DEBRE BIRHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPARTMENT OF INFORMATION SYSTEMS

PREDICTIVE MODEL FOR ACADEMIC PERFORMANCE OF STUDENTS IN ETHIOPIAN HIGHER EDUCATION ENTRANCE EXAMINATION (EHEEE)

By:

TEKLEMARIAM DEMISSIE ALEMU

JUNE05/ 2019

DEBRE BIRHAN, ETHIOPIA

DEBRE BIRHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPARTMENT OF INFORMATION SYSTEMS

PREDICTIVE MODEL FOR ACADEMIC PERFORMANCE OF STUDENTS IN ETHIOPIAN HIGHER EDUCATION ENTRANCE EXAMINATION (EHEEE)

A THESIS SUBMITTED TO THE COLLEGE OF COMPUTING OF DEBRE BIRHAN UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SYSTEMS.

By:

TEKLEMARIAM DEMISSIE ALEMU

JUNE05/2019

DEBRE BIRHAN, ETHIOPIA

DEBRE BIRHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPARTMENT OF INFORMATION SYSTEMS

PREDICTIVE MODEL FOR ACADEMIC PERFORMANCE OF STUDENTS IN ETHIOPIAN HIGHER EDUCATION ENTRANCE EXAMINATION (EHEEE)

By

TEKLEMARIAM DEMISSIE ALEMU

Name and signature of Members of the Examining Board

Name	Chair person	Signature	Date
_____	Chairperson	_____	_____
Kindie Biredagn(PhD)	Advisor	_____	_____
_____	External Examiner	_____	_____
_____	Internal Examiner	_____	_____

DECLARATION

I hereby declare that Application of Data Mining Techniques to develop a classification model which Predicts the performance of Preparatory School Students in EHEEE Result is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

TEKLEMARIAM DEMISSIE ALEMU

JUNE 05/2019

The thesis has been submitted for examination with my approval as University Advisor

---

Dr. KINDIE BIREDAGN

JUNE 05/ 2019

DEDICATION

It is a lifetime opportunity for me to dedicate this thesis work to Dear my passed away father Ato Demissie Alemu, and for all my families who are alive .Dear my families, I know all the ups and downs you passed through for me, your pray and helps in many ways, made me strong more and more. Long live for you!

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God .Because, without his assistance, neither completing my thesis work nor my effort would have been successful.

Next, my best gratitude goes to my advisor, Kindie Biredagn (PhD) for his commitment, encouragement, valuable comments, stimulating advice and constructive suggestions that made me pass the difficulties I faced to finish my thesis work. What I have achieved would not have been possible without his genuine and professional guidance. You are really my best ever!

Next my deepest gratitude goes to Mr. Kindie Alebachew who is the department head of Information System, for his commitment, encouragement, valuable comments, stimulating advice, and constructive suggestions that helped me to pass the difficulties I faced during the thesis work, I want to thank National Educational Assessment and Examination Agency principal, Ato Yoseph Abera, who gave me the necessary dataset of student in Ethiopian Higher Education Entrance examination (EHEEE).

I am also indebted to staff members of Debre Birhan City preparatory schools for their cooperation during the knowledge acquisition process, testing and evaluation of the prototype system. I would like to express my gratitude to my beloved wife Meskerem Seifu, my Families and all my friends for all their bests and patience, love and support to accomplishing my research work.

Lastly, my heartfelt thanks goes to Haile Mariam Mamo Preparatory School vice director Ato Siyamregn shewayerga and Principal Director Ato Dibabu Aderie for their full help to me during my study.

Teklemariam Demissie
JUNE 05/2019

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS &ABBREVIATION

ABE	Alternative Basic Education
ARFF	Attribute Relation File Format
BICAPG10	Biology Class Average Point in Grade 10
CACs	City Administration Councils
CAP	Class Average Points
CGPA	Class Grade Point Average
CHCAPG10	Chemistry Class Average Point in Grade 10
CICAPG10	Civics Class Average Point in Grade 10
COC	Certified of Certificate
CRISP- DM	Cross –Industry Standard Process for Data Mining
CSA	Central Statistical Agency of Ethiopia
CSV	Comma Separated Value
DBC	Debre Birhan City
DBCPS	Debre Birhan City Preparatory Schools
DM	Data Mining
ECAPG10	English Class Average Point in Grade 10
EDM	Educational Data Mining
EGSECE	Ethiopian General Secondary Education Certificate Examination
EHEEE	Ethiopian Higher Education Entrance Examination
GIGO	Garbage in Garbage Out
GPA	Grade Point Average
GUI	Graphical User Interface
KDD	knowledge Discovery in Databases
KDP	Knowledge Discovery Process
Kg	Kindergarten
MCAPG10	Mathematics Class Average Point in Grade 10
MOE	Ministry of Education
NEAEA	National Educational Assessment and Examination Agency

NRSs	National Regional States
PSLCE	Primary School Leaving Certificate Examination
SAS	Statistical Analysis System
SEMMA	Sample Explore Modify Model Assess
SIMS	Students Information Management System
SMOTE	Synthetic Minority Over Sampling Technique
TVET	Technical and Vocational Education Training
UEE	University Entrance Examination
UTM	Universal Transverse Mercator
WEKA	Waikato Environment for Knowledge Analysis

ABSTRACT

In recent years, the biggest challenges that educational institutions are facing the explosive growth of educational data and to use this data to improve the quality of managerial decisions. The number of students scoring low performance in Ethiopia Higher Education Entrance Examination (EHEEE) result is increasing from year to year as the reports of NEAEA and Ministry of Education. At the same time, many students who have low performance have joined higher institutions in order to full fill the intake capacity of universities. Moreover, many students have caused the families for more expenses. Even if there are a number of studies regarding these problems, the works were focused only on few selected attributes at preparatory and secondary schools. Those studies are also not understandable by all stakeholders and not easy to be guide by the result. Although there are studies regarding academic performance of students using data mining techniques, they are all about university students.

Thus, this study aims to apply data mining techniques to develop a classifier model which predicts the Performance of Students in Ethiopian Higher Education Entrance Examination (EHEEE) Examination in order to help new students early before they face the problem and enables managerial in making different decisions to improve the students' academic performance.

A total of 3013 student's records and 32 attributes were used to build the predictive model using J48, JRIP, REPTree and PART algorithms and the class label of the record are taken to be Excellent, Very Good, Good, Satisfactory and Fail based on the academic performance assessment system for Grade 12 National Exam. The researcher has developed classifier model by using Hybrid data mining model where PART algorithm which registered the highest accuracy of 95.37%. As the result, 157 rules have been found as the guide to perform better performance. In this case, the educational planers can identify the determinant attributes to give support at each grade level to full fill their gaps.

Finally, the system evaluation has found 96% by using the system performance and 91.2% by user acceptance tests. Further study is needed to get the better performance on the real time.

Keywords: Educational Data Mining, Classification, Students' Academic Performance

CHAPTER ONE

INTRODUCTION

1.1. Background of the Study

Education plays a great role in achieving one's country growth and development. Educational policy is mainly aimed at expanding the education for the citizen to improve quality and ensuring that educational content is coordinated with the country's economic needs [1].In order to achieve the planned goals, the quality of the education needs to be in the accurate way (Identifying the determinant factors in the field of education). The quality of the education depends on the stakeholders like instructors, students, families, society and government. This stakeholders are the main participants of quality of education. The quality of the education can be seen from the side of performances of students based on the educational institutions rules. Thus, early predicting the students' performance can help in making different and timely managerial decisions at each levels in order to improve the academic performance of students [2].Ethiopian education vision focuses on production of citizens that possess human and national responsibility, developing problem solving attitude and capacity, production of lower, middle and higher level skilled manpower that can participate in various fields of the economic sector and contribute to the country's economic growth. The educational system has been organized in consistent with the Federal Government's State Structure accordingly. In each structure of education federal state, National Regional States, zones and woredas have their own responsibilities. For example Administratively, Ethiopia is federalized (structured) into nine National Regional States/NRSs/ and two City Administration Councils /CACs/ which are under the umbrella of the Federal Government. There are 68 zones which are controlled by regions and 770 there are woredas which have their own duties and bureaus of education responsible for administrating, managing the educational system and network of management within each of them. The woreda is one of educational authority responsible for all educational institutions in its area under zone [3].

In Ethiopia, there has been a dramatic increase in admission of preparatory school students due to the increasing needs of new generation and the increasing educational institutions admission capacity in the country. In 2002 E.C. the number of students who attend preparatory school education were 243,080.This number has been increased to 425,774 in 2007 E.C. which shows that large number of increase of students since the target of admission of students to achieve in 2007 E.C. was 360,000. But many students have come with low performance (satisfactory and

poor performances) as well as fewest students were score the best performance (Excellent)[4].This leads to the increased records concerning students and invites to make analysis based on the data in order to predict the students' academic performance and helpful to identify the predetermined factors of students' performance. Thus, it is essential to find better and recent data analysis mechanisms which are data mining techniques rather than focusing on traditional ways of data analysis such as making decision without knowing . Data mining has concept which is applicable in different sectors; for instance in banking, in retail sales, and in telecommunications; starts to get an attention from educational sector and it is the process of analyzing data from different perspectives and summarizing the results as useful information. A solution to achieve this goal is to use the knowledge discovery in databases (KDD) data mining process that help to analyze the preprocessed dataset in education which is called educational data mining (EDM) focuses on applying data mining tools and techniques to educationally related data [5]. Educational Data Mining is one of application of data mining in educational environment [6] and it is a new growing research emerging discipline, concerned with data from academic field to develop various methods and to identify unique and significant factors which help to explore students' academic performance at different grade levels[6,7]. Secondary and Preparatory education have an aim of identifying students' ability and inclination in order to assign them into different fields of study in various universities. It is used to place grade 10 students in preparatory and preparatory students in universities based on their performance and choice [8].The performance of these students in EGSECE and in EHEEE result categorized into five (Excellent, Very Good, Good, Satisfactory and fail) by Ministry of Education (MOE) [9]. Public TVET institutions under the education sector were concentrating on producing middle level technical graduates at post Grade 10 level and offered five level courses, ranging from one to five years [10].All these and others educational sectors have standard educational assessments systems.

1.2. Statement of the Problem

Preparatory school is the end of the second cycle education system in which students have to take a National Examination namely Ethiopian Higher Education Entrance Examination (EHEEE) that has a purpose of selecting students for the next higher level education that is university level. But, as the National Educational Assessment and Examination Agency (NEAEA) report shows in 2007 E.C from 206,472 Grade 12 students who sat for (EHEEE)

103,492 (51.12%) students fail to join higher institutions by scoring below the cutting point .At that time, the cutting point for male and female was 343 and 320 respectively[4]. Similarly, as the report of Educational Assessment and Examination Agency (NEAEA), Debre Birhan City preparatory schools, the percentages of students who have took EHEEE and scored "low "or "satisfactory" performances (177-352 scores) are 41.2%, 41%, 42% ,36 % ,40.3% and the cutting points for male and females for the corresponding years were (294,290), (315,300), (343,320), (354,340) and (352,330) respectively [11]. Thus, as we understand from data of five years mentioned above, the percentages of the students who have took the National Examination of Grade 12 and didn't scored the passing point are many in number whose performances are 'satisfactory' [12]. On another hand, fewest (2.95%) of students have scored an 'excellent' performance, 19.11% students' have scored 'very good', and 38.2% students have scored 'Good' performances in the five years.

This problem initiates the researcher to identify the reasons why many Debre Birhan City preparatory schools students scoring low performance (satisfactory result), Good performance and few numbers of students have scored an 'excellent' and 'very good' performances. Still a days, there are many students have who scored an expected passing point (350) and above to join higher education. But, universities have took many students with low performance (satisfactory performance) less than the expected half of the total score. However, this is one of the factor which influences the educational quality [13].

1.3. Research Questions

The research attempts to answer the following questions:

1. Which attributes are the most important to identify the performance of the preparatory school students in EHEEE result?
2. Which data mining algorithm can be more appropriate to predict students' performance in Ethiopian Higher Education Entrance Examination?
3. What are the most interesting patterns or rules generated using the determinant attributes?

1.4. Objective of the Study

1.4.1. General Objectives

The general objective of this research is to developing classifier model that predicts the performance of Natural Science Preparatory students to join higher institution.

1.4.2 .Specific Objectives

To achieve the general objective, the following specific objectives are formulated to:

- Review conceptual and related literatures relevant to the study.
- Prepare the dataset in order to make it suitable for the data mining.
- Identify relevant attributes to determine academic performance of the students.
- Determine the appropriate data mining algorithm to build the predictive model.
- Evaluating the performance of different classifier models.
- Develop a classifier model for predicting performance of students in their result.
- Select the classifier model that perform higher performance.
- Develop prototype that enables the user to interact with the system easily.

1.5. Scope and Limitation of the Study

The scope of the study was limited to utilizing Debre Birhan City preparatory schools students' Ethiopian Higher Education Entrance Examination (EHEEE) result from National Educational Assessment and Examination Agency (NEAEA) and the corresponding students result from preparatory students transcript which contain Grade 9, Grade 10 Class Average points, Grade 10 Ethiopian General Secondary Education Certificate Examination (EGSECE), Grade11 and Grade12 Grade Average Points) for natural science stream students between 2005 E.C- 2009 E.C to early prediction of academic performance. They didn't include more influencing data like families of students and their economic status, health status, and Educational status, students back ground in the students' dataset at the Debre Birhan City preparatory schools.

1.6. Significance of the Study

The outcome of this research will have a great contribution for different stakeholders in different ways such as: For educational policy makers, to improve student academic performance, to offer a helpful and constructive recommendations for academic planners and to overcome the problem of low achievers of natural science stream preparatory students in grade 12 national Exam result and to ensure quality of education. This will also aid in the curriculum structure and modification of curriculum in order to improve students' academic performance. Besides, it will have an advantage to show the factors which have an impact on the performance of preparatory students result and help in giving timely managerial decisions. Furthermore, it will guide to improve the failure of preparatory students since it points out factors and will help students with problems so as to be successful students in higher education.

Using the rules, new entrant preparatory students able to identify predominant subjects which need more focus and able to decide their bests from the rules and previous experience from the

research findings as a bench. The study will enables the students to prevent him/her self and educational institutions from serious financial strains. This study also importance for the students' families and higher institutions to get students whose academic performances are early predicted and therefore making them ready accordingly. Moreover, the findings of this study could provide information for those who are interested to make further study in this similar area.

1.7. Research Methodology

A research methodology is an arrangement of condition to collect and analysis of data in a manner that aim to address the research problem. The research is carried out based on primary data extracted from the database of Ethiopian national assessment agency which is available for researchers and from Debre Birhan City preparatory schools. Secondary data for instance review important document to gain farther information related with student achievement. Accordingly, in this study both quantitative and qualitative research design; interview was used to understand the domain knowledge and to interpret the finding. The researcher has used hybrid data mining process model.

1.7.1. Design of the Study

To realize a model that yields optimum classifier of students' academic performance of an individual, a Hybrid data mining process model was applied that consumes Knowledge Discovery data base Process (KDD) and Cross-industry Standard Process (CRISP-DM) models. Hybrid data mining process model has six steps [14] namely, Understanding the Problem Domain, Understanding the Data, Preparation of the Data, Data Mining, Evaluation of the Discovered Knowledge and Use of the Discovered Knowledge. It is selected for the present study since it provides a more general, research - oriented description of the steps, emphasize the iterative aspects of the process, drawing experience from previous models, support academia and industrial data mining projects. Besides it introduces the data mining steps instead of the modeling step.

1.7.2. Literature Review

The researcher will review different articles and journals that have information about performance of student, data mining, and integrated data from different sources to get conceptual understanding about the problem on hand.

1.7.3. Knowledge Representation

The knowledge that the researcher acquired from Data mining classification technique is presented in the form of IF-THEN rules. These IF-THEN rules are used to formulate the

conditional statements that constitute knowledge base. Rule based representation is highly expressive, easy to interpret and easy to generate. In addition, it classifies new instances rapidly [15]. Hence, rules are used to represent knowledge for the Data Ming.

1.7.4. Implementation Tools

In order to extract the hidden knowledge from the pre-processed dataset and compared the performance of classifiers WEKA (Waikato Environment for Knowledge Analysis) version 3.6.13 data mining tool is used and handling missing value by using Ms- Excel and MAT LAB version R2015a. To represent rules in Hybrid process model and construct the prototype of student performance the researcher use WEKA and use Java Net Beans IDE 8.2 with JDK-8u20 [16] to develop the GUI of the propose system.

1.7.5. Evaluation Methods

The researcher has used performance metrics like True Positive rate, Precision, Recall and F-measure to evaluate the performance of the developed model. The researcher also evaluate the hybrid process model using system performance testing by preparing test cases and users' acceptance testing questionnaire which helps the researcher to make sure that whether the potential users would like to use the proposed system frequently and whether the proposed systems meets user requirements.

1.8. Ethical Clearance

The confidentiality of the data have been maintained. Private data like students' names and known IDs has been eliminated. Besides, the research is intended only for academic purpose which is for Master's Thesis for the partial fulfillment of MSc. degree in Information systems.

1.9. Organization of the Thesis

This thesis comprises seven chapters. Chapter one discusses background of the study, the problem statement and research questions, the objectives of the study, scope and limitation of the study and methodologies that the researcher uses to conduct this study. Chapter two discusses about conceptual and related works review that are relevant for this study. In this chapter, Explanation about data mining process model, data mining tasks, classification algorithm is presented. The third chapter discusses about

research methodology of the study. Here, the researcher presents data mining framework for mining knowledge from dataset and the researcher presents the data mining process model domain understanding, data understanding, data preparation, data mining, data evaluation and

knowledge discovery. The fourth chapter represents the experimental result analysis and modeling, compare in order to choose the best model and uses of discovered knowledge. The fifth chapter about implementation and results discussion and answer for the research questions and evaluation of system performance and user acceptance. The sixth chapter is about summary, conclusion and recommendation of the analysis and the rules discovered and evaluation of the discovered knowledge.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORK

2.1. Introduction

In this chapter, different literatures concerning the concept of data mining; the methods, the applicability of data mining in education sector has been discussed. Now a days, large volume of data storage has emerged due to the fact that the increasing use of information technology in different sectors. Concerning the different sectors the format of the storage also differs. Records, files, documents, images, sounds, videos are some of the formats that data stored. The data stored in different formats need to be extracted in order to gain information and knowledge so as to help for different decision making processes. Knowledge Discovery in Databases (KDD) often known as data mining is a concept which is helpful in extracting the stored data so as to get useful information and knowledge [17]. Educational records are stored in different formats in different educational systems. Thus, for the purpose of quality of education, there are different Educational System Structure in different countries. Among these the Ethiopia Educational System Structure is stated as follows:

2.2. Overview of Educational Structure in Ethiopia

The current Ethiopian educational structure consists of Kindergarten, first cycle (primary school), second cycle primary school (junior), high school, preparatory school and higher institutions.

Kindergarten: The goal of kindergarten education is to help children develop their emotional, cognitive, physical and social domains, thus encouraging their ability and interest(enthusiasm) to continue to learn in both informal and formal environments and develop their social and educational skills. This education focuses on the all-round development of children encouraging their curiosity to learn and helping them to make sense of the world around them in preparation for a full life both in and out of school (life skills and educational).

Although children aged 4–6 are at a similar stage of cognitive development (pre-operational stage), slight cognitive differences appear among them. Therefore the pre-primary education curriculum has two stages: Stage 1: 4–5 years and Stage 2: 5–6 years. There are 14 Kindergarten schools in Debre Birhan City. Kindergarten educational program take three years before they join primary school. It run by Governmental Organizations, Communities, Private Institutions,

Religious Organizations, etc. The regional government is responsible for developing curriculum and facilitating the training of teachers (who has certificate qualification) for the Kindergarten education program and providing supervision and support. The access of Kindergarten program is limited to urban areas and growth rate of enrollment highly increased. The teaching and learning process of Kindergarten program use pictures of the letters, different animals, plants and human beings as instructional media and approaches such as free and facilitated play – such as sports, dance, music, visual arts and role-play, using mother tongue as a medium of instruction and for storytelling, using hands-on activities with a creative approach, Facilitating open-ended projects with a focus on the process rather than final product thus allowing the child to experiment and discover independently, communicate through all their senses including physical, sensory, sound and visual, Teaching an integrated curriculum where all areas of learning are learnt together.

First cycle (Primary Education): The Goals of Primary Education are to: provide basic education, which is appropriate to the physical and cognitive development of the learners; familiarize the learners with the production and service giving activities within their immediate environment; provide general education that prepares the learners for further education and training and for the world of work; by equipping them with basic knowledge, skills, abilities and attitudes. It includes grade 1-4 Primary Schools which has two cycles namely first cycle (grade1-4) and Alternative basic education(ABE) which takes three years after complete grade four to learn Technical and vocational Education Training (TVET) level one which has Goals. These goals are: to Provide Basic Education, to Physical and Cognitive Development of the learners; familiarize the learners with the production and service, giving activities within their immediate environment; provide general education that prepares the learners for further education training and for the world of work; equipping them with basic knowledge, skills, abilities and attitudes. The regional government is responsible for developing curriculum and each subjects are given by different teachers. The qualification of the teachers expected are certificate and who can use visual instructional media. There are five primary schools in Debre Birhan City. In those schools students take exam at the end of the year passed to the next grade 5 whereas failures are repeat grade 4 or pass to Alternative Basic Education (ABE). At the end of Grade 4 students take an exam that enables them to continue Grade 5 if they score 50 % of exam

and Students who failed the Exam attend Alternative Basic Education (ABE) for three years or Technical and vocational Education Training (TVET) level for two years.

Second cycle (Primary Education): The Goals of Primary Education has similar goals to the First cycle; but at the end of Grade 8 students take the Ethiopian Primary School Leaving Certificate Examination (PSLCE) which is a mandate of Regional states. Passing of this examination enables students to join general secondary education which take two years or Technical and vocational Education Training (TVET) Level three (3).

Secondary schools: There are about three Secondary schools which holds Grade 9 and Grade 10. It has goals: to provide general education that enable the students to identify their needs, interests and potential so that they can choose their field of study; to enable the student to continue further education and training; to prepare students for the world of work.

At the end of Grade 10 students take the Ethiopian General Secondary Education Certificate Examination (EGSECE) which is the mandate of the Ministry of Education. This examination decides whether students join the vocational training, teacher training college or the preparatory schools [1].Students who failed the exam, students who passed the exam and the interested will attend different Technical and Vocational Education Training (TVET) levels: Level1, Level2, Level 3, Level 4, and Level 5 while students who passed the exam will join the two year preparatory education in Grade 11 and 12. The admitted students for level 1–5 can be from grade 10, grade 11 and grade 12 .But the reforms that currently underway is being made to standardize the Grade 6 national examination and unstandardized Ethiopian General Secondary Education Certificate Examination.

Preparatory schools: they are the secondary cycle schools which are prepare students for higher education. It has goals: to choose subjects or areas of training; to prepare for higher education; to prepare students for the world of work. Preparatory secondary education has two streams namely natural science and social science stream. Students which are natural science stream take subjects such as English , mathematics, physics ,chemistry ,biology and civics ethical education; general subjects like physical education and information technology while Social Science stream students take subjects such as geography ,history and economics ;general subjects such as English ,mathematics for social, physical education and information technology. At the end of Grade 12 the students take an Exam called the Ethiopian Higher Education Entrance examination (EHEEE) or University Entrance Examination (UEE) which is developed by Institute of

Educational research of the Addis Ababa University and administered by the NEAEA. Preparatory secondary education with an aim of identifying students' ability and inclination in order to assign them into different fields of study in various universities. It is used to place students in universities based on their performance and choice [8].

In Ethiopian Higher Education Entrance Examination (EHEEE) students of Natural science Stream sit for examination of seven subjects namely English, Math's for natural, Physics, Chemistry, Biology, Civics Ethical Education and Aptitude Exam. Similarly social science stream students sit for examination of seven subjects namely English, Maths for Social Science, Geography, History, Economics, Civics and Ethical Education and Aptitude Examination [3].

This study focuses on predicting the performance of Natural Science stream students in Ethiopian Higher Education Entrance examination (EHEEE) result in case of Debre Birhan City. The performance of the students categorized into five (Excellent, Very Good, Good, Satisfactory and fail) by Ministry of Education (MOE) [9]. However, there are many students who have scored low performance (Satisfactory performance) than students who have scored better performances (Excellent, Very Good and Good performances) in Ethiopian Higher Education Entrance examination (EHEEE) [1]. So identifying the determinant factors for the students' better performances is very important for preparatory schools in order to help students for better achievement. Students who didn't score the year's cutting point cannot join university; but in case they can attend different trainings at different governmental and non-governmental training centers whereas students who score the year's cutting point or above the cutting point can join Governmental and Non-Governmental Colleges or Universities and incase different trainings based on interest. For example students who failed the exam cutting point can attend vocational training (TVET) levels while the one who passed the cutting point join under graduate degree program for 1,2,3,4 and 5 and above years.

Technical and Vocational Education Training (TVET): In Ethiopia Technical and Vocational Education and Training (TVET) has been split and was delivered by different providers at various qualification levels. Public TVET institutions under the education sector were concentrating on producing middle level technical graduates at post Grade 10 level and offered five level courses, ranging from one to five years [10]. The broad objectives of the strategy of Ethiopia in TVET are: to deliver quality TVET, assure employability of trainees, improve coherence and management of training provision, promote life-long learning and improve status

and attractiveness of TVET. According to the Education and Training Policy (ETP), the formal TVET system of the Ethiopia needs completion of a tenth-grade education to get certificate, diploma and advanced diploma upon completion of the levels 10+1(certificate), 10+2 (diploma) and 10+3(advanced diploma) of the TVET program . Now the ministry of education changed the curriculum in to level system from level 1 - 5[10]. Students at each level can pass from one level to the next level if and only if they pass the Certified Of Certificate (COC).The five levels are based on Ethiopian General Secondary Education Certificate Examination (EGSECE).But the score of the students are different to attend each levels.

Training College: There is one governmental teacher training college which has different departments (fields of education), four health colleges (one is governmental and the others are of private) and there is one governmental university which is Debre Birhan University in the Debre Birhan woreda holds different fields of study. These are: School of Engineering, College of Agriculture and Natural Resource, College of Business & Economics, College of Natural and computational Sciences, College of social science& Humanities, School of Computing, School of Health, two New open field of study in 2005E.C,1College of social science& Humanities ,Sociology.

Totally there are 31 fields of studies and 2nd degree in History& Heritage Management, Mathematical Modeling, information system and Master of Public Health in collaboration with Victory College. For the above educational system in Ethiopia there is detail figural description in appendix I.

2.3. Overview of Debre Birhan City

Debre Birhan city a zonal city located in the North Shewa Zone in Amhara Region. It is has area covers 181 km² which was established in 1446 E.C year by Atse Zereaykob Government and it is currently 120 kilometers to North East of Addis Ababa on the paved highway to Dessie city. It has a latitude and longitude of 9°41′N 39°32′E Coordinates or lies between Universal Transverse Mercator (UTM) Zone 37, 552083, 563954 Easting and 1064175, 1075864 Northing and has an average elevation of 2750 meters above sea level. Today, Debre Birhan administrative City is divided in to nine intra-urban and five peri-urban kebelles .According to the traditional temperature zone classification in Ethiopia, Debre Birhan City is found in the `100% 'Dega' and weather condition' temperature zone annual mean temperature is between 10-15 Degree Celsius [10].

Based on the 2007 E.C. national survey conducted by the Central Statistical Agency of Ethiopia (CSA), Debre Birhan city total population was 65,231, of whom 31,668 are men and 33,563 women and currently in 2010 E.C there are 409,208 men ,59,617 women ,totally has 468,825 population number. The city has wool factory, which is the first factory in Ethiopia established in 1965 E.C. had been the only manufacturing industry serving the inhabitants of the city for several decades. However, since recent time many manufacturing industries like beer factories, industry zone (textile factories),metal engineering,Agroprocessing, powder factory, leather factory, ceramic factory, wood factory, shoe factory, glass factories, oil factories,etc. Generally, the income of the people are: Agricultural, monthly salary and marketing activities. The Debre Birhan City is Zonal City of Debre Birhan Woreda found at the center of Basso Woreda and shares boundary with four Kebeles of Basso Woreda. These are: In the North Sariyia Kebele of Basso Woreda, in the east Wushawushegn Kebele of Basso Woreda, in the south Kormargefiya Kebele of Basso Woreda, in the west Angolela Kebele of Basso Woreda .The major Topography (land form) components of the study area include plateau (86%), valley (10%), and some undulating hills and mountains (4%).The Map of Debre Birhan City is shown in figure 2.1.

