Developing a Predictive Model to Determine Higher Education Students' Academic Status Using Data Mining Technology

A thesis submitted

By

Sisay Girma

To

The Faculty of Informatics

Of

St. Mary's University

In Partial Fulfillment of the Requirements

For the Degree of Master of Science

In

Computer Science

March, 2019

Acceptance

Developing a Predictive Model to Determine Higher Education Students' Academic Status Using Data Mining Technology

By

Sisay Girma Endale

Accepted by faculty of informatics, St. Mary's University, In Partial Fulfillment of the Requirements for the Degree of Master of Science In

Computer Science

Thesis examination committee:

Dr. Getahun Semeon

Internal examiner

_____

Signature

Dr. Temtim Assefa

External examiner

_____

Signature

March, 2019

Declaration

I, the undersigned, declare that this thesis work is my own original work, has not been presented for a degree in this or any other universities, and all sources of material used for thesis work have been fully acknowledge.

Sisay Girma Endale

Student

_____

Signature

Addis Abeba

Ethiopia

This Thesis has been submitted for examination with my approval as advisor:

Dr. Million Meshesha

Advisor

_____

Signature

Addis Abeba

Ethiopia

March, 2019

Acknowledgement

First and for most, I would like to thank the Almighty GOD and His Mother St. Virgin Marry for giving me strength, courage, and patience in order to accomplish this research

Secondly, I would like to express my appreciation to my advisor Dr. Million Meshesha for his generous help, continual advice, constructive comments and suggestions throughout the preparation of this research paper.

Thirdly, I would like to express my sincerest graduated and heartfelt thanks to Zenebu Menberu and Meskrem Kassaye for their courteous support and helpful personality which lasts in my heart forever.

Fourthly, my sincere thanks go to my family for the lovely and tireless support it gave me throughout my life.

Finally, Grateful thanks also go to the St. Mary's university registrar staff administrative for their unreserved cooperation during the collection of the data.

Table of Contents

List of figures

List of tables

List of acronyms

CGPA Cumulative Grade Point Average

DBMS Data Base Management System

EC Ethiopian calendar

EDM Educational Data Mining

EGSECE Ethiopian General Secondary Education Certificate Examination

EHEEE Ethiopian Higher Education Entrance Examination

PHEI Private Higher Educational Institution

SRMIS Student Record Management information System

Abstract

Nowadays student attrition became a universal problem in most higher education. To improve student retention one should understand the non-trivial reason behind student attrition. Student attrition and retention in private higher institution education(PHIE) can be affected by a wide variety of factors, these factors include, demographic, social, economic, academic and institution aspects, are the major contributing aspects that leads to attrition and retention of students in higher education. The main objective of this study is to develop a predictive model using of data mining technology to determine undergraduate students' attrition or retention in higher.

In this study, the hybrid data mining process model is followed. The hybrid data mining process model has six steps such as understanding of the problem, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge and use of discovered knowledge. In this study based on the problem understanding, 15 attributes are selected and 7361 instances are used to experiment with designing a predictive model that has a capability of determining students' status. In this study, the classification algorithms such as decision tree (J48), rule induction (PART and JRIP), and Bayes classifier (naïve Bayes) are used in the model building process. And 10 fold cross-validation and 66% split test option are used to train and test the classifier model. Among the four algorithms tested, decision tree classifier (J48) algorithm scored the highest accuracy of 91.40%followed by PART, JRIP, and naïve Bayes algorithms respectively. Depending on the extracted hidden pattern using J48 algorithm, financial sources (self-sponsored and parent-sponsored, and scholarship), division (regular and extension), types of preparatory attended school (private and public), department (computer science, accounting, marketing management, hotel and tourism, and management), background of study (social and natural), and preparatory completion year, before(1994EC-2001EC) and after (2002EC-2009EC)) were identified as the major contributing factors behind student attrition and retention(graduated). The data obtained from SRMIS (student record management information system) was in two table format. So merging the two tables into one table format was the major challenge of this study. It is also difficult to get well organized, correct and quality data for the mining tasks. So we suggest educational institutions to maintain their data symmetrically for data analyses.

Keyword: - Educational data mining, status, attrition, J48 decision tree

CHAPTER ONE

Introduction

1.1 background

Knowledge has become the key of the future prosperity and social well-being of the nation and thus, education becomes one of the key sectors that contribute to the economic and social advancement of a nation [1]. Higher education system has a great advantage and brings massive benefits to both the individual and the nation, and thus, the skill and research developed through higher education have great encouragement to any society's success, increasing jobs, and developing prosperity, hence the Universities and Colleges are the ones which play a vital role in expanding this opportunity and promoting justices [2]. In one or other way, students' decision to drop their program will affect the organization, society, and the nation building, so the students' retention should be continued long term goal in all educational institutions.

Data mining is a technology used to describe knowledge discovery in the database and to search for significant relationships such as patterns, associations, and changes among variables in databases [4]. The discovery of those relationships can be examined by using statistical, and machine learning techniques to enable users to extract and identify greater information and subsequent knowledge than a simple query and analysis approach [4]. It is also the process of selection, exploration, modeling of large quantities of data to discover regulations or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the data in the database [5].

Educational data mining(EDM) is "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings with the aim of developing models to improve the learning experience and institutional effectiveness"[6]. As compared to traditional analytical studies; data mining is forward looking to individual students. For example, the predictive function from data mining can help educational institutions to plan for resources based on the knowledge of how many students with what background will take a particular course [7].

Educational data mining (EDM) has two major tasks; predictive and descriptive modeling. According to Archna et.al [5] based on the kinds of pattern data mining tasks can be classified into two categories; predictive and descriptive tasks

Steven et.al [25] presented the application of educational data mining to predict undergraduate student retention based on the student admission records and first year transcription. In this research three different predictive models were built, the first model is called pre-college model that uses pre-college datasets, to predict a graduation rate of newly admitted students and applicants, the second model is called the first term model that first semester GPA and the third model is called the full-data model that encompass all the predictors from the first and the second model. Six algorithms were used to train and test the model; logistic regression, naïve-bayes, K-nearest neighbored (KNN), random forest, multilayer perception (MLP), and decision tree. The finding on this research indicated that 70% of accuracy identifies those at-risk students as early as by the end of the first semester. The other finding of this research also indicated that, among the 38 attributes available, the unweight high school GPA (HSGPA), SAT/ACT (record of pre-college academic preparation such as SAT/ACT verbal and math scores, courses taken in high schools and GPA), the first year GPA in college, and EFC (financial data sets such as the expected family contribution) are identified as the top four informative ones for predicting students attrition.

1.2 Statement of the problem

Nowadays student attrition has become one of the most challenging problems for academic institutions [10]. Early identification of vulnerable students who are prone to drop their courses is crucial for the success of any retention strategy. This would allow educational institutions to undertake timely and proactive measures [10] [11]. To improve student retention one should understand the non-trivial reason behind the attrition and to be successful, one should accurately identify those that are at risk of dropping out [12]. The monitoring and support of the first year student are considered as a very important step at many educational institutions. Yearly student enrollment at some of the facilities can be lower than the desired enrollment when coupled with higher dropout rate students it needs an effective approach for predicting student dropout as well as identifying the factor affecting student attrition. One identified at-risk students can be then targeted with academic and administrative support to increase their chance of staying in the

program [13] [10]. Mostly the loss of students usually results in overall financial loss, lower graduation rates, and inferior school reputation in the eyes of stakeholders [10].

According to Gouws et.al [8], Student attrition has a radiating effect that could reach almost everywhere from the individual to the family, then to the community at large. Depending on the study conducted in South Africa indicated that 35.0% to 40.0% of student attrition was registered in the various tertiary institutions of the country which is much higher than the internationally acceptable rate of 10.0%

Student attrition remains to be a problem even in the developed nations. According to a study conducted by Alan [9] the attrition rates for Australian Universities in 2006, stood at 10.5%. In his study of 485,983 students selected from 32 Australian universities in 2006

According to Kassahun [53], the average student attrition rate of SMU from 1991EC to 1996EC is 23.09% and it is increasing from year to year. On the other hand, Getahun pinpointed [1]; the attrition rate of SMU is expected to be 30 to 40%.

Depending on the information obtained from the registrar office of SMU the aggregate attrition rate from 1999EC to 2008EC indicated that student's attrition in extension division (51.665%) and regular division (34%). The problem of student attrition is not only strange for St. Mary's University, but it is also common in other HEIs (higher educational institution). Therefore higher educational institutions should work hard on the major contributing factor that leads to attrition of students in higher education. To reduce attrition and promote retention, many studies have been carried out in the past by many institutions.

Based on the study of Vinayak [3], health problem, financial instability, unable to cope with an advanced subject, university examination policy, lack of attendance, lack of motivation, and policy matter of the institution are the major contributing factors that lead to attrition of students in higher education.

Getahun [1] also explored the possibility of applying a data mining technique for predicting the likelihood of students to drop out, with the intention of developing possible retention strategy and decreasing the number of dropout students.

Muluken [41] investigated the potential applicability of data mining methodology to predict student performance as success and failure cases on Debre Markos university students' database.

Alemu et.al [61] focused on predicting the performance of student at an early stage of the degree program, to help the university not only to focus more on bright students but also to initially identify students with low academic achievement and find ways to support them.

Mehlet et.al [69] developed a predictive model, which determines the number of higher education students' enrolment at the department level ahead of time using data mining approaches.

As per the literature review, in this study, we used some unique attributes that, the other researcher didn't consider it. These includes: preparatory attended region, financial sources, preparatory completion year, background of study, Year taken, types of school attended, status, and the gap (the difference between preparatory completion year and university registration year).

To this end, this study explores and answers the following research questions.

- Which set of data attributes can be collectively used to identify at risk students of attrition?
- Which DM algorithms best used to develop a model for predicting student status (attrition or retention)?
- To what extent the predictive model determine the risk status of students?

1.3Objective of the study

The general and specific objectives of this study are listed as follows

1.3.1 General objective

The general objective of this study is to develop a predictive model using data mining classification techniques so as to determine undergraduate students' status (attrition or retention) in higher education.

1.3.2 Specific objectives

To achieve the general objective of the study the following specific objective are followed.

- To review related works so as to identify suitable data mining Techniques and algorithms.
- To identify attributes that are associated with student attrition and retention using data mining technique.
- To prepare quality dataset for experimentation and constructing optimal model.
- To build a model for early prediction of student attrition and retention (graduated) using the student enrollment data.
- To develop a prototype that simplifies the use of extracted hidden knowledge in the prediction of student attrition and retention (graduated) status.
- To evaluate the performance of the model in determining the student attrition and retention(graduated)

1.4 Scope and Limitation of the Study

This research focused on identifying the prominent factors that are related to student attrition and retention (graduated) in St. Mary's University. Preparatory attended region, Types of school attended, Sex, Financial sources, Batch, 12^{th} scored result, Admission classification, Preparatory completion year, Field of study, Age, Background of study, Year taken, Employment information, gap(the difference between preparatory completion year and university registration year) and students status(attrition or retention (graduated)), were studied briefly to build the predictive model. These factors were selected based on understanding of the problem. In this research we used the data that covers from 2005 EC up to 2010 EC. Besides, a classification technique is used to construct predictive models for determining students' status.

In this research, we only used the data obtained from St. Mary's university student record management information system My-SQL database which doesn't include other private or public institutions data. Also, this study only focused on St. Mary's university undergraduate regular and extension students' status. The distance and the postgraduate program students' are not included in this study.

Due to unavailability of clear data found in the dataset, other demographic data such as the native language, marital status, and place of birth location (urban or rural), and health-related data are not included under this study.

1.5 significance of the study

The result of this study will have a great significant to St. Mary's university registrar staff administrative to early identify students status as attrition status that indicates the tendency to drop out or withdraw, suspension and dismissal so that corrective measures may be adopted early on and these students may complete their university course with success. Besides, it will also help students to early know about their academic status; hence, an educative measure can also be applied before abandoning their course.

Accordingly this study has also the following significances

- The finding of this study has important implication for the institutional stakeholders such as teachers, counselors, university curriculum designers, students, and other decision maker bodies.
- It minimizes human power such as professional teachers, cost(loss of revenue), time, insufficient utilization of resources such as classroom, laboratories, and budget
- Attempt to Develop a prototype using of C# programming language using the output generated from the highest scored algorithm
- It helps earlier in identifying the attrition status of students' and students who need special attention and allow the teacher to provide appropriate advising and counseling.
- It is used to identify students' demographic characteristics and pre-university experience before the university admits them
- To establish effective psychological and psychiatric services which monitor the symptoms of student discontinuation and resolving the psychological and academic problems that lead to attrition of students.

1.6Research Methodology

In an academic context, research is used to refer to the activity of a diligent and systematic inquiry or investigation in an area, with the objective of discovering or revising facts, and theories of applications. The goal is to discover and disseminate new knowledge [17]

1.6.1 Research design

This study follows an experimental research approach. This is because experiments that will occur to extract results from real world implementations. And it is important to restate that all the experiments and results should be reproducible [17]. Depending on the comparison made on the data mining process model in section 2.3, the hybrid data mining process model is selected. It was developed by adopting the CRISP-DM model to academic research [15]. The main feature hybrid process model include [15],

- It provides more general, research-oriented description in six step.(see fique 2.4) these include, Understanding of the problem, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of knowledge the discovered knowledge.

- It introduces several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms)
- Modification of the last step, since in the hybrid model, the knowledge discovered a particular domain may be applied in the other domain.
- It introduces a data mining step instead of the modeling step

1.6.2Understanding of the problem

This is the initial step which involves working closely with domain experts to define the problem domain. To assess the current students status (attrition and retention (graduated)) and understand the problem in St. Mary's University detailed and depth consultation (interview, direct observation, and discussion) has been carried out with St. Mary's university registrar staff administrative, facility deans, department heads, university instructors', and students in both program (regular and extension). In addition to, the discussion with the stakeholders and the experts, also detailed and extensive literature review has been carried out on the existing study area, such as documents from St. Mary's university editorials like the instructors, and graduate students handbooks (contains policies, rules and regulation of the university), bulletin of students statistics(contains students related statistics) textbooks, peer reviewed journal, and additional authenticated sources also examined to get a better insight.

1.6 .3Understanding the data

In this study, the initial raw data is extracted from St. Mary's University student record management information system My-SQL database. We used Microsoft Excel 2010 to visualize the extracted data from SRMIS (student record management information system) to check for missing values, data redundancy, data integrity, and data completeness. We also used Weka version 3.6 for visualizing and describing the attributes used for model building. This research has a total of 7361 instances and 15 attributes which is named as Preparatory attended region, Types of school attended, Sex, Financial sources, Batch, 12^{th} scored result, Admission classification, Preparatory completion year, Field of study, Age, Background of study, Year taken, Employment information, and gap (the difference between preparatory completion year and university registration year), And also it has two classes which are named as attrition and graduated.

1.6 .4Preparation of the data

In this study, the data extracted from SRMIS (student record management information system) is prepossessed and cleaned to produce a final dataset used for building a predictive model. The data extracted from SRMIS contains missing values and contains errors. In this study since we used nominal, attributes the missing values are filed with mode values. We used Microsoft Excel 2010 to handle errors and outlier by filtering mechanism. In the last step of the data preparation, the data is exported to CSV (comma delimited) files to be used in the model building process

1.6. 5Data mining for constructing predictive model

In this study, weka version, 3.6 DM software is used. It is a knowledge discovery system developed by the University of Waikato in New Zealand that implements data mining algorithms. It implements algorithms for data preprocessing, classification, clustering, and association rules. It also integrates a feature selection and visualization algorithm [73]. For this study, we used the J48 decision tree; PART and JRIP rule induction, and Bayesian classifier (naïve bayes) algorithms for constructing the model. These algorithms are selected because all they are usually used and understandable rule can be extracted.

1.6.6 Evaluation of the discovered knowledge

The evaluation phase in the hybrid model includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts

and checking the impact of the discovered knowledge [15]. In this study accuracy of the model, confusion matrix, TP rate, FP rate, precision, recall, and ROC Area are used to evaluate the performance of the discovered knowledge.

1.6.7Use of discovered knowledge

It is the final step which consists of planning where and how to use the discovered knowledge. In this study, the created model using data mining technique for the purpose of predicting the student status (attrition and retention (graduated)) is going to be disseminated to the St. Mary's University in hard copy and soft copy documents to improve the capacity of the policymakers and regulatory bodies.

In this study, we also developed a simple prototype using C# programming language. To use the hidden patterns extracted using data mining technique, a user interface is designed with the help of c# programming languages. This programming language is selected since C# is a general-purpose, type-safe, object object-oriented language [19]. The goal of the language is programmer productivity. To this end, the language balances simplicity, expressiveness, and performance. The C# language is platform-neutral, but it was written to work well with the Microsoft.net framework.

1.7Thesis organization

This thesis report is organized into five chapters. The first chapter deals with the introduction that includes the background of study, statement of a problem, research objective, scope and limitation the study, significant of the study and research design of the study.

The second chapter reviews about data mining as a technology and its applicability in real-world data, an overview of educational data mining and reviewing related works concerning student attrition and retention. The third chapter deals with the tasks related to data pre-processing, which include an understanding of the problem, understanding of the data, and preparation of the data. Chapter four deals with data mining algorithms such as the decision three (J48), rule induction (PART and JRIP) and probabilistic classifier (naïve Bayes) algorithms are going to be dealt to build the predictive model and converting the evaluation of the result into real-world application. The final chapter presents concluding remarks and recommendations of the study

CHAPTER TWO

Literature review

Deep investigation has been carried out from different literature such as books, journal articles, and the internet, to have detailed knowledge about the problem domain, and work is done in related area, a detailed conceptual understanding of data mining, data mining process model, the data mining tasks, overview of educational data mining and the tasks are presented first, this is followed by a review of different related works concerning student attrition.

2.1 Overview of data mining

According to Jiawei et al [20], the computerization of our society has substantially enhanced our capability for both generating and collecting data from diverse sources. A tremendous amount of data has flooded almost from every aspect of our lives; the explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge. This has led to the generation of a promising and flourishing frontier in computer science called data mining and its various applications [20]. Data mining, also called as knowledge discovery from data (KDD), is the automation of extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, the web, other massive information repositories, or data streams [20].

David et.al [21] noted that Data mining is the analysis of (often large) observational dataset to find the unsuspected relationship and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns. Examples include linear equation, rules, cluster, graphs, tree structure, and recurrent pattern in time series. Data mining is the study of collecting, cleaning, processing, analyzing, and gaining useful insight from data. A wide variation exists in terms of the problem domains, applications, formulation, and data representations that are encountered in real applications; therefore, “data mining” is a broad umbrella term that is used to describe these different aspects of data processing [22]. It is also,

the process of discovering insightful, interesting, and novel patterns as well as descriptive, understandable and predictive models from large-scale data [23].

Data mining is defined as the process of discovering patterns in data, the process must be automatic or (more usefully) semiautomatic, the patterns discovered must be meaningful and the data is invariably present in substantial quantities. It is a practical topic and involves learning in a practical, not a theoretical, sense [24].

2.1.1 Overview of Educational data mining

Educational data mining (EDM) relates to the interdisciplinary research that deals with the development of various methods and techniques to explore the data generated from the different educational source. With the advent of technology increasing data, volumes have been generated due to the instrumentation of educational software and e-administration by the state education authorities. Also with the use of internet in learning; new methods like e-learning and web-based education have been introduced, they almost perpetually generating data about the learning and teaching behaviors of student and educator. EDM works in the direction to better seek these repositories to develop an understanding of the underlying process, in order to make use of the data in combination with different data mining and association technique to help optimize the educational practice benefiting the end user [43].

Educational data mining is new growing research for knowledge discovery from a large amount of educational data. Across the world, different countries researchers are putting in their efforts in finding out patterns and factors that can be helpful in the betterment of education. Data mining can be used in educational filed to enhance our understanding of the learning process to focus on identifying extracting and evaluating variables related to the learning process of students. The essences of data mining concepts are used in the educational fields for the purpose of extracting useful information on student behavior in the learning process [44].

2.2 Data mining process model

A process model is the set of tasks to be performed to develop a particular element [39]. The goal of a process model is to make the process repeatable, manageable, and measurable. In this section four most popular data mining process models are discussed. These models are Knowledge Discovery Databases (KDD) process model, CRISP-DM process model, the

SEMMA process model, and the Hybrid process model. These four models are mostly practiced by the data mining and researchers.

2.2.1The knowledge discovery in database (KDD) process

KDD focuses on the overall process of knowledge discovery from data, including how the data are stored and accessed, how algorithms can be scaled to massive datasets and still run effectively, how results can be interpreted and visualized, and how the overall man-machine intersection can usefully be modeled and supported. KDD places a special emphasis on finding understandable pattern that can be interpreted as useful or interesting knowledge. It also emphasizes scaling and robustness properties of modeling algorithms for large noisy data sets [26]. As shown in figure 2.1the steps of KDD process are the following [27].

Creating target dataset: include selecting a dataset or focusing on a subset or focusing on a subset of variables of data samples on which discovery is to be performed.

Data preprocessing: includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes as well as deciding DBMS issues, such as data types, schema, and missing and unknown values.

Data reduction and projection: includes finding useful features to represent the data, depending on the goal of the tasks, and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representation to the data.

Choosing the function of data mining: includes decide the purpose of the model derived by the mining algorithms (e.g. summarization, classification, regression, and clustering).

Choosing the data mining algorithm(s): includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate (e.g. models for categorical data are different from models on vectors) and matching a particular data mining methods with overall criteria of the KDD process (e.g. the user may be more interested in understanding the model than in its predictive capability)

Data mining: includes searching for patterns of interest in a particular representation form or a set of such representations, including classification rues of trees, regression, clustering, and sequence of modeling, dependency, and line analysis.

Interpretation: includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users.

Using discovered knowledge: includes incorporating this knowledge into performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted knowledge).

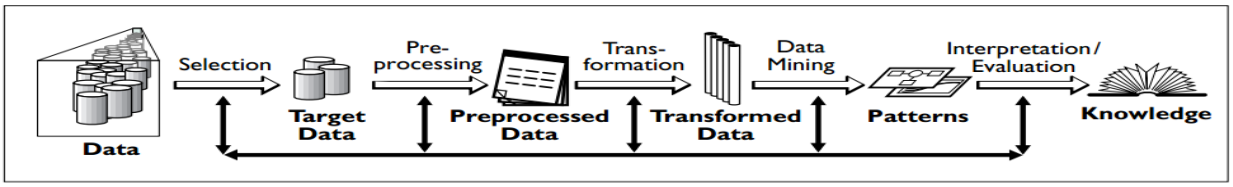


Fig 2.1 the steps constituting the KDD process [27]

2.2.2 The cross-industry standard process for data mining (CRISP-DM)

CRISP-DM is a detailed and widely used data mining methodology that aims to provide explicit guidance regarding how the various phase of a data mining project could be executed [18].

As shown in figure 2.2, CRISP-DM organizes the data mining process into six phases [26]; business understanding, data understanding, data preparation, modeling, evaluation, and deployment. These phases help organizations to understand the data mining process and provide a road-map to follow while planning and carry out a data mining project.

Business understanding: perhaps the most important phase of any data mining project, the initial business understanding phase focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives, in order to understand which data should later be analyzed, and how, it is vital for data mining practitioners fully understand the business for which they are finding a solution, this phase involves several key steps,

including determining business objective, determining the data mining goals, and producing the project plan.

Data understanding: the data understanding phase starts with an initial data collection, the analyst then proceeds to increase familiarity with the data, to identify data quality problem, to discover initial insights into the data, or to detect increasing subsets to form hypotheses about hidden information, the data understanding phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality.

Data preparation: the data preparation phase covers all activity to construct the final dataset or the data that will be fed into the modeling tools(s) from the initial raw data, the tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools, the five steps in data preparation are the selection of data, the cleaning of data, the construction of data, the integration of data, and the formatting of data.

Modeling: in this phase, various techniques are selected and applied and their parameters are calibrated to optimal values, typically, several techniques exist for the same data mining problem of data, therefore, stepping back to the data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models.

Evaluation: before proceeding to final deployment of the model built by the data analyst, it is important to more thoroughly evaluate the model and review the models construction to be certain it properly achieves the business objective, here it is critical to determine if some important business issue has not been sufficiently considered, the key steps here are the evaluation of results, the process review, and the data determination of next steps.

Deployment: the model creation is generally not the end of the project, the knowledge gained must be organized and presented in a way that the customer can use it, this often involves applying “live” models within an organization’s decision-making processes, such as the real-time personalization of web page or repeated scoring of marketing databases, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Fig 2.2 phases of the CRISP-DM reference model [26]

2.2.3THE SEMMA PROCESS MODEL

The SEMMA process was developed by SAS institute. The Acronyms SEMMA stands for Sample, explore, Modify, Model, Assess, and refers to the process of conducting a data mining project. As shown in figure 2.3, The SAS institute considers a cycle with 5 stages for the process [40];

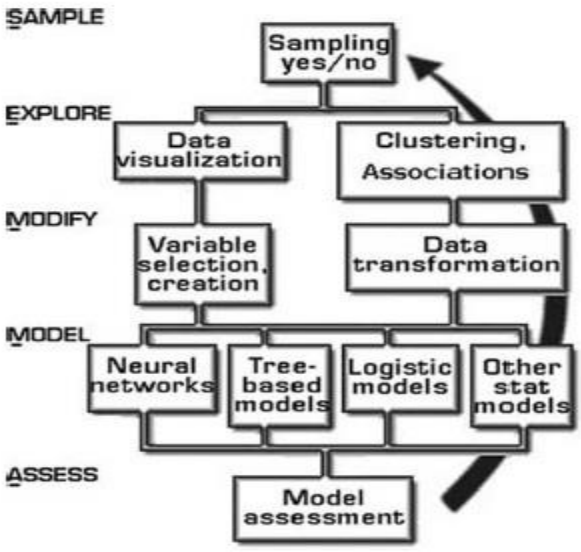


Fig 2.3 the SEMMA Analysis Cycle [30]

Sample: this stage consists on sampling the data by extracting a portion of a large dataset big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out as being optional.

Explore: this stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

Modify: this stages consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

Model: this stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predict a desired outcome.

