ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

Identifying Determinant Factors for Students' Success in Preparatory Schools Using Data Mining Techniques

By
Tariku Teklu

June 2017
Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Identifying Determinant Factors for Students' Success in Preparatory Schools Using Data Mining Techniques

A Thesis Submitted to the School of Information Science of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

Tariku Teklu

June 2017

ADDIS ABABA UNIVERSITY

SCHOOL OF

INFORMATION SCIENCE

Identifying Determinant Factors for Students' Success in Preparatory Schools Using Data Mining Techniques

By

Tariku Teklu

June 2017

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
_____	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

Declaration

I hereby declare that IDENTIFYING DETERMININANT FACTORS FOR STUDENTS' SUCCESS IN PREPARATORY SCHOOLS USING DATA MINING TECHNIQUES 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.

TARIKU TEKLU

JUNE 2017

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

MILLION MESHESHA (PhD)

JUNE 2017

Acknowledgement

First and for most I would like to thank Almighty GOD for giving me strength, courage and patience in order to accomplish this research. I would like to express my heartfelt thanks to my advisor Dr.Million Meshesha for his support and advice starting from the research work and for his constructive comments.

Next my deepest gratitude goes to National Educational Assessment and Examination Agency which sponsored me for the education and providing me the necessary data; Ato Yoseph Abera and all of his staff for their support, understanding and patience through the years of my education.

My special thanks goes to my friend Meseret Getachew for her advice and support through my research.

Finally, I would like to thank my families and my friends who believed in me and gave me support and advice.

List of Acronyms

CRISP – DM – Cross –Industry Standard Process for Data Mining

DSQ - Disqualify

EDM – Educational Data Mining

EGSECE – Ethiopian General Secondary Education Certificate Examination

EHEECE – Ethiopian Higher Education Entrance Certificate Examination

EHEEE - Ethiopian Higher Education Entrance Examination

GPA – Grade Point Average

KDD – Knowledge Discovery in Database

KDP – Knowledge Discovery Process

MoE – Ethiopian Ministry of Education

NEAEA – National Educational Assessment and Examination Agency

NG – Night Government

NNG – Night Non-Government

NPD – National Pupil Database

PR - Private

RG – Regular Government

RNG – Regular Non-Government

SEMMA – Sample, Explore, Modify, Model, Assess

SNNP – Southern Nations, Nationalities and People

UEE – University Entrance Examination

Contents

List of Tables

List of Figures

Abstract

Data from Educational Assessment and Examination Agency shows that from Addis Ababa region who took Ethiopian Higher Education Entrance examination in 2006, 2007 and 2008 E.C; 64.1 %, 54.4 %, and 42.4% respectively scores less than half of the total score. Even though the percentage decreases from 2006 E.C to 2008 E.C.by 21.7%, many students fail to score the expected half of the total score.

The study aims to apply data mining for identifying the determinant factors for the students' success in the preparatory schools to join higher education. The study focused on Addis Ababa region Natural Science stream preparatory Schools' students. Based on this, the data collected from National Educational Assessment and Examination Agency is only Addis Ababa examinees EHEEE data and correspondingly their EGSECE data. The collected dataset covers three years of data from the three years data from 2006 to 2008 E.C EHEEE.

The study uses Hybrid data mining model since it is a research oriented model and WEKA 3.8.0, Microsoft Excel 2013 and KU tools are used for data mining, for data integration and for data exploration respectively. Finally, 40328 instances and 15 attributes are selected for analysis. Additionally, 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. Association rule mining method such as Apriori and Filter Associator algorithm compared and Apriori algorithm is 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.

As a result, the study showed that scoring Very good in Physics, Civics and Biology subjects in EHEEE are determinant factors for the students 'success in the preparatory schools. Similarly scoring good in English in EHEEE is also another determinant factor. Besides, the study revealed that Regular Non-Government preparatory students are more associated with success to enter higher education than Government preparatory school students and Sub city of schools that students attend have no influence on students' success to enter higher education.

But in using Apriori algorithm, there is no standard way of setting different thresholds. This leads to missing the strong rules.

CHAPTER ONE

INTRODUCTION

1.1. Background of the study

Education plays a great role in achieving one's country growth and development. The current Ethiopian government recognizes the importance of education for national development. Policy is mainly aimed at expanding the education sector, improving quality and ensuring that educational content is harmonized with the country's economic needs [1]. In order to achieve the intended goals, the quality of the education needs to be in a right path. Even though the quality of education has many aspects and participants, the main participants are students. The quality of the education can be seen from the side of success or failure of students. Identifying the determinant factors for the students' success can help in making different and timely managerial decisions in order to improve the failure [2].

Ethiopian education vision focusses 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, each of the 9 National Regional States and the 2 City Administrations has its bureaus of education responsible for administrating and managing the educational system. Within each of these exists a network of management structure involving Zonal Educational Departments and Woreda Education offices. The Woreda is the smallest educational authority responsible for all educational institutions in its territory [3].

In Ethiopia there has been a dramatic increase in admission of preparatory secondary education students which are grade 11 and 12 .In 2002 E.C. the number of students who attend preparatory secondary education were 243,080;But in 2007 E.C. the number of students who admit preparatory secondary education become 425,774 which shows large number of increase since the target of admission to achieve in 2007 E.C. was 360,000 [4].This dramatic increasing number of students in each year which leads to the increased records and data concerning students, invites to make analysis based on the data in order to determine the students' success as well as failure

characteristics and helpful to identify the predetermined factors of students' success .On this end ,it is essential to find another data analysis mechanisms rather than focusing on traditional ways of data analysis. One of the concept which is helpful in this case is data mining. The concept which is applicable in different sectors for instance in banking, in retail sales, in bioinformatics and in telecommunications; starts to get an attention from educational sector. Data mining 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 techniques or data mining in education, called educational data mining, EDM [5].

EDM is one of application of data mining in educational environment [6].Additionally, EDM is a new growing research and emerging discipline, concerned with data from academic field to develop various methods and to identify unique patterns which will help to explore students' academic performance [7] [6].Besides, Educational Data Mining focuses on applying data mining tools and techniques to educationally related data [5].

1.2. Statement of the problem

At the end of the first cycle of secondary education which is grade 10(general secondary education), students take a national examination (Ethiopian General Secondary Education Certificate Examination) that has a purpose to certify completion of general secondary education which has an aim of selecting students for the next higher level education that is preparatory level The preparatory level is the second cycle of secondary education and prepare students for university education.

Now in Ethiopia education, grade 11 and 12 are the preparatory classes for higher education. Data shows that in 2007 E.C, 206,472 grade 12 students took Ethiopian Higher Education Entrance Examination EHEEE. From the students who took an exam 102,980 which is 49.87% joins undergraduate program in higher education [4]. In this case 50 %of the students failed to join higher education. Similarly, Data from Educational Assessment and Examination Agency shows that from Addis Ababa region who took Ethiopian Higher Education Entrance examination in 2008 E.C,42.4 % scores less than half of the total score which is below 350.Similarily from examinees of 2007 E.C ,54.4 % scores less than the half score. From 2006 E.C examinees, also, 64.1 % of students score less than half score. Even though the percentage decreases from 2006 E.C to 2008

E.C, many students fail to score the half of the total score [12].This result initiates the researcher to identify the reasons of preparatory school students' low performance. Besides it initiates to identify the factors which determines the success or failure of preparatory school students; since at present, the existing secondary curriculum is primarily designed to prepare students for university studies; if it is retained, it will not only fail students, it may also fail the country's aspirations for middle-income status [13].To identify the determinant factors, EGSECE and EHEEE data will be used as a means. The result of EGSECE, which students took at the end of grade 10 determines whether a student joins preparatory school or not. Moreover, there are common subjects which students took in grade 10 and 12.For instance Natural science stream students will take an examination of English, Math's, Physics ,Chemistry ,Biology and Civics & ethical education which are similar to grade 10 classes. Many researches in educational data mining focused on predicting students' performance. A study which focus on evaluation and prediction of students' performance in high school; tries to present superior models in predicting student's performance. At the result individual, environmental and educational factors affecting successful and unsuccessful students have been analyzed and according to them efficient models based on decision tree models have been presented [14].

Besides, there is a case study on predicting performance of students at the end of a university degree at an early stage of the degree program, to help universities in to identify students with low academic performance and to support them. The results show that it is possible to predict the graduation performance in 4^{th} year at university using only pre university marks and marks of 1^{st} and 2^{nd} year courses, no socio economic and demographic features, with a reasonable accuracy [15].

Similarly, researches in educational data mining in Ethiopia focused on predicting students' performance as well as gave emphasis in predicting success and failure using different data sets. Ethiopian Higher Education Entrance Examination score ,different subjects score ,preparatory school transcript average point used as a data set [16], [2].Besides the researches include other factors rather than academic factors such as sex, number of students in a class ,number of courses in a semester [17].

This research tries to focus on finding determinant factors for students' success in the preparatory schools using the method of description data mining with EGSECE and EHEEE score of common subjects and result data.

To this end, the research attempts to answer the following questions:

- What are the possible techniques to pre –process and prepare the raw educational data?
- Which data mining algorithm can be more suitable for the purpose of identifying determinant factors for the students' success?
- What are the possible determinant factors for the students' success in the preparatory schools?

1.3. Objective of the study

The general and specific objectives of this research are listed as follows:

1.3.1. General objectives

The general objective of this research is to identify determinant factors for the students' success in the preparatory schools.

1.3.2 .Specific objectives

To achieve the general objective, the following specific objectives are formulated.

- To use data mining methodologies to identify the factor of success in their subjects.
- To discover the hidden knowledge and patterns using the data mining techniques.
- To evaluate the performance of the discovered association rule.

1.4. Scope and Limitation of the study

The study mainly focuses on to identify the determinant factors for the success of preparatory school students in Addis Ababa region using three years of EGSECE data and EHEEE data of 2008 ,2007 and 2006 E.C. The corresponding EGSECE data will be 2006, 2005 and 2004 E.C respectively. Three years of data used in order to participate different years of results which is helpful in analyzing. Besides, Association Rules Discovery techniques is used in order to compare the students' score in the subjects common at grade 10 and grade 12 level and identify the factors

of their success. Besides, the discovered knowledge can be addressed to the domain expert by using documents only. Only two algorithms are compared to select the algorithm for the analysis.

1.5. Research Methodology

In this study, the researcher uses the data mining approach on EGSECE and EHEEE data. WEKA software is be used for the study implementation since it is freely available and widely used for research purposes in data mining.

In data mining the major standard models are CRISP-DM , KDD, Hybrid model [7].

1.5.1. Study Design

The study follows the six step Hybrid process model [18] ,which is adapted from CRISP – DM model ,since it provides a more general ,research - oriented description of the steps .Besides it introduces the data mining steps instead of the modeling step [19].The six steps are Understanding the Problem Domain ,Understanding the Data, Preparation of the Data ,Data Mining ,Evaluation of the Discovered Knowledge and Use of the Discovered Knowledge.

1.5.2. Understanding the problem domain

It is an initial step for the hybrid model and the main task in this research which accomplished by using different mechanisms. Review of documents including different manuals; policy documents, for instance reviewing the Ethiopian Education policy has an impact in understanding the problem domain and different reports; the reports which reviewed include education statistics abstract which is organized by Ministry of Education and similarly National Educational Assessment and Examination Agency's reports which contain different analysis of national exams has an impact in understanding the problem domain. Discussion with domain experts, in this case discussion with high school and preparatory school teachers concerning the overall result of grade 10 and grade 12 students has an impact in understanding the problem domain. Similarly, in 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 the two bodies to understand the problem domain. In the same way, discussion with the main participants, in this case, preparatory school students about the overall determinant factors of their success. In this case

the researcher became in a good position in order to define the problem, determine the domain objectives, to assess the problem and to determine the data mining goals, which are helpful for the next steps.

1.5.3. Understanding the data

This step includes collecting the initial students' data from National Educational Assessment and Examination Agency database. The data can be visualize using excel format. The visualization focuses on insuring the data completeness since there are students' records which has incomplete data; the fact that the students can face an absent scenario in some subjects. Similarly, the visualization can be focused on cross checking the suitability of the data concerning the data mining goals. Besides exploring the data and describe the data to identify the possible attributes is another task of understanding the data. Additionally, this step enables to determine the data mining methods and algorithms. Moreover, WEKA is another data visualization tool, which uses in describing the attributes.

1.5.4 .Preparation of the data

This step include all activates that are needed to construct the final data set. The data which is taken from National educational assessment and examination agency database preprocessed and cleaned for the application of different data mining techniques. The Ethiopian general secondary education certificate examination (EGSECE) result and the Ethiopian Higher Education Entrance Examination (EHEEE) result data contain many missing values due to the absence of students from the examination. Besides there are incomplete data which needs to be filled with values and there is a requirement of integration of result tables since the grade 10 (EGSECE) and grade 12(EHEEE) data are in a separate data table. The tools used for preprocessing are Microsoft excel, KU tools.

1.5.5 .Data Mining

This step is application of the selected data mining methods to the prepared examination result data and testing the generating rules whether they achieve the required minimum threshold. Beside it is a step of finding hidden, non – trivial and previously unknown information from the data. In this study, association rule mining approach is applicable on the processed data. WEKA is used for mining the data.

In association rule two phases are usually mentioned; finding all the frequent item sets that have support s above a predetermined minimum threshold and generate strong association rule from the frequent item sets that have a confidence c above a predetermined minimum threshold [20] .In this study the item sets are the subjects that students take in the examination of grade 10 and grade 12.

1.5.6. Evaluation of the Discovered Knowledge

This step include interpretation of the results ,cross checking the results and observing the interestingness and relationship of the discovered knowledge, review the process and another means of discovering knowledge can be assessed. Based on the review another step can be determined. Besides based on the discovered results there is a discussion with domain experts; in this case the domain experts are the data analysts from National Educational Assessment and Examination Agency and the preparatory school teachers. The discovered knowledge can be divided in to three parts namely expected and previously known which are rules that confirms user beliefs and can be used to validate the initial approach; the other is unexpected that contradicts user beliefs and which needs further investigation for its interestingness and the need for taking an action. The third one is unknown that doesn't clearly belong to any category and which needs domain specific experts to categorize [21].To evaluate the interestingness of association rules, the researcher uses measure based on correlation or lift.

Correlation (A, B) or lift = P(AUB)/P(A)P(B),where A and B are item sets .

If the correlation or lift value is less than 1, then the occurrence of A is negatively correlated with the occurrence of B.

If the correlation or lift value is greater than 1, A and B are positively correlated. This implies the occurrence of one implies (promotes) the occurrence of the other.

Similarly, if the correlation or lift value is 1, A and B are independent. This implies there is no correlation between the items.

1.6. Significance of the study

The outcome of this research will have a great contribution for different stakeholders for different decision making purposes. Besides it will have an advantage to show the factors which have an influence on the success of preparatory school students and help in giving timely managerial decisions.

Furthermore, it will guide to improve the failure of preparatory school students since it points out the factors. And will help students with problems so as to be a successful student in higher education.

Besides, similar to other assessment methods, the educational data mining approach will be a new insight in educational environment to show the solutions for many research problems.

1.7. Ethical clearance

The confidentiality of the data will be maintained. Any student names and known IDs will be eliminated and substitute by other unknown IDs. Besides, the research is intended only for academic purpose which is for Master's Thesis for the partial fulfillment of M.Sc degree in Information Science.

1.8. Thesis Organization

The thesis is divided in to six chapters; the first chapter is about introduction, which contain background about education and related concepts, statement of the problem, objectives, scope and limitation and methodology.

The second chapter reviews data mining as a technology and techniques &applications, educational data mining, and reviewing of association rule mining and related works.

The third chapter is about research design and methods.

The fourth chapter highlights the Data preparation tasks done in this study, which includes problem domain understanding, data understanding and data pre-processing.

The fifth chapter is discussion about the results of the analysis and the rules discovered and evaluation of the discovered knowledge.

The sixth chapter provides summary, concluding remarks and recommendation of the current study.

CHAPTER TWO

LITERATURE REVIEW

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 mention records, files, documents, images, sounds, videos are some of the formats that data are 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. 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 [22].

2.1. Overview of Data Mining

The emerging of digital data acquisition and the improvement of storage technology has resulted in the growth of huge data bases. This can be seen in different sectors ;for instance supermarket transactional data, telephone call details ,different governmental statistics ,credit card records, different medical records [23].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; unless they are no value without extracting efficiently in order to get information from them [24].This increase the demand of new techniques and tools which is helpful in analyzing the huge data for information and knowledge .This leads to the concept of data mining ;Data mining is often set in the broader context of knowledge discovery in data bases ,or KDD. The KDD process involves several stages; selecting the target data, preprocessing the data, transform the data, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures [18].

2.1.1. What is data mining?

Data Mining is a concept which has different meaning concerning different researchers. *In* [25], Data Mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It has been defined as the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. In this definition the term process implies there are many steps involving data preparation ,search for patterns and knowledge evaluation; the process is non- trivial in that it involves search for structure ,models ,patterns or parameters .Similarly ,the discovered patterns should be valid for the new data with some degree of certainty .besides ,the patterns should be novel for the user and potentially useful for the user and the task.at last the patterns should be understandable if not immediately ,then after some post processing [25] .However ,data mining includes different techniques from various fields including database ,machine learning ,statistics ,data visualization and others [26].Similarly in [7],**D**ata mining is comprehensively defined as "the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques". Other researcher in [27] view data mining as "the process of automatically extracting useful information and relationships from immense quantities of data. In its purest form, data mining doesn't involve looking for specific information. Rather than starting from a question or a hypothesis, data mining simply finds patterns that are already present in the data."