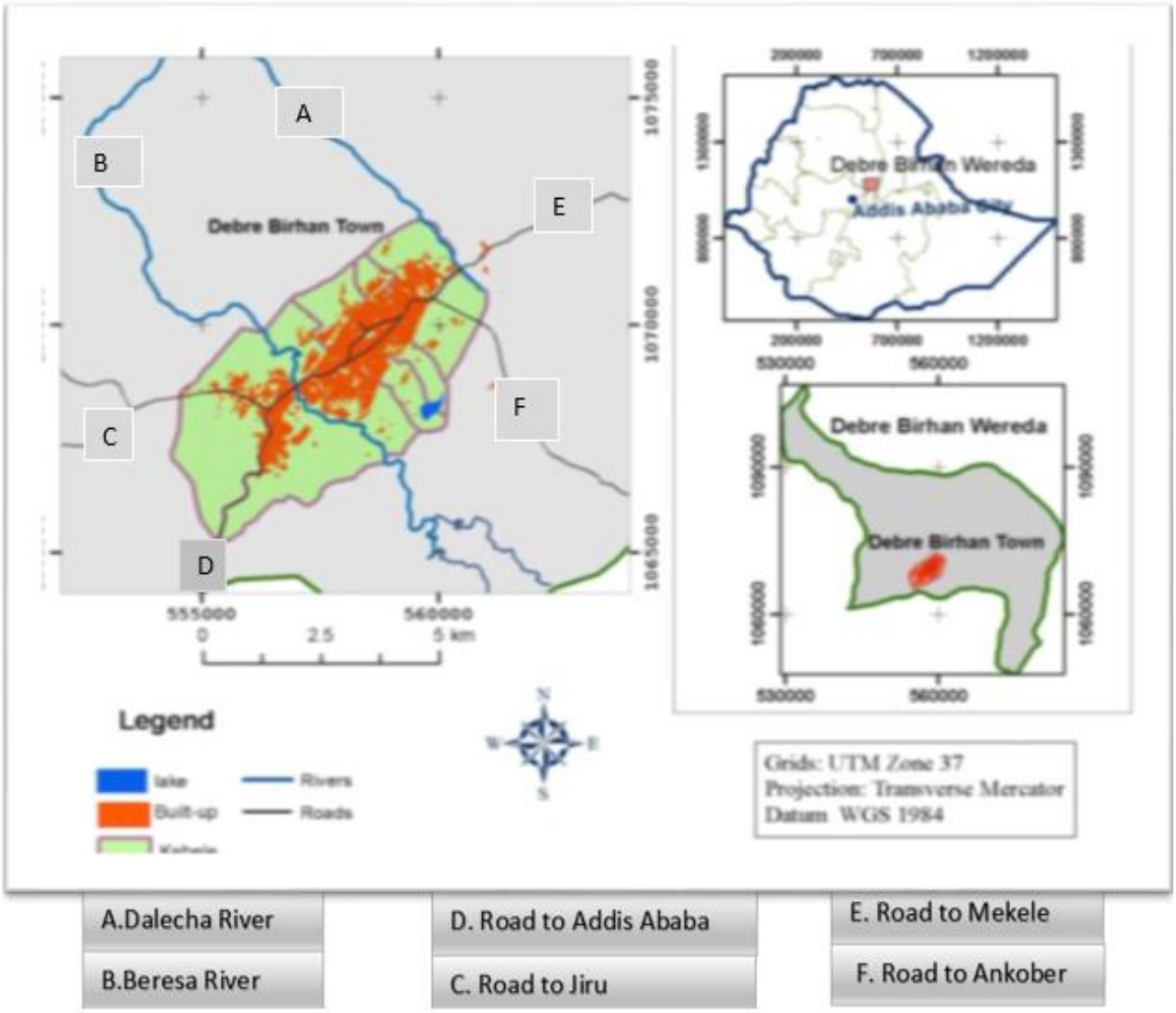


Figure 2.1.Location Map of the study area (adapted fromZewdu, 2011*)*

2.4. School Distribution in Debre Birhan City

As the observation of the researcher the School distribution in Debre Birhan City includes Governmental and Non-Governmental schools in each level except university level which is only Governmental. The Table 2.1 below shows School distribution is illustration in Debre Birhan City.

Table 2.1. School Distribution in Debre Birhan City

No	School Category	Ownership		Total	Supper visor
		Non-Government (Quantity)	Government (Quantity)		
1	Basic education (Kindergarten)	14	-	14	Woreda Education Office
2	Primary school First cycle	-	14	14	Woreda Education Office
	(Grade(1-3)	1	-	1	Woreda Education Office
	(Grade(1-4)	2	3&(6Alternative Basic Education)	11	Woreda Education Office
	(Grade1-6)	-	2	2	Woreda Education Office
	(Grade1-7)	-	1	1	Woreda Education Office
	Second Cycle Grade (1-8)	8	10	18	Woreda Education Office
3	Secondary school Grade (9-10)	-	3	3	Woreda Education Office
	Secondary and preparatory school Grade (9-12)	1	-	1	Woreda Education Office
4	Preparatory school (Grade 11-12)	-	2	2	Woreda Education Office
5	College	3	3	6	Regional Education Bureau
6	University	-	1	1	Ministry of Education
			Total	74	

Source: Debre Birhan Woreda Education Office.

2.5. Educational Data Mining

Educational Data Mining (EDM) is an emerging discipline, concerns about discovering unique and valuable hidden knowledge in large-scale data that resides in educational system repositories. The emerging of digital data gaining and the improvement of storage technology has resulted in the growth of huge dataset. This can be seen in different sectors; for example at supermarket transactional data, telephone call details, different governmental statistics, different medical records[18].Besides, the techniques and tools for data collecting, storing and transferring for different purposes has also increased. However, this huge amount of stored data needs to be extracted and suitable for gaining information and knowledge [19].This increase the demand of new techniques and tools which leads to the concept of data mining which is the process of

analyzing data from different perspectives and summarizing the results as useful information [20] often set in the broader context of Knowledge Discovery in Databases. Educational Data Mining field pursues to develop and improve data mining methods and techniques to enhance decision making process in the educational system by utilizing the discovered knowledge making the best use of it in making the decisions [21]. It aims to analyze the students' performance, understand learning behaviors, and highlight factors that affect learning process in a particular educational system in the purpose of increasing passing ratio for the students [22].Various data mining techniques can be used by Educational data mining to trim down the students failing ratio and provide recommendations to the educational system stakeholders (i.e. students, teachers, researchers and administrators), Where these recommendations might have a significant impact in improving teaching learning process level [21,22].Classification is data mining techniques used by Educational Data Mining to predict the student's performance in a particular course. This prediction can help in predicting the students' failure in earlier and take proper action to reduce or prevent the failure as much as possible [23]. Different users and stakeholders involved and can be benefited from Educational Data Mining (EDM) are discussed as follows: Learners: Educational data mining offers several advantages for students where it can help in improving students' performance by recommending tasks, activities, and resources based on pedagogical behavior. It can also make early prediction for student performance and highlight their weakness points [23]. Educators: Analyzing students' data can help teachers understand and improve learning process by determining effective activities, identifying weaknesses in their teaching methods, classifying students and detecting students who need more support and help [23]. Researchers: Educational Data Mining (EDM) helps researchers in evaluating course structure and improving course contents, and hence, increase institutional effectiveness [21]. Administrators: EDM helps administrators to make the best utilization of institutional resources (Human and materials) and enhance decision making process [24]. Similar to other assessment methods, Educational Data Mining approach is a new insight in Educational Environment to: State art regarding the prediction of academic performance using data mining techniques, Preprocess dataset consists duplicate records, attribute selection, and data integration among others, Describe model for the admitted students to identify their weak and strong side and then to enhance them accordingly and Classify model for predicting the loss of academic status due to low academic performance and show the solutions for many research problems.

2.6. Data Mining Process Models

A process model is the set of tasks to be performed to develop a particular element, as well as the elements that are produced in each task (outputs) and the elements that are necessary to do a task (inputs) [25].The data mining process models can be considered as a methodology to support the processes which leads to find the information and knowledge.

The reasons for using the process models are:

- To organize the knowledge discovery and data mining projects within a common frame work.
- To understand the knowledge discovery process and provide a roadmap while planning and carrying out the projects [26].
- To ensure that the end product will be useful for the users [20].
- To understand the process itself and to understand the concerns and need of the end users [27].End users usually lack perception of large amounts of unused and potentially valuable data and also they are not ready to devote time and resources toward formal methods of knowledge seeking.
- To providing support for managerial processes [26].

Although the models usually emphasize independence from specific applications and tools, they can be broadly divided into those that take into account industrial issues and those that do not. However, the academic models, which usually are not concerned with industrial issues, can be made applicable relatively easily in the industrial setting and vice versa. We restrict our discussion to those models that have been make popular in the literature and have been used in real knowledge discovery projects [14] are discussions is as follows:

2.6.1. Knowledge Discovery in Database (KDD) Process

According to [28] Knowledge Discovery in Database **(**KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The KDD process is an iterative fulfillment of the following steps [29] which is shown below in Figure.2.1.

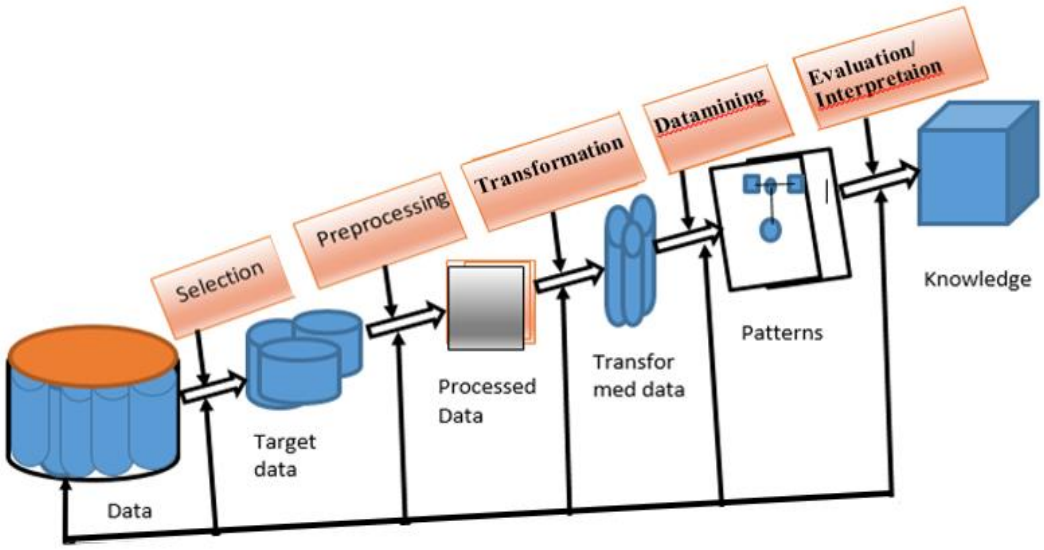


Figure.2.2. Knowledge Discovery in Database (KDD) Process

According to [29] the process of extracting knowledge from big data consists of five sequenced steps namely selection of dataset, data preprocessing, data transformation, data mining, interpretation/ evaluation and brief descriptions of each of the steps in each broadly used methodologies in data mining are as follows:

Step1. Selection of Dataset: In this stage creating a target dataset on focus of a subset of variables needed on which discovery aimed to solve the problem are selected. For discovery purposes, data relevant to the analysis task are retrieved from the database and unnecessary data (outliers) attributes should be removed, checking for errors, handling missing values, and transformation of formats.

Step 2. Data Preprocessing: In order to determine effective data mining models in terms of quality and performance, the raw dataset need to undergo preprocessing in the form of data cleaning. Because real world data are mostly dirty and unclean which need to correct bad data that encountered from data redundancy, incompleteness or missing attributes value, noise, and inconsistency in order to make knowledge searching paths ease for mining algorithms. Therefore, data quality needs to be assured in this step before ahead to next phase of knowledge discovery process in data mining.

Step 3. Data Transformation: During transformation phase, data are combined into forms appropriate for mining to reduce data size by dividing the range of attribute value into intervals each containing approximately same number of samples or to scale attribute data to fall within a

specified range. Therefore, values of attributes are changed to a new set of replacement values to ease data mining. For example, discretization of variables or production of derived variables is transformation of data.

Step 4. Data Mining: Data mining is the next essential process where intelligent methods are applied in order to extract hidden patterns in the data by using classification among major functions such as clustering, association and regression. This phase requires analysis of the preprocessed data for patterns of interest in the data depending on the business objectives and data mining requirements. Different data mining algorithms and techniques are used for searching knowledge or interesting patterns to construct predictive or descriptive models.

Step 5. Interpretation/ Evaluation: This is a post processing step in knowledge Discovery in Databases (KDD) which interprets mined patterns and relationships. If the pattern evaluated is not useful, then the process might again start from any of the previous steps, thus making knowledge Discovery in Databases (KDD) an iterative process.

Knowledge Presentation: Finally, visualization and knowledge representation are used to present the mined knowledge to the users, stored as new in the information base and incorporate it with previously known one in the area are some of the important activities during this phase.

2.6.2. Sample Explore Modify Model Assess Process

SEMMA, like CRISP-DM, grow as industrial standard and define a set of sequential steps that pretends to guide the implementation of data mining applications.

It is developed by the Statistical Analysis System (SAS) institute. The acronym SEMMA stands for Sample, Explore, Modify, Model and Assess and refers to the process of conducting a data mining project has five stages [30]. These stages has described briefly as follows.

Sample -Sampling data by extracting a portion of a large dataset big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out at being optional. Explore-Explore/search data by searching for unexpected trends and anomalies in order to gain understanding and ideas. Modify-Modifying data by creating, selecting and transforming the variables to focus the model selection process. Model-Modeling data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcomes. Assess-Assessing data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs. Although the Sample Explore, Modify, Model and Assess (SEMMA) process is independent from data mining chosen

rings, it is linked to Statistical Analysis System (SAS) Enterprise miner software and pretend to guide the user on the implementation of data mining applications. Sample Explore, Modify, Model and Assess (SEMMA) offers to understand process, allowing an organized and adequate development and maintenance of data mining project [30].

2.6.3. Industrial Models: Cross –Industry Standard Process for Data Mining (CRISP-DM) methodology is applied to build the mining models. It consists of six major phases: Business understanding, data understanding, data preparation, modeling, evaluation, and deployment. In many researches the Cross –Industry Standard Process for Data Mining (CRISP –DM) and the six step model by Cabena et al. fall in this category [31].Industrial models quickly followed academic efforts. Several different approaches were under-taken, ranging from models proposed by individuals with extensive industrial experience to models proposed by large industrial associations. The industrial six-step Cross –Industry Standard Process for Data Mining (CRISP-DM) which is developed by a large association of European companies has become the leading industrial model. The Cross –Industry Standard Process for Data Mining (CRISP-DM) model has been used in domains such as medicine, engineering, marketing, and sale [20].The CRoss-Industry Standard Process for Data Mining (CRISP-DM) consists Cycles of phases in six steps is shown in Figure 2.3.

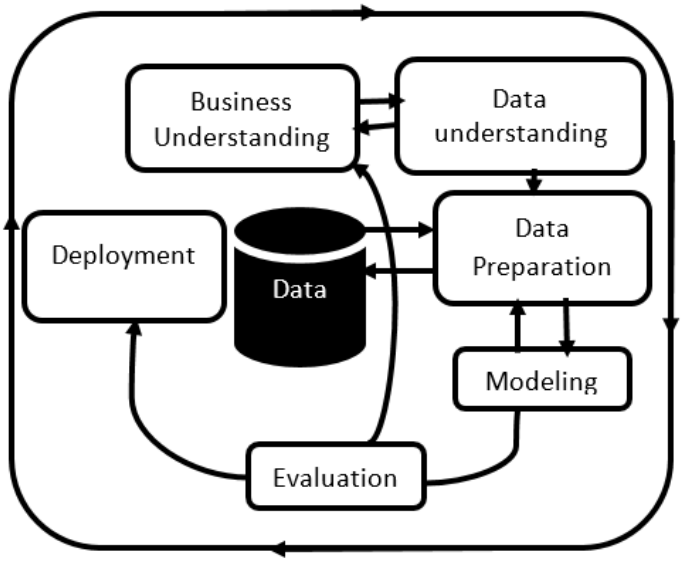


Figure 2.3. CRISP-DM Process Model

Phase.1. Business Understanding. This step focuses on the understanding of objectives and requirements from a business perspective. In the case of this study focuses on understanding the

objectives and requirements from students result data perspective, converting this knowledge into a data mining problem definition, and designing a preliminary plan to achieve the objectives. This stage sub divided in to determination of business objectives, assessment of the situation, determination of Data Mining goals, and Generation of a project plan.

Phase.2. Data Understanding: This step starts with initial data collection and familiarization with the data. Specific aims include identification of data quality problems, initial insights into the data, and detection of interesting data subsets. Data understanding is sub divided into: collection of initial data, description of data, exploration of data, and verification of data quality.

Phase.3. Data Preparation: This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into Data mining (DM) tool(s) in the next step. It includes Table, record, attribute selection, data cleaning, construction of new attributes, transformation of data, integration of data, and formatting of data.

Phase.4. Modeling: At this step, various modeling techniques are selected and applied. Modeling usually involves the use of several methods for the same Data mining (DM) problem type and the calibration of their parameters to optimal values. Since some methods may require a specific format for input data, often reiteration into the previous step is necessary. Similarly, this step is subdivided into: selection of modeling technique(s), generation of test design, creation of models, and Assessment of generated models. The modeling phase selects and applies the best models and calibrates their parameters to optimal values.

Phase.5. Evaluation: After one or more models have been built that have high quality from a data analysis perspective, the model has to be evaluated from a business objective perspective. A review of the steps executed to construct the model is also performed. A key objective is to determine whether any important business issues have not been sufficiently considered. At the end of this phase, a decision about the use of the data mining results should be reached. The key sub steps in this step include: evaluation of the results, process review, and determination of the next step. The evaluation phase evaluates the model to ensure that it achieves the business objectives.

Phase.6. Deployment: Now the discovered knowledge must be organized and presented in a way that the customer can use. Depending on the requirements, this step can be as simple as generating a report or as Complex as implementing a repeatable Knowledge Discovery Process (KDP). This step is further divided into: plan deployment, plan monitoring and maintenance,

generation of final report, and review of the process sub steps. The deployment phase specifies the tasks that are needed to use the models [32].

2.6.4. Academic Research Model

The exertions to establish a knowledge discovery process model were initiated in academia. In the mid-1990s, when the data mining field was being shaped, researchers started defining multi steps procedures to guide users of data mining tools in the complex knowledge discovery world. The main emphasis was to provide a sequence of activities that would help to execute a knowledge discovery process in an arbitrary domain. The two process models developed in 1996 and 1998 are the nine-step model by Fayyad et al. and the eight-step model by Anand and Buchner. Below we introduce the first of these, which is perceived as the leading research model. The Fayyad et al knowledge discovery process model consists of nine steps, which are outlined as follows:

The first step is developing and understanding the application domain includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge. The second step is where data miner selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset. The third step which removing outliers, noise and missing values in the data, and accounting for time sequence information and known changes.

The fourth step is which consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.

The fifth step is where data miner matches the goals defined in Step 1 with a particular Data Mining (DM) method, such as classification, regression, clustering, etc. The sixth step is at which data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate. The seventh step is Data mining that generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc. The eighth step is Interpreting Mined Patterns where analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models. This final ninth step is consolidating discovered knowledge consists of combining the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge.

This is a nine iterative process and a number of loops between any two steps are usually executed, but they give no specific details. The model provides a detailed technical description with respect to data analysis but lacks a description of business aspects. This model has become a foundation of later models.

2.6.5. Hybrid Model

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of both. One such model has six-steps such as understanding of the problem domain, understanding of the data, Preparation of the data, Data mining, Evaluation of the discovered knowledge and Use of the discovered knowledge process (KDP) model developed by [14]. It was developed based on the Cross –Industry Standard Process for Data Mining (CRISP-DM) model by adopting/take on it to academic research.

The main differences and extensions of hybrid Model:It provide more general research-oriented description of the steps, introducing a data mining step instead of the modeling step, introducing several new explicit feedback mechanisms, (the Cross –Industry Standard Process for Data Mining (CRISP-DM) model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains. Each steps of the hybrid model are illustrated in figure 2.3 and described below.

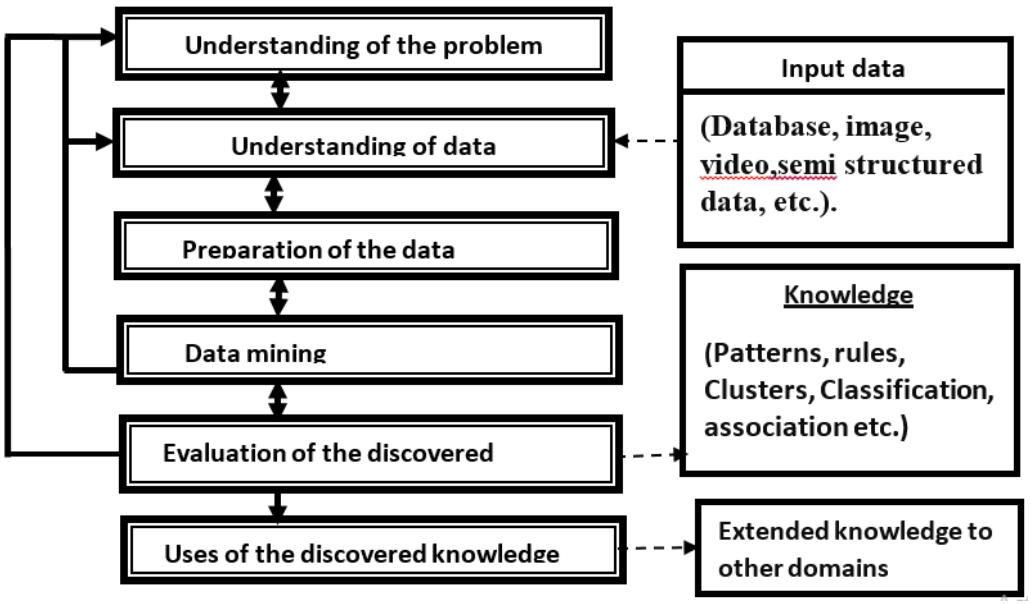


Figure 2.4. The six steps of Hybrid data mining model

Step.1. Understanding of the Problem Domain

This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. Description of the problem, including its restrictions/boundaries, is prepared. Finally, project goals are translated into Data mining (DM) goals, and the initial selection of Data mining (DM) tools to be used later in the process is performed.

Step.2. Understanding of the Data.

This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the Data mining goals.

Step.3. Preparation of the Data

The datasets were undergoes data preparation steps to confirm for completeness, redundancy, missing values and plausibility of attribute. The collected data preprocessed and cleaned in a way to fulfill the requirement of data mining software. In this step of data preparation, tasks like handling missing values, handling outliers' data, transformation of data, and data reduction were taken place. Feature selection and extraction algorithms process handled to acquire cleaned data. The results are data that meet the specific input requirements for Data Mining tools.

Step.4. Data Mining Techniques

Here the data miner uses various classification Data Mining (DM) methods to derive knowledge from the preprocessed data. Among the available algorithms in WEKA machine learning software; Decision Tree, classification rule were used in this research. These models were selected in this research due to their popularity in the recently published documents. WEKA, formally called Waikato Environment for Knowledge Learning developed at the University of Waikato in New Zealand, is open-source data mining software in java. It provides implementations of learning algorithms that can be applied to a given dataset and analyze its output to learn more about the data, and use learned models to generate predictions on new instances [16]. Another possible way to apply WEKA is that of to use the learned models to generate predictions on new instances and compare the performance of the models in order to select the best for prediction [18].

Step.5. Evaluation of the Discovered Knowledge

Evaluation includes understanding the results, checking whether the discovered knowledge is novel/new and interesting, interpretation of the results by domain experts. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared. Thus after the development of the model based on the training dataset, the accuracy of the model were tested using test datasets. A confusion matrix and the result of the model were assessed to determine the impact of the discovered knowledge. Using confusion matrix, accuracy, sensitivity, specificity, and precision were calculated to evaluate the performance of each models.

Step.6. Use of the Discovered Knowledge

This final step consists of planning where and how to use the discovered knowledge. This last step determines the success of the entire knowledge discovery process. A plan to monitor the implementation of the discovered knowledge is created and the entire project is documented. The results of this thesis work will be disseminated/spread to the stakeholders and to any interested parties. The hardcopy of this thesis results will be available to Bibliographic Library of Information System, a softcopy of this thesis will upload to Debre Birhan official website, and maximum effort will be exerted to publish the result in different journals.

2.7. Comparison of the Models

To understand and interpret the KDP models described above, a direct, side-by-side comparisons shown in Table 2.1below.

Table 2.2. KDP Models direct, side-by-side comparisons

KDD	SEMMA	ACADEMIC RESEARCH	CRISP-DM	Hybrid
Pre KDD	----------	Developing and Understanding the Application Domain	Business understanding	Domain understanding
Selection	Sample	Creating a Target Data Set	Data understanding	Data understanding
Preprocessing	Explore	Data Cleaning and Preprocessing		
Transformation	Modify	Data Reduction and Projection	Data preparation	Data preparation
Data mining	Model	Choosing the Data Mining Task	Creation of Modeles	Data mining
		Choosing the Data Mining Algorithm		
		Data Mining		
Interpretation/ Evaluation	Asses	Interpreting Mined Patterns	Evaluation of modeles	Evaluation of theDiscovered Knowledge
Post KDD	----------	Consolidating Discovered Knowledge	Deployment(or ganizing and present discovered knowledge)	Use of the Discovered Knowledge

From Table 2.1, the comparison of the models, it can be obtained that some of them follow the same steps to discovery process while others follow different steps. For example in KDD and SEMMA stages the first approach is equivalent. Sample can be identified with selection. ; explore can be identified with preprocessing; modify can be identified with transformation; model can be identified with data mining; assess can be identified with interpretation/evaluation. Cross –Industry Standard Process for Data Mining (CRISP-DM) compare to hybrid data mining is most of the components are similar but on data mining identified with modeling; knowledge discovery can be identified with deployment. Therefore CRISP-DM is mostly used to project work but hybrid is used to research work [33]. According to the comparison of models on Table 2.1, hybrid data mining process model is the best of all that have choosed for this study. Because, hybrid Data Mining process is a step which consists of methods that produce useful patterns or models from the data (detailed feedback mechanisms).

2.8. Data Mining Tasks

There are two data mining tasks which are known as primary goals of data mining (prediction and description) [34] data mining tasks. Data mining tasks like classification, regression, and prediction and Time series analysis are categorized under predictive data mining techniques

whereas clustering, association, rule discovery, summarization and sequence analysis are categorized under descriptive data mining tasks. Figure 2.4 illustrates the two category of data mining tasks. The description of each tasks presented below.

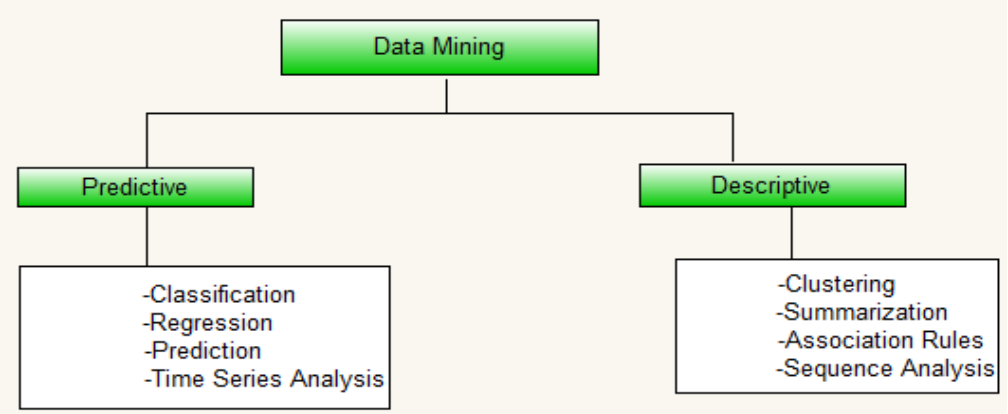


Figure.2.5. Data mining tasks.

Prediction is the objective of data mining which has an aim of predicting unknown or future values of the attributes of interest using other attributes in the databases and it is referred to as supervised learning, since calculated or estimated values are compared with known results Whereas **description** is data mining task which has an aim of describing the data in a manner understandable and interpretable to humans and its techniques are referred to as unsupervised learning which interrogate the database to identify patterns and relationship in the data. The relative importance of description and prediction depend on the use in different applications [28]. Prediction requires having labels for the output variable for a limited data set, where a label represents some trusted "ground truth" information about the output variable's value in specific cases [5].

2.8.1. Classification

Classification is to classify items into several predefined classes. It is one of the most common tasks in supervised learning, but it has not received much attention in temporal/time based data mining [35]. The classification task is characterized by a well-defined definition of the class labels, and a training set consisting of pre classified examples. The task is to develop a classifier model of some kind that can be applied to unclassified data in order to classify it. The developed model is based on the analysis of a set of training data whose class label is known and the derived model may be represented in various forms such as IF-THEN rules, decision trees, mathematical formulae, semantic network etc. [36]. Each technique employs a learning

algorithm to identify a model that best fits the relationship between the attributes set and the class label of the input data. The techniques are lists as the following [36]. Some popular classification methods include decision tree, logistic regression **.**In this study, decision tree and rule induction are discussed. A Rule-based classifier extracts a set of rules that show relationships between attributes of the data set and the class label. It uses a set of IF-THEN rules for classification. Rules are easier for humans to understand.