Assess: this stage consists on assessing the data by evaluating usefulness and reliability of the findings from the data mining process and estimate how well it performs, SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects.

2.2.4 The hybrid DM process model

The development of academic model like KDD and industrial model like CRISP-DM has led to the development of hybrid models. It was developed based on the CRISP-DM model by adopting it to academic research [15]. As presented in fig 2.4, there are six- steps of the hybrid process model [15].

Understanding of the problem domain: this initial step involves working closely with domain experts to define the problem domain, identifying key people, and learning about current solutions to the problem, it also involves learning domain-specific terminology, a description of the problem, including its restrictions, is prepared, Finally, research goals is translated in to DM goals, and the initial of DM tools to be used later in the process is performed.

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

Preparation of the data: this steps concern deciding which data will be used as input for DM methods in the subsequent step, it involves sampling, running, correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc., the cleaned data may be further processed by feature selection and extraction algorithms(to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data, the end results are data that meet the specific input requirements for the DM tools selected in step 1.

Data mining: here the data miner uses various DM methods to derive knowledge from preprocessed data. These methods includes classification, clustering and association rule discovery.

Evaluation of the discovered knowledge: evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge, only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results, a list of errors made in the process is prepared.

Use of the discovered knowledge: the final step consists of planning where and how to uses the discovered knowledge, the application area in the current domain may be extended to other domains, a plan to monitor the final implementation of the discovered knowledge is created and the entire study documented, Finally, the discovered knowledge is deployed.

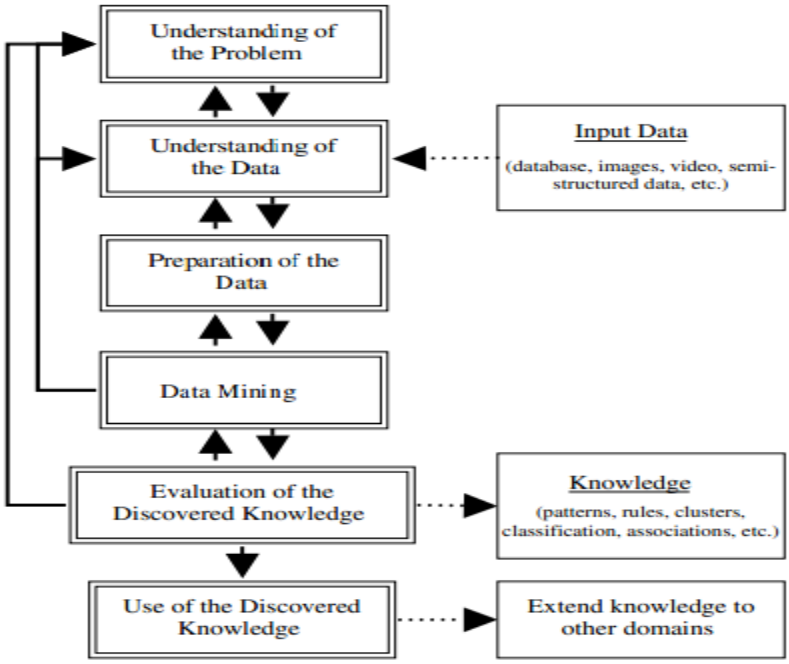


Fig 2.4 the hybrid process model [15]

2.3 Comparisons of the data mining process models

The comparison of the four data mining process model has done with the steps they have and their applicability to the academic researcher. The first comparison was carried out between KDD and SEMMA based on the step criteria they have [40].

- Sample can be identified with Selection
- Explore can be identified with Preprocessing
- Modify can be identified with Transformation,
- Model can be identified with Data Mining
- Assess can be identified with Interpretation/Evaluation

Examining it thoroughly, the five stages of the SEMMA process can be seen as a practical implementation of the five stages of the KDD process since it is directly linked to the SAS Enterprise Miner software [40].

The second comparison was made was carried out between KDD and CRISP-DM. Comparing the KDD stages with the CRISP-DM stages is not as straightforward as in the SEMMA situation [40]

- The Business Understanding phase can be identified with the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user
- The Deployment phase can be identified with the consolidation by incorporating this knowledge into the system.
- The Data Understanding phase can be identified as the combination of Selection and Preprocessing
- The Data Preparation phase can be identified with Transformation
- The Modeling phase can be identified with Data Mining
- The Evaluation phase can be identified with Interpretation/Evaluation

Based on the study of Umairet et.al [70], considering the presented analysis on CRISP-DM, KDD, and SEMMA. KDD is more appropriate for researchers and data mining experts because it is more complete and accurate. In contrast, CRISP-DM and SEMMA or mostly company oriented especially SEMMA that is used by SAS enterprise miner and integrate with their software. However, CRISP-DM is more complete as compared to SEMMA.

The third comparison was made between the CRISP-DM and Hybrid-DM. The main feature includes [15]

- Hybrid-DM Provides a more general, research-oriented description of the steps
- Hybrid-DM has a more detailed explicit feedback mechanism, whereas CRISP-DM has only three major feedback mechanisms.
- Modification of the last step, since in the hybrid model, the knowledge discovered a particular domain may be applied in the other domain.
- It introduces a data mining step instead of the modeling step

Considering the CRISP and Hybrid-DM process modeling, the one which suits for this research is Hybrid-DM modeling because it was developed based on the CRISP-DM model by adopting it

to academic research and using the discovered knowledge into other domains [15]. The summary of the four data mining process model described above depicted in Table 2.1 below

KDD	SEMMA	CRISP-DM	Hybrid
Pre KDD	----------	Business Understanding	Understanding of the problem
Selection	Sample	Data Understanding	Understanding of the data
Preprocessing	Explore		
Transformation	Modify	Data Preparation	Preparation of the data
Data mining	Model	Modeling	Data Mining
Interpretation/Evaluation	Assessment	Evaluation	Evaluation of the discovered knowledge
Post KDD	----------	Deployment	Use of the Discovered Knowledge

Table 2.1 Comparison data mining KDD, SEMMA, CRISP-DM and Hybrid-DM [40, 87 29, and 15]

2.4 Data mining tasks

It is convenient to categorize data mining into two types of tasks Such as predictive and descriptive modeling [21]. A predictive model works by making a prediction about values of data, which uses known results found from different datasets. The tasks of the Predictive data mining model includes classification, prediction, regression, and analysis of time series, whereas, descriptive model mostly identifies patterns or relationships in datasets. It serves as an easy way to explore the properties of the data examined earlier and not to predict new properties. The task of descriptive data mining model includes clustering, summarization, association rules, and sequence discovery [33].Fig 2.5 depicts data mining tasks and technology

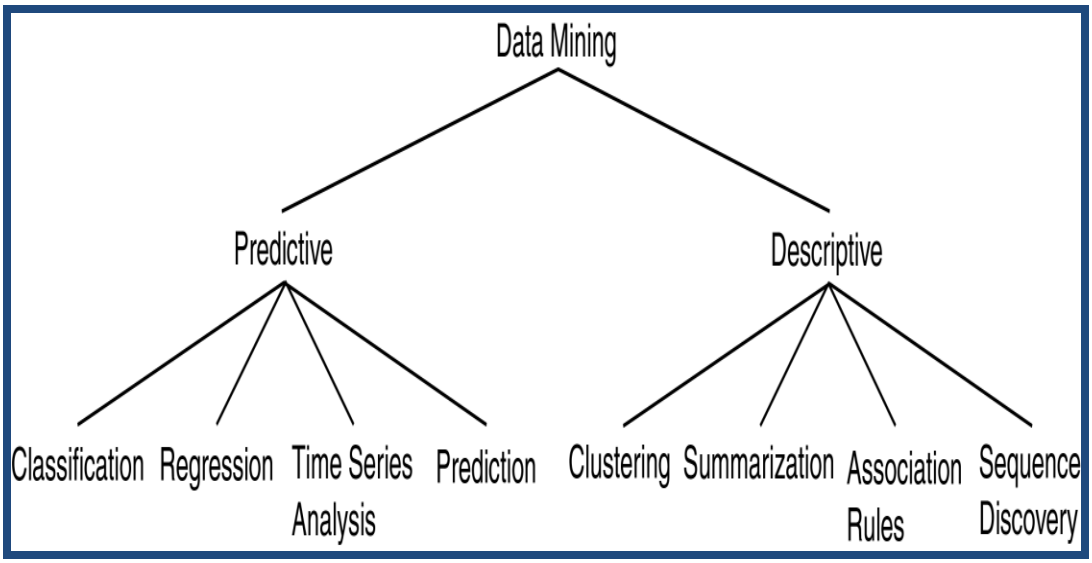


Fig 2.5 data mining main tasks with techniques [34]

According to Jiawei et al [20], the aim of predictive is to build a model that will permit the value of new instances to be predicted from the known values of other instances. In classification, the variable being predicted is categorical, while in regression the variable is quantitative. It can also be thought of as learning a mapping from an input set of vector measurement X to a scalar output y. whereas, the goal of the descriptive model is to describe all the data (or the process generating the data) [20]. Example of such description includes a model for the overall probability distribution of the data (density estimation), partitioning of the P- dimension space into groups (clustering analysis and segmentation), and models describing the relationship between variables (dependency modeling).

A descriptive model presents, in a convenient form, the main feature of the data. It is essentially a summary of the data, permitting us to study the most important aspects of the data without their being obscured by the sheer size of the dataset. In contrast, a predictive model has the specific objective of all allowing us to predict the value of some target characteristics of an object on the basis of observed values of other characteristics of the object. In this research, our aim is constructing a predictive model using classification algorithms.

2.5 Classification algorithms

According to Pooja [35], Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of the predefined set of class or group. Classification technique creates a model that can learn how to classify data items into groups.

2.5.1Decision tree algorithms

According to Joel [36], a decision tree uses a tree structure to represent a number of possible decision paths and an outcome for each path see figure 2.6. A decision tree is very easy to understand and interpret, and the process by which they reach a prediction is completely transparent. It can also easily handle a mix of numeric and categorical attributes and can even classify data for which attributes are missing. At the same time, finding an "optimal" decision tree for a set of training data is a computationally very hard problem. More important, it is very easy (and very bad) to build decision trees that are overfitted to the training data, and that doesn't generalize well to unseen data [36].

Entropy

In order to build a decision tree, we need to select the best attribute that reduces the disorder observed in the dataset. To this end, the first step is computing the information content or the disorder level with a given attribute under consideration for classification. This is followed by determining the information gain of each attribute. Entropy is used to represent the uncertainty associated with the attribute in corresponding data into the given class [36].

Given a set of S data, each member of which is labeled as belonging to one of a finite number of class C_1,C_2 …Cn. If all the data points belong to a single class, then there is no real uncertainty which means there is no or low entropy. If the data points are evenly spread across the class, there is a lot of uncertainty and there is high entropy.

Entropy D (n_+, n_-) can be calculated using equation 2.1[60]. (i.e., a set S containing a total of n examples of which n_+ are positive and n_- are negative)

$$D(n_+, n_-) = -\frac{n_+}{n}\log_2\frac{n_+}{n} - \frac{n_-}{n}\log_2\frac{n_-}{n}$$ …….equation (2.1)

Information Gain measures the unexpected reduction in entropy due to splitting on an attributes. It also measures reduction in entropy achieved because of the split.

Information gain is computed using equation 2.2 [60]. (i.e., Parent Node, S is split into k partitions; n_i is number of records in partition i)

$$GAIN_{split} = Entropy(S) - \left(\sum_{i=1}^{k}\frac{n_i}{n}Entropy(i)\right)$$-----equation (2.2)

An attribute with the highest information gain is selected to split the data at each node of the decision tree

Rule extraction from a decision tree

To extract rules from a decision tree, one rule is created for each path from the root to a leaf node. Each splitting criterion along a given path is logically ANDed to from the rules antecedent ("IF" part). The leaf node holds the class prediction, forming the consequent ("THEN" part).

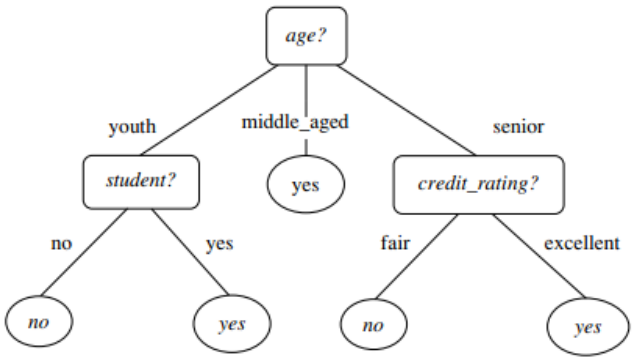


Fig 2.6 Extracting classification rules from a decision tree [20]

The following rules can be extracted from figure 2.6 following each path from the given decision

R1: IF age = youth AND student = no THEN buy computer = no

R2: IF age = youth AND student = yes THEN buy computer = yes

R3: IF age = middle aged THEN buy computer = yes

R4: IF age = senior AND credit rating = excellent THEN buy computer = yes

R5: IF age = senior AND credit rating = fair THEN buy computer = no

2.5.1.1 J48 decision tree algorithm

J48 is a decision tree algorithm which is frequently used in most application. The extra highlights of J48 are representing missing values, decision trees pruning, consistent characteristic esteem ranges, deduction of standards, and so forth [37]. In the Weka data mining apparatus, J48 is an open source java execution of the C4.5 algorithm. The Weka apparatus gives various alternatives related to tree pruning. If there is an occurrence of potential overfitting, pruning can be utilized as a device for précising. In different algorithms, the characterization is performed recursively till each and every leaf is unadulterated, that is the grouping of the data ought to be as impeccable as could be allowed [37].

Pseudo code for C4.5 (J48) decision tree algorithm is presented as follows [38]

Create a root node N;

If T belongs to same category C;

leaf node=N;

mark N as class C;

return n;

For i=1 to n;

Calculate Information gain (Ai);

ta= testing attribute;

N.ta= attribute having highest information gain;

ifN.ta==continuous;

find threshold;

For Each T' in the splitting of T

ifT' is empty

child of N is a leaf node ;

else {child of N= dtree (T')

calculate classification error rate of node N;

return N;

2.5.1.2Pruning

Tree pruning is used to cut the overfitted tree back into smaller tree by removing sub-branches that are not contributing to the generalization of accuracy and can improve the generalization performance of a decision tree, especially in noisy domains [55]. Datasets may contain little subsets of instances that are not well defined to classify them correctly; in this case pruning can be used [58]. The pruning is performed for decreasing classification errors which are being produced by specialization in the training set [59].

2.5.2 Rule–Based classification

According to Jiawei et.al [20], rule – based classifier, where the learned model is represented as a set of IF- THEN rules. Rules are a good ways of representing information or bits of knowledge. A rule – based classifier uses a set of IF - THEN rules for classification. An IF-THEN rule is an expression of the form. IF condition THEN conclusion.

The: "IF" part (or left side) of a rule is known as the rule antecedent or precondition. The "THEN" part (or right side) is the rule consequent. In the rule antecedent, the condition consists of one or more attribute test (e.g., age = youth and student = yes).

If the condition (i.e., all the attribute tests) in rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied (or simply, that the rule is satisfied) and that the rules covers the tuple.

A rule R can be assessed by its coverage and accuracy. Given a tuple, X, from a class-labeled data set, D, let n_{covers} be the number of tuples covered by R; $n_{correct}$ be the number of tuples correctly classified by R; and |R| be the number of tuples in D. the coverage and accuracy of R can be defined as[20].

$$coverage(R) = \frac{n_{covers}}{|D|}$$
$$accuracy(R) = \frac{n_{correct}}{n_{covers}}.$$

That is, a rule's coverage is percentage of tuples that are covered by the rule (i.e., their attribute values hold true for the rule's antecedent). For a rule's accuracy, we look at the tuples that it covers and see what percentage of them the rule can correctly classify. That is, a rule's coverage is the percentage of tuples that are covered by the rule (i.e., their attribute values hold true for the rule's antecedent).

If a rule is satisfied by X, the rule is said to be triggered. For example, suppose we have X = (age = youth, income = medium, student = yes, credit rating = fair). So we would like to classify X according to buys computer. X satisfies R1, which triggers the rules [20].

- If R1 is the only rule satisfy, then the rule fires by returning the class of prediction for X.
- If more than one rule is triggered, we need a conflict resolution strategy to figure out which rule gets to fire and assign its class prediction to X.

The size ordering scheme assigns the highest priority to the triggering rule that has the "toughest" requirements, where toughness is measured by the rules antecedent size. The rule ordering scheme prioritizes the rules beforehand. The ordering may be class-based or rule- based [20]. With class based ordering the classes are sorted in order of decreasing "importance" such as by decreasing order of prevalence. with rule based ordering, the rules are organized into one long priority list, according to some measure of rules quality, such as accuracy, coverage, or size(number of attributes tests in the rules antecedent), or based on advice from domain experts.

IF- THEN rules can be extracted directly from the training data (i.e., without having to generate a decision tree first) using a sequential covering algorithm. There are many sequential covering algorithms [20]. Popular variations include AQ, CN2, and the more recent RIPPER. The general strategy is as follows. Rules are learned one at a time. Each time a rules is learned, the tuples covered by the rule are removed, and the process repeats on the remaining tuples. This sequential learning of rules is in contrast to decision tree induction. Because the path to each leaf I a decision tree corresponds to a rule, we can consider decision tree induction as learning a set of rules simultaneously [20].

Given: POS (positive), NEG (negative), K (number of attributes), n (number of examples)

G_{POS} = Data Reduction (POS, k);

G^{NEG} = Data Reduction (NEG, k);

Initialize RULES = []; i=1; // where $rules_i$ denotes i^{th} rule stored in RULES

Create LIST = list of all columns in G_{POS}

Within every G_{POS} column that is on LIST, for every non missing value

A from selected column j compute sum, s_{aj}, of values of $gpos_i[k+1]$

For every row i, in which a appears and multiply s_{aj}, by the number of

Values the attribute j has

Select maximal s_{aj}, remove j from LIST, add "j=a" selector to $rules_i$

If $rules_i$ does not describe any rows in G_{NEG}

Then remove all rows described by $rules_i$ from $G_{POS,}$ i = i+1;

If G_{POS} is not empty go to 2.2, else terminate

else go to

Result: RULES describing POS

Data Reduction (D, k) // data reduction procedure for D=POS or D=NEG

Initialize G = []; i=1; tmp=d_1; g_1 =d1; $g_1[k+1] = 1$;

For j=1 to N_D // for positive/negative data; ND is NPOS or NNEG

For kk = 1 to k //for all attributes

If (d_j [kk]=/= tmp[kk] or d_j[kk] =‘∗’)

Then tmp[kk] = ‘*’; // ‘*’ denotes missing “ do not care”

if (number of non-missing values in tmp ≥2)

Then g_i =tmp; $g_i[k+1]$ ++;

else i++; g_i = dj; $g_i[k+1]$ =1; tmp=dj;

Return G;

Rule induction algorithm [15]

In this research JRIP and PART rule induction algorithm are experimented to build the predictive model

JRIP is one of the basic and most popular algorithms. classes are examined in growing size and an initial set of rules for the class is generated using incremental reduced error JRIP (RIPPER) proceeds by training data as a class, and finding a set of rules that cover next class and does the same, repeating this until all classes have covered [42]

PART is a separate and conquer rule learner. The algorithm producing sets of rules called “decision lists” which are planned set of rules. A new data is compared the class of the first matching rules. PART builds a partial C4.5 decision tree in each; iteration and makes the “best” leaf into a rule [42]

2.5.3 Probabilistic classifier

Probabilistic classifier constructs a model that quantifies the relationship between the feature variables and the target (class) variable as a probability [22]. There are many ways in which such modeling can be performed. One of the most popular models is bayes classifier [22].

2.5.3.1 Naïve bayes classifier

The bayes model is referred to as "naïve" because of the assumption of conditional independence [22]. The naive bayes classifier seems to perform quite well in practice in many domains and although it is possible to implement the bayes model using more general multivariate estimation methods; such methods can be computationally more expensive. Furthermore, the estimation of multivariate probabilities becomes inaccurate with increasing dimensionality, especially with limited training data [22].

The byes classifier is based on the bayes theorem for conditional probabilities. This theorem quantifies the conditional probability of a random variable (class variable), given known observations about the value of another set of random variables (feature variables)[22].

The Naive bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given dataset [64].The algorithm uses bayes theorem and assumes all attributes to be independent given the value of the class variable [64]

In principle, methods based on the class-conditional distributions in which the variables are all categorical are straightforward: we simply estimate the probabilities that an object from each class will fall in each cell of the discrete variables (each possible discrete value of the vector variable X), and then we use bayes theorem to produce a classification [21].

Bayes theorem

According to Mirian et.al [65], Bayesian classifier obtains the posteriori probability of each class, Ci, using bayes classifier (NBC) makes the simplifying assumption that the attributes, A, are independent given the class. So the likelihood can be obtained by the product of the individual conditional probability, (P ($C_i|A_1$… An)), is given by:

P ($C_i|A_1 \ldots A_n$) = P (C_i) P ($A_1|C_i$)….P ($A_n|C_i$), (i.e., P (posterior probability), Ai…….An (given records), Ci (the target class is to be predicted)) [60]

The naïve bayes classifier is an efficient classification model that is easy to learn and has a high accuracy in many domains [65]. However, it has two main draw backs.

- Its classification accuracy decrease when the attributes are not dependent
- It can't deal with nonparametric continuous attributes.

Input:

Training dataset T,

F= ($f_1, f_2, f_3, f_3 \ldots\ldots\ldots f_n$) // value of the predictor variable

in testing dataset.

Output:

A class of testing dataset

Read the training dataset T;

Calculate the mean and standard deviation of the

Predictor variables in each class;

Repeat

Calculate the probability of fi using the gauss

Density equation in each class;

Until the probability of all predictor variables ($f_1,f_2,f_3,f_3 \ldots\ldots\ldots f_n$) has been calculated

Calculate the likelihood for each class;

Get the greatest likelihood;

Pseudo code of naïve bayes algorithm [71]

2.6Techniques used to improve classification accuracy class-imbalanced data

Classification problems often suffer from data imbalance across classes. This is the case when the size of examples from one class is significantly higher or lower relative to the other classes [72]

2.6.1 Resampling Methods

According to Jiawei [20] both over-sampling and under-sampling change the training data distribution so that the rare (positive) class is well represented. Over-sampling works by resampling the positive tuples so that the resulting training set contains an equal number of positive and negative tuples. Under-sampling works by decreasing the number of negative tuples.it randomly eliminates tuples from the majority (negative) class until there are an equal number of positive and negative tuples

2.6.2 SMOTE (Synthetic Minority Over-sampling Technique)

The synthetic minority over-sampling technique (SMOTE) is an important approach by over-sampling the positive class or the minority class [66].The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k-minority class nearest neighbors [67]. A SMOTE force focused learning and introduces a bias towards the minority class [67].

Algorithm SMOTE (T, N, k) [68]

Input: Number of minority class samples T; Amount of SMOTEN%; Number of nearest neighbors' k

Output: (N/100)* T synthetic minority class samples

(∗ If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd. ∗)

if N<100

then Randomize the T minority class samples

T =(N/100)∗T

N = 100

endif

N =(int)(N/100)(∗ The amount of SMOTE is assumed to be in integral multiples of 100. ∗)

k = Number of nearest neighbors

numattrs = Number of attributes

sample [ ][ ]: array for original minority class samples

newindex: keeps a count of number of synthetic samples generated, initialized to 0

synthetic [ ][ ]: array for synthetic samples (∗ Compute k nearest neighbors for each minority class sample only. ∗)

for i ←1 to T

Compute k nearest neighbors for i, and save the indices in the nnarray

Populate (N, i, nnarray)

endfor

Populate (N, i, nnarray)(∗ Function to generate the synthetic samples. ∗)

While N=/= 0

Choose a random number between 1 and k, call itnn. This step chooses one of the k nearest neighbors of i.

for attr ←1 to numattrs

Compute: dif = Sample [nnarray[nn]][attr]−Sample[i][attr]

Compute: gap = random number between 0 and 1

Synthetic[newindex][attr]=Sample[i][attr]+gap∗dif

endfor

newindex++

N = N −1

endwhile

return (∗ End of Populate. ∗)

end algorithm

2.7 Model evaluation and model accuracy classifier

In this study accuracy, TP rate, FP rate, precision, recall, F-measure, the time taken and ROC are used to evaluate the performance of the discovered knowledge. 66% split and10-fold cross-validation test option are applied during experimentation.

2.7.1 Separate training and test sets

For the 'train and test' method the available data is split into two parts called a training set and a test set [83]. The training set is used to construct a classifier (decision tree, neural network etc.) and the classifier is used to predict the classification for the instances in the test set [83].

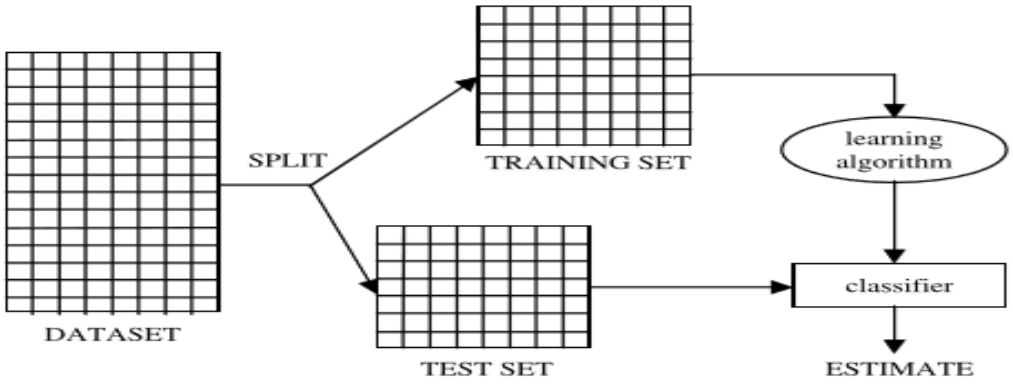


Fig 2.7 train and test [83]

2.7.2 k-fold Cross-validation

An alternative approach to 'train and test' that is often adopted when the number of instances is small is known as k-fold cross-validation [83]. It is applied if the dataset comprises N instances, these are divided into k equal parts, k typically being a small number such as 5 or 10 and each of the k parts, in turn, is used as a test set and the other k−1 parts are used as a training set [83]. In

this research, we used the default Weka 10 fold cross-validation as an alternative approach.