Figure 1. Overview of the steps constituting the KDD process

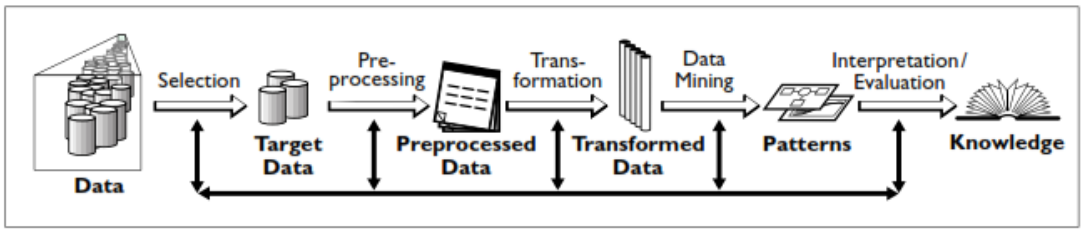

Figure 2.1.The steps constituting the KDD process

Additionally other researchers on [28] pointed out the concept data mining is a multi-disciplinary field include areas like information technology ,machine learning ,statistics ,pattern recognition ,data retrieval ,neural network ,information based system ,artificial intelligence and data visualization .

Concerning the increase in the amount of data stored, there has been a need of better, cheaper and faster way of data handling which is data mining [18] .The main difference of data mining from other approaches is that data mining is data driven while others are model driven. For instance, observing statistics and data mining; statisticians concerned with the problem of finding the smallest data size that gives sufficiently confident estimates. While in data mining, the emphasis is on large data size and the interest falls on building a data model which is small but that describes the data well [18]. Similarly, the complexity of the data and the expected results are another points which distinguish data mining from other related areas. Data mining deals with real world problems so that needs real world data which can be extracted from large data bases. The data need to be cleaned and ready for suitable format in order to make the extraction appropriate. Besides the data mining deals with different data behaviors like noise, different formats, redundant information and missing values and attributes [29] .

2.2. 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) [30].The data mining process models can be considered as a methodology to support the processes which leads to find the information and knowledge. The reason for using the process models is in order to organize the knowledge discovery and data mining projects within a common frame work. Besides the process models are helpful to understand the knowledge discovery process and provide a roadmap while planning and carrying out the projects [31].

Additionally ,the reasons of using the process models which are mentioned in a research study that to ensure that the end product will be useful for the users [25], the other reason which is pointed out by [32] is that to understand the process itself and to understand the concerns and need of the end users .end users usually lack perception of large amounts of untapped and potentially valuable data .besides they are not ready to devote time and resources toward formal methods of knowledge seeking .another reason of a need of data mining process models is mentioned by [31] is that providing support for managerial processes.

The process models are broadly divided in to academic research models and industrial models.

2.2.1. Industrial Models

In many researches the CRISP -DM and the five step model by Cabena et al. fall in this category [18] .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 consortiums. The industrial six-step CRISP-DM model, developed by a large consortium of European companies has become the leading industrial model. The CRISP-DM model has been used in domains such as medicine, engineering, marketing, and sale [25]

The CRISP-DM (Cross-Industry Standard Process for Data Mining) consists of the following six steps.

- Business understanding. This step focuses on the understanding of objectives and requirements from a business perspective. It also converts these into a data mining problem definition, and designs a preliminary project plan to achieve the objectives. It is further sub divided in to :
    - determination of business objectives,
    - assessment of the situation,
    - determination of Data Mining goals, and
    - Generation of a project plan.
- 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.
- Data preparation. This step covers all activities needed to construct the final dataset, which constitutes the data that will be fed into DM tool(s) in the next step. It includes Table, record, and attribute selection; data cleaning; construction of new attributes; and transformation of data. It is divided into
    - selection of data,

- cleansing of data,
- construction of data,
- integration of data, and
- Formatting of data.

- Modeling. At this step, various modeling techniques are selected and applied. Modeling usually involves the use of several methods for the same 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.
- Evaluation. After one or more models have been built that have high quality from a data analysis perspective, the model is 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 DM results should be reached. The key sub steps in this step include
    - evaluation of the results,
    - process review, and
    - Determination of the next step.
- 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 KDP. This step is further divided into
    - plan deployment,
    - plan monitoring and maintenance,
    - generation of final report, and
    - Review of the process sub steps.

Notes: The model is characterized by an easy-to-understand vocabulary and good documentation. It divides all steps into sub steps that provide all necessary details. It also acknowledges the strong iterative nature of the process, with loops between several of the steps. In general, it is a very successful and extensively applied model, mainly due to its grounding in practical, industrial, real-world knowledge discovery experience [41].

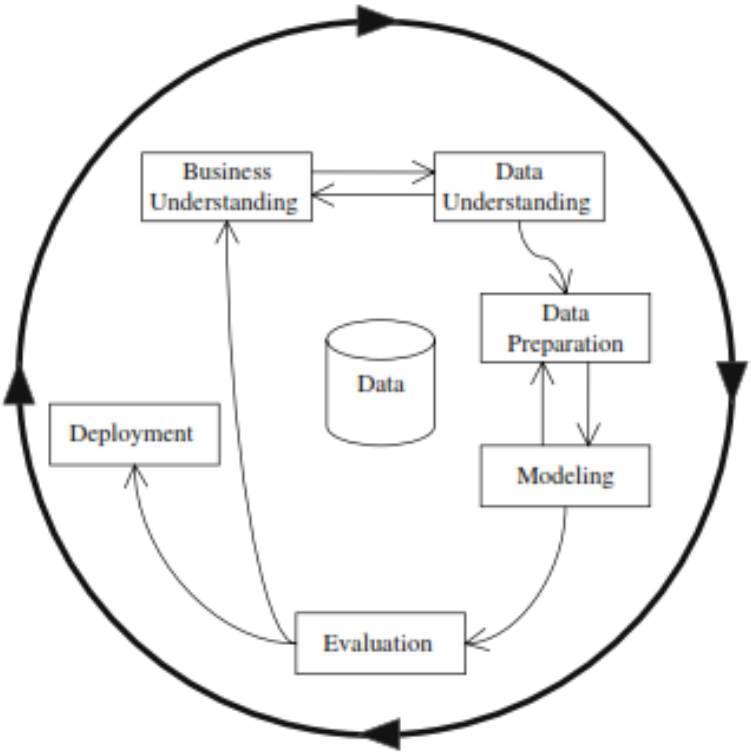

Figure 2.2. The CRISP – DM KD Process model

2.2.2. Academic research models

In the model the focus point was to provide a sequence of activates that would help to execute a KDP in an arbitrary domain. For instance the well-known model which is developed in 1996 consists of nine steps [31]. The model consists the following nine steps.

The first step is developing and understanding the application domain. This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge. The second step which is creating a target data set. Here the 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.Data cleaning and preprocessing is the third step. This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes. The fourth step is data reduction and projection. This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data. In choosing the data mining task step, the data miner matches the goals defined in Step 1 with a particular DM method, such as classification, regression, clustering, etc.in the next step the 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 other steps are data mining, interpreting mined patterns and consolidating discovered knowledge.

2.3. Hybrid Models

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 is a six-step KDP model developed by [18]. It was developed based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include

– providing 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 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.

The following are the description of the six steps.

- 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. A description of the problem, including its restrictions, is prepared. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

- 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 DM goals.
- Preparation of the data. This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in Step 1.
- Data mining. Here the data miner uses various DM methods to derive knowledge from preprocessed data.
- Evaluation of the discovered knowledge. Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. 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.
- 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 [19].

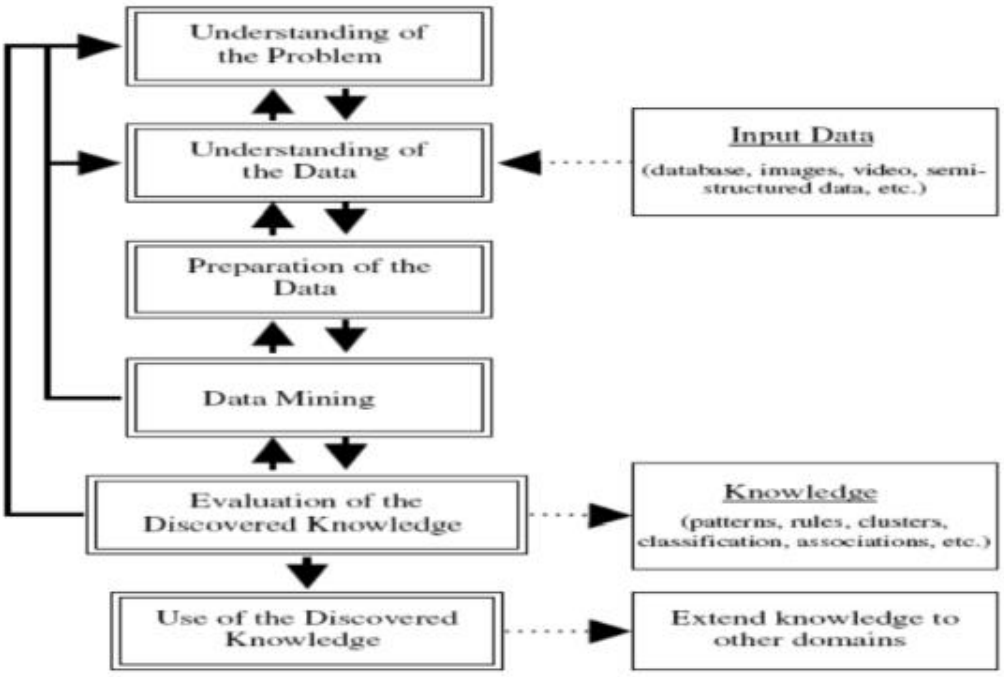

Figure 2.3 The six steps of Hybrid data mining model

2.4. Data Mining Methods

According to [33], the objective of data mining is both prediction which has an aim of predicting unknown or future values of the attributes of interest using other attributes in the databases and description which has an aim of describing the data in a manner understandable and interpretable to humans. Predicting the sale amounts of a new product based on advertising expenditure, or predicting wind velocities as a function of temperature, humidity, air pressure, etc., are examples of tasks with a predictive goal in data mining. Describing the different certain groupings that emerge in a sampling of satellite imagery is an example of a descriptive goal for a data mining task. The relative importance of description and prediction depend on the use in different applications. These two goals can be achieved by any of a number data mining tasks including: classification, regression, clustering, summarization, association dependency modeling, and deviation detection.

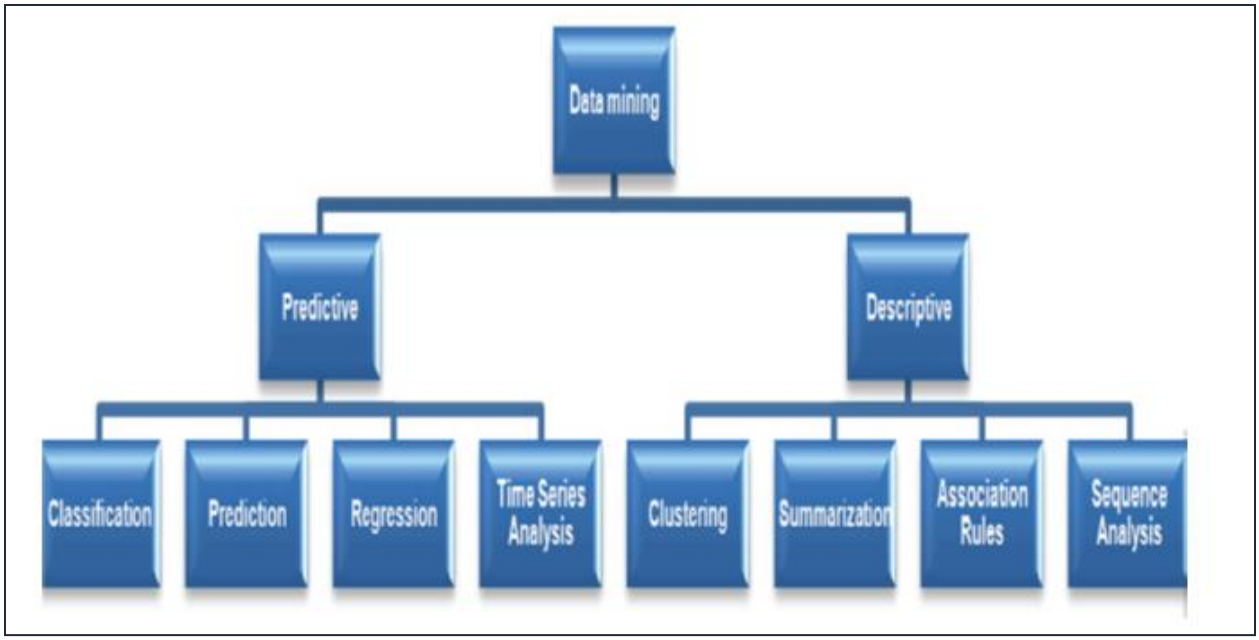

Figure 2.4 Data mining models and tasks

2.4.1. Predictive Tasks

In prediction the goal is to develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspect of the data (predictor variables).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].

Predictive data mining goals can be fulfilled the tasks of classification, regression, deviation detection.

- **Classification** – according to a research which is conducted by [14], the aim of classification is to classify items into several predefined classes. Given a collection of training samples, this type of task can be designed to find a model for class attributes as a function of the values of other attributes. Some popular classification methods include decision tree, logistic regression and support vector machine.
- **Regression** –a research which is conducted by [34],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. These tasks are studied in statistics and neural network fields .In this case some well-known regression methods include linear regression, neural network and support vector machine regression.

- **Deviation Detection** –a research which is done by [18]pointed out that the aim in this case is to discover the most significant changes in data from previously measured or normative values .Explicit information outside the data, like integrity constraints or predefined patterns, is used for deviation detection.

  [34] approached the problem from the inside of the data, using the implicit redundancy.

2.4.2. Descriptive tasks

- Clustering – according to a study which is conducted by [9], the aim of clustering is to identify a set of categories, or clusters, that describe the data .clustering is particularly useful in cases where the most common categories within the data set are not known in advance; if a set of clusters is optimal, within a category, each data point will in general be more similar to the other data points in that cluster than data points in other clusters.
- Summarization –a study of [19] pointed out that the aim of summarization is to find a concise description for a subset of data. 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 [18].

2.4.3 .Mixed tasks

There are some tasks in data mining that have both descriptive and predictive Aspects. Using these tasks, we can move from basic descriptive tasks toward higher-order predictive tasks. For instance relationship mining.

Relationship mining -In relationship mining the goal is to discover relationship between variables in a data set where large number of variables are available .in this case the emphasis is mainly on finding which variables are most strongly associated with a single variable of particular interest

[35].Association rule mining, Correlation mining, Sequential pattern mining and Causal data mining are some of the parts of relationship mining.

In association rule mining, the aim is to find if -then rules of the form that if some set of variable values is found, another variable have a specific variable.

Similarly in correlation mining, the goal is to find positive or negative linear correlation between variables .in sequential pattern mining, the goal is to find temporal association between events .in similar case, in causal data mining, the goal is to find whether one event (observed construct) was the cause of another event. Relationships found through relationship mining must satisfy two criteria: statistical significance, and interestingness. Statistical significance is generally assessed through standard statistical tests, such as F-tests. The interestingness of each finding is assessed in order to reduce the set of rules/ correlations/ causal relationships communicated to the data miner. In very large data sets, hundreds of thousands of significant relationships may be found. Interestingness measures attempt to determine which findings are the most distinctive and well-supported by the data [5].

2.5. Educational Data Mining Tasks and Approaches

Applicability of EDM can have many objectives for different stake holders .This can be viewed as academic objectives and administrative objectives .Academic objectives include person oriented which is connected with direct related to teaching – learning process; for instance student learning ,behavior ,risk and performance analysis can be an example. Department or Institution oriented is another part of academic objectives which is related to specific departments or institutions ;for instance with the aim of redesigning new courses for the department can be an example .the other part of academic objectives is domain oriented ,which is directly related to specific branch ;in this case designing methods ,tools or techniques ,knowledge discovery based decision support system for specific application or branch can be an example .Similarly an administrative objective is the other type of objective ,which is focused on administrator oriented that is related to direct involvement of higher authorities or administrators [8].

Additionally, different EDM researches have different objectives, mostly they point out predicting students' future learning behavior by creating students' model which contain detailed information as students' knowledge; discovering or improving domain models; study different kind of pedagogical relationship; as a goal [5] . Educational data mining researches use different methods;

mostly fall in to the methods of prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment [9].

In prediction, the goal of the method is to develop a model which can point out a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables).This can be applicable in predicting and understanding student educational outcome. Similarly in clustering, find data points that group together, splitting the full data set in to a set of categories. The applicability can be seen, for instance, in discovering new student behavior pattern. In relationship mining, the goal is to discover relationships between variables, in a data set with a large number of variables and this can be applied in discovering of curricular associations in course sequences .This may take the form of attempting to find out which variables are most strongly associated with a single variable of particular interest, or may take the form of attempting to discover which relationships between any two variables are strongest. Broadly, there are four types of relationship mining: association rule mining, correlation mining, sequential pattern mining, and causal data mining. In association rule mining, the goal is to find if-then rules of the form that if some set of variable values is found, another variable will generally have a specific value in relationship mining, the goal of the method is to discover relationship between variables [7].

Nowadays, the databases of most educational sectors contain so much data and information that it become complicated and difficult to analyze those data manually. To overcome the analysis of the data manually, educational data mining is a suitable technique to be used for conducting the data analysis process. It combine machine learning, statistical and visualization technique to discover and extract knowledge in such a way that humans can easily interpret [10] .The application of data mining helps to discover new knowledge about students.

Besides, the data stored in educational databases increases rapidly. These databases contain hidden information for the improvement of students' performance. Based on the data one can select frequent patterns and identify determinant factors for students' success. Similarly one can develop a model to predict the performance of students so as to identify the success and the failure [11].