Decision Tree

Decision tree technique uses tree structure to builds regression or classification models. In this technique dataset is divided into smaller subsets and at the same time an associated decision tree is incrementally developed. That result in a tree having decision node and leaf nodes. A decision node is one which has two or more branches. Leaf node represents a decision or classification. The root node known as a best predictor is the top most decision node in a tree. Decision trees handle both numerical data and categorical data [28]. The final result is a tree with leaf nodes and decision nodes where the leaf represents a decision or classification [37].The decision tree algorithm is probably the most popular data mining technique because of the fast training, performance, a high degree of accuracy, and easily understandable patterns. Splitting your data into subsets is the main idea behind the algorithm [38]. When decision tree induction is used for attributes subset selection, a tree is constructed from the given labeled data. All attributes that do not appear in the tree are assumed to be irrelevant. There is a large number of decision-tree induction algorithms described primarily in the machine-learning and applied-statistics literatures that construct decision trees from a set of input-output training samples [29]. In decision tree construction, selection of splitting attributes is necessary in order to avoid irrelevant attributes by examining the effect of each attribute for the distinct class and its likelihood for improving the overall decision performance of the tree, since the feature with minimum impact on dependent variable may distort the tree's performance and the classification accuracy.

One of the most attractive aspects of decision trees lies in their interpretability especially with respect to the construction of decision rules which is constructed from a decision tree simply by traversing any given path from the root node to any leaf [29]. Therefore, to make a decision tree model more readable, a path to each leaf can be transformed into an IF-THEN rule. Figure 2.5 illustrates the root node and leaf node as follows [39].

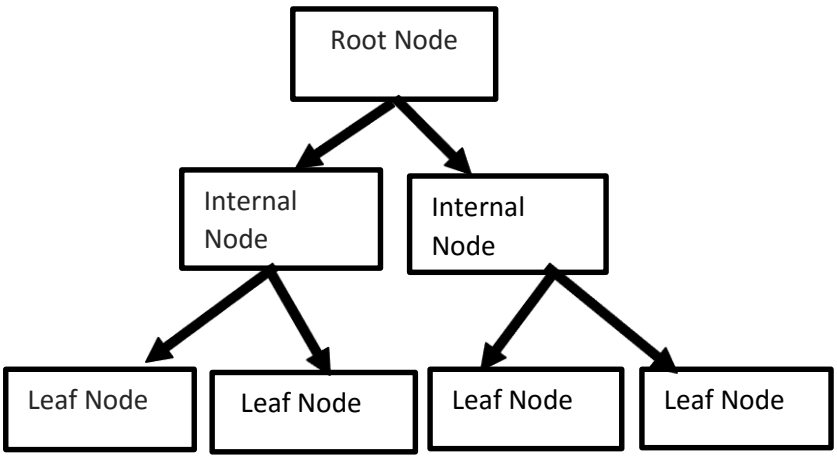


Figure 2.6. Decision tree Structure

The challenge with decision tree is over fitting. As the dataset grows larger and the number of attributes grows larger, we can create trees that become increasingly complex [36]. This potentially leads to the concept of over fitting which consequently brings the notion of pruning; this implies removing of the branches of the classification tree in order to make tree as simple and compact as possible, with as few nodes and leaves as possible. This is done through pruning a tree by halting its construction by partition the subset of training tuples at a given node or removing sub trees from a fully grown tree [36].

The main advantages of decision trees over other algorithms are that they are quick to build, efficient and easy to understand as each node is labelled in terms of the input attributes. The basic algorithm for decision tree induction is greedy algorithm that constructs decision trees in a top-down recursive divide and conquer manner [40]. The algorithm is summarized as follows:

Create a node N;

If samples are all of the same class, C then

Return N as a leaf node labeled with the class C;

If attribute-list is empty then

Return N as a leaf node labeled with the most common class in samples; select test-attribute, the attribute among attributes-list with the highest information gain; label node N with test-attribute; for each known value AI of test-attribute grow a branch from node N for the condition test-attribute= ai;

Let'si be the set of samples for which test-attribute= ai;

If si is empty then attach a leaf labeled with the most common class in samples;

else attach the node returned by Generate_decision_tree (si, attribute-list test-attribute).

Classification Algorithm

Decision Tree J48: C4.5 is an evolution of Dichotomiser ID3, presented by Quinlan J.R [41] for generating a pruned or un-pruned C4.5 tree and all the possible tests are considered during decision making based on information gain value of each attribute [42]. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets.

REP Tree: It Reduces Error Pruning (REP) Tree where Classifier is a fast type of decision tree learner which is built for the decision tree or for the regression tree by using the information gain with entropy and as in C4.5 Algorithm, which deals with the missing values by breaking the corresponding instances into pieces [43].

Rule Based Classification Algorithms

A rule is represented by the IF-THEN form, where the IF part is called the condition and the THEN part is called the action [44].The basic unit and format of knowledge in rule-based reasoning is the rule. The IF-THEN rules are quite natural for humans and are easily understood by both programmers and domain experts. However, accurate description of the domain expert's knowledge in simple rules is often difficult.

IF Condition THEN Conclusion

The rule based classifier is constructed on the concept that IF the information supplied by the user satisfies the conditions of a rule, THEN the actions of the rule are executed [44].For example, one could have the following set of rules to classify the student academic performance. According to [45], knowledge extracted from dataset, IF performance of student in English subject at every level of Grade= Excellent or very good and Maths= Excellent or very good and related subjects, THEN performance = Excellent. The advantage of IF-THEN rule is the rules are order independent i.e. regardless of the order of rules executed, the same classification of the classes is possible to reach [44]. PART and JRIP are algorithms are an example of rule based classifiers.

PART: It is a separate-and-conquer/master rule learner. The algorithm producing sets of rules called decision lists which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in its each iteration and makes the best leaf into a rule. The algorithm is a combination of C4.5 and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) rule learning [46].

JRIP: It implements a propositional rule learner. JRip proposed a Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is an inference and rules--based learner (RIPPER) that can be used to classify elements with propositional rules. The RIPPERR algorithm is a direct method used to extracts the rules directly from the data [46]. JRIP (WEKA'S implementation of the RIPPER rule learner) is a fast algorithm for learning "IF THEN" rules. Like decision trees rule learning algorithms are popular because the knowledge representation is very easy to interpret.

2.8.2. Regression

Regression, sometimes also called estimation, is a kind of statistical estimation technique which is used to map each data object to a real value provided prediction value [47]. In prediction the aim is to predict a value of a given continuously valued variable based on the values of other variables, assuming either a linear or nonlinear model of dependency [48]. That means the estimation approach has the great advantage that the individual records can be rank ordered according to the estimate. Uses of regression include prediction, modelling of causal relationships, and testing hypotheses about relationships between variables. Well suited techniques for regression tasks are (linear) regression models and none linear regression [49].

2.8.3. Time Series Analysis

In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

On the descriptive end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. Clustering, Summarization, and Visualization of databases are the main applications of descriptive data mining. The usefulness of this concept is that it enables one to generalize the data set from multiple levels of

abstraction, which facilitates the examination of the general behavior of the data, since it is impossible to deduce that from a large database.

2.8.4. Clustering

According to [50] explained, clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. Clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Clustering is commonly used to search for unique groupings within a data set. The distinguishing factor between clustering and classification is that in clustering there are no predefined classes and no examples [51]. The difference with the classification task is that clusters were unknown at the time the algorithm starts. In other words, there are no predefined classes it relies on [51]. The objects are grouped together based on self-similarity and the applicability can be seen, for instance, in discovering new student behavior pattern.

2.8.5. Summarization

The aim of summarization is to find a brief description for a subset of data [52]. Tabulating the mean and standard deviations for all fields is a simple example of summarization. There are more sophisticated techniques for summarization and they are usually applied to facilitate automated report generation and interactive data analysis (Dependency modeling – to find a model that describes significant dependencies between variables. For example, probabilistic dependency networks use conditional independence to specify the structural level of the model and probabilities or correlation to specify the strengths (quantitative level) of dependencies [28].

2.8.6. Association Rule Discovery

Association Rule is a descriptive data mining task which is one of the relationship mining like as correlation mining, sequential pattern mining, and causal data mining methods. It includes determining patterns, or associations between elements in datasets. Associations are represented in the form of rules. The association technique is used for associating tasks [49]. Examples are market basket analysis, which more or less is determining what things go together in a shopping cart at the supermarket, and cross selling programs, which helps to design attractive packages or groupings of products and services [51]. One of the tasks of data mining is association rule mining. Association rule mining finds interesting association or correlation relationships among a large dataset [44].With a massive amounts of data continuously being collected and stored, many industries became interested in mining association rules from their datasets and the discovery of interesting association among huge amount of business transaction records can help

in many business decision making processes [44]. Association rules are in the form of "If antecedent, then consequent," together with a measure of the support and confidence associated with the rule.

2.8.7. Sequence Analysis

Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend. Given is a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events.

2.9. Attribute Selection Measures

An attribute selection measure for developing decision tree is a heuristic for selecting the splitting criterion that best separates a given data partition of class-labeled training instances into individual classes. The attribute selection measure provides a ranking for each attribute describing the given training instances. The attribute that has the best score for the measure is chosen as the splitting attribute for the given instance. The tree node created for partition, let's say N, is labeled with the splitting criterion, branches are grown for each outcome of the criterion, and the instances are partitioned accordingly [36]. This section describes three popular attribute selection measures, namely information gain, gain ratio, and Gini index.

2.9.1. Information Gain

Information gain for attribute selection measure is based on the work of Claude Shannon on information theory, which studied the value or information content of messages. Iterative Dichotomiser (ID3) uses information gain for attribute selection measure. The notion used is as follows: - Let D, the data partition, be a training set of class labeled instances. Suppose the class label attribute has m distinct values defining m distinct classes, Ci (for i=1… m). The attribute with the highest information gain is selected as the splitting attribute. This attribute minimizes the information needed to classify the instances in the resulting partitions and reflects the least impurity in these partitions. Entropy (impurity) is used to measure the information content of the attributes. High entropy means the attribute is from a uniform distribution whereas low entropy means the attribute is from a varied distribution. Entropy is defined as follows. Let **pi** be the probability that an arbitrary instance, in D belongs to class Ci, estimated by $\frac{|\mathbf{Ci},\mathbf{D}|}{|\mathbf{D}|}$.Expected information (entropy) needed to classify an instance in D is given in equation 2.1:

Entropy (E (D)) =$\sum_i^m p_i log(p_i)$ ……2.1

Entropy (E (D)) - is the average amount of information needed to identify the class label of an instance in D. The smaller information required, the greater the cleanliness. At this point, the information we have is based solely on the proportions of instances of each class. A log function to the base 2 is used, because the information is encoded or measured in bits. Suppose attribute A can be used to split D into v partitions or subsets, {D1, D2… Dv}, where Dj contains those instances in D that have outcome aj of A. Information needed (after using A to split D) to classify D is shown in equation 2.2.

InfoA (Dj)$\sum_j^v = 1\ \frac{Dj}{D} * Info(Dj)$..... (2.2)

The smaller the expected information required, the greater the purity of the partitions. Information gained by branching on attribute A is given by Gain (A) =E (D)-Info A (D) as shown in equation 2.3.

Gain (A) =E (D)-Info A (D)……... (2.3)

Information gain increases with the average purity of the subsets. The attribute that has the highest information gain among the attributes is selected as the splitting attribute.

2.9.2. Gain Ratio

The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values. This may result in the selection of an attribute that is non-optimal for prediction. C4.5, a successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias. It applies a kind of normalization to information gain using a split information value defined analogously with Info (D) as shown in equation 2.4.

SplitInfoA (D) =-$\sum_j^v = 1 \frac{|Dj|}{|D|} log \frac{|Dj|}{|D|}$…… (2.4)

This value represents the potential information generated by splitting the training dataset, D, into v partitions, corresponding to the v outcomes of a test on attribute A. Note that, for each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in D. It differs from information gain, which measures the information with respect to classification that is acquired based on the same partitioning [28].

The attribute with the maximum gain ratio is selected as the splitting attribute. Note, however, that as the split information approaches 0, the ratio becomes unstable. A constraint is added to

avoid this, whereby the information gain of the test selected must be large at least as great as the average gain over all tests examined.

2.9.3. Gini Index

The Gini index is used in Classification & Regression Trees CART. Using the notation described above, the Gini index measures the impurity of D, a data partition or set of training tuples [53], as described in eq. 2.5 as follows:

Gini (D) =1-$\sum_{i}^{m} = 1\, p_i^2$ ……...… (2.5)

Where pi is the probability that a tuple in D belongs to class Ci and is estimated by$\frac{|\mathbf{Ci,D}|}{|\mathbf{D}|}$**.** The sum is computed over m classes. To determine the best binary split on A, we examine all the possible subsets that can be formed using known values of A and need to enumerate all the possible splitting points for each attribute. If A is discrete valued attribute having v distinct values, then there are 2v-v possible subsets. When considering a binary split, we compute a weighted sum of the impurity of each resulting partition. If data set D is split on A into two subsets D1 and D2, the Gini index Gini (D) is defined as [36].

GiniA (D) =$\frac{|D1|}{|D|}$Gini (D1) + $\frac{|D2|}{|D|}$Gini (D2)…….……. 2.6)

First, we calculate Gini index for all subsets of an attribute, then the subset that gives the minimum Gini index for that attribute is selected. The point, giving the minimum Gini index for a given (continuous valued) attribute is taken as the split-point of that attribute. The reduction in impurity that would be incurred by a binary split on attribute A is:

ΔGini=Gini (D) – GiniA (D)…….… (2.7)

The attribute that maximizes the reduction in impurity (or has the minimum Gini index) is selected as the splitting attribute.

To summarize, the three measures for attribute selection are used mostly. Information gain is biased towards multi valued attributes. Whereas Gain ratio tends to prefer unbalanced splits in which one partition is much smaller than the others. Gini index is biased to multi valued attributes and has difficulty when the number of classes is large. The algorithm used each attribute of the data to make decisions by splitting the data into smaller subsets. All the possible tests are considered during decision making based on the information gain value of each attribute.

2.10. Implementation Tools

In order to mine hidden knowledge from the pre-processed dataset and compare the performance of classifiers, WEKA 3.6.13 version is used. WEKA is chosen since it is proven to be powerful for data mining and used by many researchers for mining task and the researchers is familiar with the tool. It contains tools for data preprocessing, clustering, regression, classification, association rules and visualization. WEKA is written in the Java language and contains a Graphical User Interface (GUI) for interacting with data files and producing visual results. Additionally , in order to develop an application which maps the knowledge acquired from the data mining classifiers with rule based system Java Net Beans IDE 8.2 with JDK -8u20 is employed [16]. Net Beans offers easy and efficient project management, has best support for latest java technologies, and can be installed on all operating systems supporting java.

2.11. Evaluation Method

The standard classifier performance measures such as Prediction Accuracy, True Positive, False Positive, Precision, Recall and F-Measure are commonly used [54]. Confusion matrix helps to see a breakdown of a classifier‘s performance by showing how frequently instances of a class let us say class X are classified as class X or misclassified as some other class, say class Y [55]. According to Weiss and Zhang [56], a model performance evaluation should answer questions such as: How accurate is the model? How well does the model describe the observed data? How much confidence can be placed in the model‘s predictions? How understandable is the model?

The application’s results help to interpret the generated models in relation to the research question posed in the first chapter. The classification algorithm predicts the class label. The final output will be patterns which are used to find out the performance of students. Some of the performance measures are given below in Table 2.3 Confusion Matrix.

Confusion Matrix

Confusion matrix can show the selected model prediction performance by comparing it to the actual value. In evaluating the performance of a model, a confusion matrix or correct classification matrix can be used. Confusion matrix focuses on the predictive capability of a model rather than how fast takes to classify or build models, scalability, etc. [57]. In the class case prediction, the result is often displayed as a two dimensional confusion matrix with a row and column for each class.

Table 2.3. Two dimensional confusion matrix

		Predicted class	
Actual class	Class (+)	Class (+)	Class=(-)
		True Positive(TP)	**False** Negative(FN)
	Class=(-)	**False** Positive(FP)	**True** Negative(TN)

The two-class case with classes yes and no, a single prediction has the four different possible outcomes shown in table 2.3. The true positives (TP) and true negatives (TN) are correct classifications. A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive. The four outcomes of classifier and meanings are figuratively as follows.

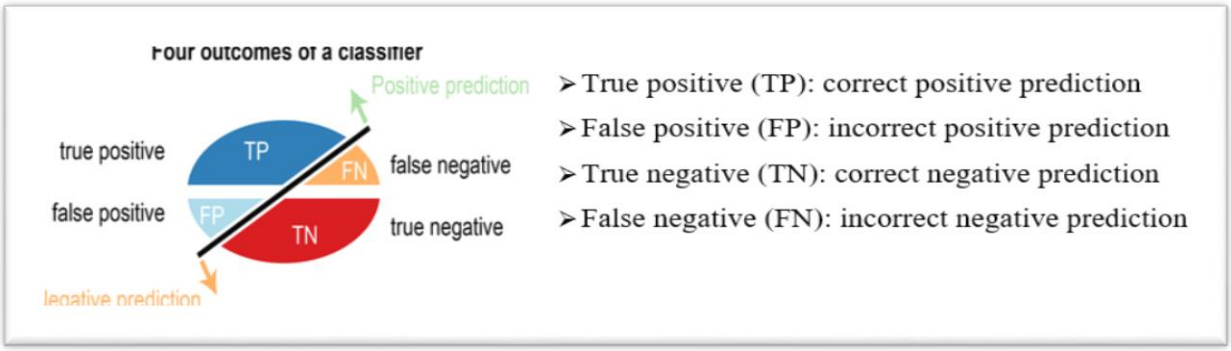


Figure 2.7.Four Outcomes of Classifier

2.11.1. True Positive Rate (TPR)

True Positive rate (TPR) is the proportion of positive or correctly classified instances as positive or correct instances. It is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true Specificity (SP). The best True Positive rate is 1.0, whereas the worst is 0.0.

$$TPR = \frac{TP}{TP+FN} \quad \text{..... (2.8)}$$

2.11.2. False Positive Rate (FPR)

The False Positive (FPR) rate is measures the proportion of negative instances that are erroneously classified as positive. It is calculated as the number of incorrect positive predictions divided by the total number of negatives. The best false positive rate is 0.0 whereas the worst is 1.0. It can also be calculated as 1 – specificity.

$$FPR = \frac{FP}{TN+FP} \quad \text{(2.9)}$$

2.11.3. Precision

Another metrics for performance evaluation of the classifier is precision which measures what percent of tuples that the classifier labelled as positive are actually positive. The best precision is 1.0, whereas the worst is 0.0.

$$Precision = \frac{TP}{TP+FP} \quad \text{(2.10)}$$

2.11.4. Recall

Recall is also another performance evaluation which measures what percent of positive tuples the classifier labelled as positive for both True and False classes. It is what percent of positive/ negative tuples the classifier labeled as positive or negative for both True and False Classes. They are summarized in the following formulas [55].The formula is:

$$Recall = \frac{TP}{TP+FN} \quad \text{(2.11)}$$

2.11.5. F-Measure

The final metric used for performance evaluation of classifiers on confusion matrix is F-measure. The F- measure is the inverse relationship between precision & recall, and calculated as the harmonic mean of recall and precision. It is the point to conclude that the precision and recall of the model are significantly balanced [55].F-Measure: is calculated as the harmonic mean of recall and precision.

$$\textbf{F-Measure (F)} = \frac{2*recall*percision}{recall+Percision} = \quad \text{(2.12)}$$

2.11.6. $$TNR = \frac{TN}{FP+TN} \quad \text{(2.13)}$$

Specificity is calculated as the number of correct negative predictions (TN) divided by the total number of negatives (N). Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best specificity is 1.0, whereas the worst is 0.0.

2.11.7. Predictive Accuracy

$$\text{Predictive Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(2.14)}$$

Among the standard performance metrics, accuracy is the most widely-used metric to check the performance of the model. The correctness accuracy for a data mining classifier is defined as the degree of closeness of its prediction to the actual values, either true or false [55].The accuracy of

a classifier is estimated by dividing the total correctly classified positives and negatives instance by the total number of samples. The accuracy [28] of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Prediction accuracy measures the proportion of instances that are correctly classified by the classifier. Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0.

2.12. Related Works

Data mining in education is a recent research field and many works have done in this area. This is because of its potential to educational institutes. Many studies have used educational data mining to analyze students learning behavior and to predict performance of students [58]. This section investigates related works done on prediction of students' academic performances.

M. Anwar [5], has developed a model that predicts the first year, first semester grade point average (GPA) of Jimma University students. The researches dataset incorporates attributes of Ethiopia Higher Education Entrance Examination (EHEEE) result, average point of preparatory school result, aptitude score, mathematics score, English score. To develop the proposed model by using WEKA, he has compared the classifier techniques J48 decision tree, PART rule induction, Naïve Bayes algorithms. The highest performance is obtained un-pruned PART rule induction algorithm with the accuracy of 90.7%. The number of instances in his dataset was used is 3749 with 17 attributes and the class labels are good, average and poor.

S. Pal [59] has developed predictive model that manages student's dropout using the Naïve Bayes classification algorithm. His proposed algorithm has been applied on the collected dataset from India that consists of 165 instances and 17 independent attributes with one of the two class label namely 'yes' and 'no' parameters. His developed classifier model shows accuracy of 87% performance result. The study also shows that the male students have greater possibilities to quit the study after a year than female students.

Muluken [58] has developed predictive model that predicts Student Success and Failure by using two classification algorithm namely J48 Classifier and Naïve Bayes. The proposed Models Compared to discover the best model for predicting student performance SAP to identify students at risk. Only one independent parameter used in their study, which is the midyear mark in order to predict the students of Computer Science final marks. The classification rule generation process is based on the decision tree and Bayes as a classification technique from

Sample dataset of 11,873 instances. The J48 classifier performed better with 92.34% of the tuples being predicted correctly compared to accuracy for Naïve Bayes. CRISP-DM (Cross Industry Standard Process for Data mining) is a data mining methodology has used by the research. Analysis is done by using WEKA3.7 application software on attributes such as Sex, age, region, Higher Education Entrance Certificate Examination result, field of study, College, Number of courses given in a semester, Total credit hours given in a semester, Number of students in a class, semester Cumulative GPA of a student. Students' performance were categorized in five groups: Very good, with a high probability of succeeding; Good students, who are above average and with a little more effort can succeed with good Grades ; Satisfactory students, who may succeed; Below Satisfactory students, who require more efforts to succeed; and Fail, who have a high probability of dropping out.

T.M. M. A., & El-Halees, A. M [60] have developed Predictive model that predicts the Students' Academic Performance of graduate students. The predictive model is developed by using two classification methods namely Rule Induction and Naïve Bayesian classifier applied on graduate students' database from college of Science and they got that Naïve Bayesian classifier is the best technique to estimating probabilities of individual variable values. The dataset was collected from the college includes 3360 records and 18 attributes for a period of fifteen years [1993 to 2007]. Finally, they got performance with acceptable accuracy of 67.5% in predicting graduate students' Grades/class (Excellent, Very good, Good, Average). The Rapid Miner software was used for applying the methods on the graduate student's data set. The GPAs categorized into five segments such as Excellent, Very good, Good, Average and Poor class.

A.A.S.Saleh [61] has proposed a student analysis system or predictive model that help to predict the students' technical and academic skills. The dataset consists of a total 1056 records and 10 attributes such as Student ID, Student Name, Campus, Gender, GPA, Term 1 mark, term 2 mark, Term 3 mark and End of year assessment. He uses classification by C4.5 algorithms ID3, CART Naïve Bayes, and KNN algorithms. The proposed system is mainly developed by applying C4.5 and justified C4.5 algorithms on student's database where bootstrapping technique is used to validate the proposed model. The best accuracy obtained in this study is 65.6% using WEKA tool.

Tariku [62] has aims to apply data mining for identifying the determinant factors for the students' success in the preparatory schools to join higher education. His study was focused on

Natural Science stream preparatory students and the dataset was collected from National Educational Assessment and Examination Agency (NEAEA) students EHEEE dataset and correspondingly their EGSECE dataset of three years 2006,2007 and 2008 E.C and 2004, 2005 and 2006 E.C respectively. He was uses Hybrid data mining model for his study since it is a research oriented model and WEKA 3.6.1.3 used for data mining, Microsoft Excel 2013 for data visualization and KU TOOLS and data integration. Finally, 40328 instances and 15 attributes (Sex, Age, Biology, Chemistry, civics, English, math, physics and class.) of Grade 10 and Grade 12 are selected for analysis and the values of some of the attributes are discretized using the assessment system of secondary education in Ethiopia which is categorized as Excellent, Very good, Good, Satisfactory and Fail and the class is yes or no. Association rule mining method such as Apriori and Filter Associator algorithm compared and Apriori algorithm was applied in order to get the results. By configuring different thresholds, different rules are achieved. The discovered rules are then evaluated using the interestingness measure lift or correlation and domain experts. From the study he has got that scoring Very good in Physics, Civics and Biology subjects in EHEEE are determinant factors for the students 'success in the preparatory schools based on lift value. Similarly scoring good in English in EHEEE is also another determinant factor. But in using Apriori algorithm, there is no standard way of setting different thresholds. This leads to missing the strong rules.

Abdellah [63] has developed model that Early predicts the academic performance of new entrant students in higher education institutions for natural science related streams to improve the student's academicals performance and skills. The dataset included all the data in the SIMS-DBU from 2006 E.C to 2008 E.C. consists of a total 22184 records and 16 attributes such as sex, region, contact, mother job, father job, stream, mathematics, English, biology, chemistry, physics, civics, aptitude, total, semester, and class. The values of some of the attributes are discretized using the assessment system of secondary education in Ethiopia which is categorized as Excellent, Very good, Good, Satisfactory and Fail and the class is (pass or fail).The researcher uses classification by J48 algorithms and PART algorithms. The proposed system is mainly developed by applying J48 and justified on student's database where bootstrapping technique is used to validate the proposed model. The system correctly classifies 94.32% and system performance of 84% prediction by using WEKA tool and user acceptance test 85.2%. The prototype has achieved the total average accuracy of 84.6%.

Solomon [64] has aims to analyses and predict the correlation between English, Mathematics and science subjects in terms of student academic result in Grade 10 and 12 NEAEA dataset by using Aprior data mining techniques which mines required information. The data used for the study was Grade 10 and 12 students' national examination result in 2005 E.C and 2007 E.C respectively. The findings of the research showed that there is a positive relationship between English, Mathematics and Science subjects excluding physics. The academic achievement result of male students in both five subjects in both Grade 10 and 12 is better than female students. This implies that government and concerned stakeholders in education sector should provide additional support for female students. The summary and comparison of the above related works is shown as follows in Table 2.3.

Table 2.4. Summary and Comparison of the Related Works.

Author &Year	Title	Techniques used	performance achieved
M. Anwar (2015)	Predicting student academic performance in higher education	J48, **PART rule**, Naïve Bayes,	90.7%
Pal, S. (2012)	Predictive Models For Students' Dropout Management	**Naïve Bayes** classification algorithm	87%
Muluken (2015).	Application Of Data Mining Techniques For Student Success And Failure Prediction	**J48**, Naïve Bayes	92.3%
T.M. M. A., &ElHalees,A .M (2013)	Propose A Framework for Predicting Students' Academic Performance	**Naïve Bayes**, rule Induction	67.5%
A.A.S. Saleh	Predict A student grade performance of Graduation.	**C4.5** algorithms, CART Naïve Bayes, and KNN	65.6%
Tariku (2017)	Apply data mining for identifying the determinant factors for the students' success in the preparatory schools to join higher education	Apriori and Filter Associator algorithm	No Mentioned
Abdellah (2016)	Early prediction of new entrant students ' academic performance in higher education institutions for natural science related streams	**J48** , PART algorithms	94.3%
Solomon (2016)	A Correlation Study of Students' Performance in Ethiopian Secondary Schools by Using Data Mining Algorithm	Aprior data mining techniques	No Mentioned

The difference and contributions of this study as compared to the related works is as follows:

- It is a work done to predict the performance of Grade 12 in Ethiopian Higher Education Entrance Examination (EHEEE) result early by using data mining techniques whereas all the related works have done on academic performance prediction of university students.
- The attributes considered for this thesis for training data set comprises of a mix of general students' data alongside academic data of all six common subjects in Grade 9, 10, 10 National Exam and grade 12 of five years of Natural science stream preparatory school students.
- Some previous researches were conducted by Using a Very Small Proportion of the Datasets, one and two algorithms and have achieved low performance compared to this research.

- This research is different in a way that answers how much the selected variables or factors have an impact on students' academic performance in terms of their national exam obtained in Grade 10 (EGSECE) result and preparatory school students national exam (EHEEE) result Natural science stream preparatory school students take common subjects like-English ,Natural mathematics ,Physics ,Chemistry ,Biology and Civics ,class average point of Grade 9, Grade 10, Grade 11 ,Grade 12. This is to help students early before they take the national exams. In addition to this, the research explore the student preference on study area based on Grade 9, Grade 10 class average point and Grade 10 national exam result which determine student's Grade 12 national exam result. So the research assist education sector planners, policy makers, and decision makers as a decision support aid in planning and implementing educational intervention programs aimed at facilitating students, improving the teaching procedures in order to improving quality of education in the regions as well as in the country and additionally the predictive model is used to assist the students timely and give decision to achieve at better performance by using GUI easily.
- The other researchers used two or three class levels that unable to identify the performances of all students in the dataset to help the students accordingly. This is the gap of the related works.