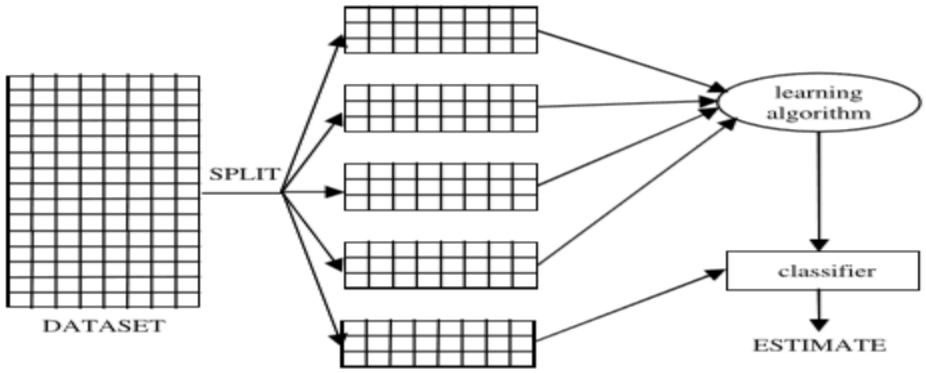


Fig 2.8 k-fold cross-validation [83]

2.7.3 Confusion matrix

The confusion matrix is a useful tool for analyzing how well the classifier can recognize tuples of different class [20].

True positive (TP):- these refer to the positive tuples that were correctly labeled by the classifier

True negative (TN):- these are the negative tuples that were correctly labeled by the classifier.

False positive (FP): these are the negative tuples that were incorrectly labeled by the classifier

False negative (FN): these are the positive tuple that were mislabeled as a negative

As a summery TP and TN tells us when the classifier is getting things right, while FP and FN tells us when the classifier is getting things wrong [20].

	Predicted class		
Actual class		Class =yes	Class = no
	Class =yes	a (TP)	B (FN)
	Class = no	c (FP)	D (TN)

Table 2.2 confusion matrix [20]

The rows correspond to the correct classifications and the columns correspond to the predicted classifications [83]. The value in the I^{th} row and j^{th} column gives the number of instances for which the correct classification is the I^{th} class which are classified as belonging to the j^{th} class [83].

Accuracy

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier [20]. Accuracy can be calculated using equation 2.3. That is,

Accuracy=TP+TN/TP+TN+FP+FN*10..……..equation 2.3

Precision

Precision can be thought of as a measure of exactness (i.e., what percentage of tuples labeled as positive are actually such) [20]. Precision can be calculated as using of equation 2.4. That is,

Precision = TP/TP+FP*100……..equation 2.4

Recall

Recall is a measure of completeness (what percentage of positive tuples are labelled as such)[20]. Recall can be calculated as using of equation 2.5. That is,

Recall = TP/TP+FN*100…….equation 2.5

F measure

An alternative way to use precision and recall is to combine them into a single measure this approach is called F measure [20].F measure can be calculated using of equation 2.7. That is,

F = 2 * precision * recall/precision + recall…….equation 2.7

2.8 Review concepts in the domain

2.8.1Education

Education can be considered as a form of learning in which the knowledge, skills, and habits of a group of people are transferred from one generation to the next through teaching, training, or

research and it is commonly divided into stages such as pre-school, primary school, secondary school, and higher learning or tertiary education [14].

Education can be used as a tool for societal development enriches individual understands of themselves and the world around them through constant improvement in their standard of living and the society at large [14]. Besides, education leads the individual to boost their productivity and creativity to promote entrepreneurship and technological advances [14].

Education nowadays has become a prominent thing as it involves most people to take part in this matter. It plays an important role in the development of a country [31]. The better the quality of education that a country has, the faster it is likely to develop. No matter what global problems that a country is facing, whether it's the elimination of poverty, the creation of peace, or environmental energy problems, the solutions will always include education [31].

2.8.2 Educational quality

Quality in education is the heart of any educational system [82]. It influences what the students learn, how well they learn and what benefit they draw from their education [82]. Quality in education is a multi-dimensional concept ,embracing all functions and activities of education ,including teaching and academic programs ,research and scholarship, staffing ,students ,building ,facilities ,equipment ,service to the community ,academic environment, taking into account national cultural values and circumstances and international dimensions such as exchange of knowledge ,interactive networking ,mobility of teachers and students and international research projects[82].

One possible path for improving the quality of education lies in the application of quality assurance in teaching and learning processes [81]. The main purposes of quality assurance include [81];

- To ensure that the education, research and consultancy provided by higher education institutions meet the development needs of the country, and where appropriate, the local community;
- To establish a monitoring system of accountability that audits the adequacy of educational provision quality;

- To secure the availability within each higher educational institution of relevant information which the institution uses actively and continuously to sustain and improve the quality of learning and teaching on programs it offers;
- To enhance quality of education by continuously improving the relevance and quality of teaching and learning; and
- To meet the expectations of students, employers and other stakeholder

2.8.3 Overview of student attrition and retention

Mesfine [51] noted that students attrition can be defined as a termination or withdrawal from an educational program run by a given academic institutions. Thus, dropping out from a program as well as delay from required time of completion is considered as attrition. Whereas retention occur when a students enroll each semester until graduation, studies full-time, and graduate in the specified time for the specific program. Attrition includes; academic dismissal, academic suspension, withdrawal and dropout [49].

2.8.4 General studies on students attrition

Asmerom et.al [74] examined the magnitude of students' dropout by faculties and subject. According to the study major cause of students' dropout can be categorized as academic and non-academic. However, "dropping out" from higher institutions is basically dependent on academic performance

Based on a study of Tsehaye et.al [75], poor preparation and commitment, mismatch area of interest and field of placement, poor social integration, and lack of appropriately developed instructional and assessment methods and Cumulative Grade Point Average (CGPA) can be regarded as determinants cause of student dropouts or persistence.

Abel[76], examine the models, which might be appropriate to represent the attrition rate of students at the Faculty of Business and Economics (FBE), Addis Abeba university to demonstrate the magnitude and pattern of academic survival rates and to compare the academic survival status of male and female students.

Fassil et.al [77] revealed that first-year students are more likely vulnerable to attrition than second and third-year students. So, to reduce the number of unfinished degrees and reduce

vulnerability to attrition, leaders of the higher institution should give due attention to students' program placement, tutorials for female students and provision of better student services.

Wudie et al [78] studied the major factors affecting the academic performance of female students' at Bahir Dar University. As the finding of this study indicated that university-related factors such as university academic and administrative rules and regulations, peer pressure, lack of female role model teachers, department choice of students and providing different supportive pieces of training and tutorial classes by the university impacts female students' academic performance.

Kwadwo et.al [79], presented the theoretical arguments made on attrition and retention in higher education. According to the study, Attrition is more pervasive for people in lower socio-economic groups.

Tino[80], developed the theory that suggests then that dropout is a multidimensional process which results from the interaction between the individual, and the institution characteristics.

Individual characteristics - it includes; the individual himself, the characteristics of his family, his educational expectancy before college entry, and his commitment to the goal of college completion.

Interaction within the college environment: - it includes; experience at the institution (academics and faculty and peer interactions), external commitments while at the institution, and integration both academically and socially.

2.9 Related works

In this study we explored different international and local related works to find out the most noticeable factors that associated with the problem of student attrition and retention.

2.9.1 Students Attrition related works

Dursun [12] explains that more students withdraw during the first year of college than during the rest of their higher education. For these study, the data of students was come from a single institution (a comprehensive public university located in the mid-west region on the United States) student database which had been used to develop the models that are capable of

predicting the problem of the student attrition. The average freshman student retention rate for the institution is about 80%, and the average 6-year graduation rate is about 60%. In the study, the researcher used 8 years of institutional data which entailed 22,224 students enrolled as freshmen from 1999 to 2006. In this work, CRISP-DM was used to provide a systematic and structured way of conducting data mining studies. The data contained variables related to student's academic, financial, and demographic characteristics. Three popular classification techniques were used namely, artificial neural networks, decision tree, and logistic regression. Based on the 10-fold cross-validation results, the artificial neural network model was able to classify students better with an overall accuracy rate of 81.91%. The finding of the study shows that, the most important predictor for student attrition are those related to the past and present educational success of the student. Accordingly, to improve the retention rates they recommend that institutions may choose to enroll more academically successful students.

The limitation of the study

- This study only considered the students dropping out at the end of the first year. That means, this study didn't consider 2nd year and above students attrition status.
- This study only considered the attrition status of students who left the institution voluntarily. That means, this study didn't include the attrition status of academically dismissal students

Vrushali et.al [47] focus on identifying slow, average and fast learners among students and displaying it by predictive data mining model using classification based algorithm. In this study also the data is collected from Sardar Patel Institute of technology college of MCA department India. The study was done on the dataset of student's academic performance and other attributes apart from academic performance were also considered. In this work, Weka version 3.6.13 is used to implement a data mining approach. The methodology of this research consists of four stages such as data collection, data preprocessing, data mining (classification) and interpretation. The Classification algorithms used for experimentation included naïve bayes, J48, zeroR, and random tree. Finally, it has been investigated that random tree technique performs best with an accuracy of 95.45%. In general, this study helps the institution to identify students who are slow learner which further provide a base for deciding special aid to them.

Carlos.et.al [48] also applied data mining techniques to predict school failure and dropout. In this study, the real data collected from 670 middle school students from Zacatecas Mexico. The aim of this study is predicting the academic failure of students using the process of knowledge discovery and data mining techniques. Similar to the previous studies school failure can be influenced by demographic, cultural, social, family, or educational background, socioeconomic status, psychological profile, and academic progress. In this study 10 data mining classification algorithms have been applied and from this 77 relevant attribute were selected. Three experiments have been applied in the first experiment using 10 classification algorithms and all available information (77 attributes). In the second experiment, only the best 15 attributes are used. The third experiment also repeated the execution by using re-balanced data files. In general, predicting student failure at school can be a difficult task not only because of its multifactor problem (in which there are a lot of personal, social, and economical that can be influential factors) but also because of the available data are normally imbalanced.

Getahun [1] explored the possibility of applying data mining techniques for predicting the likelihood of a student's dropout. He stated that Ethiopian higher education institutions are characterized by high dropout rates, but there is no system of early prediction of this problem and there is no timely corrective action being taken by the management of the institutions. In this work, the data of the students is extracted from the student information management system of the St. Mary's University College, who had got registered between the years 2004/05 GC up to 2010 GC. And the total number of degree students registered in the specified period was about 8,743 form five different departments. Form this, 2,613 were regular program while the remaining 6,130 were extension program. However many of the required attribute of the students were missing for batches 2004/05, 2005/06 and 2006/07. Therefore, most of the required attribute for building the model were generated form 2007/08. In the study, students demographic (age, sex, employment status, and income), pre-college experience (high school result, name of precious college, and previous college result) and university college experience (grade point of average or GPA of 10 terms and CGPA) were a predictor data type used for model building.

Two algorithms from decision tree (J48, and Random Forest) and one algorithm from Neural Network (Multilayer Perceptron) are selected to conduct this experiment. In terms of feature

selection algorithm, the strongest predictor of dropout was found to be CGPA, followed by GPA of Term1 and Term2 as well as Age and Previous result in college.

As a result, higher performance is demonstrated more by decision trees (J48 and random forest) than the neural network.

On the whole, the findings of the study indicate that students' dropout is more related to performance than other predictors in the study considered.

Limitation of the study

- In this study the experiment was conducted with a less rich dataset, the total dataset used in this experiment was only 1,362
- The researcher didn't use or clearly stated which data mining process model is used such as CRISP-DM, Hybrid, KDD and SEMMA
- The variable considered under this study include; students demography(age, sex, employment status, and income), pre-college experience (high school result and pre-college result), and university college experience but the researcher didn't include the financial support or sources, the type of school attended, preparatory attended region, preparatory completion year, gap(the difference between preparatory completion year and university registration year), and background of study as input variable
- The experiment was conducted by taking only graduates of 2007/08

Tariku [46] also applied data mining for identifying the determinant factors for the student success in the preparatory schools to join higher education. The study only focused on Addis Abeba region and natural science stream preparatory schools' students. In the study, EHEEE (Ethiopian Higher Education Entrance Examination) data and the corresponding, their EGSECE (Ethiopian General Secondary Education Certificate Examination) data are collected from national educational assessment and examination agency. The collected data cover from 2006 up to 2008 EHEEE. In the study, the common subject from grade 10 and grade 12 of a student which are English, mathematics, physics, chemistry, biology, and civics are selected. In addition to the subject; sub city, school type, sex are selected as an attributes. Finally, 40328 instances and 15 attributes are selected for analysis. In this work, the researcher used the hybrid data

mining process model, and also association rule mining data mining technique used to discover the rules. Apriori and filter associator algorithms are compared based on the execution time and database scan. Hence, the Apriori algorithm is selected and the analysis is undertaken. The discovered knowledge is categorized based on sex, school type, and subject attributes. From the analysis of the discovered knowledge, 40328 instances (74%) shows success and the remaining 26% shows not achieving success in entering higher education. In conclusion, scoring very good in physics, civics, biology, and scoring good in English in EHEEE, and scoring satisfactory in chemistry in EHEEE and scoring satisfactory in English in EGSECE are the determinant factors for the students' success in entering higher education.

2.9.2 Students' Academic performance related works

Muluken [41] investigated the potential applicability of data mining methodology to predict student performance as success and failure cases on Debre Markos university students' database. This research aims to prove how different factors affect a student success and failure rate in relation to other variables in students' dataset by applying data mining techniques. CRISP-DM (Cross Industry Standard Process for Data mining) is a data mining methodology to be used by the research. J48algorithm and naïve bayes algorithms are selected for classification purposes.

As 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 failure and success.

Alemu et.al [61] used data mining methodologies to design and develop a Data Mining model to predict the academic performance of students at the end of the first year degree program. The data used in this study is collected from five selected regional universities in Ethiopia, namely, Bahirdar University, Wollo University, Gondar University, D/Markos University, and Debre Berhan University. A model is built using C4.5 Decision tree learning algorithm and generates five classification rules set classifiers (predictors) in an experiment. The experiment using a test dataset produces 81.4% accuracy. The important factors identified under this study included PSGPA (preparatory school grade point average result), EUEE (Ethiopian university entrance examination result), FCI (field choice interest), FYFSA (first year first semester academic achievements) and FYSSA (first year second semester academic achievements).

Mehlet et.al [69] develop a forecasting system for higher education student enrolment at each program level or department level using data mining technology in the context of Ethiopia In this study the historical institutional data is selected from government and private sample universities such as Addis Ababa University, Mekele University, Bahirdar University, Jimma University, Harmaya University, Unity University and St. Mary University using purposive sampling technique. CRISP-DM (Cross Industry Standard Process for Data Mining) is used subsequently to conduct systematic data mining analysis. And Three predictive data mining modeling techniques, namely, Decision tree (J48 Classifier), Bayesian classifier (Naïve bayes) and Neural network (Multilayer Perception), used to address the problem. Based on this, the neural network and Multilayer perception achieved the highest accuracy as compared to other algorithms. And The Naïve bayes classifier has the lowest performance. depending on the rules generated by J48 classifier such as for one year ahead, for two years ahead and for three years ahead prediction is found for predicting higher education students' enrolment at the specified universities and departments.

Most of Ethiopian higher private or public universities are characterized by high attrition rate and low graduation rate. Despite this, they are also characterized by the poor file management system. The Ethiopian government is investing millions of dollars for expanding higher institutions all over the country to achieve the GTP plan. But the number of students who are joining higher institution is always higher than that of graduated students. Therefore, a poor student's file management system together with administrative issues has become a big obstacle to study the area well.

This research is different from the other works since it focused on student attrition and retention (graduation). Attrition includes dropout, academic dismissal, academic suspension, and withdrawal while retention occurs when a student enrolls each semester until graduation. So our aim is to predict student attrition or retention (graduation) using data mining technology, based on the data obtained from St. Mary's university student record management information system (SRMIS) database

CHAPTER THREE

Problem understanding and data preparation

3.1 understanding of the problem

3.1.1 Overview of private higher education in Ethiopia

The beginning of a full-fledged private higher education institution (PHEI) in Ethiopia dates back to 1998 which marks the year the then unity college (now University) was established. Since then, PHEIs have mushroomed in an unprecedented fashion. There are now around 46 public universities and 130 accredited non-government higher education institutions; four of which reach a full–fledged university status. Although their number is threefold of the public sectors, PHEI remains small size and account only for 15 percent of enrollment at a national level. Currently, higher education in Ethiopia includes education programs which are offered as an undergraduate degree for three, four or more years after completing secondary education and specialized degrees such as master's and PhD programs [50]. According to St. Mary's University graduate students handbook [49], St. Mary's University, among others offer conventional and distance education that is accessible to the large society through reasonable tuition focusing on quality and standard in teaching, research and outreach services**.** It has also a vision to become among the leading higher education centers of academic excellence in teaching-learning, research, publications and community service in EAST-Africa and contribute to the development of Ethiopia.

Overview of St. Mary's University

St. Mary's University, among others, provides higher education and undertakes research to promote the advancement and dissemination of knowledge and its application to the needs and aspiration of the people of Ethiopia. It blends the academic world with that of the community in order to provide a constructive intersection for the total development of the country. St. Mary's University (SMU) has evolved from St. Mary's language school, which started operation in 1991, was established as a college in 1998 [84].

St. Mary's University runs accredited undergraduate and graduate programs in diverse fields of studies. At present, it runs undergraduate programs in more than twenty fields of studies in the regular, extension and distance education divisions. At graduate level, it offers MBA in general business management, HRM, and accounting and finance, MA in rural development, agricultural economics, marketing management, project management, development economics, and MSc in computer science, MA in quality and productivity management, and MBA in impact entrepreneurship.

It has two campuses in Addis Abeba, 14 regional offices, and 130 coordination offices throughout the country. It has 206 male and 65 female full time and 9 male and 3 female part-time academic staffs in the two campuses of Addis Abeba, 81 male and 87 female academic staffs throughout the center and 26 male and 25 female administrative staffs. It guarantees to the needs of six thousand undergraduate students, twenty thousand students enrolled in distance education programs and two thousands students in graduate programs [85].

Vision and Mission

The vision of St. Mary's University is to become among the leading higher education centers of academic excellence in teaching-learning, research, publications and community services in east-Africa and contribute to the development of Ethiopia [84].

The mission of St. Mary's is to offer conventional and distance education that is accessible to the larger reasonable tuition focusing on quality and standards in teaching research and outreach services [84].

Organizational structure of St. Mary University

The Senate is the highest decision making body having roles and responsibilities spelt out in the university legislation and with the mandate given to its Standing Committees, which deliberate on and propose key issues to the Senate[86]. Members of the Senate include the President, Vice Presidents, Faculty Deans, Directors, staff and student union representatives. The President is empowered by the Senate of the University to manage, administer, and direct the affairs of the institution in areas of property, revenue, expenditure, business and other matters within the context of duties according to him/her [86]. Every academic department is empowered to decide on its budget and propose new programs to their respective academic commissions, which in turn

present their request to the Senate through the pertinent standing committee [86]. According to the organizational structure of St. Mary's University the office of the registrar is accountable for executive vice president aiming at handling student admission, issuing academic calendars and keeping academic records of enrolled students. It produces different statistical reports related to student attrition and retention [86]. The detailed organizational structure of St. Mary's University illustrated in annex 1.1.The other Duties and Responsibilities of the Office also noted as follow

- develops and implements systems to maintain student academic records
- Ensures the smooth running of the data base system of the Office of the Registrar.
- prepares original diplomas to be issued by the university;
- prepares academic calendar and presents it to the university senate for approval
- prepares and submits quarterly and annual performance reports of the Office to the Center for Educational Improvement and Quality Assurance (CEIQA)
- Controls smooth running of the data base system of the Office of the Registrar.

Prevalence of student attrition in St. Mary University

Based on the data obtained from the registrar office of SMU the aggregate attrition rates from 1999EC to 2008EC are presented in table 3.1 below. It indicates that student attrition is more prevalence in the extension division (51.665%) than the regular division (34%). And also Law students are the most affected group in both in regular division and in extension division with 54% and 91% respectively

Table 3.1 regular and extension student attrition rate form 1999 EC up to 2008 EC (sources from registrar)

Department	Total no of students registered		Attrition(W+AS+D+AD)		Percentage (%)	
	Regular	Extension	Regular	Extension	Regular	Extension
Accounting	4829	5135	1353	2405	28%	46.83%
Computer science	855	801	396	627	46%	78.27%
Law	189	291	102	267	54%	91.75%
Management	990	1940	403	1157	41%	59.42%

Marketing management	1476	1342	590	773	40%	57.60%
Information technology	158	67	28	57	18%	85.07%
Tourism and hospitality management	74	60	36	31	49%	51.665
Total	8571	9636	2908	5317	34%	55.17%

Causes of student attrition

Higher educational institutions literature suggests that for most of higher education institutions to have higher attrition rate and lower graduation rate were merely influenced by different factors such as health problem, financial instability, lack of motivation, policy mater of the institutions lack of attendance and so on [3]. But, as it has been described by most of the scholars the most noticeable factors were demography, pre-college experience and academic status [1, 48, and 12]. Kassahun [53] noted that, the important causes of student attrition in St. Mary's University are; financial problems, academic failure, lack of study and note-taking skills, lack of guidance and counseling services, grading system, frequent absenteeism from classes, large class size, unstable working conditions of employee students were responsible for students discontinuation of college education. Similarly, Mesfine [51] also pointed that, some of the factors that contributed to the student attrition in the case of St. Mary's University included; poor academic achievements, poor academic engagements, inadequacy of academic advising, reluctance in using counseling services, unavailability of courses, low motivation to study and takes courses seriously, maternity and health cases, poor financial status, and inconveniences, place of work (for employed students).

Getahun [1] also pointed out that, student dropout is more related to academic performance than other factors such as CGPA, followed by GPA of Term1 and Term2 and previous result in college based the data extracted from St. Mary's university student record management information system database.

Reason for Students' Withdrawal

As it had been discussed before student attrition is becoming a chronic problem that higher institution faces on daily bases. To understand in detail about the current students' attrition in St.

Mary's University we collect data about both undergraduate regular and extension program students who left St. Mary's University by completing the formal withdrawal process.

Table 3.2 presents summary of the survey of students' withdrawal from their higher education. Data tracing is necessary because of students withdrawal reason is not recorded in St. Mary's University student record management database. A total of around 749 data traced randomly from hard copy documents within 5 working days. The data is traced only from 2008 EC academic year

withdrawals reason/ formal process	Number	Average
Academic problem	117	15.62%
Family problem	50	6.67%
University changes	34	4.53%
Financial problem	43	5.74%
No reason	43	5.74%
Late registered	27	3.60
Health problem	87	11.61%
Distance home to school	33	4.40%
Impact of work	81	10.81%
Personal problem	164	21.89%
Out of country	23	2.07%
Pregnancy	19	2.53%
Transfer to distance program	28	3.73%
Total	749	100.00

Table 3.2 Reason for Student's Withdrawal

In general, from the above table personal problem and academic problem got the highest rank of all the reason which is 21.89 % and 15.62% respectively. And maternal cases (pregnancy) got the least reason with 3.73%

3.1.2 Factors Contributing to Student Attrition and Retention

Pre College Experience

Students that join to public or private universities are based on Ethiopian Higher Education Entrance Examination score point set by Minster of Education. Regardless of the variation in the cutting line for entrance point across the years, most students joining private HEIs are thought to be those who failed to join the public PHEIs(public higher education institutions) [51].

SMU chief of registrar staff pointed out that, most students admitted by St. Mary's University are those who are not interested to join public university or those who are not competent to get admission into the public university. Hence Poor student's academic performance in the EHEEE (Ethiopian higher education entrance examination) could have a great impact on student attrition and retention status. Depending on the discussion with SMU chief registrar staff we tried to figure out the most common and well-known factors for poor academic performance in EHEE (Ethiopian higher education entrance examination). Such as;

- Students lack of readiness for the exam
- Bad feeling or tension while the exam is conducted
- Cheating while the exam is conducted
- Lack of orientation by the examiners before the exam proceeded

Gender of Students

According to Tesfaye [53] from the total enrolled 6326 students in St. Mary's University in 2004/05 academic year's female students shared with 24.3%. Based on SMU registrar experts' male and female attrition and retention rate vary from year to year. But more or less females' students got higher ranks than males. Kassahun [54] explained the various reason that female students face in private higher institution.

- Sexual abusive behaviors of males towards girl students were widespread phenomena in many institutions and the area around the institution. Among the perpetrators are older male students, male teachers, and sugar daddies. These groups often use money and gifts to attract girls. Sexual harassment limits opportunities for women to participate actively in their education.

- Girls were also exposed to abusive behaviors by a male stranger on the way to and from college in the form of threatening or verbal assault.
- Educational services like guidance and counseling and tutorial program have been ineffective due to lack of trained guidance and counselor on the field. As a result, girls don't get advice about their personal and educational problems in order to mitigate the day to day challenges they face.

Preparatory Attended Region

Most of SMU regular and extension students attended preparatory in Addis Abeba. And the rests are from different region of the country. As per the discussion with experts More or less, students from different region are more likely to drop or withdraw than Addis Abeba students. Based on the interview held with SMU registrar staff most well-known reason are listed as follows

- Students from Addis Abeba have a better academic experience than students from the region.
- Most of the students came from a different region of the country can't get an education at early of their age as compared to Addis Abeba; so regional students are at a higher age than Addis Abeba which has an effect for the attrition and retention
- Students from Addis Abeba can easily get any information regarding their education than regional students
- Students from Addis Abeba have a better experience in writing, listening and speaking English languages than regional students; hence they need greater support in the understanding of the courses given by the university.
- Students from Addis Abeba had a better learning and teaching facility's than regional students
- Socioeconomic and cultural has also influence for regional students.

