) of image data. SKICAT can outperform humans and traditional computational techniques in classifying faint sky objects. In business, on the other hand, KDD application areas include marketing, finance (especially investment), fraud detection, manufacturing, telecommunications, and internet agents. The primary application of KDD in marketing area is database-marketing systems, which analyze customer groups and forecast their behavior. Another notable marketing application is market basket analysis systems, which find patterns such as," If customer bought z, he/she is also likely to buy Y and Z" such patterns are valuable to retailers. In investment, numerous companies use data mining, but most do not describe their systems. One exception is LBS Capital Management. Its system uses expert systems, neural nets, and genetic algorithms to manage portfolios totaling $600 million; since its start in 1993, the system has outperformed the broad stock market. In fraud detection, HNC falcon and Nestor PRISM systems are used for monitoring credit card fraud, watching over millions of accounts. The FAIS system, from the U.S. Treasury financial crimes Enforcement Network, is used to identify financial transactions that might indicate money-laundering activity. In manufacturing, the CASSIOPEE troubleshooting system was applied by three major European airlines to diagnose and predict problems for the Boeing 737.To derive families of faults, clustering methods are used. CASSIOPEE received the European first prize for innovative applications. In telecommunications, the telecommunications alarm-sequence analyzer (TASA) was built in cooperation with a manufacturer of telecommunications equipment and three telephone networks. The system uses a novel framework for locating frequently occurring alarm episodes from the alarm stream and presenting them as rules. Large sets of discovered rules can be explored with flexible information-retrieval tools supporting interactivity and iteration. In this way, TASA offers pruning, grouping, and ordering tools to refine the results of a basic brute-force search for rules.

2.9 Application of Data Mining in Agricultural Areas

Undoubtedly, digital data contents are most important when we develop agriculture information systems. Actually, fundamental and widely used data such as market information, weather information, and agricultural material information are becoming available on the Internet or other sources. In addition to the above-mentioned fundamental data, site-specific field data are definitely necessary when a farmer requires some site-specific decision support. For example, a growth model may require soil and fertilization information for accurate predictions. To obtain such information, the farmer has to record field data continuously for a considerably long period. In spite of such importance, providing tools for farmers to easily collect these data is often forgotten and neglected [23].

Recently, data collection systems like a web camera, mobile-phone-based farm-working journal, a field monitoring system called Field Server and the likes are developed by different scholars. As a result, explosion of data in the area of agriculture happen. This causes the arousal of a mechanism to study the huge amount of data to mine knowledge or to support agricultural decisions [23].

Currently, various literatures suggest that there is a huge amount of data in agricultural production and experiments that have been recorded. These long-term data may be sources of critical information that will give us new knowledge in agricultural production. At present, most of the data are still in printed materials and considerable labor is still required to digitize those data for an agricultural information system. However, once digitized, a modern technology called data mining is available to analyze the resulting huge amount of data. Data mining is based on several statistical procedures and graphical presentation of data and It is analogous with mining gold from gigantic mountains. As it is known, the background of agriculture is complicated and a huge number of factors are related to it. Therefore, data mining technology is expected to help us to mine unknown facts from huge amounts of data that will have effect on grain yield increase to contribute the country's attempt for food security.

Consequently, agricultural researchers begin to apply data mining technology in the field of agriculture. Classification system for sorting mushrooms by grade, diagnosing soybean diseases are some of the application of data mining in the field of agriculture [23].

According to Cunningham and Holmes, (1999) as cited by Armstrong L., (2007) used WEKA to develop a classification system for the sorting and grading of mushrooms. The j4.8 algorithm classifier within WEKA was used to create a model for the human inspectors and the automated system. The model created using the human rules showed that each inspector used different combinations of attributes when assigning grades to mushrooms. The application of data mining techniques provided within the WEKA software application created a model that analyzed all attributes and created a model that was faster and more accurate than the human system.

The decision tree analysis method has been used in the prediction of natural datasets in agriculture and was found to be useful in prediction of soil depth for a dataset. In Mckenzie and Ryan (1999) as cited by [3] the uses of slope angle, elevation, temperature and other factors were analyzed and models created for prediction of soil depth across a sample area. The model was tested through the use of random data sets. “at each level, trees with increasing numbers of terminal nodes were fitted 20 times with 5% of the data randomly selected and withheld to provide a test of the predictive strength of the model” (Mckenzie and Ryan,1999) as cited by [3].

Besides, [34] investigate the relative importance of soil fertility and crop management factors in determining yield variability and the gap between farmers’ maize yields and potential yields in western Kenya. Because of the complexity of the data set, classification and regression trees (CART) were used to relate crop yields to soil and management factors. CART analysis showed resource use intensity, planting date and time of planting were the principal variables determining yield, but at low resource intensity, total soil N and soil Olsen P became important yield-determining factors. Only a small group of plots with high average grain yields (2.5 t ha^{-1}; $n = 8$) was associated with use of nutrient inputs and good plant stands, whereas the largest group with low average yields (1.2 t ha^{-1}; $n = 90$) was associated with soil Olsen P values of less than 4 mg kg^{-1}.The Authors further state that this classification could be useful as a basis for targeting agronomic advice and inputs to farmers. The results suggest that soil fertility variability patterns on smallholder farms are reinforced by farmers investing more resources on already fertile fields than on infertile fields. CART proved a useful tool for simplifying analysis and providing robust models linking yield to heterogeneous crop management and soil variables.

Chapter Three: Ethiopian Agriculture and Fertilizer's Data Analysis in National Soil Testing Center

In this chapter, review of related literature regarding Ethiopian Agriculture, problems that hinder the development of Ethiopian agriculture and the solutions that are taken are presented. It also, specifically, will address the issue of soil fertility status; it is about nitrogen, phosphorus, potassium, soil pH and organic matter available in the soil. Moreover, it will state uses and productivity of the most important crops in Ethiopia such as wheat and tef in terms of production and area coverage. This will help to give a clue for understanding dependent and independent variables for improving productivity of these crops and will assist in the process of data mining application, in this research.

3.1 History of Agriculture

As history reveled, there were no agricultural communities until about 11,000 years [24]. The earliest man was a hunter of small animals and gatherer of plants. As population increases, it became difficult to find enough food from hunting and gathering. This probably led to the practice of agriculture though the change from the later life style to the former is gradual. Agriculture is a deliberate tending of crops and rearing of animals for human use [24].

As [24] stated that sites of early farms have been discovered in Thailand in about in 11,000 B.C, in the near East in about 9,000 B.C.and in Mexico in about 6,000 B.C. Then, it is extended to China, Japan and South-East Asia, Nile valley and Europe. From Nile valley, agriculture had spread southwards and westwards through Africa to the Sudan region by about 3,500 B.C.

According to Vavilov (1926) as cited by [24] China, India, Central Asia, Near East, Ethiopia, South Mexico and South/Central America are some of the centers of origin of crop plants. These regions all lie between 20^{o} and 45^{o} latitude in mountainous regions and often in areas with a temperate climate and they are separated by great deserts. Vavilov further states that Agriculture developed independently in these eight regions as evidenced by the differences in agricultural methods, implements and domestic animals.

Nowadays, most African countries are predominantly agricultural and it is recognized that increasing the productivity of their agriculture is essential to their economic development [24]. However, the report of Africa 2000, by the organization of economic cooperation and development (OECD) as cited by [24] predicts that Africa will only manage to feed 65% of its fast-growing population by the end of the century, unless radical changes take place.

3.2 Agriculture in Ethiopia

Ethiopia has a human population of 74.0 million of which 84% is rural and growing at a rate of 2.6% per annum [31]. Agriculture is the mainstay of Ethiopian economy accounting for about 53% of the GDP, more than 90% of the total export revenue and 85% of total employment [31]. Above all, it provides the largest portion of basic food supply for the urban population, raw materials for agro-industries and agricultural commodities for export. Among the different crops widely grown in Ethiopia, wheat and tef are the major ones in terms of area coverage and production.

3.2.1 Wheat and Tef Production in Ethiopia

According to [38]wheat is the most widely grown cereal in the world which is cultivated from the borders of the arctic to near the equator in areas ranging in from sea level to 4572 meter above sea level (m.a.s.l).

Wheat is the staple food of nearly 35% of the world population and it is the most important agricultural commodity in international trade and occupies approximately 20% of the world's cultivated land. It is grown on an area of 200 million hectares worldwide with annual production of 600 million tons [38].

Ethiopia is the second largest wheat producer in the sub-Saharan Africa, next to South Africa. In Ethiopia, wheat took up 14% (1.5 million ha) of the total grain crop area and accounted for 18 % (22 million quintals) of the production in Meher season of 2004 [38]. The two cultivated tetraploid wheat for which Ethiopia is known as the center of diversity are durum wheat (*T. durum*) and cultivated emmer wheat (*T .diccocum*). Bread wheat (*T. aestivum*) is produced at slightly higher elevations and on better drained soils than durum wheat [38].

Though Ethiopia is the second largest producer of wheat in the sub-Saharan Africa, the mean wheat yield is around 1.4 t/ha, which is very low compared to the yield from the experimental plot, which is 5 t/ha [38]. The low wheat yield on farmers' field is mainly because of biotic, a biotic and socio economic constraints. The major abiotic factors that limit wheat productivity are low soil fertility, water logging and moisture stresses. Diseases, weeds and insect pests are also the biotic constraints of wheat production [38].

Tef is the other widely grown cereal in Ethiopia. The fact that the genetic diversity for tef exists nowhere in the world except in Ethiopia, indicates that it is originated and was domesticated in Ethiopia [17]. [24] identified Ethiopia as the centre of origin and diversity of tef. As with several other crops, the exact date and location for the domestication of tef is unknown. However, there is no doubt that it is a very ancient crop in Ethiopia, where domestication took place before the birth of Christ.

It is stated that according to 1998/99 data, tef constitutes 31% of the land area [25] devoted to seven cereal crops, followed by maize (19.3%), sorghum (15.5%), wheat (14.6%) and barley (12.3%). In the same year, the share of tef production was 21.4% among the seven cereals surpassed by maize (31%) only [17]. Considering the country as a whole, tef is produced in seven regions to varying degrees. Amhara and Oromia have the largest acreage of tef followed by Southern Nations Nationalities and Peoples Region (SNNPR) and Tigray [17].

According to [17], tef is preferred cereal by the farmers since it has higher price as compared to other cereals, not attackable by weevils and unattached by disease epidemic. Besides, it is the mainstay of Ethiopian diet in the form of injera [28] and its straw is preferred feed for cattle, for thatched roofs and mud bricks [25]. Apart from this, it can perform under low or high moisture condition than other cereals [17].

3.2.2 Challenges in the Development of Ethiopian Agriculture

In the past, when population pressure was low, agricultural production was sustainable because soil fertility was maintained by fallowing and by rotation of legumes with cereals. After

harvesting crops, animals were kraaled overnight on arable land to add nutrients to the soil [15].Today, these indigenous techniques for soil fertility maintenance have been interrupted. Animals are no longer kept overnight on arable land because it is no longer safe to do so. Dung is collected and used as a source of fuel. Fallow land has decreased dramatically since farmlands are continuously cultivated. As a matter of fact, during the last few years, more forest and grazing lands have been converted to cropland. Legumes and oilseeds have been reduced from the rotation due to pest damage and low yield potential.

To strengthen the above idea [36] states that despite the huge working force involved in agriculture and the enormous potentials the country has, it is unable to produce enough food to feed its population and achieve self-sufficiency in food. In connection to this, according to central statistics, it is estimated that in 2001/2 croping season, 6.72 million hectares were under cereals, 0.92 million hectare under pulses and 0.51 million hectares under other annual crops. However, the average yield expected from each crop, during the season, was estimated at 1236, 775 and 67 kg/ha for cereals, pulses and other annual crops, respectively.

The poor performance of the agricultural sector is due to a number of factors, the most important of which is sever soil degradation and fertility decline [20]. The problem is most serious in the highlands (>1500 m.a.s.l) which cover 45% of the total land area (i.e. 100 million ha) and support 85% of the human and 75% of the livestock population and account for more than 90% of the regularly cultivated lands [36]. Considerably high deforestation in this high land will result alarming soil erosion and this has resulted highest nutrient depletion. Poor soil fertility status especially, deficiency of nitrogen and phosphorus among others account for this failure [40].

3.3 Fertilizers

Chemical fertilizers are compounds given to plants to promote growth. They are usually applied either through the soil, for uptake by plant roots, or by foliar feeding, for uptake through leaves. Fertilizers can be organic (composed of organic matter), or inorganic (made of simple, inorganic chemicals or minerals) [12]. They can be naturally occurring compounds such as peat

or mineral deposits, or manufactured through natural processes (such as composting) or chemical processes [8]. Similarly, [7] define fertilizer as any organic or inorganic material of natural or synthetic origin added to a soil to supply certain elements essential to the growth of plants.

3.4 Major Soil Fertility Indicators and their Concepts

Some of the major important soil fertility indicators that are used in this data mining research have been briefly discussed as follows.

3.4.1 Soil pH

The soil pH value is a measure of soil acidity or alkalinity. Its scale ranges from 0 to 14, with 7 as neutral. Numbers less than 7 indicate acidity while numbers greater than 7 indicate alkalinity. The pH value of soil is one of a number of environmental conditions that affects the quality of plant growth. The soil pH value directly affects nutrient availability for plant growth. Soil pH values above or below the neutral level may result in less vigorous growth due to nutrient deficiencies [7].

The pH of the soil solution is very important because soil solution carries in it nutrients such as Nitrogen (N), Potassium (K), and Phosphorus (P) that plants need in specific amounts to grow, thrive, and fight off diseases. If the pH of the soil solution is increased above 5.5, Nitrogen (in the form of nitrate) is made available to plants. Phosphorus, on the other hand, is available to plants when soil pH is between 6.0 and 7.0 [7].

Certain bacteria help plants obtain N by converting atmospheric nitrogen into a form of N that plants can use. These bacteria live in root nodules of legumes (like alfalfa and soybeans) and function best when the pH of the plant they live in is growing in soil within an acceptable pH range. For instance, alfalfa grows best in soils having a pH of 6.2 - 7.8, while soybean grows best in soils with a pH between 6.0 and 7.0. Peanuts grow best in soils that have a pH of 5.3 to 6.6. Many other crops, vegetables, flowers and shrubs, trees, weeds and fruit are pH dependent and rely on the soil solution to obtain nutrients. If the soil solution is too acidic plants cannot

utilize N, P, K and other nutrients they need. In acidic soils, plants are more likely to take up toxic metals and some plants eventually die of toxicity (poisoning) [7].

Herbicides, pesticides, fungicides and other chemicals are used on and around plants to fight off plant diseases and get rid of weeds and bugs that feed on plants and kill plants. Knowing whether the soil pH is acidic or basic is important because if the soil is too acidic, the applied pesticides, herbicides, and fungicides will not be absorbed (held in the soil) and they will end up in garden water and rain water runoff, where they eventually become pollutants in streams, rivers, lakes, and ground water [9].

3.4.2 Soil Organic Matter

Soil organic matter is the organic fraction of the soil that includes plant and animal residues at various stages of decomposition, cells and tissues of soil organisms and substances synthesized by the soil population, commonly determined as the amount of organic material contained in a soil sample passed through a 2-mm sieve.

Soil organic matter encompasses all the organic components of a soil: (1) living biomass (intact plant and animal tissues and microorganisms), (2) dead roots and other recognizable plant residues, as well as (3) a largely amorphous and colloidal mixture of complex organic substances no longer identifiable as tissues. Only the third category of organic material is properly referred to as soil humus [7].

Soil organic matter affects so many soil properties and soil-environment interactions. For example, adding organic mulch to the soil surface encourages earthworm activity, which in turn leads to the production of burrows and other bio-pores, which in turn increases the infiltration of water and decreases its loss as runoff, a result that finally leads to less pollution of streams and lakes [7].

Soil organic matter has influence on soil physical properties. The humic fractions in organic matters help reduce the plasticity, cohesion, and stickiness of clayey soils, making these soils

easier to manipulate. Soil water retention is also improved, since organic matter increases both infiltration rate and water-holding capacity [7].

Soil organic matter also affects the chemical property of the soil. Humus generally accounts for 50 to 90% of the cation-adsorbing power of mineral surface soils. Like clays, humus colloids hold nutrient cations (potassium, calcium, magnesium, etc.) in easily exchangeable form, wherein they can be used by plants but are not too readily leached out of the profile by percolating waters. Through its cation exchange capacity and acid and base functional groups, organic matter also provides much of the pH buffering capacity in soils. In addition, nitrogen, phosphorus, sulfur, and micronutrients are stored as constituents of soil organic matter, from which they are slowly released by mineralization [7].

The organic matter content of sub soils is even smaller. However, the influence of organic matter on soil properties, and consequently on plant growth is far greater than the low percentage would indicate.

3.4.3 Soil Nitrogen

Nitrogen is an integral component of many essential plant compounds. It is a major part of all amino acids, which are the building blocks of all proteins-including the enzymes, which control virtually all biological processes. Nitrogen is also essential for carbohydrate use within plants. A good supply of nitrogen stimulates root growth and development, as well as the uptake of other nutrients [7].

Plants respond quickly to increased availability of nitrogen, their leaves turning deep green in color. Nitrogen increases the plumpness of cereal grains, the protein content of both seeds and foliage and the succulence of such crop as lettuce and radishes. It can dramatically stimulate plant productivity, whether measured in tons of grain, volume of lumber, carrying capacity of pasture, or thickness of lawn. Healthy plant foliage generally contains 2.5 to 4.0% nitrogen, depending on the age of the leaves and whether the plant is a legume. The nitrogen content of surface mineral soils normally ranges from 0.02 to 0.5% being representative for cultivated soils [7].

