A Correlation Study of Students' Performance in Ethiopian Secondary Schools by Using Data Mining Algorithm

Thesis presented to Addis Ababa University

College of Natural science

Department Of Information Science

In Partial Fulfillment of the Requirements for the Degree Master of Science

---

By: Solomon Wendimu

Thesis Advisor: Temtim Assefa (Phd)

Submission Date: JUNE, 2016

Addis Ababa University
School Of Graduate Studies
School Of Information Science

A Correlation Study of Students' Performance in Ethiopian Secondary Schools by Using Data Mining Algorithm

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in information science

By

Solomon wendimu

JUNE 2016
ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

A Correlation Study of Students' Performance in Ethiopian Secondary Schools by Using Data Mining Algorithm

By

Solomon Wendimu

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
______________________	Examiner	______________________	______________________
______________________	Examiner	______________________	______________________
______________________	Chair person	______________________	______________________
______________________	Advisor	______________________	______________________

Declaration

I declare that the thesis is my original work and has not been presented for a degree in any other university.

Name: __________________Signature: ______________Date: _________________

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

Name: _________________Signature: ______________Date: __________________

ACKNOWLEDGEMENTS

I would like to thank all those who have helped me to accomplish this thesis. First, I gratefully express my deepest thanks to the almighty God for his guidance, help, support and who added days to my age, Glory to God.

Thank you to my professor, colleagues, family and friends whose support and guidance has helped me to accomplish my goals of completing this thesis. I would particularly like to thank Dr. Temtim Assefa for your assistance in completing this thesis and helping me in every step I took in my study and research. Your encouragement and advice is not only helpful for my thesis, but is also invaluable for my future endeavors.

Dr. Gebeyehu, thank you for giving me the chance to study in information science Department. It was great experience to work with you. I would like to thank Ato Yosefe and Ato Bereket for providing me the data and expert opinions for the study. In addition, I really appreciate and I would also like to thank my family, friends, and colleagues who have helped me to make the most of my graduate school experience. I want to thank my parents for their great love and support throughout my life.

Contents

List of Figures

List of Tables

Cases

List of Acronym

NEAEA………………. **National Educational Assessment and Examination Agency**

FDRE………………… **The Federal Democratic Republic of Ethiopia**

ESDP IV……………... **Education Sector Development Program four**

GTP………………….. **The Growth Transformation Plan**

MOE…………………. **Ministry of education**

CEICT………………. **Center for Educational Information and communication Technology**

CRISP-DM………….. **CRoss-Industry Standard Process for Data Mining**

EHEECE…………… **Ethiopian Higher Education Entrance Certificate Examination**

N.C.C.E…………….. **Nigerian Certificate of Education**

PTV………………….. **Plasma Television**

ABSTRACT

In recent years, the biggest challenges that educational institutions are facing the explosive growth of educational data and to use this data to improve the quality of managerial decisions. Educational institutions are playing an important role in our society and playing a vital role for growth and development of nation. An analysis of student's performance in educational environments is also important as well. Student's academic performance is based upon various factors like personal characteristics and psychological factors. Educational database contain useful information for predicting a students' performance, rank factor and details. The data mining techniques are more helpful in analyses of educational data and developing methods for discovering knowledge from data that comes from educational institutions. Extracting the association rules between variables has allow better decision making which can facilitate better resource utilization in terms of educational service delivery.

This study aims to analyses and predict the correlation between English, Mathematics and science subjects in terms of student academic result in grade 10 and 12 by *using* Aprior data mining techniques which mines required information. The data used for the study was grade 10 and 12 students' national examination result in 2005 E.C and 2007 E.C respectively. The findings of the research showed that there is a positive relationship between English, Mathematics and Science subjects excluding physics. The academic achievement result of male students in both five subjects in both grade 10 and 12 is better than female students. This implies that government and concerned stakeholders in education sector should provide additional support for female students. So the present education system may adopt this as input revision of its educational policy.

CHAPTER ONE

INTRODUCTION

1.1. Background

The National Educational Assessment and Examination Agency (NEAEA) documents all examination results in hard and soft copies in the agency's office. The agency has various objectives which includes determine students who are eligible to be promoted to the next level of Education through the administration of national exams, making equitable students placements with regard to students admission to public higher education institutions, as well as conducting educational assessment to evaluate the education system and assist the relevant regional office in developing their capacities. Identifying subjects eligible for national exams and up on approval implement the same; certificates of exam results to examinees; make students placements in public higher education institutions by taking into account the interests of the students, their exam results and the institutions; intake capacities, collecting service fees in accordance with the rate approved by the government; and finally undertake other related activities that are conducive to the attainment of its objectives.

The national examination is one of the forms of evaluation done on a national scale in the education world and adjusted with the national achievement standards. The national exam is carried out at high school at grade 10 and preparatory school levels on grade 12 to measure student graduate competence nationally in certain subjects as well as to map the level of achievement for students at the school level and the regional level.

NEAEA conducted three consecutive learning assessments in 2000, 2003/2004 and 2008 respectively to measure learning achievements of Grade 8 students and identify the factors that determine those achievements. It also aimed at providing comparative information on school improvement. But the target population was Grade 8 students in each study in the country.

Data mining is data analysis methodology used to identify hidden patterns in a large data set. It has been successfully used in different areas including the educational environment. Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process. Evaluating students' performance is a complex issue, which can't be restricted for the grading. Reasons of good or bad performances belong to the main interests of teachers, because they can plan and customize their teaching program, based on the feedback of policy makers. Data mining is one of the approaches, which can provide an effective assistance in revealing complex relationships behind the grades. Hence, in the present study I examined the extent to which the academic achievement of Grade 12 students is related to achievement of grade 10 and association between sciences, English and mathematics subjects by using Aprior algorithm.

1.2. Statement of the Problem

The government of Ethiopia determined and introduced what is known as a 70:30 (Ministry of Education, 2014)professional mix that is 70% science and technology and 30% social science and humanities stream at higher education. This showed that the Ethiopian government has given high concentration to science education. But the current status of science education in primary school is not the same as the expectation as (NEAEA assessment report, 2008). Hence, to study the status of science education at secondary school is essential to determine the factors and plan possible intervention to improve the quality.

The Ministry of Education is committed to improve the quality of education along with its effort to increase enrollment and access. This commitment for the improving the quality of education by including quality outcome such as student achievement score as performance monitoring indicator is reflected in education sector development program IV(ESDPIV,2008). It is also reaffirmed in the general education quality improvement program (GEQIP). Performance of students in selected grade level can indicate the strength and the weakness in student's learning within the curriculum area and can show how the intended curricula are implemented in school. Data on important student, teacher and school background, student achievement and other factors important to policy makers to identify major area of intervention.

Some developing countries have tried to assess and measure student achievement and improve their educational system, nevertheless, most countries still apply public examination for certification, selection and promotion. Generally improving student learning has remained one of the most desired goals of educational process.

Adeboye (1993) conducted a study to identify relational proficiency in English Language as a factor contributing to competency in Mathematics. Similarly he conduct another study in 1997 showed correlation study of the student' performance in English language and Science subject through analysis of student exam result.

However, all those previous were conducted by using a very small proportion of the datasets. Since the analysis made by using traditional methods focuses on problems with much more manageable number of variables and cases than may be encountered in real world datasets. Traditional methods have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional relational databases. The other reason is Data Mining techniques tend to be more robust for real-world messy data and also used less by expert users (Plate, 1997).

Thus the goal of this study is to analyses students' academic performance in term of grade 10 and 12 students' national examination result by applying association algorithms. This study is tried to build a model and addressed the factors associated with students' academic achievement using NEAEA data set by applying data mining Techniques

1.3. Research question

This study attempts to investigate and seek answers to the following research questions:

- Is there a strong relationship among English, Mathematics, Chemistry, Physics and Biology subjects?
- How demographic factors influence student achievement?
- Is there variation on student achievement across different regions?
- Does grade 10 students' result related with grade 12 result?
- Does instructional television program through plasma TV affects students' performance?

1.4. Objective of the study

1.4.1. General objective of the study

The main purpose of conducting this research is to analyses learning attainments of students by investigate the relationship students' grade 10 and 12 national examination results and identify the factors that determine those attainments.

1.4.2. Specific objective

The specific objectives of the study:-

- ➢ To show the extent of association between English and science subject's results across subgroups: gender, geographical location, age and educational satellite television program on student achievement in grade 10 and 12.
- ➢ To show the association of grade 10 and grade 12?
- ➢ Identify the association among the subjects English, mathematics, biology, chemistry and physics on student performance.

- To evaluate the effectiveness of educational media technology (plasma) to improve the students achievement.

1.5. Scope and limitations of the Research

Because of unavailability of organized data related to plasma in earlier year's data, this study will be conducted limited in result of Ethiopian student national examination of grade 10 and 12 announced last year (2005 and 2007 E.C respectively). Therefore, the scope of this research is limited to develop a model for the relationship between grade 10 and 12 student assessment result. Identify the relation between student achievement scores of grade 10 on grade 12 score. Identify the interdependency among the subjects English, mathematics, biology, chemistry and physics. Moreover the study was limited to the development of model by generating association rule, and does not incorporate the deployment phase of the model.

1.6. Significance of the Study

Data mining technology provides a user oriented approach to novel and hidden patterns in the data. The discovered patterns (knowledge) can be used by educational administrators at high level toward improving quality in all areas and in particular towards student learning achievement. Predicting the dependency of student learning achievement score of grade 12 on grade 10, determine the significance correlation between science subjects and how factors like geographical location, gender, age, teachers' qualification and instructional media technology have an impact on student academic performance and further more extract knowledge about inter dependency among the five subjects.

The primary goal of this research is to build a predictive model by using data mining techniques. The predicting of student learning achievement will be based upon demographic, parental, and pedagogical factors. The models and rules can support the national educational policy in revising the existing rules, and inducing new rules and policies. It is also possible to use the research result, the model, as a framework for improving quality of education. This type of model building might have an application provide an effective assistance in revealing complex relationships behind the grades. Education sector planners, policy makers, and decision makers as a decision support aid in planning and implementing educational intervention programs aimed at improving quality of education in the regions as well as in the country.

1.7. Organization of the Paper

The thesis is structured into six chapters. The first chapter is an introduction part, which contains background to the research work, statement of the problem addressed, objective of the research, scope and Significance of the research and methodologies adopted for the study.

The second chapter is dealt about literature review on data mining technology, methods/techniques used, and its application in the education sector.

The third chapter is devoted to give understanding about the data mining tool, techniques and algorithms that applied in the study. In this chapter, issues related to data analysis tool and how the algorithm under the technique work for the NEAEA database is addressed and discussed about the selection and preparation of data process that is undertaken in the research work.

The fourth chapter is mainly used for understanding the process in preparing the data for producing quality data using the Weka Aprior algorithm and measure of interestingness.

The fifth chapter is provided discussions about the experimentation and result analysis in different data mining steps that is undertaken in the research work. This chapter is mainly used for describing the experiment in the data prepared till discussing results. This includes extracting the most important rules and testing results obtained by using Weka software. Results are also analyzed and interpreted.

Finally, chapter six provides concluding remarks and pointers for future work to explore and discover knowledge on educational data in general and crucial statistic data in particular.

Chapter two

Literature review

2.1. Overview students' academic performance

The challenge facing Ethiopia today is not only providing equitable access to its growing student population, but also to ensure that education imparted to these students is effective in supplying them with the necessary skills to fully contribute to the development of the society and economy. Improving the quality of education require a multi directional approach including the effectiveness of the learning process by improving teaching techniques, supply better learning materials and teaching, motivating students to attend school etc. learning assessments, which are credible and objective, play important roles in this process by providing critical feedback on what and how well students are learning ESDP IV (educational sector development program IV, 2010)

Student learning assessment involve a systematic process of collecting relevant, valid and timely information about the outcome of schooling so that decision are made about the learning and development of students, curriculum, educational program and educational policy. Student learning assessment provides the necessary feedback and objective evidence required to maximize the outcome of educational efforts. Such assessment summarizes what learner knows, understand and can do in relation to some or all of learning goals determined in the curricula (Education and training policy, 2002). Over the last decade, substantial attempts have been made education, improve access, equity and efficiency in Ethiopia. Now the emphasis has also shifted toward improving quality in all areas and in particular towards student learning achievement. This study therefore, provides an indication or feedback of where students' achievements stand.

The national examination is one of the forms of evaluation done on a national scale in the education world and adjusted with the national achievement standards. The national exam is carried out at high school at grade 10 and preparatory school levels on grade 12 to measure student graduate competence nationally in certain subjects as well as to map the level of achievement for students at the school level and the regional level (Education and training policy, 2002).

Many researchers has been discussed the different factors that affects the student academic performance in their research. According to (Hussain, 2006) there are two types of factors, internal and external classroom factors, strongly affect the students' academic performance. Internal classroom factors includes students competence in English, class schedules, class size, class test results, learning facilities, homework, environment of the class, complexity of the course material, teachers role in the class, technology used in the class and exams systems. External classroom factors include extracurricular activities, family problems, work and financial, social and other problems. His Research studies shows that students' performance depends on many factors such as learning facilities, gender and age differences, etc. that can affect student academic result.

As (Galiher, 2006) research indicated to measure the student academic performance through several ways like CGPA, GPA and their test result. Most of the researcher around the word used the GPA to measure the student performance. Some other researcher, they measure student performance through the result of particular subject or the previous year result (Hijazi & Naqvi, 2006). (Noble, 2006), students' academic accomplishments and activities, perceptions of their managing strategies and positive attributions, and background characteristics (i.e., family income, parents' level of education, guidance from parents and number of negative situations in the home) were indirectly related to their composite scores, through academic achievement in high school. The students face a lot of problems in developing positive study attitudes and study habits. As (Endris, 2015) guidance is one of the factor through which a student can improve his study attitudes and study habits and is directly related to academic achievement. The students who are properly guided by their parents have performed well in the exams. The guidance from the teacher also affects the student performance. The guidance from the parents and the teachers indirectly affect the performance of the students

2.2. Gender and Academic performance

There has been a repeated argument on the issue of gender differences on math and science achievement. Different research has also shown a decline in the differences between the genders in the past few decades on standardized test, suggesting that the more exposure that women are getting to math and science classes, the better their scores (Halpern, 2000).Other research puts into questions whether gender differences still exist in academic achievement, many researchers are still finding differences in performance as well as general interest in areas related to math and science. As (National Educational Examination and Assessment Agency, 2008) educational assessment result found that Boys achieved mean scores that were higher by 4.3 % in the composite score and 3.3% to 4.8% in the five subjects (English, mathematics, physics, chemistry and biology) than girls. The differences in all subjects were statistically significant at $p < .001$. This shows that there is gender gap in academic achievement.

Thus, achievement alone cannot be the only reason for women as they make their career choices. According to (Miriam, 2008) to investigate the interest students' math's subject with their academic performance based on gender .the result shows that for young men in higher-level math tracks, math interest is much more strongly related to math school grades than for young women in the same math courses.

As Fergusson (2007) reviewed the Minister of Education's report on 1995 the New Zealand school system showed that in School Certificate examinations in mathematics and science subjects males tended to do slightly better in mathematics than females whereas females did slightly better than males in science subjects. He examined the report on Christchurch born 1,265 children who were studied from the point of school entry to the age of 18 years. The researcher used methods including parental interviews, teacher assessments, standardized testing and interviews with the children as data collection and he found that if boys and girls behaved in similar ways in the classroom setting there would have been no evidence of consistent gender related differences in school achievement.

As Arzu(2010) conducted study to identify the impact of gender on student academic performance from admission to graduation and to determine whether gender differences affected student success on the examination .The researcher took students' entering GPA, exit GPA, and first-attempt performance on the Certified Respiratory Therapy (CRT) examination and on the Written Registry for Respiratory Therapy (WRRT) examination as variables that measure academic success. the researcher used the existing records of all graduates at Georgia State University analyzed the result with SPSS and finally the study showed that there is no significant gender difference on student academic performance in RT education.

Ebenuwa's (2010) study revealed that gender, age and finance are not significant predictors of academic performance and there was no significant difference in academic performance based on age, gender and financial status rather the student's character and behaviors are basic and foremost determinants of academic performance. So the researchers recommend that counseling centers should open to handle varying problems.

Meltem and Serap (December 2004) tried to investigate in whether there are significant gender differences in academic performance among undergraduate students in a large public university in Turkey based on three indicators; university entrance scores, performance in the English preparatory school and in the program the student is majoring in. The data for this study was collected from the undergraduate student records compiled by the registrar's on 10,343 individual. The result of the research showed that a smaller number of female students manage to enter the university and when they do so, they enter with lower scores. However, once they are admitted to the university, they excel in their studies and outperform their male counterparts.

Muluken (2015) was conduct study to investigate factors that predict student performance of Debremarkos University students Models were built and tested by using a sample dataset of 11,873 regular undergraduate students. He used decision tree and Bayes as a classification technique and the generated rules were studied and evaluated. Data collected from MS_EXCEL files, and it has been preprocessed for model building. He obtained from his research result that Higher Education Entrance Certificate Examination result of a student, and sex were the main determining attributes to classify university students' academic performance as failure/success. Likewise Endris (2013) conducted study to Analyses determinants Influencing Students' Academic Achievement in English and Mathematics subjects focused on ending classes of schools (grade 8, 10 and 12) at Bahir Dar and Debre Tabor Towns at both government and non government schools in Ethiopia. The result showed that, on average, the academic achievements of female students were better than male students in mathematics at grade eight, ten and twelve.