2.6. Educational Data Mining and Related works

Educational Data Mining is one of the application Data Ming. Many researches pointed out the application data mining in education sector. Educational Data Mining (EDM) is upcoming field Knowledge discovery. Due to widespread growth of higher education, predictions related to student's performance can be accurately done through EDM. Not only predictions, classification, associations and grouping can also be done with perfection using statistical and software tools. The Education system can be equipped with more information relating to future drop out of students and their success in enrolled courses. Not only students but other stake holders could be benefitted by EDM. Nowadays interactive e - learning methods and tools have opened an opportunity to collect and scrutinize student data [34] **.**

A research which conducted by [41] ,discussed about how EDM become an emerging discipline with a suite of computational and psychological method and research approaches to understand how students learn and setting which they learn in it. The research further describes how EDM is a broader term which focuses on any educational data .Besides, it pointed out the scope includes areas that directly impacts students. Analysis of educational processes including admissions, alumni relation and course selection are the other areas of EDM.

Other research which is conducted by [42] describes about EDM which is an emerging discipline concerned with developing methods for exploring the unique type of data that came from educational settings. Besides using the methods enables to understand the problem of students.

Similarly there are different researches concerning educational data mining using different methods.

A study conducted by [43] tried to investigate the influence of factors on students' academic performance by comparing the accuracy of different classifiers. J48, Random Forest, Rep Tree and BF Tree of Decision Tree, Bayes and Naïve Bayes of Bayes Networks, Logistic and RBF Network functions and JRip rule classifiers were used. The study also shows that re-sampling of data was a critical step which is the reason of the success of the study.

Another research which done by [44] focused on the implementation of data mining methods and techniques for acquiring new knowledge from data collected by universities. The goal of the study to show the high potential of data mining applications for university management.in this case the specific objectives of the study is to find out if there are any patterns in the available data that

could be useful for predicting students' performance at the university based on their personal and pre university characteristics.

Similarly, there is a research which is conducted by [44] aimed to discover hidden knowledge and patterns about students' behavior. The research focused to develop Students' Academic Performance prediction models for the first semester Bachelor of Computer Science from University Sultan ZainalAbidin (UniSZA) by using three selected classification methods; Naïve Bayes, Rule Based, and Decision Tree. The comparative analysis is also conducted to discover the best classification model for prediction. From the experiment, the models develop using Rule Based and Decision Tree algorithm shows the best result compared to the model develop from the Naïve Bayes algorithm. Five independent parameters (gender, race, hometown, family income, university entry mode) have been selected to conduct this study. These parameters are chosen based on prior research studies including from social sciences domains. Other research [10], propose a framework for predicting students' academic performance of first year bachelor students in computer science course. Decision tree, naïve bayes and rule based classification techniques were used to the students' data to get the best students' academic performance prediction model. As the result rule based became the best model to predict comparing to other techniques by gaining highest accuracy value of 71.3%.

A research which is done [45], collected enrolled students' data from engineering institute which contain previous and current academics records like students roll no., name, date of birth, 10^{th}, 12^{th} and B.tech passing percentage & other information. They apply decision tree method for classifying students' academic performance. As a result, the information generated after the analysis of data mining techniques on student's data base is helpful for executives for training & placement department of engineering colleges. The research classifies the categories of student's performance in their academic qualifications.

There is also another research which is conducted by [46], use clustering algorithm and students' academic details to get an information. The research aims to develop a trust model using clustering algorithm. As a result, Using K-Means clustering algorithm, [44] predicted the pass percentage and fail percentage of the Overall students appeared for a particular examination. The results show the students' performance and it is seeming to be accurate. The comparison between Naïve bayes algorithm and decision stump tree technique shows that the Naïve bayes techniques produce accurate result than the other and it is measured using confusion matrix.

A research which is conducted by [47], applied three supervised data mining algorithms on assessment data to predict success in a course and the performance of the learning methods were evaluated based on their predictive accuracy .The result indicates that naïve bayes became better classifier model than neural network and decision tree.

Many researches focus on predicting students' academic performance on higher education by using Non –University academic records. But there are also researches which focus on predicting secondary school students' performance.

There was a research by [48], the research work intends to approach student achievement in secondary education using data mining techniques. Recent real world data e.g. student grades, demographic, social and school related features were collected by using school reports and questionnaires. Finally, the results showed that a good predictive accuracy can be achieved, provided that the first and /or second school period grades are available. Although students' achievement is highly influenced by past evaluations, an explanatory analysis has shown that there are also other relevant features, e.g. number of absence, parent's job and education and alcohol consumption.

In similar manner, the three researchers [49], wrote a paper which focus on evaluation and prediction of students' performance in high school. The paper tries to present superior models in predicting student's performance. At the result individual, environmental and educational factors affecting successful and unsuccessful students have been analyzed and according to them efficient models based on decision tree models have been presented

There are also different educational data mining researches using association rule data mining method. A research which is conducted by [50], investigates the application of association rule analysis to the subjects that the students choose to study at A and AS level, using data from the National Pupil Database (NPD).As a result, the analysis showed relatively little information that was not already known but uses as an illustration of the techniques and types of relationships uncovered. As a recommendation the research pointed out that the use of association is best suited to initial exploration of unfamiliar data to enable hypotheses to be formed that can be investigated using other methods.

A research which is done [52] which focused on application of association rule, also applied data mining technique using apriori algorithm to determine the pattern of relationship SBMPTN

database and students' GPA.As a suggestion they pointed that using large data enables to the rules are more diversified and useful for decision making processes.

There is also a research which uses Apriori algorithm [40]. In the research, Apriori algorithm is used which extracts the set of rules, specific to each class and analyzes the given data to classify the student based on their performance in academics. Students are classified based on their involvement in doing assignment, internal assessment tests, attendance etc., which helps to predict the performance of the student based on the pattern extracted from the educational database. From the experiment result it is found that Apriori algorithm is used to obtain minimal rules. From the extracted pattern Apriori algorithm is found to be effective in predicting the student under three categories: good, average and poor.

A research which conducted by which has an intention of mining association rule from the students' assessment data [20] , used association rule discovery techniques to discover the factors that affect the academic results so as to increase success chance of students.as a result ,the mined association rule showed various factors like student's interest, curriculum design; teaching and assessment methodologies that can affect students who have failed to attain a satisfactory level of performance in the Post-Graduation level.

In Ethiopia, there have been different researches which focused on educational data mining using different methods and techniques.

A research which is done by [54] , on students' success and failure prediction , investigated the potential applicability of data mining technology to predict student success and failure cases on University students' datasets. The research uses Classification and prediction data mining functionalities are used to extract hidden patterns from students' data. Rule generation process is based on the decision tree and Bayes as a classification technique and the generated rules were studied and evaluated. The research findings indicated that EHEECE (Ethiopian Higher Education Entrance Certificate Examination) result, Sex, Number of students in a class, number of courses given in a semester, and field of study are the major factors that affect students' performance.

The other research which is done [17], in their research, they pointed out using data mining concepts students' performance can be predicted based on students' academic record using decision tree. Based on their research from 49 attributes, 27 important rules were generated and from the generated model, specific courses, sex and academic status from 1st and 2nd year of the students determine the performance of students.

Similarly, the research which is done [16], focused his research on predicting student academic performance in higher education: the case of Jimma university by using Higher Education Entrance examination score, preparatory school transcript average point, Aptitude score, Mathematics score and English score and tries to predicate the first semester Jimma university students grade point average.J48 decision tree, PART rule induction, naïve bayes and MLP algorithms were applied. As a result, the model developed using un-pruned PART rule induction algorithm showed highest classification accuracy of 90.7%.

2.7. Factors affecting Secondary School Students' Success

Secondary School Students' success or performance is affected by different factors. The factors are different based on different environment. The academy achievements of students vary across regions, residences, across grade levels, sexes and schools. This is because that proper coverage of courses in different schools, qualification of teachers in different areas ,differences in school facilities ,organized school administration in different areas ,the way students educated beginning from lower grades are different in different areas. [58].Other studies categorize the factors in to family causal factors, academic causal factors and personal causal factors [59].In academic causal factors, the study describes the characteristics concerning the way of teachers' activities .while in personal causal factors, it describes different characteristics of students. Considering family causal factors, the study describes different characteristics of parents. In this case, the study showed that family background is the main determinant for the students' success. Here social class variables, educational and family environment of have an influence on students' success [59].

The other factors which determine the students' success are from personal variable an age of the student; from academic variable grade level and from family variable parents' education level. In this case students' performance can be predict using the variables grade level, age ,father's and mother's studies. Besides academic environment and academic motivation can predict the performance of the student [61]. Personality characteristics of the student, class size, peer group pressure, teacher personality and pressure and repeating classes are considerable factors for the students' success. In similar manner, studies showed the fact that student perception of teacher behavior, parent involvement, competence and self-worth are the predictors for the students' motivation and achievement. Besides, age and gender also were related to attitudes concerning

factors of achievement. The result supported other studies that which there is an association between achievement and student perception of teacher acceptancy, parent involvement, quality of science teaching, a supportive learning environment, and previous preparation [62].

CHAPTER THREE

METHODOLOGY

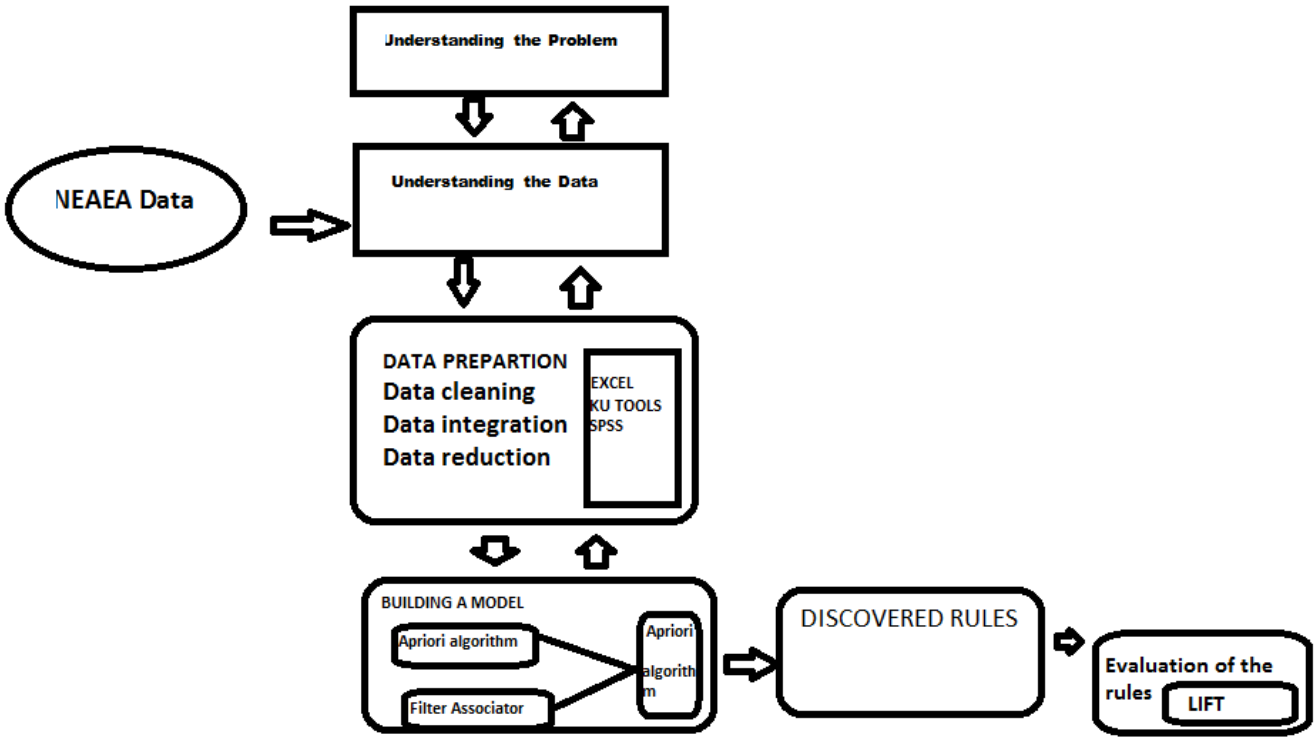

Figure 3.1. Research design

3.1. Research Design

Based on the figure the first step of the study is understanding the problem domain. This step includes overview of the preparatory school students, factors of preparatory school students' success. In understanding the data step domain specific terminologies, data description and attribute selection are included. In data preparation step, data cleaning, data integration and data reduction steps are applied. The next step is building the model based on the selected algorithm which is apriori algorithm. Using apriori algorithm the rules are discovered then the rules are evaluated using lift.

3.2. Data Mining Method

The objective of data mining is prediction and description. In prediction, one can predict unknown or future values of the attributes of interest using other attributes in the databases while description data mining has an aim of describing the data in a manner understandable and interpretable to humans [25].

This research paper uses description data mining method which is Association rule mining.

3.2.1. Association Rule Mining

Association rule mining is one of the most well studied data mining tasks. It discovers relationships among attributes in databases, producing if-then statements concerning attribute values [36] . Association rule analysis is a popular data mining method which discovers values which occur together. Its origins are in marketing and describes products which bought together by a customer. Similarly in [35], association rule mining is expressed .association rule mining is one of the data mining technique which is essential for different applications.in the research, association rules are required to fulfil the minimum support and the minimum confidence at the same time.at the start minimum support is applied to the given set of items, then using minimum confidence frequent item sets rules are formed. Association rule follow a rule of if A then B, where A and B can be items ,values or words .In this case the rule contains the antecedent or left hand side and the consequent or right hand side .it describes the relationship between support ,confidence and interestingness .other research which is discussed by [24] **,**pointed out that association can be applied to e -learning systems for traditionally association analysis that is finding correlations between items in a data set),which include for instance automatically guiding the learner's activities and intelligently generate and recommend learning materials , identifying attributes characterizing patterns of performance disparity between various groups of students , discovering interesting relationships from a student's usage information in order to provide feedback to the course author , finding out the relationships between each pattern of a learner's behavior , finding student mistakes often occurring together , guiding the search for the best fitting transfer model of student learning , optimizing the content of an e - learning portal by determining the content of most interest to the user , extracting useful patterns to help educators and web masters evaluating and interpreting on-line course activities , and personalizing e-learning based on aggregate usage profiles and a domain ontology [23].

- **Basic Concepts**

According to [18] association rule has the following concept.

Let I = {I1,I2,I3……Im} be a set of items and D be the set of transactions (transactional data set) where each transaction T ⊆ I is associated with an identifier TID and m is the number of items. Let A and B be two sets of items. A transaction T is said to contain A if and only if A ⊆ T. An association rule is an implication in the form A ⇒B where A ⊂ I, B ⊂ I, and AnB =Ø

The interestingness of an association rules describes how significant the rule is with respect to D. Two measures are used to quantify the interestingness of a rule:

Support, which indicates the frequency (probability) of the entire rule with respect to D. It is defined as ratio of the number of transactions containing A and B to the total number of transactions (the probability of both A and B co-occurring in D):

$$\text{support}(A \Rightarrow B) = P(A \cup B) = \frac{||\{T \in D | A \cup B \subseteq T\}||}{||D||}$$

Confidence, which indicates the strength of implication in the rule. It is defined as ratio of the number of transactions containing A and B to the number of transactions containing A (conditional probability of B given A):

$$\text{confidence}(A \Rightarrow B) = P(B|A) = \frac{||\{T \in D | A \cup B \subseteq T\}||}{||\{T \in D | A \subseteq T\}||}$$

Association rule is well researched method for discovering interesting relations between variables in large data bases .it is aimed to identify strong rules discovered in data bases using different measure of interestingness. Besides it has a goal to extract interesting correlations, frequent patterns, associations or causal structures among set of items in the transaction data bases. Some of the popular algorithms for generating association rules are priori algorithm, Eclat and FP - Growth which are used to mine frequent item sets [37] .

- **Apriori algorithm**

Apriori is a classic algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis. The rules generated by Apriori algorithm makes it easier for the user to understand and further apply the result. Employed the association rule method specifically Apriori algorithm to identifying novel, unpredicted and exciting samples in hospital infection control. Another study by employed Apriori algorithm to generate the frequent item sets and designed the model for economic forecasting, presented their methods on modeling and inferring user's intention via data. Association rules are usually required to satisfy a user-specified minimum support and a user specified minimum confidence at the same time [20].In the research which is conducted by [10] Apriori algorithm mentioned as one of the most important findings in the history of mining association rules since its introduction [36] and basic concepts defined before the introduction of algorithms for mining association rules. In the research Item sets included in a database is shown by Item set=,X1, X2, …Xn. Then, for each rule, two values of support and confidence is determined. Support is the probability that the transaction contain both X and While Confidence is the conditional probability that the transaction containing X and Y, also contains Y.

According to [38], Apriori algorithm is a widely used algorithm for the association rule and it is based of the rule of all sub item sets of a frequent item sets must also be frequent .By using this rule, apriori is able to prune huge amount of item sets. Besides it uses a bottom up approach.

The other research which mentioned about apriori algorithm is [26]**.**In this research apriori algorithm mentioned as the first and best known for association rule mining .It is one of the most influential Boolean association rule mining for frequent item **sets .** It is an iterative algorithm to calculate the specific length of item collection of given database to produce frequent item sets. It cut down candidate item sets using the principle that all non-empty subsets of frequent item sets are frequent too. Apriori algorithm basically works in two steps. In first step candidate item set is generated using linking process and in next step frequent item set from those candidate item set is found based on minimum support count by scanning the database.