Plants deficient in nitrogen tend to have a pale yellowish green color, have a stunted appearance and develop thin spindly stems. Nitrogen-deficient plants are the first to turn yellowish, possibly becoming prematurely senescent and dropping off [7].
When too much nitrogen is applied, excessive vegetative growth occurs; the cells of the plant stems become enlarged but relatively weak, and the top heavy plants are prone to falling over (lodging) with heavy rain or wind [7].

3.4.4 Soil Phosphorus

Neither plants nor animals can grow without phosphorus. It is an essential component of the organic compound often called the energy currency of the living cell: adenosine triphosphate (ATP). Synthesized through both respiration and photosynthesis, ATP contains a high-energy phosphate group that drives most energy-requiring biochemical processes. For example, the uptake of nutrients and their transport within the plant, as well as their assimilation into different biomolecules, are energy-using plant processes that require ATP.

Next to nitrogen, phosphorous has more widespread influence on both natural and agricultural ecosystems than any other essential element. Phosphorus-deficient plants are often severely stunted, since this element takes part in the synthesis of several essential compounds upon which all plant and animal life depends [7].

In agricultural ecosystems, phosphorus constraints are much more critical because phosphorus in the harvested crops is removed from the system with only limited quantities being returned in crops residues and animal manures. As a result, extreme phosphorus deficiencies are quite common where no supplementary sources of this element are applied to soils. Such conditions are widespread today in most countries of sub-Saharan Africa, where phosphorus-bearing fertilizers are either not available or where the cost of their being transported and applied is prohibitive. Phosphorus deficiency is one of the reasons why sub-Saharan Africa is the only major region in the world where per-captia food production has actually declined in the past three decades [39].

Adequate phosphorus nutrition enhances many aspects of plant physiology, including the fundamental processes of photosynthesis, nitrogen fixation, flowering, fruiting (including seed production) and maturation. Root growth, particularly development of lateral roots and fibrous rootlets, is encouraged by phosphorus. In cereal crops, good phosphorus nutrition strengthens structural tissues such as those found in straw or stalks, thus helping to prevent lodging (falling over). Improvement of crop quality, especially in forages and vegetables, is another benefit attributed to this nutrient [39].

3.5 Fertilizer Use in Ethiopia

Most Ethiopian soils are deficit in nutrients, especially nitrogen and phosphorus and fertilizer application has significantly increased yields of crops [4]. It is also stated that different field experiments that have been conducted through out the country tell us that fertilizer has a direct relation ship with productivity. Fertilizer plays a great role in increasing productivity [5].

In Ethiopia, organic fertilizer is rarely used because of its scarcity. It is highly used as a source of fuel than used as fertilizer. Based on this fact, scholars suggest the use of chemical fertilizer as the second best option. In connection to this, the government and its international partners attempt to promote the fertilizer sector [5].The government in 1996 launched the national fertilizer sector project (NFSP) with financial support from the World Bank and other donors. Since 1996 a total of seven organizations have participated in fertilizer importation and distribution. As a result, adequate fertilizers supply has been ensured and a significant increase in fertilizer consumption has been shown [5].

Although the fertilizer consumption has shown a tremendous increase over the last six years, the change in the quantity used per unit area has been much less impressive (about 25 kg/ha). Besides, there was no efficient use of fertilizer [5].

To promote efficient use of fertilizer on the basis of soil, crop and site-specific fertilizer recommendations, on-farm fertilizer experiments and soil test service have to be effectively promoted. Then, agricultural researchers begin to conduct field experiment to develop a guideline in adequate chemical fertilizer recommendation. However, the technique for this

guideline development is not as such sophisticated rather it is tiresome, time taking and resource demanding.

3.6 Analysis of Data in National Soil Testing Center

Researchers in the National Soil Testing Center have conducted experiments to recommend appropriate amount of fertilizer based on the given soil fertility status and crop nutrient requirements. In these experiments, first, sample plot of lands are selected in different wereda of different regions. Next, the amount of soil-pH, total nitrogen, initial phosphorus, clay content, organic matter and cation exchange capacity are measured from the soil samples collected from the selected fields. Then different proportion of chemical fertilizer from 1-18 as shown in table 1.1 together with the seeds are applied in these sample plots of lands. In the experiment, at least 15 farms in each wereda are selected. In each farm the experiment is done in two replications. Each farm is divided into two then the first half if for the first experiment and the second half is for the second experiment to make the result reliable. These and other information after sowing like height of the crops, the amount of seedling in each plot of land, amount of straw and grain yields are recorded and kept in the form of table for final data analysis. Generally, a total of 23 attributes as shown in table 4.2 are used to record data regarding the over all experiment. Finally, the analysis is done in each wereda, which is tiresome, by the help of statistical tool by the name MSTAT-C to see the relationship between the parameter grain yield and treatments. Incorporating other parameters like soil pH, total nitrogen, organic matter, initial soil phosphorus in the process of data analysis is beyond the capacity of this statistical tool. Hence, it is recommended to conduct comprehensive research with the application of advanced technology. So that, relationship among variables indicating soil fertility status with the amount of crop yield will easily be seen. Moreover, some other hidden knowledge will be investigated.

Chapter Four: Data Preparation

4.1 Overview

The fundamental purpose of data preparation is to manipulate and transform raw data so that the information content enfolded in the dataset can be exposed or made more easily accessible. Besides, data preparation involves enhancing and enriching the data in an attempt to improve knowledge discovery. There is recognition by many that there is as much art as science in data preparation. Clearly, it takes additional effort for data preparation and hence, the question of cost of doing it versus the benefits arises.

Data collection, data preprocessing (data cleaning, attribute selection, data formatting and transformation, dimensionality reduction and the like) are the most important activities under data preparation, which finally resulted in creating target data set.

4.2 Data Collection

The first and foremost step in the process of data analysis is data collection. It is unquestionable that there should be a huge data to apply data mining technology and to arrive at the result that one needs.

Collecting and structuring the data that is used for analysis is one of the tasks that needs close attention in the process of data mining. The data for this specific research was collected from National Soil Testing Center. NSTC is established to develop guideline for fertilizer recommendations based on initial soil fertility status and crop nutrient requirement as one of its major objectives. This in return achieves one of the goals of the government which is ensuring food security. In this connection, NSTC work hard with different laboratories in different parts of Ethiopia. These laboratories keep data in different soil parameters after field and laboratory experiments. In the soil, there are different nutrients in different proportion. Soil pH, nitrogen, phosphorous and organic matter are some of the soil properties (nutrients). These properties can contribute for effective growth of crops if they are in sufficient proportion, otherwise, they may cause damage on the growing crop.

So far, there are a number of records collected from experiments that have been made in different years. Among these, 2003 and 2004 data which is recent were chosen for this particular research since many of the soil parameters are there and they are also helpful to know the current situation. The record include both tef and wheat data. The wheat variety considered in the trial was bread wheat variety (a variety used to make bread) specifically called Qubsa. The tef variety used was white seeded (white color seed) known by its name called Enatite. These wheat and tef varieties were selected for the trial because they are high yielder, disease resistant, adapted to wide environments and are responsive to applied fertilizer. Because of these merits and their economic benefits, currently, most farmers are growing these wheat and tef varieties in the different parts of Ethiopia

The data under study incorporate information about the main soil parameters like soil pH, organic matter, initial available soil phosphorus and total nitrogen. It also contains information about where the experiment is performed, to whom the farm of the experiment belongs to and how much fertilizer is added to the soil while planting the crop. Above all, it incorporates different parameters that can explain the contribution of chemical fertilizer in the growth of crop; these parameters are like plant height, straw and grain yield, biomass, and seedlings. All the above information are recorded and kept during each and every experiment by NSTC. In connection to this, the data for this study, specifically, include 2887 records of wheat and 2512 records of tef. Total number of wheat and tef in each class are presented in table 4.1 and figure 4.1 as follows.

Table 4.1 Total number of wheat and tef records and their distribution in each class

	high	medium	low	Total
Wheat	1424	1154	309	2887
Tef	750	643	1119	2512
Total	2174	1797	1428	5399

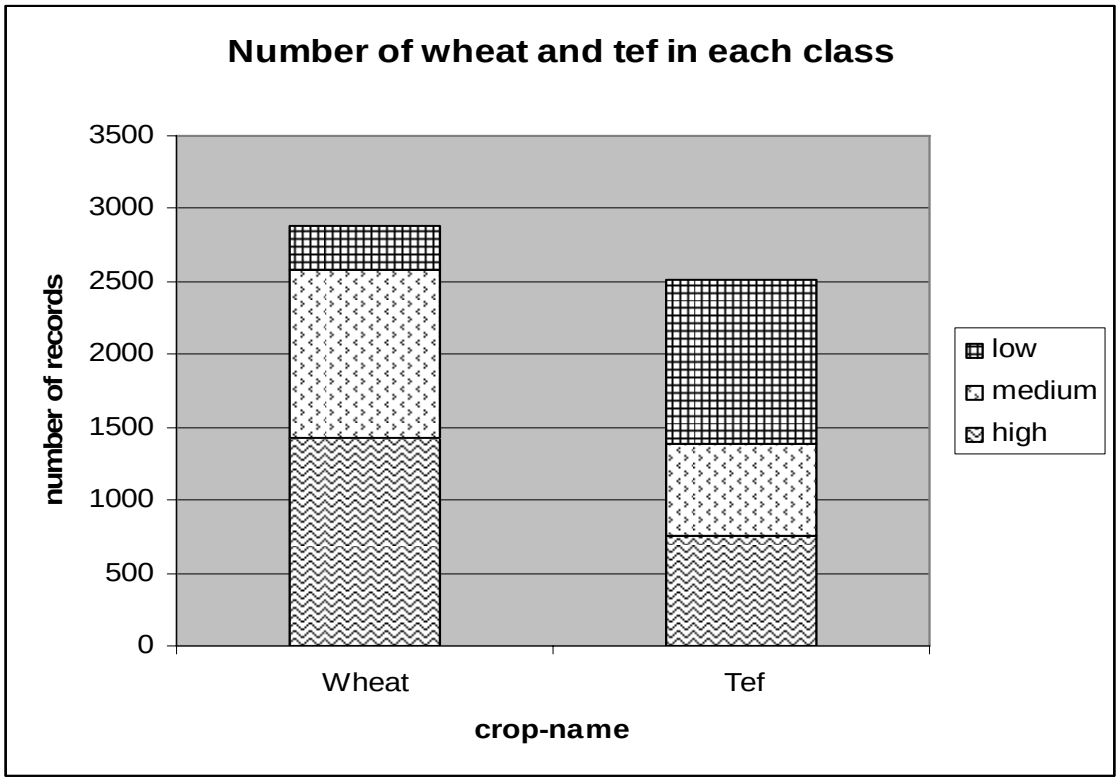


Figure 4.1Wheat and tef records that belong to the class high, medium and low.

These records in detail incorporate attributes like farmer name, plot number, site number, replication-number, treatment number, wereda, region, soil pH, organic matter, total nitrogen, phosphorus content, plant height, number of seedling, Thousand kernel weight (TKW), spike count, Olsen-p, clay content of the soil, biomass, grain yield, straw yield and CEC (cation exchange capacity) which are mentioned in table 4.2.

4.2.1 Data Understanding

As it is mentioned previously, NSTC has kept detailed information about initial fertility status of the soil, chemical fertilizer that is added, the grain and straw yield gained. This information is described through different attributes. The specific attributes that can describe the above mentioned information are plot number, replication number, treatment number, farmer name, site number, wereda, region, seedlings/m^2 , plant height (cm), spike count/m^2, Thousand kernel weight (TKW), biomass of 100 plants (gm), grain yield of 100 plants (gm), straw yield of 100 plants (gm), gain yield (Kg/ha), Phosphorous content (ppm), soil pH, organic matter (%), total

nitrogen (%), CEC (cmol+/kg soil) and clay (%) content. The attributes with their description are shown in the table below.

Table 4.2 attributes with their description and data type

S.No.	Attribute name	Data type	Description
1	Plot number	Number	Number given to the partition of the experimental land in a selected farm
2	Replication number	Number	Number given to the repetition of experiments in a farm.
3	Treatment number	Number	Different proportion and combination of applied nitrogen, phosphorous and potassium fertilizers
4	Farmer name	Text	Name of farmer whose farm is selected as experimental site
5	Site number	Number	A number to identify a selected experimental farm
6	Wereda	Text	Name given to the area where the farm is located
7	Region	Text	Area where the wereda is located
8	Seedling count	Number	Number of seedlings counted using quadrant between 15-20 days after planting
9	Plant height	Number	Height of plant measured in cm
10	Spike count	Number	Number of fertile plants per m^2 counted after heading
11	TKW	Number	Thousand Kernel (seed) weight.
12	Biomass of 100 plants	Number	Amount of both grain and straw yield of 100 plants
13	Grain yields of 100 plants	Number	Amount of grain yield per 100 plants
14	Straw yield of 100 plants	Number	Amount of straw yield per 100 plants

S.No.	Attribute name	Data type	Description
15	Grain yield	Number	Amount of grain yield in kg per hectare
16	Initial phosphorus	Number	Amount of available phosphorous in the soil before planting.
17	Soil pH	Number	pH of the soil
18	Organic matter	Number	Organic matter content of the soil
19	Total nitrogen	Number	Amount of nitrogen in the soil
20	Phosphorous	Number	Amount of available phosphorous in the soil after fertilizer is added
21	CEC	Number	Cation exchange capacity
22	Clay	Number	Clay content of the soil
23	Crop-name	text	Name of crop

In the process of understanding the data, from the attributes enumerated above, one attribute is chosen as dependent or target variable for this specific research though there are others that are likely to be dependent variable as well. This dependent attribute is labeled as grain yield in kg/ha. This variable is selected since it is worthwhile to analyze the status of crop yield which contributes greatly to the economy of the country in general and to the farmer in particular than other variables like straw yield. The remaining attributes except some of the irrelevant ones are considered as independent that determine the fate of this dependent variable. Regarding the characteristics of the dependent (the target) attribute, naturally, it has numeric value but for convenience it is converted into nominal values as: high, medium and low ;whereas many of the independent variables have continuous or numeric values and there is no conversion as far as these variables are concerned. Among others, treatment is given close attention by agricultural experts since they want to know more about the contribution of fertilizer on crop yield. In the above record treatment has numeric value that describe different proportion of fertilizer as indicated above in table 1.1.Table 4.3 and figure 4.2 summarizes about the dependent attribute and number of records in each class of the dependent attribute.

Table 4.3 Total number of records in each class

	High	medium	low
Wheat	1424	1154	309
Tef	750	643	1119
Total	2174	1797	1428

This is diagrammatically represented as follows:

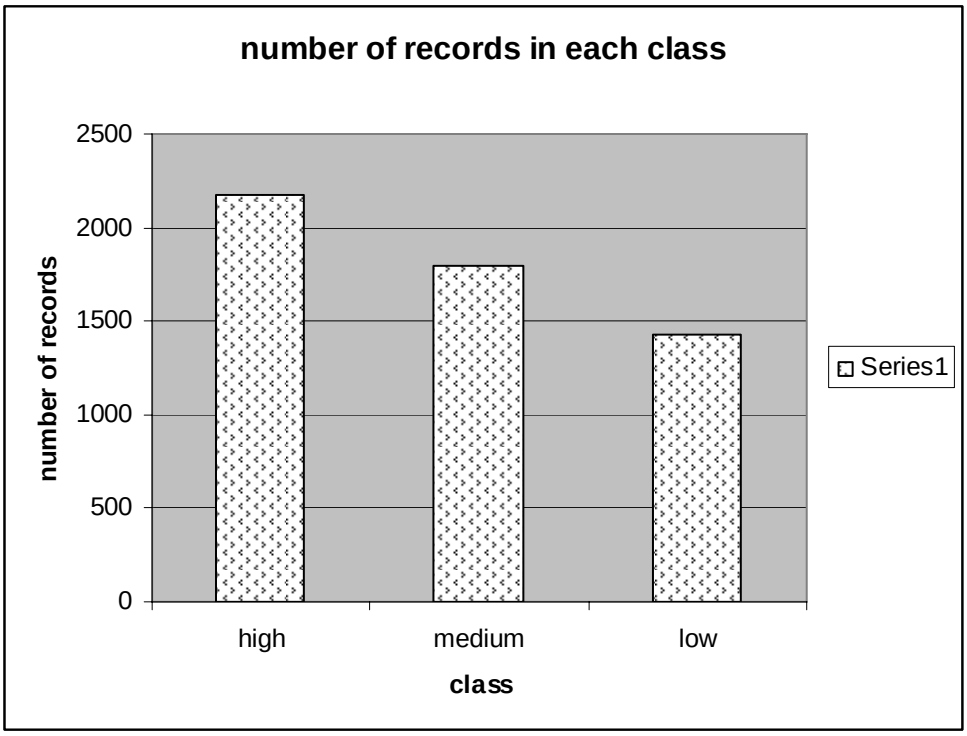


Figure 4.2 Number of records in each class labels

4.2.2 Defining the Data Mining Function

As it is explained in the literature review part, there are two main categories under the functionalities of data mining namely descriptive and predictive data mining. The focus of this research is on predictive data mining which mainly deals with classification and prediction tasks. Predictive data mining helps us to predict or classify the future result of one or more response or target variables in relation to what happens to the explanatory or input variables [16]. The researcher's aim in this study is to develop a model that can

accurately classify records into high, medium and low crop yield. And also the research will show the relationship between grain yield and the independent variables like treatments (chemical fertilizer), soil-pH, organic matter, initial phosphorous and total nitrogen content. Generally the predictive model in this study will predict the amount of crop yield based on the fertilizer added and other proportion of elements in the soil.