CHAPTER THREE

METHODOLOGY OF THE STUDY

3.1. Introduction

This chapter describes the proposed research methodology to extract hidden knowledge from the collected dataset using Hybrid data mining process by descriptive statistical method. The research is carried out based on primary data extracted from the database of Ethiopian National Assessment Examination Agency which is available for researchers. Secondary data for instance review important document to gain farther information related with student achievement ,Ethiopian General Secondary Education Certificate Examination (EGSECE) result dataset, secondary school (Grade9- Grade10) natural science stream school common subjects Class Average Points preparatory school (Grade 11 and Grade 12) natural science stream school common subjects Class Average Points. Accordingly, in this study both quantitative and qualitative research design that uses analysis by description and number; interview will be used to understand the domain knowledge and to interpret the findings.

3.2. Research Design with Respect to Hybrid Model

The proposed research frame work which is hybrid data process model consists of six steps namely problem domain understanding, data understanding, data preparation, Data mining, Evaluation of the discovered knowledge and Use of the discovered knowledge for knowledge mining from students' academic student datasets. The steps have summarized and illustrates system framework for this study in Figure 3.1 as follows:

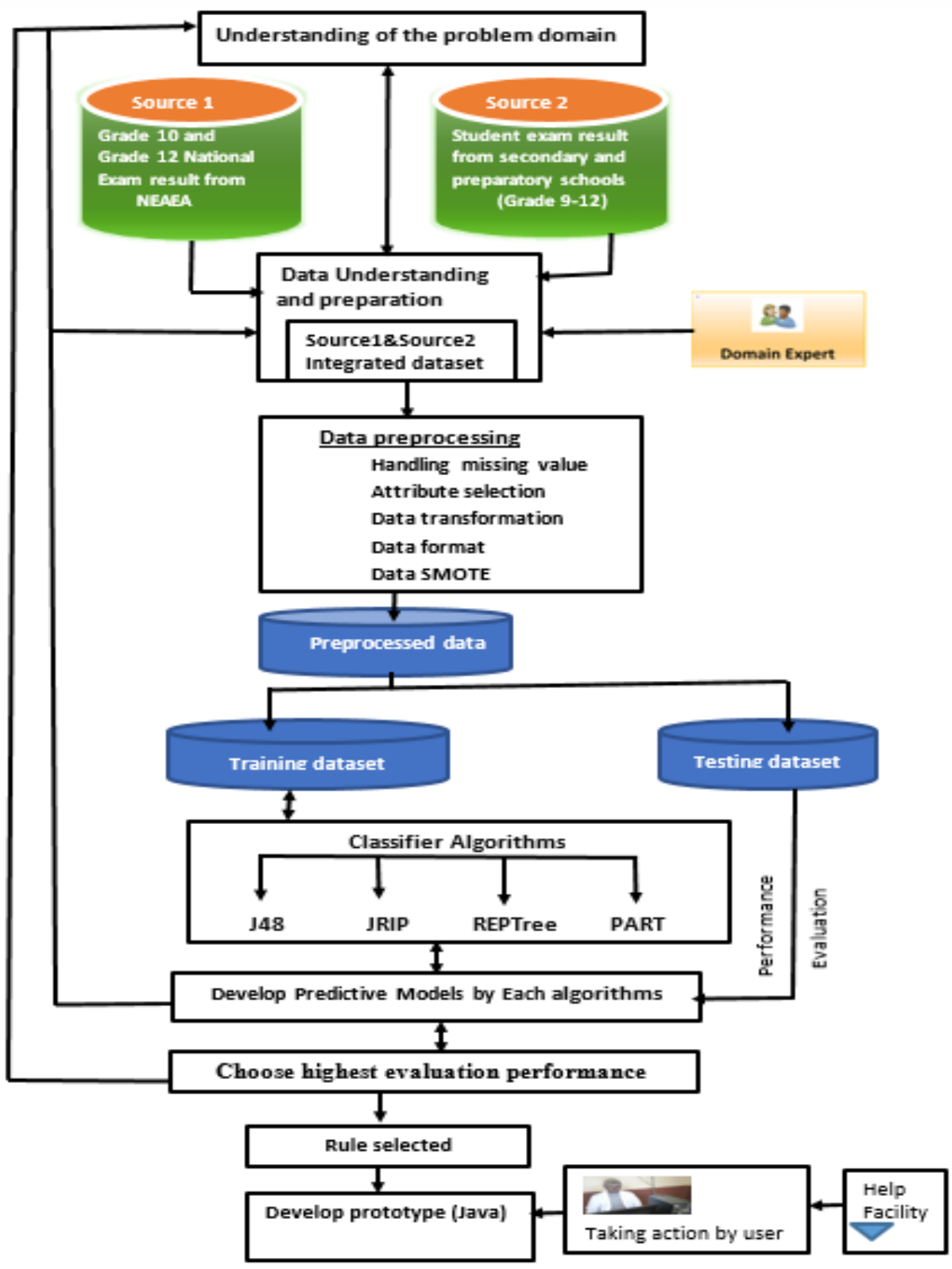


Figure 3.1.Architecture of student performance Prediction in EHEEE.

3.2.1. Understanding the Problem Domain

In any data mining task, the first step is to Understanding the Problem Domain area to be solved. It helps to know the domain problem. Thus, this initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. To understand the problem domain, observation work of the schools, interview, and discussion were also made with the senior Debre Birhan City Preparatory School (DBCPS) stakeholders. Thus, the main task in this research has accomplished by using different mechanisms such as: literature review which is performed in order to study the existing problems at preparatory schools that can be solved by the application of data mining techniques, review of documents including different manuals, Educational policy documents, for instance reviewing the Ethiopian Education policy has an impact to understanding the problem domain and different reports which include education statistics abstract organized by Ministry of Education .

Similarly National Educational Assessment and Examination Agency's (NEAEA) reports which contain different analysis of national Exams has an impact to understanding the problem domain. Discussion with domain experts (high school students and preparatory school teachers) concerning the overall result of Grade 12 students helps to understanding the problem domain. Similarly, in National Educational Assessment and Examination Agency's (NEAEA), there are experts who are direct participants in analysis of Examinees' result and there is a department which governs the students' data; so there has been a discussion with these two bodies to understand the problem domain. In the same way, discussion with the main participants (preparatory school students) about the overall significant factors of their performance. A description of the problem, including its restrictions is prepared.

As interview session determines, the institutions don't use any appropriate technique that helps them to predict the students' performance. Therefore, formal interviews with the management at the schools and departmental levels were also conducted, for finding out the specific problems at schools which have not yet been solved but are considered very important for the improvement of the students' performance and for more effective and efficient control of students. Additional concerns were gathered from informal talks with teachers and students.

Based on the outcomes of the problem understanding, the research questions and objectives were formulated. All the data concerning the student performance at the school were stored in the

preparatory schools including students' personal data. Domain understanding was achieved on appropriate issue giving due to attention on the relevant attributes [65].

In this case firstly, discussion with domain experts the researcher became in a good position in order to define the problem, determine the domain objectives; secondally, identifying key people, and learning about current solutions to the problem; thirdly, learning domain-specific terminologies and finally, project goals are translated into data mining (DM) goals, using the selected algorithms to drive knowledge.

3.2.2. Understanding of the Data and Preparation.

Data understanding is next to domain understanding in hybrid data model [28]. This step includes collecting sample data and deciding which data will be needed including data format and size. Background knowledge can be used to guide these efforts and data are checked for completeness, redundancy, missing values, acceptability of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the Data mining (DM) goals.

Thus, the initial dataset collected from two sources namely Ethiopian Higher Education Entrance Examination (EHEEE) from National Educational Assessment and Examination Agency (NEAEA) and the corresponding preparatory schools namely Haile Mariam Mamo Preparatory School and Basso in Debre Birhan city is 3206 instance and 54 attributes in the interval academic years (2005 -2009 E.C). The data collected from National Educational Assessment and Examination Agency's (NEAEA) has attributes of sex, age, school code, school name, sight, nationality, ID number, Year, Father name, Grandfather Name, Name of student, English, Maths, Physics, Biology, Chemistry, Aptitude and Civics and Ethical Education for grade 12 and Grade 10 Natural stream students whereas the data collected from corresponding preparatory schools consists of attributes like age, sex, Grade 9- Grade 12 common related courses (English, Maths, Physics, Biology, Chemistry, and Civics and Ethical Education) Grade Average Points. These two datasets are integrated with the common attributes of age, sex, English, maths for natural science stream, Physics, Chemistry, Biology, Civics.

After the integration of the dataset by using KU TOOLS software that integrates the datasets by name of students and reviewing of private attributes like Name of students, age, sex, and the common subjects for Natural Science students (English, Maths, Physics, Biology, Chemistry and Civics and Ethical Education). The Dataset consists of 3013 instances who have integrated with corresponding's name and 32 attributes by removing irrelevant attributes with domain experts.

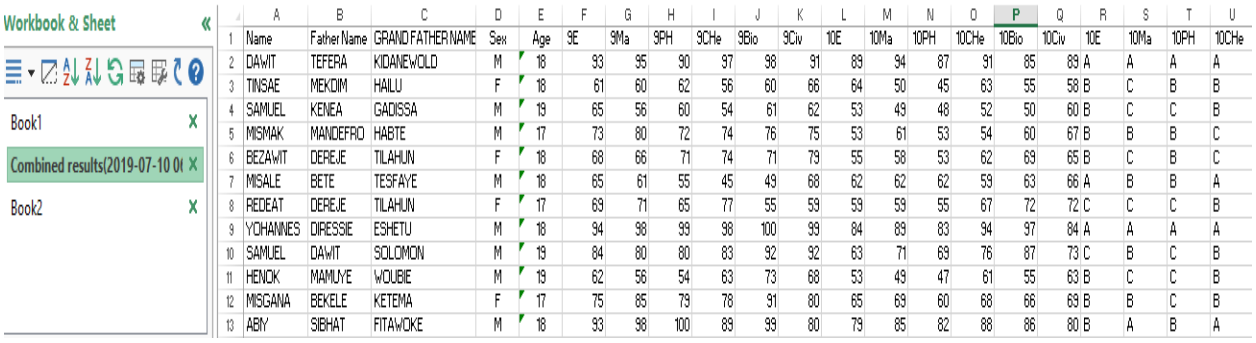
The assigned class labels are Excellent, Very Good, Good, Satisfactory and Fail based on the educational result Assessment System for different Grade levels in Ethiopia but the cutting point is different from year to year which is based on the decision given by ministry of education (MOE) by considering the intake capacity of the institutions of the country. The Assessment System of Grade 9-Grade 12 Class Average Point, Ethiopian Higher Education Entrance Examination (EHEEE), and Ethiopian General Secondary Education Certificate Examination (EGSECE) result illustrated in Table 3.5, 3.6, 3.7 and 3.8.The number of the students who took EHEEE Examination from the corresponding schools in the academic year is illustrated in Table 3.1. It shows the number of the students before and after integration. In each year students who scored the cutting point and above in the academic year join the higher institutions of the country in regular program and those score below the cutting point of the academic year didn't join the university in regular program. But in each year many students have scored satisfactory performances which ranges from 177-352 total point .This includes many scores which are less than the passing point (half of the total score which is (350).Because, the students take seven subjects like: English, Maths, Physics, Chemistry, Biology, Civics and Aptitude at EHEEE .

Table 3.1. The number of instances in the five years and integration

Categories of dataset for EHEEE, EGSECE HMMPS and BPS Students natural science common subjects of 9-12common subjects' Average points.	Before integration	Instances which haven't got the similar name in grade9, 10, 10NE, 11and 12.	After integration
	Number of Instances(before merged with name)		Number of Instances (after merged with name)
2005E.C.EHEEE, 2003 E.C EGSECE & Grade9-12common subjects Average points.	641	46	595
2006E.C.EHEEE, 2004 E.C EGSECE & Grade9-12common subjects Average points.	644	57	587
2007E.C.EHEEE, 2005 E.C EGSECE & Grade9-12common subjects Average points.	596	42	554
2008E.C.EHEEE, 2006 E.C EGSECE & Grade9-12common subjects Average points.	714	18	696
2009E.C.EHEEE, 2007 E.C EGSECE & Grade 9-12common subjects Average points.	611	30	581

As we can see from the Table 3.1 above, about 193 students haven't got similar name during the integration due to change of name. Students who have not similar at all class levels, those

instances have removed from the dataset. The following figure 3.2. Shows the actual task during the integration using 32 relevant attributes out of 54 attributes. The number of instances cannot increase than the number of grade 12 students rather than the similar names are merged, but attributes are increase horizontally for grade 9-12.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Name	Father Name	GRAND FATHER NAME	Sex	Age	9E	9Ma	9PH	9CHe	9Bio	9Civ	10E	10Ma	10PH	10CHe	10Bio	10Civ	10E	10Ma	10PH	10CHe
2	DAWIT	TEFERA	KIDANEWOLD	M	18	93	95	90	97	98	91	89	94	87	91	85	89	A	A	A	A
3	TINSAE	MEKDIM	HAILU	F	18	61	60	62	56	60	66	64	50	45	63	55	58	B	C	B	B
4	SAMUEL	KENEA	GADISSA	M	19	65	56	60	54	61	62	53	49	48	52	50	60	B	C	C	B
5	MISMAK	MANDEFRO	HABTE	M	17	73	80	72	74	76	75	53	61	53	54	60	67	B	B	B	C
6	BEZAWIT	DEREJE	TILAHUN	F	18	68	66	71	74	71	79	55	58	53	62	69	65	B	C	B	C
7	MISALE	BETE	TESFAYE	M	18	65	61	55	45	49	68	62	62	62	59	63	66	A	B	B	A
8	REDEAT	DEREJE	TILAHUN	F	17	69	71	65	77	55	59	59	59	55	67	72	72	C	C	C	B
9	YOHANNES	DIRESSIE	ESHETU	M	18	94	98	99	98	100	99	84	89	83	94	97	84	A	A	A	A
10	SAMUEL	DAWIT	SOLOMON	M	19	84	80	80	83	92	92	63	71	69	76	87	73	C	B	C	B
11	HENOK	MAMUYE	WOUBIE	M	19	62	56	54	63	73	68	53	49	47	61	55	63	B	C	C	B
12	MISGANA	BEKELE	KETEMA	F	17	75	85	79	78	91	80	65	69	60	68	66	69	B	B	C	B
13	ABY	SIBHAT	FITAWOKE	M	18	93	98	100	89	99	80	79	85	82	88	86	80	B	A	B	A

Figure 3.2. The sample of actual task during integration using name of the students.

Full explanation of Table 3.1.is on the data description section 3.2.2.1.Due to large size of the datasets to put one after the other, the researcher put the sample of dataset for one year of one class which is grade 9, before the integration as follows.

The following Table 3.2. Shows the data set class label and their instances dissemination of performance for students in the five academic year in Debre Birhan City preparatory schools.

Table 3.2. Performance distribution for complete data of students in EHEEE result in each year**.**

	Performance of Students					
	Excellent	Very Good	Good	Satisfactory	Failure	Total
2005	17	83	250	243	2	595
2006	13	137	199	235	3	587
2007	17	86	220	226	5	554
2008	29	137	281	249	0	696
2009	13	133	201	234	0	581

3.2.2.1. Data Description

After the dataset of each Grade 9,10,11,12, and Grade10 National examination has integrated with grade12 EHEEE with the corresponding years result, it consists the total 3013 instances in the five years with 32 attributes out of 3206 instances and 54 attributes by removing the 193 instances which have many incomplete values and unmatched (unable to merged) names of the

students in each year and 22 irrelevant attributes from the dataset by discussing with domain experts is shown in Table 3.3. From these independent attributes 13 attributes contributed from National Educational Assessment and Examination Agency's (NEAEA) database, 39 attributes from the corresponding preparatory school database and 2 attributes (age and sex) from both of them are merged. For each selected attributes, their descriptions, data type and attribute values are represented by other values based on Assessment system of students 'academic performance for each grade level is illustrated in Table 3.3.The assessment system of students 'academic performance for each grade level are depicted in Table 3.5,3.6,3.7 and 3.8. Relevant attributes which were selected by discussing with domain experts are depicted in Table 3.3 below.

Table 3.3. Attributes name, description, data type and values.

Attributes Name	Description of attributes	Data Type	Attributes values
Sex	Sex of the student	Nominal	{Male ,Female}
Age	Age of the student when he/she takes the exam of EHEEE.	Numeric	{16-27 }
CICAPG11	Civics class average point grade 11	Nominal	{Excellent,Verygood ,Satisfactory, Fair, poor}
MCAPG10	Maths class average point grade 10	Nominal	{Excellent ,Very good ,Satisfactory, Fair, poor}
ECAPG11	English class average point grade 11	Nominal	{Excellent ,Very good ,Satisfactory, Fair, poor}
CICAPG12	Civics class average point grade 12	Nominal	{Excellent ,Very good ,Satisfactory, Fair, poor}
CHCAPG11	Chemistry Class Average Point grade 11	Nominal	{Excellent ,Very good ,Satisfactory, Fair, poor}
PCAPG12	physics class average point grade 12	Nominal	{Excellent ,Very good ,Satisfactory, Fair, poor}
ECAPG9	English class average point grade 9	Nominal	{Excellent ,Very good ,Satisfactory, Fair, poor}
PCAPG10	physics class average point grade 10	Nominal	{Excellent ,Very good ,Satisfactory, Fair, poor}
ENEGG10	English National Exam 'grade point' in grade 10	Nominal	{Excellent ,Very good ,good, Satisfactory, Fail, }
MNEGG10	Maths National Exam 'Grade Point' in grade 10	Nominal	{Excellent ,Very good ,good, Satisfactory, Fail, }
ECAPG10	English class average point grade 10	Nominal	{Excellent ,Very good ,Satisfactory, Fair, poor}

Table 3.3 Continued…

Table 3.3 to be continued…

CHNEGG10	Chemistry National Exam ‘grade point’ in grade 10	Nominal	{Excellent ,Very good ,good, Satisfactory, Fail, }
MCAPG11	Maths class average point grade 11	Nominal	{ Excellent ,Very good , Satisfactory, Fair, poor}
CHCAPG10	Chemistry class average point grade 10	Nominal	{ Excellent ,Very good , Satisfactory, Fair, poor}
BCAPG9	Biology class average point grade 9	Nominal	{ Excellent ,Very good , Satisfactory, Fair, poor}
MCAPG12	Maths class average point grade 12	Nominal	{ Excellent ,Very good , Satisfactory, Fair, poor}
CIAPG10	Civics class average point grade 10	Nominal	{ Excellent ,Very good , Satisfactory, Fair, poor}
BNEGG10	Biology National Exam ‘grade point’ in grade 10	Nominal	{Excellent ,Very good ,good, Satisfactory, Fail, }
PCAPG11	physics class average point grade 11	Nominal	{ Excellent ,Very good , Satisfactory, Fair, poor}
PNEGG10	physics National Exam ‘grade point’ in grade 10	Nominal	{Excellent ,Very good ,good, Satisfactory, Fail, }
CHCAPG12	Chemistry class average point grade 12	Nominal	{Excellent ,Very good , Satisfactory, Fair, poor}
BCAPG11	Biology class average point grade 11	Nominal	{Excellent ,Very good , Satisfactory, Fair, poor}
BCAPG10	Biology class average point grade 10	Nominal	{Excellent ,Very good , Satisfactory, Fair, poor}
MCAPG9	Maths class average point grade 9	Nominal	{Excellent ,Very good , Satisfactory, Fair, poor}
PCAPG9	physics class average point grade 9	Nominal	Nominal {Excellent , Very good ,Satisfactory, Fair, poor}
ECAPG12	English class average point grade 12	Nominal	{Excellent ,Very good , Satisfactory, Fair, poor}
CINEGG10	Civics National Exam ‘grade point’ in grade 10	Nominal	{Excellent ,Very good ,good, Satisfactory, Fail, }
CHCAPG9	Chemistry class average point grade 9	Nominal	{Excellent ,Very good , Satisfactory, Fair, poor}
BCAPG12	Biology class average point grade 12	Nominal	{Excellent ,Very good , Satisfactory, Fair, poor}
CICAPG9	Civics class average point grade 9	Nominal	{Excellent ,Very good , Satisfactory, Fair, poor}

3.2.3. Data Preprocessing

Before applying Data mining algorithms to develop model, it is necessary to carry out some pre-processing on the dataset. The preprocessing task performed in this dataset by data cleaning, handling attribute missing value, data discretization, data transformation, data format. Since, real world data usually become noisy due too many cases such as its larger size, its origin from multiple and different sources, the targeted data need to be cleaned to get quality data and quality decisions [66]. Thus, preprocessing tasks should be performed in the dataset in order to get the targeted data set for quality work [67]. This step helps to: reduce confusion during learning, and to get better

input data for data mining techniques [68]. Attributes missing values in the dataset is most likely associated with unavailability of interesting information, lack of knowhow on the importance of data at the time of entry, misunderstanding of the data, the respondents him/her self may refuse to answer certain questions or they may not know the answer exactly or may answer in an unexpected manner [69].Missing data is the most common problem that comes up during the data analysis process. As it is suggested by Kantardzic [69], it is possible to ignore the attributes with too many missing values or more than 50% missing values. Avoiding the missing data is not time consuming and at the same time it is very easy to follow. Deleting records may result in losing some information [69].If the sample data size is large avoiding some records or attributes may not affect the results, but still we need to keep in mind we are losing something important if we ignore attributes with too many missing values without considering all aspects of the problem under study. Data Pre-processing stage was performed to improve the quality of the data set by removing the missing or incomplete values in the dataset. To do so all the dataset which is available in the Debre Birhan City preparatory schools in 2005 E.C -2009 E.C five academic years was cleaned to be the same format. As a result, the data was prepared for knowledge discovery. The researcher used the MS-Excel application to converting all nominal data type into numeric value in order to make it convenient for MATLAB matrix software to handle missing value using attribute mean. The dataset before handle missing value is found in figure.3.2.

3.2.3.1. Data Cleaning

Data cleaning involve filling missing values, smooth noisy data, identify or remove outliers etc. Thus, even though the researcher have 3013 correctly integrated instances with corresponding data sets and selected 32 attributes still there are attribute values which needs be cleaned (filling

missing value). In this study, the dataset of Ethiopian Higher Education Entrance Examination (EHEEE) contains 95 ‘age’ attribute missing values and 37 noise values like 00, 01, 02, 03 … 09. However, the researcher decided to fill the missing values and to remove noise values. Because ,in Ethiopia the primary education starts at the age of 5 or 6.In this case Grade 10 attained mostly at the age of 15 and above in cases and Grade 12 students attained mostly at the age of 19 and above in cases [1]. Thus, the age attribute with missing values in the dataset of 2005 E .C to 2009 E.C should be filled with mean values by using attribute mean using MATLAB software matrix.

The other noise attributes values have removed from the dataset which have emerged from incomplete data which is lacking attribute values; this is caused by data entry problems. Similarly there are attribute values which shows N/A which has meaning not available. This is caused by a student not taking the examination of a subject. The other noise is due to the attribute value absent which is mentioned as 'ab' in the data. This is the case when students do not take some subjects. In similar manner there is attribute values disqualify which is mentioned 'dis' in the data .This is the case where students face irregularities on exams. Zero value - this value occurs when an absent student answer sheet marked by a pencil and this answer sheet scanned with other correct answer sheets. Thus, to calculate the mean value, these noisy data have removed from the dataset.

To deal with missing values, alternatives have been suggested by T. Larose [57] and Chakrabarti et al [70]: This are: ignore the missing value, replace the missing value manually, replace the missing value with a global constant to fill in the missing value, replace the missing value with some constant, specified by the analyst, replace the missing value with the field mean (for numerical variables) or the mode (for categorical variables),replace the missing values with a value generated at random from the variable distribution observed. If missed value of attributes is 50% and above, they has to be removed (Ignored) [21], [57], [70] since it doesn't consider a large amount of data. Thus based on the percentage of missed value of attributes, the researcher used ignore the missing values and replacing the missing value with the field mean (for numerical variables) or the mode (for categorical variables) options. After the integration of instances in the data set, Grade 10 Ethiopian Grade A, B, C, D, and F have values 4, 3, 2, 1 and 0 respectively [1]. Thus the researcher substitutes the letters by numeric values by using excel

filter application in order to make the data set suitable for MAT LAB and preprocessing of the dataset as shown below in figure 3.2.

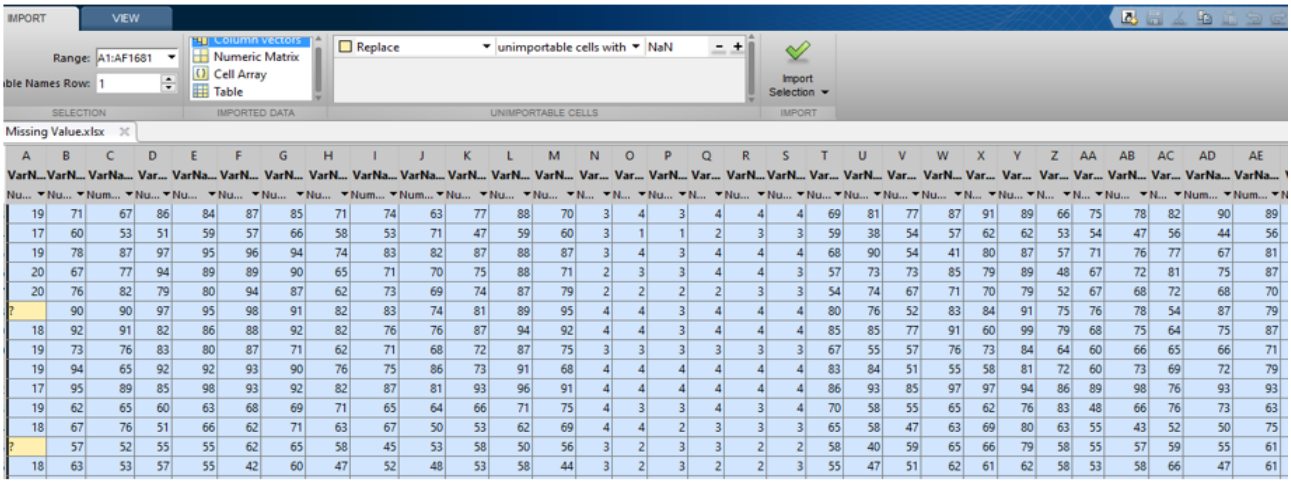


Figure 3.3. Dataset before handle missing value.

3.2.3.2. Handling Missing Values

In this research work the collected dataset has 95 missed values of age attribute among attributes namely, Sex, English, maths, Physics, Chemistry, Biology and Civic and ethical education. Thus, even if it is possible to handling those age missing values, the recommended technique is filling with attribute mean value using MATLAB matrix; this method has been used to replace missed value of 'age' attribute in order to get quality dataset for quality work. There are steps to handle missing value by attribute mean using MATLAB matrix software.

Step1.Represent the missing values by the word NAN

Step2. Calculate the mean values of each feature by ignoring NAN values

Step3. Replace the NAN value with the corresponding mean value of the feature

After the researcher has used the above steps to handle missing values of age attribute using MATLAB, the following mean value results shown in table 3.3 has found. Next, the attribute has been discretized and transformed using different methods used as standards to make suitable for mining process.The following figure 3.3 shows the dataset after handled missing value using attribute mean.

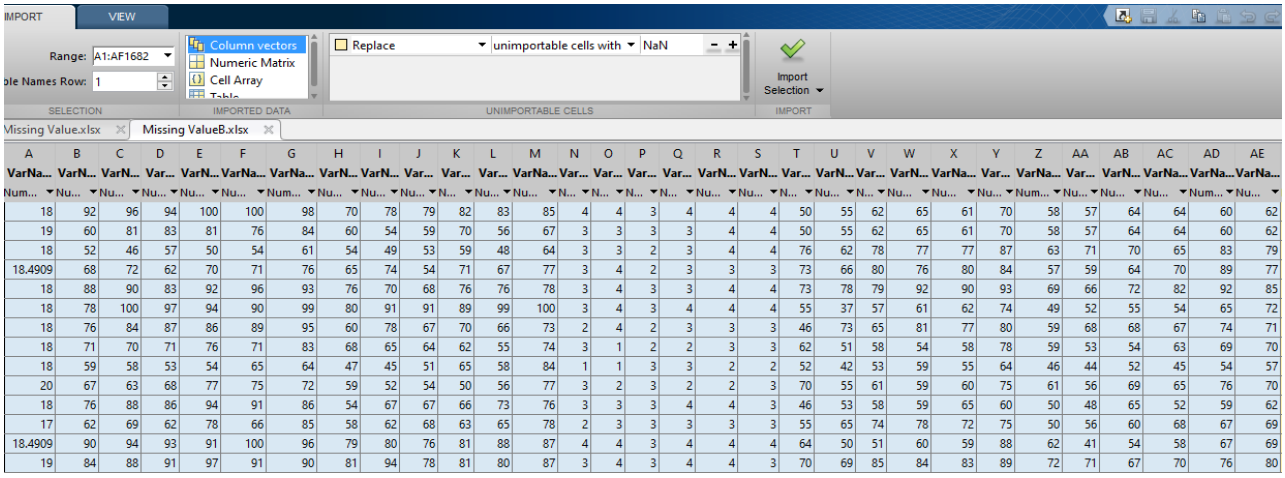


Figure 3.4. A dataset after handled missing value using attribute mean.