2.3. RELATIONSHIP BETWEEN SUBJECTS

General Education Quality Assurance And Examinations Agency, (MOE, 2010) conducted assessment to determine the various levels of students' performances at both Grades 4 and 8 in four key academic subject areas. Grade 4 pupils were tested in English, basic reading, mathematics and environmental science subjects, all prepared in the different instructional languages; and Grade 8 students were assessed in English, mathematics, chemistry, and biology subjects. They used three stage stratified random sampling design to select sample regions, schools and students at both grade levels (4 and 8). Accordingly, 256 schools for Grade4, and 136 schools for Grade 8 were sampled. About 10,506 students for Grade 4, and 5099 for Grade 8 studies were tested across ten regions of the nation. The results also indicated that schools with high achievement at Grade 4 level also tended to obtain high achievement at Grade 8. Beside this there is a positive relationship among the five subjects and the correlations were statistically significant in all cases at $p < .01$. In both grades, the study showed that student background factors, teacher variables, school structure and curricular materials, language of instruction, play a significant role in the variability of student achievement scores.

Pandey's study (March 2011) analyses the impact of language on students in class room. The main idea was to find out the support and confidence level for appropriate language and attendance in the classroom. For this purpose he used association rule to analyzed the correlation between three instructional medium English, Hindi and the mixture of two as a result he found that the Mix medium class have more impact on the student result score over Hindi and English medium class.

Adeboye (1993) carried out a co-relational study showed proficiency in English Language as a factor contributing to competency in Mathematics. He used simple random sampling technique from 1991 West Africa School Certificate published results. This was made up of fifty candidates from each sample. The result showed that there is positive significant correlation between the grades in English Language and Mathematics at school certificate level. Lack of proficiency in English Language is a factor contribution to the poor performance in mathematics also observed that students who passes the words problems are likely to pass non-word problems in mathematics.

Similarly Adegboye (1997) carried out correlation study of the student' performance in English language and Science subject through analysis of past result of candidate in science subject college of education. The sample consist of three hundred candidates of 1991 and 160 from 1992 WASC published result who offered natural science subjects. It is made up of 140 from 1991 and 160 from 1992. The coefficients of correlation between performances in the pair of five subjects under his study were computed and the coefficient of multiple correlations between performance in English Language and all the four science subjects was also obtained. His study results showed that English Language plays greater role in the study of Chemistry than it plays in the study of any other science subject; also that Mathematics plays a greater role in the study of physics. It was also observed that a multiple correlation between English on one hand and chemistry and biology on the other hand is greater than the multiple correlations between English on one hand and Mathematics and physics on the other hand.

He concluded that there is positive linear correlation between grades in English Language and the Science subjects of Biology, Chemistry, Physics and Mathematics.

2.4. The impact of plasma television on student's academic achievement

Plasma technology applied to the process of education in order to help in resolving problems that educators are challenged with. The new technologies applied to education should be considered within the socio-economic context of each Fikadu (Jnuary, 2013). Plasma technology will further change the way knowledge is developed, acquired and delivered. It is also important to note that the new technologies offer opportunities to innovate on course content and teaching methods and to widen access to learning. The Ministry of Education introduced a new educational policy, the new curriculum followed by the introduction of new technologies like plasma television at government high schools shafika (June, 2014). Plasma television program support the teaching-learning process by presenting sufficient teaching aids, organized lessons, finished content within the allotted period of time, supported the teachers teaching activity, transmitting uniform education to many students to have access to model and competent teachers, providing standardized education to all high schools, presenting abstract concepts in a simplified manner, overcoming the problem of qualified teacher, and helping to offer quality and equitable education for all children(FDRE, 2004).

The most advantages of instructional satellite television include convenience, cost reduction, economizing the use of teaching potential, and increasing the number of students to whom a teacher can have access by recording lectures Fikadu (January, 2013). It is reasoned out by UNESCO (2005) that mass media have the capacity to meet the challenge of the growth in the number of persons to be educated as they are able to transmit the same message in millions of copies. Instructional satellite television has advantages over other media in transporting the entire audience to wild, hostile or inaccessible locations, enlarging the microscopic and reducing the immense, and combining text and images, encouraging learning and understanding through different channels (Blythe-lord, 1991).

According to the argument of some researchers, satellite television instruction is not significantly different from face-to-face (teacher-based) learning. Citing a series of researches done on comparing television instruction to conventional teaching managed with the same instructors and teaching methods, Abatihun (2012) states that there was no significant difference between students' level of achievement. Heney and Ullmer (1998) have also noted that there was no significant difference between the two media of instruction. In fact, in the educational systems that have adequate teachers, teaching materials for demonstration and experiments, well-equipped libraries, and other educational inputs, television instruction may not have significant difference to the face-to-face teaching.

Ethiopia introduced educational broadcast through a satellite receiving device known as plasma display panel (PDP) which is commonly called plasma television [PTV] to all secondary schools throughout the country since 2004 whereby 10 subjects viz: Biology, Chemistry, Mathematics, Physics and Civic and Ethical Education are being broadcast and implemented for the last nine years (MOE, 2004). The assumptions held by the MOE in integrating PTV based educational broadcast into the classrooms in all the secondary schools in order to presents abstract concepts in a simplified manner, transmits uniform education to many students found in different places at the same time, enables students to have access to model and competent teachers, demonstrates laboratory equipment found in one place (classroom) to other learning classrooms and brings the "real world" into a classroom.

As Gebeyehu (2014) research indicated that Ethiopian Ministry of Education has made efforts in utilizing educational media such as satellite TV program for improving quality of secondary education for last decades of years. However, there was lack of national investigation in the aspect of systemic evaluation, for measuring the effectiveness of satellite TV program. The aim of his research was to investigate the actual practice and effectiveness of satellite TV program in Ethiopian secondary schools. He used questionnaire to collect data and sampling method at two secondary schools and selected 228 students (Grade 9-12).. The results indicated that Ethiopian students utilizing satellite TV program scored highly in the evaluation areas of context, input, process, and product of the program. The research revealed that learning demand, learning content, and class management were factors affecting the satisfaction in satellite TV program. They also found that satellite TV program could play an important role in improving the quality of in Ethiopian secondary education.

The successful integration of media technology, like plasma, in education has the great potential to boost up the quality, quantity, relevance and equity. When properly utilized, media technology exceedingly helps students' learning and hence, improves students' academic performances. Corroborating these points, Gillani (2005) claimed that the principal role of media technology is to help improve the overall efficiency of the teaching/learning process. She further elaborated that in education and training, improved efficiency can manifest itself in many ways, for example increasing the quality of learning, or the degree of mastery, decreasing the time taken for learners to attain desired goals and increasing the efficiency of teachers in terms of numbers of learners taught without reducing the quality of learning.

According to Gara (November 2012,) research result indicated that methods of teaching (traditional and PTV instructions) PTV didn't significantly contribute to the better understanding of the lesson compared to the conventional method of teaching. However, the researcher noted that proper implementation or utilization of the instructional media technologies such as PTV could significantly help students' learning and improve their academic performances.

2.5. Location and achievement

As (National Educational Examination and Assessment Agency, 2008) educational assessment result Students from rural schools performed better than those from urban ones in all subjects except in English. This implies In English students from urban schools performed better than those in rural ones.

Regions as factor and the composite score as dependent variable were carried out in order to detect the existence of significant differences. Tigray scored highest with a mean difference of 11.3 % from Gambella which achieved the least score. The mean scores of Oromia (37.1%) and Amhara (36.9 %) were slightly higher than the national mean (35.6%). All other regions achieved mean scores equal to or less than the national mean.

Students from non-metropolitan areas are more likely to have lower educational outcomes in terms of academic performance and retention rates than students from metropolitan areas Zappala (2002). Despite an adequate number of educational facilities in rural and remote Australia, school children from these areas remain disadvantaged by other factors include costs, the availability of transport and levels of family income support. In addition, inequity exists with regard to the quality of the education that rural students receive, often as a result of restricted and limited subject choice. Furthermore, students may also have limited recreational and educational facilities within their school.

As (Musibau, 2013) study revealed that school location had no significant influence on students' academic performance. This finding implies that whether a student attends rural or urban secondary school it does not make a difference in student's academic performance.

2.6. Data Mining

Technology now allows us to capture and store vast quantities of data. The amount of data in the world, in our lives, seems to be increasing and there's no end in sight (Witten and Frank, 2005). Rapid advances in data collection and storage technology have enabled organizations to accumulate vast amount of data. We are

overwhelmed by data - educational, medical data, demographic data, financial data and marketing data (Han and Kamber, 2006). It has been estimated that the amount of data stored in the world's databases doubles every 20 months (Witten and Frank, 2005). To undertake large data analysis project, researchers and practitioners have adopted established algorithms from statistics, machine learning, neural networks, and databases and have also developed new methods targeted at large data mining problems (Hand, Heikki and Smyth, 2001).
Witten and Frank (2005) states lack of data is no longer a problem at the current stage. However the inability to generate useful information from data is the problem. As the volume of data increases unstoppably, the proportion of people understands decreases, alarmingly. Lying in hidden data, in all these data potentially useful information, i.e. rarely made explicit or taken advantage of Data can be form simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. To take complete advantage of these data; data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is 'Data Mining' (Deshpande and Thakare, 2010).

According to Berry and Linoff (2004), data mining usually makes sense when there is large amount of data. For this reason most of the algorithms developed for data mining purpose requires large volume of data so as to build and train models that are responsible for different tasks of data mining such as classification, clustering, prediction, association and the like. Moreover, the need for bulky data can be explained by a couple of reasons. Primarily in the case of small databases, it is feasible to capture appealing trends and relationships by introducing traditional tools such as spreadsheets and database query. The second reason is that most data mining tools and algorithms demand large amount of training data (data used for building a model) in order to generate unbiased models. The rationale is simple and straight forward, small training data results in unreliable generalizations based on chance patterns.

2.7. Application of data mining in education

According to Praveen (Feb. 2013), educational data mining is a study that make use of statistical, machine-learning, and data-mining algorithms over the different types of educational data in order to analyze these types of data to resolve educational research issues. Even if there are many approach, the researcher applied Bayesian Network classification methods up on student database to predict the students' academic performance. The result of the study will help the students to improve their performance and also it helps teacher to identify those students which needs a special attention to reduce failing ratio and taking appropriate action at right time. Based on the Experimental Results he concludes that AODEsr Algorithm predict more accuracy than any other Algorithms.

Based on Maharashtra (March 2014) Educational data mining (EDM) is a field that exploits machine-learning, statistical and data-mining algorithms over the different types of educational data. Its main objective is to analyses data in order to resolve educational research issues. EDM is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn. This data helps to understand learners and learning and to develop computational approaches that combine data and knowledge to transform practice to benefit learners.

As Muluken(2015) analysis of student data using by data mining tools helpful to the academic planners in universities of learning to enhance their decision making process. The research found the results offer helpful and constructive recommendations to the academic planners in universities of learning to enhance their decision making process, aid in the curriculum structure and modification in order to improve students' academic performance. Students able to decide about their field of study before they are enrolled in specific field of study based on the previous experience taken from the research-findings. 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 affecting the student performances. So, on the bases of the research findings the level of student success will increase and it is possible to prevent educational institutions from serious financial strains.

2.7.1. Classification and prediction

Classification, one of the most common data mining methods, seems to be a human imperative. Human beings usually classify, categorize or grading in order to understand and communicate about the world (Berry and Linoff, 2004). Many of the data mining applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes. It is a supervise learning because the classes are predefined before the examination of the target data (Han and Kamber, 2006; Two crows corporations, 2005). Pang-Ning et al (2006) also stated that classification is the task of learning a target function f that maps each attribute X to one of the predefined class label Y.

According to Two Crows Corporations (2005), classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. Han and Kamber (2001) describing the classification task stated that data classification is a two-step process. In the first step, a model is constructed by analyzing database tuples described by the attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. In the second step the model is used for classification. The holdout method is a simple technique that uses a test set of class-labeled samples. These samples are randomly selected and are independent of the training samples. Then the accuracy of a model on a given test set is evaluated.

Any of the techniques used for classification can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior (Berry and Linoff, 2004).

For this research, an association task is to be carried out since a model is to be built by extracting the most important and meaningful association rules; 2005EC and 2007EC grade 10 and 12 national examination student result data respectively in Ethiopia. Although the selection of techniques suitable for association is mainly depends on the type of data used for mining and the expected outcome of the mining process. The domain experts play a significant role in the selection of algorithm for data mining (Deshpande and Thakare, 2010)

2.7.2. Association Rule in education

Data mining is the process of discovering interesting knowledge from large amount of data stored in database, data warehouse or other information repositories. It includes various tasks such as classification, clustering, association rule etc.

Association rules are used to show the relationship between data items. Mining association rules allows finding rules of the form: If antecedent then (likely) consequent where antecedent and consequent are item sets which are sets of one or more items. Association rule generation consists of two separate steps: First, minimum support is applied to find all frequent item sets in a database. Second, these frequent item sets and the minimum confidence constraint are used to form rules. Support & confidence are the normal method used to measure the quality of association rule. Support for the association rule X->Y is the percentage of transaction in the database that contains XUY. Confidence for the association rule is X->Y is the ratio of the number of transaction that contains XUY to the number of transaction that contain X (kamper, 2006). Association rule can be used in educational data mining and teacher's evaluation system for analysing the learning data. According to (SuchitaBorkar, 2013) student's performance is evaluated using association rule mining algorithm. The research has been done on assessing student's performance based on various attributes graduation mark, test mark, attendance and university result .In this study important rules were generated to measure the correlation among selected attributes which would help to improve the student's academic performance. Experiment is conducted using Weka and NEAEA dataset available in the organization database. The researcher analyse student's examination result in different subject at grade 10 and 12, gender, age, students 'residence etc.

2.8. RELATED WORKS

Data mining in education is a recent research field, there are many works are already done in this area. That is because of its potential to educational institutes. Various study that used educational data mining to analyze students learning behavior Muluke(2015) gave a case study that used educational data mining to identify behavior of failing students to warn students at risk before final exam. Some Used educational data A Comparative Study for Predicting Student's Academic Performance www.iosrjen.org.

Mining to identify and then enhance educational process in higher educational system which can improve their decision making process. (Al-Najjar, 2006) applied the classification as data mining technique to evaluate students' performance. The goal of their study is to predict the final grade of the students. The outcome of their results indicated that Decision tree model had better prediction than other models. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising.

Anuradha (2015) Applied the classification as data mining technique to evaluate student "performance, they used decision tree method for classification. Anuradha

(2015) applied the association rule mining analysis based on students' failed courses to identify students' failure patterns. The goal of their study is to identify hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low performing students. Durairaj(2014) used Bayesian Classification Method as a data mining technique and concluded that students grade in senior secondary exam, living location, medium of teaching, mother's qualification, students other habits, family annual income and students family status were highly correlated with the student academic performance. Beikzadeh,(2005) used simple linear regression analysis and it was found that the factors like mother's education and student's family income were highly correlated with the student academic performance. Inayat(2005) conducted a study on the student performance using association rule technique and they find the interestingness of student in opting class teaching language.

Different researchers researched on various variables and a lot of different variables were studied. All those previous were conducted by using a very small proportion of the datasets and the analysis was made by using traditional methods. There for limited capacity to discover new and unanticipated patterns and relationships. In this study the researcher increase the dataset and used data mining tools in order to prove the result of the previous researchers.

This research is different in a way that answers how much the selected variables or factors have an impact on students' academic performance in term of their national exam result obtained on grade 10 and 12. In addition to this the research explore the student preference on study area based on grade 10 national exam result determine student's grade 12 national exam result. So the research assist education sector planners, policy makers, and decision makers as a decision support aid in planning and implementing educational intervention programs aimed at facilitating students , improving the teaching procedures in order to improving quality of education in the regions as well as in the country.

CHAPTER THREE

Methodology

3.1. Study area

The research is carried out based on primary data extracted from the database of Ethiopian national assessment agency which is available for researchers. Secondary data for instance review important document to gain farther information related with student achievement. Accordingly, in this study both quantitative and qualitative research design; interview will be used to understand the domain knowledge and to interpret the finding.

3.2. Study design

In order to achieve the above stated objectives, the researcher has used the CRoss-Industry Standard Process for Data Mining (CRISP-DM) model, which contains six phases.

3.2.1. CRISP–DM

This section deals with the overviews of the data source, data cleaning and data transformation of the data employed in this study. In general the researcher has followed the steps of data mining process mentioned in the chapter one. In this study the methodology adapted is CRISP-DM.

CRISP-DM (CRoss-Industry Standard Process for Data Mining), a nonproprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit. Developed by industry leaders with input from more than 200 data mining users and data mining tool and service providers, CRISP-DM is an industry tool and application neutral model. This model encourages best practices and offers organizations the structure needed to realize better, faster results from data mining. The CRISP–DM demands that data mining be seen as an entire process, from communication of the business problem through data collection and management, data preprocessing, model building, model evaluation, and finally, model deployment (Chapman et al, 2000). Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it (Chapman et al, 2000).

According to Chapman et al (2000), a given data mining project has a life cycle consisting of six phases. Note that the phase sequence is adaptive. That is, the next phase in the sequence often depends on the outcomes associated with the preceding phase. Depending on the behaviour and characteristics of the model, we may have to return to the data preparation phase for further refinement before moving forward to the model evaluation phase (Larose, 2005). CRISP-DM is complete and well documented. All the stages are properly organized, structured and defined, allowing that a project could be easily understood or revised (Santos &Azevedo, 2005).