4.3 Data Pre-Processing

There are several problems related to the data collected from the real world. Some of these problems are data with missing, out of range or corrupt elements, noisy data, data from several levels of granularity, large data sets, data dependency and irrelevant data, multiple sources of data. These problems will jeopardize the task of data mining since it needs quality data. Unless there is quality data, the result of data mining will be garbage in-garbage out. Data preprocessing is, hence, working on the quality of data. It is through data preprocessing that one can increase the accuracy of data mining. Besides, it is believed that above 80% of the time in the process of knowledge discovery should be spent in the task of preprocessing [18]. Otherwise, garbage out is expected. Data preprocessing involves data cleaning, attribute selection, data transformation, integration, reduction and the like. Among this, the following are needed for this specific research.

4.3.1 Data Cleaning

It refers to fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

Fill in missing attribute or class values can be performed by using the attribute mean (or majority nominal value) or by using the attribute mean (majority nominal value) for all samples belonging to the same class. The other way of filling missing values is predicting the missing value by using a learning algorithm that consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value. Above all, ignoring the tuple can be done if majority of the attributes values or the class labels of the tuple are missed.

Identify outliers and smooth out noisy data is another technique of data cleaning. It is performed using binning, clustering and regression methods. Cleaning data can also be done by correcting inconsistent data.

Obviously, real world data is full of noise, inconsistency, missed values and other inconveniences because of technological and/or manual problems. As any other real world data, the data of this research share these problems. Hence it needs cleaning; without cleaning the data, analysis become under question and one may arrive at defective output which is a big loss in the research work. In this specific research, data cleaning was performed by consulting the domain expert as well as based on the researcher's own observation. There were a number of records whose values of many attributes were missed. In this case, it is unadvisable to fill all these values of attributes. As a result, the researcher was forced to chop these records. Besides, there were attributes whose values for the majority of the records were missed hence these attributes are also ignored. Attributes with higher percentage of missing value are displayed below in table 4.4.

Table 4.4 attributes with missing values

S.No	Attribute name	Missing value in percent
1	Clay (%)	40%
2	CEC(cmol+/kg soil	20%

Moreover, the records that belong to one wereda have no value for the attribute soil pH, organic matter, total nitrogen, initial phosphate, clay and CEC. Filling these values will bring unreliable result. Hence, the record as a whole was ignored.

Regarding inconsistency, since the data of this research was integrated from different work book of Excel and hard copies, naming inconsistency was there. Among these inconsistencies, some are mentioned as follows: name of one of the regions is written as Amahara in one record whereas it is written as Amhara in the other record. The researcher resolved this problem using one of the applications of Excel.

4.3.2 Data Integration

Merging data whose source are from different place and different format is the necessary step that needs attention in the process of data preparation for analysis.

Without merging the data, one cannot perform analysis and arrive at generalized result. In this research, some of the data were kept in Ms-Excel in different workbook and the others were in hard copy. All these were merged for the preparation of the final dataset.

4.3.3 Data Reduction

Under this technique tasks like reducing number of attributes, number of attribute values and number of tuples are incorporated. These tasks again can be performed by different methods. Data cube aggregation, removing irrelevant attributes and principal component analysis (numeric attributes only) are categorized under the task reducing number of attributes ;whereas binning (histograms), clustering, aggregation or generalization are under the category of reducing number of attribute values. Sampling, on the other hand, belongs to the method of reducing number of tuples.

In this study, some of these tasks were performed. Numbers of attributes were reduced from 23 to 9 by removing irrelevant attributes. Attributes, in this research, are considered as irrelevant if they don't contribute to the development of the model or if they are not relevant to solve the problem under study.

4.3.3.1 Attribute/Variable Selection

In the construction of the final dataset that is subjected to analysis, attribute selection is another thing that should be considered. Attributes that are irrelevant for the research goal should be swept away. To accomplish this task, the selected software Weka has the power to select the relevant attributes after calculating information gain and the like parameters of the attributes. However, this step (process) is unnecessary to take place since it isn't beyond the capacity of the domain expert and the researcher to identify the irrelevant attributes before that step took place. Hence, by taking into account the domain experts' idea and the researcher's own observation, the following attributes were chosen as relevant for this specific research.

Table 4.5 selected attributes for model development

S No.	Attribute name	Type	Description
1	Treatment-number	Number	Different proportion of N, P, K
2	Crop-name	Text	Name of the crop
3	Wereda	Text	Name given to the area where the farm is Located
4	Region	Text	Area where the wereda is located
5	Grain yield	Number	Amount of grain yield in kg per hectare
6	Soil pH	Number	pH of the soil
7	Organic matter	Number	Organic matter content of the soil
8	Total nitrogen	Number	Amount of nitrogen in the soil
9	Initial phosphorous	Number	Amount of phosphorous before fertilizer is added.

4.3.3.2 Data Discretization

It is one way of data reduction and it is done by replacing numerical attributes with nominal ones. There are two types of discretization: unsupervised and supervised. Under unsupervised discretization class variable is not used and it is done either using equal-interval (equi-width)binning that is splitting the whole range of numbers in intervals with equal size or by using equal-frequency (equi-depth) binning that is using intervals containing equal number of values. On the other hand, the supervised discretization uses the values of the class variable and it is done either by placing breakpoints between values belonging to different classes or by using entropy (information)-based discretization or by generating concept hierarchies.

In this research, there was an attempt to discretize the class value which is numeric in nature into nominal. As a result, the class value has three distinct values namely: high, medium and low. This kind of discretization is made with the intention that it can make the result of the analysis more interpretable and understandable. And to make the data suitable to the technique J48 selected in the Weka software. Moreover, other attributes like soil-pH, total-nitrogen and organic-matter were discritized to make the results (rules) of the analysis sound, reliable and interpretable and also to reduce the size of the decision tree. All these discretization were done by consulting the domain experts and by using range of number acceptable by the domain experts. Table 4.6 shows how discritization is done for class value. The discretization for the other values that is for soil pH and total nitrogen is shown in the analysis of the results.

Table 4.6 Discritized value of the class grain in kg/ha

Kg/ha	wheat	Kg/ha	Tef
>=3500	high	>=1500	high
1500-3500	Medium	1000-1500	Medium
<=1500	Low	<=1000	Low

4.3.4 Data Transformation and Aggregation

This is the other core point that should be considered in the process of data preprocessing. Data transformation is mainly appropriate where there are outliers and when the numeric values range varies significantly. Hence, not to create problem in the classification process, values will be transformed to the range between 0 and 1. In this research, data transformation is not needed since the above reason for transformation was not in the data.

Chapter Five: Model Building

This chapter portrays the experimentation and evaluation of the study. Specifically, building and training different models to select the best one in supporting decision makings is explained in detail. The chapter also presents discussions on the interpretation, validation and evaluation of such models.

5.1 Overview

Data classification is a two step process: one is model building and the other is using the model to predict the class label of new records [18]. The aim of this specific research is to understand the determinant variables affecting the amount of grain yield than in simply classifying the amount of yield as high, medium and low. In this chapter, different models were built and evaluated so as to meet the goal of building best decision tree model. Hence, decision tree that generate the soundest rules based on domain experts opinion than that predict the class label of new record was selected.

5.2 Selection of Modelling Techniques

This is the part that selection of modeling techniques is performed. As it is explained in the literature part, there are different data mining techniques that are capable to involve in the process of modeling. Among other, decision tree, neural network and regression are the major ones. In this research decision tree modeling technique is selected. According to many scholars including [6] suggest that decision tree is the most powerful modeling technique for both classification and prediction tasks. Since decision tree can be represented in the form of rules, it is easily understandable and readable. Similarly, [27] pointed out that decision tree is recommended for predictive tasks that require a classification-oriented model. Above all, decision tree is useful to depict the relationship of large number of input variable to target variable. Hence it is recommended to be used as the first step in the modeling process even when building the final model using some other technique [6].All the above facts make decision tree the most appropriate technique for the classification and prediction tasks that were carried out in this research.

Neural network technique next to decision tree is the other widely used technique in the task of classification. However, it has limitation with regard to its applicability for this study. Limitation of neural network for this study is its complexity to understand the induction process. According to [14] the induction process here is a black box. As a result it is not possible to generate rules from neural networks. Because of these limitations, neural network is not selected as a technique of analysis for this specific research.

Weka comprises decision tree algorithms of J48 and Id3 sorts under classifiers package. J48 implements a later and slightly improved version of C4.5; it is capable of handling numeric (continuous value) data. As a result, it is selected for this specific research since majority of the attributes under study have numeric values.

As to the evaluation techniques applied by this algorithm, there are more than three evaluation technique that is supported in Weka. Among these, percentage splits, cross-validatiion, training set are the major ones. In this research, two of these evaluation techniques namely cross-validation and training set were applied for comparison sake. After applying these two evaluation techniques in the process of building model, the one with best accuracy and sound rule in the eyes of the domain expert was selected for final interpretation of the result.

5.3 The Experiment

As it is mentioned previously, in this research, the process of building model was supported by Weka software. Weka software has different classification packages. Among others, tree was selected for the purpose of this research. Tree classifier as its name indicates comprises both decision tree and rule generation facilities. Under this tree classifier, there are different packages (algorithms) like Id3, J48. In this research J48 is preferred since it is the latest and capable of analyzing numerical data than Id3 or others (26).Weka can read files saved as .csv,.arff extensions. Hence, the Excel data of this research was transported to csv and/or arff format. Apart from the different classification technique that Weka supports, it has preprocessing facilities by the name filter that is found in the ‘choose’ menu of Weka user interface and under this, one can get various

techniques of preprocessing like discrtizetion, replacing missing value etc. Here is the screen shot of Weka with its preprocessing facilities.

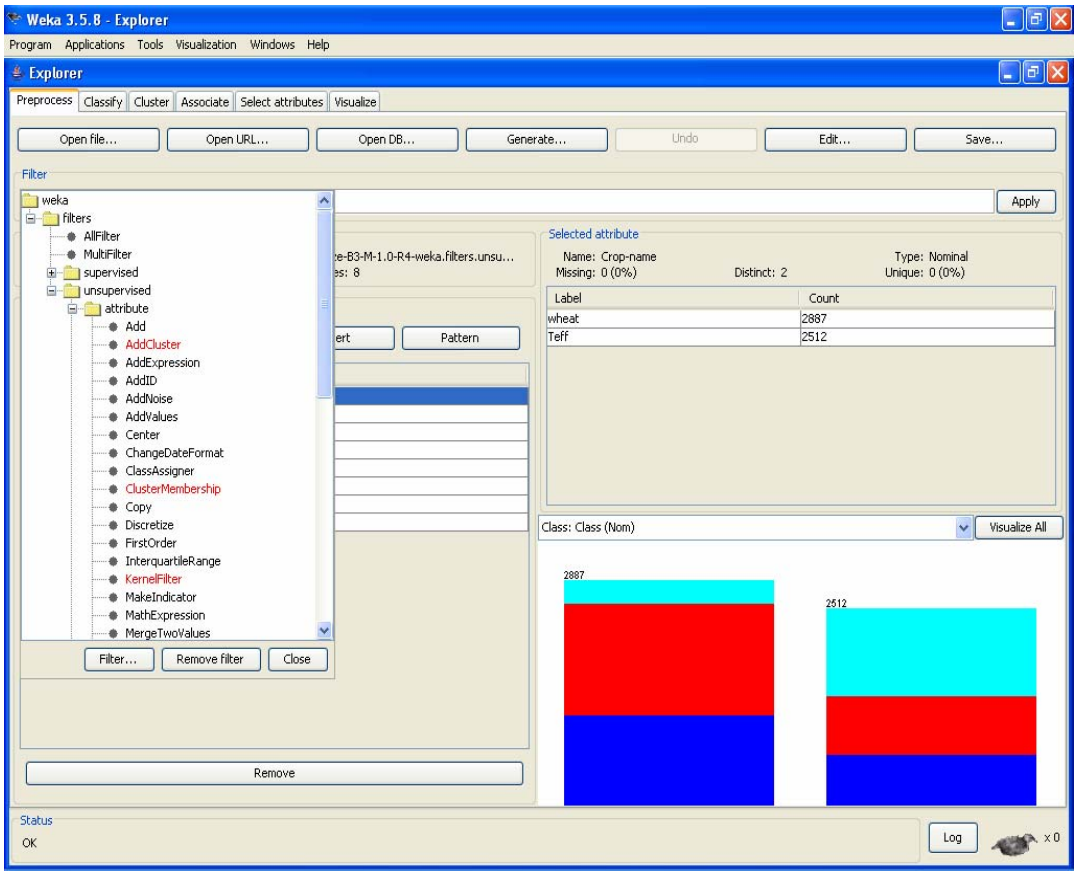


Figure 5.1 Preprocessing (filtering) dialog box in Weka.

Moreover, Weka has evaluation mechanism by the name use training set, cross-validation, supplied test set and percentage split. They are displayed as follows.

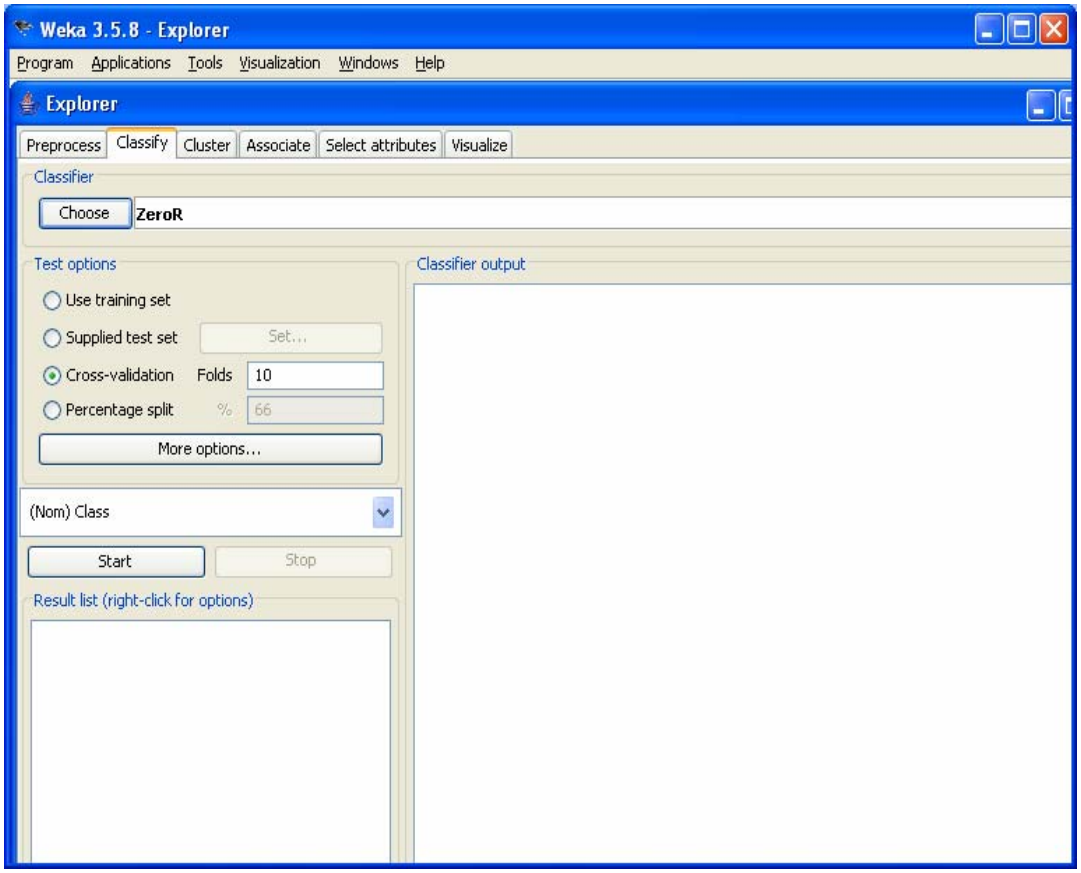


Figure 5.2 Evaluation dialog box in Weka.

In this research both training set and cross validation techniques were selected and applied for comparison of the results. Finally, a result with sound rule, better accuracy and more interpretability was chosen as a best model. The reason why these evaluation techniques are chosen is because they take sample for training and testing dataset automatically but the others need users to specify the number of data in the sample which is error-prone.

5.3.1 Model Building

It is the next step after preprocessing and selection of modeling technique and algorithm. In the processes of building the model, two distinct procedures were adopted. In the first case the model was built using the default values of the algorithm. In the other case, some parameters were modified when found necessary. On this premises the following experiments were conducted.

Decision Tree Model Building

In this research, there was an attempt to do as many experiments as possible. However some of the experiments were failed. In these experiments, Initial-p was descritized like the other attributes but it lowers the accuracy of the model and the attribute region was included by excluding the attribute wereda. Again, this makes the decision tree unsound (uninformative and more condensed). As a result, three of the experiments that are displayed below were chosen based on the soundness of the rules and the accuracy that they have.