3.2.3.3. Data Discretization

Data discretization techniques applied in order to get the reduced representation of the data set so as which is useful in mining efficiently and get the same analytical result of the original data [31]. In this study the researcher is used discretization (binned) mechanism to reduce data representation. Discretization is a concept in which raw data values for attributes are replaced by ranges or higher conceptual levels [71].In this study the subject attributes and an age attribute discretized since it has a continuous value and used for efficient later processing, simplified data description and understanding for data mining results.

The continuous width method divides the data in to a fixed number of intervals of equal or almost equal length. The following Table 3.4 illustrates value an attribute age discretized using equal width portioning discretization method [52].

The Interval of the age Attribute is calculated as Shown Below:

$W=\frac{(\max-\min)}{k}$, where w is the width of interval/between each and k is the number of interval

Max is the maximum value of the age and min is the minimum value of the age. Thus in the dataset the max age is 27 and the minimum age is 16 years old. So that, $W=\frac{(27-16)}{4}$ =2.75 and the interval boundaries are: min+w, min+2w…min+ (k-1) w. The number of interval (K) is based on the interval between the largest age and the minimum age in the dataset. If the age boundaries is more near to the minimum age , it is the more correct interval for partitioning the data into equal length.so the boundaries are 16+2.75, 16+2(2.75) and 16+3(2.75): 18.75, 21.5 and 24.25 [52].

Table 3.4. Age discretization

Attribute Name	Original value	Discretized value	Transformed value
Age	16 – 27	16 – 18	Age1 (category one)
		19 – 21	Age 2 (category two)
		22– 24	Age 3 (category three)
		25 –27	Age 4 (category four)

3.2.3.4. Data Transformation

As per Han et al [72]. Data transformation is about transforming or consolidating the data to make it appropriate for mining. Data transformation can involve the following operations: Smoothing; this techniques include binning, regression, and clustering which works to remove noise from the data. Generalization of the data; where low-level raw data are replaced by higher-level concepts. For example numeric attribute of age may be generalized to youth, middle age, and old age. Normalization; this operations performed on attribute to scale the value to fall within a small specified range. Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute in to equal intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels there by reduces and simplifies the original data. This leads to a concise, easy to use, knowledge-level representation of mining results.

Therefore, the researcher performed discretization on the attributes by using a binning method. The attribute "Age" is binned into four levels (Age1, Age2, Age3 and Age4) ,the attribute 'EGSECE' natural science common subjects scores" is binned to (Excellent, very good, Good ,satisfactory and Fail)," " and "Grade 9-12 natural science stream common subjects Grade average point" is binned into five levels (Excellent, very good, satisfactory ,Fair and poor) Grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) natural science stream common subjects total score" is binned into five levels (Excellent, very Good, Good ,satisfactory and Fail) as class by using the Assessment System of secondary and preparatory education in Ethiopia by Ministry of Education (MOE). The other attributes which discretized are the subjects in EGSECE, Ethiopian Higher Education Entrance Examination (EHEEE) that shown below in

3.5, 3.6 and 3.7. The subjects are continuous value and discretized using the assessment system of secondary and preparatory education in Ethiopia [1].

Table 3.5. Assessment system of secondary education subjects in EGSECE

In percentages	Letter Grade	Number	Meaning(performance)
90 – 100	A	4	Excellent
80 – 89	B	3	Very good
70 – 79	C	2	Good
50 – 69	D	1	Satisfactory
Under 50	F	0	Fail

In the Table 3.8. Above the assessment system of Secondary education in Ethiopia but Physics subject discretized using equal width partitioning due the subject problematic condition than the other subjects. The width of the interval will be W =$\frac{(B-A)}{N}$, where W – width, B – the maximum value, A – the minimum value, N – number of interval. Therefore W =$\frac{(90-3)}{5}$ =18.

Table 3.6. Assessment system of secondary school EGSECE for physics value.

In percentages	Letter Grade	Numeric	Meaning(performance)
81 – 100	A	4	Excellent
62 – 80	B	3	Very good
43 – 61	C	2	Good
24 – 42	D	1	Satisfactory
< = 23	F	0	Fail

Table 3.7. The assessment system of Preparatory EHEEE in Ethiopia

In percentages	Total result	Performance
75 – 100	527-700	Excellent
63– 75	440-527	Very good
47 –63	352-440	Good
25– 47	177-352	Satisfactory
Under 25	Under 177	Fail

Table 3.8. Assessment system of students 'performance in Grade1-12 in Ethiopia

In percentages	Meaning(performance)
90 – 100	Excellent
80– 89	Very good
60 –79	Satisfactory
50–59	Fair
Under 50	poor

Table 3.9. Nominal attribute values after integration including class lable.

Sex	Age	ECAPG9	MCAPG9	PCAPG9	CHCAPG9	BCAPG9	CICAPG9	ECAPG10	MCAPG10	PCAPG10
M	AGE-1	VGOOD	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	SATISFACTORY	SATISFACTORY	VGOOD
F	AGE-1	VGOOD	SATISFACTORY	SATISFACTORY	SATISFACTORY	VGOOD	VGOOD	SATISFACTORY	FAIR	FAIR
F	AGE-1	FAIR	VGOOD	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	SATISFACTORY	FAIR	EXCELLENT
M	AGE-1	VGOOD	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT
F	AGE-1	SATISFACTORY	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	VGOOD	VGOOD
M	AGE-1	SATISFACTORY	SATISFACTORY	VGOOD	EXCELLENT	VGOOD	EXCELLENT	FAIR	VGOOD	FAIR
M	AGE-1	SATISFACTORY	SATISFACTORY	SATISFACTORY	VGOOD	VGOOD	SATISFACTORY	FAIR	POOR	SATISFACTORY
M	AGE-1	FAIR	SATISFACTORY	SATISFACTORY	SATISFACTORY	POOR	SATISFACTORY	FAIR	FAIR	POOR
M	AGE-1	FAIR	FAIR	SATISFACTORY	FAIR	SATISFACTORY	SATISFACTORY	FAIR	SATISFACTORY	POOR

Table 3.9 Continued

CHCAPG10	BCAPG10	CICAPG10	ENEGG10	MNEGG10	PNEGG10	CHNEGG10	BNEGG10	CINEGG10	ECAPG11	MCAPG11
EXCELLENT	VGOOD	SATISFACTORY	EXCELLENT	EXCELLENT	VGOOD	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	POOR
SATISFACTORY	SATISFACTORY	SATISFACTORY	EXCELLENT	VGOOD	VGOOD	VGOOD	EXCELLENT	VGOOD	VGOOD	SATISFACTORY
EXCELLENT	SATISFACTORY	EXCELLENT	VGOOD	EXCELLENT	VGOOD	EXCELLENT	EXCELLENT	EXCELLENT	SATISFACTORY	SATISFACTORY
EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	SATISFACTORY
EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	EXCELLENT	VGOOD	VGOOD
SATISFACTORY	SATISFACTORY	FAIR	GOOD	VGOOD	VGOOD	VGOOD	VGOOD	VGOOD	VGOOD	VGOOD
SATISFACTORY	SATISFACTORY	SATISFACTORY	VGOOD	GOOD	GOOD	EXCELLENT	EXCELLENT	EXCELLENT	SATISFACTORY	SATISFACTORY
FAIR	FAIR	SATISFACTORY	GOOD	GOOD	VGOOD	GOOD	VGOOD	VGOOD	FAIR	FAIR
FAIR	FAIR	SATISFACTORY	EXCELLENT	EXCELLENT	VGOOD	EXCELLENT	EXCELLENT	VGOOD	SATISFACTORY	POOR
SATISFACTORY	FAIR	SATISFACTORY	SATISFACTORY	GOOD	GOOD	VGOOD	VGOOD	GOOD	FAIR	SATISFACTORY

Table 3.9 Continued

Table 3.9 Continued

PCAPG11	CHCAPG11	BCAPG11	CICAPG11	ECAPG12	MCAPG12	PCAPG12	CHCAPG12	BCAPG12	CICAPG12	CLASS
SATISFACTORY	VGOOD	VGOOD	VGOOD	POOR	FAIR	SATISFACTORY	SATISFACTORY	FAIR	EXCELLENT	VGOOD
FAIR	SATISFACTORY	SATISFACTORY	VGOOD	SATISFACTORY	POOR	FAIR	SATISFACTORY	SATISFACTORY	VGOOD	GOOD
VGOOD	SATISFACTORY	VGOOD	FAIR	FAIR	FAIR	FAIR	SATISFACTORY	SATISFACTORY	SATISFACTORY	SATISFACTORY
SATISFACTORY	VGOOD	VGOOD	EXCELLENT	SATISFACTORY	SATISFACTORY	EXCELLENT	FAIR	SATISFACTORY	EXCELLENT	EXCELLENT
SATISFACTORY	EXCELLENT	VGOOD	SATISFACTORY	VGOOD	VGOOD	SATISFACTORY	SATISFACTORY	VGOOD	EXCELLENT	VGOOD
VGOOD	EXCELLENT	VGOOD	SATISFACTORY	SATISFACTORY	SATISFACTORY	SATISFACTORY	SATISFACTORY	VGOOD	EXCELLENT	GOOD
SATISFACTORY	SATISFACTORY	SATISFACTORY	SATISFACTORY	POOR	SATISFACTORY	FAIR	SATISFACTORY	SATISFACTORY	VGOOD	SATISFACTORY
SATISFACTORY	VGOOD	FAIR	FAIR	POOR	POOR	FAIR	FAIR	FAIR	VGOOD	SATISFACTORY
SATISFACTORY	VGOOD	FAIR	FAIR	POOR	FAIR	FAIR	SATISFACTORY	SATISFACTORY	VGOOD	GOOD
SATISFACTORY	SATISFACTORY	SATISFACTORY	SATISFACTORY	FAIR	SATISFACTORY	FAIR	FAIR	SATISFACTORY	FAIR	FAIL

3.2.3.5. Data Format Conversion for WEKA software

In this study, the preprocessed dataset is used MATLAB and Microsoft Excel for identified and took necessary measure to handle missing values, to replace numeric in to nominal and experimented using WEKA tool for the purpose of extracting classification rule. The final integrated dataset changed in to the format and data type which is suitable for WEKA software should be prepared. WEKA data mining tool accept ARFF (Attribute-Relation File Format) file formats. This format contained attribute values whose values are separated by comma. The file extension for the file format ARFF is ".arff"[16].Initially the data collected from the original hard copy exported to an excel file format and the preprocessed technique have been applied. Then the excel file format converted to comma separated value (CSV) file format. Next the CSV file format sample opened with WEKA and then saved with .arff file extension suitable for mining using WEKA 3.6.13 is shown in figure.3.4. The following figure 3.4. Illustrates the sample of the changed ARFF format. After data format conversion it should be balanced if there is unbalanced instances in the dataset [73].

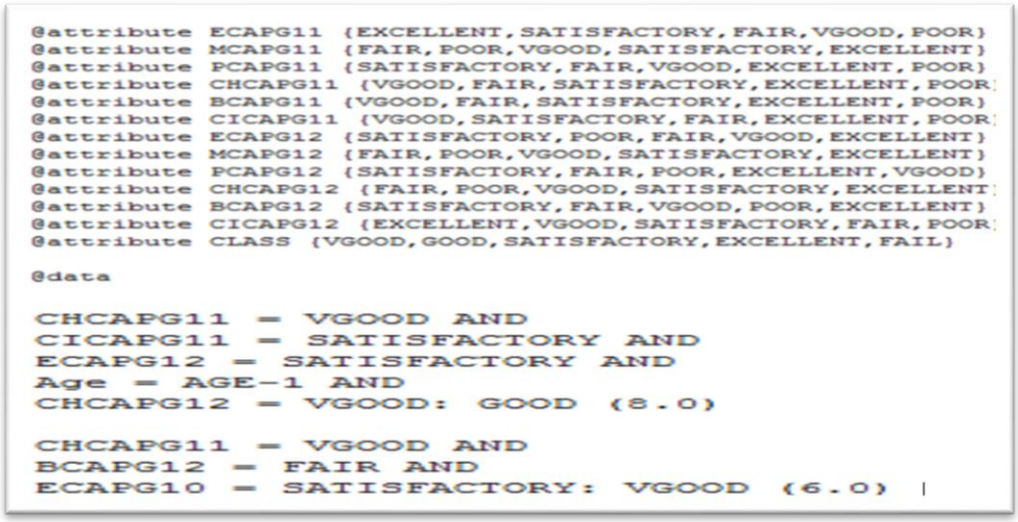


Figure 3.5. Sample of Machine understandable ARFF format dataset in WEKA.

3.2.4. Learning Process (Classification)

Here the data minor uses various data mining methods to derive knowledge from preprocessed data [14]. At this stage, a particular data mining method that matches the goal of the data mining process defined in the first step is selected. However, the detail of developing and training the model vary from technique to technique and hence, there are no blueprint procedures. During the modeling stages for this study, we use classification algorithms grouped in different classifying methods, including a common decision tree algorithms JRIP, PART, J48 and REPTree. These models algorithms were selected in this research due to their popularity in the recently published documents. WEKA data mining tools have used to implement the KDP. WEKA formally called Waikato Environment for Knowledge Learning developed at the University of Waikato in New Zealand, is open-source data mining software in java. WEKA provides implementations of learning algorithms that can be applied to a given dataset and analyze its output to learn more about the data, and use learned models to generate predictions on new instances [74].These four classifiers are discussed below.

3.2.4.1. Classification Using JRIP

JRIP is a prepositional rule learner, i.e. Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [46]. Rules in this algorithm are generated for every class in the training set and are then pruned. The discovered knowledge is represented in the form of IF-THEN prediction rules, which have the advantage of being a high-level and symbolic knowledge

representation contributing towards the comprehensibility of the discovered knowledge. JRIP is based on the construction of a rule set in which all positive instances are covered by partitioning the current set of training instances into two subsets namely a growing set and a pruning set. The rule is constructed from instances in the growing set. Initially, the rule set is empty and the rules are added incrementally to the rule set until no symptom instances are covered. Following this the algorithm substitutes or revises individual rules by using reduced error pruning in order to increase the accuracy of the rules. To prune a rule the algorithm takes into account only a final sequence of conditions of the rule and sorts the deletion that maximizes the function [46].

3.2.4.2. Classification Using J48

C4.5 is an evolution of ID3, presented by Quinlan J.R [41]. It uses gain ratio as splitting criteria. The splitting ceases when the number of instances to be split is below a certain threshold. C4.5 can handle numeric attributes. J48 is an implementation of Quilan algorithm (C4.5). J48 classifier develops a decision tree for the given dataset, whose nodes represent discrimination rules acting on selective features by recursive partitioning of data using depth- first strategy. The algorithm used each attribute of the data to make decision by splitting the data into smaller subjects. All the possible tests are considered during decision making based on information gain value of each attribute as stated in equation 2.1 [41].For an event with probability *p*, the average amount of information in bits required to transmit the result is -log2*p*. For variables with several outcomes, we simply use a weighted sum of the log2*pi'*s, with weights equal to the outcome probabilities. Therefore, the mean information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, explain as equation 2.1.

At each decision node, C4.5 uses the attribute with the maximum gain ratio as the splitting attribute and recursively visits each decision node, selecting the optimal split, until no further splits are possible. J48 also used the same concept to construct the decision tree and it supports both numeric and nominal predictors and nominal class attribute [42]. Once the tree is constructed, it is possible to generate the rule in order to apply it for new instances which are independent of the training tuples. Figure 2.5 illustrates the decision tree constructed from which one can easily generate the decision rule.

The J48 algorithm gives several options related to tree pruning. Many algorithms attempt to "prune", or simplify, their results. Pruning produces fewer, more easily interpretable results. More importantly, pruning can be used as a tool to correct for potential over fitting. The basic

algorithm described above recursively classifies until each leaf is pure, meaning that the data has been categorized as perfectly as possible [41]. This process ensures maximum accuracy on the training data, but it may create excessive rules that only describe particular peculiarities of that data. When tested on new data, the rules may be less effective. Pruning always reduces the accuracy of a model on training data.

3.2.4.3. Classification Using REPTree

REP Tree: Reduces Error Pruning (REP) Tree where Classifier is a fast type of decision tree learner which is built for the decision tree or for the regression tree by using the information gain with entropy and as in C4.5 Algorithm, which deals with the missing values by breaking the corresponding instances into pieces [43].

3.2.4.5. Classification Using PART

According to [46], many learning techniques look for structural descriptions of what is learned, descriptions that can become fairly complex and are typically expressed as sets of rules. Because they can be understood by people, these descriptions serve to explain what has been learned and explain the basis for new predictions. Classifications rules are a popular alternative to decision trees in representing the structures that learning methods produce. The antecedent, or precondition, of a rule is a series of tests just like the tests at nodes in decision trees, and the consequent, or conclusion, gives the class or classes that apply to instances covered by that rule, or perhaps gives a probability distribution over the classes. Generally, the preconditions are logically joined together by "AND", and all the tests must succeed if the rule is to work. It is also easy to read a set of rules directly off a decision tree. One rule is generated for each leaf. The antecedent of the rule includes a condition for every node on the path from the root to that leaf, and the consequent of the rule is the class assigned by the leaf. One reason why rules are popular is that each rule seems to represent an independent “nugget” of knowledge. PART is a class for generating decision list in WEKA. It builds a partial C4.5 decision tree in each iteration and makes the best leaf into a rule. Rules or decision lists which are generated using PART algorithm are more clear and understandable. As a result the researcher has also used this algorithm for modeling.

3.2.5. Performance Evaluation

The pre-processed dataset is divided in to training and testing set .The training set is used for learning process (classification) to develop classifier model by applying the techniques such as

J48, PART, JRIP and REP Tree .The performance of developed classifier has to be evaluated by using the testing set. The predicted class of the test set has to be checked with the corresponded class label on the comparison result is counted and recorded in the confusion matrix with the means of Ture positive, False positive, True negative .The performance metrics used to evaluate the performance of each classifier are accuracy, precision, recall, F-measure, TP rate, FP rate [54, 75].

The classified model performs the highest performance will be selected and it will be used as knowledge base to develop the prototype. This will done and discussed detail in the next on chapter four

UNIT FOUR

EXPERMENTAL ANALYSIS AND RESULT

4.1. Introduction

This chapter discuss about the experimentation phase applied on the preprocessed dataset using with the selected classification techniques. The chapter has consisting of experimental setup, experimental analysis, experimental result, and comparison of the result. The experimental part describes how the preprocessed dataset is partitioned for training and testing purpose. Beside, this tells us the parameters values for an experiments of the selected classifier and finally, the model with the best performance is selected.

4.2. Balancing Instances

The class imbalance problem is prevalent in many applications, including: fraud/intrusion detection, risk management, text classification, and medical diagnosis/monitoring, etc. It typically occurs when, in a classification problem, there are many more instances of some classes than others. In such cases, standard classifiers tend to be overcome by the large classes and ignore the small ones. Particularly, they tend to produce high predictive accuracy over the majority class, but poor predictive accuracy over the minority class [73].A number of solutions to the class-imbalance problem were proposed both at the data and algorithmic levels. At the data level, these solutions include many different forms of re-sampling such as over-sampling and under-sampling. Synthetic Minority Over-sampling Technique (SMOTE) is an over-sampling approach which generates synthetic examples in a less application specific manner. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors [86].WEKA 3.6.13 have a functionality of ‘SMOTE’ to resample and balance the number of class labels for those attributes used as target class in model building process. The figure 4.1 Illustrates imbalance dataset before SMOTE is shown below.

No.	Label	Count
1	VGOODI	576
2	GOOD	1151
3	SATISFACTORY	1187
4	EXCELLENT	89
5	FAIL	10

Class: CLASS (Nom) Visualize All

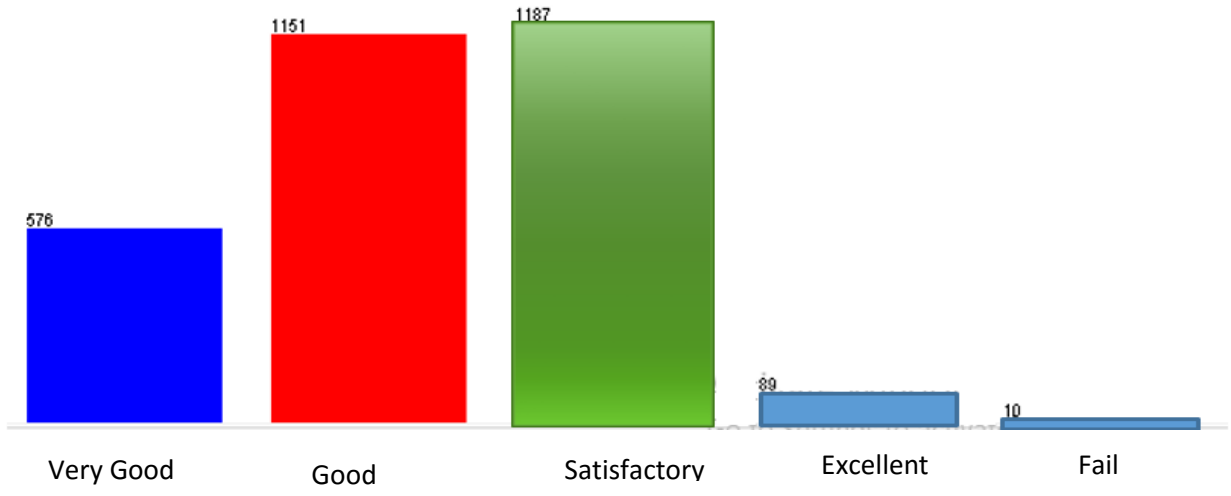
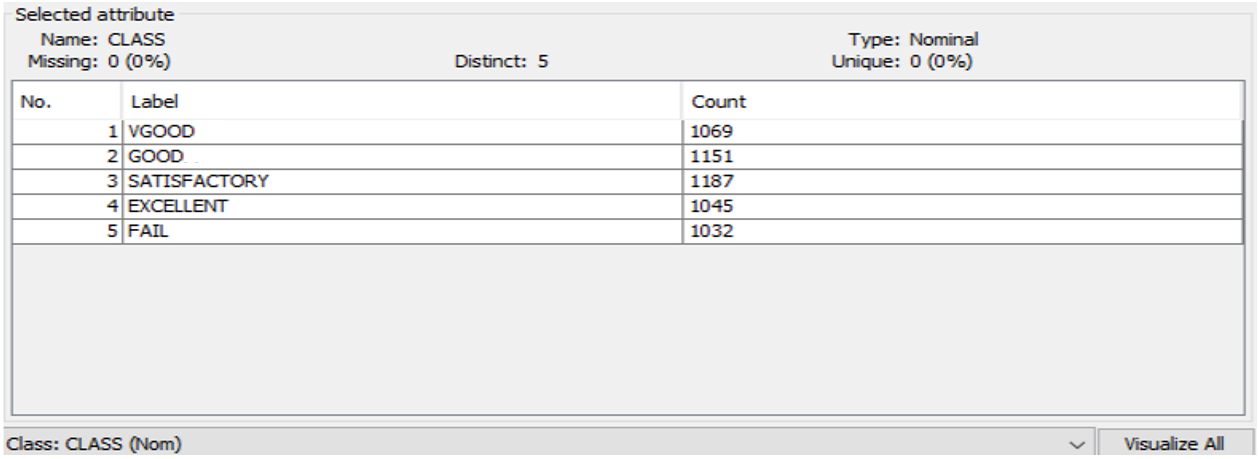


Figure 4.1. Imbalance dataset before SMOTE.

As we can observe in figure 4.1 above, the classes are imbalance and to avoid this, the researcher SMOTE the data set and discretize it before conducting the experiments and as a result the dataset increases from 3013 records to 5484 records.



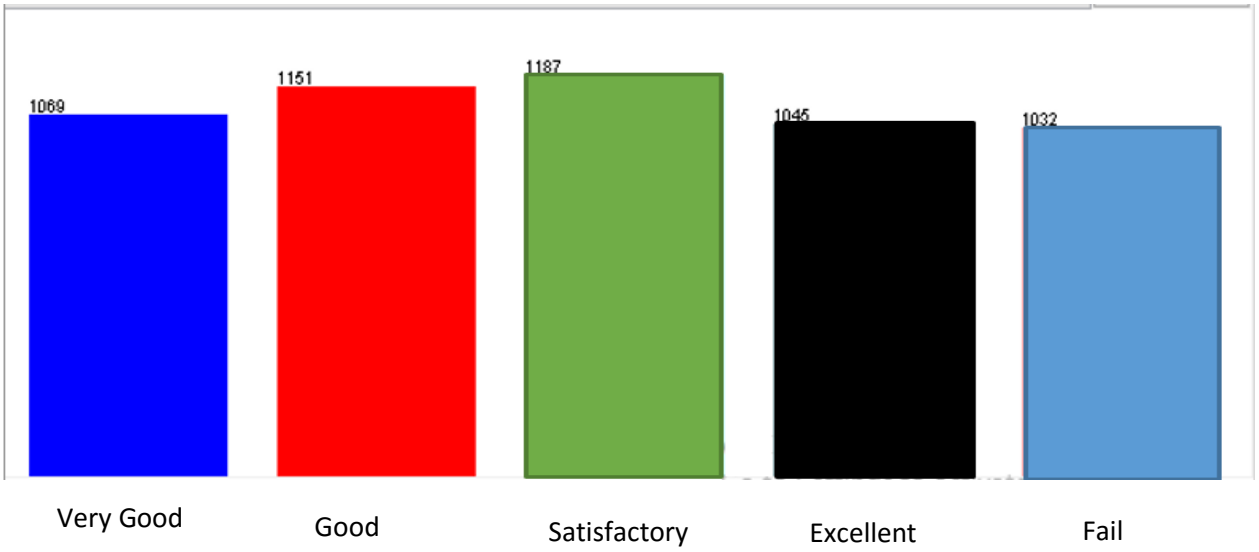


Figure. 4.2. Balanced Dataset after SMOTE

Table 4.1. Dataset class label and the number instances

Class lable	Number of instances Before SMOTE	Number Of Instances After SMOTE
Excellent	89	1045
Very good	576	1069
Good	1151	1151
Satisfactory	1187	1187
Fail	10	1032

4.3. Attribute Selection

The attribute selection process is performed using the ranker selection attribute values by using info gain ratio attribute selection measure to select the most relevant attributes, whole attributes .Out of 32 ranked attributes using info gain ratio, 18 attributes were selected based on relevance value and decided the thresh hold 0.33. The following Table 4.2 depicts the 32 attribute selection ranker using info gain ratio from largest to lowest.

Table 4.2.The Experiment of the attribute selected ranker.

Rank	Attributes	Info gain value	Rank	Attributes	Info gain value	Rank	Attributes	Info gain value
1	CICAPG11	1.33	12	CHNEGG10	0.39	23	BCAPG10	0.28
2	MCAPG10	1.20	13	MCAPG11	0.38	24	MCAPG9	0.27
3	ECAPG11	1.09	14	CHCAPG10	0.37	25	PCAPG9	0.26
4	CICAPG12	0.90	15	BCAPG9	0.36	26	ECAPG12	0.25
5	CHCAPG11	0.88	16	MCAPG12	0.35	27	CINEGG10	0.24
6	PCAPG12	0.81	17	CICAPG10	0.34	28	CHCAPG9	0.23
7	ECAPG9	0.79	18	BNEGG10	0.33	29	BCAPG12	0.22
8	PCAPG10	0.63	19	PCAPG11	0.32	30	CICAPG9	0.19
9	ENEGG10	0.55	20	PNEGG10	0.31	31	Sex	0.18
10	MNEGG10	0.50	21	CHCAPG12	0.30	32	Age	0.05
11	ECAPG10	0.49	22	BCAPG11	0.29			

Next, the researcher decide to perform subsequent experiments based on both whole 32 attributes in Table 4.2 and 18 selected attributes which are shown below in Table 4.X .The Experiment of the selected attribute with thresh hold 0.33.