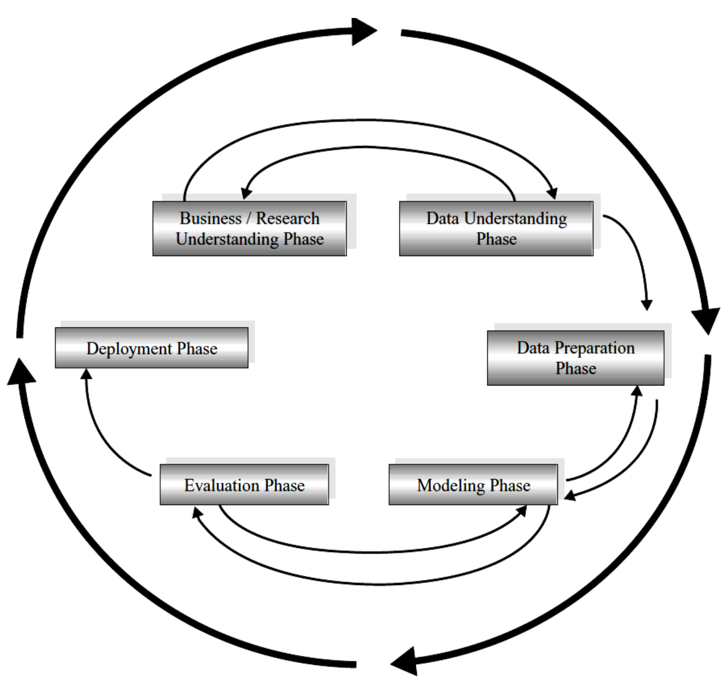

Figure 3.1 Phases of the CRISP-DM reference model Source Laros (2005)

3.3. Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives. In this study in order to understand the business and the application, we work closely with domain experts in NEAEA and review related literatures on the research topic. The source of data for this research taken from 2005 grade10 and 2007 grade 12 Ethiopian students National Examination result database. It provides critical information for the monitoring and evaluation of the country's Plan for Accelerated and Sustained Development to End Poverty (PASDEP). The education sector general education quality improvement policies and programs assist in the monitoring of the progress towards meeting the Millennium Development Goals (MDGs).

NEAEA conducted three consecutive learning assessments in 2000, 2003/2004 and 2008 respectively to measure learning achievements of Grade 8 students and identify the factors that determine those achievements. It also aimed at providing comparative information on school improvement. But the target population was Grade 8 students in each study in the country.

3.4. Data understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets. The primary data source for this study mainly are 2005 and 2007 NEAEA data sets, Center for Information and Communication Technology and Educational Management Information system Directorate under Ministry of Education data set in excel format, which contains details information on student name, address of student, student academic result of grade 10 and 12, school type, age, gender, teachers' qualification, availability of media technology (plazma) at school etc. The data are dispersed on different offices in unorganized form. Hence to get familiar with the data, to identify data quality problems, to have insights into the data and to detect interesting subsets of the data to be used different literatures have been surveyed. Besides, careful analysis of the data and its structure has been done together with the domain experts.

3.5. Data preparation

Data preparation is the most important phases of the data analysis activity which involves the construction of the final data set (data that will be fed into the analysis tool) from the initial raw data. Data preparation generates a dataset smaller than the original one, which can significantly improve the efficiency of data mining. This task includes: attribute selection, filling missed values, correcting errors, or removing outliers (unusual or exceptional values), resolve data conflicts using domain knowledge or expert decision to settle inconsistency. Since there is no missing value occurred on the data set, no missing value replacement technique applied on this research. Collected excel format prepared in some Weka understandable formats. Then preprocessing activities are performed and the file is saved into Weka acceptable comma separated values (CSV) or comma delimited file format. Weka native data format is known as the ARFF (Attribute Relation File Format). It is basically a CSV (comma separated value) format with some extra headers to specify what type each attribute is (numerical, binary, nominal). The CSV file format is converted into ARFF by using Weka mining software, to take advantage of easier data manipulation and also compatible interaction with Weka software. During scan of the preprocessed data some basic statistics summary was produced for each attributes into a form acceptable by the selected data mining software Weka. 3.6 .4

3.6. Modeling

In this phase various modeling techniques are selected and applied, and their parameters are adjusted to optimal values. Although the choice of data mining techniques for association tasks seems to be strongly dependent on the application, one of the data mining techniques that are frequently employed for analysis tasks is association rule. As it is indicated previously, the purpose of this research is to develop association rules model. Association rule data mining technique is applied in predicting correlation between grade 10 and 12 students' academic result.

Once the data preprocessing task completed, the researcher Apriority algorithm, which is the default algorithm selected. However, in order to change the parameters for this run (e.g., support, confidence, etc.) WEKA allows the resulting rules to be sorted according to different metrics such as confidence, leverage, and lift. In this research, the researcher has selected lift as the criteria entered certain value as the minimum value for lift (or improvement) is computed as the confidence of the rule divided by the support of the right-hand-side (RHS). In a simplified form, given a rule L => R, lift is the ratio of the probability that L and R occur together to the multiple of the two individual probabilities for L and R, i.e.

$$lift = Pr(L,R) \ / \ Pr(L).Pr(R).$$

If this value is 1, then L and R are independent. The higher this value, the more likely the existence of L and R together in a transaction is not just a random occurrence, but because of some relationship between them. Here the researcher also change the default value of rules (10) to be 20; this indicates that the program will report no more than the top 20 rules (in this case sorted according to their lift values). The upper bound for minimum support is set to 1.0 (100%) and the lower bound to 0.1 (10%). Apriori in WEKA starts with the upper bound support and incrementally decreases support (by delta increment which by default is set to 0.05 or 5%). The algorithm halts when either the specified numbers of rules are generated, or the lower bound for min. support is reached. The significance testing option is only applicable in the case of confidence and is by default not used (-1.0). The rules were discovered based on the specified threshold values for support and lift. For each rule, the frequency counts for the LHS and RHS of each rule is given, as well as the values for confidence, lift, leverage, and conviction.

3.7. Evaluation

At this stage, the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives. Evaluation is the key to making real progress in data mining. After building a model, we must evaluate its results and interpret their significance (Two Crows Corporation, 2005). This stage consists of the interpretation and evaluation of the mined patterns. This indicates interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing irrelevant patterns, and translating the useful ones into terms understandable by users. The evaluation process is also carried out to identify interesting patterns representing knowledge based on some lift as an interestingness measures.

Both objective and subjective measures have been applied during the association rule analysis. The subjective method requires additional expert knowledge or input, which is not fully available during this study; the researcher used lift as objective measure. Therefore Improvement and further analysis can be made in this area.

3.8. Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The sequence of the phases is not rigid. Moving back and forth between different phases is usually required and possible (Chapman et al, 2000). As it has been mentioned the scope of this study is to build a model for students' academic performance. The successful model can support educational planner and decision maker for analyzing education system and decision making system. As data mining consider powerful tool to analyze the data and it has the capability to merge with any system. Therefore the proposed data mining model can be integrating with the existing DSS system. The practical implementation of the model can be helpful for the decision makers to modify and adapt according to the organization.

DM TOOL, TECHNIQUES AND ALGORITHMS

3.9. Data Mining Tool Selection

Data mining or knowledge discovery refers to the process of finding interesting information in large repositories of data. The term data mining also refers to the step in the knowledge discovery process in which special algorithms are employed in hopes of identifying interesting patterns in the data. These interesting patterns are then analyzed yielding knowledge (Bowen, 2006).

Discovering knowledge in data presents data mining as a well-structured standard process, intimately connected with managers, decision makers, and those involved in deploying the results (Larose, 2005). Therefore both novices and data-mining specialists need assistance in knowledge discovery processes (Deshpande and Thakare, 2010). Many good data mining software products are being used, ranging from well-established, Enterprise Miner by SAS and Intelligent Miner by IBM, CLEMENTINE by SPSS, PolyAnalyst by Megaputer, and many others in a growing and dynamic industry (Witten and Frank, 2005). WEKA (from the University of Waikato in New Zealand) is an open source tool with many useful machine learning methods (David and Delen, 2008).

Data mining tools need to be versatile, scalable, capable of accurately predicting responses between actions and results, and capable of automatic implementation (Chackrabarti, et al, 2009). Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to the business strategies, knowledge bases, scientific and medical research (Han and Kamber, 2006).

Data mining tools predict future trends and behaviors and help organizations to make practical knowledge-driven decisions (Larose, 2005). The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were more time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations (Deshpande and Thakare, 2010). In this research, Weka 3.6.4 software is used as a mining tool. In addition Microsoft Excel for data cleaning and for converting the original file to CSV file format, and Microsoft Word for documentation purpose have been used.

There are factors that contribute to the usefulness of data mining tools or software to the intended data mining tasks: The tool selected should be able to provide the required data mining functions. The data mining functionality that the researcher will intended to carry out in this research is describe the relation between the factors variables and students academic result. In addition the methodologies used by the data mining software to perform each of the data mining functions are also important factor to consider. The researcher chooses association rules in order to analyze the correlation between the subjects, student's performances related with other subgroups and the relationship of grade 10 and 12 academic performance of the student.

In addition, the selected tool can comfortably operate on windows operating system and stand alone environment. Hence Windows 7 operating system on a standalone machine has been utilized. Another important consideration in tool selection is visualization capabilities. The variety, quality and flexibility of visualization tools may strongly influence the usability, interpretability, and attractiveness of data mining systems. Weka has a facility to visualize its output in this regard (Witten and Frank, 2005). The other reason behind the selection of Weka for this study is familiarity of the researcher with the tool, its comprehensiveness for this study requirements and the ease of availability of the tool; Weka provides a number of data mining functionalities, Such as classification, clustering, association, attribute selection and visualization. Weka is developed at the University of Waikato, in New Zealand. 'Weka' stands for the Waikato Environment, Knowledge Analysis. The system is written in JAVA an object oriented programming language, and has been tested under Linux, windows and Macintosh operating systems. Weka includes varieties of tools, for preprocessing a data set, such as attribute selection, attribute filtering and attribute transformation (Witten and Frank, 2005).

As it is mentioned in the methodology section in chapter one, the objective of this research is to find association among variables. Hence, it is important to extract the association rule implementations for model building and experiments to be carried out in the data mining process, which also involve data mining tool selection and algorithms used for modeling. To demonstrate real practicality in any data mining process, selecting the potential mining tool is important to understand clearly the techniques and algorithms to be implemented, and describe them specifically based on the tool used for the research work.

So, in this study the researcher chooses to use the Weka 3.6.4 software. The reason why this tool is specially selected is that it is the only toolkit that has gained widespread adoption and survived for an extended period of time and it is freely available for download (i.e. it is an open source software) and as well it offers many powerful features (sometimes not found in commercial data mining software), it has become one of the most widely used data mining systems.

3.10. Data Mining or Knowledge Discovery Process

Data mining or knowledge discovery refers to the process of finding interesting information in large repositories of data. The term data mining also refers to the step in the knowledge discovery process in which special algorithms are employed in hopes of identifying interesting patterns in the data. These interesting patterns are then analyzed yielding knowledge (Bowen, 2006). Discovering knowledge in data presents data mining as a well-structured standard process, intimately connected with managers, decision makers, and those involved in deploying the results (Larose, 2005). Therefore both novices and data-mining specialists need assistance in knowledge discovery processes (Deshpande and Thakare, 2010).

Data mining tasks

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories, descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions (Han and Kamber, 2006).

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined (Deshpande and Thakare, 2010). Descriptive models belong to the realm of unsupervised learning. Such models interrogate the database to identify patterns and relationships in the data. Clustering (segmentation) algorithms, pattern recognition models, visualization methods, among others, belong to this family of descriptive models (Han and Kamber, 2006).

In predictive modeling tasks, one identifies patterns found in the data to predict future values. Predictive modeling consists of several types of models such as classification, regression and Artificial Intelligence based models. Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. Whereas descriptive techniques are sometimes referred to as unsupervised learning because there is no already-known result to guide the algorithms (Two Crows Corporations, 2005). According to Fayyad, et al (1996) the goals of prediction and description can be achieved using a variety of particular data-mining methods. The data mining methods are broadly categorized as: On-Line Analytical Processing (OLAP), Classification, Clustering, and Association Rule mining, etc. These methods use different types of algorithms and data. The data source can be data warehouse, database, flat file or text file. The algorithms may be Statistical Algorithms, Decision Tree based, Nearest Neighbor, Neural Network based, Genetic Algorithms based, Ruled based, Support Vector Machine etc. (Deshpande and Thakare, 2010). The Data mining methods that are used for this study and the behavior of the patterns it discovers is described below.

3.11. The Weka tool

Once the Weka software download and install to the machine where the mining process carried out and optional files have been set up, everything is ready to use Weka. To employ Weka for data mining tasks, there is a graphical user interface (shown below), which consists of buttons and menu commands. This main window has four panels on its interface, where each panel is used to perform different tasks.

Figure 3.2 Weka GUI Application Main Window

The first step after we run Weka is then to select the application (Explorer, Experimenter, etc) to process for the data mining tasks using the application panel on the window available. After the explorer window is chosen and the application data file is opened, different techniques can be selected to perform data mining tasks as shown in the following figure. The Weka software contains different techniques for data pre-process, classify, cluster, associate, etc.

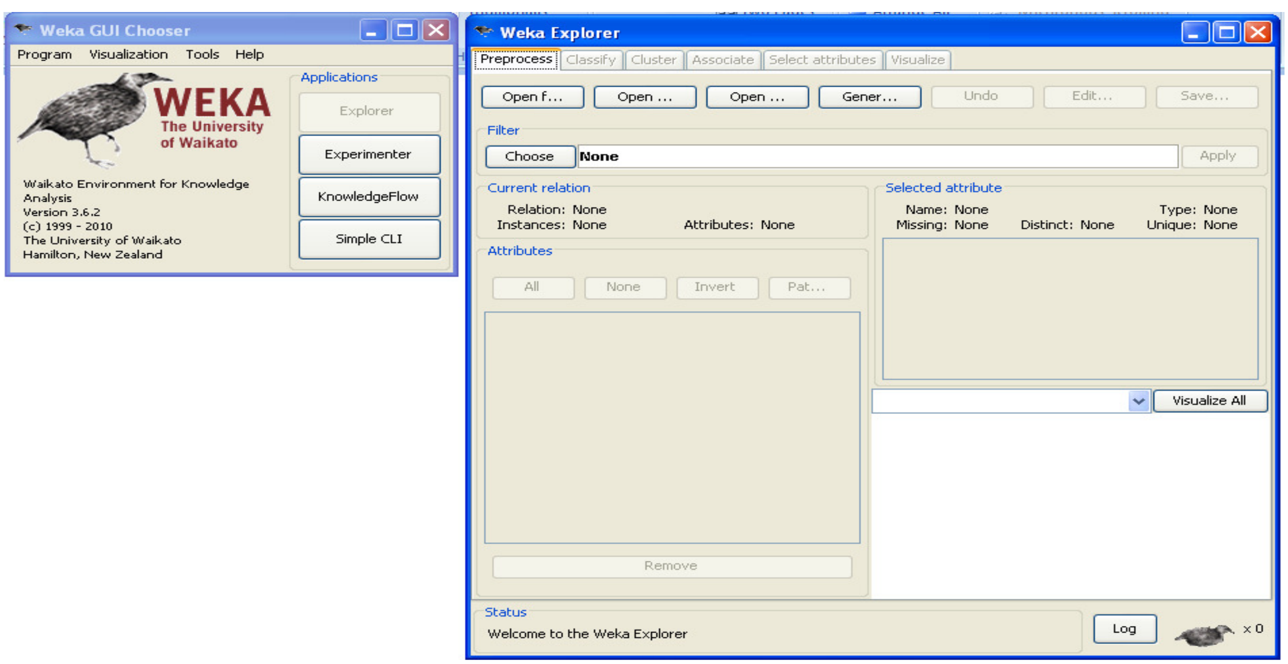

Figure 3.3 Weka Explorer Windows

In the pre-process tab using Weka explorer window, we can load data into the system by clicking on "Open file..." or other tabs in the window where the data is available and well-suited for tool in the ARFF format or CVS format. Converting data formats through Weka can be done by clicking on the "Save..." at the explorer window. Editing data in Weka can also be done by clicking on the "Edit..." tab in the window. We then examine the data thoroughly its attribute type, properties and class (last attribute) distribution. Finally, we can visualize the data graphical what looks like before we continue to use other techniques in the explorer window for further analysis presented. That is, the histogram shows the distribution of each attribute with coloring by class grouping.

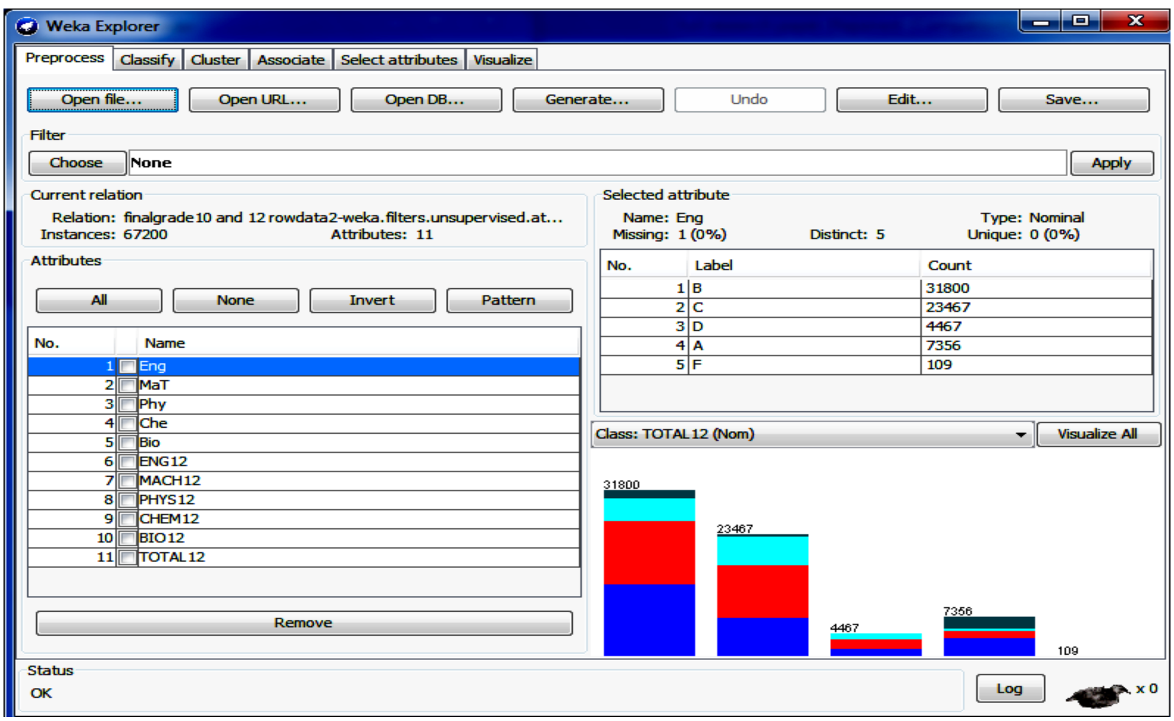

Figure 3.4 visualize the data graphical *weka explorer windows*

In addition to pre-processing, a number of data mining methods/tools are implemented in the Weka software. Some of them are based on Apriori algorithm. Before we get run through Weka, we should get into the specific details of each method and we should understand what model strives to accomplish what type of data and what goals model attempts to accomplish. If association rule panel of

Weka's Explorer window is used, it provides several options that affect the type of associate that Weka produces and the way that it is constructed. The construct associate rule button on the toolbar displays a dialog box, which consists of construction options displayed on the following figure.