Experiment One

The data provided for the purpose of classification task in this experiment has 5399 records and 9 attributes. Regarding the rest attributes, they weren't relevant to discriminate the records into the predefined classes (high, medium and low grain yield).All of the parameters in this experiment were set to their default and the evaluation technique chosen was 10-fold cross validation which is the best technique sited in different literatures. Output of experiment one in the form of confusion matrix is displayed below in table 5.1

Table 5.1 Output of experiment one in the form of confusion matrix

Actual	Predicted				
	High	medium	Low	Total	Score
High	1284	282	231	1797	71.4%
Medium	284	1864	26	2174	85.7%
Low	242	12	1174	1428	82.2%
Total	1810	2158	1431	5399	80.05%

The confusion matrix in table 5.1 depicts that out of the total records supplied to the classification algorithm 4322 records which is 80% are classified correctly. Out of which 1284, 1864, 1174 records were classified correctly as high, medium and low respectively. On the other hand, 1077 records which represent 19.9% are classified incorrectly. Specifically, 282 and 231 instances should have been classified as high but they are classified incorrectly as medium and low respectively. Similarly, 284 and 26 instances should have been classified as medium but they are wrongly classified as high and low respectively. Besides, 242 and 12 instances should have been classified as low but they are classified wrongly as high and medium respectively. Moreover, from the score column one can deduce that the class medium is classified correctly with minimum error as compared to the other class values. Above all, records that are incorrectly classified are very few. This indicates that the model performs very well.

However, the decision tree for the above experiment is bushy and complex in understanding since it has 250 numbers of leaves and 448 size of tree. Besides, since the discritization was done by the software itself, the rules generated were not meaningful in the eyes of the domain experts. As a result, based on the experts' opinion, the second experiment comes into existence with some adjustment of parameters to curb the problem that is faced by this experiment. Part of the output of the algorithm is displayed in the form of rules and decision tree below.

```
J48 pruned tree
------------------
Treatment <= 5
|   Crop-name = wheat
|   |   Organic-matter <= 1.41
|   |   |   Treatment <= 4
|   |   |   |   Soil-pH <= 7.09: medium (104.0/13.0)
|   |   |   |   Soil-pH > 7.09
|   |   |   |   |   Total-nitrogen <= 0.05: low (8.0)
|   |   |   |   |   Total-nitrogen > 0.05: medium (32.0/4.0)
|   |   |   Treatment > 4: medium (36.0/5.0)
```

```
|  |  Organic-matter > 1.41
|  |  |  Soil-pH <= 8.2
|  |  |  |  Region = Amahara
|  |  |  |  |  Soil-pH <= 5.52
|  |  |  |  |  |  Total-nitrogen <= 0.18
|  |  |  |  |  |  |  Treatment <= 2
|  |  |  |  |  |  |  |  Soil-pH <= 5.39: low (8.0)
|  |  |  |  |  |  |  |  Soil-pH > 5.39
|  |  |  |  |  |  |  |  |  Wereda  = Wogera: low (8.0/1.0)
|  |  |  |  |  |  |  |  |  Wereda  = Woreilu: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Debreilias: medium (8.0/1.0)
|  |  |  |  |  |  |  |  |  Wereda  = Yilmana-Densa
|  |  |  |  |  |  |  |  |  |  Treatment <= 1: medium (2.0)
|  |  |  |  |  |  |  |  |  |  Treatment > 1: low (2.0)
|  |  |  |  |  |  |  |  |  Wereda  = Lume: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Hetosa: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Dejen: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Achefer: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Tahtaykoraro: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Alefa: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Kuyu: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Wara-Jarso: medium (0.0)
|  |  |  |  |  |  |  Treatment > 2: medium (47.0/17.0)
|  |  |  |  |  |  Total-nitrogen > 0.18: medium (42.0/8.0)
|  |  |  |  |  Soil-pH > 5.52
|  |  |  |  |  |  Total-nitrogen <= 0.06
|  |  |  |  |  |  |  Treatment <= 3
|  |  |  |  |  |  |  |  Treatment <= 1: medium (2.0)
|  |  |  |  |  |  |  |  Treatment > 1: high (4.0/1.0)
|  |  |  |  |  |  |  Treatment > 3: medium (4.0)
```

Figure 5.3 Part of the rules generated by the default values of the program

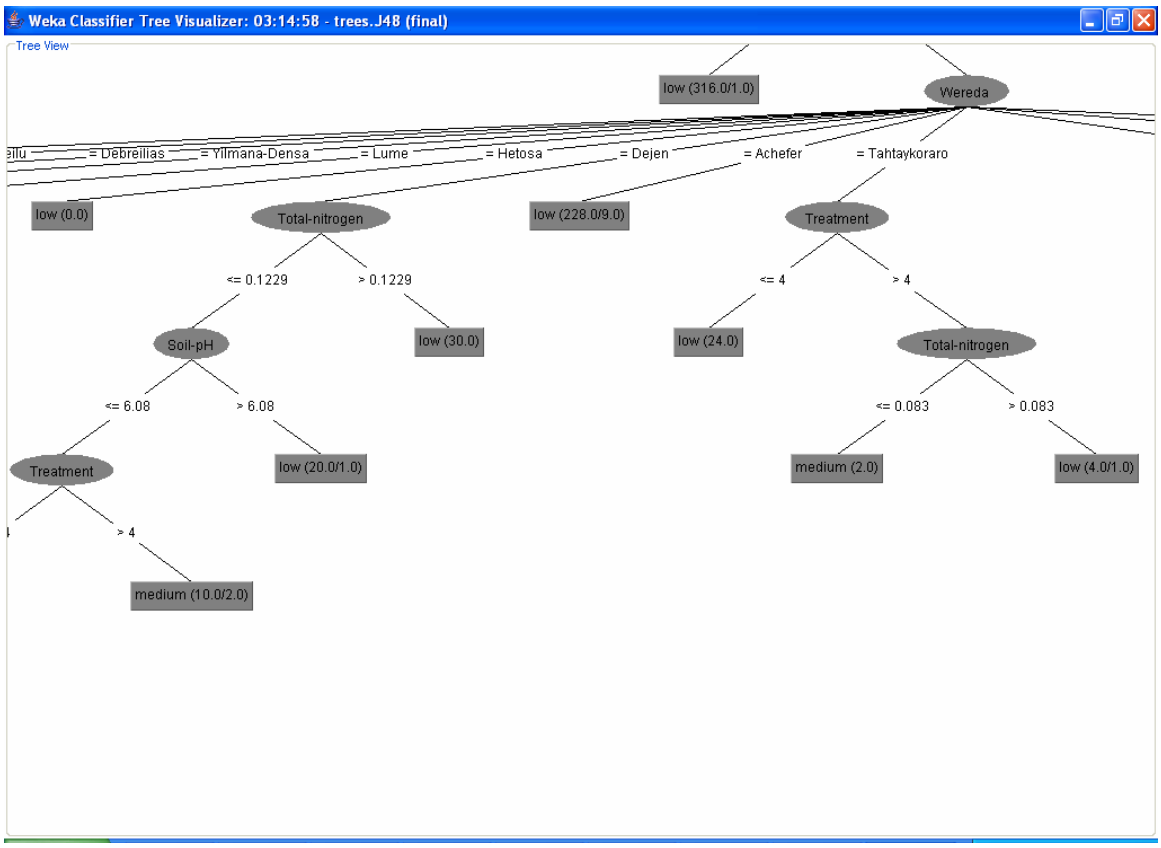


Figure 5.4 part of a decision tree generated by the default values of the program

Experiment Two

This experiment was conducted with some parameter modifications. These modifications were performed with the intention to reduce the size of the tree and number of leaves and to make the rule more sound, interpretable and understandable. To achieve this objective, the following tasks were performed. Discritization was applied in the parameter soil-pH, total-nitrogen and organic matter. This was done by consulting the domain experts and based on the classification of these parameters in the office of National Soil Testing Center. However, one of the parameters that is initial-P was not discritized since it was shown in different experiments that discritizing it reduces the accuracy of the results in that experiment. Because of this, it is preferred to leave Initial-P as it is, without discritization.

Above all, in this experiment, the attribute region was removed since its capacity to discriminate the records into the predefined classes was not that much. And also it can be represented, specifically, by the attribute wereda.

Having these modifications, slight reduction in the size of tree is achieved. Number of leaves in this case is increased to 256 but size of the tree is reduced to 442. Regarding the accuracy of the model, correctly classified data is almost the same as the previous experiment (with slight difference). In connection to this, the domain experts suggest that though there is insignificant difference in the accuracy of the two experiments, the decision tree for the later experiment has generated sound and more interpretable rules.

Table 5.2 Output of experiment two in the form of confusion matrix.

Actual	Predicted				
	High	Medium	Low	Total	Score
High	1297	291	209	1797	72.1%
Medium	277	1875	22	2174	86.2%
Low	270	27	1131	1428	79.2%
Total	1844	2193	1362	5399	79.6%

The confusion matrix in Table 5.2 depicts that out of the total records supplied to the classification algorithm 4303 records that is approximately 80% are classified correctly. Out of which 1297, 1875, 1131 records were classified correctly as high, medium and low respectively. On the other hand, 1096 records that represent almost 20% are classified incorrectly. Specifically, 291 and 209 instances should have been classified as high but they are classified incorrectly as medium and low respectively. Similarly, 277 and 22 instances should have been classified as medium but they are wrongly classified as high and low respectively. Besides, 270 and 27 instances should have been classified as low but they are classified wrongly as high and medium respectively. Moreover, from the score column one can deduce that the class medium is classified correctly with minimum error as compared to the other class values.

J48 pruned tree

Treatment <= 5

| Crop-name = wheat

| | Wereda = Wogera

| | | Initial-P <= 13.342

| | | | Total-nitrogen = <0.14

| | | | | Soil-pH = <5: low (10.0/3.0)

| | | | | Soil-pH = [5-8.5]: medium (10.0/3.0)

| | | | | Soil-pH = >8.5: medium (0.0)

| | | | Total-nitrogen = [0.14-0.24]

| | | | | Initial-P <= 5.2

| | | | | | Treatment <= 3: low (6.0/2.0)

| | | | | | Treatment > 3: medium (4.0)

| | | | | Initial-P > 5.2: low (40.0/6.0)

| | | | Total-nitrogen = >0.24: low (10.0/1.0)

| | | Initial-P > 13.342: medium (20.0/7.0)

| | Wereda = Woreilu

| | | Organic-matter = <2.04

| | | | Treatment <= 4

| | | | | Soil-pH = <5: medium (0.0)

| | | | | Soil-pH = [5-8.5]

| | | | | | Initial-P <= 4.4: medium (24.0/7.0)

| | | | | | Initial-P > 4.4

| | | | | | | Initial-P <= 22: low (16.0/2.0)

| | | | | | | Initial-P > 22

| | | | | | | | Treatment <= 3: medium (6.0/2.0)

| | | | | | | | Treatment > 3: low (2.0)

| | | | | Soil-pH = >8.5

| | | | | | Initial-P <= 7.09: low (8.0/2.0)

| | | | | | Initial-P > 7.09: medium (16.0/6.0)

Figure 5.5 Part of rules generated with some modified value of the program

Experiment Three

This was done with slight modification from experiment two. Here, the evaluation technique used is training set. The other parameters were as they were in the experiment two. This experiment is conducted with the intention to improve the accuracy and interpretability of the decision tree model. As it is shown in the confusion matrix below, the accuracy of the model has registered dramatic increase from 80% to 85.6%. However, regarding the rules generated, it is as sound as the previous experiment.

Table 5.3 Output of experiment three in the form of Confusion Matrix.

Actual	Predicted				
	High	Medium	Low	Total	Score
High	1427	230	140	1797	79.4%
Medium	190	1963	21	2174	90.2%
Low	177	19	1232	1428	86.2%
Total	1794	2212	1393	5399	85.6%

The confusion matrix in Table 5.3 depicts that out of the total records supplied to the classification algorithm 4622 records that is approximately 86% are classified correctly. Out of which 1427, 1963, 1232 records were classified correctly as high, medium and low respectively. On the other hand, 777 records that represent almost 14% are classified incorrectly. Specifically, 230 and 140 instances should have been classified as high but they are classified incorrectly as medium and low respectively. Similarly, 190 and 21 instances should have been classified as medium but they are wrongly classified as high and low respectively. Besides, 177 and 19 instances should have been classified as low but they are classified wrongly as high and medium respectively. Moreover, from the score column one can deduce that the class medium is classified correctly with minimum error as compared to the other class values.

This decision tree model was selected as a best model to generate working rules and to mine knowledge since the rules, here, are sound, accurate and interpretable. Besides, the numbers of records which were misclassified are very few.

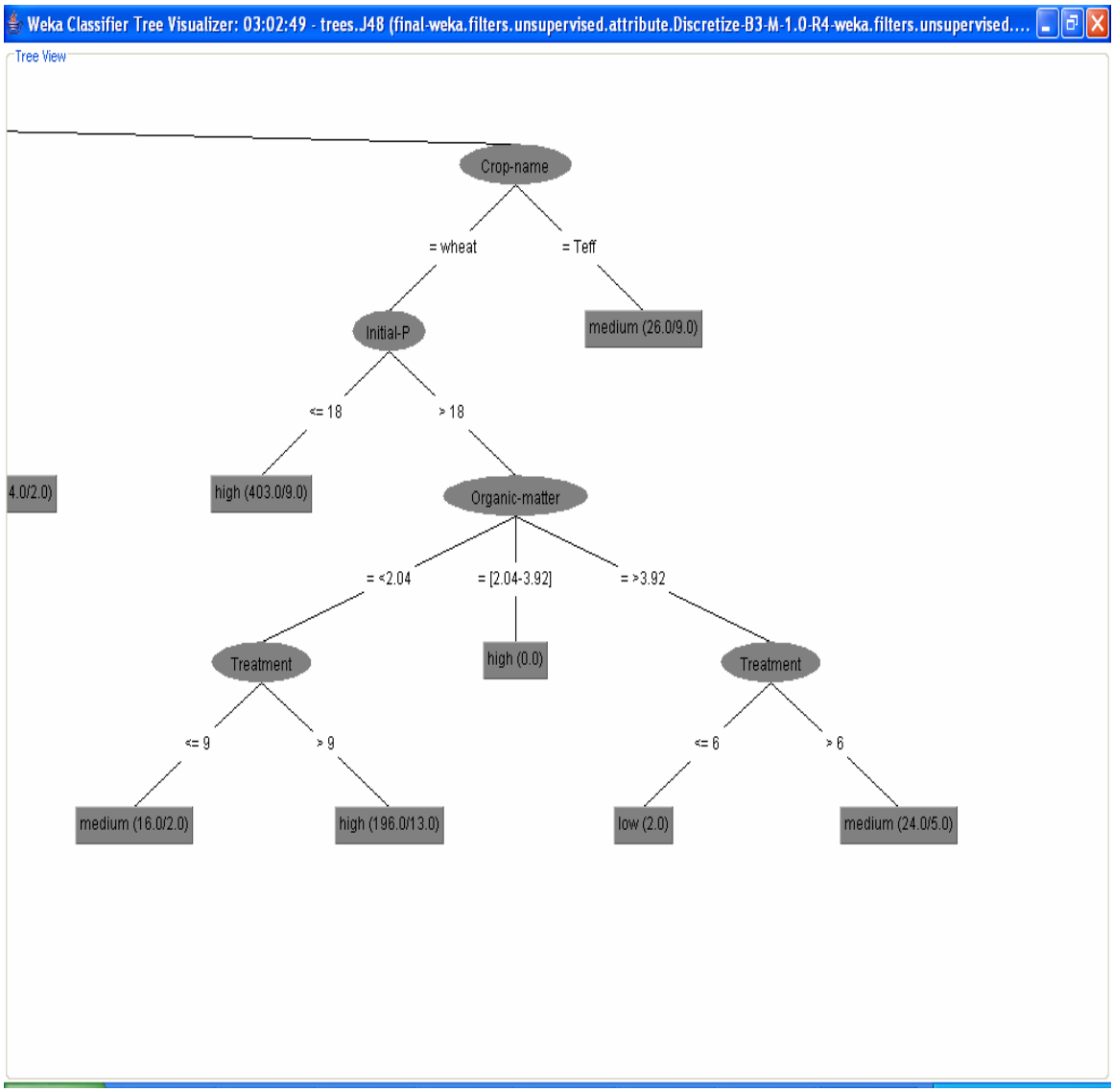


Figure 5.6 Part of decision tree generated by modifying some of the parameters in the program

Generating Rules

Rules were generated from the last decision tree built with better accuracy and better interpretability. Generating rules from decision tree comes after successive experiments have been done in building the best decision tree model. Rules can be generated by traversing through the decision tree. According to (Bao, 2003) as cited by [22] it is possible to find out a number of rules simply by traversing the decision tree and generating a rule for each leaf and making combination of all the tests found on the path from the root to the leaf node. Rules show the relationship between independent variables and dependent variables. The following are some of the rules that are taken out of the decision tree developed in the third experiment. In these rules, numbers in brackets show correctly and incorrectly classified data respectively.

Rule 1

If treatment <=5 and crop-name= wheat and wereda=wogera and initial-p= (5-13] and Total nitrogen (0.14-0.24] then low **(40.0/6.0)**

Rule 2

If treatment <=5 and crop-name = tef and wereda =Tahtaykoraro then low yield **(346.0/4.0)**

Rule 3

If treatment <=5 and crop-name = tef and wereda = Achefer then low **(228.0/9.0)**

Rule 4

If treatment >5 and Initial-P<=7.2 and wereda=wogera and Total-nitrogen = <0.14 and treatment >11 then high yield **(142.0/2.0)**

Rule 5

If treatment >5 and wereda = woreilu and Initial-P<=9.65 then medium (**230.0/29.0)**

Rule 6

If treatment >5 and wereda = woreilu and Initial-P>9.65 then high **(52.0/18.0)**

5.3.2 Results and Discussion

Analyzing rules that are obtained from the selected, best decision tree model is the next step that needs attention in the process of knowledge discovery. The analysis is categorized into three groups. The first one is analysis of rules that are categorized under high class label. The second category is about rules categorized under medium class label and the third is rules under low class labels. They are discussed in the following ways.