Table 4.3.The Experiment of the selected attributes

Rank	Attributes	Info gain value	Rank	Attributes	Info gain value
1	CICAPG11	1.33	12	CHNEGG10	0.39
2	MCAPG10	1.20	13	MCAPG11	0.38
3	ECAPG11	1.09	14	CHCAPG10	0.37
4	CICAPG12	0.90	15	BCAPG9	0.36
5	CHCAPG11	0.88	16	MCAPG12	0.35
6	PCAPG12	0.81	17	CICAPG10	0.34
7	ECAPG9	0.79	18	BNEGG10	**0.33**
8	PCAPG10	0.63			
9	ENEGG10	0.55			
10	MNEGG10	0.50			
11	ECAPG10	0.49			

4.4. Experimental Setup

As mentioned in data preprocessing the integrated dataset has to pass through the data preprocess task to clean dataset. The missing value has been imputed with the mean value of corresponding attribute by applying the steps mentioned on 3.2.3.2. (Handling missing values). In any data mining research before developing a model, we should generate a mechanism to test the model performance. For instance, in the supervised data mining task, such as classification, it is common to use classification accuracy measure, True Positive rate (TP), precision, recall and F-measure of the experts are used as to measure the performance of the developed data mining model. Each algorithms was conducted in different ways using attributes (whole and selected attributes) based on info gain ratio values in WEKA tool with balanced dataset. Thus, the whole attributes and selected attributes have experimented with the four classifies (J48, REPTree, JRip and PART) in the manner shown in Table 4.3.

Table 4.4. Experimentation ways of whole attributes and selected attributes

No	Classifier techniques	Data partitioning	Parameter	Remark
1	J48	10 cross validation Test	whole features	Experment I
			selected features	Experment II
		80/20 Training/Testing	whole features	Experment III
			selected features	Experment IV
2	JRIP	10 cross validation Test Training/Testing	whole features	Experment V
			selected features	Experment VI
		80/20 Training/Testing	whole features	ExpermentVII
			selected features	ExpermentVIII
3	REPTree	10 cross validation Test	whole features	Experment IX
			selected features	Experment X
		80/20 Training/Testing	whole features	ExpermentXI
			selected features	Experment XII
4	PART	10 fold cross validation Test Training/Testing	whole features	Experment XIII
			selected features	Experment XIV
		80/20 Training/Testing	whole features	Experment XV
			selected features	Experment XVI

4.5. Experimental Results

Totally, 16 Experimentation have been performed on the preprocessed dataset. Each of this experiments and the result explained below.

4.5.1. Experiment I: Using J48 with 10-fold cross-validation applied on whole features**.**

This experiment conducts under 10-fold cross-validation test option with default parameters of WEKA and the algorithm generates a model as J48 .As it shown below, out of the 5484 records of the dataset 5185(80%) records are correctly classified and the model has an accuracy of 94.55%, 299(5.45%) incorrectly Classified Instances .

Table 4.5.Confusion Matrix of J48 Algorithm

	Predicted class					
Actual class	Very good	Good	Satisfactory	Excellent	Fail	
	998	58	0	13	0	Very good
	75	**999**	77	0	0	Good
	0	62	**1123**	1	1	Satisfactory
	12	0	0	**1033**	0	Excellent
	0	0	0	0	**1032**	Fail

Table 4.6. Detail Analysis Result of J48 Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.934	0.02	0.92	0.934	0.927	Very good
	0.868	0.028	0.893	0.868	0.88	Good
	0.946	0.018	0.936	0.946	0.941	Satisfactory
	0.989	0.003	0.987	0.989	0.988	Excellent
	1	0	0.999	1	1	Fail
Weighted Avg.	0.945	0.014	0.945	0.945	0.945	

4.5.2. Experiment II: Using J48 with 10-fold cross-validation applied on the selected features**.**

This experiment conducts under 10-fold cross-validation test option with selected attributes default parameters of WEKA and the algorithm generates a model as J48 and Correctly Classified Instances are 5184 which means 94.53 %, 300 (5.47%) incorrectly classified instances and taking 0.13 seconds to build the model.

Table 4.7. Confusion Matrix of J48 Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual Class	Very good	**998**	62	0	9	0
	Good	68	**999**	84	0	0
	Satisfactory	0	64	**1122**	0	1
	Excellent	12	0	0	**1033**	0
	Fail	0	0	0	0	**1032**

Table 4.8. Detail analysis result of J48 algorithm

Classifier	TP Rate	FP- Rate	Precision	Recall	F-Measure	Class
J48	0.934	0.018	0.926	0.934	0.93	Very good
	0.868	0.029	0.888	0.868	0.878	Good
	0.945	0.02	0.93	0.945	0.938	Satisfactory
	0.989	0.002	0.991	0.989	0.99	Excellent
	1	0	0.999	1	1	Fail
Weighted Avg.	0.945	0.014	0.945	0.945	0.945	

4.5.3. Experiment III: Using J48 with 80/20 Percentage Split Test applied on the whole features.

In this experiment 80% of the dataset are taken as a training set and the other 20% are used for testing set. Percentage split option to train the classification model. As it can be seen from the summary the total number of data used for testing set is 1097(it is 20% of the dataset planned testing the trained dataset).From those 1028 instances are classified correctly (93.71%) and 69(6.29%) instances are incorrectly classified and taking 0.13 seconds to build the model.

Table 4.9.Confusion Matrix of J48 Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**198**	10	0	5	0
	Good	18	**187**	20	0	0
	Satisfactory	0	15	**234**	0	0
	Excellent	1	0	0	**191**	0
	Fail	0	0	0	0	**218**

Table 4.10. Detail analysis result of J48 algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.930	0.021	0.912	0.93	0.921	Very good
	0.831	0.029	0.882	0.831	0.856	Good
	0.94	0.024	0.921	0.94	0.93	Satisfactory
	0.995	0.006	0.974	0.995	0.985	Excellent
	1	0	1	1	1	Fail
Weighted Avg.	0.937	0.016	0.936	0.937	0.937	

4.5.4. Experiment IV: Using J48 with 80/20 Percentage Split Test applied on the selected features

This experiment conducts under 80/20 Percentage Split Test Option with selected attributes default parameters of WEKA and the algorithm generates a model as J48 and as it can be seen from the summary the total number of data used for testing set is 1097(it is 20% of the data set planned testing the trained data).From those 1034 instances are classified correctly (93.26%) and 63(5.74%) instances are incorrectly classified and taking 0.11 seconds to build the model.

Table 4.11.Confusion Matrix of J48 Algorithm

		Predictive class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**197**	14	0	2	0
	Good	16	**194**	15	0	0
	Satisfactory	0	16	**233**	0	0
	Excellent	0	0	0	**192**	0
	Fail	0	0	0	0	**218**

Table 4.12. Detail analysis result of J48 algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.925	0.018	0.925	0.925	0.925	Very good
	0.862	0.034	0.866	0.862	0.864	Good
	0.936	0.018	0.94	0.936	0.938	Satisfactory
	1	0.002	0.99	1	0.995	Excellent
	1	0	1	1	1	Fail
Weighted Avg.	0.943	0.015	0.942	0.943	0.942	

Experiment I, II, III and IV shows that the Classification accuracy of the models based on the above four methods respectively. The first experiment was performed based on 10-fold cross validation with whole attributes method and classifies with 94.55% accuracy rate, the second experiment performed based on 10-fold cross validation with selected attributes and accuracy of 94.53%,the third experiment was performed based on 80/20 percentage split Test Option with whole attributes method and classifies with 93.71% accuracy rate and the fourth experiment performed based on 80/20 percentage split Test Option with whole selected attributes with 94.26 % correctly classified instances .So to sum up, when we compared the four experiments the first

experiment performed based on 10-fold cross validation test option on whole attributes has a better accuracy than the others.

4.5.5. Experiment V: Using REPTree with 10-fold cross-validation applied on the whole features**.**This experiment conducts under 10-fold cross-validation test option with default parameters of WEKA and the algorithm generates a model as REPTree and Correctly Classified Instances are 5110 which means 93.18 % ,374(6.82%) incorrectly classified instances from the total 5484 and taking : 0.47 seconds to build the model.

Table 4.13.Confusion Matrix of REPTree Algorithm

		Predictive class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**992**	61	1	15	0
	Good	90	**966**	94	1	0
	Satisfactory	5	92	**1086**	1	3
	Excellent	11	0	0	**1034**	0
	Fail	0	0	0	0	**1032**

Table 4.14. Detail Analysis Result of REPTree Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.928	0.024	0.903	0.928	0.916	Very good
	0.839	0.035	0.863	0.839	0.851	Good
	0.915	0.022	0.92	0.915	0.917	Satisfactory
	0.989	0.004	0.984	0.989	0.987	Excellent
	1	0.001	0.997	1	0.999	Fail
Weighted Avg.	0.932	0.018	0.931	0.932	0	

4.5.6. Experiment VI: Using REPTree with 10-fold cross-validation applied on the selected features

This experiment conducts under 10-fold cross-validation test option with selected attributes default parameters of WEKA and the algorithm generates a model as REPTree and Correctly Classified Instances are 5088 which means 92.78 %,396(7.22%) incorrectly classified instances from total 5484 instances and taking 0.09 seconds to build the model.

Table 4.15.Confusion Matrix of REPTree Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**976**	76	0	17	0
	Good	87	**968**	96	0	0
	Satisfactory	6	104	**1074**	0	3
	Excellent	7	0	0	**1038**	0
	Fail	0	0	0	0	**1032**

Table 4.16. Detail Analysis Result of REPTree Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.913	0.023	0.907	0.913	0.91	Very good
	0.841	0.042	0.843	0.841	0.842	Good
	0.905	0.022	0.918	0.905	0.911	Satisfactory
	0.993	0.004	0.984	0.993	0.989	Excellent
	1	0.001	0.997	1	0.999	Fail
Weighted Avg.	0.928	0.019	0.928	0.928	0.928	

4.5.7. Experiment VII: Using PERTRee with 80/20 Percentage Split Test applied on the whole features

This experiment conducts under 80/20 Percentage Split Test Option with selected attributes default parameters of WEKA and the algorithm generates a model as REPTree and as it can be seen from the summary the total number of data used for testing set is 1097(it is 20% of the data set planned testing the trained data).From those 994 instances are classified correctly (90.61%) and 103(9.49%) instances are incorrectly classified and taking 0.16 seconds to build the model.

Table 4.17.Confusion Matrix of REPTree Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**181**	25	0	7	0
	Good	21	**187**	17	0	0
	Satisfactory	1	31	**217**	0	0
	Excellent	1	0	0	**191**	0
	Fail	0	0	0	0	**218**

Table 4.18. Detail Analysis Result of REPTree Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.850	0.026	0.887	0.85	0.868	Very good
	0.831	0.064	0.77	0.831	0.799	Good
	0.871	0.02	0.927	0.871	0.899	Satisfactory
	0.995	0.008	0.965	0.995	0.979	Excellent
	1	0	1	1	1	Fail
Weighted Avg.	0.906	0.024	0.908	0.906	0.907	

4.5.8. Experiment VIII: Using PERTRee with 80/20 Percentage Split Test applied on the selected features

This experiment conducts under 80/20 Percentage Split Test Option with selected attributes default parameters of WEKA and the algorithm generates a model as REPTree and as it can be seen from the summary the total number of data used for testing set is 1097(it is 20% of the data set planned testing the trained data).From those 1001instances are classified correctly 91.25% and 96 (8.75 %) instances are incorrectly classified and taking 0.13 seconds to build the model.

Table 4.19.Confusion Matrix of REPTree Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	182	24	0	7	0
	Good	21	188	16	0	0
	Satisfactory	1	27	221	0	0
	Excellent	0	0	0	192	0
	Fail	0	0	0	0	218

Table 4.20. Detail Analysis Result of REPTree Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.854	0.025	0.892	0.854	0.873	Very good
	0.836	0.058	0.787	0.836	0.81	Good
	0.888	0.019	0.932	0.888	0.909	Satisfactory
	1	0.008	0.965	1	0.982	Excellent
	1	0	1	1	1	Fail
Weighted Avg.	0.912	0.022	0.914	0.912	0.913	

As we can see from the REPTree experiment, it has low capacity to classify instances as correctly rather than classifies as incorrect.

Using Rule Induction

4.5.9. Experiment IX: Using JRiP with 10-fold cross-validation applied on the whole features

In this experiment JRIP rule induction algorithm is employed. Therefore, to generate IF-THEN rules from the experimental intrusion dataset JRIP algorithm with its default values of the parameter and 10-fold cross-validation test mode is employed with default parameters of WEKA and the algorithm generates a model as JRIP and Correctly Classified Instances are 5188 which means 94.60 % and Incorrectly Classified Instances are 296 which means 5.40 % from total number of 5484 of Instances and taking 4.11 seconds to build the model.

Table 4.21.Confusion Matrix of JRIP Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**1011**	47	0	11	0
	Good	75	**1028**	47	1	0
	Satisfactory	1	108	**107**	0	0
	Excellent	6	0	0	**1039**	0
	Fail	0	0	0	0	**1032**

Table 4.22. Detail Analysis Result of JRIP Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.946	0.019	0.925	0.946	0.935	Very good
	0.893	0.036	0.869	0.893	0.881	Good
	0.908	0.011	0.958	0.908	0.933	Satisfactory
	0.994	0.003	0.989	0.994	0.991	Excellent
	1	0	1	1	1	Fail
Weighted Avg.	0.946	0.014	0.947	0.946	0.946	

4.5.10. Experiment X: Using JRiP with 10-fold cross-validation applied on the selected features

This experiment conducts under 10-fold cross-validation test option with selected attributes default parameters of WEKA and the algorithm generates a model as JRIP and Correctly Classified Instances are 5161 which means 94.11 % ,323(5.89%) instances incorrectly classified from total number of 5484 of Instances and taking 3.53 seconds to build the model.

Table 4.23.Confusion Matrix of JRIP Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**1009**	47	1	12	0
	Good	65	**1031**	54	1	0
	Satisfactory	2	133	**1052**	0	0
	Excellent	8	0	0	**1037**	0
	Fail	0	0	0	0	**1032**

Table 4.24. Detail Analysis Result of JRIP Algorithm.

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.944	0.017	0.931	0.944	0.937	Very good
	0.896	0.042	0.851	0.896	0.873	Good
	0.886	0.013	0.95	0.886	0.917	Satisfactory
	0.992	0.003	0.988	0.992	0.99	Excellent
	1	0	1	1	1	Fail
Weighted Avg.	0.941	0.015	0.942	0.941	0.941	

4.5.11. Experiment XI: Using JRiP with 80/20 Percentage Split Test applied on the whole features

This experiment conducts under 80/20 Percentage Split Test Option with selected attributes default parameters of WEKA and the algorithm generates a model as JRIP and as it can be seen from the summary the total number of data used for testing set is 1097(it is 20% of the data set planned testing the trained data).From those 1031instances are classified correctly 93.98 % and 66 (6.02 %)instances are incorrectly classified and taking 3.11 seconds to build the model.

Table 4.25.Confusion Matrix of JRIp Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**205**	7	0	1	0
	Good	11	**199**	15	0	0
	Satisfactory	0	31	**218**	0	0
	Excellent	1	0	0	**191**	0
	Fail	0	0	0	0	**218**

Table 4.26. Detail Analysis Result of JRIP Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.962	0 .014	0.945	0.962	0.953	Very good
	0.884	0.044	0.84	0.884	0.861	Good
	0.876	0.018	0.936	0.876	0.905	Satisfactory
	0.995	0.001	0.995	0.995	0.995	Excellent
	1	0	1	1	1	Fail
Weighted Avg.	0.94	0.016	0.941	0.94	0.94	

4.5.12. Experiment XII: Using JRiP with 80/20 Percentage Split Test applied on the selected features

This experiment conducts under 80/20 Percentage Split Test Option with selected attributes default parameters of WEKA and the algorithm generates a model as JRIP and as it can be seen from the summary the total number of data used for testing set is 1097(it is 20% of the data set planned testing the trained data).From those 1032 instances are classified correctly 94.07 %, and 65 (5.92%) instances are incorrectly classified and taking 2.69 seconds to build the model

Table 4.27.Confusion Matrix of JRIP Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**203**	5	1	4	0
	Good	14	**198**	13	0	0
	Satisfactory	0	27	**221**	1	0
	Excellent	0	0	0	**192**	0
	Fail	0	0	0	0	**218**

Table 4.28. Detail Analysis Result of JRIP Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.953	0.016	0.935	0.953	0.944	Very good
	0.88	0.037	0.861	0.88	0.87	Good
	0.888	0.017	0.94	0.888	0.913	Satisfactory
	1	0.006	0.975	1	0.987	Excellent
	1	0	1	1	1	Fail
Weighted Avg.	0.941	0.015	0.941	0.941	0.941	

4.5.13. Experiment XIII: Using PART with 10-fold cross-validation applied on the whole features

In this experiment PART rule induction algorithm is employed. Therefore, to generate IF-THEN rules from the experimental intrusion dataset PART algorithm with its default values of the parameter and 10-fold cross-validation test mode is employed with default parameters of WEKA and the algorithm generates a model as PART and Correctly Classified Instances are 5230 which means 95.37 % from total number of 5484 of Instances and take 0.55 seconds to build the model.

Table 4.29.Confusion Matrix of PART Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**996**	58	0	15	0
	Good	57	**1035**	59	0	0
	Satisfactory	2	54	**1130**	0	1
	Excellent	8	0	0	**1037**	0
	Fail	0	0	0	0	**1032**

Table 4.30. Detail Analysis Result of PART Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.932	0.015	0.937	0.932	0.934	Vgood
	0.899	0.026	0.902	0.899	0.901	Good
	0.952	0.014	0.95	0.952	0.951	Satisfactory
	0.992	0.003	0.986	0.992	0.989	Excellent
	1	0	0.999	1	1	Fail
Weighted Avg.	0.954	0.012	0.954	0.954	0.954	

5.14. Experiment XIV: Using PART with 10-fold cross-validation applied on the selected features

This experiment conducts under 10-fold cross-validation test option with selected attributes default parameters of WEKA and the algorithm generates a model as PART and Correctly Classified Instances are 5208 which means 94.97%, 273(5.03%) incorrectly classified instances from total number of 5484 of Instances and take 0.45 seconds to build the model.

Table 4.31.Confusion Matrix of PART Algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**989**	60	3	17	0
	Good	62	**1026**	63	0	0
	Satisfactory	1	58	**1127**	0	1
	Excellent	11	0	0	**1034**	0
	Fail	0	0	0	0	**1032**

Table 4.32. Detail Analysis Result of PART Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.925	0.017	0.93	0.925	0.928	Very good
	0.891	0.027	0.897	0.891	0.894	Good
	0.949	0.015	0.945	0.949	0.947	Satisfactory
	0.989	0.004	0.984	0.989	0.987	Excellent
	1	0	0.999	1	1	Fail
Weighted Avg.	0.95	0.013	0.95	0.95	0.95	

4.5.15. Experiment XV: Using PART with 80/20 Percentage Split Test applied on the whole features

This experiment conducts under 80/20 Percentage Split Test Option with selected attributes default parameters of WEKA and the algorithm generates a model as PART and as it can be seen from the summary the total number of data used for testing set is 1097(it is 20% of the data set planned testing the trained data).From those 1041 instances are classified correctly (94.90 %) and 56 (5.1048 %) instances are incorrectly classified and taking 0.55 seconds to build the model.

Table 4.33.Confusion matrix of PART algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**194**	15	0	4	0
	Good	11	**200**	14	0	0
	Satisfactory	0	12	**237**	0	0
	Excellent	0	0	0	**192**	0
	Fail	0	0	0	0	**218**

Table 4.34. Detail Analysis Result of PART Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.911	0.012	0.946	0.911	0.928	Very good
	0.889	0.031	0.881	0.889	0.885	Good
	0.952	0.017	0.944	0.952	0.948	Satisfactory
	1	0.004	0.98	1	0.99	Excellent
	1	0	1	1	1	Fail
Weighted Avg.	0.949	0.013	0.949	0.949	0.949	

4.5.16. Experiment XVI: Using PART with 80/20 Percentage Split Test applied on the selected features

This experiment conducts under 80/20 Percentage Split Test Option with selected attributes default parameters of WEKA and the algorithm generates a model as PART and as it can be seen from the summary the total number of data used for testing set is 1097(it is 20% of the data set planned testing the trained data).From those 1033 instances are classified correctly 94.17 % and 64(5.83 %) instances are incorrectly classified and taking 0.39 seconds to build the model.

Table 4.35.Confusion matrix of PART algorithm

		Predicted class				
		Very good	Good	Satisfactory	Excellent	Fail
Actual class	Very good	**195**	9	0	9	0
	Good	15	**197**	13	0	0
	Satisfactory	0	18	**231**	0	0
	Excellent	0	0	0	**192**	0
	Fail	0	0	0	0	**218**

Table 4.36. Detail Analysis Result of PART Algorithm

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.915	0.017	0.929	0.915	0.922	Very good
	0.876	0.031	0.879	0.876	0.878	Good
	0.928	0.015	0.947	0.928	0.937	Satisfactory
	1	0.01	0.955	1	0.977	Excellent
	1	0	1	1	1	Fail
Weighted Avg.	0.942	0.015	0.941	0.942	0.941	

4.6. Performance of Comparison

As the researcher observed the PART algorithm under 10-fold cross-validation test option experiment it results the highest accuracy of 95.37 %. Develop classifier model by comparing the Selected Classifier Performance. In order to meet my objective, I need a model that can classify my dataset with a better performance from different viewpoints. So the below Table 4.36 has compare the output of all the four model based on accuracy of the model, the time it take to build the model, correctly classified instances and incorrectly classified instance based on the 10-fold cross-validation test option with whole attributes, 10fold Cross-validation with selected attributes , 80/20 percentage split with whole attributes and 80/20 percentage split with selected attributes on which they have highest value using four algorithms(J48,JRip,REPTree and PART were applied for the purpose.

Table 4.37. Performance Comparison

No	algorithm	Parameter	Accuracy (%)	Error analysis (%)
1	J48	10fold Cross-validation with whole attributes	94.55	5.45
		10fold Cross-validation with selected attribute	94.53	5.47
		80/20 percentage split with whole attribute	94.71	5.29
		80/20 percentage split with selected attribute	94.26	5.74
2	JRIP	10fold Cross-validation with whole attributes	94.60	5.40
		10fold Cross-validation with selected attribute	94.11	5.89
		80/20 percentage split with whole attribute	93.98	6.02
		80/20 percentage split with selected attribute	94.07	5.93
3	REPTree	10fold Cross-validation with whole attributes	93.18	6.82
		10fold Cross-validation with selected attribute	92.78	7.22
		80/20 percentage split with whole attribute	90.61	9.39
		80/20 percentage split with selected attribute	91.25	8.75
4	PART	10fold Cross-validation with whole attributes	**95.37**	**4.63**
		10fold Cross-validation with selected attribute	94.97	5.03
		80/20 percentage split with whole attribute	94.90	5.10
		80/20 percentage split with selected attribute	94.17	5.83

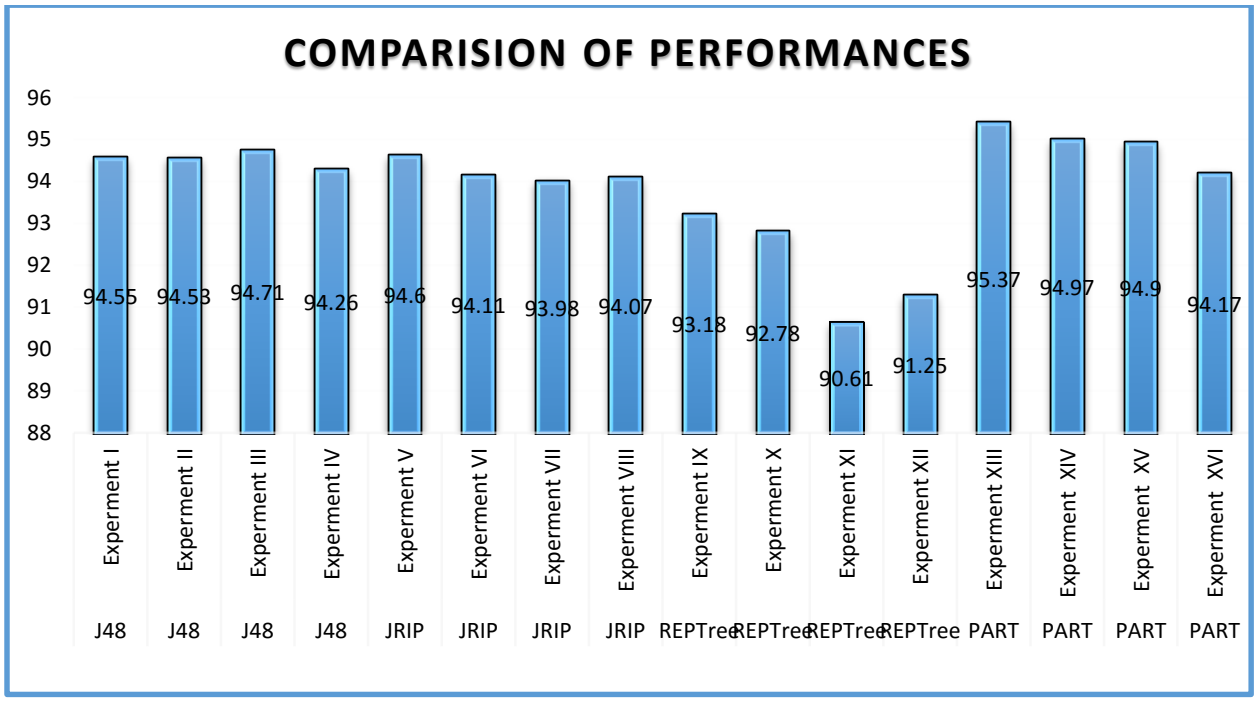


Figure 4.3. Comparison of performances

The above experiments were performed on WEKA experiment by using training sets in WEKA test option and the experiments were performed by using two methods namely 10-fold cross validation and percentage split test option. In these experiments four algorithms are used, namely REPTree, J48 decision tree PART and JRIP rule based. From each methods totally twelve models are developed based on the two methods. As shown on the above comparison table (Table 4.36), all the results were almost all closely equal but the difference lays on the execution period or the time taken to build the model. Thus, as we can see from the comparison, 10fold Cross-validation with whole attributes using PART algorithm is the best of all and has taken for the development of the system.

4.7. Rule Extraction from PART Classification Algorithm

Having generated rules using PART classifier, the next task is building or constructing the knowledge base. For this study we devised an automatic construction of knowledge base aligned with the data mining task. The overall task of the application is to extract rules from Preparatory school students' academic data set using PART classifier. The sample rules generated by PART shows in appendix II. In the rule generated, the number after slash reveal the wrongly classified instances and therefore it is possible to dedicate the probability of occurrence of the rule [36]. Although there are 157 generated rules, only some of the interesting rules which includes the five

have classes (Excellent, Very Good, Good, Satisfactory and Fail) are explained as sample in the **narrative form** as follows:

Students who score Grade nine (9) English subject an "Excellent" performance (90 – 100) Class Average Point , Grade ten (10) Ethiopian General Secondary Education Certificate Examination (EGSECE) mathematics subject an "Excellent" performance (81–100), Grade ten(10) physics subject "Fair " performance(50-59) Class Average Point and Grade eleven (11) English subject an "Excellent" performance (90 – 100) Class Average Point have tendency to score an "Excellent" performance(75 –100 in percentage or 527-700 total score) in Grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) result.

Students who score "very Good" performance (80 –89) Class Average Point in Grade nine (9) English subject, an "Excellent" performance (90 – 100) Class Average Point in Grade eleven(11) English subject, an "Excellent" performance (90 – 100) Class Average Point in Grade eleven (11) chemistry subject , "very Good" performance (80 –89) Class Average Point in Grade twelve(12) physics subject and Age1(Age category one (16–18)) have tendency to score an "Excellent" performance (75 – 100 in percentage or 527-700 total score) in grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) result.

Students who score "very Good" performance (80 –89) Class Average Point in Grade nine (9) English subject, an "Excellent" performance (90 – 100) Class Average Point in Grade eleven(11) English subject , "satisfactory" performance (60 – 79) in Grade ten (10) Ethiopian General Secondary Education Certificate Examination (EGSECE) mathematics subject ,an "Excellent " performance (90 –100) Class Average Point in Grade twelve(12) Civics subject have tendency to score "very Good" performance (63– 75 in percentage or 440-527 total score) in grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) result.

Students who score "satisfactory" performance (60 – 79) Class Average Point in Grade eleven (11) English subject , "very Good" performance (80 – 89) in Grade ten (10) Ethiopian General Secondary Education Certificate Examination (EGSECE) mathematics subject , "satisfactory" performance (60 –79) Class Average Point in Grade twelve(12) physics subject have the tendency to score "Good" performance (47 –63 in percentage or 352-440 total score) in grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) result.

Students who score "Satisfactory" performance (60 – 79) Class Average Point in Grade twelve (12) Civics subject , "Satisfactory" performance (60 – 79) Class Average Point in Grade ten(10)

physics subject , “Fair” performance (50-59) Class Average Point in Grade ten (10) mathematics subject , “Satisfactory” performance (60 – 79) Class Average Point in Grade twelve(12) in chemistry subject have the tendency to score “satisfactory” performance (25– 47 in percentage or177-352 total score) in grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) result.

Students who score “poor” performance (<50) Class Average Point in Grade twelve (12) Biology subject, “poor” performance (<50) Class Average Point in Grade eleven (11) English subject, “Fair” performance (50-59) Class Average Point in Grade twelve (12) mathematics subject have the tendency to score “Fail” performance Under 25 in percentage or Under 177 total score) in grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) result. Some top/best generated rules or rules that unique and newly discovered using data mining are depicted below.