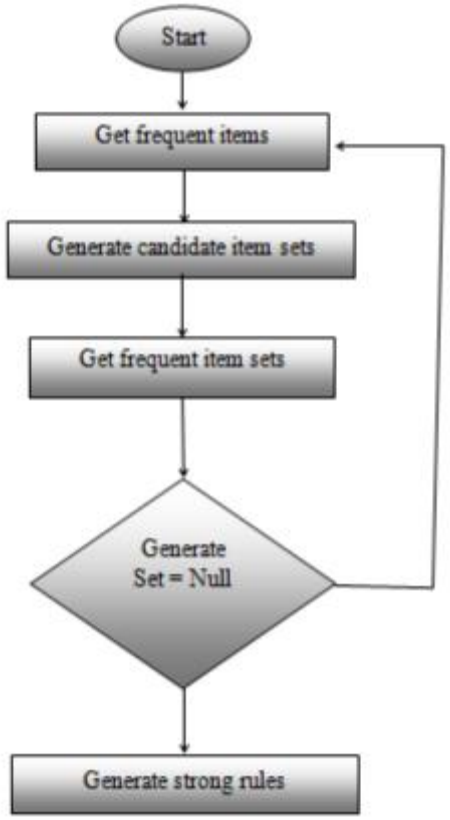

Fig 3.2. Flow Chart of Apriori algorithm

- FP-Growth

According to [39], FP-Growth is a new approach, which has a property in pre-processing step that derives a highly condensed representation of the transaction data, so called FP tree. The generation of FP -tree is done by counting occurrences .FP -growth uses the FP tree to derive the support values of all frequent item sets.

- Eclat

The algorithm Eclat applies the optimization called fast intersections .Whenever we intersect two tide lists, then we are only intersecting in the resulting tide lists if it is cordially reaches min support.in this case the need to break off each intersection as soon as it is sure that it will not achieve the threshold. Éclat originally generates only frequent item sets of size greater or equal to three [40].

- Filtering Associator

  The lift of an association **I**=>J is defined as: lift = p(J/I)/P(J), P(J) = (support of j)/(no of transactions).

  If lift > 1, then I and j are positively correlated; if lift < 1, then I and J are negatively correlated. If lift =1, I and J are independent [36].

3.3. Pre- processing techniques

Real world data usually become noisy due to its larger size and its origin from multiple and different sources .The fact that quality decisions need quality data; the targeted data need to be cleaned [19].In this research, from different pre - processing techniques data cleaning, data integration and data reduction can be applied in order to get the targeted data set.

3.3.1. Data Cleaning

Data cleaning can be applied to remove the noise .the noise which is emerged from the age attribute .both data which are EGSECE (Grade 10) and EHEEE (Grade 12) data in the age attribute contains the age value 00 to 14. In Ethiopia the primary education starts at the age of 5 or 6.in this case Grade 10 and Grade 12 will be attained mostly at the age of 15 and above. [1].the other noise which emerges 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 an attribute in this case 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 an attribute value disqualify which is mentioned 'dis' in the data .this is the case where students face irregularities on exams.

3.3.2. Data Integration

The data which comes from National Educational Assessment and Examination Agency (NEAEA) are in different tables. The three years 2004 E.C, 2005 E.C,2006 E.C. EGSECE(Grade 10) data and the corresponding 2006 E.C. ,2007 E.C. and 2008 E.C. EHEEE(Grade 12) respectively data need to integrate in to the new table.

3.3.3. Data Reduction and transformation

The EGSECE (Grade 10) data contains from 30 to 36 attributes in each year while the EHEEE (Grade 12) data contains 21 attributes in each year; this needs removing the irrelevant attributes. Similarly the attribute age and the other subject attributes contain continuous values which needs data discretization.

CHAPTER FOUR

DATA PREPARATION

This include problem domain understanding, data understanding and data pre - processing. Problem domain understanding contains overview of preparatory school and students, secondary school students' success and factors which affect secondary school students' success.

4.1. Problem Domain Understanding

Primary education has a duration of 8 years(age groups 6-12) and divided in to parts with 4 year cycles ;the first part which goes to grade 4 and the second part from grade 5 to grade 8.At the end of grade 4 students will take an exam and scoring 50 percent enables them to continue grade 5.At the end of grade 8 students take the National primary School Certificate Examination which is accomplished by in each region of education bureaus .Passing this examination enables students to join general secondary education which take two years .At the end of grade 10 ,students take the EGSECE which is administered by the National Educational Assessment and Examination Agency. Students who failed the exam will attend different vocational training while the one who passed will join the two year preparatory classes grade 11 and 12 .At the end of grade 12 the students will take an exam called the EHEEE.

The second phase of preparatory secondary education has two streams which are natural science and social science stream. Students which are natural science stream will take special subjects ;mathematics ,chemistry ,biology and physics ;general subjects English ,civics &ethical education ,physical education and information technology; and will take optional subjects for instance ethnic language ,national language or foreign language .While Social Science stream students will take specific subjects ;geography ,history and economics ;general subjects English ,mathematics for social, physical education and information technology; and will take optional subjects foreign language ,ethnic language or national language .In EHEEE, students of Natural Stream will take an exam of subjects English ,Math's for natural ,physics ,chemistry ,biology ,civics &ethical education and an Aptitude exam. Similarly Social Science stream students will take an examination of English, Math's for social, geography, history, economics, civics ethical education and an Aptitude examination [3].

4.1.1. Overview of Preparatory School Students

In Ethiopia education system, since 2001 G.C, the 6 plus 2 plus 4 structure; that is six years of primary schooling followed by two years of junior secondary school then followed by four years of senior secondary education; is changed in to a 4 plus 4 plus 2 plus 2 structure [1].The primary education which consists of age group of 6 up to 14 students lasts in 8 years. It is divided in to 2 four year cycles. The first cycle goes from grade 1 to grade 4 and the second cycle is from grade 5 to grade 8.These cycles now followed by a 2 year general secondary education ,which are Grade 9 and Grade 10.At the end of grade 10 ,students take an examination called EGSECE which is administered by National Educational Assessment and Examination Agency.This examination decides whether students join the vocational training or the two general upper grades or preparatory schools [1].Thus ,the general secondary education (Grade 9 and Grade 10) aims to prepare students to identify areas of interest for further education and training. Then the next 2 years are for preparatory secondary education (Grade 11 and Grade 12). Grade 11 and Grade 12 are the preparatory levels for higher education with an aim of prepare students for higher education or choosing a career [13]. The preparatory class contain two streams which are Natural science and Social science stream.

Considering enrollment growth rate of the country, the first cycle secondary school students which are grade 9 and 10 growth enrollment rate has more than doubled since 2000 G.C. [4];even though there are challenges considering secondary education in Ethiopia; including a low primary education completion rate. The other challenge is inequitable access considering rural population and female participation. Besides, there is a challenge of student learning achievement that are disappointedly low [4].In 2006 E.C. EHEEE ,A.A region preparatory school students are 25885 ,which are 12.9% of the country's examinees. From this, 42.6% are male students and 57.4 % are female students. Similarly, in 2007 E.C. EHEEE, 25879 which are 12.2 % examinees are A.A region preparatory students. From this 43.2 % are male while the rest 56.8 % are female students. In similar manner in 2008 E.C., A.A region examinees are 27686 which are 10.9 % of the country's examinees; from which 42.4 % are male and the rest 57.6 % are female students [13].

4.1.2. Preparatory School Students' Success

In Ethiopia, It is known that grade 11 and grade 12 are the preparatory classes for higher education. The students' success to higher education depends on the cutting points set by MOE for each year

.The cutting points consider different cases, which is shown on the following table. Similarly the cutting points of Natural and Social streams are different since 70% of students are Natural science stream students [

This study focuses on Natural Science stream subjects and so as students. Considering this stream there are different cutting point cases.

Cases	Year					
	2006 E.C.		2007 E.C.		2008 E.C.	
	M	F	M	F	M	F
Regular &Night students	315	300	343	320	354	340
Pastoralists and less developed regions	305	300	332	315	340	335
Hearing disability students	300	300	297	297	297	297
Blind students	-	-	-	-	-	-
Non regular and night students	330	320	353	331	360	355
For private higher education centers	265	265	275	275	275	275

Source: NEAEA Students' higher education placement department

Table 4.1. Cutting points to enter higher education

- **Case 1** Regular & Night students - these are governmental and non-governmental schools students in which the classes are regularly in a day and night time.in this case the cutting points are different from other cases.

- **Case 2** These are pastoralists and less developed region students. Ethiopian Somali, Benishangul, Afar region and some schools in South Nations, Nationalities, and People (SNNP) students are considered in this case. The cutting points are less than other region students.
- **Case 3** and **Case 4** these are the cases for hearing disability students and blind students the cutting point for this case is less than other cases.
  Blind students take only social science subjects. There is no blind student in natural science stream.
- **Case 5** These are students who will join private higher education centers in any programs. These students are the one who fails to score government higher education cutting points.

Based on the cutting points ,from Natural Science Stream students different success in entering in higher education is observed .In 2006 E.C. EHEEE data , 68% of male preparatory class students which contain government regular ,non-government regular ,government night students succeeded in joining government higher education ;meanwhile from female and similar categories 57 % of students joined government higher education. Similarly in this year from private students who take an exam, less than 1 % joined government higher education.

In 2007 E.C EHEEE data ,from natural science stream students ,66% of male and 59 % of female students which are in the category of government regular, non-government regular and government night preparatory school students succeeded in entering government higher education centers. But less than 1 % of male and female private students succeeded in joining government higher education.

Similarly, in 2008 E.C EHEEE data , 72% of male government regular, non-government regular and government night preparatory school students succeeded in joining government higher education .Considering female government regular, non-government regular and government night preparatory school students 62 % of them joined government higher education. From private students 1 % of male students succeeded in joining higher government education centers but less than 1 % of female students joined higher government education centers.

The goals of preparatory schools are preparing students for higher education. There achievement can be seen in students' success in entering to higher government colleges or universities. The

students who score the year's cutting point or above the cutting point will join the government higher education centers. So identifying the determinant factors for the students' success is very important for preparatory schools in order to help students for better achievement.

4.2. Data Understanding

According to [18], next to domain understanding in hybrid data model is data understanding.

The data for this study is collected from NEAEA's examinees result data base. The data contains Addis Ababa region of EHEEE data of 2006 E.C, 2007 E.C and 2008 E.C data and similarly EGSECE data of these students .So 2004 E.C Grade 10 data of 2006 E.C Grade 12 students, 2005 E.C Grade 10 data of 2007 E.C Grade 12 students and 2006 E.C Grade 10 data of 2008 E.C Grade and 2008 E.C Grade 12 students data included.

4.2.1. Domain Specific Terminologies

In EHEEE and EGSECE data; there are terminologies which are used in domain experts in this case NEAEA data analysts. In NEAEA the examinees are categorized in to 12 regions namely Tigray region ,Afar region , Amhara region , Oromia region, Somali region , Benishangul region ,South Nations ,Nationalities and People(SNNP) region , Gambella region , Harari region ,Dire Dawa region and Abroad. In each region the examinees are placed in to different zones, school type and school number. To represent the region, zone, school type and school number; there is a unique code .The examinees code is placed as AA BB CC DD; where

AA - represents a two digit number which denotes region number.

BB - represents a two digit number which shows the zone in a region.

CC - represents a two digit number which shows school type.

DD - represents a two digit number which shows the school number in the zone.

For instance in 14 01 01 01;

14 - Represents a region code which is Addis Ababa.

01 - Represents the zone code, in Addis Ababa the zones are the sub-cities.in this case which is Addis Ketema sub-city.

01 - Represents school type which is regular.

01 - Represents school number in the zone.

Considering Addis Ababa region, the zones are sub-cities. Each sub-city has its code.

SUB-CITY	SUB-CITY CODE
Addis Ketema	01
Akaki Kaliti	02
Arada	03
Bole	04
Gulele	05
Kirkos	06
Kolfe Keranyo	07
Ledeta	08
Nefas seleke	09
Yeka	10

Table 4.2. The code of Sub Cities in Addis Ababa region

In National Educational Assessment and Examination Agency (NEAEA), considering Addis Ababa students; there are 5 school type codes.

SCHOOL TYPE	CODE
Regular Government School students (RG)	01
Regular Non-Government School students (RNG)	02
Night Government School students (NG)	03
Night Non-Government School students (NNG)	04
Private students who sat only for examination (PR)	05

Table 4.3. School Type and its code

RG – Preparatory students who attend in Government preparatory schools regularly.

RNG – preparatory students who attend in Non-Government preparatory schools regularly.

NG – preparatory students who attend in Government preparatory schools at night section.

NNG – preparatory students who attend in Non –Government preparatory schools at night section.

PR – students who do not attend classes but sat only for the examination.

4.2.2. Data Description (exploration)

In 2006 E.C EHEEE data ,there are 25884 instances; from which 18971 which is 73.3% are Natural Science stream students while the rest 8447 which is 26.7% are Social Science students. Similarly this data contains 21 attributes.

In 2007 E.C EHEEE data contains 25879 instances and 21 attributes. From 25879 instances, 19074 which is 73.7% are Natural Science stream students; the rest 6926 which is 26.3% are Social Science stream students.

In 2008 E.C EHEEE data, there are 27686 instances and 22 attributes. From the instances 19544 which is 70.6% are Natural Science stream students; while 8142 which is 29.4% are Social Science stream students.

Each sub city has different number of students based on the school types. The school type includes RG,NG,RNG,NNG and PR .The following table summarizes EHEEE data of 2006 E.C , 2007 E.C and 2008 E.C

Count of Mat. Nat.	School Type												
	NG			PR			RG			RNG			Gr. Tot.
Sub City	F	M	Tot.	F	M	Tot.	F	M	Tot.	F	M	Tot.	
Addis Ketema	20	25	45	41	85	126	513	503	1016	236	133	369	1556
Akaki				34	55	89	276	277	553	202	193	395	1037
Arada	26	28	54	110	78	188	777	547	1324	656	470	1126	2692
Bole	8	11	19	70	61	131	596	502	1098	363	394	757	2005
Gulele	19	29	48	67	51	118	700	753	1453	389	363	752	2371
Kirkos	12	13	25	89	65	154	327	337	664	257	332	589	1432
Kolfe				145	98	243	677	571	1248	342	277	619	2110
Ledeta				41	40	81	220	186	406				487
Nefas Seleke				141	84	225	730	580	1310	844	786	1630	3165
Yeka				103	76	179	635	483	1118	377	442	819	2116
Gr. Tot.	**85**	**106**	**191**	**841**	**693**	**1534**	**5451**	**4739**	**10190**	**3666**	**3390**	**7056**	**18971**

Table 4.4. 2006 E.C. EHEEE data which shows Sub-City by School Type of Natural Science stream students

Count of Mat. Nat.	School Type															
	NG			NNG			PR			RG			RNG			Gr. Tot.
Sub city	F	M	Tot.	F	M	Tot.	F	M	Tot.	F	M	Tot.	F	M	Tot.	
Addis Ketema	15	22	37	4	12	16	53	65	118	477	500	977	237	128	365	1513
Akaki							60	48	108	306	297	603	186	197	383	1094
Arada	14	15	29				133	112	245	715	584	1299	649	488	1137	2710
Bole	8	8	16				129	81	210	489	446	935	434	474	908	2069
Gulele							77	68	145	622	685	1307	344	281	625	2077
Kirkos	16	23	39				122	68	190	336	252	588	260	377	637	1454
Kolfe							132	117	249	630	528	1158	431	378	809	2216
Ledeta							40	34	74	234	181	415				489
Nefas Selke							112	79	191	619	537	1156	994	978	1972	3319
Yeka							99	75	174	608	487	1095	403	461	864	2133
Gr. Tot.	**53**	**68**	**121**	**4**	**12**	**16**	**957**	**747**	**1704**	**5036**	**4497**	**9533**	**3938**	**3762**	**7700**	**19074**

Table 4.5. 2007 E.C. EHEEE data which shows Sub-City by School Type of Natural Science stream students

Count of Mat. Nat.	SCHOOL TYPE												
	NG			PR			RG			RNG			Gra.Tot.
SUB CITY	F	M	Tot.	F	M	Tot.	F	M	Tot.	F	M	Tot.	
Addis Ketema	10	15	25	56	45	101	588	497	1085	207	132	339	1550
Akaki K				57	72	129	349	322	671	230	242	472	1272
Arada	13	21	34	169	111	280	586	495	1081	626	475	1101	2496
Bole	6	11	17	104	94	198	502	432	934	562	492	1054	2203
Gulele				70	82	152	609	665	1274	365	346	711	2137
Kirkos	13	19	32	99	57	156	315	223	538	233	322	555	1281
Kolfe				167	125	292	533	513	1046	601	498	1099	2437
Ledeta				44	30	74	208	114	322				396
Nefas Seleke				84	91	175	635	536	1171	1003	1048	2051	3397
Yeka	2	3	5	121	83	204	642	514	1156	485	525	1010	2375
Gra.Tot.	**44**	**69**	**113**	**971**	**790**	**1761**	**4967**	**4311**	**9278**	**4312**	**4080**	**8392**	**19544**

Table 4.6. 2008E.C. EHEEE data which shows Sub-City by School Type of Natural Science stream students

Count of Bio G10	School Type												
	NG		NG Total	PR		PR Total	RG		RG Total	RNG		RNG Total	Grand Total
Sub City	F	M		F	M		F	M		F	M		
Addis Ketema	14	16	30	37	71	108	1188	1122	2310	604	414	1018	3466
Akaki							230	207	437	183	167	350	787
Akaki K				17	32	49	434	456	890	282	317	599	1538
Arada	14	24	38	77	53	130	1364	1065	2429	1589	1210	2799	5396
Bole	6	7	13	33	40	73	1102	1009	2111	1164	1142	2306	4503
Gulele	12	18	30	46	36	82	1283	1387	2670	948	891	1839	4621
Kirkos	7	11	18	62	39	101	649	594	1243	606	821	1427	2789
Kolfe				1		1	829	700	1529	847	762	1609	3139
Kolfe				112	81	193	472	409	881	250	200	450	1524
Ledeta				30	25	55	403	304	707	47	34	81	843
Nefas Selke				93	44	137	1584	1339	2923	2077	2111	4188	7248
Yeka				41	34	76	1237	977	2214	1077	1108	2185	4474
Grand Total	53	76	129	550	455	1005	10775	9569	20344	9674	9177	18851	40328

Table 4.7. EGSECE data which shows Sub - City by School Type

4.2.3. Attribute Selection

The EHEEE data contains 21 attributes. Based on the domain experts ,Natural science stream preparatory school students take an exam of English ,Natural mathematics ,Physics ,Chemistry ,Biology ,Aptitude and Civics subjects. Therefore the other subject attributes such as History,

Geography, Economics, and Mathematics for social are irrelevant attributes for this study. Besides an aptitude subject exam score also irrelevant since the study takes common subjects that students take in Grade 10 and Grade 12.Similarily the students' full name ,registration number ,school name ,nationality of the student are irrelevant attributes for the study. In similar manner an attribute, sight of the student, is not an irrelevant attribute since the study takes natural stream students so as blind students only take social stream subjects. From different literatures, different attributes are observed in determining students' success in secondary schools. Regions ,residence ,sex, grade level ,age of the student ,the way of students educated from lower grades or previous preparation of students can be considered in determining students success [61], [62],

[58]..Based on the above ideas the study select the following attributes.