Among the rules generated by Weka in the third experiment, the following **rules are likely to bring high yield** in which majority of the records, above 80%, are classified correctly.

1. Treatment >11 and crop-name = wheat and initial-P <=7.2 and wereda = Wogera and total nitrogen < 0.14 (142/2) **(98.6%).**
2. Treatment > 7 and crop-name = wheat and wereda=Yilmana-Densa (198/47) **(81%)**
3. Treatment > 13 and crop-name = tef and wereda = Dejen (164/19) (90%)
4. Treatment >5 and crop-name = tef and wereda = Warajarso and Initial-p>4.47(244/2) **(99%)**
5. Treatment >5 and crop-name = wheat and Initial-P = (15.7-18] (403/9) **(98%)**
6. Treatment > 9 and crop-name = wheat and Initial-P >18 and organic matter < 2.04 (196/13) **(94%)**
7. Treatment < 5 and wereda = Kuyu and initial-P > 6 and Organic matter = [2.04-3.92](10/2) **(83%)**
8. Treatment < 5 and wereda = Hetosa and Organic matter [2.04-3.92] and soil-pH > 8.5 and initial-P = (0.15-0.17] (10) (**100%**)
9. Treatment >5 and wereda = Hetosa and initial-p<=0.2 and total nitrogen = [0.14-0.24] and organic matter = [2.04-3.92] and soil-pH >8.5 (88/9) (**90%**)
10. Treatment >5 and wereda = Hetosa and initial-p = (0.18-0.2] and total nitrogen = [0.14-0.24] and organic matter = [2.04-3.92] and soil-pH [5-8.5] (22/2) (**91%**)

From the above rules one can predict the amount of fertilizer to be added to produce high yield in the following ways.

- If crop-name = wheat, initial-p<=7, wereda = wogera and total nitrogen<0.14, then treatment >11.
- If crop-name = wheat and wereda=Yilmana-Densa, then Treatment > 7
- If crop-name = tef and wereda = Dejen ,then Treatment > 13
- If crop-name = tef and wereda = Warajarso and Initial-p>4.47, then Treatment >5.
- If crop-name = wheat and Initial-P = (15.7-18], then Treatment >5.
- If crop-name = wheat, Initial-P >18 and organic matter < 2.04, then Treatment > 9.
- If wereda = Kuyu, initial-P > 6 and Organic matter = [2.04-3.92], then Treatment < 5.
- If wereda = Hetosa, Organic matter= [2.04-3.92], soil-pH > 8.5 and initial-P = (0.15-0.17], then Treatment < 5.

It is apparent from the above rules and other rules attached in Annex C that, in many of the weredas, if treatment is greater than five, the likely hood of getting high yield irrespective of the kind of grain produced is above 80%.This fact is clearly indicated by the numbers in brackets shown in the above rules. For example in rule 1 above, out of 144 records that are likely to be classified as high,142 records which is about 99% is classified correctly as high but the rest 2 are classified incorrectly by the selected best model. The other fact deduced from the above rules is that average level of initial-p and organic matters have their own contribution to enhance both wheat and tef yield, this is particularly seen in rule number 4, 5 and 6.

On the contrary, rule number 7 and 8 show that one can produce high wheat and tef yield by applying treatment that is less than five. This is particularly true in wereda Hetosa and Kuyu. This idea further strengthens that initial-p, organic-matter and also soil-pH, indicated in rule 7 and 8, have their own contribution in the development of these crops. To add more, rule 8 illustrates that average organic-matter, initial-P content and soil pH >8.5 can produce high yield though treatment level is less than 5.

Rules that are likely to Classify Records as <u>Medium</u>

These selected rules classify majority of the records, above 80%, correctly.

1. Treatment <=5 and crop-name=wheat and wereda=Yilman-Densa (100/19) **(84%)**
2. Treatment >5 and crop-name = wheat and Initial-p<=9.65(230/29) **(89%)**
3. Treatment < = 4 and crop-name = wheat and wereda = Lume and Initial-p >10 (32/8) **(80%)**
4. Treatment <= 5 and crop-name = wheat and wereda = Hetosa and Organic matter > 3.92(55/10) **(85%)**
5. Treatment < = 4 and crop-name = wheat and wereda = Woreilu and organic-matter < 2.04 and soil-pH= [5-8.5] and Initial-p <= 4.4 (24/7) **(77%)**
6. Treatment < =5 and crop-name = wheat and wereda = Wogera and Initial-p >13.3 (20/7) **(74%)**
7. Treatment = [4-5] and crop-name = wheat and wereda = woreilu and organic matter<2.04(18/2) **(90%)**
8. Treatment <=5 and crop-name = tef and wereda = Wara-Jarso and Initial-p >4.74 and soil-pH [5-8.5] (8) **(100%)**

From these rules also, one can predict the amount of fertilizer to be added to get medium yield in the following ways:

- If crop-name=wheat and wereda=Yilman-Densa, then Treatment <=5.
- If crop-name = wheat and Initial-p<=9.65, then Treatment >5.
- If crop-name = wheat and wereda = Lume and Initial-p >10, then Treatment < = 4.
- If crop-name = wheat and wereda = Hetosa and Organic matter > 3.92, then Treatment <= 5.
- If crop-name = wheat and wereda = Woreilu and organic-matter < 2.04 and soil-pH= [5-8.5] and Initial-p <= 4.4, then Treatment < = 4.
- If crop-name = wheat and wereda = Wogera and Initial-p >13.3, then Treatment < =5.
- If crop-name = wheat and wereda = woreilu and organic matter<2.04,then Treatment = [4-5].
- If crop-name = tef and wereda = Wara-Jarso and Initial-p >4.74 and soil-pH [5-8.5] Treatment <=5.

Majority of the above rules depict that application of treatment less than 5 can produce medium yield if soil-pH, initial-p and organic matter are in a normal condition. It is indicated that soil pH = [5-8.5] is a normal condition for crops to grow and to be fertile as it is also suggested by literatures. This is proved in different weredas, particularly, in Wara-Jarso and Woreilu. Besides this, it is illustrated that high amount of initial-p can compensate less amount of fertilizer that is <5 to get medium yield. It is also deduced that high treatment and less amount of initial-p can generate medium yield as it is shown in rule number 2.

Rules that are Likely to Classify Records as <u>Low</u>

1. Treatment <= 5 and crop-name = wheat and wereda = Wegera and Initial-p = (5-13] and Total nitrogen = (0.14-0.24] (40/6) **(87%)**
2. Treatment <= 5 and crop-name = tef and wereda = Kuyu and Initial-p <= 6 (20/1) **(95%)**
3. Treatment <= 5 and crop-name = tef and wereda = Achefer (228/9) **(96%)**
4. Treatment < = 5and crop-name = tef and wereda = Tahtaykoraro (346/4) **(99%)**
5. Treatment < =5 and crop-name = tef and wereda = Dejen and soil-pH < 5 (40/4) **(91%)**
6. Treatment < = 4 and crop-name = tef and wereda = Dejen and soil-pH [5-8] (48/14) **(77%)**
7. Treatment = [5-10] and crop-name = tef and wereda = Achefer and Organic matter = [2.04-3.92] and Initial-p < 4.62 (52/5) **(91%)**
8. Treatment = [5-16) and crop-name = tef and wereda =Achefer and Organic matter = [2.04-3.92] and initial-p = (7-9] (22/2) **(92%)**
9. Treatment = [5-10) and crop-name = tef and wereda = Tahtaykoraro and initial-p <=4.28 and soil-pH= [5-8.5] and initial-p <= 0.38 (10) **(100%)**
10. Treatment >5 and crop-name = tef and wereda = Achefer and Organic matter = [2.04-3.92] and initial-p < 4.52 (52/5) **(91%)**
11. Treatment = (5-6) and crop-name = tef and wereda = Dejen (14/3) **(82%)**
12. Treatment > 5 and crop-name = wheat and wereda = Wogera and initial-p > 9 and Organic matter >3.92 (30/7) **(81%)**

As it is demonstrated above, application of treatment less than 5 bring about low yield in different weredas. This thing again will be aggravated if other factors like initial-p, soil-pH and organic matter are in lesser amount. Rule 2 and 5 prove this fact. Different literatures also suggest that soil-pH < 5 is not favorable for crop to be grown especially for tef and wheat in this case study.

The other thing that is shown is producing low yield having treatment greater than 5 and very less amount of initial-p and medium level organic matter. This is true particularly to Achefer and Tahtaykoraro suggesting that initial-p and organic matter are important in yield increase for tef in these two weredas.

The above analysis can be supported by the following graph.

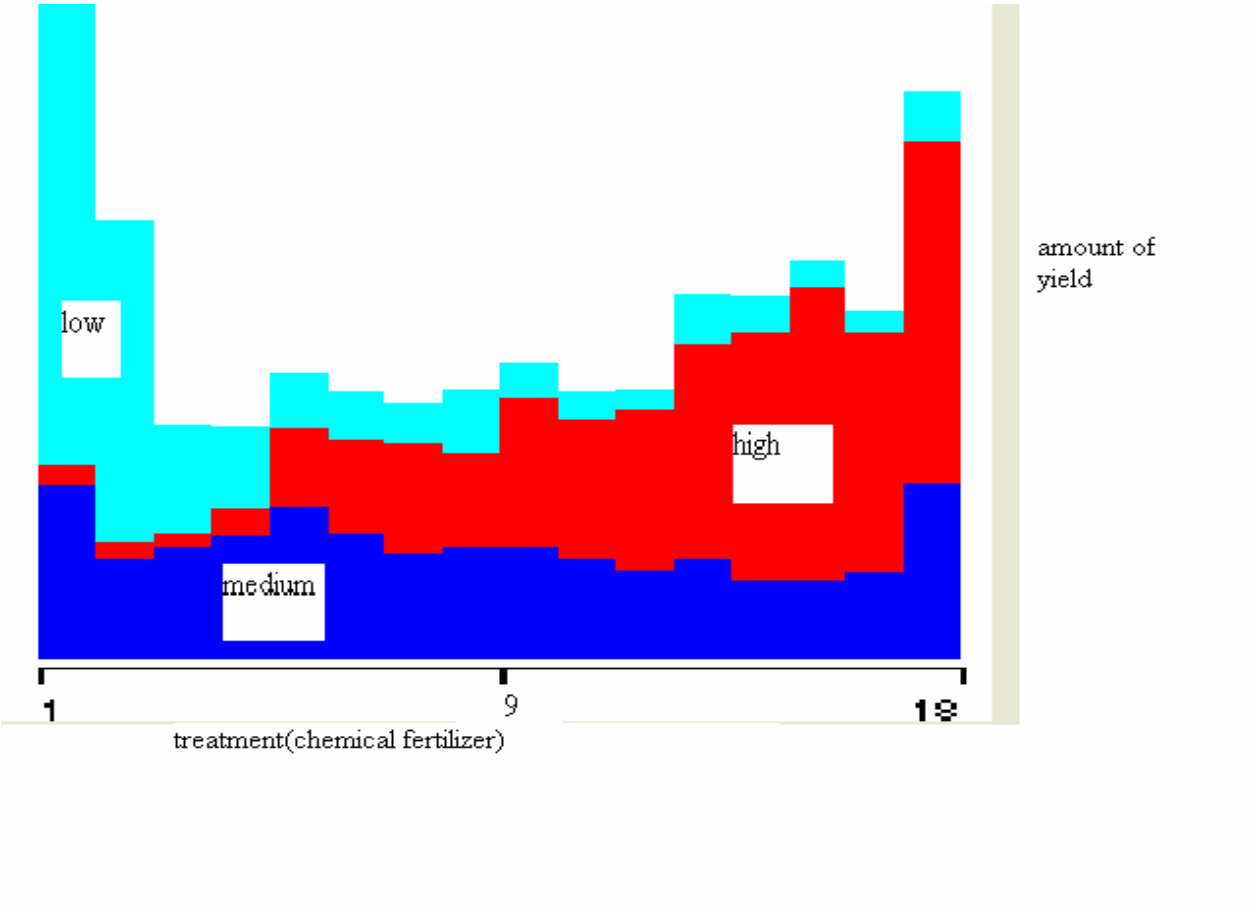


Figure5.7 treatments and their effect in the amount of grain yield

The graph in figure5.7 sums up the idea mentioned in the result and discussion session. As it is observed, low grain yield covers the highest percentage as compared to high and medium yield in cases of treatments less than five. On the other hand, high yield is in the highest percentage in cases of treatments greater than 9. Besides, medium yield is in the highest percentage if treatment is about in between of five and nine. Generally, what is observed here is, as the amount of treatment increases, the amount of grain yield also increases and also grain yield decreases if the amount of treatment decreases though there are other factors that aggravate this situations as it is indicated above. From this, one can conclude that treatment and grain yield have direct relationship especially in case of high and low level of grain yield. Regarding medium yield, it can be gained in both the situation of low level of treatment and high level of treatment. Even the mount of medium yield is high in case of treatment less than 5 that is because of other factors like initial-p, organic-matter and soil-pH as it is shown in the result of the decision tree.

5.4 Evaluation

After building and validating models, the next step was assessing the degree to which the model meets the business objectives. Besides, in this section of the research, steps executed to construct the model have been reviewed. In this research, the model building session has two parts one is building decision tree that can generate the soundest rule sets and the other is validating the rules generated using the data that were assigned for validation purpose and the result of validation test is presented in the form of confusion matrix and accuracy level. In the decision tree building part, different experiments were conducted only to generate rules that are sound to the business objective. Hence, according to the domain expert, the variables used to construct the best tree were worthwhile to determine status of grain yield and the rules in the constructed, best decision tree were very helpful to develop guideline for fertilizer recommendation. Regarding validation of the decision trees constructed in different experiments, 10-fold cross validation and use training set techniques were used and these techniques automatically classify the whole dataset into training and validation dataset. Finally, the one with better accuracy and sound rule sets was taken for analysis purpose.

The decision tree depicted in figure 5.6 was the best decision tree in that it supplied the soundest rules to the problem domain and it reduced the number of misclassified records of different level

of grain yield. But, it is the researcher belief that more experimentation could yield more meaningful trees that can generate more sound rules. The validation test conducted on the results of the selected decision tree revealed that the over all accuracy of classifying grain yield is 85.6%.

From the research result, it is both the researcher and the domain expert belief that decision tree is a very helpful tool in detecting important variables that can classify grain yield into high, medium and low. In the process of determining adequate chemical fertilizer, the decision maker in the office of National Soil Testing Center can trace the decision tree developed to come across all the relevant combination of soil parameters and treatments that can bring high, medium and low grain yield. Besides, the decision maker will be in a position to know how much fertilizer (treatment) has to be added having the different level of soil-pH, organic matter, initial available phosphorous and so on.

Generally, it is apparent from the above discussion that application of data mining technology is helpful in the analysis of fertilizer related data. It can show the importance of initial soil fertility status that is amount of soil pH, organic matter and others mentioned above as well as different level of treatments (fertilizer) to produce grain yield. This in turn helps decision makers in the process of developing guidelines for fertilizer recommendation.

Chapter Six: Conclusion and Recommendation

6.1 Conclusion

Agriculture is the foundation of Ethiopian economy and accounts for half of gross domestic product (GDP), 60% of exports, and 80% of total employment. However, it is plagued by periodic drought, soil degradation caused by overgrazing, deforestation, high population density and the like. Above all, nutrient depletion as a result of cultivating mono-crop throughout different years is the aggravated problem, here in Ethiopia. To curb problems of this sort, different measures have been taken like applying chemical fertilizer.

National Soil Testing Center is the responsible organ in the management of fertilizer or in developing guideline for fertilizer recommendation based on soil fertility status so as to minimize the risk related to low and excess fertilizer use and arriving at unwanted result. To achieve this objective a number of field-experimental data has been collected, stored and analyzed. However, the analysis technique is not sophisticated.

Data mining is an advanced technique in the analysis of huge datasets and it can easily shows the relationship between variables.The objective of this research was hence to explore the potential application of data mining technology at National Soil Testing Center, for developing decision tree based classification model to classify wheat and tef grain yield as high, medium and low. This classification was done not for the sake of classifying objects rather to identify the determinant factors that affect high, medium and low grain yield. This supports decision makers in the process of identifying best proportion of chemical fertilizer and other factors like soil-pH, initial-p, and organic-matter to get high yield.

A dataset having 5399 records of fertilizer related issue was used to build and test decision tree classification models. Different models were built for comparison sake and one with high accuracy and sound rule is chosen as a best model. From this best model, the determinant variables that affect grain yield were identified. These are different level of ’soil-pH’, ’organic-matter’, ’initial-P’, ‘total-nitrogen’ and ‘treatment’. The classification accuracy of the decision tree was tested, and it showed an accuracy of 85%.

From the result, one can conclude that data mining particularly decision tree is a very fantastic tool to show the relationship between variables and identify determinant variable that affect grain yield. The result also indicates that to obtain higher yield, application of treatment number more than 5 is crucial though the amount of treatment that should be applied varies in different wereda. This shows that more nitrogen fertilizer is needed to increase grain yield since treatment level above 5 holds more nitrogen content than phosphorus. By doing this, it is manageable to assure the country's attempt for food security. Therefore, more nitrogen fertilizer should be imported and distributed to the farmers than phosphorus fertilizers. Besides, it is observed that the lower total nitrogen or organic matter content of the soil, the higher treatment number required to be applied to obtain higher grain yield.

6.2 Recommendation

In this research work an attempt was made to apply data mining technology to classify grain yield into high, medium and low. Based on the results found in this research, recommendations in relation to fertilizer management analysis are suggested.