Some an interesting unique and newly discovered rules for an “Excellent” performance are :

IF CICAPG11 = EXCELLENT AND MCAPG10 = EXCELLENT AND ENEGG10 = EXCELLENT **THEN** EXCELLENT.

IF ECAPG11 = EXCELLENT AND PCAPG12 = EXCELLENT AND ECAPG9 = VGOOD **THEN** EXCELLENT

IF ECAPG11 = EXCELLENT AND CHCAPG11 = EXCELLENT AND PCAPG12 = VGOOD AND Age = AGE-1 AND ECAPG9 = VGOOD **THEN** EXCELLENT

IF ECAPG11 = EXCELLENT AND ECAPG9 = EXCELLENT AND MNEGG10 = EXCELLENT AND PCAPG10 = FAIR **THEN** EXCELLENT

IF MCAPG10 = VGOOD AND CICAPG10 = EXCELLENT AND BCAPG10 = EXCELLENT AND Age = AGE-1 AND MCAPG9 = EXCELLENT **THEN** EXCELLENT

IF MCAPG10 = EXCELLENT AND ECAPG11 = EXCELLENT AND ECAPG9 = EXCELLENT **THEN** EXCELLENT

IF CICAPG12 = EXCELLENT AND CICAPG11 = EXCELLENT AND PNEGG10 = VGOOD **THEN** EXCELLENT

IF CICAPG12 = EXCELLENT AND CICAPG11 = EXCELLENT AND CHCAPG10 = SATISFACTORY **THEN** EXCELLENT

Some an interesting unique and newly discovered rules for “very Good” performance are:

IF ECAPG11 = SATISFACTORY AND MCAPG10 = VGOOD AND ECAPG10 = VGOOD **THEN** VGOOD

IF CICAPG12 = VGOOD AND PCAPG12 = VGOOD AND ENEGG10 = EXCELLENT **THEN** VGOOD

IF CICAPG12 = VGOOD AND CICAPG11 = EXCELLENT AND ECAPG11 = EXCELLENT **THEN** VGOOD

IF ECAPG11 = VGOOD AND MCAPG10 = EXCELLENT AND PNEGG10 = VGOOD **THEN** VGOOD

IF ECAPG11 = VGOOD AND MCAPG10 = SATISFACTORY AND PCAPG12 = VGOOD AND CICAPG11 = SATISFACTORY **THEN** VGOOD

IF ECAPG11 = SATISFACTORY AND CHCAPG11 = VGOOD AND PNEGG10 = VGOOD **THEN** VGOOD

IF ECAPG11 = VGOOD AND MCAPG10 = VGOOD AND CICAPG12 = EXCELLENT AND MCAPG12 = POOR AND ENEGG10 = EXCELLENT **THEN** VGOOD

IF ECAPG11 = VGOOD AND MCAPG10 = VGOOD AND MCAPG12 = VGOOD **THEN** VGOOD

IF ECAPG11 = VGOOD AND MCAPG10 = VGOOD AND ECAPG12 = FAIR **THEN** VGOOD

IF ECAPG11 = VGOOD AND MCAPG10 = VGOOD AND PCAPG12 = SATISFACTORY AND Age = AGE-1 **THEN** VGOOD

IF ECAPG11 = EXCELLENT AND CHCAPG11 = EXCELLENT AND Sex = F AND ENEGG10 = EXCELLENT **THEN** VGOOD

IF ECAPG11 = EXCELLENT AND MCAPG10 = SATISFACTORY AND ECAPG9 = VGOOD AND CICAPG12 = EXCELLENT **THEN** VGOOD

IF ECAPG11 = EXCELLENT AND MCAPG10 = SATISFACTORY AND ECAPG9 = SATISFACTORY AND PCAPG12 = SATISFACTORY AND CHCAPG11 = VGOOD AND Age = AGE-1 **THEN** VGOOD

IF MCAPG10 = VGOOD AND ECAPG9 = SATISFACTORY AND MNEGG10 = EXCELLENT **THEN** VGOOD

4.8. Use of the Discovered Knowledge

This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the

implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed [5].

The top and newly discovered rules or results of this thesis work have listed above will be disseminated to the following stakeholders and to any interested parties through the following ways:

- A softcopy of this thesis will upload to Debre Birhan University official website.
- Maximum effort will be exerted to publish the result in different journals (Publishing in local and international journals).
- Putting the hardcopy and the softcopy of this thesis results will be available in the preparatory schools library and other libraries. So that, interested readers can get access to the research output to be used for decision support, take action or to use it as a base for further research in the area or for any other applicable reasons.

CHAPTER FIVE

IMPLEMENTATION, EVALUATION and DISCUSSION of RESULTS

5.1. Introduction

Findings of data mining are the base for the development of knowledge through the WEKA software and java Neat Beans applications. Facts we get from Data mining can be represented as a rule. After knowledge acquisition is done using a PART rule induction algorithm, which performs on data collected from Debre Birhan City Preparatory schools and from ministry of education (MOE) the facts extracted is represented in the knowledge system. In addition to this, the researcher used in document analysis and interviewing domain experts who are work in Debre Birhan City Preparatory Schools(DBCPS) and domain experts who work at National Educational Assessment and Examination Agency (NEAEA) about the general idea of data as mentioned in previous sessions. Before applying Data mining(DM) algorithms it is necessary to carry out some pre-processing tasks such as cleaning, integration, discretization and variable transformation in order to get the targeted data set [67].The last step is developing the knowledge base using the knowledge acquired from data mining, domain expert and document analysis. Besides to this, since knowledge is always dynamic, the researcher tried to narrow the gap by automatically the data mining results of PART algorithm for generate rule, interviewed and document analysis the prototype is performed using WEKA software and Java Neat Beans. Then the main challenge here how to use the knowledge extracted from data mining in the developed prototype is discussed detail in the following sessions.

5.2. Prototype Development

All trained models in this research could be used to predict students' academic performance. Though, PART algorithm perform classification and prediction students' academic performance with minimal rate of error. As results, prototype has developed based on the extracted knowledge of PART algorithm. This prototype is based on the rule identified in section 7.4. These rules represent more than 80% of the training and testing dataset used in this study. Because, data classification is a two phase process in which first step is the training phase where the classifier algorithm builds a classifier with the training set of tuples and the second phase is classification

phase where the model is used for classification and its performance is analyzed with the testing set of tuples [76].

5.3. User Interface

The user interface is a channel for communication between the system and the end user. The prototype graphical user interface is developed based on the model generated by PART classifier with whole attributes selected based on the relevancy by the help of domain experts .Thus, the researcher used the PART algorithm due to its higher performance than the other three algorithms (J48, REPTree and JRIP) to design the graphical user interface by using 157 rules that briefly discussed in chapter three and four respectively.

After the prototype displays the greeting page by using help facility description, a user can interact directly with the system by selecting the predicted result (Excellent, very Good, Good, Satisfactory and Fail) and then the predictors. Here, especially students used this user interface during the teaching and learning process at each grade level in order to perform the better or best performance by choosing their bests to get their weak side and fill the gaps using the predictors (most useful determinants) accordingly for the five categories of performances if the other conditions are satisfied. Moreover, teachers and families can simply follow and help their students early before they face the problems which causes economic strain on all the concerned and the country. For example, the following figure 5.1.Shows the sample dialogue windows between the user and the system to display the most predominant predictors among the six common subjects (English, Maths, Physics, Chemistry, Biology and Civics) in grade 9,Grade 10,grade 10 National Exam, Grade 11 and grade 12 Class Average Points(CAP).

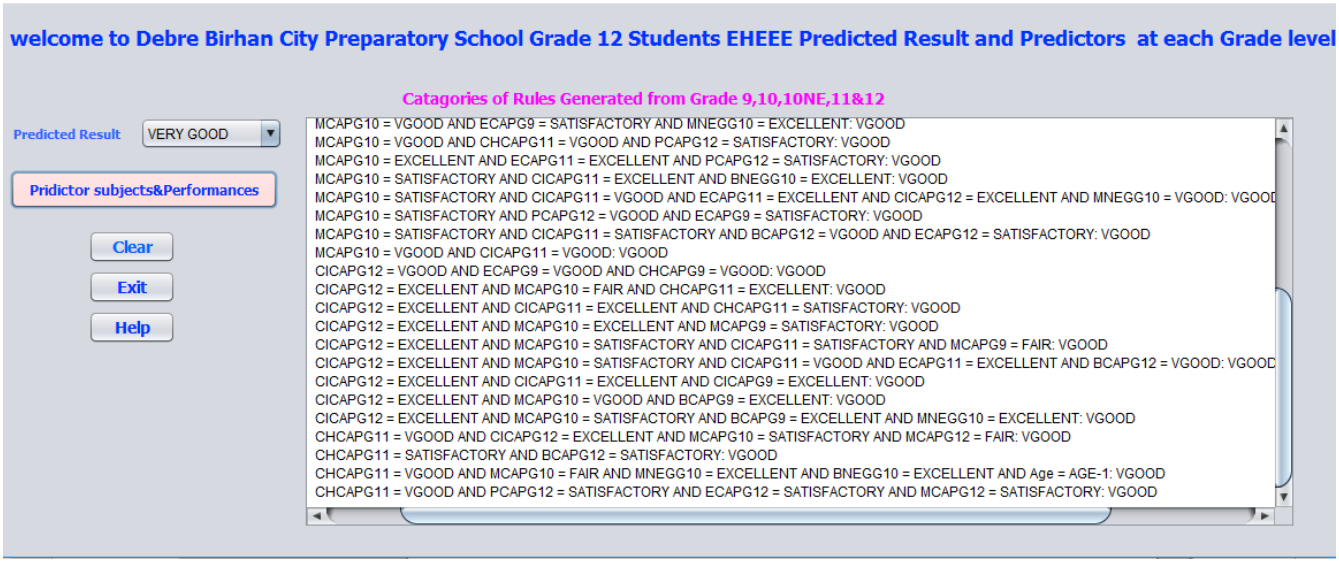


Figure 5.1. The GUI to show the most pre predictors of students' performance in EHEEE result.

5.4. Help Facility

The help facility describes when a new user use this system may confuse to click the combo box button and see drop down menu like Excellent, very good, good, satisfactory, and fail, not fast to understand what does it mean click help and select one of them and so on. As we can see in the above figure 5.1 the system supports its options of rules for every five performances under categories of rules generated from grade 9,10,10NE,11and 12. This helps the students, educational experts and educational planners to follow up the progressive teaching and learning work based on the subjects' performances appeared in the rules to select best rules accourdglly.

5.5. Evaluation of the Prototype

In order to assure that the prototype meets the requirement it is developed for, it has to be tested. Testing was held to have sure conformity of the system with user acceptance testing, domain experts, system requirement and has to validate by test cases before take it into action by the domain experts [77]. System performance testing is basically used to measure the accuracy of the system. 50 test cases are from natural science stream students 'common subjects dataset in grade 9-12 and their grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) result performances in 2010 E.C was extracted and given to the predictive model. True positive, F-measure, recall and Precision measure how much the system is accurate. Confusion matrix is

used for measure the performance of the system. The other way of evaluation is user acceptance testing to check the proposed system meets the potential user requirement.

5.5.1. System Performance Testing

The system performance evaluation is preparing test cases from 2010 E.C Grade 12 Ethiopian Higher Education Entrance Examination(EHEEE) result performance from Ministry of Education (MOE) exam agency and their Grade 9- Grade 12 class average points as well as Grade 10 Ethiopian General Secondary Education Certificate Examination (EGSECE) results from Debre Brian City Preparatory Schools .The test cases(instances) which are unlabeled(removed) performances in Ethiopian Higher Education Entrance Examination (EHEEE) result which have labeled by NEAEA delivered to the system to label them as Excellent, Very good, Good, Satisfactory, and Fail/ i.e. poor and the researcher and domain experts(teachers, principal and students as well as zone education domain experts) checked how much the system is strong to predict the of Grade 12 students' performance Ethiopian Higher Education Entrance Examination (EHEEE) result. Thus, the researcher have used the test cases which include samples of student's instances from Debre Brian City Preparatory School Students class result average points of Grade 9-Grade 12, Ethiopian General Secondary Education Certificate Examination (EGSECE) result and Ethiopian Higher Education Entrance Examination (EHEEE) result dataset of Grade 12 natural science for the students in 2010 E.C. The instances include 32 attributes with their respective values to compare the performance of the system and the results of the students. In this study, four educational planning instructors from North shewa zone educational office and one principal from NEAEA were selected for the purpose of testing the performance of prototype system by referring the performance of the students with the system predicted performances. The criteria for selecting the evaluators were because of the fact that the instructors are in the domain area and the NEAEA principal is also concerned. About 50 test cases which include 32 attributes with their respective values are taken for the testing purpose. The set of test instances are provided to performances predictive system and the outputs are compared to the Confusion matrix is used for measure the performance of the system. Thus, True positive, F-measure, Recall and Precision measure how accurate the system is. So that, the questionnaire format for feedbacks of domain experts on system interactions and sample of test cases from Grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) result performances in 2010 E.C are shown on appendix III. The Confusion matrix used for

comparing the performance of Grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) result in 2010 E.C and interview questions feedbacks/suggestions from domain experts on system interactions is shown below on Table 5.3.

Table 5.1. Confusion matrix for evaluation of proposed system

2010 E.C Grade 12 Natural Science Stream common subjects (Grade 9-12 attributes values) and performances in EHEEE result	Predictive model feed back					Total
	Types of performances	Fail	Satisfactory	Good	Very Good	Excellent	
	Fail	**8**	1	0	0	0	9
	Satisfactory	0	**17**	1	0	0	18
	Good	0	0	**11**	0	0	11
	Very Good	0	0	0	**6**	0	6
	Excellent	0	0	0	0	**6**	6
	Total	8	18	12	6	6	50

The confusion matrix on Table 5.3 shows matrix of test cases evaluation by Performances predictive system and Grade 12 EHEEE result performances in 2010 E.C. The rows illustrate evaluation of Grade 12 EHEEE result performances in 2010 E.C and the columns illustrate result of Performances predictive system. The entries under column Satisfactory the system testified that 18 instances are Satisfactory Performance. But 17 of the instances are correctly identified as Satisfactory Performance and one (1) instance is identified as Fail Performance. The entries in the confusion matrix under Good column describe that 12 instances are correctly identified by the system as Good. But 11 instances are correctly classified as Good and one (1) instance is classified as Satisfactory Performance. The entries for Fail, Very Good and Excellent columns show that the system has correctly identified 8 instances as Fail Performances, 6 instances as Very Good and again 6 instances are as Excellent respectively. The entries in the confusion matrix under Good column describe that 1 satisfactory performance instance incorrectly identified by the system as Good. With regard to Good prediction the system achieved the lowest result as compared to the others. But as shown in the confusion matrix, one (1) Satisfactory Performance instance is incorrectly identified as good and one Fail instance is incorrectly identified as Satisfactory Performance. The researcher believes that the system‘s identification of instances to isolate correctly Satisfactory from Fail or Good from Satisfactory is a little

difficulty. The system has correctly identified 48 test instances out of 50 to their correct class. This means the system has 96% classified correctly and two (2) instances out of fifty 50 are incorrectly classified which is 4%.

As clearly illustrated in table 5.4 bellow, the system‘s performance is evaluated in terms of TP rate, FP rate, Precision, Recall and F-measure which enables us to view in detail how accurate is the system in perdition types of performance.

Table 5.2. System performance evaluation by 2010 E.C EHEEE result.

	TP rate	FP rate	Precision	Recall	F-Measure	Type of performance
	88.89%	0	100%	88.89%	94.12%	Fail
	94.44%	3.13%	94.44%	94.44%	94.44%	Satisfactory
	100%	2.63%	91.67%	100%	95.65%	Good
	100%	0	100%	100%	100%	Very good
	100%	0	100%	100%	100%	Excellent
Weighted Average	96.67%	2.88%	97.22%	96.67%	96.84%	

According to the TP Rate and recall in table 5.4, the type of performances that are correctly classified out of all performances shows that Good, Very good and excellent class scores the highest TP Rate and recall 100%, followed by the Satisfactory Performance class 94.44% and Fail Performance class 88.89%. However, taking only the TP Rate and recall for measuring the performance of prediction model can be confusing. Instead, the researcher should take the commonly used measuring parameters for measuring the performance of any classifier such as weighted average of precision and f-measure in order to measure the performance of prediction model for performance testing. The performance of the system achieves weighted average of 96.67% TP Rate, & recall, precision of 97.22% ,F-Measure of 96.84% ,and 2.88% FP Rate. However, the system score FP rate 3.13%, and 2.63% for Satisfactory and Good performance respectively. This shows to prediction satisfactory and Good performance by the system is a little difficulty of classification of performance.

5.5.2. User Acceptance Testing

The user acceptance testing ensures how the users or domain experts view the system on the bases of the rules and performance to predict the students result. Thus, in this study 24 users were selected and had been given the chance to use and interact with the system. The user selected were two education experts from north showa zone, ten Natural Science Stream teachers

and Twelve Grade 12 Natural Science Stream students from Debre Birhan City preparatory schools. The interview questions format for feedbacks of domain experts on system interactions are shown on appendix IV.

The evaluators assessed prediction model by using the following standards [78].

- Easiness to use and interact with the system
- Attractiveness of the system
- Efficiency in time
- The accuracy of the system in reaching a decision to identify the performance
- The ability of the system to make right conclusion and recommendation
- Importance of the model in the domain area

Different researchers have used different types of user acceptance testing evaluation criteria [53]. But for this study, the evaluation criteria suggested by [79] and [80] have been used such as Excellent = 5, Very Good =4, Good =3, Fair =2 and Poor =1.Therefore, evaluators were allowed to the following closed ended questions.

Table 5.3. The feedbacks of domain experts on system interactions.

Criteria of evaluation	Respondents on					Average
	Excellent	Very good	Good	Fair	Poor	
Easiness to use and interact with the system	13	8	3	0	0	4.41
Attractiveness of the system	14	6	4	0	0	4.42
Efficiency in time	15	9	0	0	0	4.63
The ability of the system to make right conclusion and recommendation	14	8	2	0	0	4.50
The accuracy of the prototype in reaching a decision to predict the performance of students	16	8	0	0	0	4.67
Importance of the prototype in the domain area	17	7	0	0	0	4.70
	Average					4.56

As table 5.5 points out, the highest number that 13 divided by 24 becomes 54.17 % of the respondents reply easiness use and to interact with the prototype as Excellent, 33.33% rated easiness to use and to interact with the prototype as very good, 12.5% rated easiness to use and to

interact with the prototype as good. In case of attractiveness of the prototype, 58.33% of the respondents reply the prototype knowledge base system as Excellent, the highest number which is 25 % rated the attractiveness of the prototype knowledge base system as very good, 16.67% of the respondents reply the prototype knowledge base system as good. In case of time, efficiency, the same number of evaluators 37.5 % rated the prototype knowledge base system as very good and 62.5% as excellent. Regarding the ability of the prototype system in to make right conclusion and recommendation also 33.33% system as very good and 58.33% as Excellent, 8.33% of respondents reply the prototype knowledge based system as good. In case of the accuracy of the prototype in reaching a decision to predict the performance of students 66.67 % of the respondents evaluate the prototype as Excellent, 33.33% % of the respondents evaluate the prototype as very good. In case of Importance of the Prototype in the domain area, 70.83% as Excellent, 29.17 % as very good. In sum, 61.81% of the respondents assessed the prediction model by using the evaluation standards as Excellent, 31.94% of the respondents assessed the prediction model as 'very good' and 6.25% of the respondents assessed the prediction model as Good. Based on the results obtained the overall average performance prediction system with user's point of view is 4.56 out of five categories is 91.2%. This implies that the modeled prototype was performs Excellent in making right decisions on the student performance prediction from the domain Expert point of view. The results obtained from domain experts during interaction with the system and closed ended question are mostly the same. All of the evaluators evaluated the system as good, very Good and Excellent and none of them respond poor and fair for any section of the suggestion. Table 5.6 indicates the number of responses obtained for each of the options that the respondents' rate to evaluate the prototype in Closed ended questions.

Table 5.4. Domain experts' response in closed ended question.

Evaluators who responds as	value	Total number of responses for each option of all 6 questions.			Percentage
		Zone education expert	Teachers	Students	
Excellent	5	2	5	6	54.17%
Very good	4	1	4	5	41.67%
Good	3	0	1	1	8.33%
Fair	2	0	0	0	0%
Poor	1	0	0	0	0%

As shown in the above table 5.6, based on closed ended questions as evaluation criteria, the domain experts reply 2 times by Zone education expert, 5 times by Teachers, 6 times by Students the prototype questions as excellent 13 times which is 54.17% and so on for very good, good, fair and poor responses for the prototype evaluation criteria as shown above in table 5.6. The least response is good 2 times 8.33% regarding the prototype on its easiness, attractiveness of the system, the ability of the system to make right conclusion and recommendation.

Therefore, the performance of the prototype in both cases has got "very good"in predicting the performance of the students. This prototype has also great value in predicting performance of students with time efficiency and cost effectiveness.

5.6. Results Discussion

In this study, an attempt to identify those attributes causing the students failure upon 5484 Ethiopian Higher Education Entrance Examination (EHEEE) and Ethiopian General Secondary Education Certificate Examination (EGSECE) records stored in NEAEA and Class Average Point (CAP) of corresponding students in Debre Birhan City Preparatory Schools (DBCPS). After this data passed through the pre- processing step, the researcher have done the followings in order to answer the research questions.

1. The researcher have took the domain expert's interview at both sources of datasets for clarifying the datasets which are unclear (abbreviated, new attributes for the researcher) and for choosing the common subjects and information gain method were used for find out that all instances of the rule generated to answer the question **"Which data attributes are significant in achieving the most significant prediction analysis that will help to improve course performance**?". Thus, the most significant prediction analysis that will help to improve course performance of students in preparatory schools are 32 relevant attributes based on the info gain ratio values as shown in Table 4.2. are English, Natural mathematics, Physics, Chemistry, Biology and Civics in Grade 9, Grade 10, Grade 11 and Grade 12 including Grade 10 Ethiopian General Secondary Education Certificate Examination (EGSECE) attributes, Age, Sex, and preparatory school students Ethiopian Higher Education Entrance Examination (EHEEE) result as class. Grade 11 Civics, Grade 10 mathematics, Grade 11 English, Grade 12 Civics, Grade 11 Chemistry, Grade 12 Physics, Grade 9 English, grade 10 Physics, Ethiopian General Secondary Education Certificate Examination (EGSECE) English, Ethiopian General Secondary Education Certificate Examination (EGSECE) mathematics which are the top ten determinants to perform

better performances than the fail and satisfactory performances of the students in Ethiopian Higher Education Entrance Examination (EHEEE) result. From the analysis, the determinant factors for the students' Excellent performance are scoring Excellent In Grade 11 Civics, Grade 10 mathematics, and Ethiopian General Secondary Education Certificate Examination (EGSECE) English, Grade 11 English, Grade 12 Physics, Grade 11 Chemistry, Grade 9 English, Ethiopian General Secondary Education Certificate Examination (EGSECE) mathematics, grade10 civics, Grade 10 biology, Age = Age1 , , Grade 9 mathematics, grade12 civics, Grade 9 English = Very good, Grade 12 Physics = Very good, Physics, Grade 10 = Very good, Grade 10 mathematics = Very good And Grade 10 Chemistry = Satisfactory

Based on the rules generated, students can adjust the better performance he/she will perform at in grade 12 Ethiopian Higher Education Entrance Examination (EHEEE) result by giving attention to the determinant subjects in each grade levels.

2. Sixteen experiments have done b/y using four classification algorithms namely J48, JRIP, PART and REPTree under 10-fold Cross-Validation test option, whole attribute and percentage split 80/20 test option were conducted under each classification algorithms to answer question "**which data mining algorithm can be more appropriate to develop model that predict significant factors for Students performances in preparatory schools?".** So as the experiments showed that PART classification algorithm is the best classification algorithm to develop the prediction model that can predict the performance of the students. Because the PART algorithm registered better performance with 95.37% evaluation result and the researcher used it for further in the development of prototype. Based on this more appropriate algorithm, the researcher have got the performance of the proposed system achieved 96% by test cases respectively and 91.2% by using user acceptance evaluation standards.

3. Some most an interesting rules for an "Excellent" performance are the followings. Thus," **What are the most an interesting patterns or rules generated using the determinant attributes?"** is answered with the following rule generated.

IF CICAPG11 = EXCELLENT AND MCAPG10 = EXCELLENT AND ENEGG10 = EXCELLENT **THEN** EXCELLENT

IF ECAPG11 = EXCELLENT AND PCAPG12 = EXCELLENT AND ECAPG9 = VGOOD **THEN** EXCELLENT

IF ECAPG11 = EXCELLENT AND CHCAPG11 = EXCELLENT AND PCAPG12 = VGOOD AND Age =

AGE-1 AND ECAPG9 = VGOOD **THEN** EXCELLENT

IF ECAPG11 = EXCELLENT AND ECAPG9 = EXCELLENT AND MNEGG10 = EXCELLENT AND PCAPG10 = FAIR **THEN** EXCELLENT

IF MCAPG10 = VGOOD AND CICAPG10 = EXCELLENT AND BCAPG10 = EXCELLENT AND Age = AGE-1 AND MCAPG9 = EXCELLENT **THEN** EXCELLENT

IF MCAPG10 = EXCELLENT AND ECAPG11 = EXCELLENT AND ECAPG9 = EXCELLENT **THEN** EXCELLENT

IF CICAPG12 = EXCELLENT AND CICAPG11 = EXCELLENT AND PNEGG10 = VGOOD **THEN** EXCELLENT

IF CICAPG12 = EXCELLENT AND CICAPG11 = EXCELLENT AND CHCAPG10 = SATISFACTORY **THEN** EXCELLENT

Therefore, educational planners, domain experts can help the students by make decision to follow these rules trend in order to enable the students to score the better performance in their academic results. All the main educational quality participants can use the graphical user interface to know the rules that predict the performances of UEE.

CHAPTER SIX
Summary, Conclusion and Recommendation

6.1. SUMMARY

Data mining is data analysis methodology used to identify hidden patterns in a large data set. It has been successfully used in different areas including the educational environment whereas Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process.

The objective of the study was to develop a Classification model which predicts the Performance of Natural Science Preparatory School Students in Ethiopian Higher Education Entrance Examination (EHEEE) result using data mining techniques. In order to achieve the objective, the study uses Ethiopian Higher Education Entrance Examination (EHEEE) data, corresponding Grade 9, Grade 10, Grade 11and Grade 12 natural science common subjects class average point and Ethiopian General Secondary Education Certificate Examination (EGSECE) dataset. The dataset from an agency which is National Educational Assessment and Examination Agency responsible for giving an examination, correcting the examination and make an analysis for Grade 10 and Grade 12 students throughout the country.

Before the analysis, data pre-processing techniques are applied in order to get unbiased results. Data which creates noise are eliminated by data cleaning. Five years Ethiopian Higher Education Entrance Examination (EHEEE) result dataset from NEAEA is integrated with its corresponding five years Ethiopian General Secondary Education Certificate Examination (EGSECE) and the common natural science stream subject's class average point's data of Grade 9, Grade 10, Grade 11and Grade 12 preparatory schools in Debre Birhan City(DBC). Common subjects of natural science stream are English, Math's for natural, physics, chemistry, Biology, civics ethical education subjects are selected. The results of the common subjects' attributes are discretized using the assessment system of secondary education in Ethiopia.

From the data mining models, the hybrid data model is applied in order to achieve the objective of the research. This includes problem understanding, data understanding, data preparation, data mining and evaluation of the model. The study uses classification rule mining to discover the rules which classifies performances of the students in Ethiopian Higher Education Entrance Examination (EHEEE). J48, REPTree, PART and JRIP algorithms are compared based on the

execution/carrying out time, Accuracy, TP Rate, Precision, Recall, and F-measure. Using different algorithms different rules are generated. The rules are then evaluated using the interestingness measure accuracy. Thus, PART algorithm is selected to generate rule and the analysis is undertaken to develop prototype which enables domain experts to predict the academic performance of the students in Ethiopian Higher Education Entrance Examination (EHEEE).

The findings of the research showed that the most predominant subjects are English, Mathematics and Science subjects excluding physics. The academic achievement result of male students in both five subjects in both grade 10 and 12 is better than female students. This implies that government and concerned stakeholders in education sector should provide additional support for female students. So the present education system may adopt this as input revision of its educational policy.

6.2. CONCLUSIONS

As a summary to this research Education plays a great role in achieving one's country growth and development. The quality of the education can be seen from the side of performances of students based on the educational institutions rules. Thus, early prediction of the students' performance can help in making different and timely managerial decisions at each levels in order to improve the academic performance of students. Students' low academic performance has been a long standing problem in preparatory schools.

The objective of this research undertaken was to investigate the possible application of data mining technology in Ethiopian preparatory school education context, particularly on Debre Birhan City preparatory schools students' data set in order to develop a predictive model that could help preparatory schools to identify students will perform . Thus, they can be treated before the condition escalate into students' academic dismissal and wastage of resources. Moreover, curriculum specialist can also benefit from EDM in identifying strengths and weaknesses in materials used for the course, and based on that, proper interventions can be made to these courses.