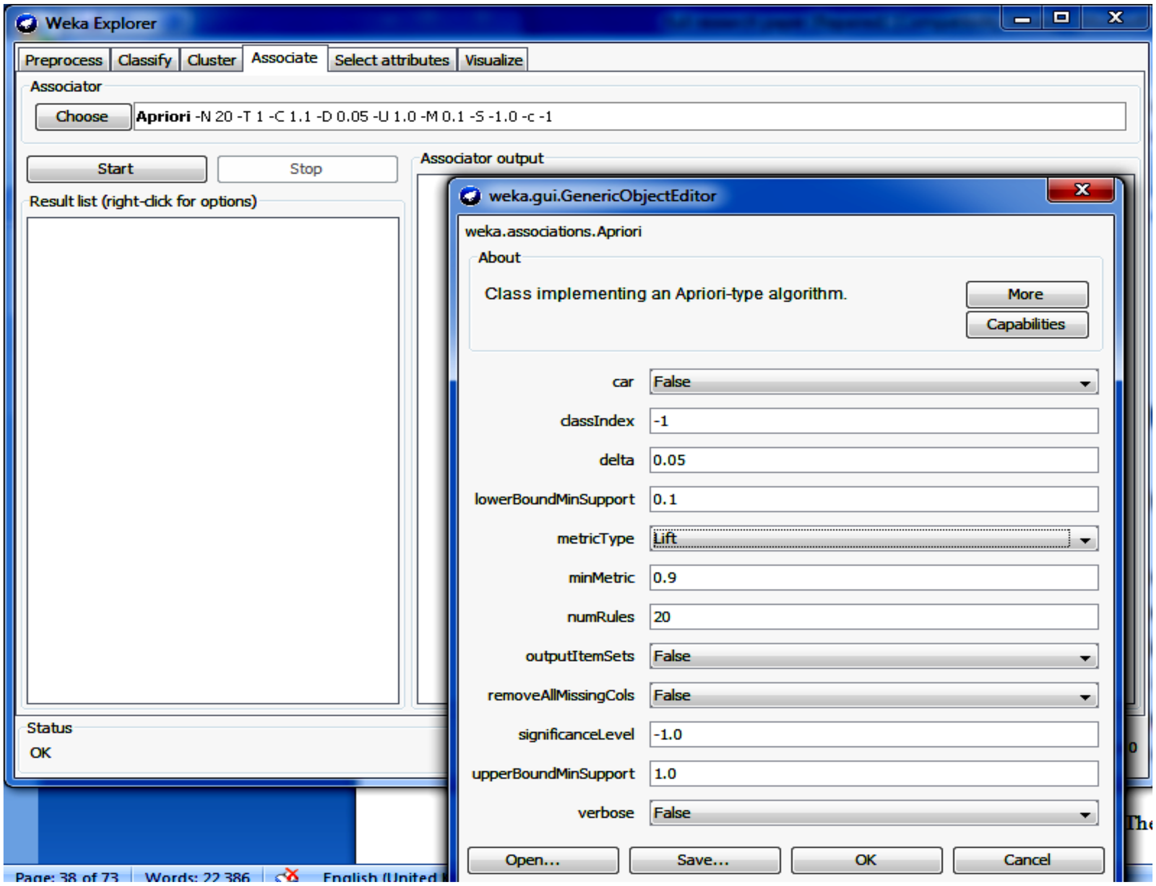

Figure 3.4 Weka Explorer with association rule Evaluation Options dialog box

3.12. Association rule in weka

There are a number of machine intelligent techniques that are available in the market but at the same time not all tools are the best for all problems in the data set. Different data sets produce different results based on the algorithms used. In this study we are testing one algorithm by adjusting different parameters value based on Apriori method association rule. Our aim is to find the best model. The association rule mining can be considered as a two-step process (Kamber ,2006)

1) Find all frequent itemsets: Each of these itemsets will occur at least as frequently as a predetermined minimum support count.

2) Generate strong association rules from the frequent itemsets: The rules must satisfy minimum support and confidence. These rules are called strong rules.

Association rule learning is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness; detect relationships or associations between specific values of categorical variables in large data sets. Which means in association we are interested in finding out association between items with no particular focus on a target one, whereas in classification we basically map the set of record (attributes, variables) into the class attribute. For this, the researcher used the NEAEA dataset which is provided as a common dataset from 2005 and 2007. 67,200 records or cases from the total of 210,000 records are used for this association analysis. Association rule mining has been applied to find correlations between items in a dataset, including identifying attributes characterizing patterns of performance disparity between various groups of students, discovering interesting relationships from a student's academic result, finding out the relationships between each pattern of subjects. Association rules 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 (Kumar, 2012).

IF-THEN rules are one of the most popular ways of knowledge representation, due to their simplicity and comprehensibility (Han and Kapler , 2005). There are different types of rules according to the data mining technique used, for example: classification, association, sequential pattern analysis, prediction, causality induction, optimization, etc. In the area of knowledge discovery in databases the most studied ones are association rules, classifiers and predictors. An example of a generic rule format IF-THEN in Extend Backus Naur Form (EBNF) notation it is shown in the table below.

IF-THEN rule format	
<rule>	::= IF <antecedent> THEN <consequent>
<antecedent>	::= <condition> +
<consequent>	::= <condition> +
<condition>	::= <attribute> <operator> <value>
<attribute>	::= Each one of the possible attributes set
<value>	:= Each one of the possible values of the attributes set
<operator>	::= = | > | < | ≥ | ≤ | ≠

Table 3.1 IF-THEN rule Extend Backus Naur Form

Before beginning the study in depth of the main association rule mining algorithms, I first formally define what an association rule (Han and Kapler , 2005) is. Let I = {i1, i2 , ..., im} be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that T C I. Each transaction is associated with a unique identifier, called its TID. We say that a transaction T contains X, a set of some items in I, if X C I. An association rule is an implication of the form X -> Y, where X C I, Y C I, and X C Y=Ø.

The rule X->Y holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y . The rule X -> Y has support s in the transaction set D if s% of transactions in D contains X U Y. An example association rule is: "IF a diaper is in a transaction, THEN beer is in the transaction as well in 30% of the cases. Diapers and beer are bought together in 11 % of the rows in the database". In this example, support and confidence of the rule are:

s = P(X U Y) =11%

c = P(Y/X) = P(X UY) / P(X) = s (X U Y) / s (X) = 30%

Given a set of transactions D, and user-specified thresholds minimum support (called minsup) and minimum confidence (called minconf), the problem of mining association rules is to generate all association rules that have support and confidence greater than minsup and minconf respectively. My discussion is neutral with respect to the representation of D. For example, D could be a data file, a relational table, or the result of a relational expression.

3.13. **Aprior algorithm**

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules. Some key concepts for Apriori algorithm are :

- Frequent Itemsets: The sets of item which has minimum support (denoted by Li for ith-Itemset).
- Apriori Property: Any subset of frequent itemset must be frequent.
- Join Operation: To find Lk , a set of candidate kitemsets is generated by joining Lk-1 with itself.

The Apriori algorithm is the first and simplest Association rule mining algorithm. According to the result obtained using datamining tool author find that Apriori Association algorithm performs better than the Predictive Apriori Association Rule and Tertius Association Rule Algorithms (Kanwal, June 2013). Apriori algorithm the first pass of the algorithm simply counts item occurrences to determine the frequent item sets of size (cardinality) a frequent item set of size k is in literature also called a large k-item set. All subsequent passes, consists of two phases. First, the large itemsets Lk found in the (k)th pass are used to generate the candidate itemsets Ck+1 and Next, the database is scanned and the support of candidates in Ck+1 is counted.

The Apriori algorithm performs a breadth first search in the search space by generating candidate k+1- itemsets from frequent k itemsets

3.13.1. Pseudo-code of APRIOR Algorithm

Ck: Candidate itemsets of size k;

Lk : Frequent itemsets of size k

L1 = {frequent items};

for (k = 1; Lk != Ø; k++) do begin

    Ck+1 = {candidates generated from Lk};

    for each transaction t in the database do

     increment the count of all candidates in Ck+1

    that are contained in t;

    Lk+1 = {candidates in Ck+1 with minsup}

end

return U k Lk;

Let us consider a set L3={abc, abd, acd, ace, bcd}

The join step:

- Ck+1 is generated by joining Lk with itself

Self-joining: L3*L3 (Items must be in lexicographic order.)

- abcd from abc and abd
- acde from acd and ace

The prune step:

- Any k-itemset that is not frequent cannot be a subset of a frequent (k+1)-itemset
- Pruning C4: acde is removed because ade is not in L3

The final candidate set: C4={abcd}

Finding the appropriate parameter settings of the mining algorithm Association rule mining algorithms need to be configured before they are executed. So, the user has to give appropriate values for the parameters in advance (often leading to too many or too few rules) in order to obtain a good number of rules. Most of this algorithm requires the user to set two thresholds, the minimal support and the minimal confidence, and then find all the rules that exceed the thresholds specified by the user. Therefore, the user must possess a certain amount of expertise in order to find the right settings for support and confidence to obtain the best rules. One possible solution to this problem can be to use a parameter-free algorithm. For example, the Weka (Machine Learning with WEKA, 2004) package implements an Apriori-type algorithm that solves this problem partially. This algorithm reduces iteratively the minimum support, by a factor delta support (Δs) introduced by the user, until a minimum support is reached or a required number of rules has been generated.

Chapter four

DATA PREPARATION AND PRE-PROCESSING

4.1. Overview of Data Pre-processing

It is well known that success of every data mining algorithm is strongly dependent on a quality. Data pre-processing task could be critical and a very complicated task. Sometimes, the data pre-processing takes more than half of the total time spent on solving data problems, because incomplete, noisy and inconsistent data are commonplace properties of large real-world databases and data warehouses (Laros, 2005). Thus, data preprocessing is an important and critical step in the data mining process, and it has a huge impact on the success of a data mining project. The purpose of data preprocessing is to clean the noisy data, extract and merge the data from different sources, and then transform and convert the data into a proper format (Laros, 2005). It is an important step in data mining, because quality decisions must be based on quality data.

4.2. Description of the original data

The raw data usually has a great deal of noise, incompleteness and inconsistency. Raw data cannot be used directly for processing, with the machine-learning algorithms. They first need to be preprocessed into machine understandable format. The database of National Educational Examination and Assessment Agency, 2005 and 2007 is considered as below to demonstrate Preprocessing. The data types of the attributes with the raw data are given below.

Attribute		Description	Type	Width
Reg.No		registration number of the student	double	8
NAME		full name of the student who seat on the exam	Char	20
Sex		gender of the student	char	1
Age		the age of the student	int	2
Nationality		nationality of the student	char	15
Disability		the visual impair of the student	Boolean	1
School code		the representation of the school which contain region, exam center station, of the school located	text	20
School name		the name of the school	text	20
Instructional media Technology		currently the school receive and use the satellite television program in the class room	Boolean	1
Qualification		the academic qualification of teachers in each school	char	2
Subject	Amh	Amharic subject result	int/char	2
	Eng	English subject result	int/char	2
	Phy	physics subject result	int/char	2
	Chem	chemistry subject result	int/char	2
	Bio	biology subject result	int/char	2
	MatN	mathematics subject for natural science result	int/char	2
	MatS	mathematics subject for social sciences result	int/char	2
	Civ	civic and ethical education subject result	int/char	2
	Geo	Geography subject result	int/char	2
	His	History subject result	int/char	2
	Gee	Geez subject result	int/char	2
	Fre	French language subject result	int/char	2
	Tig	Tigrigna language subject result	int/char	2
	Oro	Affan oromo language	int/char	2
	Har	Harreri language	int/char	2
	Agn	Agew language	int/char	2
	Nue	Nuer language	int/char	2

	Wol	Wolayita language	int/char	2
	Kaf	Kaffa language	int/char	2
	Had	Haddyia language	int/char	2
	Sid	Siddama language	int/char	2
	Sil	Sillti language	int/char	2
	Som	Somali language	int/char	2
	Ged	Gediyo language	int/char	2
	Daw	Dawrro language	int/char	2
	Gam	Gammu language	int/char	2
	Gof	Goffa language	int/char	2
total		the sum of student mark at grade 12 in each subject	int	2
barcode		an identification number given to the student identification card by the machine	float	8

Table 4.1 set of attributes the original data file

This data set consists of 785,180 records of students who were sat on grade 10 national exam and 230,000 records of students who were sat on grade 12 national exam in 2005 E.C and 2007 E.C respectively. Data selection method is the first step which includes raw data extraction and selecting the target dataset that are used for the study problem. In a case where we have millions of data points, we have to select a subset of the database to perform the required knowledge discovery steps. Therefore, selection is the process of choosing on the right data from the database on which the tools in data mining can be used to extract useful information, knowledge and pattern from the provided raw data (Kambler ,2005).

4.3. Attribute Selection

Deciding on the data that will be used for the analysis is based on several criteria, including its relevance to the data mining goals, as well as quality and technical constraints such as limits on data volume or data types (Shearer, 2000). Selecting relevant features (attributes) in any data mining task is important for increasing the efficiency of the algorithm. Liu and Motoda (1998) mentioned that “the abundance of potential features constitutes a serious obstacle to the efficiency of most learning algorithms". Therefore, eliminating some attributes, which are assumed to be irrelevant to build the model can increase the accuracy of the rule, save the computational time, and simplify results obtained. Therefore, in this thesis the attributes are selected with the help of domain expert and extensive literature review. Because taking all the variables in the data base we have, feed them to the data mining tool and find those which are the best predictors may not work very well. One reason is that the time it takes to build a model increases with the number of variables. Another reason is that blindly including extraneous columns can lead to incorrect models (Two Crows Corporations, 2005). Thus, it is necessary to leave out those attributes that are not important for analysis with the help of domain experts in order to simplify the task of modeling.

The NEAEA data set contains many attributes and to decide on the relevant attributes for this study, the researcher has discussed with domain expert in the area. As described in table, the following attributes are selected: student ID, student address, Region, school type, age, gender, teachers' academic qualification, school's media technology. The final selected attributes were prepared and preprocessed as stated in the following section, before starting rule generating process. Based on expert's opinion, only some of them are considered as relevant for the specific task to be undertaken in this research work. In this regard, the original data set consisted of a total of 210,000 records from grade 10 and 12 student name and school name as relationship keys, from which 20 attributes for the target data set for this research work has been selected. In this research, the main focus is association between different factors (attributes) based on the research objectives. The researcher therefore used those attributes related to student academic performance. After this step, the researcher able to filter out those irrelevant or less relevant attributes like any subject score except the five selected subjects excluded. The data mining tasks thus become simpler and more accurate. Hence the researcher decides to apply dimensionality reduction on the data set created for analysis by selecting the minimum set of features (attributes) that are associated. A total of 17 candidate attributes are selected as relevant features for analysis and test the required result. Selection of these relevant features is conducted in consultation with domain experts and personal experience about the data set of NEAEA database. In the database the attributes that are not contributed any information towards the scope of this research have been removed from all the cases within the database. The following table gives a brief description of the candidate attributes selected as relevant features to build and test the required models.

<table>
<tr><th colspan="2">attribute</th><th>description</th><th>type</th><th>width</th></tr>
<tr><td colspan="2">sex</td><td>gender of the student</td><td>char</td><td>1</td></tr>
<tr><td colspan="2">age</td><td>the age of the student</td><td>int</td><td>2</td></tr>
<tr><td colspan="2">school code</td><td>the representation of the school which contain region, exam center station, of the school located</td><td>text</td><td>20</td></tr>
<tr><td colspan="2">instructional media technology</td><td>currently the school receive and use the satellite television program in the class room</td><td>Boolean</td><td>1</td></tr>
<tr><td rowspan="10">subjects</td><td>Eng 10</td><td>English subject result of grade 10 student</td><td>char</td><td>2</td></tr>
<tr><td>Phy 10</td><td>physics subject result of grade 10 student</td><td>char</td><td>2</td></tr>
<tr><td>Chem 10</td><td>chemistry subject result of grade 10 student</td><td>char</td><td>2</td></tr>
<tr><td>Bio 10</td><td>biology subject result of grade 10 student</td><td>char</td><td>2</td></tr>
<tr><td>Maths 10</td><td>mathematics subject for of grade 10 student</td><td>char</td><td>2</td></tr>
<tr><td>Eng 12</td><td>English subject result of grade 12 student</td><td>int</td><td>2</td></tr>
<tr><td>Phy 12</td><td>physics subject result of grade 12 student</td><td>int</td><td>2</td></tr>
<tr><td>Chem 12</td><td>chemistry subject result of grade 12 student</td><td>int</td><td>2</td></tr>
<tr><td>Bio 12</td><td>biology subject result of grade 12 student</td><td>int</td><td>2</td></tr>
<tr><td>MatN 12</td><td>mathematics subject for natural science result of grade 12 student</td><td>int</td><td>2</td></tr>
<tr><td colspan="2">total</td><td>the sum of student mark at grade 12 student in each subject</td><td>int</td><td>2</td></tr>
</table>

Table 4.2 Attributes selected from the original data file

4.3. Creating new table

Starting from the grade 10 and 12 the researcher have created a new summarization tables that aggregate the information at the required level and integrates the most important information for the research objective. In order to create this table it is necessary to do several technique to the database in excel format to compare student name column values from both grade 10 and 12 tables and merge the corresponding field value in order to obtain the information of the desired students (student subject result values grade 10 table) and (student subject result values grade12table).