Even though, this research was conducted as an academic exercise its results are found to be promising to be applied in addressing practical problems related to fertilizer data analysis. Hence, based on the findings of this study, the following recommendations are forwarded:

- National soil testing center is recommended and encouraged to collect more experimental data to get reliable result.
- Using the dataset of this research and other additional datasets one can investigate determinant factors other than the one mentioned in this research that can bring high yield.
- Application of data mining technology in the area of agriculture is very promising though it is a new discipline in this area. So it needs others to work on it.
- The present experiment provides a way to only classify grain yield as high, medium and low. Further experiments should be able to find out other interesting pattern of the data set.
- Further tests by extending this research or viewing it from different angel should also be made on the same dataset to see if other techniques like neural network or a combination of one or more techniques will result in a better classification performance.

References

1. Agriculture in Ethiopia (2008). Available URL: http://en.wikipedia.org/wiki/Agriculture_in_Ethiopia (visited at August 20, 2008)
2. Agriculture (2000). Available URL:http://countrystudies.us/ethiopia/87.htm, (Visited at August 5, 2008)
3. Armstrong L.J., Diepeveen D. and Maddern R. (2007). The Application of Data Mining Techniques to Characterize Agricultural Soil Profiles. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 70. Peter Christen, Paul Kennedy, Jiuyong Li, Inna Kolyshkina and Graham Williams, Ed. Australian Computer Society, Inc.Gold Coast, Australia. Available URL: http://crpit.com/confpapers/CRPITV70Armstrong.pdf. (Visited at Oct 5, 2008)
4. Balesh Tulema, (2005). Integrated Plant Nutrient Management in Crop Production in The Central Ethiopian High Land. Department of international environment and development studies, Noragric. Norway. Available URL: http://www.umb.no/noragric/publications/phdtheses/Balesh_introduction.pdf (visited at August 20, 2008)
5. Belay Semane, (2004). Fertilizer Demand and Status of Regional Soil Testing Laboratories. *Optimizing Fertilizer Use in Ethiopia*. Proceeding of the workshop on phosphorous soil test calibration study in Hetosa Wereda, Arsi Zone,December 10,2002, Addis Ababa, Ethiopia:Sasakawa Global 2000.
6. Berry, M. J. and Linoff, G. (2004). *Data Mining Techniques for Marketing, Sales and Customer Relationship Management*. Wiley publishing,Inc.Indiana polis,Indiana.
7. Brandy N. and Weil R. (2002). *The Nature and Properties of Soils*. Thirteenth Edition.
8. Chemical Fertilizer or Organic Fertilizer (1998). Available URL: http://www.ecochem.com/t_faq9.html, (visited at August 10, 2008)
9. Christy Spector (2001). About Soil pH. Available at URL: http://soil.gsfc.nasa.gov/soil_pH/plant_pH.htm.(visited at Oct 10, 2008)
10. CRISP-DM. (2000).CRISP-DM 1.0: Step-by-Step Data Mining Guide. Available URL: http://www.crisp-dm.org. (Visited at August 25, 2008)

11. Fayyad, Usma, Piatetsky-shpiro, G. and Smyth, P. (1996). From Data Mining to Knowledge Discovery an overview. *In advances in knowledge discovery and data mining*. Fayyad Et al (Eds) MIT press. Available URL: http://citeseer.nj.nec.com/fayyad96from.html (Visited at August 17, 2008).
12. Fertilizer (2008).Available URL:http://en.wikipedia.org/wiki/Fertilizer
13. Gashaw Mulatu, (2004). *Application of Data Mining Technology to Support Insolvency Prediction at Ethiopian Telecommunication Corporation*. Master's Thesis. Addis Ababa University. Addis Ababa.
14. Girma Belew (2004). *The Role of Data mining in the Risk Assessment Custom(with special reference to Ethiopian Customs Authority).* Master's Thesis. Addis Ababa University. Addis Ababa.
15. Girma Gebre Kidane (2004).Review of Fertilizer Trial Activities of the Crop Production and Regulatory Department (CPPRD) of the Ministry of Agriculture (MoA).*Optimizing fertilizer use in Ethiopia*. Proceeding of the workshop on phosphorous soil Test Calibration study in Hetosa Wereda, Arsi Zone, December 10, 2002, Addis Ababa, Ethiopia: Sasakawa Global 2000.
16. Giudici P. (2003). *Applied Data mining Statistical Method for Business and Industry*. John Wily & Sons LTD, Faculty of Economics. University of Pavia. Italy
17. Hailu Tefra and Seyfu Ketema (2001). Production and Importance of Tef in Ethiopian Agriculture. *Narrowing the Rift, research and development*. In:Hailu T.,Getachew B.and Mark S. (ed.).Ethiopian Agricultural Organization. Proceedings of the international Workshop on genetics and improvement, Debre Zeit,Ethiopia.16-19 October 2000.
18. Han F. and Kamber M.,(2001). *Data Mining, Concepts and Techniques*, Morgan Kaufmann publishers, academic press. San Francisco
19. Hand D., Mannila H., Smyth P. (2001). *Principles of Data Mining.* The MIT press.Cambridge.
20. Julio H. and Carljos B. (1999). *Estimating Rates of Nutrient Depletion in Soils of Agricultural Lands of Africa*. International Fertilizer Development Center. Muscle shoals, Alabama 35662, U.S.A.

21.Klerfors, D. (1998). Artificial Neural Network. Available at URL:http://www.burks.brighton.ac.uk/burks/foldoc/42/7.htm (Visited at September 25, 2008)

22. Leul Woldu. (2003). *The Application of Data Mining in Crime Prevention: The case of Oromia Police Commission.* Master's Thesis. Addis Ababa University. Addis Ababa.

23. Ninomiya S. (2004). Successful Information Technology (IT) for Agriculture and Rural Development. National Agricultural Research Centre. National Agricultural Research Organization Kannondai, Tsukuba, Ibaraki 305-8666, Japan. Available URL: http://www.agnet.org/library/eb/549/(Visited at Augest 20, 2008)

24. Onwveme I.C. and Sinsha T.D, (1991). *Field Crop Production in Tropical Africa*.

25. Piccinin D.(2002). More about Ethiopian Food: Teff. Department of Nutrition and Food Service from an interview with Tsegazeab Woldetatios, PhD, Agriculture Conctract Interpeter at Harborview Medical Center. In: Christine Wilson (ed). Available URL:http://www.ethnomed.org/cultures/ethiop/teff.html/ (Visited at Augest 27, 2008)

26. Richard K and Eibe F. (2007). Weka Explorer User Guide for Version 3-4.Available URL: http://www.cs.waikato.ac.nz/ml/weka.(Visited at September 20, 2008)

27. Seidman C. (2001). *Data mining with Microsoft SQL server 2000*.Microsoft press.Washington.

28. Seyfu Ketema (1997).Promoting The Conservation And Use Of Underutilized And Neglected Crops; International Plant Genetic Resources Institute, Addis Abeba, Ethiopia.AvailableatURL: http://www.bioversityinternational.org/fileadmin/bioversity/publications/pdfs/279.pdf. (Visited at September 6, 2008)

29. Singnal A. (2007). *Data Warehousing and Data mining Techniques for Cyber Security.* NIST, Computer Security Division.USA

30. Taye Bekele, Yesuf Assen, Sahlemedhin Sertsu, Amanuel Gorfu, Mohammed Hassena, D.G. Tanner, Tesfay Tessemma and Takele Gebre,(2002). *Optimizing fertilizer use in Ethiopia: Correlation of soil analysis with fertilizer response in Hetosa Wereda,Arsi Zone*. Addis Ababa:Sasakawa-Global 2000.

31. Tesfay Teklay,(2005). *Organic Inputs from Agroforestry Trees on Farms for Improving soil Quality and crop productivity in Ethiopia.* Doctorial thesis No.2005: 122 faculty of forest sciences.

32. Thearling, K. (2003). An Introduction to Data Mining.USA. http://www.thearling.com/text/dmwhite/dmwhite.htm. (Visited at September 10, 2008).

33. Tibebe Beshah(2005). *Application of Data Mining Technology to Support Road Traffic Accident Severity Analysis at Addis Ababa Traffic Office.* Master's Thesis. Addis Ababa University. Addis Ababa.

34. Tittonella P., Shepherd K.D., Vanlauwe B. and Giller K.E (2007). Unraveling The Effects of Soil and Crop Management on Maize Productivity in Smallholder Agricultural Systems of Western Kenya—an application of classification and regression tree analysis. Agriculture, Ecosystems, Environment. Volume 123, Issues 1-3, January 2008, pages 137-150.Available at URL: http://www.sciencedirect.com/science?(Visited at November 20,2008)

35. Two Crows Corporation (1999). Introduction to Data Mining and Knowledge Discovery. Available at URL: http://www.twocrows.com/intro-dm.pdf. (Visited at October 3, 2008)

36. Wassie Haile (1999).*Isolation and Characterization of Phosphate Solublizing bacteria from some Ethiopian soils and their effect on the Growth of FABA bean*. Thesis

37. Witten, I.H and Frank, E. (2000). *Data Mining practical machine learning tools and techniques with java implementations.* Morgan Kaufmann publishers. San-Francisco.

38. Worku Denbel (2008).*Survey of Wheat Stem Rust in Central and South Eastern Shewa and Identification of Sources of Resistance in Commercial Cultivates and Local Accessions of Wheat.* M.Sc. Thesis.

39. Yesuf Assen (2004). Principles and Concept of Soil Testing. *Optimizing fertilizer use in Ethiopia.* Proceeding of the workshop on phosphorous soil Test Calibration study in Hetosa Wereda, Arsi Zone, December 10,2002, Addis Ababa, Ethiopia: Sasakawa Global 2000.

40. Yesuf Assen, (2006). Nitrogen Fertilizer Requirement of Bread Wheat (*Triticum aestivum L.*) *After Legume pre-cursor crops in Arsi Zone of Ethiopia*. Ethiopian Journal of Natural Resources. 8 (2):217-228.

Appendices

Annex A: Rules from experiment one

Annex A

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:final-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R4-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R6-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R7-weka.filters.unsupervised.attribute.Remove-R2
Instances: 5399
Attributes: 8
Crop-name
Wereda
Soil-pH
Initial-P
Total-nitrogen
Organic-matter
Treatment
Class
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

Treatment <= 5
| Crop-name = wheat
| | Wereda = Wogera
| | | Initial-P <= 13.342
| | | | Total-nitrogen = <0.14
| | | | | Soil-pH = <5: low (10.0/3.0)
| | | | | Soil-pH = [5-8.5]: medium (10.0/3.0)
| | | | | Soil-pH = >8.5: medium (0.0)
| | | | Total-nitrogen = [0.14-0.24]
| | | | | Initial-P <= 5.2
| | | | | | Treatment <= 3: low (6.0/2.0)
| | | | | | Treatment > 3: medium (4.0)
| | | | | Initial-P > 5.2: low (40.0/6.0)
| | | | Total-nitrogen = >0.24: low (10.0/1.0)
| | | Initial-P > 13.342: medium (20.0/7.0)
| | Wereda = Woreilu
| | | Organic-matter = <2.04
| | | | Treatment <= 4
| | | | | Soil-pH = <5: medium (0.0)
| | | | | Soil-pH = [5-8.5]
| | | | | | Initial-P <= 4.4: medium (24.0/7.0)
| | | | | | Initial-P > 4.4
| | | | | | | Initial-P <= 22: low (16.0/2.0)
| | | | | | | Initial-P > 22
| | | | | | | | Treatment <= 3: medium (6.0/2.0)

Annex B: Rules from experiment two

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     final
Instances:    5399
Attributes:   9
              Crop-name
              Region
              Wereda
              Soil-pH
              Initial-P
              Total-nitrogen
              Organic-matter
              Treatment
              Class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Treatment <= 5
|  Crop-name = wheat
|  |  Organic-matter <= 1.41
|  |  |  Treatment <= 4
|  |  |  |  Soil-pH <= 7.09: medium (104.0/13.0)
|  |  |  |  Soil-pH > 7.09
|  |  |  |  |  Total-nitrogen <= 0.05: low (8.0)
|  |  |  |  |  Total-nitrogen > 0.05: medium (32.0/4.0)
|  |  |  Treatment > 4: medium (36.0/5.0)
|  |  Organic-matter > 1.41
|  |  |  Soil-pH <= 8.2
|  |  |  |  Region = Amahara
|  |  |  |  |  Soil-pH <= 5.52
|  |  |  |  |  |  Total-nitrogen <= 0.18
|  |  |  |  |  |  |  Treatment <= 2
|  |  |  |  |  |  |  |  Soil-pH <= 5.39: low (8.0)
|  |  |  |  |  |  |  |  Soil-pH > 5.39
|  |  |  |  |  |  |  |  |  Wereda  = Wogera: low (8.0/1.0)
|  |  |  |  |  |  |  |  |  Wereda  = Woreilu: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Debreilias: medium (8.0/1.0)
|  |  |  |  |  |  |  |  |  Wereda  = Yilmana-Densa
|  |  |  |  |  |  |  |  |  |  Treatment <= 1: medium (2.0)
|  |  |  |  |  |  |  |  |  |  Treatment > 1: low (2.0)
|  |  |  |  |  |  |  |  |  Wereda  = Lume: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Hetosa: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Dejen: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Achefer: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Tahtaykoraro: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Alefa: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Kuyu: medium (0.0)
|  |  |  |  |  |  |  |  |  Wereda  = Wara-Jarso: medium (0.0)
|  |  |  |  |  |  |  Treatment > 2: medium (47.0/17.0)
```

```
|   |   |   |   |   |   Total-nitrogen > 0.18: medium (42.0/8.0)
|   |   |   |   |   Soil-pH > 5.52
|   |   |   |   |   |   Total-nitrogen <= 0.06
|   |   |   |   |   |   |   Treatment <= 3
|   |   |   |   |   |   |   |   Treatment <= 1: medium (2.0)
|   |   |   |   |   |   |   |   Treatment > 1: high (4.0/1.0)
|   |   |   |   |   |   |   Treatment > 3: medium (4.0)
|   |   |   |   |   |   Total-nitrogen > 0.06
|   |   |   |   |   |   |   Treatment <= 4
|   |   |   |   |   |   |   |   Wereda  = Wogera
|   |   |   |   |   |   |   |   |   Total-nitrogen <= 0.14: medium (8.0/2.0)
|   |   |   |   |   |   |   |   |   Total-nitrogen > 0.14: low (48.0/7.0)
|   |   |   |   |   |   |   |   Wereda  = Woreilu
|   |   |   |   |   |   |   |   |   Soil-pH <= 6.52: medium (8.0/3.0)
|   |   |   |   |   |   |   |   |   Soil-pH > 6.52
|   |   |   |   |   |   |   |   |   |   Initial-P <= 11.084: low (32.0/4.0)
|   |   |   |   |   |   |   |   |   |   Initial-P > 11.084: medium (16.0/7.0)
|   |   |   |   |   |   |   |   Wereda  = Debreilias: low (35.0/15.0)
|   |   |   |   |   |   |   |   Wereda  = Yilmana-Densa: low (0.0)
|   |   |   |   |   |   |   |   Wereda  = Lume: low (0.0)
|   |   |   |   |   |   |   |   Wereda  = Hetosa: low (0.0)
|   |   |   |   |   |   |   |   Wereda  = Dejen: low (0.0)
|   |   |   |   |   |   |   |   Wereda  = Achefer: low (0.0)
|   |   |   |   |   |   |   |   Wereda  = Tahtaykoraro: low (0.0)
|   |   |   |   |   |   |   |   Wereda  = Alefa: low (0.0)
|   |   |   |   |   |   |   |   Wereda  = Kuyu: low (0.0)
|   |   |   |   |   |   |   |   Wereda  = Wara-Jarso: low (0.0)
|   |   |   |   |   |   |   Treatment > 4
|   |   |   |   |   |   |   |   Wereda  = Wogera
|   |   |   |   |   |   |   |   |   Initial-P <= 14.28: low (12.0/4.0)
|   |   |   |   |   |   |   |   |   Initial-P > 14.28: medium (2.0)
|   |   |   |   |   |   |   |   Wereda  = Woreilu
|   |   |   |   |   |   |   |   |   Organic-matter <= 1.95: medium (12.0/2.0)
|   |   |   |   |   |   |   |   |   Organic-matter > 1.95: low (2.0)
|   |   |   |   |   |   |   |   Wereda  = Debreilias: medium (8.0/1.0)
|   |   |   |   |   |   |   |   Wereda  = Yilmana-Densa: medium (0.0)
|   |   |   |   |   |   |   |   Wereda  = Lume: medium (0.0)
|   |   |   |   |   |   |   |   Wereda  = Hetosa: medium (0.0)
|   |   |   |   |   |   |   |   Wereda  = Dejen: medium (0.0)
|   |   |   |   |   |   |   |   Wereda  = Achefer: medium (0.0)
|   |   |   |   |   |   |   |   Wereda  = Tahtaykoraro: medium (0.0)
|   |   |   |   |   |   |   |   Wereda  = Alefa: medium (0.0)
|   |   |   |   |   |   |   |   Wereda  = Kuyu: medium (0.0)
|   |   |   |   |   |   |   |   Wereda  = Wara-Jarso: medium (0.0)
|   |   |   |   Region = Oromiya
|   |   |   |   |   Soil-pH <= 6.79
|   |   |   |   |   |   Initial-P <= 0.12
|   |   |   |   |   |   |   Treatment <= 4: medium (8.0)
|   |   |   |   |   |   |   Treatment > 4: high (2.0)
|   |   |   |   |   |   Initial-P > 0.12
|   |   |   |   |   |   |   Treatment <= 4
|   |   |   |   |   |   |   |   Initial-P <= 0.19: low (16.0/3.0)
|   |   |   |   |   |   |   |   Initial-P > 0.19
|   |   |   |   |   |   |   |   |   Initial-P <= 0.22: medium (8.0/2.0)
|   |   |   |   |   |   |   |   |   Initial-P > 0.22: low (12.0/4.0)
|   |   |   |   |   |   |   Treatment > 4: medium (9.0/3.0)
```