S.No.	Attribute name	Description
1.	Zone (sub city)	Sub cities in Addis Ababa region
2.	School type	The type of schools students take an examination
3.	Sex	Sex of the student
4.	Age	Age of the student
5.	EGSECE english	The score of English subject in EGSECE
6.	EHEEE english	The score of English subject in EHEEE
7.	EGSECE maths natural	The score of Mathematics natural subject in EGSECE
8.	EHEEE maths natural	The score of Mathematics subject in EHEEE
9.	EGSECE physics	The score of Physics subject in EGSECE
10.	EHEEE physics	The score of Physics subject in EHEEE
11.	EGSECE chemistry	The score of Chemistry subject in EGSECE
12.	EHEEE chemistry	The score of Chemistry subject in EHEEE

13.	EGSECE biology	The score of Biology subject in EGSECE
14.	EHEEE biology	The score of Biology subject in EHEEE
15.	EGSECE civics	The score of Civics subject in EGSECE
16.	EHEEE civics	The score of Civics subject in EHEEE
17.	Success	Preparatory school students' success in entering to higher education

Table 4.8. The list of Attributes

Description of the Attribute based on Data Mining goal

- Zone (Sub city) denotes the schools' sub cites in Addis Ababa region .Its attribute type is nominal {Addis Ketema , Akaki Kaliti, Arada ,Bole ,Gulele ,Kirkos ,Kolfe ,Ledeta ,Nefas seleke ,Yeka }
- School Type denotes the type of school the preparatory students attend. Its attribute type is nominal {Regular Government (RG) ,Night Government (NG) ,Regular Non-Government (RNG) , Private(PR)}
- Sex denotes the sex of the student. Its attribute type is nominal {Male ,Female}
- Age denotes the age of the student when he/she takes the EHEEE examination. Its attribute type is numeric.
- EGSECE English denotes the score of English subject in Ethiopian General Secondary Education Certificate Examination.
- EHEEE English denotes the score of English subject in Ethiopian Higher Education Entrance Examination.
- EGSECE Math Natural denotes the score of Mathematics subject in Ethiopian General Secondary Education Certificate Examination.
- EHEEE Math Natural denotes the score of Mathematics subject in Ethiopian Higher Education Entrance Examination.
- EGSECE Physics denotes the score of Physics subject in Ethiopian General Secondary Education Certificate Examination.

- EHEEE Physics denotes the score Physics subject in Ethiopian Higher Education Entrance Examination.
- EGSECE Chemistry denotes the score of Chemistry subject in Ethiopian General Secondary Education Certificate Examination.
- EHEEE Chemistry denotes the score of Chemistry subject in Ethiopian Higher Education Entrance Examination.
- EGSECE Biology denotes the score of Biology subject in Ethiopian General Secondary Education Certificate Examination.
- EHEEE Biology denotes the score of Biology subject in Ethiopian Higher Education Entrance Examination
- EGSECE Civics denotes the score of Civics subject in Ethiopian General Secondary Education Certificate Examination.
- EHEEE Civics denotes the score of Civics subject in Ethiopian Higher Education Entrance Examination

Attribute Name	Attribute Type
Zone (Sub city	Nominal {Addis Ketema , Akaki Kaliti, Arada ,Bole ,Gulele ,Kirkos ,Kolfe ,Ledeta ,Nefas seleke ,Yeka }
School Type	Nominal {Regular Government (RG) ,Night Government (NG) ,Regular Non-Government (RNG) , Private(PR)}
Sex	Nominal {Male ,Female}
Age	Numeric
EGSECE English	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}
EHEEE English	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}
EGSECE Math Natural	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}

EHEEE Math.Natural	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}
EGSECE Physics	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}
EHEEE Physics	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}
EGSECE Chemistry	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}
EHEEE Chemistry	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}
EGSECE Biology	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}
EHEEE Biology	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}
EGSECE Civics	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}
EHEEE Civics	Nominal {Excellent ,Very good ,Good, Satisfactory, Fail}

Table 4.9. Attribute Type

4.3. Preparation of the Data and Pre-Processing

4.3.1. Data Cleaning

Data cleaning in this study applied to remove the noise; the noise which is emerged from the age attribute .Both data which are EGSECE and EHEEE data in the age attribute contains the age value 00 to 14. In Ethiopia the primary education starts at the age of 5 or 6.in this case Grade 10 and Grade 12 will be attained mostly at the age of 15 and above. [1].The other noise which emerges 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 an attribute in this case 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 an attribute value disqualify which is mentioned 'dis' in the data .this is the case where students face irregularities on exams.

The age attribute - this attribute contains 00 - 14 age values which creates noise. In Ethiopian education context ,the EHEEE is undertaken at the age greater than 15 years old [1].

Year	Attribute	Need to be Cleaned	Number of Instances
2006 E.C. EHEEE data	Age	Age values from 00-14	117
2007 E.C. EHEEE data	Age	Age values from 00-14	118
2008 E.C. EHEEE data	Age	Age values from 00-14	143

Table 4.10. Number of Instances of an Age attribute which needs cleaning

The age value which creates noise replaced by the most frequent value which is mode [19].In 2006 E.C EHEEE data the most frequent value is 18 ;while in 2007 E.C. EHEEE data the most frequent value is 18 .Similarly in 2008 E.C. data the most frequent value is 18.

The subject attribute value - In each EHEEE data the subject attribute contains different values which creates noise.

- 'Abs' value - this value denotes the students who do not take an exam; so known as absent. In the examination a student can be absent from taking all exams or specific subjects.
- 'n/a' value - this denotes the not available value. Each EHEEE data contains both Natural science and Social science streams examination scores and results. A student who is a Natural science stream student take an exam of English, Math natural, Aptitude, Physics, Chemistry, Biology and Civics. The Social science scores for this student becomes in the data n/a. Similarly, a Social science student in its Natural science score column will have a value of n/a.
- 'dsq' value - this denotes the disqualify value. A Student who obey rules and regulations of an examination face a value of disqualify in its subject score.
- 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.

  The following table illustrates each year's data 'abs', 'n/a', 'dsq' and zero value.

  The 2006 E.C. EHEEE data contains 25884 instances and 21 attributes .From 21 attributes Natural Science stream students take an examination of English, Math Natural ,Physics

,Chemistry ,Biology and Civics subjects which are similar to EGSECE data of the instances.

Year	Attribute	Need to be cleaned	Number of Instances	%
2006 E.C. EHEEE data	English	Absent ,n/a, disqualify, zero value	433	1.67%
	Maths. Natural	Absent ,n/a, disqualify, zero value	441	1.70%
	Physics	Absent ,n/a, disqualify, zero value	448	1.73%
	Chemistry	Absent ,n/a, disqualify, zero value	6882	26.6%
	Biology	Absent ,n/a, disqualify, zero value	6882	26.6%
	Aptitude	Absent ,n/a, disqualify, zero value	457	1.77%
	Civics	Absent ,n/a, disqualify, zero value	457	1.77%

Table 4.11. Subject attribute of 2006 EHEEE data which needs cleaning

Similarly 2007 E.C. EHEEE data contains 25879 instances and 21 attributes. From this the following table illustrates the attributes and the noise value.

Year	Attribute	Need to be cleaned	Number of Instances	%
2007 E.C EHEEE data	English	Absent, n/a, disqualify, zero value	533	2.06%
	Maths.Natural	Absent, n/a, disqualify, zero value	6757	26.11%
	Physics	Absent, n/a, disqualify, zero value	6760	26.12%
	Chemistry	Absent, n/a, disqualify, zero value	6781	26.20%
	Biology	Absent, n/a, disqualify, zero value	6767	26.15%
	Aptitude	Absent, n/a, disqualify, zero value	548	2.11%
	Civics	Absent, n/a, disqualify, zero value	550	2.13%

Table 4.12. Subject attribute of 2007 EHEEE data which needs cleaning

Moreover, the 2008 E.C. EHEEE data contains 27686 instances and 22 attributes. The following table illustrates the attribute value which contain noise value.

Year	Attribute	Need to be Cleaned	Number of Instances	%
2008 E.C. EHEEE data	English	Absent,n/a,disqualify,zero value	519	1.87%
	Maths.Natural	Absent,n/a,disqualify,zero value	8102	29.26%
	Physics	Absent,n/a,disqualify,zero value	8112	29.30%
	Chemistry	Absent,n/a,disqualify,zero value	8119	29.33%
	Biology	Absent,n/a,disqualify,zero value	8109	29.29%
	Aptitiude	Absent,n/a,disqualify,zero value	544	1.96%
	Civics	Absent,n/a,disqualify,zero value	540	1.95%

Table 4.13. Subject attribute of 2008 EHEEE data which needs cleaning

4.3.2 Data Integration

The EHEEE and the corresponding EGSECE data integrated in new table. Therefore, 2006 E.C EHEEE score of each subject of a student correspondingly integrated with its 2004 E.C EGSECE score of common subject of similar student. Similarly 2007 E.C EHEEE integrated with 2005 E.C of EGSECE. The other data which is 2008 E.C EHEEE integrated with 2006 E.C of EGSECE. The integration takes place using the name of the students in each examination.

For integration purpose, KU TOOLS software is used. Using the KU TOOLS the EHEEE data is merged with the corresponding EGSECE data and then identifies the score of the common subjects by using common or similar name of a student. The number of students who are in EHEEE are

smaller than the EGSECE data since the students are failed in EGSECE. Similarly after integration, the number of students (instances) are decreased since the students name in EHEEE do not match with their name in EGSECE.

After integration the number of instances of each year's EHEEE data decreased since there are names of instances which are in EHEEE data but not in EGSECE; this is because of name changing of students in EHEEE.

The following table illustrates the number of instances before and after integration.

Data	Number of Instances before integration	Number of Instances after integration
2006 EHEEE data	18971	12933
2007 EHEEE data	19074	13422
2008 EHEEE data	19544	13976

Table 4.14. Number of instances before and after integration of EHEEE and EGSECE data

Sub City	School Type	Sex	Age	Bio G10	Bio G12	Chem G10	Chem G12	Civ G10	Civ G12	Eng G10	Eng G12	Math G10	Math G12	Phy G10	Phy G12	success
Nefas Selke	RNG	M	19	51	57	43	40	64	61	57	64	42	72	27	46	yes
Kirkos	RNG	M	18	68	67	56	36	64	49	62	78	64	82	53	62	yes
Kirkos	RNG	M	17	88	74	76	61	76	94	76	74	53	63	36	84	yes
Ledeta	RG	F	18	33	51	31	41	41	78	43	35	36	29	33	30	no
Kolfe	RG	M	19	27	55	49	43	38	57	35	32	51	34	22	30	no
Arada	RG	M	19	39	54	44	38	30	75	35	48	37	43	11	24	no
Nefas Selke	RNG	M	19	66	81	64	63	63	63	68	63	61	77	36	72	yes
Kolfe	RG	M	18	47	55	38	51	45	69	35	33	27	34	33	34	no
Addis Ketema	RG	F	19	78	83	75	68	66	80	62	63	54	55	33	44	yes
Kolfe	RG	M	18	50	67	66	51	55	76	30	44	36	66	33	62	yes
Ledeta	RG	M	19	56	58	28	33	69	83	43	45	34	29	24	34	no
Kolfe	RG	M	18	59	68	36	43	48	58	39	48	31	34	20	28	no
Ledeta	RG	M	18	60	57	30	38	61	51	34	43	46	40	24	24	no
Nefas Selke	RG	M	17	69	76	54	50	57	79	56	61	44	68	40	54	yes
Kolfe	RNG	M	19	53	72	51	54	55	67	49	52	64	45	27	46	yes
Nefas Selke	RNG	M	18	61	66	60	50	64	80	43	80	80	85	47	60	yes
Gulele	RG	M	21	38	39	34	35	34	73	48	56	44	34	29	28	no
Yeka	RG	M	19	80	82	71	56	74	82	63	54	53	66	38	52	yes
Kolfe	RG	M	20	57	59	69	49	50	66	43	36	44	72	38	58	yes
Yeka	RG	M	19	50	57	30	34	64	76	56	48	56	40	47	48	yes
Akaki K	RG	M	19	67	58	55	41	54	61	48	38	44	32	36	48	no
Kolfe	RG	M	18	54	53	31	53	57	51	38	44	17	38	27	32	no
Arada	RNG	M	18	86	75	83	46	89	95	86	78	85	57	62	64	yes
Ledeta	RG	M	17	56	65	38	34	61	65	43	48	29	38	29	32	no
Yeka	RG	M	17	52	60	56	36	45	70	43	34	39	32	36	42	no
Yeka	RG	M	19	77	66	54	36	73	90	67	68	56	65	24	48	yes
Addis Ketema	RG	M	19	68	77	40	56	55	74	43	56	25	29	20	56	yes
Kolfe	RG	M	18	54	54	49	49	50	62	43	27	32	66	24	58	yes
Bole	RG	M	20	47	59	35	45	39	68	29	38	34	34	31	40	no
Gulele	RG	M	21	46	66	43	51	50	53	38	41	29	68	42	50	yes
Kirkos	RG	M	18	73	59	40	36	69	70	66	49	24	26	36	36	no
Kirkos	RNG	M	18	78	80	43	51	65	87	70	67	71	77	38	74	yes

Figure 4.1. The numerical attribute values after integration

4.3.3. Data Reduction and Transformation

Data reduction 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 [18].One of the mechanism used in this study is discretization. Discretization is a concept in which raw data values for attributes are replaced by ranges or higher conceptual levels [24].In this study the subject attributes and an age attribute discretized since it has a continuous value.

The continuous value an attribute age discretized using equal width portioning discretization method [19].The equal width method divides the data in to a fixed number of intervals of equal or almost equal length. The following table illustrates how the age value is discretized.

The interval of the age attribute is calculated as shown below:

W = (max-min)/k , where w is the width of interval and k is the number of interval

Max is the maximum value of the age and min is the minimum value of the age.

W =(53-15)/4 = 9.5 and the interval boundaries are : min+w, min+2w,…min+(k-1)w

So the boundaries are 15+9.5,15+2(9.5) and 15+3(9.5): 24.5 ,34 and 43.5

Attribute Name	Original value	Discretized value	Transformed value
Age	15 – 53	15 – 24.5	Age 1
		25 – 34.5	Age 2
		35– 44.5	Age 3
		44 – 53.5	Age 4

Table 4.15. Age discretization

The other attributes which discretized are the subjects in EHEEE and EGSECE. The subjects are continuous value and discretized using the assessment system of secondary education in Ethiopia [1].The following table 4.16 illustrates the assessment system of secondary education.

Letter grade	In percentages	Meaning
A	90 – 100	Excellent
B	80 – 89	Very good
C	70 – 79	Good
D	50 – 69	Satisfactory
F	Under 50	Fail

Table 4.16. The assessment system of Secondary education in Ethiopia

But Physics subject discretized using equal width partitioning.

The width of the interval will be W = (B-A)/N, where W – width, B – the maximum value, A – the minimum value, N – number of interval

Therefore W = (91-2)/5 =17.8 ≈18

In percentage	Meaning
81-100	Excellent
62-80	Very good
43-61	Good
24-42	Satisfactory
<=23	Failure

Table 4.17. Discretization of Physics subject

Data Format Conversion for Weka software

The final integrated data changed in to the format which is suitable for WEKA. The following figure 4.2 illustrates the sample of the changed ARFF format.