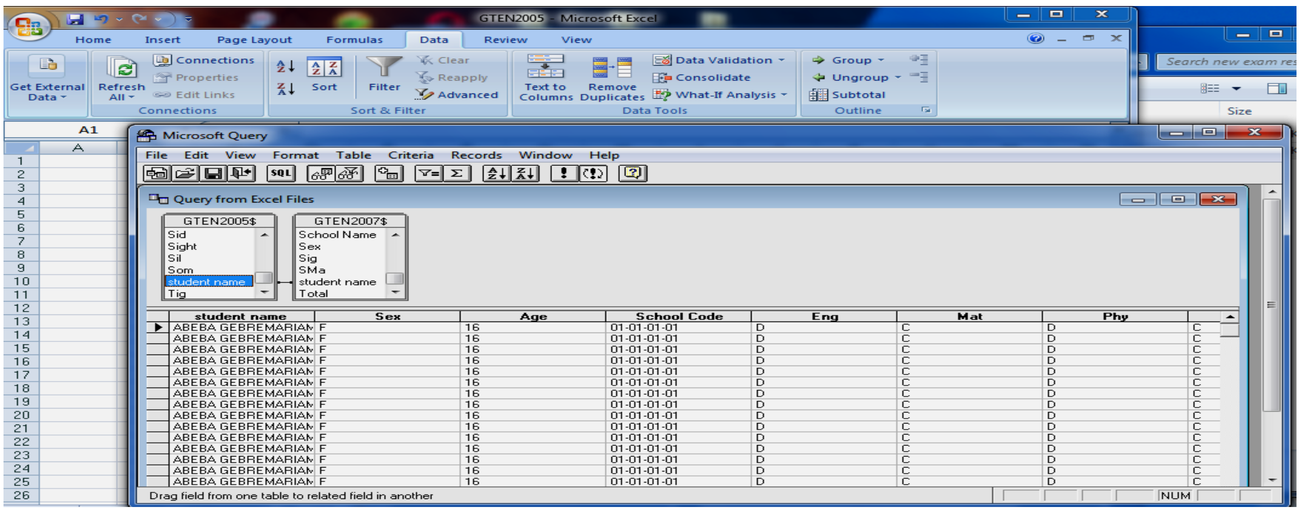

Fig 4.1 new table created by relationship between two row datas of grade 10 and 12 student result

4.4. Discretizing data

Some techniques, such as association rule mining, can only be performed on categorical data. This requires performing discretization on numeric or continuous attributes. The researcher performs a discretization of numerical values in order to increase the interpretation and comprehensibility. Discretization divides the data in categorical classes that are easier to understand. In this experiment, all the numerical values are discretized in a new table (except for student identification code). There are several unsupervised global methods for transforming continuous attributes into discrete attributes, such as the equal-width method, equal-frequency method or the manual method (in which you have to specify the cut-off points). In this case, The WEKA discretization filter, can divide the ranges blindly, or used various statistical techniques to automatically determine the best way of partitioning the data. In this case, simple binning was used. Method for the subject mark attribute of grade 12 and the equal-width method for the others attribute. The labels that we have used in the mark attribute are Under_basic, Basic, proficient and advance .

4.5.Data decoding and attribute transformation

It is often necessary to transform values of attribute to another for making the attributes useful in the prediction modeling process. Particularly, if numbers represent unique concepts and not values within a continuous range of some quantity or rating, they should be converted into symbols or separate columns (Kamber,2006). In the original data set created for this data mining task, the values of some attributes are represented using numerical codes. Such numerical codes that used to represent unique concepts had to be converted into its representation or columns to avoid any confusion during training and testing of the model as follows: The attribute "school code" which stands for "the student region, exam center station, name and school type" is represented as alphanumerical codes using eight different values. So, for the purpose of this research work, those numerical codes were converted into their respective symbolic values. The following table shows the registration code used in the original data file and the reformatted name of the school code.

School code	**Value represented**			
	Region	Zone /sub city	School type	School number
140090201	14(addis ababa)	zone 09	02(non-government)	01
01010113	01(tigray region)	zone 01	01(government)	13

Table 4.3 Actual Values of each registration code and its value represent

The evaluations of the students in their respective subjects differ from grade10 to grade12. On grade 12 have marking system while on grade 10 have decide on grading system. In the dataset of this research, which consist of students who have completed from these two grades have their results in both forms, so the researcher have streamlined the data by converting all the results in Grade 12 from numeric to nominal form. The students' achievement is classified into four standards as: under_basic, Basic, proficient and advance. The classification is based on norm referenced ability scores using Item Response Theory (IRT) (NEAEA, 2008). It used scaled scores instead of raw scores where: Proficient is greater than or equal to 1 standard deviation above the mean, Basic and proficient are within 1 standard deviation above the mean and Below Basic is below the mean score. In addition the following data transformation and reformatting operations are employed in order to create new variables like school location (region and zone/woreda), school type, from the existing ones and to reformat the original values of some attributes data set selected for analysis.

Finally, the researcher has transformed the data to the required format by using the data mining algorithm. In this case, exported the new created table to a text file with ARFF format used by Weka. An Attribute-Relation File Format (ARFF) file is an ASCII text file that describes a list of instances sharing a common attributes (Laros, 2005).

Chapter five

Experimentation and result analysis

5.1. Overview of data preparation

In this research very important statistical data is mined to extract patterns related to student academic performance issues with the help of Weka tool. The data mining process, the flow of the system starts with the collection of raw data. This data is first preprocessed and converted into formats understood by the Weka tool that are used in the DM process as we discussed in chapter three of this paper.

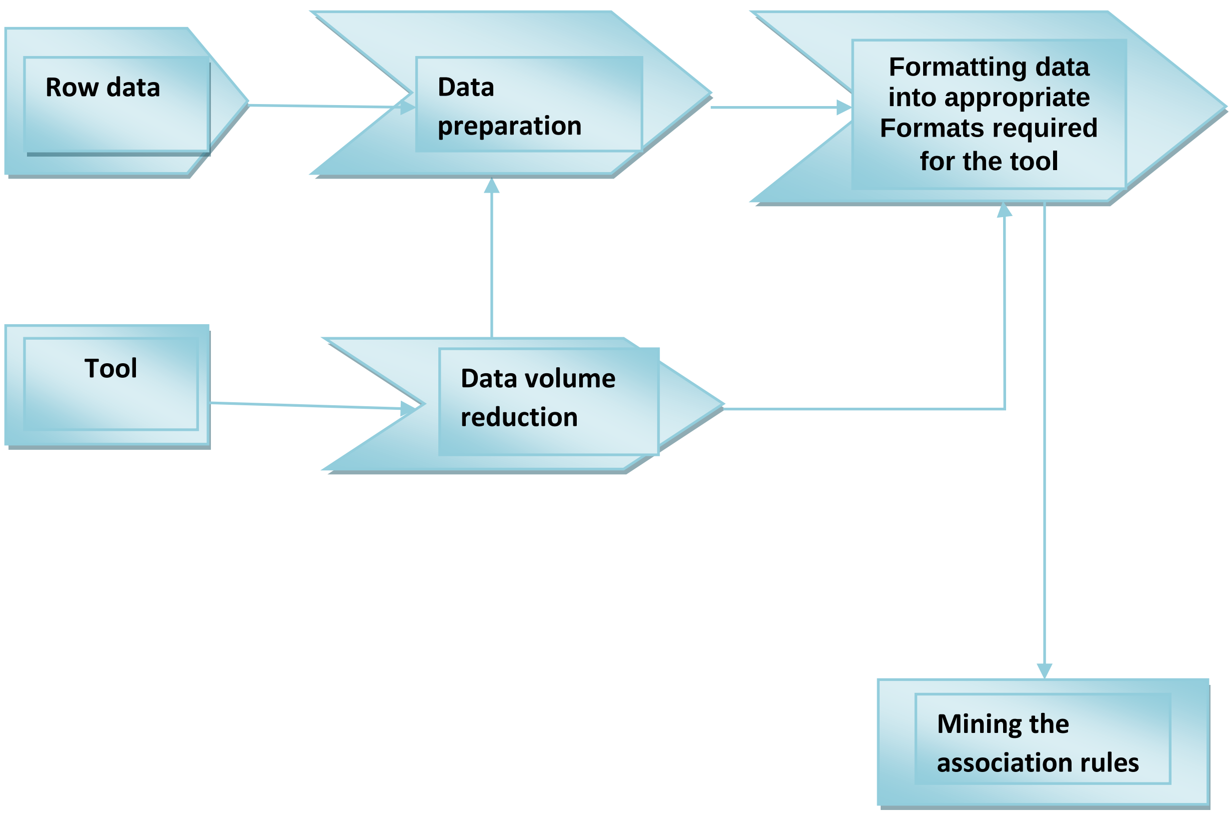

Figure 5.1 show an overview of data preparation

5.2. Experimentations

Association rule mining suits data sets that have no single category that needs to be predicted. Rather, this technique suits best very large datasets from which unexpected associations between any fields of the data are looked for, perform association rule mining on it. Association rules are increasingly used in educational data mining. However, measuring the interestingness of a rule can be problematic, as explained in Borkar(2013). Two measures, support and confidence, are commonly used to extract association rules. However it is well known that even rules with a strong support and confidence may in fact be uninteresting. This is why, once the association rule X→Y has been extracted, it is wise to double check how much X and Y are related. Unfortunately, no measure is better than all the others in all situations, though measures tend to agree when support is high. Association rules are useful in Educational Data Mining for analyzing learning data. This technique requires not only that adequate thresholds be chosen for the two standard parameters of support and confidence, but also that appropriate measures of interestingness be considered to retain meaningful rules and filter uninteresting ones out. In this study, different experiments are conducted altering parameters of the Aprior algorithm to generate the best association rule model. An association rule is rated uninteresting by the lift if its value is around or less than 1.

Lift is designed to measure the reliability of the rule. The difference between confidence and lift lies in their formulation and the corresponding limitations. Confidence is sensitive to the probability of consequent (Y). Higher frequency of Y will ensure a higher confidence value even if there is no true relationship between X and Y. But if we increase the threshold of the confidence value to avoid this situation, some important pattern with relatively lower frequency may be lost. In contrast to confidence, lift is not vulnerable to the rare items problem. It is focused on the ratio between the joint probability of two item sets with respect to their expected probabilities if they are independent.

Even item sets with lower frequency together can have high lift values. As Merceron (2008) states, it is sufficient to focus on a combination of support, confidence, and either lift or leverage to quantitatively measure the "quality" of the rule. However, the real value of a rule, in terms of usefulness and action ability is subjective and depends heavily of the particular domain and business objectives. The researcher has selected lift as the criteria. Furthermore, the researcher entered 0.9 as the minimum value for lift (or improvement) is computed as the confidence of the rule divided by the support of the right-hand-side (RHS).

Before selecting the values of parameters the researcher conducted different experiment, which was important to produce optimal and accurate association rules in order to analyze the correlation among different subjects of students' academic result, by altering the values of minimum metric value. The best rule extracted where the value of minimum metric value is 0.9.

Here also change the default value of rules (10) to be 20; this indicates that the program will report no more than the top 20 rules (in this case sorted according to their lift values). The upper bound for minimum support is set to 1.0 (100%) and the lower bound to 0.1 (10%). Apriori in WEKA starts with the upper bound support and incrementally decreases support (by delta increments which by default are set to 0.05 or 5%). The algorithm halts when either the specified numbers of rules are generated, or the lower bound for minimum support is reached. The significance testing option is only applicable in the case of confidence and is by default not used (-1.0). Once the parameters have been set, We click on start button to run the program in weka.

5.1. Subjects correlation

5.1.1. Correlation between five subjects in grade 10

Association rule extracted between subjects	Confidence value	Lift value	Leverage value	Conviction value
English = B => math = B	0.5	1.06	0.01	1.04
English = C => math = C	0.46	1.2	0.03	1.14
English = B=> physics = B	0.48	1.0	0	1
English = C=> physics = C	0.45	1.5	0.01	1.04
English = B=> chemistry = B	0.37	1.04	0.01	1.02

English = C=> chemistry = C	0.37	1.27	0.01	1.13
English = B => biology = A	0.55	1.15	0.02	1.16
English = B => biology = B	0.51	1.07	0.01	1.07
English = C => biology = C	0.42	1.35	0.04	1.19
biology = A => English =B	0.55	1.35	0.04	1.16
Math = B => physics = B	0.39	1.02	0	1.01
Math = C => physics = C	0.46	1.09	0.01	1.07
Math = A => chemistry = A	0.69	2.5	0.07	2.34
Math = B => chemistry = B	0.4	1.15	0.02	1.09
Math =C => chemistry = B	0.38	1.08	0.01	1.05
Math = C => chemistry = C	0.39	1.34	0.04	1.16
Math =A => biology = A	0.66	2.2	0.06	2.07
Math =B => biology = A	0.31	1.03	0	1.01
Math =B => biology = B	0.36	1.08	0.01	1.04
Math =C => biology = C	0.39	1.26	0.03	1.13
Physics = B => chemistry = A	0.38	1.37	0.04	1.17
Physics = C=> chemistry =B	0.38	1.08	0.01	1.04
Physics = C=> chemistry =C	0.31	1.07	0.01	1.03
Physics = B => biology = A	0.39	1.31	0.04	1.15
Physics = C => biology = B	0.36	1.08	1.01	1.04
Physics = C => biology = C	0.33	1.05	0.01	1.02
Chemistry = A => biology = A	0.68	2.26	0.1	2.18
Chemistry = B => biology = B	0.46	1.31	0.04	1.21
Chemistry = C => biology =C	0.48	1.53	0.05	1.32

Table 5.1. Association rule extracted on grade 12students result on science subject

The above table shows that the probability of students who score “A” in mathematics have got “A” in Biology is higher. The same wise a student who get “B” in mathematics cannot get “C” in chemistry. When the result shows that a student who scores “A” in math does not to score “B” or “C” in biology. Students who score “A” in physics do not have any association with chemistry result but the physics result score “B” have some association with chemistry result. In other words a student who scores “A” in chemistry cannot get a result less than “C” in physics. A student who scores “A” in physics do not have any association with biology subject result. However, students’ result in physics at least “B” do not have a probability to get less than “A” in biology subject. As we can see from the above table mathematics student results have a strong relationship with both chemistry and biology subject student result. In addition to these students results on chemistry subject and biology subject have strong correlation than other subjects.

```
Size of set of large itemsets L(1): 13

Size of set of large itemsets L(2): 15

Best rules found:

 1. Che=A 18588 ==> Bio=A 12630    conf:(0.68) < lift:(2.26)> lev:(0.1) [7052] conv:(2.18)
 2. Bio=A 20163 ==> Che=A 12630    conf:(0.63) < lift:(2.26)> lev:(0.1) [7052] conv:(1.94)
 3. Che=B 23717 ==> Bio=B 10409    conf:(0.44) < lift:(1.31)> lev:(0.04) [2476] conv:(1.19)
 4. Bio=B 22475 ==> Che=B 10409    conf:(0.46) < lift:(1.31)> lev:(0.04) [2476] conv:(1.21)
 5. Eng=C 23462 ==> Mat=C 10716    conf:(0.46) < lift:(1.2)> lev:(0.03) [1759] conv:(1.14)
 6. Mat=C 25651 ==> Eng=C 10716    conf:(0.42) < lift:(1.2)> lev:(0.03) [1759] conv:(1.12)
 7. Eng=B 31800 ==> Bio=A 10991    conf:(0.35) < lift:(1.15)> lev:(0.02) [1448] conv:(1.07)
 8. Bio=A 20163 ==> Eng=B 10991    conf:(0.55) < lift:(1.15)> lev:(0.02) [1448] conv:(1.16)
 9. Mat=C 25651 ==> Phy=C 11922    conf:(0.46) < lift:(1.09)> lev:(0.01) [952] conv:(1.07)
10. Phy=C 28735 ==> Mat=C 11922    conf:(0.41) < lift:(1.09)> lev:(0.01) [952] conv:(1.06)
11. Phy=C 28735 ==> Bio=B 10386    conf:(0.36) < lift:(1.08)> lev:(0.01) [774] conv:(1.04)
12. Bio=B 22475 ==> Phy=C 10386    conf:(0.46) < lift:(1.08)> lev:(0.01) [774] conv:(1.06)
13. Phy=C 28735 ==> Che=B 10907    conf:(0.38) < lift:(1.08)> lev:(0.01) [764] conv:(1.04)
14. Che=B 23717 ==> Phy=C 10907    conf:(0.46) < lift:(1.08)> lev:(0.01) [764] conv:(1.06)
15. Eng=B 31800 ==> Bio=B 11402    conf:(0.36) < lift:(1.07)> lev:(0.01) [765] conv:(1.04)
16. Bio=B 22475 ==> Eng=B 11402    conf:(0.51) < lift:(1.07)> lev:(0.01) [765] conv:(1.07)
17. Eng=B 31800 ==> Mat=B 12245    conf:(0.39) < lift:(1.06)> lev:(0.01) [707] conv:(1.04)
18. Mat=B 24378 ==> Eng=B 12245    conf:(0.5) < lift:(1.06)> lev:(0.01) [707] conv:(1.06)
19. Eng=C 23462 ==> Phy=C 10497    conf:(0.45) < lift:(1.05)> lev:(0.01) [463] conv:(1.04)
20. Phy=C 28735 ==> Eng=C 10497    conf:(0.37) < lift:(1.05)> lev:(0.01) [463] conv:(1.03)
```