```
|  |  |  |  |  Soil-pH > 6.79: medium (70.0/12.0)
|  |  |  |  Region = Tigray: medium (0.0)
|  |  |  Soil-pH > 8.2
|  |  |  |  Soil-pH <= 8.56: medium (10.0/1.0)
|  |  |  |  Soil-pH > 8.56: high (10.0)
|  Crop-name = f
|  |  Total-nitrogen <= 0.05: low (316.0/1.0)
|  |  Total-nitrogen > 0.05
|  |  |  Wereda  = Wogera: low (0.0)
|  |  |  Wereda  = Woreilu: low (0.0)
|  |  |  Wereda  = Debreilias: low (0.0)
|  |  |  Wereda  = Yilmana-Densa: low (0.0)
|  |  |  Wereda  = Lume: low (0.0)
|  |  |  Wereda  = Hetosa: low (0.0)
|  |  |  Wereda  = Dejen
|  |  |  |  Total-nitrogen <= 0.1229
|  |  |  |  |  Soil-pH <= 6.08
|  |  |  |  |  |  Treatment <= 4: low (40.0/17.0)
|  |  |  |  |  |  Treatment > 4: medium (10.0/2.0)
|  |  |  |  |  Soil-pH > 6.08: low (20.0/1.0)
|  |  |  |  Total-nitrogen > 0.1229: low (30.0)
|  |  |  Wereda  = Achefer: low (228.0/9.0)
|  |  |  Wereda  = Tahtaykoraro
|  |  |  |  Treatment <= 4: low (24.0)
|  |  |  |  Treatment > 4
|  |  |  |  |  Total-nitrogen <= 0.083: medium (2.0)
|  |  |  |  |  Total-nitrogen > 0.083: low (4.0/1.0)
|  |  |  Wereda  = Alefa
|  |  |  |  Initial-P <= 2.46
|  |  |  |  |  Treatment <= 1: low (16.0)
|  |  |  |  |  Treatment > 1
|  |  |  |  |  |  Organic-matter <= 3.32
|  |  |  |  |  |  |  Total-nitrogen <= 0.19
|  |  |  |  |  |  |  |  Treatment <= 4
|  |  |  |  |  |  |  |  |  Organic-matter <= 1.24: low (12.0/4.0)
|  |  |  |  |  |  |  |  |  Organic-matter > 1.24
|  |  |  |  |  |  |  |  |  |  Treatment <= 2: low (2.0)
|  |  |  |  |  |  |  |  |  |  Treatment > 2: medium (4.0)
|  |  |  |  |  |  |  |  Treatment > 4: low (6.0)
|  |  |  |  |  |  |  Total-nitrogen > 0.19: medium (16.0/6.0)
|  |  |  |  |  |  Organic-matter > 3.32: low (24.0/3.0)
|  |  |  |  Initial-P > 2.46: medium (20.0/3.0)
|  |  |  Wereda  = Kuyu
|  |  |  |  Soil-pH <= 5
|  |  |  |  |  Initial-P <= 6.94: low (20.0/1.0)
|  |  |  |  |  Initial-P > 6.94
|  |  |  |  |  |  Total-nitrogen <= 0.244
|  |  |  |  |  |  |  Treatment <= 3: low (18.0/5.0)
|  |  |  |  |  |  |  Treatment > 3
|  |  |  |  |  |  |  |  Total-nitrogen <= 0.1773
|  |  |  |  |  |  |  |  |  Treatment <= 4: medium (2.0)
|  |  |  |  |  |  |  |  |  Treatment > 4: low (2.0)
|  |  |  |  |  |  |  Total-nitrogen > 0.1773
|  |  |  |  |  |  |  |  Soil-pH <= 4.7: medium (4.0/1.0)
|  |  |  |  |  |  |  |  Soil-pH > 4.7: high (4.0/1.0)
|  |  |  |  |  |  Total-nitrogen > 0.244: low (30.0/11.0)
```

```
|  |  |  |  Soil-pH > 5
|  |  |  |  |  Soil-pH <= 5.2: high (10.0/2.0)
|  |  |  |  |  Soil-pH > 5.2
|  |  |  |  |  |  Treatment <= 2: low (4.0)
|  |  |  |  |  |  Treatment > 2: medium (6.0/1.0)
|  |  |  Wereda  = Wara-Jarso
|  |  |  |  Soil-pH <= 5.33: low (10.0/2.0)
|  |  |  |  Soil-pH > 5.33
|  |  |  |  |  Treatment <= 4
|  |  |  |  |  |  Soil-pH <= 6.28
|  |  |  |  |  |  |  Treatment <= 2: high (4.0/1.0)
|  |  |  |  |  |  |  Treatment > 2: medium (4.0/1.0)
|  |  |  |  |  |  Soil-pH > 6.28: medium (8.0)
|  |  |  |  Treatment > 4: high (4.0)
Treatment > 5
|  Initial-P <= 15.702
|  |  Wereda  = Wogera
|  |  |  Organic-matter <= 2.89
|  |  |  |  Treatment <= 11
|  |  |  |  |  Treatment <= 6: high (2.0)
|  |  |  |  |  Treatment > 6: medium (10.0/3.0)
|  |  |  |  Treatment > 11: high (142.0/2.0)
|  |  |  Organic-matter > 2.89
|  |  |  |  Soil-pH <= 5.52
|  |  |  |  |  Treatment <= 12: medium (14.0)
|  |  |  |  |  Treatment > 12: high (12.0/2.0)
|  |  |  |  Soil-pH > 5.52
|  |  |  |  |  Treatment <= 10
|  |  |  |  |  |  Organic-matter <= 3.27: medium (10.0/2.0)
|  |  |  |  |  |  Organic-matter > 3.27: low (50.0/12.0)
|  |  |  |  |  Treatment > 10
|  |  |  |  |  |  Organic-matter <= 3.27: medium (16.0/3.0)
|  |  |  |  |  |  Organic-matter > 3.27
|  |  |  |  |  |  |  Organic-matter <= 3.7: low (16.0/2.0)
|  |  |  |  |  |  |  Organic-matter > 3.7: medium (64.0/11.0)
|  |  Wereda  = Woreilu
|  |  |  Initial-P <= 9.65: medium (230.0/29.0)
|  |  |  Initial-P > 9.65: high (52.0/18.0)
|  |  Wereda  = Debreilias
|  |  |  Treatment <= 10
|  |  |  |  Initial-P <= 6.42: medium (45.0/7.0)
|  |  |  |  Initial-P > 6.42
|  |  |  |  |  Soil-pH <= 5.39: medium (18.0/4.0)
|  |  |  |  |  Soil-pH > 5.39: high (27.0/10.0)
|  |  |  Treatment > 10
|  |  |  |  Treatment <= 17
|  |  |  |  |  Treatment <= 16
|  |  |  |  |  |  Treatment <= 14
|  |  |  |  |  |  |  Treatment <= 12
|  |  |  |  |  |  |  |  Treatment <= 11
|  |  |  |  |  |  |  |  |  Total-nitrogen <= 0.1643: medium (6.0/1.0)
|  |  |  |  |  |  |  |  |  Total-nitrogen > 0.1643: high (14.0/2.0)
|  |  |  |  |  |  |  |  Treatment > 11: high (20.0/2.0)
```

Number of Leaves : 250

Size of the tree : 448

Time taken to build model: 2.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	4322	80.0519 %
Incorrectly Classified Instances	1077	19.9481 %
Kappa statistic	0.6965	
Mean absolute error	0.1732	
Root mean squared error	0.3192	
Relative absolute error	39.5308 %	
Root relative squared error	68.2079 %	
Total Number of Instances	5399	

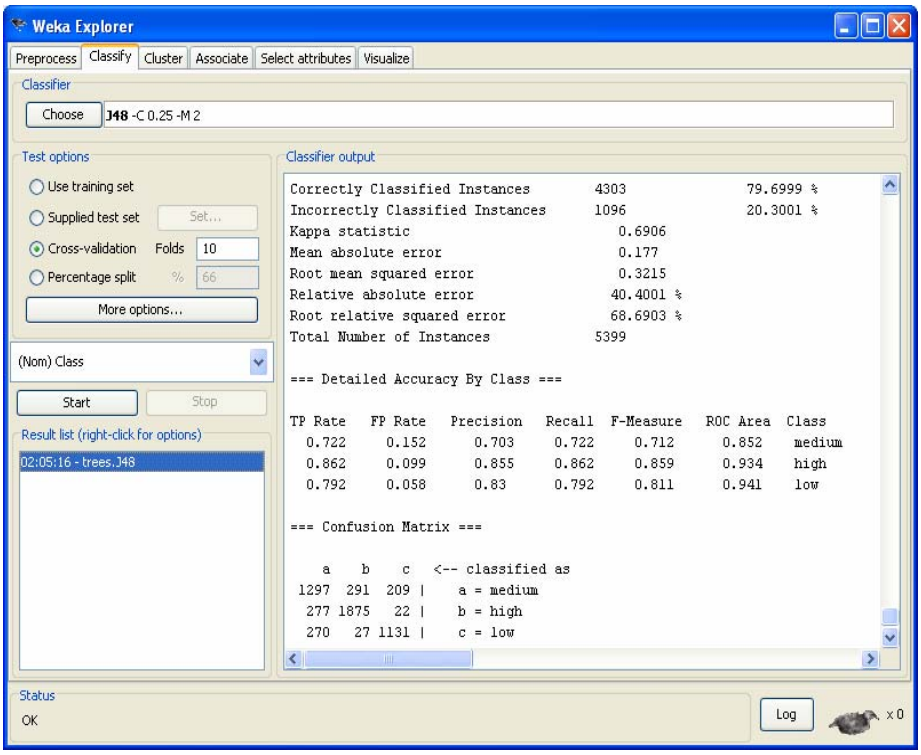
=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.715	0.146	0.709	0.715	0.712	0.851	medium
0.857	0.091	0.864	0.857	0.861	0.932	high
0.822	0.065	0.82	0.822	0.821	0.949	low

=== Confusion Matrix ===

a	b	c	<-- classified as
1284	282	231	a = medium
284	1864	26	b = high
242	12	1174	c = low

Here is the screen shot for the result of experiment two



Annex C: Rules from experiment three

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: final-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R4-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R6-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R7-weka.filters.unsupervised.attribute.Remove-R2
Instances: 5399
Attributes: 8
Crop-name
Wereda
Soil-pH
Initial-P
Total-nitrogen
Organic-matter
Treatment
Class
Test mode: evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree

Treatment > 5
| Initial-P <= 15.702
| | Wereda = Wogera
| | | Initial-P <= 7.15

```
|   |   |   |   Total-nitrogen = <0.14
|   |   |   |   |   Treatment <= 11
|   |   |   |   |   |   Treatment <= 6: high (2.0)
|   |   |   |   |   |   Treatment > 6: medium (10.0/3.0)
|   |   |   |   |   Treatment > 11: high (142.0/2.0)
|   |   |   |   Total-nitrogen = [0.14-0.24]
|   |   |   |   |   Initial-P <= 5.2
|   |   |   |   |   |   Treatment <= 12: medium (14.0)
|   |   |   |   |   |   Treatment > 12: high (12.0/2.0)
|   |   |   |   |   Initial-P > 5.2
|   |   |   |   |   |   Treatment <= 10
|   |   |   |   |   |   |   Treatment <= 6: medium (2.0)
|   |   |   |   |   |   |   Treatment > 6: low (8.0/3.0)
|   |   |   |   |   |   Treatment > 10: medium (16.0)
|   |   |   |   Total-nitrogen = >0.24: high (0.0)
|   |   |   Initial-P > 7.15
|   |   |   |   Initial-P <= 9.32: low (26.0/2.0)
|   |   |   |   Initial-P > 9.32
|   |   |   |   |   Treatment <= 10
|   |   |   |   |   |   Organic-matter = <2.04: low (0.0)
|   |   |   |   |   |   Organic-matter = [2.04-3.92]: medium (10.0/2.0)
|   |   |   |   |   |   Organic-matter = >3.92: low (30.0/7.0)
|   |   |   |   |   Treatment > 10: medium (64.0/14.0)
|   |   Wereda  = Woreilu
|   |   |   Initial-P <= 9.65: medium (230.0/29.0)
|   |   |   Initial-P > 9.65: high (52.0/18.0)
|   |   Wereda  = Debreilias
|   |   |   Treatment <= 10
|   |   |   |   Initial-P <= 6.42: medium (45.0/7.0)
|   |   |   |   Initial-P > 6.42
|   |   |   |   |   Initial-P <= 8.48
|   |   |   |   |   |   Initial-P <= 7.5
|   |   |   |   |   |   |   Initial-P <= 6.53
|   |   |   |   |   |   |   |   Treatment <= 8: high (10.0/2.0)
|   |   |   |   |   |   |   |   Treatment > 8: medium (8.0/3.0)
|   |   |   |   |   |   |   Initial-P > 6.53: medium (9.0/2.0)
|   |   |   |   |   |   Initial-P > 7.5: high (9.0/3.0)
|   |   |   |   |   Initial-P > 8.48: medium (9.0/2.0)
|   |   |   Treatment > 10
|   |   |   |   Treatment <= 17
|   |   |   |   |   Treatment <= 16
|   |   |   |   |   |   Treatment <= 14
|   |   |   |   |   |   |   Treatment <= 12: high (40.0/9.0)
|   |   |   |   |   |   |   Treatment > 12
|   |   |   |   |   |   |   |   Treatment <= 13: medium (20.0/4.0)
|   |   |   |   |   |   |   |   Treatment > 13
|   |   |   |   |   |   |   |   |   Initial-P <= 8.48: high (18.0/5.0)
|   |   |   |   |   |   |   |   |   Initial-P > 8.48: medium (2.0)
|   |   |   |   |   |   Treatment > 14: high (40.0/7.0)
|   |   |   |   |   Treatment > 16
|   |   |   |   |   |   Initial-P <= 6.42: medium (10.0/1.0)
|   |   |   |   |   |   Initial-P > 6.42
|   |   |   |   |   |   |   Initial-P <= 8.48: high (8.0/2.0)
|   |   |   |   |   |   |   Initial-P > 8.48: medium (2.0)
|   |   |   |   Treatment > 17: high (20.0/2.0)
|   |   Wereda  = Yilmana-Densa
```