Impact of Works

Work has its own impact on student attrition and retention. Depending on the interview held with the instructors, most of the extension program students are being at the workplace hence, they didn't attend the class properly because they fill tiredness while the class is conducted. As per

the consultation with SMU faculty of business administration instructors, some of the impacts of work are summarized as follows.

- Filling tiredness while the class is conducted
- Difficult to get spare time to study their subject and to attend tutorial classes
- Difficult to manage huge tasks in a very short period of time after staying without work for a week
- Lately, begin the class because of the work burden
- distance from the workplace to the university so they are late from class
- Some students go to field work for a week and even for months hence thy missed exam and assignments.
- Unstable working conditions such as conflict with bosses and change of workplace

Division

At present, SMU runs an undergraduate program in regular and extension division. In the regular division, the program is given during day time while in extension division the program is given during the night time. Based on the data obtained from office registrar the aggregate attrition rate of SMU from 1999EC to 2008EC is presented in table 3.1. It indicates that student attrition is more prevalence in the extension division (51.665%) than the regular division (34%). According to SMU chief of registrar staff, some of the reasons are listed why extensions program students' attrition rate is high

- According to SMU instructors, most of the extension students are employed and mostly their achievement is based on their works burdens. Some students had work burden that is why they become unsuccessful.
- The continuous assessments and the grading system given by SMU is not considered the extension program students; that means, the grading system of the university is very tough as compared to regular program students. Extension program students had work burden and didn't have enough time to study hard as compared to regular students.
- Most of extensions programs students focused on getting the Degree rather than focusing on the knowledge because most of them are employed and they need it only to upgrade their work level.

Financial Sources

According to SMU registrar, most of the regular program students are sponsored by their family while extension program students also sponsored by their family and by themselves. In case students who are sponsored by their family any problem happen to students family that are sponsoring them, students are forced to withdrew or dropout from the program they attended. Some of the common reasons are summarized as follows.

- Students' family breakdown and divorce
- Family death (mother, father, brothers and any other family missing)
- Sickness
- Family problems, such as having a conflict with the parents

In the case of students who are sponsored by themselves and the problem happened in their own personal life affect their retention programs, the major reasons include the following

- Fired from their job
- Personal problem, such as sickness
- Workplace change
- Marriage
- Maternal cases (pregnancy)

Graduation stream

Table 3.1 indicates that law department students are the most affected group in both regular and extension, with 54% of regular and 91.75% extension students' attrition rate According to SMU faculty of informatics instructors' student should get enough information in their field study they prefer to learn. They also suggested that students should learn based on their interest and capability and needs. Hence students' wrong choose of department affects also their retention program. Some of the reasons that affect their department selection are listed below.

- Family influence in students choices of department
- Lack of brief information in the field they intended to study
- Lack of guidance and counseling by the institutions before students register or choice to join the department

- Lack of orientation about the nature of the program and the challenges it has.

Currently, student attrition becomes a serious threat for St. Mary's University as it had been described in table 3.1 the ten years attrition rate of SMU is reached to 55.1% in extension division. In addition to the literature review, we found additional factors behind student attrition. Those reasons are, personal problem, out of the country, distances from home to school, late registered and transfer to distance program.

In conclusion, developing a predictive model using data mining technology is a very useful technique for early prediction of students who may have shown attrition or retention in their study. The outcome of this research also has a great impact on higher education institution to early adjust its resources and to minimize its expenditures.

3.2 understanding of the data

Data understanding focuses on initial data collection and working on the data to be familiar with it to identify data quality problem to discover first insight into the data, or to detect interesting subset to form hypotheses for hidden information [28].

3.2.1 Data collection

The data source of this study is SRMIS (student record management information system) of St. Mary's University. The database was filled up with different attributes which are interrelated to students. Among those attributes, depending on our problem understanding we selected attributes that are related to student status (attrition and retention (graduated)). In this study considering factors identified under problem understanding we selected 15 contributing factors such as, Preparatory attended region, Types of school attended, Sex, Financial sources, Batch, 12^{th} scored result, Admission classification, Preparatory completion year, Field of study, Age, Background of study, a Year taken, Employment information, gap(the difference between preparatory completion year and university registration year) and students status(attrition or graduated), are selected to build the predictive model. For the data analysis we considered students data covering from 2005EC to 2010 EC. The original dataset obtained from SRMIS (student record management information system) has 12 attributes and more than 8000 instances. Before we obtained the final dataset we removed incomplete and missed data from each record. In this

work, attributes such as program, and student ID are removed because they are identifiers of students in this study.

3.2.2 Data description

During this phase, the data analyst examine the "gross" or "surface " properties of the acquired data and report the results, examining issues such as the format of the data, the quality of the data, the number of records and fields in each table, the identity of the field, and any other surface features of the data [28]. The obtained data from SRMIS was in two table form. The first table includes; pre-college experience, preparatory completion year, preparatory attended region, batch, and admission classification. The second table includes; sex, the background of study, graduation stream, the year taken, financial sources, and status. The two tables are merged into one table by cross-matching based on students ID numbers. Finally 15 attributes are selected. The descriptions of the selected attributes are listed below in table 3.3 and 3.4.

Table 3.3 Selected attributes with their description from SRMIS dataset

No	Variables	Descriptions	Data types	Possible values	Remark
1	Preparatory Attended Region	A region where students attended preparatory school	Nominal	{Addis Abeba , Amhara, SNNP, Oromia, Tigray, Benishagule Gumuz ,Dire Dawa, Gambella, Harari, and Somali}	Original
2	Sex	Students sex	Nominal	{male, female}	Original
3	Financial Sources	The financial sources of the students	Nominal	{parent-sponsored , self-sponsored scholar-ship}	Original
4	Batch	A year students join university	nominal	{2004EC, 2005, 2006,…… 2010EC}	Original
5	12th scored result	EHEEE results	Nominal	{Satisfactory, Good, very good and excellent}	Original
6	Admission Classification	The choice of the students intended to learn	nominal	{regular, extension}	Original
7	Preparatory completion year	A year students completed Preparatory		{1994,1995,1996……..2009}EC	Original

8	Field of study	A stream the students intended to learn	nominal	{Computer science, accounting, management, marketing management, and hotel and truism}	Original
9	Background of study	Students background study at preparatory school	Nominal	{Social, natural}	Original
10	Year taken	Total years take to finish the course	nominal	{3,4, and 5}	Original
11	Status	The academic status	Nominal	{graduated , attrition}	Original

Table 3.4Selected derived attributes with their description

No	Variables	Descriptions	Data types	Possible values	Remark
1	Types of School Attended	The types of school students took EHEEE(Ethiopian higher education entrance exam)	Nominal	{private, public}	Derived
2	Age	Age of students	nominal	{19,20,…….30}	Derived
3	Gap	The difference between preparatory completion year and university registration year	nominal	{1, 2,3,……11}	Derived
4	Employment Information	The current employment status of the student	nominal	{yes, no}	Derived

3.3 preparation of the data

The data preparation covers all the activities that are required for preparing the final dataset. The activities of the data preparation phase heavily depend on the features and the quality of the original raw data [16]. As it had been depicted in table 3.4, in this study four attributes are derived from the existing original dataset. These are;

- **Types of School Attended attribute**: it is derived from name attended institutions. We derived this attribute because it could help us to categorize students preparatory school type as private and public
- **Age attribute**: it is derived from the differences of preparatory completion year attribute and university registration year. As per the discussion with experts, we assumed that

students begin their education at the age of 7 and when they joined the university their age became 19. We derived this attribute because it could help us to determine students' status with respect to their age.

- **Gap attribute**: it is also derived from the differences of preparatory completion year attribute and university registration year. we also derived this attribute to determine students gap with respect to their status
- **Employment information**: it derived from the division attribute. As per the discussion with experts, most of the extension division program students are employed. Hence, we derived this attribute to determine the impact of work on students' status.

3.3.1 Data cleaning

Data cleaning is routine work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies [20].Without clean data, the result of a data mining analysis is in question. In this study we applied data cleaning to handle incorrect, inconsistent, and missing entry values which emerged from attended result(12^{th} score results), attended region, attended year(preparatory completion year), Gap(the difference between preparatory completion year and university registration year), and status(graduated or attrition) attributes.

3.3.1.1 Handling the incorrect or inconsistent values

The incorrect and inconsistent values have been noticed in an attended result (12^{th} score results), region, and attended year (preparatory completion year) attributes.

The key methods that are used for correcting the incorrect and inconsistent entries are as follows [22];

- **Inconsistency detection:** this typically done when the data is available from a different source in different format.
- **Domain knowledge:** A significant amount of domain knowledge is often available in terms of the ranges of the attributes or rules that specify the relationships across different attributes

- **Data-centric methods:** In these cases, the statistical behavior of the data is used to detect outliers. Data-centric methods for cleaning can sometimes be dangerous because they can result in the removal of useful knowledge from the underlying system

Table 3.5 summery of incorrect or inconsistent values and method for handling them

No	Attributes	Total instances	No of incorrect or inconsistent values	%	Handling mechanisms
1	12^{th} score result	7331	30	0.4%	Use Domain knowledge
2	Attended year	7333	28	0.0038%	Use Domain knowledge
3	Attended region	7346	15	0.002%	Use Domain knowledge

3.3.1.2 Handling missing values

Missing data could be caused by varied factors such as faulty equipment or incorrect measurements [51]. However the most serious problems of missing values are that it can result in loss of efficiency, less information extracted from data or conclusions statistically less strong, a complication in handling and analyzing the data where methods are in general not prepared to handle them and may have an impact on modeling and sometimes they can destroy it. Furthermore missing data can introduce bias resulting from differences between missing and complete data [51]. Many data mining algorithms and statistical techniques are generally tailored to draw inferences from complete datasets. It may be difficult or even inappropriate to apply these algorithms and statistical techniques on an incomplete dataset. However, any missing data treatment methods should not change the data distribution and the relationship among the attributes should be retained. The missing data problem arises when values for one or more variables are missing from recorded observations [51].

As suggested by Jiawei et.al [20], some of the methods to handle the missing values are presented as follows.

- Filling the missing values manually, it is time-consuming and may not be feasible given in large dataset with many missing values.
- Use attributes mean to fill in the missing values, in case of continuous values
- Use attributes mean for all samples belonging to the same class as the given tuple

- Use the most probable values to fill in the missing values, in case of nominal values.

As it had been presented in table 3.6 most predominantly four attributes (preparatory attended region, 12th score, gap, and status) missing values were observed and thus, we used the most probable values using mode to handle these missing values as depicted in table 3.6

Table 3.6 summery of missing values and method for filing them

No	Attributes	Total instances	No of missing values	%	Method for filing
1	Preparatory attended region	7240	121	1.67%	the most probable values using mode
2	Gap	7310	51	0.69%	the most probable values using mode
4	12th score result	7219	142	1.967%	the most probable values using mode
5	Status	7342	19	0.258%	the most probable values using mode

3.3.2 Data discretization

Jiawei et.al [20] noted that Data discretization can be used to replace a numeric attribute by interval labels or conceptual labels. The labels, in turn, can be recursively organized into high-level concepts, resulting in concept hierarchy for the numeric attributes. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data. This leads to concise, easy to use, knowledge-level representation of mining results. In this study five attributes are discretized into their respective represented values these are;

Age: Age attribute contains numeric instances of students' age from 19 to 30. According to SMU quality of education experts, we discretized age attributes into the following

- Age_1: Students are joined the University at the age of 19.

- Age_2: students are joined the University in between age 20 and 22.
- Age_3: students are joined the University at the age of 23 or 24.
- Age_4: students are joined the University in between age 25 and 30

Batch: This attribute also contains a numeric values from 2005EC to 2010 EC. As per the discussion with experts, we discretized these values into the following,

- Batch_1: students joined the University in 2005EC
- Batch_2: students joined the University in 2006EC
- Batch_3: students joined the University in 2007EC
- Batch_4: students joined the University in 2008EC
- Batch_5: students joined the University in 2009EC
- Batch_6: students joined the University in 20010EC

Gap: this attribute contains numeric values from 0 up to 11. As per the discussion with experts we also discretized these numeric instances into the following intervals.

- Gap_0: it indicates that there is no gap between the preparatory completion year and University registration year. Therefore after students completed preparatory they directly joined the university
- Gap_1: it indicates the gap (between preparatory completion year and university registration year) is 0ne.
- Gap_2: it indicates the gap(between preparatory completion year and university registration year) is two to three years
- Gap_3: it indicates the gap(between preparatory completion year and university registration year) is four to five years
- Gap_4: it indicates the gap(between preparatory completion year and university registration year) is six to seven years
- Gap_5: it indicates the gap(between preparatory completion year and university registration year) is in between eight to eleven years

Preparatory attended year: it contains numeric values from 1994EC up to 2009EC. Hence we discretized these values into two intervals as before (19994EC-2001EC) and as after (2002EC-2009EC).

Attend result: this attributes also continuous numeric values. But as per the discussion with the quality of education experts we also discretized this value into the following four intervals,

- Satisfactory: it contains students 12^{th} scored results in between 125 to 300
- Good : it contain students 12^{th} scored results in between 301 to 325
- Very Good : it contain students 12^{th} scored results in between 326 to 375
- Excellent : it contain students 12^{th} scored results greater than 376

In general, Depending on the discussion quality of education experts we discretized 12^{th} score result, Gap (the difference between preparatory completion year and university registration year), batch, and preparatory completion year values.

Table 3.7 discretized attributes and their values

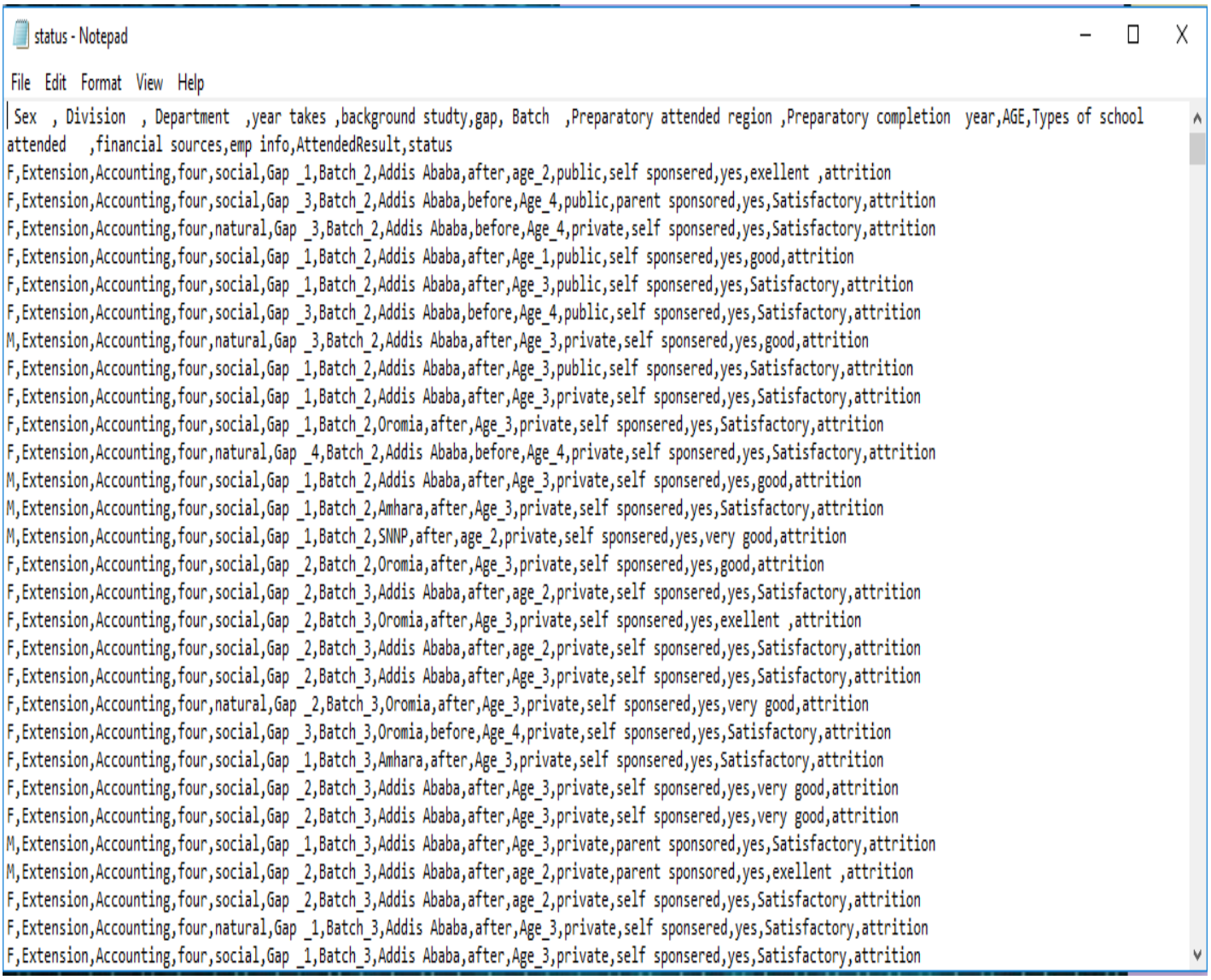
Attribute name	Values	Represented values
Age	age<=19	Age_1
	20<=age<=22	Age_2
	23<=age<=24	Age_3
	25<=age<=30	Age_4
Batch	batch<=2005	batch _1
	2005< batch <=2006	batch _2
	2006< batch <=2007	batch _3
	2007< batch <=2008	batch _4
	2008< batch<=2009	Batch_5
	2009< batch<=2010	Batch_6

Gap	gap<=0	gap _0
	0< gap <=1	gap _1
	2<= gap <=3	gap _2
	4<= gap <=5	gap _3
	6<= gap <=7	gap _4
	8<= gap <=11	gap _5
preparatory completion year	1994<=attended year<=2001	Before
	2001<attended year<=2009	After
Attended result	Attend result <=300	Satisfactory
	300< Attend result <=325	Good
	325< Attend result <=375	Very Good
	Attend result >375	Excellent

Finally after we preprocessed the original dataset we obtained 15 attributes and 7361 instances which is used for model building

3.3.3Data Format

In this study, we extracted the original data form student record management information system My-SQL database of SMU. We preprocessed the original data using a diffident preprocessing technique using Microsoft Excel. Before we apply data mining algorithm we need to convert the data into CSV (Comma Delimited) file format which is Weka acceptable format. Figure 3.1 shows sample dataset saved in CSV format for the data mining tasks using classification algorithms


status - Notepad

File Edit Format View Help

```
Sex  , Division  , Department  ,year takes ,background studty,gap, Batch  ,Preparatory attended region ,Preparatory completion  year,AGE,Types of school
attended   ,financial sources,emp info,AttendedResult,status
F,Extension,Accounting,four,social,Gap _1,Batch_2,Addis Ababa,after,age_2,public,self sponsered,yes,exellent ,attrition
F,Extension,Accounting,four,social,Gap _3,Batch_2,Addis Ababa,before,Age_4,public,parent sponsored,yes,Satisfactory,attrition
F,Extension,Accounting,four,natural,Gap _3,Batch_2,Addis Ababa,before,Age_4,private,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,social,Gap _1,Batch_2,Addis Ababa,after,Age_1,public,self sponsered,yes,good,attrition
F,Extension,Accounting,four,social,Gap _1,Batch_2,Addis Ababa,after,Age_3,public,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,social,Gap _3,Batch_2,Addis Ababa,before,Age_4,public,self sponsered,yes,Satisfactory,attrition
M,Extension,Accounting,four,natural,Gap _3,Batch_2,Addis Ababa,after,Age_3,private,self sponsered,yes,good,attrition
F,Extension,Accounting,four,social,Gap _1,Batch_2,Addis Ababa,after,Age_3,public,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,social,Gap _1,Batch_2,Addis Ababa,after,Age_3,private,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,social,Gap _1,Batch_2,Oromia,after,Age_3,private,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,natural,Gap _4,Batch_2,Addis Ababa,before,Age_4,private,self sponsered,yes,Satisfactory,attrition
M,Extension,Accounting,four,social,Gap _1,Batch_2,Addis Ababa,after,Age_3,private,self sponsered,yes,good,attrition
M,Extension,Accounting,four,social,Gap _1,Batch_2,Amhara,after,Age_3,private,self sponsered,yes,Satisfactory,attrition
M,Extension,Accounting,four,social,Gap _1,Batch_2,SNNP,after,age_2,private,self sponsered,yes,very good,attrition
F,Extension,Accounting,four,social,Gap _2,Batch_2,Oromia,after,Age_3,private,self sponsered,yes,good,attrition
F,Extension,Accounting,four,social,Gap _2,Batch_3,Addis Ababa,after,age_2,private,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,social,Gap _2,Batch_3,Oromia,after,Age_3,private,self sponsered,yes,exellent ,attrition
F,Extension,Accounting,four,social,Gap _2,Batch_3,Addis Ababa,after,age_2,private,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,social,Gap _2,Batch_3,Addis Ababa,after,Age_3,private,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,natural,Gap _2,Batch_3,Oromia,after,Age_3,private,self sponsered,yes,very good,attrition
F,Extension,Accounting,four,social,Gap _3,Batch_3,Oromia,before,Age_4,private,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,social,Gap _1,Batch_3,Amhara,after,Age_3,private,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,social,Gap _2,Batch_3,Addis Ababa,after,Age_3,private,self sponsered,yes,very good,attrition
F,Extension,Accounting,four,social,Gap _2,Batch_3,Addis Ababa,after,Age_3,private,self sponsered,yes,very good,attrition
M,Extension,Accounting,four,social,Gap _1,Batch_3,Addis Ababa,after,Age_3,private,parent sponsored,yes,Satisfactory,attrition
M,Extension,Accounting,four,social,Gap _2,Batch_3,Addis Ababa,after,age_2,private,parent sponsored,yes,exellent ,attrition
F,Extension,Accounting,four,social,Gap _2,Batch_3,Addis Ababa,after,age_2,private,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,natural,Gap _1,Batch_3,Addis Ababa,after,Age_3,private,self sponsered,yes,Satisfactory,attrition
F,Extension,Accounting,four,social,Gap _1,Batch_3,Addis Ababa,after,Age_3,private,self sponsered,yes,Satisfactory,attrition
```

Figure 3.1 sample dataset in CSV format

CHAPTER FOUR

Experimentations and analysis

In this research four classification algorithm, such as decision tree (J48), rule induction (PART and JRIP) algorithms, and probabilistic classifier (naïve bayes) are used to build the student attrition and (graduation) retention predictive model. In this study, a total of 30 experiments are conducted using the four classification algorithm stated above using 66% percentage split test and 10-fold cross-validation test options

4.1 overview

We used Weka version 3.6 software. According to Priyanga et.al [56], Weka is open-source software developed at the University of Waikato and the programming language is based on Java. Weka has 4 different applications, Explorer, Experimenter, Knowledge Flow, and Simple CLI. Knowledge Flow is a node and linked based interface and Simple CLI is the command line prompt version where each algorithm is run by hand. In our study, we used Explorer applications of the Weka. WEKA contains many inbuilt algorithms for data mining and machine learning. Weka implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes visualization tools.

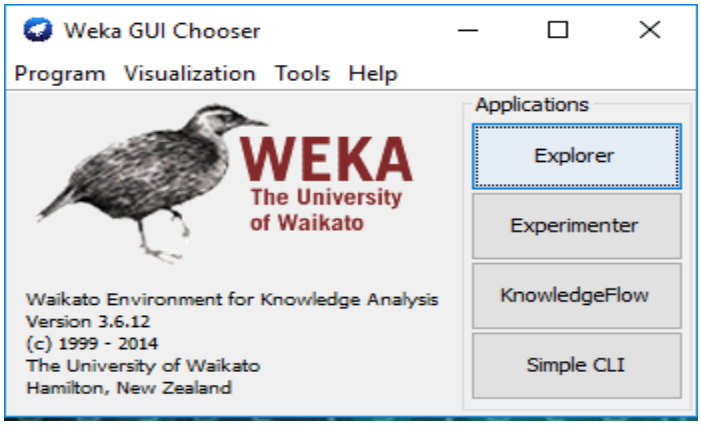


Fig 4.1 Weka interface

In this study, a total of 30 experiments are conducted using the four classification algorithm stated above using 66% percentage split test option and 10-fold cross-validation test options

4.2Balancing the Dataset

SMOTE and resampling technique also employed to solve class imbalance problem and to improve the classification accuracy.