Fig 5.2. Association rule extracted on grade 10 students’ science result by prior algorithm

5.1.2. **Correlation between five subjects in grade 12**

Association rule extracted between subjects	Confidence value	Lift value	Leverage value	Conviction value
English_score =”under_basic” => Nmath_score = “under_basic”	0.78	1.05	0.03	1.1
Nmath_score =”under_basic” => English_score = “under_basic”	0.71	1.05	0.03	1.12
English_score =”under_basic” => physics_score = “under_basic”	0.86	1.03	0.02	1.17
physics_score =”under_basic” => English_score = “under_basic”	0.69	1.03	0.01	1.06
English_score =”underbasic”=> chemistry_score = “under_basic”	0.26	1.09	0.01	1.03
chemistry_score =”underbasic”=> English _score = “under_basic”	0.73	1.09	0.01	1.02
English_score =”underbasic”=> chemistry_score = “_basic”	0.61	1.01	0	1.02
chemistry_score =”basic”=> English _score = “under_basic”	0.68	1.09	0	1.02
English_score = “basic” => biology_score = “advance”	0.77	1.26	0.03	1.18
biology_score = “advance” => English_score = “basic”	0.39	1.26	0.03	1.13
English_score = “under_basic => biology_score= “under_basic”	0.76	1.13	0.01	1.37
Math_score = “under_basic” => physics_score = “under_basic”	0.91	1.08	0.05	1.17
physics_score = “under_basic” => Math_score = “under_basic”	0.8	1.08	0.05	1.31
Math_score = “under_basic => chemistry_score = “under_basic”	0.3	1.25	0.04	1.08
chemistry_score = “under_basic =>	0.93	1.25	0.04	3.57

Math_score = “under_basic”				
Math_score=”basic”=> biology_score = “advance”	0.62	1.67	0.06	1.66
biology_score=”addvance” => Math_score = “basic”	0.74	1.67	0.06	1.27
Math_score=”under_basic”=> biology_score = “under_basic”	0.17	1.23	0.02	1.04
biology_score=”under_basic” => Math_score = “under_basic”	0.91	1.67	0.06	2.95
Physics_score = “under_basic => chemistry_score = “under_basic”	0.27	1.13	0.03	1.04
chemistry_score = “under_basic => Physics_score = “under_basic”	0.94	1.08	0.01	2.9
Physics_score = “under_basic”=> biology_score =”under_basic”	0.15	1.11	0.01	1.02
biology_score = “under_basic”=> Physics_score =”under_basic”	0.15	1.11	0.01	1.02
Chemistry_score = “advance” => biology_score = “advance”	0.73	1.96	0.06	2.31
biology_score = “advance” => Chemistry_score = “advance”	0.61	1.96	0.06	1.22

Table 5.2. Association rule extracted on grade 12 students’ result on science subject

From the table we can understand that chemistry and biology have better correlation than other subjects. On one hand English and mathematics student result have a better correlation with biology subject student result than either physics or chemistry subject result. The results showed that English Language and mathematics plays greater role in the study of biology than it plays in the study of any other science subject; it also that Mathematics plays a greater role in the study of biology. It was also observed that a multiple correlation between English on one hand and chemistry and biology on the other hand is greater than the multiple correlations between English on one hand and Mathematics and physics on the other hand. From this research finding the correlation between Physics and Mathematics is very weak both in grade 10 and 12, in contrary to other researcher finding Mathematics plays a greater role in the study of physics (Adegboye, 1997). One of the causes for this weak connection might be cheating during exam but it needs further investigation to reach the conclusion.

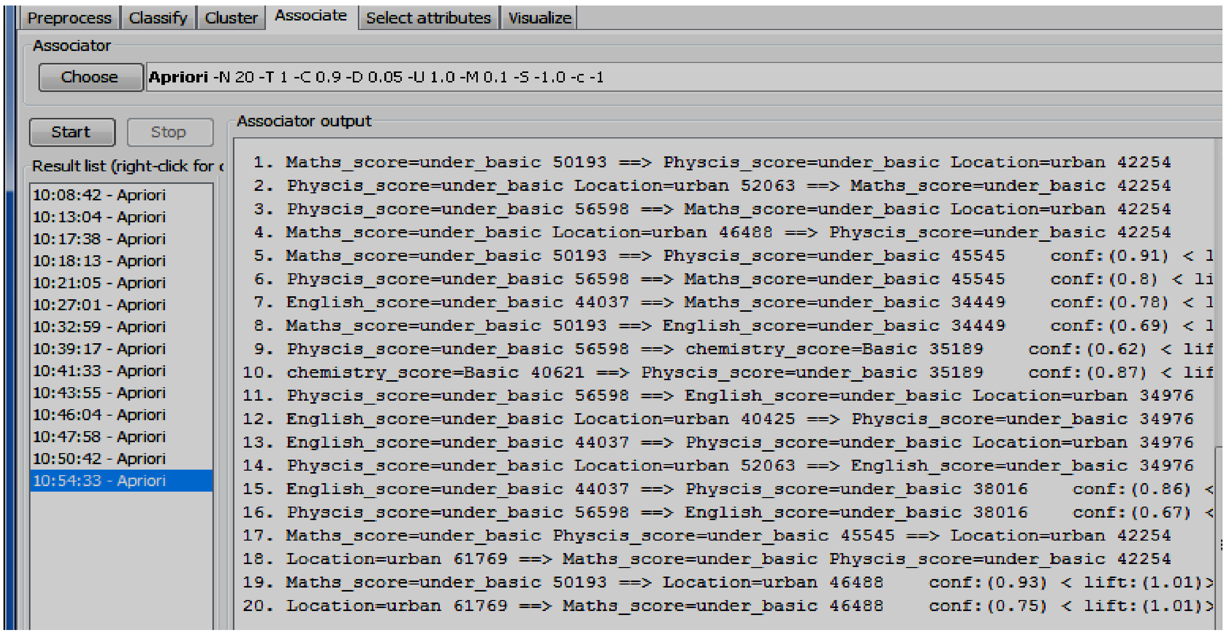

Fig5.2. association rule extracted among grade 12 science subject student result by prior algorithm

The above table also showed that even though not strong, there is relationship among Mathematics, Chemistry and Biology subjects. Out of the total number of students 4,298 students score advance, 11,374 students score proficient, 25,555 students score basic and the remaining 25,973 students score under basic. But total mark affected by civic and aptitude student exam result. 38.6% of the total students achieved less than basic score and only 6.39% achieved advance level. This result is alike with NEAEA assessment result that was the achievement level of students on composite score of the five subjects' namely English, Mathematics, Biology, Chemistry and Physics was (35.6%) by far less than the 50% achievement level expected by the education and training policy of Ethiopia. And also based on the composite score only 13.9% of pupils achieved mastery level and 24.0% were at the basic level and the remaining 62.1% were below the basic level. Tilahun (2010) also indicated that students' achievement of Biology, Chemistry and Physics was below 50.

5.2. Correlation between five subjects and gender difference

5.2.1. Correlation between five subjects and gender deference on grade 10 students

subject	sex	Association rule extracted	Confidence value	Lift value	Leverage value	Conviction value
English	M	Sex =M => English = B	0.66	1.26	0.04	1.16
	F	Sex =M => English = B	0.44	1.26	0.04	1.16
mathematics	M	Sex =M => math = A	0.22	1.33	0.33	1.07
		Math = A => sex = M	0.78	1.33	0.33	1.9
	F	Sex = F => math = C	0.48	1.23	0.04	1.17
Physics	M	Sex =M => physics = B	0.4	1.1	0.02	1.06
		Sex =M => physics = B	0.65	1.1	0.02	1.07
	F	Sex =F => physics = C	0.45	1.06	0.01	1.05
Chemistry	M	Sex =M => chemistry = A	0.37	1.34	0.05	1.15
		chemistry =A => sex = M	0.79	1.34	0.05	1.92
		Sex =M => chemistry = B	0.37	1.04	0.01	1.03
		chemistry = B=> sex = M	0.61	1.04	0.01	1.07
	F	Sex =F => chemistry = C	0.41	1.4	0.05	1.2
		chemistry =C => sex = F	0.58	1.4	0.05	1.39
Biology	M	Sex =M => biology = A	0.4	1.33	0.06	1.16

	biology =A => sex = M	0.78	1.33	0.06	1.87
	Sex =M => biology = B	0.36	1.07	0.01	1.04
	biology =A => sex = M	0.63	1.07	0.01	1.11
F	Sex =F => biology = C	0.46	1.46	0.06	1.26
	biology =C => sex = F	0.6	1.46	0.06	1.48

Table 5.3. Association rule extracted on grade 10 students' result on science subject and gender

Association rules between gender differences

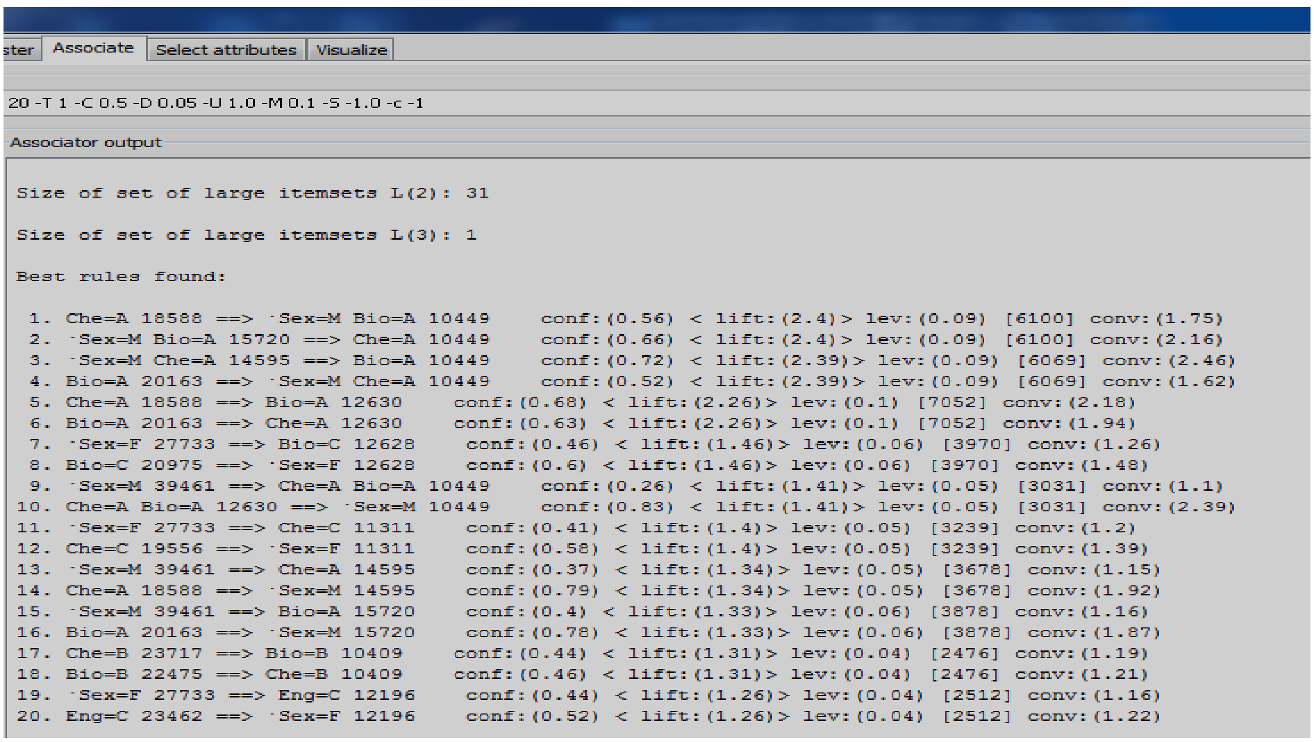

Figure 5.3. Association rule extracted on grade 10 students' result on science subject and gender

Fig 5.3 shows 41.26% of the students are female and the remaining 58.73% are male. From the above table we found that being male student and score grade "B" in English, chemistry and physics subjects has a correlation and also male students and chemistry and biology subject grade "A" have a strong correlation relative to in English and physics student result. Therefore we can clearly see male student relatively have better performance than female students in all five subjects on grade 10 students. This implies that there is a remarkable gender difference in academic achievements in these subjects. This finding supports previous research findings at various educational levels found by different researcher study result conduct in this area.

5.2.2. Correlation between five subjects and gender deference on grade 12 students

subject	sex	Association rule extracted	Confidence value	Lift value	Leverage value	Conviction value
English	M	Sex =M => English_mark = “basic”	0.64	1.04	0.01	1.02
	F	Sex =F => English_mark = “under_basic”	0.67	1.04	0.01	1.16
mathematics	M	Sex =M => math|_score = “basic”	0.61	1.03	0	1.05
	F	Sex = F => math_score = “under_basic”	0.76	1.02	0.01	1.05
Physics	M	Sex =F => physics_score = “under_basic	0.85	1.01	0	1.05
Chemistry	M	Sex =M => chemistry_score = “basic”	0.61	1.01	0	1.01
		chemistry_score = “basic”=> Sex =M	0.59	1.01	0	1.01
		Sex =F => chemistry_score = “under_basic”	0.47	1.07	0	1.03
Biology	M	Sex =M => biology_score = “advance”	0.39	1.04	0.01	1.03
		biology_score=advance” => Sex =M	0.61	1.04	0.01	1.03
		Sex =F=> biology_score = “basic”	0.50	1.02	0	1.01
Total_mark	M	Sex =M => total_score = “basic”	0.62	1.04	0.01	1.04
		Sex=”F”=> total_score = “under_basic”	0.51	1.06	0.01	1.06

Table 5.4. Association rule extracted on grade 12students result on science subject and gender

Likewise grade 10 students results have gender difference, the above table showed, even though it’s weak, there is a correlation between the student academic results of grade 12 on gender difference. Generally male students achieve good result in all Science, Mathematics and English subjects than female students. As Methasbiya (2013) found male students have better GPA than that of the female students in her research on private higher education. She mentioned that in developing countries like Ethiopia females are more accountable and pressured in household activities especially in rural areas. This study finding supports national educational assessment result, which is, boys achieve in the composite score and in the five subjects better than girls. The differences in all subjects were statistically significant at $p < .001$. This shows that the gender gap in academic achievement is persistent in Ethiopia. Tilahun(2010) found that boys performed better than girls on science subject. These research findings do confirm the existence of a significant gender gap in educational achievement among students. This is also common for students from low socioeconomic status as Metasbiya (2013) found. This suggests that the current policy should focus on improving the performance of female students. It is also pointed out that any programs cater for female from low socio economic families.

5.3. Association between regions and student academic result

```
 1. Che=A 18588 ==> Bio=A 12630    conf:(0.68) < lift:(2.26)> lev:(0.1) [7052] conv:(2.18)
 2. Bio=A 20163 ==> Che=A 12630    conf:(0.63) < lift:(2.26)> lev:(0.1) [7052] conv:(1.94)
 3. Che=B 23717 ==> Bio=B 10409    conf:(0.44) < lift:(1.31)> lev:(0.04) [2476] conv:(1.19)
 4. Bio=B 22475 ==> Che=B 10409    conf:(0.46) < lift:(1.31)> lev:(0.04) [2476] conv:(1.21)
 5. Eng=C 23462 ==> Mat=C 10714    conf:(0.46) < lift:(1.2)> lev:(0.03) [1758] conv:(1.14)
 6. Mat=C 25648 ==> Eng=C 10714    conf:(0.42) < lift:(1.2)> lev:(0.03) [1758] conv:(1.12)
 7. Eng=B 31800 ==> Bio=A 10991    conf:(0.35) < lift:(1.15)> lev:(0.02) [1448] conv:(1.07)
 8. Bio=A 20163 ==> Eng=B 10991    conf:(0.55) < lift:(1.15)> lev:(0.02) [1448] conv:(1.16)
 9. Mat=C 25648 ==> Phy=C 11920    conf:(0.46) < lift:(1.09)> lev:(0.01) [952] conv:(1.07)
10. Phy=C 28733 ==> Mat=C 11920    conf:(0.41) < lift:(1.09)> lev:(0.01) [952] conv:(1.06)
11. Phy=C 28733 ==> Bio=B 10386    conf:(0.36) < lift:(1.08)> lev:(0.01) [775] conv:(1.04)
12. Bio=B 22475 ==> Phy=C 10386    conf:(0.46) < lift:(1.08)> lev:(0.01) [775] conv:(1.06)
13. Phy=C 28733 ==> Che=B 10907    conf:(0.38) < lift:(1.08)> lev:(0.01) [765] conv:(1.04)
14. Che=B 23717 ==> Phy=C 10907    conf:(0.46) < lift:(1.08)> lev:(0.01) [765] conv:(1.06)
15. Eng=B 31800 ==> Bio=B 11402    conf:(0.36) < lift:(1.07)> lev:(0.01) [765] conv:(1.04)
16. Bio=B 22475 ==> Eng=B 11402    conf:(0.51) < lift:(1.07)> lev:(0.01) [765] conv:(1.07)
17. Eng=B 31800 ==> Mat=B 12245    conf:(0.39) < lift:(1.06)> lev:(0.01) [708] conv:(1.04)
18. Mat=B 24378 ==> Eng=B 12245    conf:(0.5) < lift:(1.06)> lev:(0.01) [708] conv:(1.06)
19. Eng=C 23462 ==> Phy=C 10495    conf:(0.45) < lift:(1.05)> lev:(0.01) [462] conv:(1.04)
20. Phy=C 28733 ==> Eng=C 10495    conf:(0.37) < lift:(1.05)> lev:(0.01) [462] conv:(1.03)
```

Fig5.4. association rule extracted among science subjects by Aprior algorithm

According to this particular study Aprior algorithm can't extract the rules which indicate the significance relationship between the student academic results among the regions. In contrast to many academic research findings, geographical location has impact on student academic performance. However, this research finding does not support such a conclusion. This implies that geographical location was not a significant predictor of academic achievement, whether children live in Amhara region or Oromiya region is not found to be significant unless there are other factors.