Figure 4.2. The sample arff format.

```
@relation  FINALDATA
@attributeSubCity{AddisKetema,Akaki,Arada,Bole,Kirkos,Kolfe,Ledeta,Yeka,NefasSel
ke,Gulele}
@attribute  SchoolType{RNG,RG,NG,PR}
@attribute  Sex{M,F}
@attribute  BioG10{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  BioG12{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  ChemG10{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  ChemG12{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  CivG10{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  CivG12{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  EngG10{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  EngG12{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  MathG10{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  MathG12{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  PhyG10{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  PhyG12{Average,Verygood,Satisfactory,Excellent,Failure}
@attribute  Success{yes,no}

@DATA

Bole,RG,F,Failure,Average,Failure,Failure,Failure,Failure,Failure,Failure,Failure,Failure
Failure,Failure ,no
Arada,RG,M,Average,Verygood,Failure,Verygood,Satisfactory,Satisfactory,Average,Av
erage,Average,Average,Failure,Average,yes
Kolfe,RG,F,Failure,Failure,Failure,Failure,Failure,Failure,Failure,Failure,Failure,
Failure,Failure ,yes
```

Finally before analysis 16 attributes and 40328 instances are available

- Sub City - Its attribute type is nominal {AddisKetema , Akaki , Arada, Bole ,Gulele ,Kirkos ,Kolfe ,Ledeta ,NefasSeleke ,Yeka }
- School Type - Its attribute type is nominal {Regular Government (RG) ,Night Government (NG) ,Regular Non-Government (RNG) , Private(PR)}
- Sex - Its attribute type is nominal {Male(M) ,Female(F)}
- EGSECE English (EngG10) _ Its attribute type is nominal{Excellent, Very good, Satisfactory, Average, Failure}
- EHEEE English(EngG12) - Its attribute type is nominal{Excellent, Very good, Satisfactory, Average, Failure}
- EGSECE Math. Natural(MathG10) - Its attribute type is nominal{Excellent, Very good, Satisfactory, Average, Failure}
- EHEEE Math Natural(MathG12) – Its attribute type is nominal {Excellent, Very good, Satisfactory, Average, Failure}
- EGSECE Physics(PhyG10) - Its attribute type is nominal {Excellent, Very good, Satisfactory, Average, Failure}
- EHEEE Physics(PhyG12) - Its attribute type is nominal {Excellent, Very good, Satisfactory, Average, Failure}
- EGSECE Chemistry(ChemG10) - Its attribute type is nominal {Excellent, Very good, Satisfactory, Average, Failure}
- EHEEE Chemistry(ChemG12) - Its attribute type is nominal {Excellent, Very good, Satisfactory, Average, Failure}
- EGSECE Biology(BioG10) - Its attribute type is nominal {Excellent, Very good, Satisfactory, Average, Failure}
- EHEEE Biology(BioG12) - Its attribute type is nominal {Excellent, Very good, Satisfactory, Average, Failure}
- EGSECE Civics(CivG10) - Its attribute type is nominal {Excellent, Very good, Satisfactory, Average, Failure}
- EHEEE Civics(CivG12) - Its attribute type is nominal {Excellent, Very good, Satisfactory, Average, Failure}
- Success – Its attribute type is nominal(yes, no)

CHAPTER FIVE

EXPERIMENTAION AND RESULTS OF ANALYSIS

In Hybrid data mining method, the next step after data preparation is mining the data. In this case the selected data mining method which is association rule is applied to the prepared examination data and test whether it achieve the required minimum threshold.

From 40328 instances 29861 (74.%) shows success and 10467(26%) shows not achieving a success in entering to higher education

5.1. Comparing Different Algorithms

The data mining in this study is to identify the determinant factors for the students' success in the preparatory schools based on the rules achieved using the attributes on the analysis.

In association rule mining, the number of possible association rules can be very large even with small data sets. In order to reduce the count of rules found and so as to get the most interesting rules; setting minimum threshold on support and confidence values is important. [64].To compare algorithms by using WEKA based on execution time and data base scan; different parameters are used. The parameters are number of instances, confidence and support levels. [65].Here the compared algorithms are Apriori and Filtered Associator algorithm.

- Based on the number of instances .The number of instances when changing to the delta value .The delta values are 0.001,0.005,and 0.05

Delta value	Execution time per second	
	Apriori	Filter Associator
0.001	114	233
0.005	48	48
0.05	7	8

Table 5.1. The execution time using different delta value

- Based on the different confidence level. Here the confidence levels are 30%,40%,50%,60%,70% ,80% and 90%

Confidence level	Execution time per second	
	Apriori	Filter Associator
0.3	2	2
0.4	2	2
0.5	3	3
0.6	3	3
0.7	7	7
0.8	7	7
0.9	7	7

Table 5.2. The execution time using different confidence level.

- Based on different support level. The support levels are 0.20,0.15 and 0.10

Support level	Execution time per second	
	Apriori	Filter Associator
0.20	2	2
0.15	3	3
0.10	7	7

Table 5.3. The execution time using different support level

From the above records, the completion time of the analysis of the two algorithms are almost similar. But considering different support levels, the amount of time taken by Apriori algorithm to run completion is less than the amount of time needed by Filter associator. Thus, this study takes an apriori algorithm for analysis.

5.2. Attributes during the analysis

Using WEKA 3.8.0 the attributes are selected before the minimum threshold of the confidence is set.Attributes are selected during the analysis using CorrelationAttributeEval; since it evaluates the value of an attribute by measuring the correlation between an attribute and the class

Based on CorrelationAttributeEvaluation, the following table illustrates the ranking of the attributes during the analysis.

Correlation value	Attribute Name	Ranked order
0.27365	MathG12	1st
0.26977	ChemG12	2nd
0.26783	PhyG12	3rd
0.23145	EngG12	4th
0.22759	EngG10	5th
0.22016	MathG10	6th
0.21845	ChemG10	7th
0.19846	BioG10	8th
0.18201	BioG12	9th
0.16715	CivG12	10th
0.12792	PhyG10	11th
0.1274	CivG10	12th

0.06466	SchoolType	13th
0.0622	Sex	14th
0.00755	SubCity	15th

Table 5.4. Ranking of an attributes

5.3. Discovering the rules

Depending on the choice of the thresholds, the algorithm can become very slow and generate an extremely large amount of results or generate none or too few results, omitting valuable information [52].

5.3.1. Experimental setup to discover the rules

- Configuring minimum support 0.10, and minimum confidence of 0.9 ;some of the rules are :

Rules	Support	Confidence
PhyG12=Verygood 4368 ==> success=yes 4300	0.1	0.98
CivG12=Verygood 5594 ==> success=yes 5330	0.1	0.95
BioG12=Verygood 6294 ==> success=yes 5985	0.1	0.95
BioG12=Failure ChemG12=Failure 6377 ==> MathG12=Failure	0.1	0.94

- Configuring minimum support 0.1 and minimum confidence of 0.7;some of the discovered rules are :

Rules	Support	Confidence
Sex=F EngG12=Failure 8376 ==> MathG12=Failure 7269	0.1	0.87
MathG12=Satisfactory 10657 ==> success=yes 9587	0.1	0.90
Sex=M MathG10=Satisfactory 5918 ==> success=yes 5137	0.1	0.87
CivG12=Good 9256 ==> success=yes 8242	0.1	0.89

- Configuring minimum support of 0.15 and minimum confidence of 0.5;some of the following rules are discovered:

Rules	Support	Confidence
EngG12=Good 6992 ==> success=yes 6523	0.15	0.93
success=no 10467 ==> MathG12=Failure 9307	0.15	0.89
EngG10=Satisfactory EngG12=Satisfactory 7728 ==> success=yes 6691	0.15	0.87
Sex=F EngG12=Failure 8376 ==> MathG12=Failure 7269	0.15	0.87
SchoolType=RNG Sex=F 9674 ==> success=yes 7176	0.15	0.74

- Configuring minimum support of 0.2 and minimum confidence of 0.3;the following rules are discovered:

Rules	Support	Confidence
MathG12=Satisfactory 10657 ==> success=yes 9587	0.2	0.9
EngG12=Failure 14000 ==> MathG12=Failure 11722	0.2	0.84
CivG10=Good 9311 ==> success=yes 8139	0.2	0.87
SchoolType=RG 20343 ==> MathG12=Failure 14008	0.2	0.69

Configuring Minimum support 0.2 and minimum confidence 0.2; some of the following rules are discovered.

Rules	Support	Confidence
SchoolType=RNG 18851 ==> success=yes 14558	0.2	0.77
Sex=F 21051 ==> MathG12=Failure 14173	0.2	0.67
ChemG12=Satisfactory 17974 ==> success=yes 15465	0.2	0.86
Sex=M 19277 ==> success=yes 14823	0.2	0.77

5.3.2. Evaluation of the discovered knowledge

Based on the hybrid data model, the next step after mining of the data is evaluation of the discovered knowledge or rules. To evaluate the interestingness of the discovered rules, the study uses interestingness measure based on correlation or lift and evaluation using domain experts. [18]. Based on the value of the lift and the domain experts, the following rules are selected.

Rules	Confidence	Lift
BioG12=Failure ChemG12=Failure 6377 ==> MathG12=Failure	0.94	1.50
success=no 10467 ==> MathG12=Failure 9307	0.89	1.42
Sex=F EngG12=Failure 8376 ==> MathG12=Failure 7269	0.87	1.39
EngG12=Failure 14000 ==> MathG12=Failure 11722	0.84	1.34
PhyG12=Verygood 4368 ==> success=yes 4300	0.98	1.33
CivG12=Verygood 5594 ==> success=yes 5330	0.95	1.29
BioG12=Verygood 6294 ==> success=yes 5985	0.95	1.28
EngG12=Good 6992 ==> success=yes 6523	0.93	1.26

5.4. Findings of the study

From the analysis the discovered knowledge are categorized based on the sex, school type and subject attributes. Based on the sex attribute, 77 % of male preparatory students are associated with the success to enter higher education. While 87 % of Female preparatory students who failed in English examination in EHEEE are associated with failure in Mathematics examination in EHEEE.

Concerning School type, 77% of Non – Government preparatory school students associated with success to enter higher education. Similarly, 74% of Female Non-government preparatory school students associated with success to enter higher education. Likewise, 69 % of Government preparatory school students fail in Mathematics examination in EHEEE.

Based on the subject attribute, 98% of students who score very good in physics in EHEEE associated with success to enter higher education. Similarly 95% of students who score very good in Civics and Biology subjects' examination in EHEEE; associated with success to enter higher education. Moreover 93% of students who score good in English subject examination in EHEEE associated with success to enter higher education. In similar manner, 86% of students who score satisfactory in Chemistry subject examination in EHEEE associated with success to enter higher education. Besides a student who score good in Civics examination in EGSECE is associated with success to enter higher education. Concerning English subject, a student who score satisfactory in EGSECE and EHEEE is associated with success to higher education.

Besides the above discovered rules, 84% of students who fail in English subject also fail in Mathematics subject examination in EHEEE.

CHAPTER SIX

SUMMARY, CONCLUSION and RECOMMENDATION

6.1. Summary

The objective of the study was to identify the determinant factors of success in the preparatory schools using data mining techniques. In order to achieve the objective, the study uses EHEEE data and correspondingly their EGSECE data. The collected data is from National Educational Assessment and Examination Agency, an agency which is 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. Through data cleaning ,data which creates noise are eliminated; through integration the three year EHEEE data are integrated with its corresponding three years EGSECE data and the common subjects from Grade 10 and Grade 12 of a student which are English, Mathematics ,Physics ,Chemistry ,Biology and Civics are selected. Besides the subjects; Sub city, School Type, Sex are selected as an attributes. Finally, 40328 instances and 15 attributes are selected for analysis. Additionally, the results of the 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. This includes business understanding, data understanding, data preparation, data mining and evaluation of the discovered knowledge. The study uses Association rule mining to discover the rules. Apriori and Filter Associator algorithms are compared based on the execution time and data base scan. The parameters for this purpose are number of instances, confidence value and support value. Then Apriori algorithm is selected and the analysis is undertaken. Using different thresholds different rules are discovered. The rules are then evaluated using the interestingness measure lift and domain experts.

6.2. Conclusion

The success to enter higher education of preparatory students is based on the cutting points set by MoE. 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 EHEEE is also another determinant factor. In similar manner, scoring satisfactory in Chemistry in EHEEE and scoring satisfactory in English in EGSECE are the determinant factors for the students' success in entering higher education.

Based on the sex, male preparatory students are more associated with success to enter higher education than female preparatory students.

Concerning School type, Regular Non-Government preparatory students are more associated with success to enter higher education than Government preparatory school students. Sub city of schools that students attend have no influence on students' success to enter higher education. Similarly strong rules are achieved using apriori algorithm which can be easily understand by domain experts and other stakeholders. But in using Apriori algorithm, there is no standard way of setting different thresholds. This leads to missing the strong rules.

6.3. Recommendation

Since the preparatory classes have an aim of preparing students for higher education, from this study different stake holders can achieve important points which are helpful in strengthen the students' success.

The study uses association rule mining method and Apriori algorithm to identify the determinant factor for the students' success; it will be essential if other researches use other algorithm since there is no standard way of setting the threshold and this leads to missing of important rules.

Likewise, both male and female students of Government as well as Non-Government 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.

- Future Research

This study is focused on Addis Ababa region of preparatory school students, it will be important if other study works on other 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.

ANNEXES

ANNEX I: Sample rules discovered using Minimum support: 0.1 (4033 instances) Minimum metric <confidence>: 0.9

Best rules found:

1. PhyG12=Verygood 4368 ==> success=yes 4300 <conf:(0.98)> lift:(1.33) lev:(0.03) [1065] conv:(16.43)

2. BioG12=Failure ChemG12=Failure EngG12=Failure 4610 ==> MathG12=Failure 4400 <conf:(0.95)> lift:(1.53) lev:(0.04) [1520] conv:(8.2)

3. BioG12=Failure ChemG12=Failure PhyG12=Satisfactory 4744 ==> MathG12=Failure 4524 <conf:(0.95)> lift:(1.53) lev:(0.04) [1560] conv:(8.06)

4. CivG12=Verygood 5594 ==> success=yes 5330 <conf:(0.95)> lift:(1.29) lev:(0.03) [1187] conv:(5.48)

5. BioG12=Verygood 6294 ==> success=yes 5985 <conf:(0.95)> lift:(1.28) lev:(0.03) [1324] conv:(5.27)

6. ChemG12=Failure MathG10=Failure PhyG12=Satisfactory success=no 4915 ==> MathG12=Failure 4671 <conf:(0.95)> lift:(1.52) lev:(0.04) [1600] conv:(7.53)

7. ChemG12=Failure PhyG10=Satisfactory PhyG12=Satisfactory success=no 4362 ==> MathG12=Failure 4141 <conf:(0.95)> lift:(1.52) lev:(0.04) [1415] conv:(7.37)

8. ChemG12=Failure PhyG12=Satisfactory success=no 5868 ==> MathG12=Failure 5566 <conf:(0.95)> lift:(1.52) lev:(0.05) [1900] conv:(7.27)

9. ChemG10=Failure ChemG12=Failure PhyG12=Satisfactory success=no 4490 ==> MathG12=Failure 4257 <conf:(0.95)> lift:(1.52) lev:(0.04) [1451] conv:(7.2)

10. ChemG10=Failure ChemG12=Failure EngG10=Failure MathG12=Failure success=no 4363 ==> MathG10=Failure 4136 <conf:(0.95)> lift:(1.57) lev:(0.04) [1503] conv:(7.59)

11. ChemG12=Failure CivG12=Failure EngG12=Failure 4497 ==> MathG12=Failure 4262 <conf:(0.95)> lift:(1.52) lev:(0.04) [1452] conv:(7.15)

12. Sex=F ChemG12=Failure success=no 4695 ==> MathG12=Failure 4449 <conf:(0.95)> lift:(1.52) lev:(0.04) [1515] conv:(7.13)

13. ChemG10=Failure ChemG12=Failure EngG10=Failure success=no 4626 ==> MathG10=Failure 4380 <conf:(0.95)> lift:(1.57) lev:(0.04) [1589] conv:(7.43)

14. ChemG10=Failure EngG10=Failure EngG12=Failure success=no 4487 ==> MathG10=Failure 4248 <conf:(0.95)> lift:(1.57) lev:(0.04) [1540] conv:(7.42)

15. Sex=F BioG12=Failure ChemG12=Failure 4315 ==> MathG12=Failure 4083 <conf:(0.95)> lift:(1.51) lev:(0.03) [1387] conv:(6.95)

16. BioG10=Failure ChemG10=Failure EngG10=Failure MathG12=Failure success=no 4501 ==> MathG10=Failure 4258 <conf:(0.95)> lift:(1.57) lev:(0.04) [1542] conv:(7.32)

17. BioG10=Failure ChemG10=Failure EngG10=Failure success=no 5004 ==> MathG10=Failure 4727 <conf:(0.94)> lift:(1.57) lev:(0.04) [1708] conv:(7.14)

18. ChemG10=Failure ChemG12=Failure MathG10=Failure success=no 5359 ==> MathG12=Failure 5062 <conf:(0.94)> lift:(1.51) lev:(0.04) [1714] conv:(6.75)

19. ChemG12=Failure MathG10=Failure success=no 6394 ==> MathG12=Failure 6039 <conf:(0.94)> lift:(1.51) lev:(0.05) [2044] conv:(6.74)

20. SchoolType=RG ChemG12=Failure success=no 4383 ==> MathG12=Failure 4139 <conf:(0.94)> lift:(1.51) lev:(0.03) [1400] conv:(6.71)

21. ChemG10=Failure ChemG12=Failure EngG10=Failure MathG10=Failure success=no 4380 ==> MathG12=Failure 4136 <conf:(0.94)> lift:(1.51) lev:(0.03) [1399] conv:(6.71)

22. BioG10=Failure ChemG12=Failure MathG10=Failure success=no 4647 ==> MathG12=Failure 4384 <conf:(0.94)> lift:(1.51) lev:(0.04) [1480] conv:(6.61)

23. ChemG10=Failure ChemG12=Failure success=no 5812 ==> MathG12=Failure 5483 <conf:(0.94)> lift:(1.51) lev:(0.05) [1852] conv:(6.61)

24. ChemG10=Failure ChemG12=Failure EngG10=Failure success=no 4626 ==> MathG12=Failure 4363 <conf:(0.94)> lift:(1.51) lev:(0.04) [1473] conv:(6.58)

25. ChemG12=Failure EngG10=Failure MathG10=Failure success=no 5028 ==> MathG12=Failure 4742 <conf:(0.94)> lift:(1.51) lev:(0.04) [1600] conv:(6.57)

26. ChemG12=Failure MathG10=Failure PhyG10=Satisfactory success=no 4903 ==> MathG12=Failure 4623 <conf:(0.94)> lift:(1.51) lev:(0.04) [1559] conv:(6.55)

27. ChemG10=Failure EngG10=Failure MathG12=Failure success=no 5525 ==> MathG10=Failure 5209 <conf:(0.94)> lift:(1.56) lev:(0.05) [1875] conv:(6.91)

28. Sex=F ChemG12=Failure EngG12=Failure PhyG12=Satisfactory 4510 ==> MathG12=Failure 4252 <conf:(0.94)> lift:(1.51) lev:(0.04) [1434] conv:(6.53)