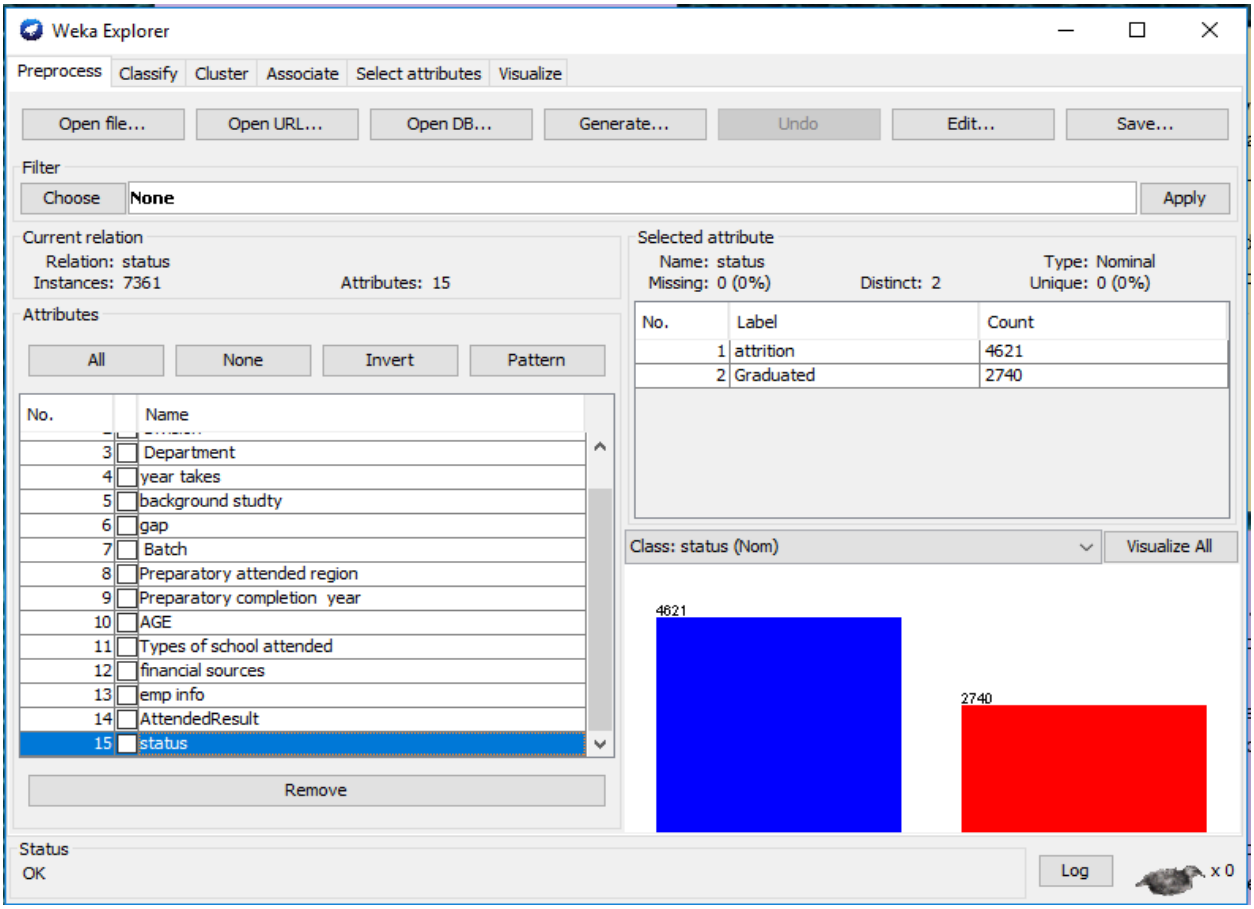


Fig 4.2 loaded data view with Weka interface

As depicted in figure 4.2, in this research we used 7361 instances with 14 independent attributes and one dependent attribute that have two classes, named as attrition and graduated. The original records are unbalanced with attrition (4621) greater than graduated (2740). In this study to balance instances of the two classes, we apply SMOTE in Weka as shown in figure 4.3 with 69% SMOTE as a result of which the size of records increased to 9251(4621 attrition and 4630 graduated).

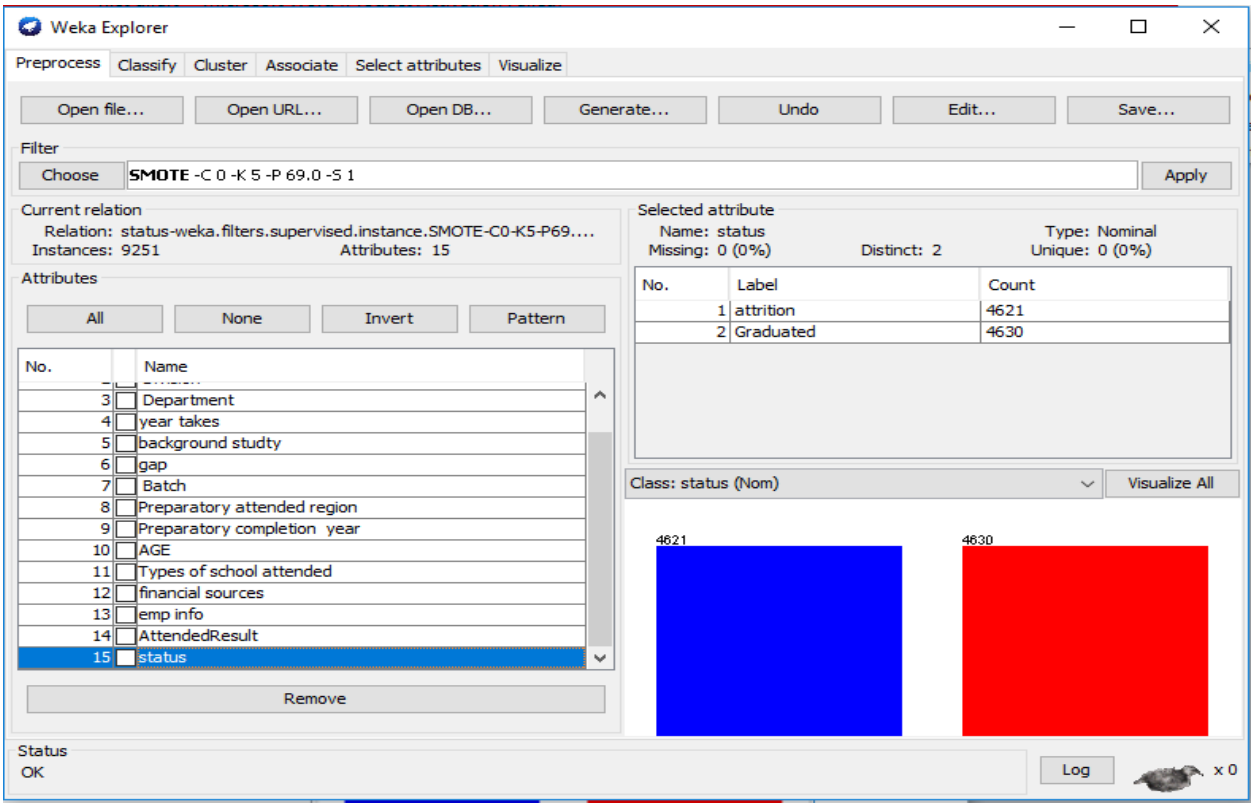


Fig 4.3 the Weka interface using SMOTE

To further increase the accuracy of the classification algorithm and to solve the problem of class imbalance we applied supervised resampling technique by changing the default values of (biasTouniformClass(0.0) into (1.0)). According to Eibe et.al [24] always predicting the majority outcome rarely says anything interesting about the data. The problem is that raw accuracy, measured by the proportion of correct predictions, is not necessarily the best criterion of success .As shown in figure 4.4 after we applied to resample the two dependent classes become (4641) attrition and (4641) graduated.

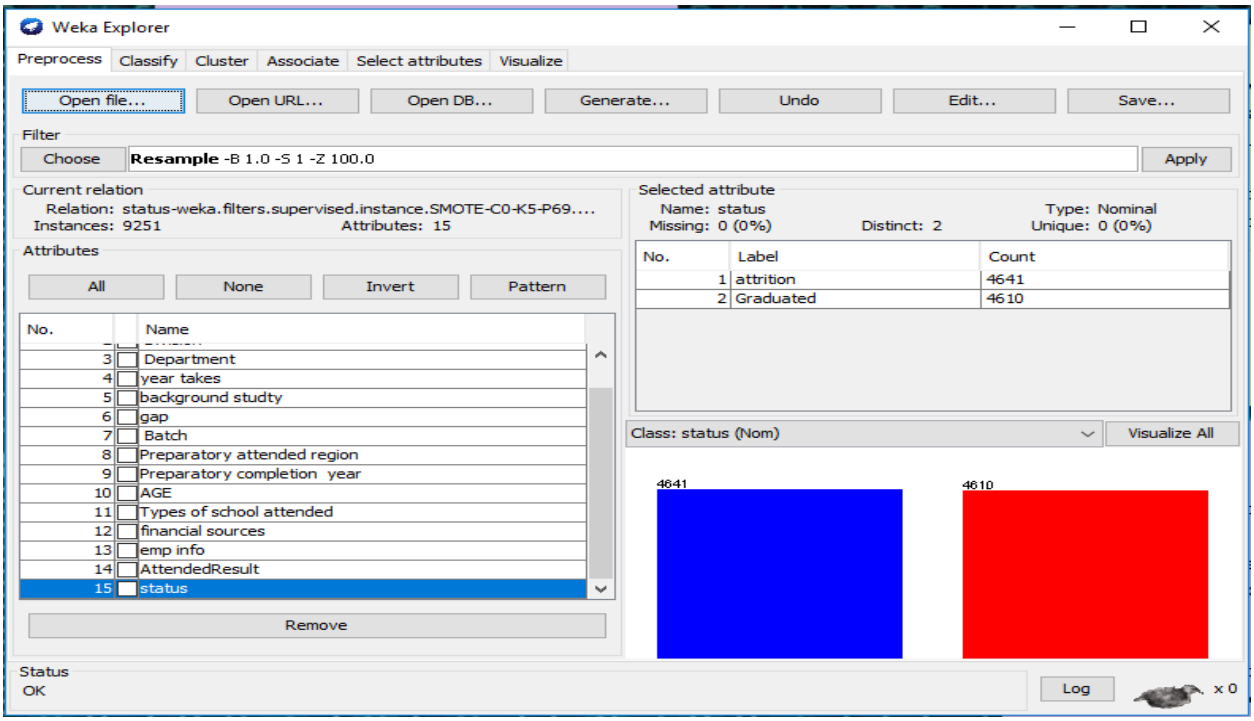


Fig 4.4 the Weka interface using resampling technique

4.3 Experimenting Decision tree classifier

Decision trees are powerful and popular tools for classification. A decision tree is a tree-like structure, which starts from root attributes, and ends, with leaf nodes [57]. Generally, a decision tree has several branches consisting of different attributes, the leaf node on each branch representing a class or a king of class distribution [57]. The advantages of decision trees are that they represent rules which could easily be understood and interpreted by users, don't require complex data preparation, and perform well for numerical and categorical variables. The Weka J48 classification filter is applied to the dataset during the experimental study. It is based on the C4.5 decision tree algorithm, building decision trees from a set of training data using the concept of information entropy [57]. In this study, we used the J48 decision tree algorithm. This algorithm is selected because of its popularities in recently published papers and has the highest accuracy as compared to other algorithms [1].The additional features of j48 are accounting for missing values, decision tree pruning, continuous attribute value ranges, and derivation of rules. It is also used to create classification model. J48 Classifier uses the normalized Information Gain Ratio for building trees as the splitting criteria. It has both reduced error pruning and normal

C4.5 pruning option [56]. Figure 4.5 depicts the default parameters of the j48 decision tree in Weka

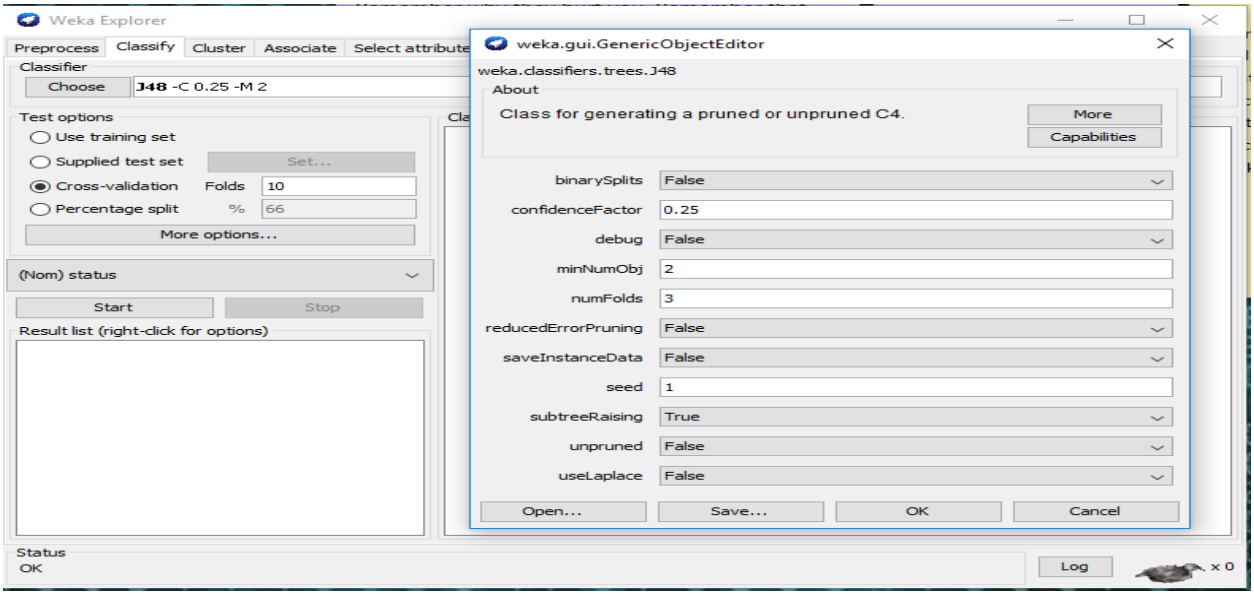


Fig 4.5 the default parameter for Weka j48 setting

Using J48 decision tree algorithm a total of eight experiments are conducted

- Experiment 1 = pruned J48 algorithm with 10 fold cross-validation test option
- Experiment 2 = pruned J48 algorithm with 66% split test option.
- Experiment 3 = unpruned J48algorithm with10 fold cross-validation test option
- Experiment 4 = unpruned J48algorithm with66% split test option
- Experiment 5 = pruned J48 algorithm with SMOTE and 10 fold cross-validation test option
- Experiment 6 =pruned J48 algorithm with SMOTE and 66% split test option.
- Experiment 7 = pruned J48 algorithm using resampling and 10 fold cross-validation test option
- Experiment 8 = pruned J48 algorithm using resampling and 66% split test option

Summary of experimental result using J48 decision tree algorithm is presented in table 4.1 below.

Table 4.1 Summery of experimental results using J48 algorithm

Experiment	Accuracy	TP rate	FP rate	precision	Recall	f- measure	ROC area	Test model
1	87.83%	0.878	0.143	0.878	0.878	0.878	0.94	10 fold cross-validation
2	88.25%	0.883	0.15	0.883	0.883	0.881	0.946	66% split
3	87.18%	0.872	0.148	0.872	0.872	0.872	0.912	10 fold cross-validation
4	87.02 %	0.87	0.152	0.87	0.87	0.87	0.915	66% split
5	88.76%	0.888	0.112	0.888	0.888	0.888	0.946	10 fold cross-validation
6	88.17%	0.882	0.118	0.883	0.882	0.882	0.94	66% split
7	91.40 %	0.914	0.086	0.915	0.914	0.914	0.959	10 fold cross-validation
8	90.24%	0.902	0.098	0.903	0.902	0.902	0.955	66% split

As shown in table 4.1 experiment 7(pruned J48 algorithm using resampling and 10 fold cross-validation test option) scored the highest accuracy of 91.40%,which means that out of the total 9251 instances 8455(91.40%) are correctly classified. While experiment 8(pruned J48 algorithm using resampling and 66% split test option) and experiment 5(pruned J48 algorithm with SMOTE and 10 fold cross-validation test option) comes 2nd and 3rd place respectively.

4.4 Rule induction algorithm

Rule induction (generation) is distinct from the generation of the decision tree [15]. While it's trivial to write a set of rules given a decision tree, it's more complex to generate rules directly from the data. The rules, however, have many advantages over decision tree, namely [15]

- They are easy to comprehend
- Their output can be easily written in the first order logic format or directly used as a knowledge base expert system
- Single rule can be understood without reference to other rule

Their disadvantage is that they don't show relationships between the rules (since the rules are independent).

In this study PART and JRIP rule induction algorithm are experimented because rule induction algorithm produces accurate and fairly compact rule sets that are easier to interpret [24].

4.4.1Model building using JRIP algorithm

JRIP implements proportional rule leaner, repeated incremental pruning to produce error reduction (RIPPER) [63]. JRIP is an interface and rule-based learner (RIPPER) that tries to come up with proportional rules which can be used to classify elements [63].

A total of eight experiments are conducted using JRIP rule induction algorithm.

- Experiment 1 = JRIP algorithm with 10 fold cross-validation test option
- Experiment 2 = JRIP algorithm using of 66% split test option
- Experiment 3 = unpruned JRIP algorithm using10 fold cross-validation test option
- Experiment 4 = unpruned JRIP algorithm using of 66% split test option
- Experiment 5 = JRIP algorithm using SMOTE and 10 fold cross-validation test option
- Experiment 6 = JRIP algorithm using SMOTE and of 66% split test option
- Experiment 7 = JRIP algorithm using resampling and 10 fold cross-validation test option
- Experiment 8 = JRIP algorithm using resampling and 66% split test option

Table 4.2 below depicts summary of experimental result using JRIP rule induction algorithm

Table 4.2 all JRIP algorithm experiments and performance result

Experiment	Accuracy	TP rate	FP rate	precision	Recall	f- measure	ROC area	Test mode
1	85.40%	0.854	0.196	0.854	0.854	0.851	0.848	10 fold cross-validation
2	85.30 %	0.853	0.195	0.854	0.853	0.85	0.844	66% split
3	80.48%	0.805	0.32	0.838	0.805	0.786	0.747	10 fold cross-validation
4	79.54%	0.795	0.321	0.831	0.795	0.776	0.738	66% split
5	87.10 %	0.871	0.129	0.875	0.871	0.871	0.894	10 fold cross-validation
6	87.47 %	0.875	0.125	0.878	0.875	0.874	0.893	66% split

7	88.33 %	0.883	0.117	0.883	0.883	0.883	0.919	10 fold cross validation
8	88.84 %	0.888	0.112	0.888	0.888	0.888	0.918	66% split

Overall experiment 8 (JRIP algorithm using resampling and 66% split test option) scored the highest accuracy of 88.84%which means, that out of the total 3145 records used for testing 2794(88.84%) instances are correctly classified. While experiment 7(JRIP algorithm using resampling and 10 fold cross-validation test option) and experiment 6(JRIP algorithm using SMOTE and of 66% split test option) takes 2nd and 3rd position respectively.

4.4.2 Model building using PART algorithm

PART(partial decision tree) adopts the divide – and – conquer strategy of RIPPER and combines it with the decision tree approach of C4.5 [62]. PART generates a set of rules by applying the divide- and –conquer strategy, removing all instances from the training collection that are covered by this rule and proceeds recursively until no instances remain [62]. To generate a single rule, PART builds a partial decision tree for the current set of instances and chooses the leaf with the largest covered as a new rule [62].

Eight experiments are conducted using PART rule induction algorithm

- Experiment1 = PART algorithm with 10 fold cross-validation test option
- Experiment2 = PART algorithm with 66% split test option
- Experiment3 = unpruned PART algorithm with10 fold cross-validation test option
- Experiment4 = unpruned PART algorithm with 66% split test option
- Experiment5 = PART algorithm with SMOTE and of 10 fold cross-validation test option
- Experiment6 = PART algorithm with SMOTE and 66% split test option
- Experiment7 = PART algorithm using resampling and 10 fold cross-validation test option
- Experiment 8 = PART algorithm using resampling and 66% split test option

A Summary of experimental result using PART rule induction algorithm is shown in table 4.3 below

Table 4.3 all PART algorithm experiments and performance result

Experiment	Accuracy	TP rate	FP rate	precision	Recall	f- measure	ROC area	Test mode
1	86.51%	0.865	0.156	0.865	0.865	0.865	0.938	10 fold cross-validation
2	87.14%	0.871	0.153	0.871	0.871	0.871	0.93	66% split
3	85.72 %	0.857	0.169	0.857	0.857	0.857	0.891	10 fold cross-validation
4	85.94 %	0.859	0.17	0.859	0.859	0.858	0.894	66% split
5	88.54%	0.885	0.115	0.886	0.885	0.885	0.948	10 fold cross-validation
6	87.50%	0.875	0.125	0.876	0.875	0.875	0.941	66% split
7	91.17 %	0.912	0.088	0.912	0.912	0.912	0.96	10 fold cross-validation
8	90.00%	0.9	0.101	0.9	0.9	0.899	0.948	66% split

Overall the seventh experiment (PART algorithm using resampling and 10 fold cross-validation test option) scored the highest accuracy of 91.17%, That is out of 9251 instances used for experiment 8434(91.17%) are correctly classified, while experiment 8 (PART algorithm using resampling and 66% split test option) and experiment 5(PART algorithm with SMOTE and of 10 fold cross-validation test option) comes 2nd and 3rd respectively.

4.5 Model building using naïve bayes algorithm

The Bayes classifier is used to model the probability of each value of the target variable for a given set of feature variables [22].

Six experiments are conducted using naïve bayes algorithm

- Experiment 1 = naïve bayes algorithm using 10 fold cross-validation test option
- Experiment 2 = naïve bayes algorithm using 66% split test option

- Experiment 3 = naïve bayes algorithm using SMOTE and 10 fold cross-validation test option
- Experiment 4 = naïve bayes algorithm using SMOTE and 66% split test option.
- Experiment 5 = naïve bayes algorithm using resampling and 10 fold cross-validation test option
- Experiment 6 = naïve bayes algorithm using resampling and 10 fold cross-validation test option

The different experiments conducted and the results obtained using naïve bayes algorithm is presented in table 4.4 below.

Table 4.4 Summery of naïve bayes algorithm experiments and performance result

Experiment	Accuracy	TP rate	FP rate	precision	Recall	f- measure	ROC area	Test mode
1	73.71%	0.737	0.271	0.747	0.737	0.74	0.827	10 fold cross-validation
2	74.55%	0.746	0.263	0.753	0.746	0.748	0.84	66% split
3	72.98 %	0.73	0.27	0.731	0.73	0.73	0.83	10 fold cross-validation
4	73.51%	0.735	0.265	0.736	0.735	0.735	0.83	66% split
5	73.11 %	0.731	0.269	0.732	0.731	0.731	0.831	10 fold cross-validation
6	72.02 %	0.72	0.28	0.72	0.72	0.72	0.823	66% split

As depicted in table 4.4 the second experiment of naïve bayes algorithm with 66% split scored the highest accuracy of 74.55% that means, out of 2503 the total instances used for testing in this experiment 1866(74.55%) are correctly classified, while experiment1 (naïve bayes algorithm using 10 fold cross-validation test option) and experiment4 (naïve bayes algorithm using SMOTE and 66% split test option) comes 2nd and 3rd respectively.

4.6 Comparison among classification algorithms

In this study depending on the problem understanding totally 15 attributes are selected. We used four classification algorithms to build a model that can predict student status. The highest scored result obtained with 10 fold cross-validation test option and 66% split test option is presented in table 4.5

Table 4.5 comparison of selected experiment with 10 fold cross-validation test option and 66% split test option

Types Algorithm	Test option	Accuracy	TP rate	FP rate	precision	Recall	f- measure	ROC area
J48	10-fold cross validation	91.40 %	0.914	0.086	0.915	0.914	0.914	0.959
	66% split	90.24 %	0.902	0.098	0.903	0.902	0.902	0.955
JRIP	10-fold cross validation	88.33 %	0.883	0.117	0.883	0.883	0.883	0.919
	66% split	88.84%	0.888	0.112	0.888	0.888	0.888	0.918
PART	10-fold cross validation	91.17 %	0.912	0.088	0.912	0.912	0.912	0.96
	66% split	90.00 %	0.9	0.101	0.9	0.9	0.899	0.948
Naïve bayes	10-fold cross validation	73.71%	0.737	0.271	0.747	0.737	0.74	0.827
	66% split	74.55%	0.746	0.263	0.753	0.746	0.748	0.84

As it had been presented in table 4.5 J48 and PART classifiers scored the highest accuracy as compared to naïve bayes, and JRIP algorithm. Overall pruned J48 algorithm using resampling and 10 fold cross-validation test option scored the highest accuracy of 91.40%. PART, JRIP, and naïve bayes comes, 2nd, 3rd, and 4th places respectively. The confusion matrix of decision tree J48 algorithm with 10 fold cross-validation test option is presented in table 4.6

Confusion matrix		
A	B	Classified
4147	494	a = attrition
302	4308	b = graduated

Table 4.6 confusion matrix for the punned J48 algorithm using resampling and 10 fold cross validation

The confusion matrix shown in table 4.6 proved that out of 9251 instances 4147 instances are correctly classified as “attrition” while 4308 instances are classified as “graduated”. On the other hand, this classifier incorrectly classified 302 instances as “attrition” and 494 instances as “graduated”. As per the discussion with experts, the reason for the misclassification of the two classes was if attrition status occurs there is also a possibility that graduated status to have occurred

4.7 Rules generated by J48 algorithm

The following are inserting rules generated by J48 algorithm. The detailed are presented in annex 4.1

Rule 1:- If financial sources = self-sponsored and division = extension and preparatory completion year = after and type of school attended = public then graduated (18.0/1.0)

If students’ financial sources = self-sponsored and program followed = extension division and preparatory completion year = after 2001EC and preparatory attended school type = public then the probability of student status is graduated

Rule 2:- If financial sources = self-sponsored and division = extension and preparatory completion year = after and types of school attended = private then attrition (4.0)

If students’ financial sources = self- sponsored and program followed = extension division and preparatory completion year = after (2002EC-2009EC) and preparatory attended school type = private then the probability of students status is attrition

Rule 3:- If financial sources = self-sponsored and division = extension and preparatory completion year = before then Graduated (155.0/12.0)

If students'' financial sources = self-sponsored and program followed = extension division and preparatory completion year = before (1994EC-2001EC) then the probability of students' status is graduated

Rule 4:- If financial sources = self-sponsored and division = regular then graduated (146.0)

If students' financial sources = self-sponsored and program followed = regular division then the probability of students' status is graduated

Rule 5:- If financial sources = parent-sponsored and division = extension then graduated (44.0)

If students' financial sources = parent-sponsored and program followed = extension division then the probability of students' status is graduated

Rule 6:- If financial sources = parent-sponsored and division = regular and department = accounting and background of study = social then attrition (63.0/4.0)

If students' financial sources = parent-sponsored and program followed = regular division and attended department = accounting and background of study = social science then the probability of students status is attrition

Rule 7:- If financial sources = parent-sponsored and division = regular and department = accounting and background of study = natural then Graduated (2.0)

If students' financial sources = parent- sponsored and program followed = regular division and attended department = accounting and background of study = natural science then the probability of students' status is graduated

Rule 8:- If financial sources = parent-sponsored and division = regular and department = management then attrition (3.0)

If students' financial sources = parent-sponsored and program followed = regular division and attended department = management then the probability of students status is attrition.