5.4. Association between student location and academic achievement

subject	Association rule extracted	Confidence value	Lift value	Leverage value	Conviction value
English	English= "A"=> location = "urban"	0.94	1.02	0	1.3
	Location="urban"=> English = "A"	0.11	1.02	0	1
Mathematics	math = "C" => location = "urban"	0.92	1.01	0	1.08
	Location = "urban" => math = "C"	0.38	1.01	0	1
Physics	Physics="D" =>location = "urban"	0.92	1	0	1.03
	location = "urban=> Physics ="D"	0.12	1	0	1.03
	Physics="B" => location = "urban"	0.92	1	0	1
	location = "urban=> Physics ="B"	0.38	1	0	1
Chemistry	Chem = "C" => location ="urban"	0.93	1.01	0	1.14
	Location ="urban " => chem ="C"	0.29	1.01	0	1
Biology	biology="A" => location ="urban"	092	1	0	1.05
	Location="urban"=>biology ="A"	0.30	1	0	1
	biology="C" => location ="urban"	0.92	1	0	1.01
	Location="urban" => biology ="C"	0.31	1	0	1

Table 5.5. Association rule extracted on grade 12students result on science subject and student location

subject	Association rule extracted	Confidence value	Lift value	Leverage value	Conviction value
English	English_score = ”under_basic“=> location = “urban”	0.92	1	0	1.02
Mathematics	math = ”C“ => location = “urban”	0.93	1.01	0.01	1.13
Physics	Physics_score=”under_basic” =>location = “urban”	0.92	1	0	1.02
Chemistry	Chem_score = “under_basic” => location =”urban”	0.95	1.03	0	1.14
Biology	Biology_score=“basic” => location =”urban”	0.94	1.02	0	1
Total_mark	Location=”urban”=>total_mark =”basic”	0.49	1.02	0	1.02

Table 5.6. Association rule extracted on grade 12students result on science subject and student location

This data in the table above revealed that school location had no significant influence (strong relationship) on the student academic result. On contrary to NEAEA assessment report on grade 8, Students from rural schools performed better than those from urban ones in all subjects except in English. In English students from urban schools performed better than those in rural ones, This finding revealed that whether a student attends rural or urban secondary school it does not associated with his/her academic result., while urban schools have better facility than rural schools, unless schools utilized the available resources properly, facilities couldn’t improve the student result by themselves. According to World Bank report on accomplishment performance of ESDP IV, in almost all secondary schools, basic minimum facilities, like laboratories equipment, chemicals, apparatus, library facilities and furniture are not in place or out-dated. Tilahun (2010) also stated that 50% percent of students in Addis Ababa city administration primary schools do not know the availability or presences of practical and science kits in the schools. Even most of them did not conduct practical activities during their science lesson. This study result support what Joseph (2010) found on his research, that there is no significant difference between rural and urban secondary school in terms of academic performance in agriculture science because of both rural and urban schools almost have the same school facilities like laboratory, text books and library. But in Ethiopian school context, there is proper utilization problem of the existing facilities in urban areas which enhance the student performance and academic results.

```
 1. Che=C 19556 ==> location=urban 18129    conf:(0.93) < lift:(1.01)> lev:(0) [207] cor
 2. location=urban 61579 ==> Che=C 18129    conf:(0.29) < lift:(1.01)> lev:(0) [207] cor
 3. Mat=C 25651 ==> location=urban 23661    conf:(0.92) < lift:(1.01)> lev:(0) [153] cor
 4. location=urban 61579 ==> Mat=C 23661    conf:(0.38) < lift:(1.01)> lev:(0) [153] cor
 5. Bio=A 20163 ==> location=urban 18563    conf:(0.92) < lift:(1)> lev:(0) [84] conv:(1
 6. location=urban 61579 ==> Bio=A 18563    conf:(0.3) < lift:(1)> lev:(0) [84] conv:(1)
 7. Bio=C 20975 ==> location=urban 19247    conf:(0.92) < lift:(1)> lev:(0) [24] conv:(1
 8. location=urban 61579 ==> Bio=C 19247    conf:(0.31) < lift:(1)> lev:(0) [24] conv:(1
 9. Phy=B 25710 ==> location=urban 23565    conf:(0.92) < lift:(1)> lev:(0) [3] conv:(1)
10. location=urban 61579 ==> Phy=B 23565    conf:(0.38) < lift:(1)> lev:(0) [3] conv:(1)
11. Eng=C 23462 ==> location=urban 21487    conf:(0.92) < lift:(1)> lev:(0) [-14] conv:(
12. location=urban 61579 ==> Eng=C 21487    conf:(0.35) < lift:(1)> lev:(0) [-14] conv:(
13. Phy=C 28735 ==> location=urban 26297    conf:(0.92) < lift:(1)> lev:(0) [-36] conv:(
14. location=urban 61579 ==> Phy=C 26297    conf:(0.43) < lift:(1)> lev:(0) [-36] conv:(
15. Eng=B 31800 ==> location=urban 28956    conf:(0.91) < lift:(0.99)> lev:(0) [-186] co
16. location=urban 61579 ==> Eng=B 28956    conf:(0.47) < lift:(0.99)> lev:(0) [-186] co
17. Bio=B 22475 ==> location=urban 20459    conf:(0.91) < lift:(0.99)> lev:(0) [-137] co
18. location=urban 61579 ==> Bio=B 20459    conf:(0.33) < lift:(0.99)> lev:(0) [-137] co
19. Che=A 18588 ==> location=urban 16920    conf:(0.91) < lift:(0.99)> lev:(0) [-114] co
20. location=urban 61579 ==> Che=A 16920    conf:(0.27) < lift:(0.99)> lev:(0) [-114] co
```

5.5. **Association between grade 10 and 12 student academic result**

Fig 5.5. association rule extracted among science subjects and student location by Aprior algorithm

5.6. Association between Plasma and students' national exam result

5.6.1. Association between Instructional media (PLASMA) and grade 10 students national exam result

subject	Association rule extracted	Confidence value	Lift value	Leverage value	Conviction value
English	No rule extracted	-	-	-	-
maths	Math="A" => plasma ="yes"	0.9	1.02	0	1.18
	Plasma "yes" => Math = "A"	0.17	1.02	0	1
Physics	Physics ="B" => plasma ="yes"	0.89	1.01	0	1.08
	Plasma "yes" => physics = "B"	0.39	1.01	0	1.01

Chemistry	chemistry ="A" => plasma ="yes"	0.89	1.01	0	1.07
	Plasma "yes" => chemistry = "A"	0.28	1.01	0	1
Biology	biology ="A" => plasma ="yes"	0.9	1.02	0.01	1.21
	Plasma "yes" => biology = "A"	0.31	1.02	0.01	1.01
	biology ="B" => plasma ="yes"	0.88	1	0	1.01
	Plasma "yes" => biology = "B"	0.34	1	0	1.01

Table 5.7. Association rule extracted on grade 10 students' result on science subject and instructional media technology (PLASMA)

Association between Instructional media (PLASMA) and students result on grade 10

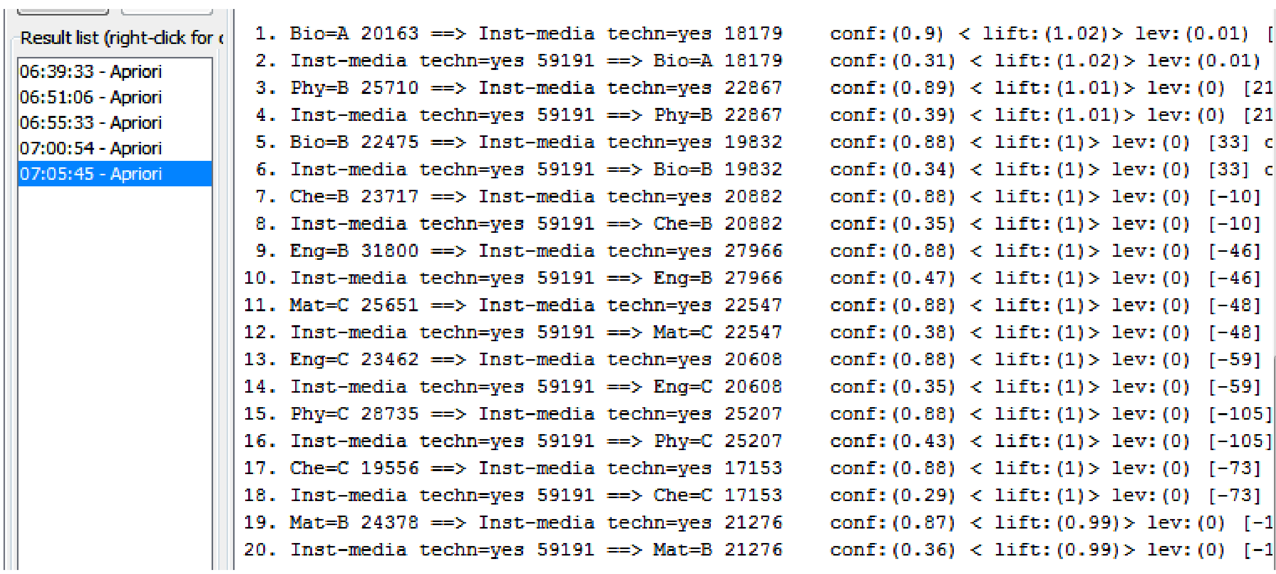

Fig5.6. association rule extracted among grade 10 science subject student result by prior algorithm

5.6.2. Association between Instructional media (PLASMA) and students result on grade 12

subject	Association rule extracted	Confidence value	Lift value	Leverage value	Conviction value
English	English _mark= “basic”=> plasma =”yes”	0.89	1.01	0	1.05
	English _mark= “basic”=> plasma =”yes”	0.8	1	0	1
maths	Math_score=”basic” => plasma =”yes”	0.88	1	0	1
	Plasma ”yes” => Math_score = ”under_basic”	0.74	1	0	1
Physics	Physics_score =”basic” => plasma =”yes”	0.88	0.99	0	0.96
	Plasma ”yes” => physics = ”B”	0.39	1.01	0	1.01
Chemistry	Chemistry_score =”above_basic” => plasma =”yes”	0.88	1	0	0.99
	Chemistry_score = ”basic” => Plasma ”yes”	0.88	1	0	097
Biology	Biology_score=”basic” => plasma =”yes”	0.88	1	0	1.03
	Biology_score =”above_basic” => plasma =”yes”	0.88	0.99	0	0.96
Total_mark	Total_score ”basic” => plasma = ”yes”	0.5	1	0	1

Table 5.8. Association rule extracted by Aprior algorithm on grade 12 student result and plasma technology

```
Minimum support: 0.1 (6739 instances)
Minimum metric <lift>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Size of set of large itemsets L(2): 2

Best rules found:

 1. instractional media tech=yes 59373 ==> total_score=under_basic 28573    conf:(0.48) < lift:(1)> lev:(0) [76] conv:(1)
 2. total_score=under_basic 32344 ==> instractional media tech=yes 28573    conf:(0.88) < lift:(1)> lev:(0) [76] conv:(1.02)
 3. instractional media tech=yes 59373 ==> total_score=Basic 29904    conf:(0.5) < lift:(1)> lev:(0) [-107] conv:(1)
 4. total_score=Basic 34064 ==> instractional media tech=yes 29904    conf:(0.88) < lift:(1)> lev:(0) [-107] conv:(0.97)
```

Fig 5.7. Association rule extracted by Aprior algorithm on grade 12 student result and plasma technology

As we can see from the result on the above table there is a correlation between student academic result and the availability of instructional media technology (plasma) in Biology and physics subjects though it is weak. But there is no correlation between other subject's student result and availability of plasma is very weak. This indicates that schools might not utilize the technology properly in teaching and learning process. According to Gara(2012) teaching students through plasma didn't contribute a lot to the better understanding of the biology lesson compared to the conventional method of teaching. This might be due to improper implementation or utilization of the instructional plasma. But instructional technology could significantly help students' learning and improve their academic performances. Birhanu (2011) mentioned that some of the reasons for this are teachers lack of using and integrating plasma technology with annual plan and ,lack of awareness about plasma TV program can improve students result and help reduce teachers burden to explain abstract ideas specially on science subjects curriculum. Therefore this leads to affect negatively both teachers and learner in their teaching learning process. In parallel to the Birhanu's work, Abathun(2010) also found that the mathematics achievement score of plasma instructed students is less than that of the teacher-based instructed students: Some of the reasons were most students have difficulty to understand the mathematics lesson due to the pronunciation of plasma teacher and not sufficient time to give to do exercise in the class work and interact with plasma teacher. So most student had negative attitude to learn through plasma. This leads to low performance of students result and lack of interest to open and follow the lesson broadcasted from the center.

5.7. Association between grade 10 and grade 12 student result

subject	Association rule extracted	Confidence value	Lift value	Leverage value	Conviction value
English	English =C => English 12="under_basic"	0.84	1.24	0.05	1.78
	English =B=> English 12="Basic"	0.53	1.14	0.02	1.06
maths	Math="c" => maths12="under_basic"	0.78	1.05	0.01	1.18
	Math="B" => maths12="under_basic"	0.76	1.02	0.01	1.06
Physics	Physics ="D" => physics12 ="under_basic"	0.85	1.01	0	1.11
	Physics ="C" => physics12 ="under_basic"	0.85	1.02	0	1.08
Chemistry	chemistry ="C" => plasma ="under_basic"	0.89	1.0	0	1.07
Biology	biology ="A" => Biology12 ="above_basic"	0.47	1.27	0.03	1.19
	biology ="B" => Biology12 =" basic"	0.52	1.27	0.01	1.18
	biology ="C" => Biology12 =" basic"	0.54	1.04	0.01	1.04

Table 5.9. Association rules extracted by Aprior algorithm on grade 10 and grade 12 student results.

5.8. Modeling

Association between grade 10 and 12 subjects student academic result

Here The researcher cfsubsetEval filtering in order to select the best attributes to generate the most important rules.   As a result chemistry students' result of grade 10, English and Biology grade 12 subjects result are selected based on the predictive ability of each feature with the class attribute that is grade 12 total marks.

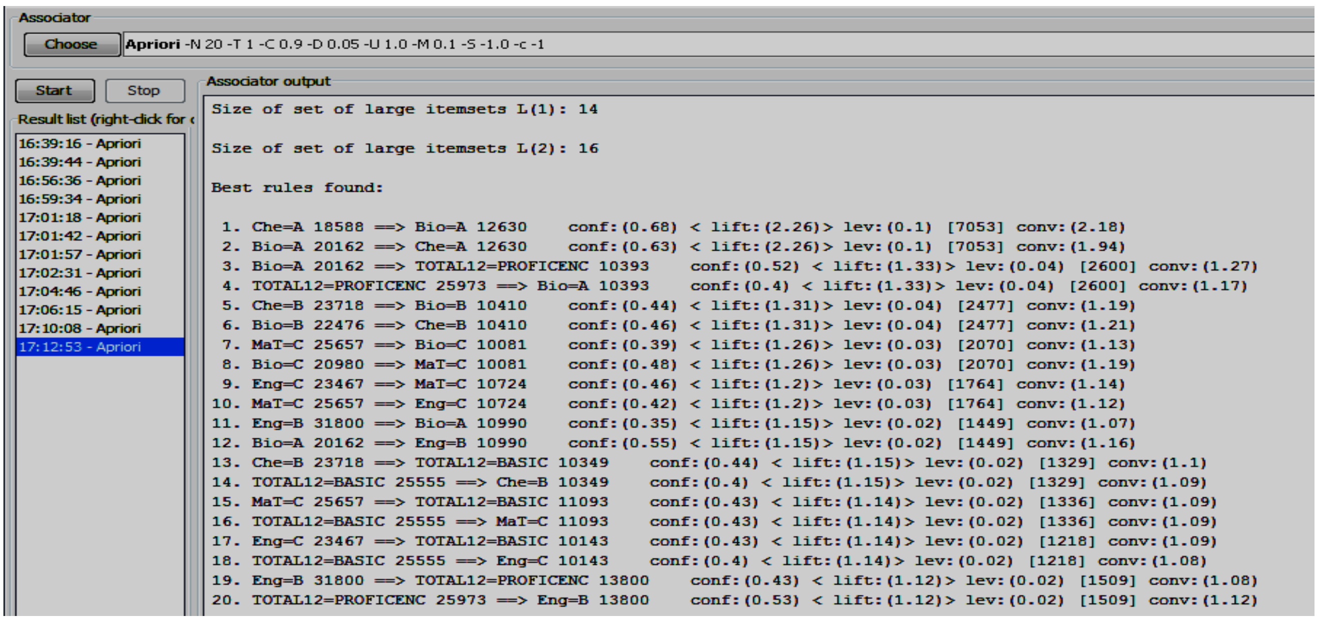

Fig5.8. association rule extracted among grade 12 science subject student result by prior algorithm

The proposed model is used by the concerned body for further actions. The teacher can use the discovered information for making decisions about the students' exam achievement of the subject in order to improve the students' learning. Next, the researcher is going to explain the meaning of some obtained rules. Figure 5.8 shows how a huge number of association rules can be discovered. There are also a lot of uninteresting rules, like a great number of already extracted rules during subjects association (rules with a generalization of relationships among subjects like rules 1, 2, and rule 5, 6, 7 … and12). But there are also rules that show unexpected relationships which can be very useful for the teachers and education planner in decision making about the activities and detecting students with learning achievements, like rules 3, 13 15, 17and 19 are rules that show relevant information for educational purposes, conforming relationships. If a student has got grade “A” in Chemistry or grade “B” in English at grade 10, he/she has a higher probability to gate “proficient” result in cumulative mark in grade 12. Since Biology subject students’ result has better correlation with English, Maths and Chemistry subject in grade 10 and 12 based on this research finding, it is logical. if the a student have got minimum “B” in Biology or “C” in English and Mathematics subject, then the final mark at grade 12 obtained is “basic”. Starting from this information, the curriculum planner and other concerned bodies can pay more attention to these finding because it gives clue what is the relationships between grade 10 and grade 12 subjects integration for further analysis.