29. BioG10=Failure ChemG10=Failure ChemG12=Failure success=no 4473 ==> MathG12=Failure 4216 <conf:(0.94)> lift:(1.51) lev:(0.04) [1421] conv:(6.51)

30. ChemG10=Failure EngG10=Failure PhyG10=Satisfactory success=no 4612 ==> MathG10=Failure 4347 <conf:(0.94)> lift:(1.56) lev:(0.04) [1564] conv:(6.88)

31. ChemG12=Failure EngG12=Failure success=no 5305 ==> MathG12=Failure 5000 <conf:(0.94)> lift:(1.51) lev:(0.04) [1685] conv:(6.51)

32. Sex=F ChemG10=Failure EngG10=Failure EngG12=Failure 4671 ==> MathG10=Failure 4401 <conf:(0.94)> lift:(1.56) lev:(0.04) [1582] conv:(6.84)

33. ChemG12=Failure EngG12=Failure MathG10=Failure success=no 4554 ==> MathG12=Failure 4290 <conf:(0.94)> lift:(1.51) lev:(0.04) [1444] conv:(6.45)

34. Sex=F BioG10=Failure ChemG10=Failure EngG10=Failure MathG12=Failure 4935 ==> MathG10=Failure 4648 <conf:(0.94)> lift:(1.56) lev:(0.04) [1670] conv:(6.8)

35. ChemG10=Failure ChemG12=Failure PhyG10=Satisfactory success=no 4392 ==> MathG12=Failure 4136 <conf:(0.94)> lift:(1.51) lev:(0.03) [1392] conv:(6.41)

36. BioG10=Failure ChemG12=Failure success=no 5148 ==> MathG12=Failure 4846 <conf:(0.94)> lift:(1.51) lev:(0.04) [1629] conv:(6.38)

37. ChemG12=Failure success=no 7615 ==> MathG12=Failure 7168 <conf:(0.94)> lift:(1.51) lev:(0.06) [2410] conv:(6.38)

38. SchoolType=RG ChemG12=Failure EngG12=Failure PhyG12=Satisfactory 4476 ==> MathG12=Failure 4213 <conf:(0.94)> lift:(1.51) lev:(0.04) [1416] conv:(6.36)

39. EngG12=Failure PhyG12=Satisfactory success=no 5085 ==> MathG12=Failure 4786 <conf:(0.94)> lift:(1.51) lev:(0.04) [1609] conv:(6.36)

40. ChemG10=Failure EngG10=Failure success=no 6114 ==> MathG10=Failure 5754 <conf:(0.94)> lift:(1.56) lev:(0.05) [2065] conv:(6.72)

ANNEX II: Sample rules discovered using Minimum support: 0.1 (4033 instances) Minimum metric <confidence>: 0.7

Best rules found:

1. PhyG12=Verygood 4368 ==> success=yes 4300 <conf:(0.98)> lift:(1.33) lev:(0.03) [1065] conv:(16.43)

2. BioG12=Failure ChemG12=Failure EngG12=Failure 4610 ==> MathG12=Failure 4400 <conf:(0.95)> lift:(1.53) lev:(0.04) [1520] conv:(8.2)

3. BioG12=Failure ChemG12=Failure PhyG12=Satisfactory 4744 ==> MathG12=Failure 4524 <conf:(0.95)> lift:(1.53) lev:(0.04) [1560] conv:(8.06)

4. CivG12=Verygood 5594 ==> success=yes 5330 <conf:(0.95)> lift:(1.29) lev:(0.03) [1187] conv:(5.48)

5. BioG12=Verygood 6294 ==> success=yes 5985 <conf:(0.95)> lift:(1.28) lev:(0.03) [1324] conv:(5.27)

6. ChemG12=Failure MathG10=Failure PhyG12=Satisfactory success=no 4915 ==> MathG12=Failure 4671 <conf:(0.95)> lift:(1.52) lev:(0.04) [1600] conv:(7.53)

7. ChemG12=Failure PhyG10=Satisfactory PhyG12=Satisfactory success=no 4362 ==> MathG12=Failure 4141 <conf:(0.95)> lift:(1.52) lev:(0.04) [1415] conv:(7.37)

8. ChemG12=Failure PhyG12=Satisfactory success=no 5868 ==> MathG12=Failure 5566 <conf:(0.95)> lift:(1.52) lev:(0.05) [1900] conv:(7.27)

9. ChemG10=Failure ChemG12=Failure PhyG12=Satisfactory success=no 4490 ==> MathG12=Failure 4257 <conf:(0.95)> lift:(1.52) lev:(0.04) [1451] conv:(7.2)

10. ChemG10=Failure ChemG12=Failure EngG10=Failure MathG12=Failure success=no 4363 ==> MathG10=Failure 4136 <conf:(0.95)> lift:(1.57) lev:(0.04) [1503] conv:(7.59)

11. ChemG12=Failure CivG12=Failure EngG12=Failure 4497 ==> MathG12=Failure 4262 <conf:(0.95)> lift:(1.52) lev:(0.04) [1452] conv:(7.15)

12. Sex=F ChemG12=Failure success=no 4695 ==> MathG12=Failure 4449 <conf:(0.95)> lift:(1.52) lev:(0.04) [1515] conv:(7.13)

13. ChemG10=Failure ChemG12=Failure EngG10=Failure success=no 4626 ==> MathG10=Failure 4380 <conf:(0.95)> lift:(1.57) lev:(0.04) [1589] conv:(7.43)

14. ChemG10=Failure EngG10=Failure EngG12=Failure success=no 4487 ==> MathG10=Failure 4248 <conf:(0.95)> lift:(1.57) lev:(0.04) [1540] conv:(7.42)

15. Sex=F BioG12=Failure ChemG12=Failure 4315 ==> MathG12=Failure 4083 <conf:(0.95)> lift:(1.51) lev:(0.03) [1387] conv:(6.95)

16. BioG10=Failure ChemG10=Failure EngG10=Failure MathG12=Failure success=no 4501 ==> MathG10=Failure 4258 <conf:(0.95)> lift:(1.57) lev:(0.04) [1542] conv:(7.32)

17. BioG10=Failure ChemG10=Failure EngG10=Failure success=no 5004 ==> MathG10=Failure 4727 <conf:(0.94)> lift:(1.57) lev:(0.04) [1708] conv:(7.14)

18. ChemG10=Failure ChemG12=Failure MathG10=Failure success=no 5359 ==> MathG12=Failure 5062 <conf:(0.94)> lift:(1.51) lev:(0.04) [1714] conv:(6.75)

19. ChemG12=Failure MathG10=Failure success=no 6394 ==> MathG12=Failure 6039 <conf:(0.94)> lift:(1.51) lev:(0.05) [2044] conv:(6.74)

20. SchoolType=RG ChemG12=Failure success=no 4383 ==> MathG12=Failure 4139 <conf:(0.94)> lift:(1.51) lev:(0.03) [1400] conv:(6.71)

21. ChemG10=Failure ChemG12=Failure EngG10=Failure MathG10=Failure success=no 4380 ==> MathG12=Failure 4136 <conf:(0.94)> lift:(1.51) lev:(0.03) [1399] conv:(6.71)

22. BioG10=Failure ChemG12=Failure MathG10=Failure success=no 4647 ==> MathG12=Failure 4384 <conf:(0.94)> lift:(1.51) lev:(0.04) [1480] conv:(6.61)

23. ChemG10=Failure ChemG12=Failure success=no 5812 ==> MathG12=Failure 5483 <conf:(0.94)> lift:(1.51) lev:(0.05) [1852] conv:(6.61)

24. ChemG10=Failure ChemG12=Failure EngG10=Failure success=no 4626 ==> MathG12=Failure 4363 <conf:(0.94)> lift:(1.51) lev:(0.04) [1473] conv:(6.58)

25. ChemG12=Failure EngG10=Failure MathG10=Failure success=no 5028 ==> MathG12=Failure 4742 <conf:(0.94)> lift:(1.51) lev:(0.04) [1600] conv:(6.57)

26. ChemG12=Failure MathG10=Failure PhyG10=Satisfactory success=no 4903 ==> MathG12=Failure 4623 <conf:(0.94)> lift:(1.51) lev:(0.04) [1559] conv:(6.55)

27. ChemG10=Failure EngG10=Failure MathG12=Failure success=no 5525 ==> MathG10=Failure 5209 <conf:(0.94)> lift:(1.56) lev:(0.05) [1875] conv:(6.91)

28. Sex=F ChemG12=Failure EngG12=Failure PhyG12=Satisfactory 4510 ==> MathG12=Failure 4252 <conf:(0.94)> lift:(1.51) lev:(0.04) [1434] conv:(6.53)

29. BioG10=Failure ChemG10=Failure ChemG12=Failure success=no 4473 ==> MathG12=Failure 4216 <conf:(0.94)> lift:(1.51) lev:(0.04) [1421] conv:(6.51)

30. ChemG10=Failure EngG10=Failure PhyG10=Satisfactory success=no 4612 ==> MathG10=Failure 4347 <conf:(0.94)> lift:(1.56) lev:(0.04) [1564] conv:(6.88)

31. ChemG12=Failure EngG12=Failure success=no 5305 ==> MathG12=Failure 5000 <conf:(0.94)> lift:(1.51) lev:(0.04) [1685] conv:(6.51)

32. Sex=F ChemG10=Failure EngG10=Failure EngG12=Failure 4671 ==> MathG10=Failure 4401 <conf:(0.94)> lift:(1.56) lev:(0.04) [1582] conv:(6.84)

33. ChemG12=Failure EngG12=Failure MathG10=Failure success=no 4554 ==> MathG12=Failure 4290 <conf:(0.94)> lift:(1.51) lev:(0.04) [1444] conv:(6.45)

34. Sex=F BioG10=Failure ChemG10=Failure EngG10=Failure MathG12=Failure 4935 ==> MathG10=Failure 4648 <conf:(0.94)> lift:(1.56) lev:(0.04) [1670] conv:(6.8)

35. ChemG10=Failure ChemG12=Failure PhyG10=Satisfactory success=no 4392 ==> MathG12=Failure 4136 <conf:(0.94)> lift:(1.51) lev:(0.03) [1392] conv:(6.41)

36. BioG10=Failure ChemG12=Failure success=no 5148 ==> MathG12=Failure 4846 <conf:(0.94)> lift:(1.51) lev:(0.04) [1629] conv:(6.38)

37. ChemG12=Failure success=no 7615 ==> MathG12=Failure 7168 <conf:(0.94)> lift:(1.51) lev:(0.06) [2410] conv:(6.38)

38. SchoolType=RG ChemG12=Failure EngG12=Failure PhyG12=Satisfactory 4476 ==> MathG12=Failure 4213 <conf:(0.94)> lift:(1.51) lev:(0.04) [1416] conv:(6.36)

39. EngG12=Failure PhyG12=Satisfactory success=no 5085 ==> MathG12=Failure 4786 <conf:(0.94)> lift:(1.51) lev:(0.04) [1609] conv:(6.36)

40. ChemG10=Failure EngG10=Failure success=no 6114 ==> MathG10=Failure 5754 <conf:(0.94)> lift:(1.56) lev:(0.05) [2065] conv:(6.72)

ANNEX III: Sample rules discovered using Minimum support: 0.15 (6049 instances) Minimum metric <confidence>: 0.5

Best rules found:

1. ChemG12=Failure success=no 7615 ==> MathG12=Failure 7168 <conf:(0.94)> lift:(1.51) lev:(0.06) [2410] conv:(6.38)

2. ChemG12=Failure EngG12=Failure PhyG12=Satisfactory 6974 ==> MathG12=Failure 6544 <conf:(0.94)> lift:(1.5) lev:(0.05) [2187] conv:(6.07)

3. EngG12=Good 6992 ==> success=yes 6523 <conf:(0.93)> lift:(1.26) lev:(0.03) [1345] conv:(3.86)

4. ChemG10=Failure EngG10=Failure EngG12=Failure 7185 ==> MathG10=Failure 6691 <conf:(0.93)> lift:(1.54) lev:(0.06) [2356] conv:(5.76)

5. ChemG10=Failure ChemG12=Failure EngG10=Failure 7304 ==> MathG10=Failure 6800 <conf:(0.93)> lift:(1.54) lev:(0.06) [2393] conv:(5.74)

6. BioG10=Failure ChemG10=Failure EngG10=Failure MathG12=Failure 7661 ==> MathG10=Failure 7123 <conf:(0.93)> lift:(1.54) lev:(0.06) [2501] conv:(5.64)

7. EngG12=Failure success=no 6688 ==> MathG12=Failure 6215 <conf:(0.93)> lift:(1.49) lev:(0.05) [2036] conv:(5.29)

8. PhyG12=Satisfactory success=no 7688 ==> MathG12=Failure 7129 <conf:(0.93)> lift:(1.48) lev:(0.06) [2326] conv:(5.15)

9. Sex=F ChemG10=Failure EngG10=Failure 7410 ==> MathG10=Failure 6868 <conf:(0.93)> lift:(1.54) lev:(0.06) [2397] conv:(5.41)

10. BioG10=Failure ChemG10=Failure EngG10=Failure 9782 ==> MathG10=Failure 9054 <conf:(0.93)> lift:(1.53) lev:(0.08) [3152] conv:(5.32)

11. BioG10=Failure ChemG10=Failure EngG10=Failure PhyG10=Satisfactory 7358 ==> MathG10=Failure 6796 <conf:(0.92)> lift:(1.53) lev:(0.06) [2356] conv:(5.18)

12. ChemG10=Failure EngG10=Failure MathG12=Failure PhyG12=Satisfactory 6955 ==> MathG10=Failure 6416 <conf:(0.92)> lift:(1.53) lev:(0.06) [2220] conv:(5.11)

13. SchoolType=RG ChemG10=Failure EngG10=Failure 6961 ==> MathG10=Failure 6420 <conf:(0.92)> lift:(1.53) lev:(0.06) [2220] conv:(5.09)

14. SchoolType=RG ChemG12=Failure PhyG12=Satisfactory 7267 ==> MathG12=Failure 6700 <conf:(0.92)> lift:(1.48) lev:(0.05) [2160] conv:(4.8)

15. ChemG10=Failure EngG10=Failure MathG12=Failure 9644 ==> MathG10=Failure 8889 <conf:(0.92)> lift:(1.53) lev:(0.08) [3070] conv:(5.06)

16. BioG10=Failure ChemG12=Failure PhyG12=Satisfactory 6863 ==> MathG12=Failure 6325 <conf:(0.92)> lift:(1.48) lev:(0.05) [2037] conv:(4.78)

17. ChemG10=Failure EngG10=Failure PhyG12=Satisfactory 7994 ==> MathG10=Failure 7361 <conf:(0.92)> lift:(1.53) lev:(0.06) [2538] conv:(5)

18. ChemG10=Failure EngG10=Failure MathG12=Failure PhyG10=Satisfactory 7301 ==> MathG10=Failure 6718 <conf:(0.92)> lift:(1.53) lev:(0.06) [2313] conv:(4.96)

19. Sex=F ChemG12=Failure PhyG12=Satisfactory 7512 ==> MathG12=Failure 6882 <conf:(0.92)> lift:(1.47) lev:(0.05) [2189] conv:(4.47)

20. EngG10=Failure MathG12=Failure success=no 6624 ==> MathG10=Failure 6066 <conf:(0.92)> lift:(1.52) lev:(0.05) [2069] conv:(4.7)

21. ChemG12=Satisfactory PhyG12=Good 6989 ==> success=yes 6397 <conf:(0.92)> lift:(1.24) lev:(0.03) [1221] conv:(3.06)

22. ChemG10=Failure EngG10=Failure 12568 ==> MathG10=Failure 11484 <conf:(0.91)> lift:(1.51) lev:(0.1) [3901] conv:(4.6)

23. ChemG10=Failure EngG10=Failure PhyG10=Satisfactory 9528 ==> MathG10=Failure 8702 <conf:(0.91)> lift:(1.51) lev:(0.07) [2953] conv:(4.57)

24. ChemG10=Failure MathG12=Failure success=no 6980 ==> MathG10=Failure 6372 <conf:(0.91)> lift:(1.51) lev:(0.05) [2160] conv:(4.55)

25. EngG10=Failure success=no 7366 ==> MathG10=Failure 6723 <conf:(0.91)> lift:(1.51) lev:(0.06) [2279] conv:(4.54)

26. Sex=F BioG10=Failure EngG10=Failure 6980 ==> MathG10=Failure 6367 <conf:(0.91)> lift:(1.51) lev:(0.05) [2155] conv:(4.51)

27. ChemG10=Satisfactory ChemG12=Satisfactory 6879 ==> success=yes 6273 <conf:(0.91)> lift:(1.23) lev:(0.03) [1179] conv:(2.94)

28. ChemG10=Failure ChemG12=Failure MathG10=Failure PhyG12=Satisfactory 6780 ==> MathG12=Failure 6180 <conf:(0.91)> lift:(1.46) lev:(0.05) [1944] conv:(4.23)

29. ChemG10=Failure ChemG12=Failure PhyG12=Satisfactory 7680 ==> MathG12=Failure 6996 <conf:(0.91)> lift:(1.46) lev:(0.05) [2198] conv:(4.21)

30. BioG10=Failure ChemG12=Failure EngG10=Failure 6733 ==> MathG10=Failure 6133 <conf:(0.91)> lift:(1.51) lev:(0.05) [2070] conv:(4.44)

31. ChemG10=Failure success=no 7745 ==> MathG10=Failure 7049 <conf:(0.91)> lift:(1.51) lev:(0.06) [2376] conv:(4.41)

32. ChemG12=Failure EngG12=Failure 9550 ==> MathG12=Failure 8687 <conf:(0.91)> lift:(1.46) lev:(0.07) [2720] conv:(4.15)

33. BioG10=Satisfactory ChemG12=Satisfactory 6930 ==> success=yes 6301 <conf:(0.91)> lift:(1.23) lev:(0.03) [1169] conv:(2.86)