Rule 9:- if financial sources = parent-sponsored and division = regular and department = computer science then attrition (0.0)

If students' financial sources = parent-sponsored and program followed = regular division and attended department = computer science then the probability of students' status is attrition

Rule 10:- If financial sources = parent-sponsored and division = regular and department= marketing management then Graduated (37.0/2.0)

If students' financial sources = parent- sponsored and program followed = regular and attended department = marketing management then the probability of students' status is graduated

Rule 11:- If financial sources = parent-sponsored and division = regular and department = hotel and tourism then attrition (0.0)

If students' financial sources = parent-sponsored and program followed = regular division and attended department = hotel and tourism management then the probability of students' status is attrition.

Rule 12:- If financial sources = parent-sponsored and division = extension then Graduated (44.0)

If students' financial sources = parent- sponsored and program followed = extension division then the probability of students' status is graduated

Rule 13:- If financial sources = scholarship then scholar ship: Graduated (14.0)

If students' financial sources = scholarship then the probability of student stats is graduated

4.7.1 Discussion on the major findings

Depending on the rules generated by the J48 algorithm we selected six attributes to build the predictive model. These attributes are financial sources, division, department and preparatory school type, preparatory completion year, and background of the study.

The first factor identified in this study is financial sources. As we tried to explore in the problem understanding one of the financial sources is self-sponsored students themselves. This is to mean that, it is expected that self-sponsored students are probably an employee of the organization or they may run their own business. Therefore, any challenges happened in their personal life affect their retention strategy. These challenges are described in problem understanding. The same thing also true in case of parent sponsored-students that is, as we described in the problem

understanding there are also parent sponsored students. Therefore any problem happens in their parents greatly affects their retention strategy. Hence, parents should follow up with their child whether he/she is attending his/her program carefully. So we think that SMU should be aware of these factors so as to take timely action as much as possible.

A division was the second prominent factors identified in this study. As we studied under problem understanding student attrition is more prevalence to extension division than a regular division based on the data obtained from the SMU office of the registrar. In this research, it had also been observed that division has still become the major contributing factor behind student attrition.

The third factor identified in this study is the department. As we discussed before students wrong choice of department, and lack of brief orientation in the field they intended to study affect their retention. Hence SMU should have a body that is responsible for guiding and counseling students before students register or choice the department to learn.

Preparatory completion year was another factor identified under study. Preparatory completion year before and after 200l EC has also a great contribution to students' success. This is to mean that, from 1994EC up to 2001 EC the number of the subject taken at EHEEE was five but after 2002EC the number of the subject taken at EHEEE become seven. Therefore the ongoing natures of the preparatory curriculum change along with the content of the subject taken at EHEEE affect students' status.

Preparatory school type is also another finding of this study. This is to mean that students' success can also be influenced by the type preparatory completion school. This is can be associated with the availabilities of learning material or facilities, and the administrative nature of the school allows students to access to relevant information, and the availabilities of professional teachers are main indicators of students' success.

The last factor identified in this study was also a background of the study. As we observed in the generated rules social stream students has the probability of attrition status whereas natural stream students have the probability of graduate status.

In this research we tried to revealed different related works concerning the issue of student attrition and retention.

Based on the study of Getahun [1], students demography (age, sex, employment status, and income), pre-college experience (high school result and pre-college result), and university college experience ware considered as an input variable during the study. So according to his study, student dropout is more related to academic performance than other predictors.

Muluken [41] 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 failure and success.

Alemu et.al [61], also indicated that PSGPA (preparatory school grade point average result), EUEE (Ethiopian university entrance examination result), FCI (field choice interest), FYFSA (first year first semester academic achievements) and FYSSA (first year second semester academic achievements).are the major factors behind students' performance.

In general, under this study depending on the knowledge generated by the decision tree J48 algorithm, it has been found that financial sources, division, department and preparatory school type, preparatory completion year, and background of study become a new finding that the other researcher didn't consider it.

4.8 User interface design

Using the 13 rules generated by J48 algorithm we designed a user interface using C# programming language. As it had been described in section 4.7 the rules are listed as IF–THEN rule. The designed user prototype accept user query and suggest students Status. Figure 4.5 below shows the user interface that enables user interaction with the user interface

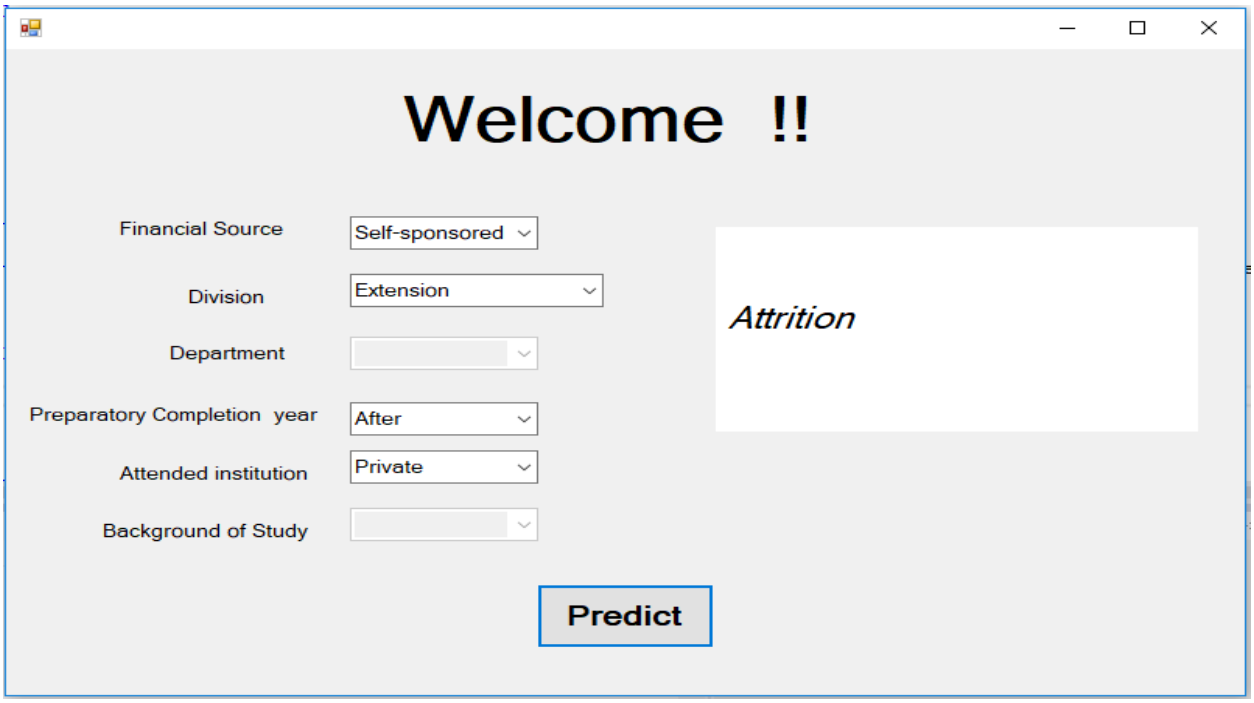


Fig 4.6 Sample prototype for suggesting student status

As it had been shown in figure 4.6, we used six attributes to build student status prototype. This user interface accepts the data of student such as financial sources, division, department, preparatory completion year, preparatory school type and background of study from the user and it displays students' status as attrition and graduated

4.8.1Validity of user acceptance testing

After we developed a prototype that can predict students status, we, prepared a questioner (attached in Annex 4.4) to check the validity of the prototype. This questioner consists of four items and five scales ranging from strongly agree to strongly disagree. The respondents of these questioners are selected from SMU registrar staff, academic deans, department heads, and instructors' from the faculty of business and school of informatics' totally five people are involved to evaluate the prototype.

Questionnaires	Strongly Agree (5)	Agree (4)	Undecided (3)	Disagree (2)	Strongly Disagree (1)
Efficiency: • In terms of time. • In terms of accuracy .	90%	10%	-	-	-
	85%	15%			
Effectiveness: • In terms of the output result. • In terms of performance	80%	10%	10%	-	-
	90%	10%			
Easy to understand • In terms of the feature of user interface. • In terms of platform	80%	10%	10%	-	-
	85%	15%			
Easy to Remember: • In terms remembering the way to use the prototype	80%	20%		-	-

Table 4.8validity on student status prototype

As it had been illustrated in Table 4.8 most of the respondents have a positive attitude towards the validity of the prototype. Most of the respondents also explained their thought that predicting student status is a new technique that the university didn't use before. Of course, in this research, we introduced new trends for St. May's University to use the historical institutional students' records for determining students' status by applying data mining technique. Therefore St. Mary's University can use this system to effectively identify students' status as either of attrition or graduated (retention).

When we came to the weakness suggested by the respondents, one respondent from undergraduate register member explained her thought as follow,

"I found this user interface so interesting in terms of predicting students' status early as attrition and graduated since we didn't use such data mining technique before to early identify our students status so I believe this prototype we make our job easier since we can identify students'

status early as much as possible, but I have one suggestion to this prototype that is, if other factors also be included it will be more and more better"

Another respondent, from the member of department head also explained his thought as follow,

"nowadays student attrition is becoming a big headache for our institution, even the university is just facing as such a difficult issue corrective attempt is not taking timely by the university chief executive staffs, more or less the university didn't conduct any remarkable study concerning the issue of student attrition, so I believe this prototype will minimize cost expenditure of the university and minimize attrition rate of the university so I found that it is so impressive but much research also need be done in terms feature of user interface that is another feature also be included"

CHAPTER FIVE

CONCULUSION AND RECOMEDATION

Conclusion

Student attrition is a universal problem in the academic arena. It has both educational and cost implications. St. Mary's University among others PHEI (private higher educational institution) offer conventional and distances education that is accessible to the large society through reasonable tuition focusing on quality and standard in learning, research and outreach services. Despite this, student attrition also becomes a chronic problem that the university faces on a daily bases and losing unexpected cost expenditure every year.

The aim of this study is to develop a predictive model using data mining classification techniques so as to determine undergraduate students' status in higher education.

As it had been indicated, in this study, there are various aspects that can have a great potential contribution to student attrition and retention (graduation) status. A variable considered under this study includes, Preparatory attended region, Types of school attended, Sex, Financial sources, Batch, 12^{th} scored result, Admission classification, Preparatory completion year, Field of study, Age, Background of study, Year taken, Employment information, gap(the difference between preparatory completion year and university registration year), and students status(attrition or graduated) were used to build the predictive model.

This study was conducted by obtaining the historical institutional data from St. Mary's university student record management information system My-SQL database to develop a model that is capable of determining students' status using data mining technology. In this study, we used five years of data that covers from 2005EC to 2010EC.

The hybrid data mining process model and the Weka version 3.6 data mining tool were employed to undertake the experiment. In this study, 30 experiments have been carried out using four classification algorithms such as decision tree classifier (J48), rule induction (JRIP and PART), and probabilistic classifier (naïve bayes). Hence, among the four algorithms tested, the J48 decision tree classifier algorithm scored the highest accuracy followed by PART, JRIP, and

naïve bayes algorithms. In addition to this, to solve data imbalanced problem (overfitting) and to increase classification accuracy we also applied SMOTE (synthetic minority oversampling technique) and resampling technique.

Based on, the extracted hidden pattern using J48 algorithm, financial sources, division, department, preparatory completion year, preparatory school type, and background of study are identified as the major contributing factors of student status. Also depending on the rule generated by J48 algorithm user interface is designed that can accept user query and suggest students Status

Due to unavailability of clear data found in the dataset, other demographic data such as the native language, marital status, place of birth location (urban or rural), and health-related data are not included under this study. In addition, the data obtained from SRMIS (student record management information system) was in two table format. So merging the two tables into one table format was the major challenge of this study.

Recommendation and future works

Based on the finding of the study, we recommended the following as future research direction

- In this study, we used a decision tree, rule induction, and probabilistic data mining classifiers. To improve the performance of the predictive model, further research is needed using other models
- As it had been indicated, in this research the survey has been carried out to find out the major causes of student attrition in St. Mary's University. Depending on it we recommended other researchers to investigate on other causes such as family-related problem, health problem, distance related issues from home to school, transferring program from regular to extension or distance program, university change, late registered, and out of the country
- It is difficult to get well organized, correct and quality data for the mining tasks. We suggest educational institutions to analyze their data symmetrically for data analyses.

- In this study the determination of student status using data mining techniques was only done for St. Mary's University but we proposed other researchers to explore the other private or public university trends of student status

References

[1] G.Semeon, “using data mining technique to predict student dropout in St. Mary’s university college: its implication to quality of education”, UNECA Conference Center Addis Abeba, Ethiopia, pp.51-76.2011

[2]A. AL-Hawaj, W. Elali, and E.T wizell,”higher education in the twenty-first century: issues and challenges”, in higher education in the twenty-first century: issues and challenges, KINGDOM OF BAHARIN,2007, pp,3-97

[3] V. Hegde, “Dimensionality Reduction Technique for Developing Undergraduate Student Dropout Model using Principal Component Analysis through R Package,” 2016.

[4] S. Pal, “mining educational data to reduce dropout rates of engineering students”, I.J information engineering & electronics business, vol.2, no.1 pp.1-7, 2012

[5] A. Bajaj and D.Saxena,”use of data mining techniques in accessing student and faculty needs”, international journal of innovative research and development, vol.6, no.3, pp.199-202, 2017.

[6] A.Dutt, M.Isemail and T. Herawan,”A Systematic Review on educational Data Mining”, IEEE Access, vol,pp, 15991-16005,2017.

[7] Y.Zhang and S.Oussena, ”use data mining to improve student retention in higher education a case study”, ICEIS-12th international conference on enterprise information system, pp.8-12, 2010.

[8] D.Gouws and H. Wolmarans,” Quality cost in tertiary education: Making internal failure cost visible”, Meditari Accountancy Research Vol.10, 2002

[9] A. Olsen,” Staying the Course: Retention and Attrition in Australian Universities”, National Audit Office 2007 Staying the course: the retention of students in higher education, report by the Comptroller and Auditor General, 26 July 2007

[10] G.Topf, A.M & J.schuh, “institutional selectivety and institutional expenditures: examining organizational factors that contribute to retention and graduation”, research in higher education, vol,47(6), pp,613-642,2006.

[11] Z.Kovacic, “predicting student success by mining enrollment data”, research in higher education journal, pp.1-20.June 2010.

[12].D.Delen,"predicting student attrition with data mining method", J. college student retention,vol.13(1),pp.17-35,2011-2012.

[13] G.Dekker, M.Pechenizkiy and M.Vleeshouwers, "predicting students dropout: A case study", educational data miming", pp.41-50, 2009.

[14] M. Adane,"Realities and Challenges of Private Higher Education Institutions to promote Quality education for sustainable development in Ethiopia: The case of St. Mary's University",Held at African Union Conference Centre, Addis Ababa, Ethiopia,, 13th July, 2016

[15] K. Cios, W.pedrycz, R.Swiniarki and L.Kurgan, the knowledge discovery approach, springer, 2007.

[16].[2]"Data Mining Methodology | DATASKILLS", Dataskills.it, 2018. [Online]. Available: https://www.dataskills.it/en/data-mining/methodology/. [Accessed: 21- Oct- 2018].

[17] M .Ayash, "Research Methodologies in Computer Science and Information Systems",Alquds Open University College of Technology & Applied Science,[accessed 18: jan-2019].

[18] R. Pressman," Software Engineering: A Practitioner's Approach", McGraw-Hill, New York, 2015.

[19]J. Albahari and B. Albahari, C# 5.0 in a nutshell. Beijing: O'Reilly, 2012.

[20] J.Han, M.Kamber and J.pei, data mining concepts and techniques, 3rd ed. Waltham: Elsevier,2012

[21] D. Hand, H. Mannila and P. Smyth, principle of data mining. New Delhi: MIT press, Cambridge,MA, USA, 2004, pp. 1-3.

[22] CH.Aggrwal, data mining the text book. Switzerland: springer, p.1,2015

[23] M.Zaki and W. Meira, data mining and analysis. New York: Cambridge university press, 2014, p.4.

[24] I. Witten, E. Frank, M. Hall, C. Pal, data mining practical machine learning tools and technique s, 4th ed, Hampshire: Elsevier, 2017, p 5-9.

[25] S. Lehr, H. Liu, S.Klinglesmith, and A.Konyha,"Use Educational Data Mining to Predict Undergraduate Retention",IEEE 16th International Conference on Advanced Learning Technologies, pp.428-430, July 2016

[26] U.Fayyad, G.Shapiro, and P. Smyth," from data mining to knowledge discovery in databases", AL Magazine, vol.17, no.2, 1996.

[27] U. Fayyad, G. Shapiro and P. Smyth, "the KDD process for extracting useful knowledge from volumes of data", communication of the ACM, vol.39, no.1, November 1996.

[28] C.Shearer,"the crip-dm model: the new blueprint for data mining", journal of data warehousing, vol.5, no.4, pp.13-14, 2000.

[29] O.Marban, G.Mariscal, and J.segovia," A Data Mining & Knowledge Discovery Process Model", Julio Ponce and Adem Karahoca, pp. 438, February 2009.

[30]"image for data mining semma process model - Google Search", Google.com, 2018. [Online]. Available:https://www.google.com/search?q=image+for+data+mining+semma+process+model&tbm=isch&source=iu&ictx=1&fir=0I0LKez3R3NxIM%253A%252CjrTXfy3W51KSTM%252C_&usg=AI4_-kQIwvt_ILXiQqdNVFnqvHQBhI_Npw&sa=X&ved=2ahUKEwjb-ozQhpfeAhXE6FMKHYmYDawQ9QEwCXoECAUQFg#imgrc=0I0LKez3R3NxIM:.[Accessed :21-Oct- 2018].

[31]R. Johan & J. Harlan, "EDUCATION NOWADAYS", International Journal of Educational Science and Research (IJESR),Vol. 4, Issue 5, pp.51-56, Oct 2014.

[32] St. Mary college's, prospectus 2005-2006, Addis abeba, January, 2005.

[33] K.Agyapong, J. Acquah,,and M. Asante, "An Overview of Data Mining Models (Descriptive and Predictive)",International Journal of Software & Hardware Research in Engineering, vol.4, no.5, pp.53-60 ISSN-2347-4890 2016.,

[34]"image for data mining data minig tasks - Google Search", Google.com, 2018. [Online]. Available: https://www.google.com/search?tbm=isch&sa=1&ei=MTDMW5OgB9CazwK8s6OQCw&q=image+for+data+mining+data+minig+tasks&oq=image+for+data+mining+data+minig+tasks&gs_l=img.3...190995.207104.0.207408.39.34.2.0.0.0.517.5123.2-17j1j1j1.20.0....0...1c.1.64.img..22.0.0....0.W1h4QYJkEt0#imgrc=0gifOGAbPHIQWM:.[Accessed: 21- Oct- 2018].

[35] P. Sharma, data mining and data warehousing. New Delhi: SHREE publication and distributors, 2013, p. 101.

[36] J. Grus, data science from scratch: first principle with python. United States of America: O'Reilly media, inc.,1005, Gravenstein Highway North, Sebastopol, CA,95472, 2015, pp. 201-2005.

[37]N. Saravanan and V. Gayathric, "CLASSIFICATION OF DENGUE DATASET USING J48 ALGORITHM AND ANT COLONY BASED AJ48 ALGORITHM", International Conference on Inventive Computing and Informatics, 2017.

[38] S. Sahu and B. Mehte, "Network Intrusion Detection System Using J48 Decision Tree", IEEE, p. 2024, 2015

[39] S.Sharma, K.Muata, and O.Bryson," Organization-Ontology Based Framework for Implementing the Business Understanding Phase of Data Mining Projects", the 41st Hawaii International Conference on System Sciences, Virginia Commonwealth University,2008.

[40]A. Azevedo and M. Santos,"KDD, SEMMA and CRISP-DM: a parallel overview", IADIS European conference on data mining, pp.184, 2008.

[41] M.Yehuala,"Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University)",INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, pp.91-93, VOLUME 4, ISSUE 04, ISSN 2277-8616,APRIL 2015

[42]V. Parsania and N. H Bhalodiya, "Applying Naïve bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis", INTERNATIONAL JOURNAL OF DARSHAN INSTITUTE ON ENGINEERING RESEARCH & EMERGING TECHNOLOGIES, p. 61, 2014.

[43] K. Sukhija, M. Jindal, and N. Aggarwal,"the recent state of educational data mining: A survey and futer visions",2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), IEEE access,pp.354-358,2015.

[44] A.Mehata, and N.Buch, "depth and breadth of educational data mining: researchers' point of view", IEEE access.

[45] L. Khanna, D. Singh and D. Aset," educational data mining and its role in determining factors affecting students' academic performance: A systematic review", information processing(IICIP), 2016 1st India international conference, 2017.

[46] T. Teklu, "Identifying Determinant Factors for Students' Success in," AAU respository, Addis Abeba, 2017.

[47]V.Mhetre and M. Mhetre, "Classification based data mining algorithms to predict slow, average and fast learners in educational system using Weka.," in Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication((ICCMC), mumbai, 2017.

[48] C.Vera, C.Morales, and S.Soto,"Predicting School Failure and Dropout by Using Data Mining Techniques",IEEE journal of Latin- American Learning Technologies,Vol. 8, No. 1,2013.

[49] St. Mary University, "graduate student hand book", 2nd ed, Addis Abeba, SMU printing press, 2014/2015.

[50]W. Tamrat, "A Glimpse of non- for-profit higher education institution in Ethiopia", the teacher, pp. 2-3, 2018.

[51]R. Houari, A. Bounceur, A. Tari and M. Kechadi, "Handling Missing Data problem with sampling methods", *research gate*, 2014.

[52]M. Tekleab,"Student Attrition: Factors and Possible Ways of Management in Private Higher Education Institutions", Major Theme: Charting the Roadmap to Private Higher Education in Ethiopia,UN Conference Center, Addis Ababa, Ethiopia,August 29, 2009.

[53] K. Habtamu,"An investigation of the Major Causes of Student Attrition in St. Mary's University College",Major Theme: Private Higher Education in Ethiopia at the turn of the Ethiopian Millennium,Proceedings of the Fifth National Conference on Private Higher Education Institutions (PHEIs) in Ethiopia,UN Conference Center Addis Ababa Ethiopia,August 25, 2007.

[53]T. Semela,"Academic, Social and Psychological Correlates of Gender Disparity in Higher Education: The Case of Debub University",Proceedings of the Fourth National Conference on Private Higher Education in Ethiopia, August 18 & 19, 2006.

[54]K. Tegegne, "Factors Affecting Gender Equality in Private Higher Education of Ethiopia : The Case of North Gondar",Proceedings of the Fourth National Conference on Private Higher Education in Ethiopia, August 18 & 19, 2006.

[55] L.Rokach, and O.Maimon,"decision tree", the data mining and knowledge discovery handbook, pp.165-192, January,2005.

[56] P. Chandrasekar, K.i Qian, H. Shahriar, P. Bhattacharya,"Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing", IEEE 41st Annual Computer Software and Applications Conference,2017

[57] BULGARIAN ACADEMY OF SCIENCES,"Predicting Student Performance by Using Data Mining Methods for Classification", Volume 13, No 1,2013

[58] N. Bhargava, G. Sharma, R. Bhargava and M. Mathuria,"Decision Tree Analysis on J48 Algorithm for Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[59] G.Kaur,A.Chhabra,"Improved J48 Classification Algorithm for the Prediction of Diabetes",International Journal of Computer Applications (0975 – 8887),Volume 98 – No.22, July 2014

[60] pp,"data mining and data warehousing ", 2017-18.

[61] Alemu et.al,"Educational Data Mining for Students' Academic Performance Analysis in Selected Ethiopian Universities",Journal of Information and Knowledge Management Vol. 9 (2) Pg 1 - 15 ISSN: 2141 – 4297 (print) ISSN: 2360 – 994X (e-version),2018.

[62] M.Berger, D.Merkl, and M.Dittenbach, "Exploiting Partial Decision Trees for Feature Subset Selection in e-Mail Categorization" ACM, April, 2006

[63] C.Devasena, T.Sumathi, V.Gomathi, and M.Hemalatha, "effectiveness evaluation of rule – based classifiers for the classification or Iris dataset", bonfring international journal of am machine interface, vol.1, dec.2011

[64] T.Patil,M S.Sherekar,"Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification",International Journal Of Computer Science And Applications,Vol. 6, No.2, April 2013.

[65]M.Arroy, and L.Sucar,"Learning an Optimal Naive Bayes Classifier", IEEE Acess,The 18th International Conference on Pattern Recognition (ICPR'06),2006.

[66]J. Wang, M. Xu, H. Wang, and J. Zhang,"Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding", ICSP2006 Proceedings,Shanghai,2006.

[67]N. Chawl,DATA MINING FOR IMBALANCED DATASETS: AN OVERVIEW, pp.860-861.

[68] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer,"SMOTE: Synthetic Minority Over-sampling Technique",Journal of Artificial Intelligence Research 16 (2002) 321–357,AI Access Foundation and Morgan Kaufmann Public

[69] Mahlet et.al,"Higher Education Students' Enrolment Forecasting System Using Data Mining Application in Ethiopia",Higher Education Students' Enrolment Forecasting System Using Data Mining Application in Ethiopia,HiLCoE Journal of Computer Science and Technology, Vol. 2, No. 2, pp.37-43.

[70] U. Shafique and H. Qaiser,"A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)",International Journal of Innovation and Scientific Research,Vol. 12 No. 1, pp. 217-222 Nov. 2014,

[71] "Pseudo code-of-naive-bayes-algorithm", researchgate.com, 2019. [Online]. Available:https://www.researchgate.net/figure/Pseudocode-of-naive-bayes-algorithm_fig2_325937073[accessed: 16:-Jan-2019]

[72] A. More,"Survey of resampling techniques for improving classification performance in unbalanced datasets",arXiv:1608.06048v1 [stat.AP] 22 Aug 2016.

[73]S.Aksenova, "machine learning with weka explorer tutorials for weka version 3.4.3", school of engineering and computer science department of computer science California state university, sacramento California, 95819,pp 1-44,2004

[74] A. Kidane, L. Wolde, T. Makonnen, Y. Yusuf and O. Abdi, ”students drop-out in institutions of higher learning in Ethiopia magnitude causes and cures”, the Ethiopian journal of education vol.x, no.2, 1989

[75] Ts. Weldegiorgis and Y. Awel, “Determinants of Student Attrition at College of Business and Economics, Mekelle University: Econometric Investigation”, Proceedings of the National Symposium on Establishing, Enhancing & Sustaining Quality Practices in Education.