```
|  |  |  Treatment <= 7: medium (36.0/9.0)
|  |  |  Treatment > 7: high (198.0/47.0)
|  |  Wereda  = Lume
|  |  |  Initial-P <= 5.2: high (26.0/1.0)
|  |  |  Initial-P > 5.2
|  |  |  |  Soil-pH = <5: high (0.0)
|  |  |  |  Soil-pH = [5-8.5]
|  |  |  |  |  Initial-P <= 8: high (52.0/8.0)
|  |  |  |  |  Initial-P > 8
|  |  |  |  |  |  Initial-P <= 10
|  |  |  |  |  |  |  Treatment <= 15: medium (20.0/2.0)
|  |  |  |  |  |  |  Treatment > 15
|  |  |  |  |  |  |  |  Treatment <= 16: high (2.0)
|  |  |  |  |  |  |  |  Treatment > 16: medium (4.0/1.0)
|  |  |  |  |  |  Initial-P > 10
|  |  |  |  |  |  |  Initial-P <= 10.6: high (26.0/1.0)
|  |  |  |  |  |  |  Initial-P > 10.6
|  |  |  |  |  |  |  |  Treatment <= 10: medium (20.0/5.0)
|  |  |  |  |  |  |  |  Treatment > 10: high (32.0/5.0)
|  |  |  |  Soil-pH = >8.5: medium (52.0/17.0)
|  |  Wereda  = Hetosa
|  |  |  Initial-P <= 0.2
|  |  |  |  Total-nitrogen = <0.14
|  |  |  |  |  Treatment <= 7: low (4.0/1.0)
|  |  |  |  |  Treatment > 7
|  |  |  |  |  |  Treatment <= 15: medium (16.0/1.0)
|  |  |  |  |  |  Treatment > 15: high (2.0)
|  |  |  |  Total-nitrogen = [0.14-0.24]
|  |  |  |  |  Organic-matter = <2.04: high (0.0)
|  |  |  |  |  Organic-matter = [2.04-3.92]
|  |  |  |  |  |  Soil-pH = <5: high (0.0)
|  |  |  |  |  |  Soil-pH = [5-8.5]
|  |  |  |  |  |  |  Initial-P <= 0.18: medium (22.0/2.0)
|  |  |  |  |  |  |  Initial-P > 0.18: high (22.0/2.0)
|  |  |  |  |  |  Soil-pH = >8.5: high (88.0/9.0)
|  |  |  |  |  Organic-matter = >3.92
|  |  |  |  |  |  Initial-P <= 0.12: high (22.0/3.0)
|  |  |  |  |  |  Initial-P > 0.12
|  |  |  |  |  |  |  Treatment <= 9: medium (24.0/2.0)
|  |  |  |  |  |  |  Treatment > 9: high (42.0/17.0)
|  |  |  |  Total-nitrogen = >0.24: high (0.0)
|  |  |  Initial-P > 0.2: medium (33.0/5.0)
|  |  Wereda  = Dejen
|  |  |  Treatment <= 11
|  |  |  |  Treatment <= 9
|  |  |  |  |  Initial-P <= 11.181
|  |  |  |  |  |  Treatment <= 6: low (14.0/3.0)
|  |  |  |  |  |  Treatment > 6
|  |  |  |  |  |  |  Initial-P <= 6.8
|  |  |  |  |  |  |  |  Initial-P <= 6.66: medium (6.0/2.0)
|  |  |  |  |  |  |  |  Initial-P > 6.66: low (6.0/1.0)
|  |  |  |  |  |  |  Initial-P > 6.8: medium (30.0/12.0)
|  |  |  |  |  Initial-P > 11.181: medium (24.0/2.0)
|  |  |  |  Treatment > 9
|  |  |  |  |  Soil-pH = <5: medium (16.0/4.0)
|  |  |  |  |  Soil-pH = [5-8.5]
```

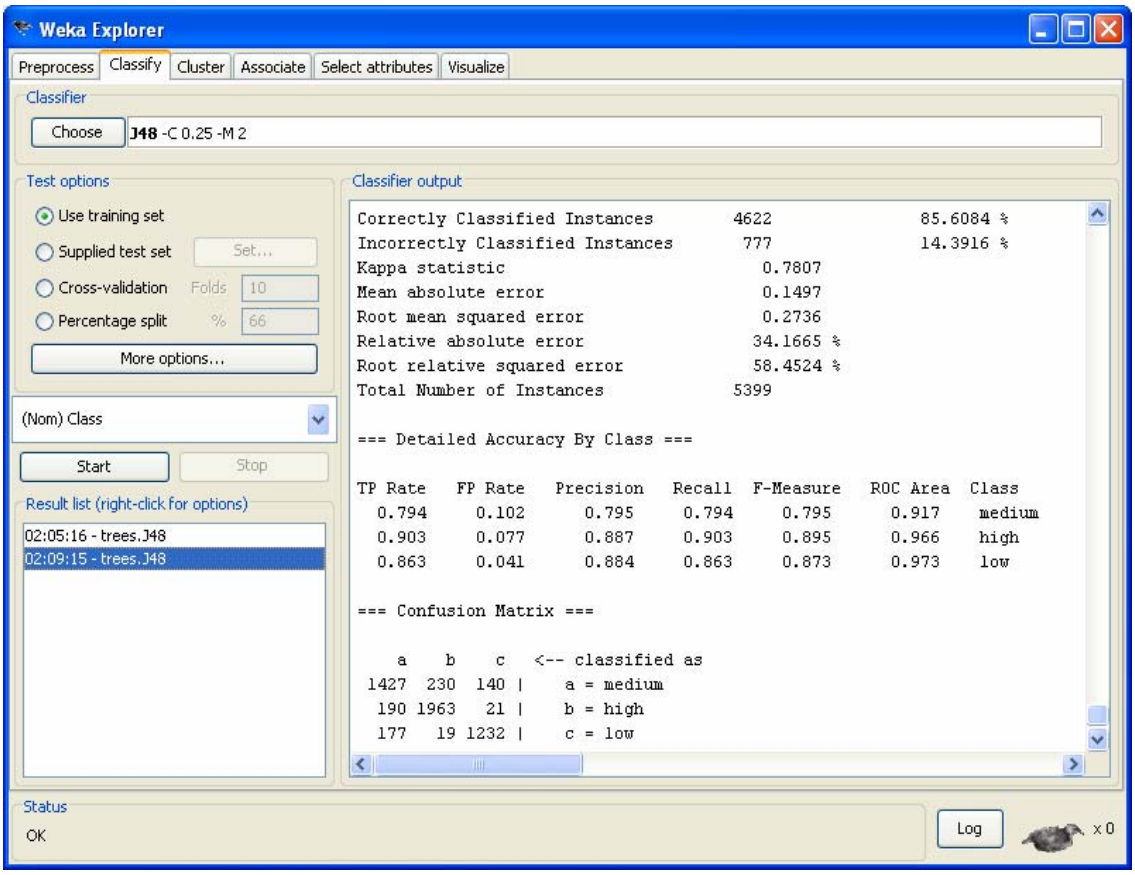
```
|  |  |  |  |  |  Organic-matter = <2.04
|  |  |  |  |  |  |  Initial-P <= 14.28: medium (12.0/3.0)
|  |  |  |  |  |  |  Initial-P > 14.28: high (4.0/1.0)
|  |  |  |  |  |  Organic-matter = [2.04-3.92]: high (8.0/3.0)
|  |  |  |  |  |  Organic-matter = >3.92: medium (0.0)
|  |  |  |  |  Soil-pH = >8.5: medium (0.0)
|  |  |  Treatment > 11
|  |  |  |  Treatment <= 13
|  |  |  |  |  Soil-pH = <5
|  |  |  |  |  |  Treatment <= 12
|  |  |  |  |  |  |  Total-nitrogen = <0.14: medium (6.0/2.0)
|  |  |  |  |  |  |  Total-nitrogen = [0.14-0.24]: high (2.0)
|  |  |  |  |  |  |  Total-nitrogen = >0.24: medium (0.0)
|  |  |  |  |  |  Treatment > 12
|  |  |  |  |  |  |  Initial-P <= 11.084
|  |  |  |  |  |  |  |  Initial-P <= 6.66: medium (2.0)
|  |  |  |  |  |  |  |  Initial-P > 6.66: low (4.0/1.0)
|  |  |  |  |  |  |  Initial-P > 11.084: medium (2.0/1.0)
|  |  |  |  |  Soil-pH = [5-8.5]
|  |  |  |  |  |  Treatment <= 12
|  |  |  |  |  |  |  Initial-P <= 14.28: high (10.0/2.0)
|  |  |  |  |  |  |  Initial-P > 14.28: medium (2.0)
|  |  |  |  |  |  Treatment > 12
|  |  |  |  |  |  |  Initial-P <= 9.06: medium (4.0/1.0)
|  |  |  |  |  |  |  Initial-P > 9.06: high (8.0)
|  |  |  |  |  Soil-pH = >8.5: high (0.0)
|  |  |  |  Treatment > 13: high (164.0/19.0)
|  |  Wereda  = Achefer
|  |  |  Organic-matter = <2.04
|  |  |  |  Treatment <= 9: medium (16.0/6.0)
|  |  |  |  Treatment > 9: high (36.0/6.0)
|  |  |  Organic-matter = [2.04-3.92]
|  |  |  |  Initial-P <= 4.62: low (52.0/5.0)
|  |  |  |  Initial-P > 4.62
|  |  |  |  |  Initial-P <= 7.5
|  |  |  |  |  |  Initial-P <= 6.7
|  |  |  |  |  |  |  Treatment <= 8: medium (18.0/6.0)
|  |  |  |  |  |  |  Treatment > 8
|  |  |  |  |  |  |  |  Treatment <= 9: low (6.0)
|  |  |  |  |  |  |  |  Treatment > 9
|  |  |  |  |  |  |  |  |  Treatment <= 12: medium (18.0/3.0)
|  |  |  |  |  |  |  |  |  Treatment > 12
|  |  |  |  |  |  |  |  |  |  Treatment <= 13: low (6.0/1.0)
|  |  |  |  |  |  |  |  |  |  Treatment > 13: medium (30.0/9.0)
|  |  |  |  |  |  Initial-P > 6.7: high (26.0/7.0)
|  |  |  |  |  Initial-P > 7.5
|  |  |  |  |  |  Initial-P <= 9.32
|  |  |  |  |  |  |  Treatment <= 16: low (22.0/2.0)
|  |  |  |  |  |  |  Treatment > 16: medium (4.0/1.0)
|  |  |  |  |  |  Initial-P > 9.32
|  |  |  |  |  |  |  Treatment <= 16: medium (22.0/6.0)
|  |  |  |  |  |  |  Treatment > 16: high (4.0)
|  |  |  Organic-matter = >3.92: medium (0.0)
|  |  Wereda  = Tahtaykoraro
|  |  |  Initial-P <= 12
|  |  |  |  Treatment <= 13
```

```
|  |  |  |  |  Initial-P <= 4.28
|  |  |  |  |  |  Soil-pH = <5: low (0.0)
|  |  |  |  |  |  Soil-pH = [5-8.5]
|  |  |  |  |  |  |  Treatment <= 12
|  |  |  |  |  |  |  |  Treatment <= 10
|  |  |  |  |  |  |  |  |  Initial-P <= 0.38: low (10.0)
|  |  |  |  |  |  |  |  |  Initial-P > 0.38
|  |  |  |  |  |  |  |  |  |  Treatment <= 8: medium (12.0/4.0)
|  |  |  |  |  |  |  |  |  |  Treatment > 8: low (8.0/2.0)
|  |  |  |  |  |  |  |  Treatment > 10: medium (12.0/1.0)
|  |  |  |  |  |  |  Treatment > 12: low (6.0)
|  |  |  |  |  |  Soil-pH = >8.5: low (16.0/1.0)
|  |  |  |  |  Initial-P > 4.28
|  |  |  |  |  |  Initial-P <= 4.74: medium (16.0/1.0)
|  |  |  |  |  |  Initial-P > 4.74
|  |  |  |  |  |  |  Soil-pH = <5: medium (0.0)
|  |  |  |  |  |  |  Soil-pH = [5-8.5]: medium (32.0/6.0)
|  |  |  |  |  |  |  Soil-pH = >8.5
|  |  |  |  |  |  |  |  Treatment <= 6: medium (2.0)
|  |  |  |  |  |  |  |  Treatment > 6: low (14.0/6.0)
|  |  |  |  Treatment > 13
|  |  |  |  |  Soil-pH = <5: medium (0.0)
|  |  |  |  |  Soil-pH = [5-8.5]: medium (51.0/17.0)
|  |  |  |  |  Soil-pH = >8.5
|  |  |  |  |  |  Initial-P <= 4.28
|  |  |  |  |  |  |  Treatment <= 14: low (2.0)
|  |  |  |  |  |  |  Treatment > 14
|  |  |  |  |  |  |  |  Treatment <= 16: medium (3.0)
|  |  |  |  |  |  |  |  Treatment > 16: low (4.0/1.0)
|  |  |  |  |  |  Initial-P > 4.28
|  |  |  |  |  |  |  Initial-P <= 4.74
|  |  |  |  |  |  |  |  Treatment <= 17: medium (8.0/1.0)
|  |  |  |  |  |  |  |  Treatment > 17: high (2.0)
|  |  |  |  |  |  |  Initial-P > 4.74: low (10.0/4.0)
|  |  |  Initial-P > 12: low (26.0/2.0)
|  |  Wereda  = Alefa
|  |  |  Organic-matter = <2.04
|  |  |  |  Initial-P <= 2.46
|  |  |  |  |  Treatment <= 14
|  |  |  |  |  |  Treatment <= 12
|  |  |  |  |  |  |  Treatment <= 10
|  |  |  |  |  |  |  |  Treatment <= 8
|  |  |  |  |  |  |  |  |  Soil-pH = <5
|  |  |  |  |  |  |  |  |  |  Treatment <= 6: medium (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  Treatment > 6: high (8.0/1.0)
|  |  |  |  |  |  |  |  |  Soil-pH = [5-8.5]: medium (12.0/2.0)
|  |  |  |  |  |  |  |  |  Soil-pH = >8.5: medium (0.0)
|  |  |  |  |  |  |  |  Treatment > 8: medium (16.0/6.0)
|  |  |  |  |  |  |  Treatment > 10: high (16.0/4.0)
|  |  |  |  |  |  Treatment > 12
|  |  |  |  |  |  |  Treatment <= 13: low (8.0/3.0)
|  |  |  |  |  |  |  Treatment > 13
|  |  |  |  |  |  |  |  Total-nitrogen = <0.14
|  |  |  |  |  |  |  |  |  Soil-pH = <5: medium (2.0)
|  |  |  |  |  |  |  |  |  Soil-pH = [5-8.5]: low (2.0)
|  |  |  |  |  |  |  |  |  Soil-pH = >8.5: medium (0.0)
```

```
|  |  |  |  |  |  |  |  Total-nitrogen = [0.14-0.24]: high (4.0)
|  |  |  |  |  |  |  |  Total-nitrogen = >0.24: high (0.0)
|  |  |  |  |  Treatment > 14
|  |  |  |  |  |  Soil-pH = <5
|  |  |  |  |  |  |  Treatment <= 17: high (12.0/4.0)
|  |  |  |  |  |  |  Treatment > 17: medium (4.0)
|  |  |  |  |  |  Soil-pH = [5-8.5]
|  |  |  |  |  |  |  Treatment <= 17
|  |  |  |  |  |  |  |  Total-nitrogen = <0.14: medium (6.0/2.0)
|  |  |  |  |  |  |  |  Total-nitrogen = [0.14-0.24]
|  |  |  |  |  |  |  |  |  Treatment <= 16: high (4.0)
|  |  |  |  |  |  |  |  |  Treatment > 16: medium (2.0)
|  |  |  |  |  |  |  |  Total-nitrogen = >0.24: medium (0.0)
|  |  |  |  |  |  |  Treatment > 17: high (4.0)
|  |  |  |  |  |  Soil-pH = >8.5: high (0.0)
|  |  |  |  Initial-P > 2.46: medium (26.0/7.0)
|  |  |  Organic-matter = [2.04-3.92]: medium (26.0/10.0)
|  |  |  Organic-matter = >3.92
|  |  |  |  Total-nitrogen = <0.14: medium (26.0/9.0)
|  |  |  |  Total-nitrogen = [0.14-0.24]: low (52.0/7.0)
|  |  |  |  Total-nitrogen = >0.24
|  |  |  |  |  Treatment <= 16: low (22.0/8.0)
|  |  |  |  |  Treatment > 16: medium (4.0)
|  |  Wereda  = Kuyu
|  |  |  Organic-matter = <2.04
|  |  |  |  Treatment <= 14
|  |  |  |  |  Treatment <= 12: high (14.0/6.0)
|  |  |  |  |  Treatment > 12: medium (4.0)
|  |  |  |  Treatment > 14: high (8.0)
|  |  |  Organic-matter = [2.04-3.92]
|  |  |  |  Total-nitrogen = <0.14
|  |  |  |  |  Initial-P <= 10: high (26.0/2.0)
|  |  |  |  |  Initial-P > 10: low (26.0/4.0)
|  |  |  |  Total-nitrogen = [0.14-0.24]
|  |  |  |  |  Treatment <= 15: low (20.0)
|  |  |  |  |  Treatment > 15
|  |  |  |  |  |  Treatment <= 16: medium (2.0)
|  |  |  |  |  |  Treatment > 16
|  |  |  |  |  |  |  Treatment <= 17: low (2.0)
|  |  |  |  |  |  |  Treatment > 17: medium (2.0)
|  |  |  |  Total-nitrogen = >0.24: low (0.0)
|  |  |  Organic-matter = >3.92
|  |  |  |  Total-nitrogen = <0.14: high (0.0)
|  |  |  |  Total-nitrogen = [0.14-0.24]
|  |  |  |  |  Initial-P <= 12: high (26.0)
|  |  |  |  |  Initial-P > 12
|  |  |  |  |  |  Treatment <= 7: medium (4.0)
|  |  |  |  |  |  Treatment > 7: high (22.0/6.0)
|  |  |  |  Total-nitrogen = >0.24
|  |  |  |  |  Initial-P <= 12
|  |  |  |  |  |  Initial-P <= 6.94
|  |  |  |  |  |  |  Treatment <= 11: medium (12.0/3.0)
|  |  |  |  |  |  |  Treatment > 11: high (14.0/6.0)
|  |  |  |  |  Initial-P > 6.94
|  |  |  |  |  |  Treatment <= 7: medium (4.0)
|  |  |  |  |  |  Treatment > 7: high (22.0/3.0)
```

```
|  |  |  |  |  Initial-P > 12: medium (26.0/9.0)
|  |  Wereda  = Wara-Jarso
|  |  |  Initial-P <= 4.74
|  |  |  |  Treatment <= 7: medium (4.0)
|  |  |  |  Treatment > 7: high (22.0/5.0)
|  |  |  Initial-P > 4.74: high (244.0/2.0)
|  Initial-P > 15.702
|  |  Crop-name = wheat
|  |  |  Initial-P <= 18: high (403.0/9.0)
|  |  |  Initial-P > 18
|  |  |  |  Organic-matter = <2.04
|  |  |  |  |  Treatment <= 9: medium (16.0/2.0)
|  |  |  |  |  Treatment > 9: high (196.0/13.0)
|  |  |  |  Organic-matter = [2.04-3.92]: high (0.0)
|  |  |  |  Organic-matter = >3.92
|  |  |  |  |  Treatment <= 6: low (2.0)
|  |  |  |  |  Treatment > 6: medium (24.0/5.0)
|  |  Crop-name = f: medium (26.0/9.0)

Number of Leaves  :  256

Size of the tree :        442

Time taken to build model: 19.23 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        4622               85.6084 %
Incorrectly Classified Instances       777               14.3916 %
Kappa statistic                          0.7807
Mean absolute error                      0.1497
Root mean squared error                  0.2736
Relative absolute error                 34.1665 %
Root relative squared error             58.4524 %
Total Number of Instances             5399

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.794    0.102      0.795     0.794     0.795      0.917    medium
                 0.903    0.077      0.887     0.903     0.895      0.966    high
                 0.863    0.041      0.884     0.863     0.873      0.973    low


=== Confusion Matrix ===

    a    b    c   <-- classified as
 1427  230  140 |    a = medium
  190 1963   21 |    b = high
  177   19 1232 |    c = low
```

Here is the screen shot for the result of experiment three



Declaration

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

_____

Zebiba Ali Abegaz

January, 2009

The thesis has been submitted for examination with my approval as university advisor

_____

Dr. Manoj V.N.V

January, 2009