The original data which was 3013 instances imbalanced number of attributes are changed to balanced datasets (SMOTE) 5484 instances was used to minimize the impact of majority class data sets on degree of accuracy. Hence the performance of decision tree classifier on the resampling data set was increased than the unbalanced dataset. The dataset is preprocessed using

MS excel and MATLAB and made suitable experiment using J48, REP Tree, PART and JRip classifier algorithms for extracting hidden knowledge. Hence, data mining classifier, PART is engaged for knowledge acquisition step since it has performed best among the selected classifiers with an accuracy of 95.37% which means that the model is successfully predicting the final Grade of 5230 students out of 5484 have been successfully classified. In this study, various predictive models have been developed and validated for the purpose of trimming down the student failure rate in natural science stream courses in EHEEE at preparatory school, by making pre prediction of the students' marks in exams of each Grade level (secondary and preparatory school cycles).

Using WEKA 3.6.13 version tool, 32 combinations of variables have been selected by info gain feature selection method, to be used as predictors and total of 157 rules were generated. So, the result of Classification model reveals that specific courses, student academic status in five years, age and sex are determinant of students' performance. So that, teacher, students and their parents can improve the result of student who are likely to pass in low Grade through proper counseling by using the generated rule. It also assist students before they reached risk of failure, effective resource utilization and cost minimization, helping and guiding administrative officers to be successful in management and decision making. Finally, system performance evaluation testing and user acceptance testing were conducted. As a result, the proposed system could perform 96% of system performance evaluation. This result indicates that the study was effective in acquiring knowledge through data Ming. The user acceptance testing is achieved 91.2% performance based on six criteria of evaluation. Selected domain experts are trained and used the system to evaluate how much the prototype or the system meets their requirements.

In general, the results obtained from this research have shown the potential applicability of data mining technology to classify preparatory students' academic performance as Excellent, Very Good, Good, Satisfactory and Failure. It was possible to identify the main determining attributes/variables and their values for the performances of students in a specific schools; number of common courses given in a semester at each class level, Ethiopian General Secondary Education Certificate Examination (EGSECE), Higher Education Entrance Certificate Examination (EHEEE) result of a student, age, and sex were the main determining attributes obtained from this research result.

As the experiment of the study shows both male and female students of Debre Birhan City preparatory schools failed to score a good result in Physics subject. This shows that the schools need to give an attention to physics subject so as to improve the students 'performance. From the analysis, the determinant factors for the students' excellent performance are scoring Very Good in Physics, Civics and Biology subjects in Ethiopian General Secondary Education Certificate Examination (EGSECE). Similarly scoring good in English in Ethiopian General Secondary Education Certificate Examination (EGSECE) is also another determinant factor. In similar manner, scoring satisfactory in Chemistry in Ethiopian Higher Education Entrance Examination (EHEEE) and scoring satisfactory in English in ECAPG11 are the determinant factors for the students' better performance in Ethiopian Higher Education Entrance Examination (EHEEE).

MOE uses the students' total score to set the cutting point. From the analysis, the determinant factors for the students' success are scoring Very good in Physics, Civics and Biology subjects in EHEEE. Similarly scoring good in English in Ethiopian Higher Education Entrance Examination (EHEEE) is also another determinant factor. In similar manner, scoring satisfactory in Chemistry in Ethiopian Higher Education Entrance Examination (EHEEE) and scoring satisfactory in English in Ethiopian General Secondary Education Certificate Examination (EGSECE) are the determinant factors for the students' success in entering higher education. Likewise strong rules are achieved using classification algorithm which can be easily understand by domain experts and other stakeholders.

The hidden knowledge (findings) of the research are: The percentage of students who have scored better performances are few whereas students who have scored low performances are large in number as shown in Table 3.2, Grade 9 English, Mathematics, Grade 10 English, Mathematics, Civics, Mathematics of National Exam, English of National Exam, Biology, Physics, Grade 11civics, English, Chemistry, Grade12 physics, Civics and Age1 are the Most Predominant for 'Excellent' Performance. Grade 9 English, Biology, Chemistry, Grade 10 English, Maths,Grade11english,Civics,Chemistry,Sex-Female,Age-1,Grade12 ,Civics, Physics, Maths, English and Grade 10 National Exam English, Physics, Maths and Biology are the Most Pre Dominants for 'Very Good' Performance.

6.3. RECOMMENDATION

Develop computerized data collection method is highly recommended for data handlers. Hence, the future researcher easily access the dataset to address any educational research problem.

Based on the findings of this research work, the researcher would like to make the following recommendations:

- In this research work, an attempt has been made to assess the applicability of data mining technology to predict the likelihood of student academic performance in the Preparatory schools by using some set of variables/attributes that were considered as important by different literatures. For a number of other variables, in education area of Ethiopia; especially student performance versus health related problems, financial resource problems, family background, academic schedule and assessment method, qualification of lecturers and much more, it remains to investigate further the effect of those variables to build models with better accuracy and performance than the models built in this research work.
- Even though there are many data mining techniques, the researcher done experiment by classification only. But the other data mining techniques which were not tested by the researcher might reveal important patterns in relation to age, sex and academic factors which affecting student performance in Ethiopian Higher Education Entrance Examination (EHEEE).
- Teacher, students and their parents can improve the result of student who are likely to pass in low Grade through proper counseling by using the generated rule. They should also assist early the students before they reached risk of failure or low performances, effective resource utilization and cost minimization, helping and guiding administrative officers to be successful in management and decision making. All stake holders in education should treat the students with all possibilities before the condition escalate into students' academic low performance, dismissal and wastage of resources in order to have high performance in Ethiopian Higher Education Entrance Examination (EHEEE).
- Since the preparatory classes have an aim of preparing students for higher education.so that, different stake holders can achieve important points from this study which are helpful in strengthen the students' performance.
- The study uses classification rule mining method and PART algorithm to identify the determinant factor for the students' performance in Ethiopian Higher Education Entrance Examination (EHEEE); it will be essential if other researches use other algorithm since there is no standard way of setting the experiment and this leads to important rules. Both male and female students of Government Preparatory schools failed to score a good result in Physics

subject. So, the Preparatory schools need to give an attention to physics subject so as to improve the students' performance and in Ethiopian Higher Education Entrance Examination (EHEEE).

- Policy makers and education service providers should use and implement the findings to real field practices and provisions. It is also recommended that the identified model should be applied in the field and monitored to validate the finding.
- Developing the prototype system to access from anywhere and anytime where the connection is available using mobile as well as personal computer platform.

6.4 Future Research

This study is focused on Debre Birhan City preparatory schools students' academic performance in Ethiopian Higher Education Entrance Examination (EHEEE); it will be important if other study works on other areas (Cities, zones and regions) with the same data format. Moreover, the study focuses on natural science stream preparatory students, it will be essential if other research works by including both streams. It will be also important if other study works on Student performance by considering attributes like: Student performance versus health related problems, financial resource problems, family background, academic schedule and assessment method, qualification of teachers, facility of school and much more, number of the students in the class, it remains to investigate further the effect of those variables to build models with better accuracy and performance than the models built in this research work.

References

[1] O. Negassa, ″Ethiopian students' achievement challenges in science education: implications to
Policy formulation," *International Journal of Computer Applications,* vol. 131, no. 5, 2014.

[2] M .Tesfa , "The Validity of university entrance examination and high school Grade point average for predicting first year university students' academic performance," University of Twente, Faculty of Behavioural science, " 2013.

[3] C. Romero, ″Educational Data Mining*:* A Review of the State of the Art,″ *IEEE transaction on systems, MAN, and Cybernetics-part c: Application and Reviews,* vol. 40, no. 6, 2010.

[4] W.Rouse, ″ Need to know—information, knowledge, and decision making,″ *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 32, no. 4, pp. 282-292, 2002.

[5] M. Anwar, "knowledge mining in supervised and unsupervised assessment," *2nd International conference on Networking and Information,* vol. 17, 2015.

[6] N. S. Shah, "predicting factors that affect students' academic performance by using data mining techniques," *Pakistan business review, 2012.*

[7] M. A. Yehuala, "Application of data mining technique for student success and failure prediction (the case of Debre Markos university," *International Journal of scientific &technology research,* vol. 4, no. 04, 2015.

[8] M. Adell, "Stategies for improving academic performance in adolecents, spain, Madrid," 2002.

[9] N. S. H. E. P. Department, "Students' higher education placement," *MOE, ADDIS ABEBA, 2005 -2009.*

[10] Z. Alebachew, Analysis of Urban Growth And Sprawl Mapping Using Remote Sensing And Geographic Information System, Debre Birhan: Debre Birhan Town., 2011.

[11] D. Zam, "Annual Education statistics," Ministry of Education, Addis Ababa, 2014.

[12] K. Mehmed and B. John, "Data mining concepts, Models, Methods, and Algorithms," *Wiley -IEEE Press, USA, 2003.*

[13] F. Ahmad, "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques," vol. 1, no. 1, 2015.

[14] K. J. Cios, P. Witold, S. Roman W. and K. Lukasz A., Data Mining: A Knowledge

Discovery Approach, USA: Springer Science & Business Media, 2007.

[15] A .Selam, "Predicting the Occurrence of Measles Outbreak in Ethiopia Using DM Technology," MSc. Thesis Addis Ababa University, Addis Ababa, Ethiopia, 2011.

[16] Net Beans, 06/04/2015. [Online]. Available: https://netbeans.org/features/.

[17] K. J. Sathick, "Extraction of Actionable knowledge to predict students' academic performance using data mining technique, an experimental study," vol. 1, no. 1, 2013.

[18] P. Cortez, "Using Data mining to predict secondary school students' performance," 2014.

[19] F. T., "Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining*," International Conference on Information Acquisition*, vol. 1, no. 1, 2006.

[20] T. Sutch, "Using association rules to understand subject choice at AS/A level." Cambridge, 2015.

[21] B. K., & Pal, S., Baradwaj "Mining educational data to analyze students' performance," *22200 Jertih, Terengganu,* vol. vol, no. arXiv., p. 1201.3417, 2015.

[22] A. Azwa, "First Semester Computer Science Academic Performances Analysis by using Data Mining Classification algorithms," *AICS,* vol. 1, no. 1, pp. 15-16, 2014.

[23] S. K. Yadav, Bharadwaj, B., & Pal, S., "Data mining applications: A comparative study for predicting student's performance", vol. 202.4815, arXiv preprint arXiv, 2012.

[24] R. Naqvi, "Data Mining in Educational Settings," *Pakistan Journal of Engineering,* vol. 1, no. 4(2), p. 201.4615, 2015.

[25] B. R.S, "Data Mining for Education," *International Encyclopedia of Education,* vol. 1, no. 1, 2011.

[26] R. Pressman, *"Software Engineering: A Practitioner's Approach,"* vol. 1, no. 1, 2005.

[27] C. P., "Data Mining: A Knowledge Discovery Approach", New York, 2000.

[28] J. Han, Kamber, M. and Pei, J., "Data Mining*: Concepts and Techniques, third Edition ed". 225 Wyman Street, Waltham, USA: Morgan Kaufmann Publishers is an imprint of Elsevier, 2011.*

[29] H. Tesfahun, "Application of Data Mining For Predicting Adult Mortality," Master Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2012.

[30] S. Institute, *SAS Enterprise* Minor SEMMA, 2016.

[31] R. Agrawal, "Mining Association Rules between Sets of Items in Large Databases.,"in *In Proceedings of SIGMOD, 20716, 1993.According to (Sutch, T. (2015). Using association*

rules to understand subject choice at AS/A level, 2015.

[32] K. Charly, "Data Mining for the Enterprise," *31st Annual Hawaii Int. Conf. on System Sciences,* vol. 7, pp. 295-304, 1998.

[33] G. Awoke, "Predicting HIV Infection Risk Factor using Voluntary Counseling and Testing Data," Addis Ababa University , Addis Ababa, Ethiopia, 2012.

[34] A.Olani, "Predicting First year University Students' Academic Success," *Institute for Educational Research, 2008.*

[35] S. V.Sathiya and N. Dr.Sai Satya, "Data Mining Tasks Performed By Temporal Sequential Pattern," *International Journal of Research and Computational Technology,* vol. 2, no. 3, pp. 1-6, 2012.

[36] H. Jiawei, K. Micheline and P. Jian,, "Data Mining Concepts and Techniques," Morgan Kaufmann Publishers is an imprint of Elsevier, New York, 2012.

[37] S.O., Danso "An Exploration of Classification prediction techniques in data mining the insurance domain," Bournemouth university. Bournemouth, 2006.

[38] M. Vincent, "Introducing a data mining process framework to enable consultants to determine effective data analytics tasks," Master Thesis, University of Technology, Delft, 2012.

[39] C. d. R. Bruno and T. d. S. J. Rafael, "Identifying Bank Frauds using CRISP-DM and Decision Tree," *International journal of computer science & information Technology (IJCSIT),* vol. 2, no. 5, pp. 162-169, 2010.

[40] P. Thair Nu, "Survey of Classification Techniques in Data mining," *Survey of Classification Techniques in Data mining,* vol. 18, no. 20, pp. 978-988., 2009.

[41] Q. J., "C4.5 Programs for Machine Learning", Los Altos: Morgan Kaufmann, 1993.

[42] M. K. J.B., Data Mining-Concepts, Models, Methods, and Algorithms, John Wiley, USA: Sons Publication Inc, 2003.

[43] T. T. R. a. F. J. Hastie, "the Elements of Statistical Learning" (2nd ed.). pp., 1- 764, 2008, Second Edition ed., Springer. USA, 2008.

[44] G., J. D. J. Z. Calivn, “Mastering Data Mining and Scince of Customer Relationship Management," *A computer- Based herbsicide Injury Diagnostic Expert System. Weed Technology, vol,* vol. 19, no. 7, pp. 486-491, 2005.

[45] S. Singh, "classification of students’ data using data mining techniques for training and

placement department in technical education," *International Journal of Computer Science and Network (IJCSN),* vol. 1, no. 4, 2012.

[46] A. A. G. Aditi Mahajan, "Performance Evaluation of Rule Based Classification Algorithms," *International Journal of Advanced Research in Computer Engineering & Technology* , vol. 3, no. 10, pp. 12-19, 2014.

[47] C. G. Carrier and O. Povel, "Characterising data mining software," *Intelligent Data Analysis,* vol. 7, pp. 181 - 192., 2003.

[48] J. Luan, "Data mining and knowledge management in higher education potential applications." *the association, for institutional research,* vol. 1, no. 1, 2002.

[49] A. B. Michael J. and S. L. Gordon, "Data Mining Techniques for Marketing, Sales, and Customer Relationship Management", Second ed., Indianapolis, Indiana: Wiley Publishing, Inc., 2004.

[50] A.Meseret ,"A Combined Reasoning System For Knowledge Based Network Intrusion Detection," Master thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2016.

[51] E. W. T. Ngai, L. Xiu and D. C. K. Chau, "Application of data mining techniques in customer relationship management," *A literature review and classification. Expert Systems with Applications,* vol. 36, pp. 2592-2602, 2009.

[52] V. Kumar, "Mining Association Rules in Student's Assessment Data," *International Journal of Computer Science Issues*, vol. 9, no. 5, 2012.

[53] A. Mohammed, "Towards Integrating Data Mining with Knowledge Based System: The Case of Network Intrusion detection," M.Sc. Thesis, Addis Ababa University, 2013.

[54] T .Verbraken, Bravo, C., Weber, R., & Baesens, B. , "Development and application of consumer credit scoring models using profit-based classification measures," *European Journal of Operational Research,* vol. 238(2), pp.505-513, (2014).

[55] B. Max, *Principles of Data Mining,* UK: Portsmouth Springer, 2007.

[56] M .Weiss, Sholom, Zhang, Tong. "Performance analysis and evaluation". InYe Nong, editor. The Hand book of data mining. New Jeresy, USA: Lawerence Erlbaum Associates Inc., 2003.

[57] Witten, I. H. and Frank, E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco, CA, Morgan Kaufman, 2005.

[58] A. Muluken, "Application of Data Mining Techniques for Student Success and Failure

Prediction," *International Journal of Scientific & Technology Research*, vol. 4, no. 4, 2015.

[59] S. Pal, "Mining Educational Data Using Classification to Decrease Dropout Rate of Students*," International Journal of Multidisciplinary Sciences and Engineering,* vol. 3(5), p. 35–39. (2012).

[60] T.M. M. A., & El-Halees, A. M., "Mining Educational Data to Improve Students 'Performance: A Case Study, *"International Journal of Information and Communication Technology Research,* vol. 2(2), p. 140–146, (2012).

[61] A.A. S. Saleh, "Education Data Mining to Predict Student Exam Grades in Vocational Institutes," British university, dubi, 2015.

[62] T. Tariku, "Identifying Determinant Factors for Students' Success in Preparatory Schools Using Data Mining Techniques," Adiss Ababa University, Adiss Ababa, Ethiopia, 2017.

[63] B. Abdellah, "Early prediction of new entrant students' academic Performance in higher education institutions for natural science related streams," Debre Birhan University, Debre Birhan, 2016.

[64] W. Solomon, "A Correlation Study of Students' Performance in Ethiopian," Addis Ababa University, Addis Ababa, 2016.

[65] B.Yadav, "Mining Educational data to predict students retention" *IJCSIS,* vol.10, no. 2, pp.19-34, 2012.

[66] H. Esfahan, "Application of Data Mining For Predicting Adult Mortality," Master Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2012.

[67] S. Sembiring, M. Zarlis, D. Hartama, & E. Wani, "Prediction of student academic performance by an application of data mining techniques", *2011 International Conference on Management and Artificial Intelligence, 6 (2011). 110–114.*

[68] S. Huang, & N. Fang, Work in Progress - "Prediction of Students' Academic Performance in an Introductory Engineering Course", *In 41st ASEE/IEEE Frontiers in Education Conference, (2011), 11–13. http://dx.doi.org/10.1109/fie.2011.6142729*

[69] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms". New York: John Wiley & Sons, 2003.

[70] Milley, A. "Healthcare and data mining. Health Management Technology", 2000. 21(8), 44-47.

[71] F.T., "Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data

mining," *International Conference on Information Acquisition,* vol. 1, no. 1, 2006.

[72] R. S. J. Baker, Data Mining for Education, Advantages Relative to Traditional Educational Research Paradigms, (2010).

[73] Y. Chen, "Learning classifiers from imbalanced", only positive and unlabeled data Sets, USA: Department of Computer Science, Iowa State University, 2009.

[74] J. Han, "Data Mining concepts and techniques", Illinios: Elsevier Inc., 2012.

[75] A. Shanavas, "An Analysis of students' performance using classification algorithms," *Journal of Computer Engineering,* vol. 16, no. 1, 2014.

[76] . K. Dr. D. Ashok and G. R., "Performance and Evaluation of Classification Data Mining Techniques in Diabetes," *International Journal of Computer Science and Information Technologies,* vol. 6, no. 2, pp. 1312-1319, 2015.

[77] Hassan M., Ghaziri Elias and M.Awad, "Google Book," [online].

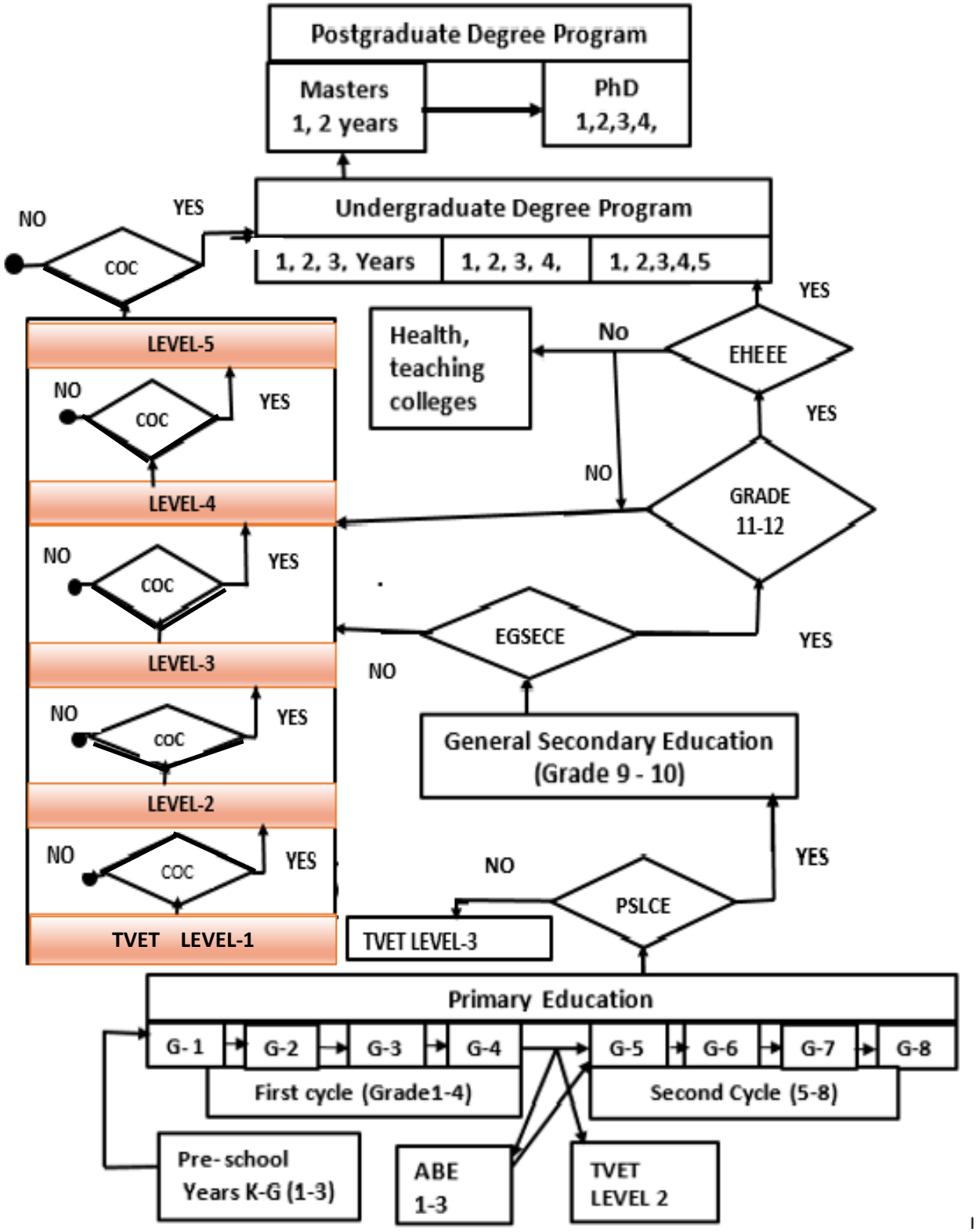
[78] T. Betelhem, "Insulin Resistance and Dyslipidemia in Type 2Diabetic Patients: A Cross-Sectional Study at the Diabetic Clinic of Tikur Anbessa Specialized Teaching Hospital," Master thesis addis Ababa University, Addis Ababa, Ethiopia, 2015.

[79] G. Solomon, "A Self-Learning Knowledge Based System for Diagnosis and Treatment of Diabetes," Master's Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2013.

[80] A. Tagel, "Knowledge Based System for Pre-Medical Triage Treatment at Adama hospital," Master's Thesis, Addis Ababa university, Addis Ababa, Ethiopia, 2013.

Appendix I

Educational Structure in Ethiopia



Appendix II

10-Fold Cross-Validation Sample Rule Generation by PART Algorithm

IFECAPG11 = VGOOD AND MCAPG10 = EXCELLENT AND PNEGG10 = VGOOD **THEN** VGOOD (9.0)
IFCICAPG12 = VGOOD AND PCAPG12 = VGOOD AND ENEGG10 = EXCELLENT **THEN** VGOOD (20.0/1.0)
IFCICAPG12 = VGOOD AND PCAPG12 = FAIR AND MCAPG10 = POOR AND Age = AGE-1 **THEN** SATISFACTORY (12.0)
IFECAPG11 = SATISFACTORY AND PCAPG12 = SATISFACTORY AND MNEGG10 = VGOOD **THEN** GOOD (29.0)
IFECAPG11 = VGOOD AND MCAPG10 = POOR AND CICAPG11 = SATISFACTORY **THEN** GOOD (8.0)
IFECAPG11 = VGOOD AND MCAPG10 = SATISFACTORY AND PCAPG12 = VGOOD AND CICAPG11 = SATISFACTORY **THEN** VGOOD (16.0)
IFECAPG11 = VGOOD AND MCAPG10 = SATISFACTORY AND CICAPG11 = FAIR AND CICAPG12 = EXCELLENT **THEN** GOOD (24.0)
IFECAPG11 = VGOOD AND MCAPG10 = SATISFACTORY AND CICAPG11 = SATISFACTORY AND CHCAPG11 = VGOOD **THEN** GOOD (36.0/2.0)
IFMCAPG10 = SATISFACTORY AND CICAPG11 = VGOOD AND ECAPG11 = EXCELLENT AND CICAPG12 = EXCELLENT AND MNEGG10 = VGOOD **THEN** VGOOD (30.0)
IFMCAPG10 = SATISFACTORY AND PCAPG12 = VGOOD AND ECAPG9 = SATISFACTORY **THEN** VGOOD (37.0)
IFMCAPG10 = SATISFACTORY AND CICAPG11 = FAIR AND Sex = M **THEN** GOOD (4.0)
IFMCAPG10 = VGOOD AND CICAPG10 = EXCELLENT AND BCAPG10 = EXCELLENT AND Age = AGE-1 AND MCAPG9 = EXCELLENT **THEN** EXCELLENT (62.0/1.0)
IFMCAPG10 = VGOOD AND CICAPG11 = VGOOD **THEN** VGOOD (53.0)
IFMCAPG10 = SATISFACTORY AND CICAPG11 = SATISFACTORY AND ECAPG12 = FAIR **THEN** GOOD (17.0)
IFMCAPG10 = EXCELLENT AND ECAPG11 = EXCELLENT AND ECAPG9 = EXCELLENT **THEN** EXCELLENT (20.0)
IFCICAPG12 = VGOOD AND ECAPG9 = SATISFACTORY AND Age = AGE-1 **THEN** GOOD (16.0/1.0)
IFECAPG11 = POOR AND MCAPG12 = FAIR AND BCAPG10 = POOR **THEN** FAIL (99.0/1.0)
IFMCAPG10=FAIR AND ECAPG11 = POOR AND CICAPG11 = POOR AND CHCAPG11 = FAIR: FAIL (15.0

Appendix III

System Performance Testing Questions

From fifty test cases, some sample of test cases are provided for the prototype developed to test the system performance from dataset Grade 12, 2010 E.C EHEEE and the corresponding Grade 9, Grade 10, Grade 11, Grade12 common subjects Class Average Point as well as Grade 10 EGSECE result to compare the performances of each students against the performances predicted by the system as Excellent, Very Good, Good, Satisfactory and Fail for Grade 12, 2010 E.C EHEEE.

IFGrade 11 English = satisfactory and Grade 11 Chemistry = vgood and Grade 11 physics = satisfactory **THEN** ----------?
IFGrade 11 English = vgood and Grade 10 maths = vgood and Grade 11civics = excellent and Grade 12 maths = poor and EGSECE English = excellent **THEN** ----------?
IFGrade 10 maths = satisfactory and Grade 12 physics = vgood and Grade 9 English = satisfactory **THEN**----------?
IFGrade 11 English = vgood and Grade 10 maths = vgood and Grade 12 maths THEN ----------?
IFGrade 12 Civics = excellent and Grade11 Civics = excellent and Grade 11 Chemistry = satisfactory **THEN** ----------?
Grade 11 English = vgood and Grade 10 maths = vgood and EGSECE Biology **THEN**----------?
IFGrade 11 English = vgood and Grade 10 maths = fair and Grade 11 Chemistry = vgood and EGSECE civics = vgood and Grade 9 physics = satisfactory **THEN** ----------?
IFGrade 11 English = vgood and Grade 10 maths = satisfactory and Grade 11 Civics = satisfactory and Grade 11 Chemistry = satisfactory **THEN** ----------?
IFGrade 12 Civics = excellent and Grade 10 maths = satisfactory and Grade 11Civics = satisfactory and Grade 9 Civics = vgood **THEN** ----------?
IFGrade 12 Civics = excellent and Grade 10 maths = satisfactory and Grade 11 Civics = vgood and Grade 11 English = excellent and Grade 12 Biology = vgood **THEN** ----------?

Appendix IV

Domain expert evaluation form/query

Dear Evaluator,

The importance of this evaluation form is to evaluate to what extent the prototype system is usable by the end-users in the domain area. Therefore, you are kindly requested to evaluate the system by labeling (X) symbol on the space provided for the corresponding attributes values for each criteria of evaluation.

I would like to appreciate your collaboration in providing the information. Note: - the values for all attributes in the table are related as excellent=5, Very good=4, Good=3, Fair=2, and Poor=1.

No	Criteria of evaluation	poor	Fair	Good	Very Good	Excellent	Average
1	Simplicity to use and interact with the system						
2	Attractiveness of the system						
3	Efficiency in time						
4	The accuracy of the system in reaching a decision to identify the types of Diabetes						
5	The ability of the system to make right conclusion and recommendation						
6	Importance of the KBS in the domain area						