[76] A. Tilahun, “Assessment of Student Attrition at the Faculty of Science, Addis Ababa University”, AAU repository, May, 2003

[77]F. Eshetu, A. Guye, G. Kelemework and S. Abebe, “Determinants of students’ vulnerability to attrition in higher education: Evidence from Arba Minch University, Ethiopia”,Educational Research and Reviews, Vol. 13(15), pp. 570-581, 10 August, 2018.

[78] W. Atnaf, T. Petros, “Factors affecting female students’ academic performance at higher education: The case of Bahir Dar University, Ethiopia”, African Educational Research Journal Vol. 2(4), pp. 161-166, December 2014.

[79]K. Asante & D. Doh, “Students’ attrition and retention in higher education: A conceptual discussion”,star conference proceedings 29, July, 2016

[80] V. Tinto, and J. Cullen, “dropout in higher education: a review and theoretical syntheses of recent research”, Teachers College Columbia University, June 30, 1973

[81] Y. Woldetensae, “Enhancement of Quality in Teaching and Learning: Implications to Ethiopian Fourth National Conference on Private Higher Education in Ethiopia, August 18 & 19, 2006.

[82] Abiy .Fetene,”an assessment of the status of quality of education in government secondary schools of bole sub-city in Addis abeba city administration”,AAU repository, 2005.

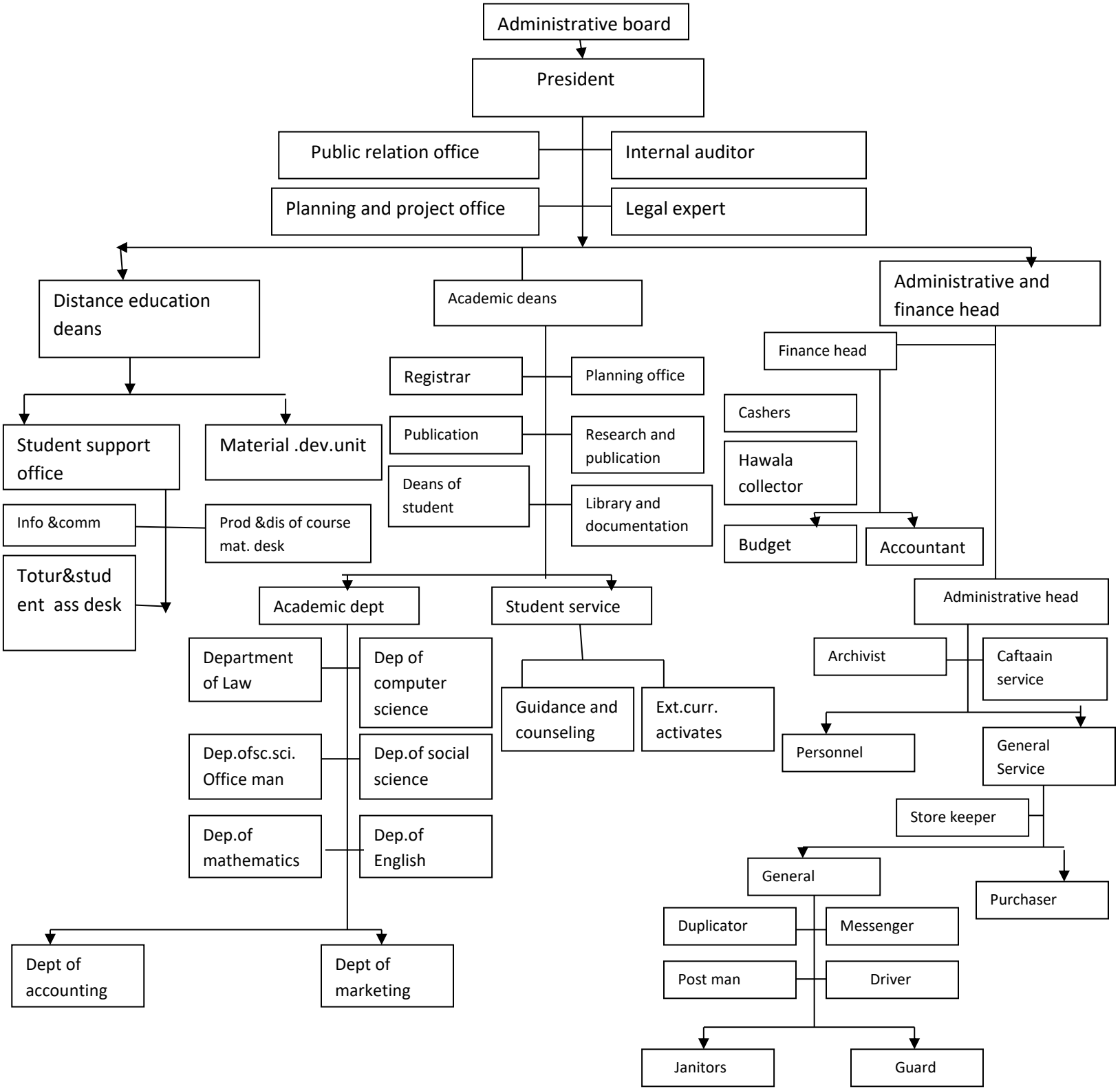
[83] M. Bramer, principle of data mining undergraduate topic in computer science, 2nded, London, springer, 2013.

[84] St. Mary University, “graduate student hand book”, 2nded, Addis Abeba, SMU printing press, 2014/2015.

[85] Z. Ayalew, "overview of St. Mary university", St.Mary university school of graduate, 2018.

[86]"OFFICES AND CENTERS - st. Mary's University", Smuc.edu.et, 2018. [Online]. Available: http://www.smuc.edu.et/index.php/servicgovernancees. [Accessed: 21- Oct- 2018].

Annex 3.1 organizational structure of St. Mary University [32]



Annex: 4.1 rule generated by J48 algorithm

J48 pruned tree

Batch = Batch_2

| background studty = social

| | financial sources = self sponsered

| | | year takes = four

| | | | Division = Extension

| | | | | AGE = age_2

| | | | | | Preparatory attended region = Addis Ababa: attrition (21.0/1.0)

| | | | | | Preparatory attended region = Oromia

| | | | | | | gap = Gap _1: Graduated (4.0)

| | | | | | | gap = Gap _3: attrition (1.0)

| | | | | | | gap = Gap _4: Graduated (0.0)

| | | | | | | gap = Gap _2

| | | | | | | | Department = Accounting

| | | | | | | | | AttendedResult = exellent : attrition (0.0)

| | | | | | | | | AttendedResult = Satisfactory: attrition (16.0/6.0)

| | | | | | | | | AttendedResult = good: attrition (0.0)

| | | | | | | | | AttendedResult = very good: Graduated (2.0)

| | | | | | | | Department = Computer Science: Graduated (0.0)

| | | | | | | | Department = Management: Graduated (2.0)

| | | | | | | | Department = Marketing Management: Graduated (7.0)

| | | | | | | | Department = Tourism and Hospitality Management: Graduated (0.0)

| | | | | | | gap = Gap _5: Graduated (0.0)

| | | | | | | gap = Gap _0: Graduated (0.0)

| | | | | | Preparatory attended region = Amhara: attrition (10.0)

| | | | | | Preparatory attended region = SNNP: attrition (4.0)

| | | | | | Preparatory attended region = Gambella: attrition (7.0)

| | | | | | Preparatory attended region = Tigray: Graduated (9.0/3.0)

| | | | | | Preparatory attended region = Somali: attrition (0.0)

| | | | | | Preparatory attended region = Dire Dawa

| | | | | | | Types of school attended = public: Graduated (2.0)

| | | | | | | Types of school attended = private: attrition (2.0)

| | | | | | Preparatory attended region = Harari: attrition (0.0)

| | | | | | Preparatory attended region = somali: attrition (0.0)

| | | | | | Preparatory attended region = Benishagule Gumuz: attrition (0.0)

| | | | | AGE = Age_4

| | | | | | Preparatory attended region = Addis Ababa

| | | | | | | Types of school attended = public

| | | | | | | | Preparatory completion year = after

| | | | | | | | | gap = Gap _1: attrition (3.0)

| | | | | | | | | gap = Gap _3: Graduated (3.0)

| | | | | | | | | gap = Gap _4: attrition (0.0)

| | | | | | | | | gap = Gap _2: attrition (2.0)

| | | | | | | | | gap = Gap _5: attrition (0.0)

| | | | | | | | | gap = Gap _0: attrition (1.0)

| | | | | | | | Preparatory completion year = before

| | | | | | | | | AttendedResult = exellent : Graduated (0.0)

| | | | | | | | | AttendedResult = Satisfactory: Graduated (106.0/22.0)

| | | | | | | | | AttendedResult = good

| | | | | | | | | | Sex = F: attrition (6.0/1.0)

| | | | | | | | | | Sex = M: Graduated (3.0)

| | | | | | | | | AttendedResult = very good: Graduated (0.0)

| | | | | | | Types of school attended = private

| | | | | | | | Department = Accounting: attrition (17.0/1.0)

| | | | | | | | Department = Computer Science: attrition (0.0)

| | | | | | | | Department = Management: Graduated (4.0/1.0)

| | | | | | | | Department = Marketing Management: Graduated (1.0)

| | | | | | | | Department = Tourism and Hospitality Management: attrition (0.0)

| | | | | Preparatory attended region = Oromia

| | | | | | | Types of school attended = public: attrition (17.0/2.0)

| | | | | | | Types of school attended = private

| | | | | | | | AttendedResult = exellent : attrition (2.0)

| | | | | | | | AttendedResult = Satisfactory: Graduated (26.0/7.0)

| | | | | | | | AttendedResult = good: Graduated (0.0)

| | | | | | | | AttendedResult = very good: attrition (1.0)

| | | | | | Preparatory attended region = Amhara

| | | | | | | Sex = F: Graduated (10.0/2.0)

| | | | | | | Sex = M: attrition (6.0)

| | | | | | Preparatory attended region = SNNP

| | | | | | | AttendedResult = exellent : attrition (0.0)

| | | | | | | AttendedResult = Satisfactory: attrition (8.0/1.0)

| | | | | | | AttendedResult = good: Graduated (2.0)

| | | | | | | AttendedResult = very good: attrition (0.0)

| | | | | | Preparatory attended region = Gambella: Graduated (0.0)

| | | | | | Preparatory attended region = Tigray: attrition (18.0/5.0)

| | | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | | Preparatory attended region = Dire Dawa: Graduated (3.0)

| | | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | | AGE = Age_1

| | | | | | AttendedResult = exellent : attrition (2.0)

| | | | | | AttendedResult = Satisfactory

| | | | | | | Sex = F

| | | | | | | | Department = Accounting

| | | | | | | | | Preparatory attended region = Addis Ababa

| | | | | | | | | | Types of school attended = public: Graduated (17.0/7.0)

| | | | | | | | | | Types of school attended = private: attrition (5.0)

| | | | | | | | | Preparatory attended region = Oromia: attrition (10.0/3.0)

| | | | | | | | | Preparatory attended region = Amhara: Graduated (12.0/1.0)

| | | | | | | | | Preparatory attended region = SNNP: Graduated (4.0/1.0)

| | | | | | | | | Preparatory attended region = Gambella: Graduated (0.0)

| | | | | | | | | Preparatory attended region = Tigray: attrition (4.0/1.0)

| | | | | | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | | | | | Preparatory attended region = Dire Dawa: Graduated (0.0)

| | | | | | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | | | | | Department = Computer Science: Graduated (0.0)

| | | | | | | | Department = Management: Graduated (5.0)

| | | | | | | | Department = Marketing Management: attrition (8.0/2.0)

| | | | | | | | Department = Tourism and Hospitality Management: Graduated (0.0)

| | | | | | | Sex = M

| | | | | | | | Types of school attended = public: attrition (16.0)

| | | | | | | | Types of school attended = private: Graduated (8.0/3.0)

| | | | | | AttendedResult = good: attrition (11.0/1.0)

| | | | | | AttendedResult = very good: attrition (11.0)

| | | | | AGE = Age_3

| | | | | | Preparatory attended region = Addis Ababa

| | | | | | | Department = Accounting

| | | | | | | | Types of school attended = public

| | | | | | | | | AttendedResult = exellent : Graduated (0.0)

| | | | | | | | | AttendedResult = Satisfactory

| | | | | | | | | | Sex = F: Graduated (13.0/4.0)

| | | | | | | | | | Sex = M: attrition (5.0)

| | | | | | | | | AttendedResult = good: Graduated (24.0/4.0)

| | | | | | | | | AttendedResult = very good

| | | | | | | | | | gap = Gap _1: attrition (0.0)

| | | | | | | | | | gap = Gap _3: Graduated (3.0/1.0)

| | | | | | | | | | gap = Gap _4: attrition (0.0)

| | | | | | | | | | gap = Gap _2: attrition (2.0)

| | | | | | | | | | gap = Gap _5: attrition (0.0)

| | | | | | | | | | gap = Gap _0: attrition (0.0)

| | | | | | | | Types of school attended = private

| | | | | | | | | Preparatory completion year = after: attrition (23.0/6.0)

| | | | | | | | | Preparatory completion year = before: Graduated (4.0)

| | | | | | | Department = Computer Science: Graduated (0.0)

| | | | | | | Department = Management: Graduated (6.0)

| | | | | | | Department = Marketing Management

| | | | | | | | Sex = F

| | | | | | | | | gap = Gap _1: attrition (0.0)

| | | | | | | | | gap = Gap _3: Graduated (2.0)

| | | | | | | | | gap = Gap _4: attrition (0.0)

| | | | | | | | | gap = Gap _2: attrition (2.0)

| | | | | | | | | gap = Gap _5: attrition (0.0)

| | | | | | | | | gap = Gap _0: attrition (0.0)

| | | | | | | | Sex = M: Graduated (11.0)

| | | | | | | Department = Tourism and Hospitality Management: Graduated (0.0)

| | | | | | Preparatory attended region = Oromia

| | | | | | | Types of school attended = public

| | | | | | | | AttendedResult = exellent : attrition (1.0)

| | | | | | | | AttendedResult = Satisfactory: attrition (17.0)

| | | | | | | | AttendedResult = good: Graduated (3.0)

| | | | | | | | AttendedResult = very good: attrition (1.0)

| | | | | | | Types of school attended = private

| | | | | | | | Preparatory completion year = after: attrition (3.0)

| | | | | | | | Preparatory completion year = before: Graduated (40.0/9.0)

| | | | | | Preparatory attended region = Amhara

| | | | | | | Sex = F

| | | | | | | | Types of school attended = public: attrition (8.0)

| | | | | | | | Types of school attended = private

| | | | | | | | | Department = Accounting: Graduated (9.0/1.0)

| | | | | | | | | Department = Computer Science: Graduated (0.0)

| | | | | | | | | Department = Management: Graduated (0.0)

| | | | | | | | | Department = Marketing Management

| | | | | | | | | | gap = Gap _1: Graduated (0.0)

| | | | | | | | | | gap = Gap _3: attrition (3.0)

| | | | | | | | | | gap = Gap _4: Graduated (0.0)

| | | | | | | | | | gap = Gap _2: Graduated (4.0)

| | | | | | | | | | gap = Gap _5: Graduated (0.0)

| | | | | | | | | | gap = Gap _0: Graduated (0.0)

| | | | | | | | | Department = Tourism and Hospitality Management: Graduated (0.0)

| | | | | | | Sex = M: attrition (13.0)

| | | | | | Preparatory attended region = SNNP

| | | | | | | Department = Accounting

| | | | | | | | AttendedResult = exellent : Graduated (2.0)

| | | | | | | | AttendedResult = Satisfactory

| | | | | | | | | gap = Gap _1: Graduated (4.0)

| | | | | | | | | gap = Gap _3: attrition (5.0/1.0)

| | | | | | | | | gap = Gap _4: attrition (0.0)

| | | | | | | | | gap = Gap _2

| | | | | | | | | | Sex = F: Graduated (3.0/1.0)

| | | | | | | | | | Sex = M: attrition (2.0)

| | | | | | | | | gap = Gap _5: attrition (0.0)

| | | | | | | | | gap = Gap _0: attrition (0.0)

| | | | | | | | AttendedResult = good: Graduated (0.0)

| | | | | | | | AttendedResult = very good: Graduated (0.0)

| | | | | | | Department = Computer Science: Graduated (0.0)

| | | | | | | Department = Management: attrition (2.0)

| | | | | | | Department = Marketing Management: Graduated (4.0)

| | | | | | | Department = Tourism and Hospitality Management: Graduated (0.0)

| | | | | | Preparatory attended region = Gambella: attrition (2.0)

| | | | | | Preparatory attended region = Tigray

| | | | | | | Department = Accounting

| | | | | | | | AttendedResult = exellent : attrition (0.0)

| | | | | | | | AttendedResult = Satisfactory: attrition (14.0/3.0)

| | | | | | | | AttendedResult = good: Graduated (2.0)

| | | | | | | | AttendedResult = very good: attrition (0.0)

| | | | | | | Department = Computer Science: attrition (0.0)

| | | | | | | Department = Management: Graduated (3.0)

| | | | | | | Department = Marketing Management: Graduated (5.0/1.0)

| | | | | | | Department = Tourism and Hospitality Management: attrition (0.0)

| | | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | | Preparatory attended region = Dire Dawa: Graduated (2.0)

| | | | | | Preparatory attended region = Harari: attrition (1.0)

| | | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | Division = Regular: attrition (18.0)

| | | year takes = three

| | | | Department = Accounting

| | | | | AGE = age_2: Graduated (197.0/4.0)

| | | | | AGE = Age_4: Graduated (64.0/4.0)

| | | | | AGE = Age_1: Graduated (302.0)

| | | | | AGE = Age_3

| | | | | | gap = Gap _1

| | | | | | | Sex = F

| | | | | | | | Types of school attended = public: Graduated (6.0/1.0)

| | | | | | | | Types of school attended = private: attrition (3.0)

| | | | | | | Sex = M: attrition (7.0)

| | | | | | gap = Gap _3: Graduated (36.0)

| | | | | | gap = Gap _4: Graduated (0.0)

| | | | | | gap = Gap _2: Graduated (87.0/9.0)

| | | | | | gap = Gap _5: Graduated (0.0)

| | | | | | gap = Gap _0: Graduated (0.0)

| | | | Department = Computer Science: Graduated (0.0)

| | | | Department = Management: Graduated (40.0/2.0)

| | | | Department = Marketing Management: attrition (42.0)

| | | | Department = Tourism and Hospitality Management: Graduated (0.0)

| | | year takes = five: Graduated (0.0)

| | financial sources = parent sponsored

| | | year takes = four

| | | | Department = Accounting: Graduated (222.0/19.0)

| | | | Department = Computer Science: Graduated (0.0)

| | | | Department = Management: Graduated (50.0)

| | | | Department = Marketing Management: attrition (54.0/4.0)

| | | | Department = Tourism and Hospitality Management: Graduated (22.0/1.0)

| | | year takes = three

| | | | Department = Accounting

| | | | | AGE = age_2

| | | | | | Sex = F

| | | | | | | Preparatory attended region = Addis Ababa

| | | | | | | | AttendedResult = exellent : attrition (10.0/1.0)

| | | | | | | | AttendedResult = Satisfactory

| | | | | | | | | Types of school attended = public: attrition (11.0)

| | | | | | | | | Types of school attended = private: Graduated (4.0)

| | | | | | | | AttendedResult = good: Graduated (2.0)

| | | | | | | | AttendedResult = very good

| | | | | | | | | gap = Gap _1

| | | | | | | | | | Types of school attended = public: Graduated (10.0/3.0)

| | | | | | | | | | Types of school attended = private: attrition (3.0)

| | | | | | | | | gap = Gap _3: Graduated (0.0)

| | | | | | | | | gap = Gap _4: Graduated (0.0)

| | | | | | | | | gap = Gap _2: Graduated (2.0)

| | | | | | | | | gap = Gap _5: Graduated (0.0)

| | | | | | | | | gap = Gap _0: Graduated (0.0)

| | | | | | Preparatory attended region = Oromia: attrition (9.0)

| | | | | | Preparatory attended region = Amhara: attrition (1.0)

| | | | | | Preparatory attended region = SNNP: attrition (2.0)

| | | | | | Preparatory attended region = Gambella: attrition (0.0)

| | | | | | Preparatory attended region = Tigray: attrition (3.0)

| | | | | | Preparatory attended region = Somali: attrition (0.0)

| | | | | | | Preparatory attended region = Dire Dawa: attrition (0.0)

| | | | | | | Preparatory attended region = Harari: attrition (0.0)

| | | | | | | Preparatory attended region = somali: attrition (0.0)

| | | | | | | Preparatory attended region = Benishagule Gumuz: attrition (0.0)

| | | | | | Sex = M: attrition (20.0)

| | | | | AGE = Age_4: attrition (64.0)

| | | | | AGE = Age_1

| | | | | | gap = Gap _1

| | | | | | | Preparatory attended region = Addis Ababa

| | | | | | | | AttendedResult = exellent : attrition (2.0)

| | | | | | | | AttendedResult = Satisfactory

| | | | | | | | | Sex = F

| | | | | | | | | | Types of school attended = public: Graduated (27.0/8.0)

| | | | | | | | | | Types of school attended = private: attrition (4.0/1.0)

| | | | | | | | | Sex = M

| | | | | | | | | | Types of school attended = public: attrition (4.0)

| | | | | | | | | | Types of school attended = private: Graduated (2.0)

| | | | | | | | AttendedResult = good

| | | | | | | | | Sex = F: Graduated (2.0)

| | | | | | | | | Sex = M: attrition (5.0/1.0)

| | | | | | | | AttendedResult = very good

| | | | | | | | | Sex = F: attrition (3.0)

| | | | | | | | | Sex = M: Graduated (2.0)

| | | | | | | Preparatory attended region = Oromia

| | | | | | | | Sex = F: attrition (4.0/1.0)

| | | | | | | | Sex = M: Graduated (4.0)

| | | | | | | Preparatory attended region = Amhara: Graduated (3.0)

| | | | | | | Preparatory attended region = SNNP: attrition (1.0)

| | | | | | | Preparatory attended region = Gambella: Graduated (0.0)

| | | | | | | Preparatory attended region = Tigray: attrition (1.0)

| | | | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | | | Preparatory attended region = Dire Dawa: Graduated (0.0)

| | | | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | | | gap = Gap _3: attrition (0.0)

| | | | | | gap = Gap _4: attrition (0.0)

| | | | | | gap = Gap _2: attrition (1.0)

| | | | | | gap = Gap _5: attrition (0.0)

| | | | | | gap = Gap _0: attrition (10.0)

| | | | | AGE = Age_3: attrition (69.0/5.0)

| | | | Department = Computer Science: attrition (0.0)

| | | | Department = Management: attrition (38.0)

| | | | Department = Marketing Management

| | | | | Sex = F

| | | | | | Preparatory attended region = Addis Ababa

| | | | | | | AttendedResult = exellent

| | | | | | | | gap = Gap _1: attrition (3.0)

| | | | | | | | gap = Gap _3: Graduated (0.0)

| | | | | | | | gap = Gap _4: Graduated (0.0)

| | | | | | | | gap = Gap _2: Graduated (10.0)

| | | | | | | | gap = Gap _5: Graduated (0.0)

| | | | | | | | gap = Gap _0: Graduated (2.0)

| | | | | | | AttendedResult = Satisfactory: Graduated (35.0/3.0)

| | | | | | | AttendedResult = good

| | | | | | | | AGE = age_2: Graduated (13.0)

| | | | | | | | AGE = Age_4: attrition (4.0)

| | | | | | | | AGE = Age_1: Graduated (0.0)

| | | | | | | | AGE = Age_3: Graduated (4.0)

| | | | | | | AttendedResult = very good

| | | | | | | | Types of school attended = public: Graduated (4.0/1.0)

| | | | | | | | Types of school attended = private: attrition (6.0)

| | | | | Preparatory attended region = Oromia

| | | | | | Types of school attended = public

| | | | | | | | AGE = age_2: Graduated (6.0)

| | | | | | | | AGE = Age_4: attrition (2.0)

| | | | | | | | AGE = Age_1: Graduated (0.0)

| | | | | | | | AGE = Age_3: attrition (1.0)

| | | | | | | Types of school attended = private: attrition (7.0)

| | | | | | Preparatory attended region = Amhara: attrition (2.0)

| | | | | | Preparatory attended region = SNNP: Graduated (4.0/1.0)

| | | | | | Preparatory attended region = Gambella: Graduated (0.0)

| | | | | | Preparatory attended region = Tigray: attrition (4.0/1.0)

| | | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | | Preparatory attended region = Dire Dawa: Graduated (0.0)

| | | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | | Sex = M

| | | | | | Types of school attended = public

| | | | | | | AttendedResult = exellent : Graduated (8.0/2.0)

| | | | | | | AttendedResult = Satisfactory: Graduated (10.0/4.0)

| | | | | | | AttendedResult = good: Graduated (1.0)

| | | | | | | AttendedResult = very good: attrition (5.0/1.0)

| | | | | | Types of school attended = private: attrition (25.0/1.0)

| | | | Department = Tourism and Hospitality Management: attrition (0.0)

| | | year takes = five: Graduated (0.0)

| | financial sources = scholar ship: Graduated (14.0)

| background studty = natural

| | year takes = four: attrition (108.0/2.0)

| | year takes = three : attrition (7.0/1.0)

| | year takes = five

| | | financial sources = self sponsered

| | | | Sex = F

| | | | | gap = Gap _1: attrition (0.0)

| | | | | gap = Gap _3: Graduated (3.0)

| | | | | gap = Gap _4: attrition (0.0)

| | | | | gap = Gap _2: attrition (3.0)

| | | | | gap = Gap _5: attrition (0.0)

| | | | | gap = Gap _0: attrition (2.0)

| | | | Sex = M: Graduated (23.0/1.0)

| | | financial sources = parent sponsored: attrition (9.0)

| | | financial sources = scholar ship: Graduated (0.0)