34. ChemG12=Failure MathG10=Failure PhyG12=Satisfactory 8952 ==> MathG12=Failure 8139 <conf:(0.91)> lift:(1.46) lev:(0.06) [2546] conv:(4.13)

35. ChemG12=Failure PhyG12=Satisfactory 12509 ==> MathG12=Failure 11369 <conf:(0.91)> lift:(1.45) lev:(0.09) [3554] conv:(4.11)

36. ChemG12=Failure PhyG10=Satisfactory PhyG12=Satisfactory 9038 ==> MathG12=Failure 8213 <conf:(0.91)> lift:(1.45) lev:(0.06) [2566] conv:(4.11)

37. ChemG12=Failure MathG10=Failure PhyG10=Satisfactory PhyG12=Satisfactory 6945 ==> MathG12=Failure 6306 <conf:(0.91)> lift:(1.45) lev:(0.05) [1967] conv:(4.07)

38. BioG10=Failure ChemG10=Failure EngG12=Failure 6916 ==> MathG10=Failure 6272 <conf:(0.91)> lift:(1.5) lev:(0.05) [2099] conv:(4.25)

39. BioG10=Failure ChemG10=Failure ChemG12=Failure MathG12=Failure 6844 ==> MathG10=Failure 6206 <conf:(0.91)> lift:(1.5) lev:(0.05) [2076] conv:(4.25)

40. EngG12=Failure PhyG12=Satisfactory 9538 ==> MathG12=Failure 8646 <conf:(0.91)> lift:(1.45) lev:(0.07) [2687] conv:(4.01)

ANNEX IV: Sample rules discovered using Minimum support: 0.2 (8066 instances) Minimum metric <confidence>: 0.3

Best rules found:

1. BioG10=Failure ChemG10=Failure EngG10=Failure 9782 ==> MathG10=Failure 9054 <conf:(0.93)> lift:(1.53) lev:(0.08) [3152] conv:(5.32)

2. ChemG10=Failure EngG10=Failure MathG12=Failure 9644 ==> MathG10=Failure 8889 <conf:(0.92)> lift:(1.53) lev:(0.08) [3070] conv:(5.06)

3. ChemG10=Failure EngG10=Failure 12568 ==> MathG10=Failure 11484 <conf:(0.91)> lift:(1.51) lev:(0.1) [3901] conv:(4.6)

4. ChemG10=Failure EngG10=Failure PhyG10=Satisfactory 9528 ==> MathG10=Failure 8702 <conf:(0.91)> lift:(1.51) lev:(0.07) [2953] conv:(4.57)

5. ChemG12=Failure EngG12=Failure 9550 ==> MathG12=Failure 8687 <conf:(0.91)> lift:(1.46) lev:(0.07) [2720] conv:(4.15)

6. ChemG12=Failure MathG10=Failure PhyG12=Satisfactory 8952 ==> MathG12=Failure 8139 <conf:(0.91)> lift:(1.46) lev:(0.06) [2546] conv:(4.13)

7. ChemG12=Failure PhyG12=Satisfactory 12509 ==> MathG12=Failure 11369 <conf:(0.91)> lift:(1.45) lev:(0.09) [3554] conv:(4.11)

8. ChemG12=Failure PhyG10=Satisfactory PhyG12=Satisfactory 9038 ==> MathG12=Failure 8213 <conf:(0.91)> lift:(1.45) lev:(0.06) [2566] conv:(4.11)

9. EngG12=Failure PhyG12=Satisfactory 9538 ==> MathG12=Failure 8646 <conf:(0.91)> lift:(1.45) lev:(0.07) [2687] conv:(4.01)

10. BioG10=Failure EngG10=Failure MathG12=Failure 9021 ==> MathG10=Failure 8134 <conf:(0.9)> lift:(1.49) lev:(0.07) [2691] conv:(4.03)

11. MathG12=Satisfactory 10657 ==> success=yes 9587 <conf:(0.9)> lift:(1.21) lev:(0.04) [1695] conv:(2.58)

12. BioG10=Failure EngG10=Failure 11620 ==> MathG10=Failure 10400 <conf:(0.9)> lift:(1.48) lev:(0.08) [3389] conv:(3.78)

13. BioG10=Failure ChemG10=Failure MathG12=Failure 10409 ==> MathG10=Failure 9275 <conf:(0.89)> lift:(1.48) lev:(0.07) [2995] conv:(3.64)

14. CivG12=Good 9256 ==> success=yes 8242 <conf:(0.89)> lift:(1.2) lev:(0.03) [1388] conv:(2.37)

15. success=no 10467 ==> MathG12=Failure 9307 <conf:(0.89)> lift:(1.42) lev:(0.07) [2767] conv:(3.38)

16. ChemG12=Failure EngG10=Failure 9217 ==> MathG10=Failure 8180 <conf:(0.89)> lift:(1.47) lev:(0.06) [2619] conv:(3.52)

17. EngG10=Failure EngG12=Failure 9100 ==> MathG10=Failure 8076 <conf:(0.89)> lift:(1.47) lev:(0.06) [2585] conv:(3.52)

18. ChemG10=Failure ChemG12=Failure MathG12=Failure 9400 ==> MathG10=Failure 8319 <conf:(0.89)> lift:(1.47) lev:(0.07) [2647] conv:(3.45)

19. Sex=F EngG10=Failure 9359 ==> MathG10=Failure 8255 <conf:(0.88)> lift:(1.46) lev:(0.06) [2608] conv:(3.36)

20. BioG10=Failure ChemG10=Failure 13636 ==> MathG10=Failure 12025 <conf:(0.88)> lift:(1.46) lev:(0.09) [3798] conv:(3.36)

21. BioG10=Failure ChemG12=Failure 9599 ==> MathG12=Failure 8455 <conf:(0.88)> lift:(1.41) lev:(0.06) [2458] conv:(3.15)

22. BioG10=Failure ChemG10=Failure PhyG10=Satisfactory 10344 ==> MathG10=Failure 9100 <conf:(0.88)> lift:(1.46) lev:(0.07) [2859] conv:(3.3)

23. ChemG10=Failure ChemG12=Failure 10915 ==> MathG10=Failure 9595 <conf:(0.88)> lift:(1.46) lev:(0.07) [3009] conv:(3.28)

24. SchoolType=RNG ChemG12=Satisfactory 9455 ==> success=yes 8310 <conf:(0.88)> lift:(1.19) lev:(0.03) [1309] conv:(2.14)

25. CivG10=Good 9311 ==> success=yes 8139 <conf:(0.87)> lift:(1.18) lev:(0.03) [1244] conv:(2.06)

26. EngG10=Failure MathG12=Failure 12593 ==> MathG10=Failure 11002 <conf:(0.87)> lift:(1.45) lev:(0.08) [3404] conv:(3.14)

27. EngG10=Failure MathG12=Failure PhyG10=Satisfactory 9602 ==> MathG10=Failure 8388 <conf:(0.87)> lift:(1.45) lev:(0.06) [2595] conv:(3.14)

28. EngG10=Failure PhyG12=Satisfactory 10505 ==> MathG10=Failure 9146 <conf:(0.87)> lift:(1.44) lev:(0.07) [2808] conv:(3.06)

29. BioG10=Failure EngG10=Failure MathG10=Failure 10400 ==> ChemG10=Failure 9054 <conf:(0.87)> lift:(1.75) lev:(0.1) [3886] conv:(3.88)

30. SchoolType=RG ChemG12=Failure 10174 ==> MathG12=Failure 8849 <conf:(0.87)> lift:(1.39) lev:(0.06) [2493] conv:(2.88)

31. ChemG10=Failure MathG12=Failure PhyG12=Satisfactory 10298 ==> MathG10=Failure 8944 <conf:(0.87)> lift:(1.44) lev:(0.07) [2731] conv:(3.01)

32. BioG12=Satisfactory ChemG10=Failure 9585 ==> MathG10=Failure 8318 <conf:(0.87)> lift:(1.44) lev:(0.06) [2535] conv:(3)

33. MathG10=Satisfactory 11441 ==> success=yes 9925 <conf:(0.87)> lift:(1.17) lev:(0.04) [1453] conv:(1.96)

34. ChemG10=Failure ChemG12=Failure MathG10=Failure 9595 ==> MathG12=Failure 8319 <conf:(0.87)> lift:(1.39) lev:(0.06) [2324] conv:(2.82)

35. ChemG10=Failure MathG12=Failure 14538 ==> MathG10=Failure 12586 <conf:(0.87)> lift:(1.43) lev:(0.09) [3815] conv:(2.95)

36. ChemG10=Failure MathG12=Failure PhyG10=Satisfactory 11145 ==> MathG10=Failure 9646 <conf:(0.87)> lift:(1.43) lev:(0.07) [2922] conv:(2.95)

37. PhyG12=Good 12549 ==> success=yes 10851 <conf:(0.86)> lift:(1.17) lev:(0.04) [1559] conv:(1.92)

38. ChemG10=Failure PhyG12=Satisfactory 12037 ==> MathG10=Failure 10402 <conf:(0.86)> lift:(1.43) lev:(0.08) [3140] conv:(2.92)

39. SchoolType=RG PhyG12=Satisfactory 11670 ==> MathG12=Failure 10082 <conf:(0.86)> lift:(1.38) lev:(0.07) [2791] conv:(2.76)

40. Sex=F ChemG12=Failure 10657 ==> MathG12=Failure 9205 <conf:(0.86)> lift:(1.38) lev:(0.06) [2547] conv:(2.75)

ANNEX V: Attributes Discretization

Attribute name	Original value	Discretized value	Transformed value	Number of instances	%
Biology G12 (BioG12)	14 – 99	90 – 100	Excellent	1366	3.3%
		80 – 89	Very good	6294	15.6%
		70 – 79	Good	8670	44.1%
		50 – 69	Satisfactory	15944	17.0%
		Under 50	Failure	8054	20.0%
Biology G10 (BioG10)	11 – 99	90 – 100	Excellent	527	1.3%
		80 – 89	Very good	2814	7.0%
		70 – 79	Good	4826	12.0%
		50 – 69	Satisfactory	14771	36.6%
		Under 50	Failure	17390	43.1%
ChemistryG12 (ChemG12)	13 – 98	90 – 100	Excellent	77	0.2%
		80 – 89	Very good	857	2.1%
		70 – 79	Good	3347	8.3%
		50 – 69	Satisfactory	17974	44.6%
		Under 50	Failure	18073	44.8%
ChemistryG10 (ChemG10)	13 – 100	90 – 100	Excellent	455	1.1%
		80 – 89	Very good	2044	5.1%
		70 – 79	Good	3870	9.6%

		50 – 69	Satisfactory	13920	34.5%
		Under 50	Failure	20039	49.7%
CivicsG12 (CivG12)	16 – 98	90 – 100	Excellent	1186	2.9%
		80 – 89	Very good	5594	13.9%
		70 – 79	Good	9256	23.0%
		50 – 69	Satisfactory	16398	40.7%
		Under 50	Failure	7894	19.6%
CivicsG10 (CivG10)	13 – 98	90 – 100	Excellent	497	1.2%
		80 – 89	Very good	4756	11.8%
		70 – 79	Good	9311	23.1%
		50 – 69	Satisfactory	18384	45.6%
		Under 50	Failure	7380	18.3%
EnglishG12 (EngG12)	4 – 96	90 – 100	Excellent	38	.09%
		80 – 89	Very good	2075	5.1%
		70 – 79	Good	6992	17.3%
		50 – 69	Satisfactory	17223	42.7%
		Under 50	Failure	14000	34.7%
English10 (EngG10)	13 – 96	90 – 100	Excellent	386	1.0%
		80 – 89	Very good	3394	8.4%
		70 – 79	Good	6096	15.1%
		50 – 69	Satisfactory	13860	34.4%

			Under 50	Failure	16592	41.1%
			80 – 89	Very good	1577	3.9%
			70 – 79	Good	2682	6.7%
			50 – 69	Satisfactory	10657	26.4%
			Under 50	Failure	25203	62.5%
	MathsG10 (MathG10)	13 – 100	90 – 100	Excellent	314	0.8%
			80 – 89	Very good	1412	3.5%
			70 – 79	Good	2827	7.0%
			50 – 69	Satisfactory	11441	28.4%
			Under 50	Failure	24334	60.3%
	PhysicsG12 (PhysG12)	4 – 98	90 – 100	Excellent	88	0.2%
			80 – 89	Very good	708	1.8%
			70 – 79	Good	1711	4.2%
			50 – 69	Satisfactory	9383	23.3%
	PhysicsG10 (PhysG10)	2 – 93	Under 50 90 – 100	Failure Excellent	28438 4	70.5% .001%
			80 – 89	Very good	24	0.1%
			70 – 79	Good	225	0.6%
			50 – 69	Satisfactory	3765	9.3%
			Under 50	Failure	36310	90.0%

References

[1] Oli Negasa, "ethiopian students' achievement challenges in science education:implications to policy formulation".

[2] M. Tesfa, "The Validity of university entrance examination and high school grade point average for predicting first year university students' academic performance," 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. 2. 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] Dr. Pranav Patil, "a study of students' academic performance using data mining techniques," vol. 3, no. 9, 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] A. L. Diaz, "Personal family and academic factors affecting low achievement in secondary school," *electronic journal of reserach in educational psychology and psycopedagogy.*

[10] D. Abebe, "predicting students' academic performance in higher education: the case of Jimma university," Jimma, 2013.

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

[12] A. Merceron, "Interestingness Measures for Association Rules in Educational Data," 2014.

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

[14] A. Bala, "Performance Analysis of Apriori and FP-Growth Algorithms(Association rule mining)," *International Journal of Computer Technology & Applications,* vol. 7, no. 2, pp. 279-293, 2016.

[15] S. S. Ubha, "Data Mining for Prediction of students' performance in the secondary schools of the state of Punjab," *International Journal of Innovative research in computer and communication engineering,* vol. 4, no. 8, 2016.

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

[17] C. Bambrah, "mining association rules in student assesement data," vol. 3, no. 3, 2014.

[18] 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.

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

[20] K. Hernawati, "Application Of Association Rules With Apriori Algorithm To Determine The Pattern Of The Relationship Between SBMPTN Database And Student's Grade Point Average," 2014.

[21] H. Graham, "modeling the KDD process," 1996.

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

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

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

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

[26] "Education system Ethiopia," Nuffic, Nedereland, 2015.

[27] K. Charanjeet, "Association rule mining using apriori algorithm : a survey," *international journal of advanced research in computer engineering and technology(IJARCE),* vol. 2, no. 6, 2013.

[28] D. Hand, Principles of Data Mining, London, 2004.

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

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

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

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

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

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

[35] M. Durairaj, "Educational Data mining for Prediction of Student Performance Using Clustering Algorithms," *International Journal of Computer Science and Information Technologies,* vol. 5, no. 4, 2014.

[36] F. E.Harrell, Regression modeling startegies, Springer International, 2001.

[37] F. Berhanu, "Students' Performance Prediction Based on their Academic Record," *International Journal of Computer Applications,* vol. 131, no. 5, 2015.

[38] L. Kurgan, Trends in Data Mining and Knowledge Discovery, London, 2005.

[39] A. Marchesi, "Evaluation in secondary education ,snap shot from controversial era," *International Journal of Research and Method in Education,* vol. 4, no. 4, 2002.

[40] M. Fabunmi, "Class Factors as Determinants of Secondary School Student's Academic Performance in Oyo State, Nigeria,," *Journal of Social Science,* vol. 14, no. 3, pp. 243-247, 2007.

[41] J. Jha, "Educational Data Mining Using Improved Apriori algorithm," vol. 3, no. 5, 2013.

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

[43] B. Minaei-Bidgol, "DATA MINING FOR A WEB-BASED EDUCATIONAL SYSTEM," Michigan, 2004.

[44] Mohammadreza, "Prediction of students' performance in high school by data mining classification techniques," *International Journal of Science and Engineering,* vol. 2, no. 7, pp. 25-33, 2015.

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

[46] K. Verma, "A review on predicting students' performance using data ming methods," vol. 3, no. 1, 2016.

[47] R. J. a. A. Verspoor, "Secondary Education in Ethiopia :supporting growth and transformation," Addis Ababa, 2012.

[48] M. D. Vijitha, "Educational Data mining for Prediction of Student Performance Using Clustering Algorithms," *International Journal of Computer Science and Information Technologies,* vol. 5, no. 4, 2014.

[49] W. Novitasari, "A Method of Discovering Interesting Association Rules from Student Admission Dataset," *International Journal of Software Engineering and Its Applications,* vol. 9, no. 8, 2015.

[50] E. Osmanbegovic, "Data Mining approach for Predicting Students Performance," *economic review – journal of economics and business,* vol. 10, no. 1, 2012.

[51] U. Fayyad, "from data mining to knowledge discovery databases AI Magazine," vol. 17, no. 3, 1996.

[52] S. Martin, "Data Mining - Association Analysis," vol. 1, no. 1, 2009.

[53] A. Dagnaw, "Impact of School Climate on Students' Academic Achievement in Bahir Dar Secondary Schools: Ethiopia," Bahir Dar, 2014.

[54] A. D. Magdalene, "Association rule generation for student performance analysis using apriori algorithm," vol. 1, no. 1, 2013.

[55] R. Asif, "Predicting student academic performance at degree level," *International Journal of Intelligent systems and applications,* vol. 01, no. 1, pp. 49-61, 2015.

[56] L. A. Kurgan, "The Knowledge Engineering Review," vol. 21, no. 1, pp. 1-24, 2012.

[57] 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.

[58] S. Ougiaroglou, "association rule mining from the educational data of ESOG web based application," vol. 3, no. 3, 2013.

[59] UNESCO, "The development of education national report of Ethiopia," Addis Ababa, 2001.