5.9. Evaluation

In this study to evaluate the grade 12 students exam result based on grade 10 national exam result, the researcher used the confidence and lift value between antecedent and consequent and the occurrence of one implies the occurrence of the other. For example the first rule with lift is 1.3 means there is a positive correlation between the antecedent grade 10 student biology results and the consequent total mark on grade 12 student. With the lift value, we can measure the importance of a rule. The first rule, with the highest lift which means highest correlation is the most important, and so on. To interpret the rules in the association rules model, the second rule means that of the government secondary school students sat on national exam, 68% (confidence or certainty) have scored chemistry and Biology mark “A”.

5.12. Knowledge representation by weka tools

In this research The researcher Knowledge Flow interface in Weka, which allows for assembling of various types of modules that are a part of the data mining process. First loaded a file in Arff format, viewed it using a text viewer module for validating purposes, ran it through the association rule finding module, out of which it flowed to the text viewer to show the results, as shown in Fig 5.9 below

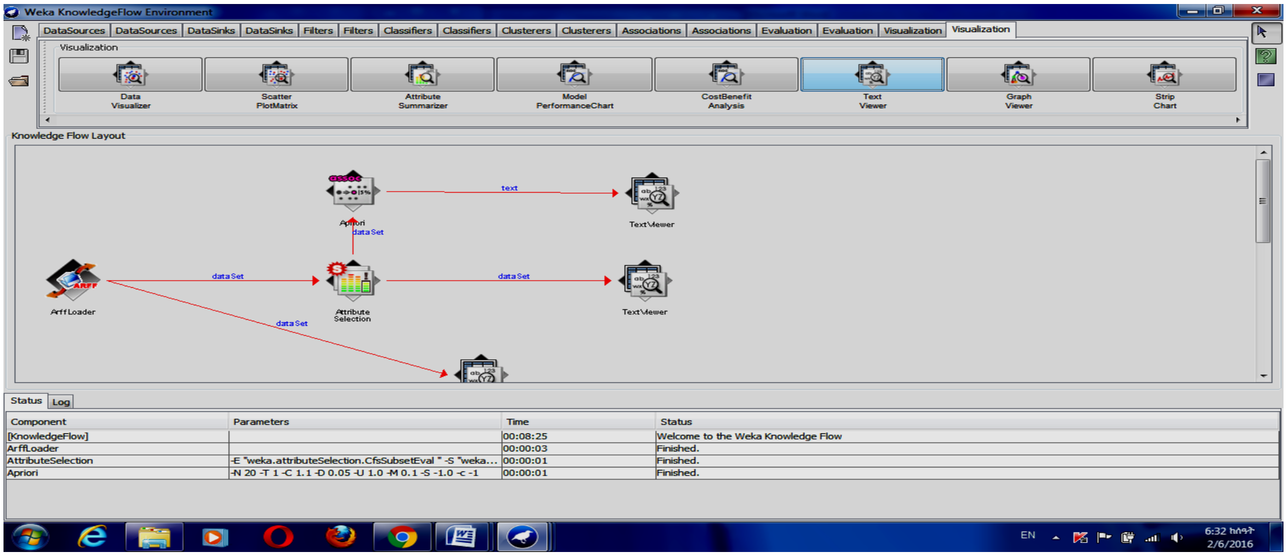

Figure 5.9 knowledge representation of extracting association rule by apripor algorithm

The researcher built an Arff loader that performs to extracted important rules by Aprior algorithm. This showed that the overall data preprocessing and rule generating procedure during this particular study by using weka data mining tool.

5.10. Challenges in understanding the rules

Considering association rules, in order to identify those that are truly interesting i.e those that give potentially useful information about the behavior of the education dataset. The main obstacles the researcher faced were as follows.

1. The number of potentially interesting rules generated by Apriori algorithm that the researcher would not identify as truly interesting was overwhelming. Browsing through so many rules that seemed not to give any useful information, it was quite difficult to find rules that could potentially be truly interesting.
2. The interestingness measures were not reliable enough to reflect the true interestingness of the rules for instance there is weak correlation between the subject mathematics and physics in term of students result in contrary to other researcher finding.
3. Since the difference in the age of students is very narrow, the association rule extracted by Aprior algorithm was not meaningful.

4. It was difficult to associate student academic result with the qualification of teachers due to the information stored in the database inadequate and incomplete.

Chapter six

Conclusion and Recommendation

In Ethiopia, huge progress has been made at all levels of education in the past years. Enrollments are increased in both primary and secondary schools. But it is difficult to translate these achievements in to concrete improvements in students' competencies. The truly important measure is not only enrollment and attendance but how learning is going on and how the students' academic performance is improved in terms of academic result. Based on the finding of this study, the following conclusion and recommendation are made.

6.1. Conclusion

The academic achievement result of the male students in both five subjects namely English, Mathematics, Physics, Chemistry and Biology on both grade 10 and 12 have a better correlation than female students.
Based on the rating scale put on educational training policy (MOE, 2008) The student achievements score in grade 12 were dividing in to four standards as: under basic, Basic, proficient and advance. The rate scale is less than 50, between 50 and 60, greater than 60 to 70 and above 70 respectively. Among the total number of students sat on the grade 12 exam which are included in this study, around 1.5% were scored above basic,50.5 % achieved basic and 48% were below basic. This indicates that most student still under basic academic performance.

Both in grade 10 and 12, school using plasma system were identified and the achievement score of students using plasma system and those not using the system were compared. There is no observable strong correlation between students' achievement in the subjects and the two groups (using and not using). In grade 10 and 12 there exists positive relationships between the five subjects. Moreover Mathematics, English, Biology and Chemistry have better correlation than others. This shows that students performing well in one subject, they also do the same in the other subject. Looking at instructional language English and other subjects it was the highest with biology (English="B"=>biology="A" con=0.77 and lift=1.26). Subjects such as biology when compared with other subject are more influenced by language ability. The same like chemistry and biology are strongly correlated each other.

According to this particular study geographical location was not a significant predictor of academic achievement. This means whether children live in Amhara region or Oromiya region have no significant impact on students achievement unless there are other factors. As most of previous research reveal that urban locations have better infrastructure and facilities than rural areas and this will have influence on student academic achievement. Development of geographical location of a certain school has relation with the overall teaching learning process and student achievement directly or indirectly, unless students utilized educational resources and facilities available in schools properly and effectively, students' achievement may not vary just only by their geographical location.

6.2. Recommendation

Based on the findings of this study the following recommendations are made.

The gender gap in secondary school is persistent and there is a need to support female students at low level of academic achievement and make further investigation that would help to narrow the gap between male and female students. Understanding the barriers will guide the design of appropriate polices to enhance the achievements of female student result. Therefore concerned bodies should address provision of additional support for female students.

Intervention should be taken to raise the achievement level of grade 10 and 12 students at all levels. The performance of the student might be attributed to some factors outside of the scope of this study. This may both inside and outside school factors like low parental involvement in school administration, poor motivation of teachers due to low salary, lack of student interest for different subject and some socioeconomic background of the students. So creating conducive atmosphere for effective teaching and learning process help to ensure better academic performance.

Motivational instructional design should be considered for improving students' active participation in satellite TV program. In order to resolve the intractability and fixed schedule problem of the existing broadcasting system the ministry and other stakeholder should design and develop new technology alternatives to deliver to school according to their demand like two-ways communication, individual, self-directed, self-paced learning environments. Besides this there should be countermeasures for raising teacher's integration of the satellite TV program to the class. No matter how difficult it is, great awareness should be made for teachers, students and the community at large concerning plasma TV. Likewise all education stakeholders should find means of accessibility and encourage students to utilize effectively ICT in science education, micro-scale laboratory apparatus, low cost apparatus and chemicals

In order to bring about tangible results on the quality of science learning in Ethiopia multi faceted efforts and comprehensive interventions should put into the system, like interventions on curriculum, instruction, assessment, teacher training, learning conditions, and instructional materials. This study only stick on the student national examination result of grade 10 and 12 students' data. The researcher recommended that further study should be conducted both inside and outside school factors which have not seen under this study.

There is a gender gap in academic achievement. Therefore Ministry of Education should give special support for female students to succeed in their academic career.

6.3. Limitation of Study

The applicability of data mining to NEAEA database is limited to develop and test the model instead of deploying the model at the MOE. Since the study is carrying out for academic achievement. That is, the scope of the current experimental research undertaking is strictly limited to evaluate the potential applicability of data mining technology to support analysis activities at the NEAEA study area.

Although several techniques and algorithms are available in data mining tasks, the researcher concentrate on applying Aprior association rule.

Another limitation for this study is that only 67,200 of records are taken due to the scalable constraint of Weka heap size (151.5 MB memory size), spelling error of student name during registration at different time (grade 10 and 12) and this study focused on only government regular students result. Such kind of constriction forces the researcher to be limited with only one technique and an algorithm instead to make further more different experiments for discovering knowledge.

6.4. Future study

Though Aprior association rules easy to implementation than others, PedictiveApriori Association Rule Algorithm, FP-Growth and FilteredAssociator, it scan database repeatedly for searching frequent itemsets, so more time and resource are required in large number of scans so it is inefficient in large datasets. Association rules produced by the abovementioned algorithms can be combined for better results and an efficient algorithm for any real life application. Thus it would be a future research area to minimize cost and time when applied for any size data sets. Thus if this is implemented successfully, it reduces cost and time.

References

- Abatihun Alehegn, Sewagegn,A (October 12, 2012) *Comparative Study on Mathematics Achievement of Grade Ten Students as Observed in Two Methods of Instruction*.
- Adegboye, A.O. (1993). *Proficiency in English Language As A Factor Contributing To Competency In Mathematics*. Education Today, 6 (2), 9-13.
- Adegboye, A.O. (1997). *The Study of the Relationship Between Students' Performance In English Language And Science*. Pro Science Education Journal. 1&2. 73-83.
- Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. (2006) *Mining Student Data Using Decision Trees, The 2006 International Arab Conference on Information Technology (ACIT'2006).*
- Anuradha .C and Velmurugan . T. (2015). *A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance, Indian Journal of Science and Technology, Vol 8(15), July 2015*
- Beikzadeh,M. and Delavari, N., (July 2005) *A New Analysis Model For Data Mining Processes In Higher Educational Systems. On the proceedings of the 6th Information Technology Based Higher Education and Training 7-9*
- Berhanu Abera. (August 2012). *The Plasma-based Instruction in Ethiopia*
- Berry, J.A.M., and Linoff, G. (2007) Data mining techniques-for marketing, sales and customer support, New York: Wiley.
- Bharadwaj B.K. and Pal .S, (2011). *"Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140,*
- Chakrabarti .S , L.,Thomas P. , Toby J. and Witten H. (2009). *Data mining know it all*. Morgan Kaufmann Publishers 30 Corporate Drive, Suite 400 Burlington, Unite State
- Chapman. P, Julian C., Randy K.,Thomas K., Thomas R.,Colin S., and Rüdiger W. (2000). CRISP-DM 1.0 – Step -by-step data mining guide. Available at www.spss.ch/file.php?file=/upload/1107356429_CrispDM1.0.pdf.
- Daniel T.Larose(2005). *DISCOVERING KNOWLEDGE IN DATA*, An Introduction to Data Mining.
- Deshpande. S. P and Thakare .V. M, (2010). *Data Mining System and Applications: A Review*. International Journal of Distributed and Parallel systems (IJDPS) Volume 1, Number 1: 445-463
- Durairaj .M, Vijitha .C, (2014). *Educational Data mining for Prediction of Student Performance Using Clustering Algorithms, International Journal of Computer Science and Information Technologies*
- Ebenuwa-Okoh. (2010) *Influence of Age, Financial Status, and Gender on Academic Performance among Undergraduates*
- Endris,A.E, Mulugeta,A.D. (2015). *Analysis of Determinants Influencing Students' Academic Achievement in English and Mathematics Subjects of Ending Classes of Schools at Bahir Dar and Debre Tabor Towns*, Ethiopia.
- Fayyad U., Piatetsky-Shapiro G., and Smyth P. (1996). *From Data Mining to Knowledge Discovery in Databases.* Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.42.1071&rep=rep1&type=pdf.

- Fergusson DM, Horwood LJ.(2007) *Gender Differences In Educational Achievement In A New Zealand Birth Cohort*. New Zealand Journal of Educational Studies
- Frank, E and Witten, H. (2005). Data mining Text book: *Practical Machine Learning Tools And Techniques*. 2nd Ed. Morgan Kaufmann Publishers, San Francisco
- Gare a Latchanna and GarkeboBasha (November 2012), *Effectiveness Of Plasma Television Integration In Teaching Secondary School Biology* Vol. 2, No. 2 pp. 215-230
- Han,J. and Kamber. M., (2006)."*Data Mining Concepts And techniques*", 2nd edition,
- Hijazi S. T., and Naqvi R. (2006), "*Factors Affecting Student's Performance*: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1,
- Hussain, Ch. Abid. (June 2006). *Effect Of Guidance Services On Study Attitudes*, Study Habits and Academic Achievement of Secondary School Students. Bulletin of Education and Research, vol. 28, No. 1 (35-45).
- Kim.S and Bogale.Gebeyehu. (2013) *Does Satellite Television Program Satisfy Ethiopian Secondary School Students?*
- Kumar,V, Anupama,C. (September 2012). *Mining Association Rules In Student's Assessment Data*, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3,
- Lori Bowen Ayre. (2006). *Data Mining for Information Professionals. Available at http://techessence.info/files/Ayre_DataMiningForInformationProfessionals_June2006.pdf*
- Maharashtra.(March 2014). *India International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 3.*
- Metasebia Bekele.(2013*).Gender Difference in Academic Achievement at Private Higher Education Institutions:* The case of selected PHE Institutions in Addis Ababa.
- Ministry of Education. (February, 2002). *The Ethiopian Education And Training Policy And Its Implementation*.
- Ministry of Education. (2010). *Educational Sector Development Program* IV (2010-2014)
- Ministry of Education, (July 2008). *Ethiopian Third National Learning Assessment Of Grade Eight Students* report, Ababa, Ethiopia.
- Miriam R. Linver. (2010) *Influences of Gender on Academic Achievement Teachers College*, Columbia University Pamela E. Davis-Kean University of Michigan Jacquelynne S. Eccles University of Michigan
- *Meltem Dayıo_lu and Serap Türüt-A_ık. (December 2004). Gender Differences in Academic Performance in a Large Public University in Turkey, ERC Working Papers in Economics 04/17*
- Muluken AlemuYehuala,(2015) *Application Of Data Mining Techniques For Student Success And Failure Prediction* :The Case Of Debre_Markos University.
- MS Shweta, Dr .Kanwal Garg,(June 2013) *MiningEfficient Association Rules through AprioriAlgorithm Using Attribute and comparativeAnalysis of Various Association RuleAlgorithm* , In: volume 3, Issue 6
- *New Media, Democratization and 'Going with the Grain* http://hornaffairs.com/en/2014/07/06/ethiopia-new-media-democratization

- Noble, Julie P,Roberts, William L. and Sawyer Richard L. (October 2006). '*Student Achievement, Behavior, Perceptions, and Other Factors Affecting ACT Scores'*. ACT Research Report Series 200ining6 - 1.
- Plate T., Band P.r, Joel B. and John G. (1997). *A comparison between neural networks and other statistical techniques for modeling the relationship*. Advances in Neural Information Processing, MIT Press.
- *Praveen.V., Sunda. (Feb. 2013), 42a Comparative Study For Predicting Student's Academic Performance Using Bayesian Network Classifiers Vol. 3, Issue 2 PP 37-*
- Ruth Zuzovsky. (2003). *Changed Teachers' Qualifications And Their Impact On Student Achievement Findings From Data In Israel*
- Santos M.F. and Azevedo C. (2005). *Data Mining*: Descoberta de Conhecimento em Bases de Dados, FCA Editora, Lisbon.
- Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, (2010). *Data mining model for higher education system, Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29*
- SuchitaBorkar, K. Rajeswari. *(2013). Predicting Students Academic Performance Using Education Data Mining.*
- Svetlana S. Aksenova. (2004). *Machine Learning with WEKA Explorer Tutorial* for WEKA Version 3.6.4, School of Engineering and Computer Science Department of Computer Science California State University,
- Tilahun Tesfaye, Sileshi Yitbarek, and Anteneh Tesfaye *(December, 2010). Status Of Science Education In Primary Schools Of Addis Ababa*, By City Government of Addis Ababa. *Education Bureau*
- Two Crows Corporation. (2005). *Introduction to Data Mining and Knowledge Discovery. 3rd* Ed. Two Crows Corporation. 500 Falls Road, Potomac, USA. available at www.twocrows.com
- *Wakshum Fekadu (January 2013). Teaching - Learning Through Plasma Television and Its Challenges In Selected Secondary Schools Of Addis Ababa,*
- UNESCO(2005). Education for sustainable development Goal online available at *https://www.google.com.et/?gws_rd=cr,ssl&ei=6MSxVtz8Jor7ygPaxaTYBw#q=unesco+2005+education+for+sustainable+development*
- Umesh Kumar Pandey, S. Pal. (March 2011). A *Data Mining View On Class Room Teaching Language IJCSI.* International Journal of Computer Science Issues, Vol. 8, Issue 2, *www.IJCSI.org*
- www.iosrjen.org