Batch = Batch_3

| Division = Extension

| | Department = Accounting: attrition (450.0/14.0)

| | Department = Computer Science

| | | year takes = four: Graduated (2.0)

| | | year takes = three : attrition (0.0)

| | | year takes = five: attrition (21.0)

| | Department = Management: Graduated (43.0/12.0)

| | Department = Marketing Management

| | | Preparatory completion year = after

| | | | Types of school attended = public

| | | | | financial sources = self sponsered: attrition (16.0/1.0)

| | | | | financial sources = parent sponsored: Graduated (8.0/2.0)

| | | | | financial sources = scholar ship: attrition (0.0)

| | | | Types of school attended = private: attrition (65.0/2.0)

| | | Preparatory completion year = before

| | | | Sex = F

| | | | | AGE = age_2: Graduated (17.0/1.0)

| | | | | AGE = Age_4: Graduated (14.0/1.0)

| | | | | AGE = Age_1: attrition (7.0/1.0)

| | | | | AGE = Age_3: attrition (2.0)

| | | | Sex = M: attrition (40.0/6.0)

| | Department = Tourism and Hospitality Management: attrition (0.0)

| Division = Regular

| | financial sources = self sponsered

| | | Department = Accounting: Graduated (1125.0/1.0)

| | | Department = Computer Science: Graduated (49.0/2.0)

| | | Department = Management: Graduated (56.0/3.0)

| | | Department = Marketing Management: attrition (19.0)

| | | Department = Tourism and Hospitality Management: attrition (3.0)

| | financial sources = parent sponsored

| | | Department = Accounting: attrition (504.0/51.0)

| | | Department = Computer Science

| | | | AttendedResult = exellent : attrition (13.0)

| | | | AttendedResult = Satisfactory

| | | | | Sex = F: attrition (3.0)

| | | | | Sex = M

| | | | | | Types of school attended = public: Graduated (9.0/1.0)

| | | | | | Types of school attended = private

| | | | | | | Preparatory completion year = after: attrition (5.0)

| | | | | | | Preparatory completion year = before: Graduated (2.0)

| | | | AttendedResult = good: attrition (9.0/4.0)

| | | | AttendedResult = very good

| | | | | Sex = F: Graduated (3.0/1.0)

| | | | | Sex = M: attrition (11.0)

| | | Department = Management: attrition (26.0)

| | | Department = Marketing Management

| | | | AGE = age_2: Graduated (138.0/24.0)

| | | | AGE = Age_4

| | | | | Preparatory completion year = after: attrition (28.0)

| | | | | Preparatory completion year = before

| | | | | | gap = Gap _1: attrition (0.0)

| | | | | | gap = Gap _3: attrition (2.0)

| | | | | | gap = Gap _4

| | | | | | | Preparatory attended region = Addis Ababa: attrition (3.0)

| | | | | | | Preparatory attended region = Oromia: attrition (0.0)

| | | | | | | Preparatory attended region = Amhara: Graduated (3.0)

| | | | | | | Preparatory attended region = SNNP: attrition (0.0)

| | | | | | | Preparatory attended region = Gambella: attrition (0.0)

| | | | | | | Preparatory attended region = Tigray: attrition (1.0)

| | | | | | | Preparatory attended region = Somali: attrition (0.0)

| | | | | | | Preparatory attended region = Dire Dawa: attrition (0.0)

| | | | | | | Preparatory attended region = Harari: attrition (0.0)

| | | | | | | Preparatory attended region = somali: attrition (0.0)

| | | | | | | Preparatory attended region = Benishagule Gumuz: attrition (0.0)

| | | | | | gap = Gap _2: attrition (0.0)

| | | | | | gap = Gap _5: Graduated (3.0)

| | | | | | gap = Gap _0: attrition (0.0)

| | | | AGE = Age_1: attrition (12.0)

| | | | AGE = Age_3

| | | | | gap = Gap _1: attrition (7.0)

| | | | | gap = Gap _3

| | | | | | Preparatory completion year = after: attrition (5.0)

| | | | | | Preparatory completion year = before: Graduated (3.0)

| | | | | gap = Gap _4: Graduated (0.0)

| | | | | gap = Gap _2: Graduated (41.0/8.0)

| | | | | gap = Gap _5: Graduated (0.0)

| | | | | gap = Gap _0: attrition (7.0)

| | | Department = Tourism and Hospitality Management

| | | | Types of school attended = public: Graduated (29.0/2.0)

| | | | Types of school attended = private

| | | | | AGE = age_2

| | | | | | Sex = F

| | | | | | | Preparatory attended region = Addis Ababa: Graduated (7.0)

| | | | | | | Preparatory attended region = Oromia: attrition (2.0)

| | | | | | | Preparatory attended region = Amhara: Graduated (3.0)

| | | | | | | Preparatory attended region = SNNP: Graduated (0.0)

| | | | | | | Preparatory attended region = Gambella: Graduated (0.0)

| | | | | | | Preparatory attended region = Tigray: Graduated (0.0)

| | | | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | | | Preparatory attended region = Dire Dawa: Graduated (0.0)

| | | | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | | | Sex = M: attrition (3.0)

| | | | | AGE = Age_4: attrition (9.0)

| | | | | AGE = Age_1: attrition (3.0)

| | | | | AGE = Age_3: Graduated (7.0/3.0)

| | financial sources = scholar ship: Graduated (12.0)

Batch = Batch_4

| Division = Extension: attrition (219.0)

| Division = Regular

| | Department = Accounting

| | | AGE = age_2

| | | | Preparatory completion year = after

| | | | | Sex = F

| | | | | | Preparatory attended region = Addis Ababa

| | | | | | | gap = Gap _1: Graduated (54.0/8.0)

| | | | | | | gap = Gap _3: Graduated (0.0)

| | | | | | | gap = Gap _4: Graduated (0.0)

| | | | | | | gap = Gap _2: Graduated (24.0/8.0)

| | | | | | | gap = Gap _5: Graduated (0.0)

| | | | | | | gap = Gap _0: attrition (11.0/3.0)

| | | | | | Preparatory attended region = Oromia

| | | | | | | Types of school attended = public: Graduated (11.0/3.0)

| | | | | | | Types of school attended = private: attrition (3.0)

| | | | | | Preparatory attended region = Amhara: Graduated (9.0)

| | | | | | Preparatory attended region = SNNP: Graduated (5.0/1.0)

| | | | | | Preparatory attended region = Gambella: Graduated (0.0)

| | | | | | Preparatory attended region = Tigray: attrition (7.0/2.0)

| | | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | | Preparatory attended region = Dire Dawa: Graduated (0.0)

| | | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | | Sex = M

| | | | | | Types of school attended = public

| | | | | | | AttendedResult = exellent : Graduated (9.0/2.0)

| | | | | | | AttendedResult = Satisfactory: attrition (5.0/1.0)

| | | | | | | AttendedResult = good: attrition (7.0/3.0)

| | | | | | | AttendedResult = very good: Graduated (12.0/3.0)

| | | | | | Types of school attended = private: attrition (26.0/4.0)

| | | | Preparatory completion year = before

| | | | | Types of school attended = public: attrition (6.0/2.0)

| | | | | Types of school attended = private: Graduated (82.0/8.0)

| | | AGE = Age_4

| | | | Preparatory completion year = after: attrition (63.0/3.0)

| | | | Preparatory completion year = before

| | | | | emp info = yes

| | | | | | Sex = F: attrition (8.0/1.0)

| | | | | | Sex = M

| | | | | | | Types of school attended = public: attrition (3.0/1.0)

| | | | | | | Types of school attended = private: Graduated (5.0)

| | | | | emp info = no: Graduated (7.0)

| | | AGE = Age_1

| | | | Types of school attended = public: Graduated (382.0/36.0)

| | | | Types of school attended = private

| | | | | Preparatory attended region = Addis Ababa: Graduated (150.0/44.0)

| | | | | Preparatory attended region = Oromia

| | | | | | gap = Gap _1: Graduated (43.0/3.0)

| | | | | | gap = Gap _3: Graduated (0.0)

| | | | | | gap = Gap _4: Graduated (0.0)

| | | | | | gap = Gap _2: Graduated (0.0)

| | | | | | gap = Gap _5: Graduated (0.0)

| | | | | | gap = Gap _0: attrition (7.0/2.0)

| | | | | Preparatory attended region = Amhara

| | | | | | AttendedResult = exellent : Graduated (4.0/1.0)

| | | | | | AttendedResult = Satisfactory: Graduated (7.0/1.0)

| | | | | | AttendedResult = good: attrition (4.0)

| | | | | | AttendedResult = very good: Graduated (0.0)

| | | | | Preparatory attended region = SNNP

| | | | | | gap = Gap _1: Graduated (4.0)

| | | | | | gap = Gap _3: attrition (0.0)

| | | | | | gap = Gap _4: attrition (0.0)

| | | | | | gap = Gap _2: attrition (0.0)

| | | | | | gap = Gap _5: attrition (0.0)

| | | | | | gap = Gap _0: attrition (6.0/1.0)

| | | | | Preparatory attended region = Gambella

| | | | | | AttendedResult = exellent : Graduated (2.0)

| | | | | | AttendedResult = Satisfactory: Graduated (1.0)

| | | | | | AttendedResult = good: Graduated (2.0)

| | | | | | AttendedResult = very good: attrition (4.0)

| | | | | Preparatory attended region = Tigray: Graduated (0.0)

| | | | | Preparatory attended region = Somali: attrition (5.0)

| | | | | Preparatory attended region = Dire Dawa: attrition (3.0)

| | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | AGE = Age_3

| | | | Preparatory completion year = after

| | | | | Preparatory attended region = Addis Ababa

| | | | | | gap = Gap _1: attrition (4.0/1.0)

| | | | | | gap = Gap _3

| | | | | | | AttendedResult = exellent : Graduated (5.0)

| | | | | | | AttendedResult = Satisfactory: attrition (2.0)

| | | | | | | AttendedResult = good: Graduated (0.0)

| | | | | | | AttendedResult = very good: attrition (1.0)

| | | | | | gap = Gap _4: Graduated (0.0)

| | | | | | gap = Gap _2: Graduated (79.0/21.0)

| | | | | | gap = Gap _5: Graduated (0.0)

| | | | | | gap = Gap _0

| | | | | | | AttendedResult = exellent : attrition (0.0)

| | | | | | | AttendedResult = Satisfactory: attrition (3.0)

| | | | | | | AttendedResult = good

| | | | | | | | Sex = F: Graduated (2.0)

| | | | | | | | Sex = M: attrition (2.0)

| | | | | | | AttendedResult = very good: attrition (0.0)

| | | | | Preparatory attended region = Oromia: attrition (3.0/1.0)

| | | | | Preparatory attended region = Amhara: attrition (8.0/1.0)

| | | | | Preparatory attended region = SNNP: attrition (4.0/1.0)

| | | | | Preparatory attended region = Gambella: Graduated (0.0)

| | | | | Preparatory attended region = Tigray

| | | | | | AttendedResult = exellent : attrition (5.0)

| | | | | | AttendedResult = Satisfactory: Graduated (3.0)

| | | | | | AttendedResult = good: attrition (0.0)

| | | | | | AttendedResult = very good: Graduated (1.0)

| | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | Preparatory attended region = Dire Dawa: Graduated (0.0)

| | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | Preparatory completion year = before: Graduated (74.0)

| | Department = Computer Science: attrition (85.0)

| | Department = Management

| | | financial sources = self sponsered: Graduated (49.0/1.0)

| | | financial sources = parent sponsored: attrition (31.0/1.0)

| | | financial sources = scholar ship: Graduated (0.0)

| | Department = Marketing Management

| | | financial sources = self sponsered: attrition (29.0)

| | | financial sources = parent sponsored

| | | | AGE = age_2

| | | | | gap = Gap _1

| | | | | | Preparatory attended region = Addis Ababa

| | | | | | | Sex = F: Graduated (9.0)

| | | | | | | Sex = M: attrition (3.0/1.0)

| | | | | | Preparatory attended region = Oromia: Graduated (0.0)

| | | | | | Preparatory attended region = Amhara: Graduated (0.0)

| | | | | | Preparatory attended region = SNNP: attrition (3.0)

| | | | | | Preparatory attended region = Gambella: Graduated (0.0)

| | | | | | Preparatory attended region = Tigray: Graduated (0.0)

| | | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | | Preparatory attended region = Dire Dawa: Graduated (0.0)

| | | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | | gap = Gap _3: Graduated (0.0)

| | | | | gap = Gap _4: attrition (2.0)

| | | | | gap = Gap _2

| | | | | | Preparatory attended region = Addis Ababa

| | | | | | | AttendedResult = exellent : attrition (6.0/2.0)

| | | | | | | AttendedResult = Satisfactory

| | | | | | | | Types of school attended = public: Graduated (2.0)

| | | | | | | | Types of school attended = private: attrition (2.0)

| | | | | | | AttendedResult = good

| | | | | | | | Types of school attended = public: Graduated (2.0)

| | | | | | | | Types of school attended = private: attrition (9.0/3.0)

| | | | | | | AttendedResult = very good: Graduated (5.0)

| | | | | | Preparatory attended region = Oromia: attrition (3.0)

| | | | | | Preparatory attended region = Amhara: attrition (3.0/1.0)

| | | | | | Preparatory attended region = SNNP: Graduated (5.0)

| | | | | | Preparatory attended region = Gambella: Graduated (0.0)

| | | | | | Preparatory attended region = Tigray: Graduated (0.0)

| | | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | | Preparatory attended region = Dire Dawa: Graduated (0.0)

| | | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | | gap = Gap _5: Graduated (0.0)

| | | | | gap = Gap _0: Graduated (55.0/6.0)

| | | | AGE = Age_4

| | | | | Preparatory completion year = after: attrition (70.0/11.0)

| | | | | Preparatory completion year = before

| | | | | | Preparatory attended region = Addis Ababa: Graduated (5.0/1.0)

| | | | | | Preparatory attended region = Oromia: Graduated (2.0)

| | | | | | Preparatory attended region = Amhara: attrition (4.0/1.0)

| | | | | | Preparatory attended region = SNNP: Graduated (0.0)

| | | | | | Preparatory attended region = Gambella: Graduated (0.0)

| | | | | | Preparatory attended region = Tigray: attrition (1.0)

| | | | | | Preparatory attended region = Somali: Graduated (0.0)

| | | | | | Preparatory attended region = Dire Dawa: Graduated (0.0)

| | | | | | Preparatory attended region = Harari: Graduated (0.0)

| | | | | | Preparatory attended region = somali: Graduated (0.0)

| | | | | | Preparatory attended region = Benishagule Gumuz: Graduated (0.0)

| | | | AGE = Age_1: attrition (18.0)

| | | | AGE = Age_3

| | | | | gap = Gap _1: attrition (8.0)

| | | | | gap = Gap _3: attrition (10.0/2.0)

| | | | | gap = Gap _4: attrition (0.0)

| | | | | gap = Gap _2

| | | | | | Types of school attended = public: attrition (16.0/4.0)

| | | | | | Types of school attended = private: Graduated (31.0/5.0)

| | | | | gap = Gap _5: attrition (0.0)

| | | | | gap = Gap _0: attrition (29.0/2.0)

| | | financial sources = scholar ship: attrition (0.0)

| | Department = Tourism and Hospitality Management

| | | gap = Gap _1: attrition (4.0/1.0)

| | | gap = Gap _3: attrition (2.0/1.0)

| | | gap = Gap _4: Graduated (2.0)

| | | gap = Gap _2

| | | | AttendedResult = exellent : Graduated (3.0)

| | | | AttendedResult = Satisfactory: Graduated (3.0)

| | | | AttendedResult = good: attrition (0.0)

| | | | AttendedResult = very good: attrition (7.0)

| | | gap = Gap _5: attrition (1.0)

| | | gap = Gap _0

| | | | AGE = age_2: Graduated (0.0)

| | | | AGE = Age_4: Graduated (28.0)

| | | | AGE = Age_1: Graduated (0.0)

| | | | AGE = Age_3: attrition (4.0)

Batch = Batch_5: attrition (854.0)

Batch = Batch_6: attrition (607.0)

Batch = Batch_1

| financial sources = self sponsered

| | background studty = social

| | | Division = Extension

| | | | Preparatory completion year = after

| | | | | Types of school attended = public: Graduated (18.0/1.0)

| | | | | Types of school attended = private: attrition (4.0)

| | | | Preparatory completion year = before: Graduated (155.0/12.0)

| | | Division = Regular: Graduated (146.0)

| | background studty = natural

| | | gap = Gap _1: Graduated (3.0)

| | | gap = Gap _3: Graduated (0.0)

| | | gap = Gap _4: attrition (4.0)

| | | gap = Gap _2

| | | | Preparatory attended region = Addis Ababa: Graduated (1.0)

| | | | Preparatory attended region = Oromia: attrition (5.0)

| | | | Preparatory attended region = Amhara: Graduated (2.0)

| | | | Preparatory attended region = SNNP: attrition (0.0)

| | | | Preparatory attended region = Gambella: attrition (0.0)

| | | | Preparatory attended region = Tigray: attrition (0.0)

| | | | Preparatory attended region = Somali: attrition (0.0)

| | | | Preparatory attended region = Dire Dawa: attrition (0.0)

| | | | Preparatory attended region = Harari: attrition (0.0)

| | | | Preparatory attended region = somali: attrition (0.0)

| | | | Preparatory attended region = Benishagule Gumuz: attrition (0.0)

| | | gap = Gap _5: Graduated (7.0)

| | | gap = Gap _0: Graduated (0.0)

| financial sources = parent sponsored

| | Division = Extension: Graduated (44.0)

| | Division = Regular

| | | Department = Accounting

| | | | background studty = social: attrition (63.0/4.0)

| | | | background studty = natural: Graduated (2.0)

| | | Department = Computer Science: attrition (0.0)

| | | Department = Management: attrition (3.0)

| | | Department = Marketing Management: Graduated (37.0/2.0)

| | | Department = Tourism and Hospitality Management: attrition (0.0)

| financial sources = scholar ship: Graduated (14.0)

Annex: 4.2 J48 algorithm experiments using of 10 fold cross validation test option

```
Correctly Classified Instances        8455               91.3955 %
Incorrectly Classified Instances       796                8.6045 %
Kappa statistic                          0.8279
Mean absolute error                      0.1215
Root mean squared error                  0.264
Relative absolute error                 24.2997 %
Root relative squared error             52.7946 %
Total Number of Instances             9251

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.894     0.066      0.932     0.894     0.912      0.959    attrition
                 0.934     0.106      0.897     0.934     0.915      0.959    Graduated
Weighted Avg.    0.914     0.086      0.915     0.914     0.914      0.959

=== Confusion Matrix ===

    a    b   <-- classified as
 4147  494 |    a = attrition
  302 4308 |    b = Graduated
```

Annex:4.3 C# sources code for predicting student status

```
using System;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Drawing;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Windows.Forms;

namespace WindowsFormsApplication1
{
publicpartialclassForm1 : Form
    {
public Form1()
        {
InitializeComponent();
this.outP.Visible = false;
this.divisonbox.Enabled = false;
this.departementbox.Enabled = false;
this.gapselection.Enabled = false;
this.departementbox.Enabled = false;
this.comboback.Enabled = false;
```

```
this.attenddbox.Enabled = false;
        }

privatevoid label1_Click(object sender, EventArgs e)
        {

        }



privatevoid Form1_Load(object sender, EventArgs e)
        {

        }

privatevoid label1_Click_1(object sender, EventArgs e)
        {

        }
privatevoid label5_Click(object sender, EventArgs e)
        {

        }

privatevoid comboBox2_SelectedIndexChanged(object sender, EventArgs e)
        {
if (this.Finresourcebox.Text == "Scolarship")

            {
this.divisonbox.Enabled = false;
this.departementbox.Enabled = false;
this.gapselection.Enabled = false;
this.comboback.Enabled = false;
this.attenddbox.Enabled = false;
            }
elseif (this.Finresourcebox.Text == "Self-sponsored")
            {
this.divisonbox.Enabled = true;
this.departementbox.Enabled = false;
this.gapselection.Enabled = false;
this.comboback.Enabled = false;
this.attenddbox.Enabled = false;
            }
elseif (this.Finresourcebox.Text == "Parent-sponsored")
            {
this.divisonbox.Enabled = true;
this.departementbox.Enabled = false;
this.gapselection.Enabled = false;
this.comboback.Enabled = false;
this.attenddbox.Enabled = false;
            }
else
            {
this.divisonbox.Enabled = true;
this.departementbox.Enabled = true;
this.gapselection.Enabled = true;
this.comboback.Enabled = true;
```

```
this.attenddbox.Enabled = true;
            }
        }

privatevoid comboBox3_SelectedIndexChanged(object sender, EventArgs e)
        {
if (this.departementbox.Text == "Accounting"&&this.Finresourcebox.Text == "Parent-
sponsored"&&this.divisonbox.Text == "Regular")
            {
this.comboback.Enabled = true;
            }
else
            {
this.comboback.Enabled = false;
            }
        }

privatevoid label4_Click(object sender, EventArgs e)
        {

        }

privatevoid comboBox1_SelectedIndexChanged(object sender, EventArgs e)
        {
if (this.Finresourcebox.Text == "Self-sponsored"&&this.divisonbox.Text == "Regular")
            {
this.divisonbox.Enabled = true;
this.departementbox.Enabled = false;
this.gapselection.Enabled = false;
this.comboback.Enabled = false;
this.attenddbox.Enabled = false;

            }
elseif (this.Finresourcebox.Text == "Self-sponsored"&&this.divisonbox.Text ==
"Extension")
                        {
this.divisonbox.Enabled = false;
this.departementbox.Enabled = false;
this.gapselection.Enabled = true;
this.comboback.Enabled = false;
this.attenddbox.Enabled = false;

                        }
elseif (this.Finresourcebox.Text == "Parent-sponsored"&&this.divisonbox.Text ==
"Extension")
                        {
this.divisonbox.Enabled = true;
this.departementbox.Enabled = false;
this.gapselection.Enabled = false;
this.comboback.Enabled = false;
this.attenddbox.Enabled = false;

                        }
elseif (this.Finresourcebox.Text == "Parent-sponsored"&&this.divisonbox.Text ==
"Regular")
                        {
this.divisonbox.Enabled = true;
```

```
this.departementbox.Enabled = true;
this.gapselection.Enabled = false;
this.comboback.Enabled = false;
this.attenddbox.Enabled = false;

                        }
        }

privatevoid textBox1_TextChanged(object sender, EventArgs e)
        {

        }

privatevoid label3_Click(object sender, EventArgs e)
        {

        }

privatevoid label2_Click(object sender, EventArgs e)
        {

        }

privatevoid comboBox4_SelectedIndexChanged(object sender, EventArgs e)
        {

        }

privatevoid Predictbtn_Click(object sender, EventArgs e)
        {
String Divison=divisonbox.Text;
String Finresource = Finresourcebox.Text;
String departement = departementbox.Text;
String attendd = attenddbox.Text;

this.outP.Visible = true;
if (Finresource == "Scolarship")
          {
              lbloutput.Text = "Graduated";
          }
elseif (Finresource == "Self-sponsored")
          {
if (Divison == "Regular")
{ lbloutput.Text = "Graduated"; }
elseif (Divison == "Extension")
              {
if (gapselection.Text == "Before")
                  {
                      lbloutput.Text = "Graduated";
                  }
elseif (gapselection.Text == "After")
                  {
if (attendd == "Private")
                      {
                          lbloutput.Text = "Attrition";
                      }
elseif (attendd == "Public")
```

```
                    {
                        lbloutput.Text = "Graduated";
                    }
else
{ lbloutput.Text = "Invalid Attended institution"; }
                }
else
{ lbloutput.Text = "Invalid Selection"; }
            }

        }
elseif (Finresource == "Parent-sponsored")
        {
if (Divison == "Extension")
{ lbloutput.Text = "Graduated";}
elseif (Divison == "Regular")
            {
if (departement == "Computer science")
{ lbloutput.Text = "Attrition"; }
elseif (departement == "Management")
{ lbloutput.Text = "Attrition"; }
elseif (departement == "Marketing management")
{ lbloutput.Text = "Graduated"; }
elseif (departement == "Hotel and tourism")
{ lbloutput.Text = "Attrition"; }
elseif (departement == "Accounting")
                {
if (comboback.Text == "Social")
{ lbloutput.Text = "Attrition"; }
elseif (comboback.Text == "Natural")
{ lbloutput.Text = "Graduated"; }
else
{ lbloutput.Text = "Invalid Selection"; }
                }
else
{ lbloutput.Text = "Invalid Selection"; }
            }
else
            {lbloutput.Text = "Invalid Selection"; }
        }
else
{ lbloutput.Text = "Invalid Selection"; }


    }

privatevoid lbloutput_Click(object sender, EventArgs e)
    {

    }

privatevoid label6_Click(object sender, EventArgs e)
    {

    }

privatevoid gapselection_SelectedIndexChanged(object sender, EventArgs e)
```

```
        {
if (this.gapselection.Text == "Before")
            {
this.divisonbox.Enabled = true;
this.departementbox.Enabled = false;
this.gapselection.Enabled = true;
this.departementbox.Enabled = false;
this.comboback.Enabled = false;
this.attenddbox.Enabled = false;

            }
if (this.gapselection.Text == "After")
            {
this.divisonbox.Enabled = true;
this.departementbox.Enabled = false;
this.gapselection.Enabled = true;
this.departementbox.Enabled = false;
this.comboback.Enabled = false;
this.attenddbox.Enabled = true;

            }
        }

    }
}
```

Annex 4.4 questioner form

St. Mary's university school of graduate studies, faculty of informatics department of computer science

Direction:

Dear respondents,

This questionnaire is designed to check for the validity of the prototype which I designed for predicting students' academic status. The truthfulness of your responses will donate much to the validity of this prototype. So you are requested to be honest to give accurate information. No need to write your name on any part of this questioner. Thank you for your cooperation

Gender of the respondents 1 female 4 male

Questionnaires	Strongly Agree (5)	Agree (4)	Undecided (3)	Disagree (2)	Strongly Disagree (1)
Efficiency: • In terms of time. • In terms of accuracy .					
Effectiveness: • In terms of the output result. • In terms of performance					
Easy to understand • In terms of the feature of user interface. • In terms of platform					
Easy to Remember: • In terms remembering the way to use the prototype					