



Use of Data Mining For Determining Higher Education
Students' Performance

A thesis submitted

By

Solome Samson

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

August, 2019

ACCEPTANCE

Use of Data Mining For Determining Higher Education

Students' Performance

By

Solome Samson Altaye

**Accepted by the 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

Date

Dr. Temtim Assefa

External Examiner

Signature

Date

August 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.

Solome Samson Altaye

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

August, 2019

ACKNOWLEDGEMENT

First I would like to thank God for giving me strength, hope and for giving me the ability to be where I am.

I would like to express my gratitude and heartfelt thanks to my advisor, Dr. Million Meshesha for the guidance, support, and direction showed me throughout my thesis writing. This thesis would not have been completed successfully without his continuous, constructive comments and critical readings of the study.

I would like to express my sincerest gratitude and heartfelt thanks to Dr. Getahun Semone for allowing me to carry out this research using the required data and for his support. I forward my sincere gratitude to St. Mary's University staff for allowing me to carry out this research using the required dataset from the St. Mary's University databases and for their general comments on the nature of the dataset.

Finally, my appreciation and thankfulness also go to my family who unrestricted support and encouragement since the time of my admission to the completion of my study. To all others who helped me I would like to sincerely recognize and appreciate the efforts of people who are not mentioned have contributed greatly in one way or the other towards the successful completion of this study, I am equally grateful.

Table of Contents

ACRONYMS	x
ABSTRACT	xi
CHAPTER ONE	5
INTRODUCTION	5
1.1. Background.....	5
1.2. Statement of the Problem	6
1.3. Objective of the Study	8
1.3.1. General Objective.....	8
1.3.2. Specific Objectives.....	8
1.4 Scope and Limitation of the Study	9
1.5 Significance of the study	9
1.6 Organization of the Study.....	11
CHAPTER TWO	12
LITERATURE REVIEW	12
2.1. Data Mining and Knowledge Discovery	12
2.1.1. Educational Data Mining	13
2.2. Data Mining Tasks.....	14
2.3. Predictive modeling using classification algorithm.....	16
2.3.1. Decision Trees.....	17
2.3.1.1. J48 Decision Tree Algorithm	19
2.3.2. Rule-Based Classifiers	21
2.3.2.1. PART Rule induction	21
2.3.3. Bayesian Classification	22
2.3.3.1. Naïve Bayesian	23
2.3.4. Regression	25
2.3.4.1. Logistic Regression	25
2.3.5. Support Vector Machines.....	25
2.3.5.1. Sequential Minimal Optimization	26
2.3.6. Neural Networks	26
2.3.6.1. Multilayer Perception	27

2.4.	Data Mining Tool Selection	28
2.5.	Related works	29
CHAPTER THREE		35
Methodology		35
3.1.	Research Design	35
3.1.1.	Understanding of the problem domain.....	37
3.1.2.	Understanding of the data	37
3.1.3.	Preparation of the data.....	37
3.1.4.	Data mining	38
3.1.5.	Evaluation of the discovered knowledge	38
3.1.6.	Use of the discovered knowledge.....	39
3.2.	Data Mining Process Models.....	39
3.2.1.	KDD Process model	40
3.2.2.	The CRISP-DM process model.....	41
3.2.3.	Hybrid Models.....	43
3.2.4.	Comparison of Data Mining process Models.....	45
3.3.	Model Evaluation	46
3.4.	User Acceptance Testing	49
3.4.1.	Effectiveness	49
3.4.2.	Efficiency	49
3.4.3.	Engaging.....	49
3.4.4.	Easy to Learn.....	49
CHAPTER FOUR.....		50
PROBLEM UNDERSTANDING AND DATA PREPARATION.....		50
4.1.	Understanding of the problem domain	50
4.1.1.	Overview of the organization, St. Mary's University	52
4.1.2.	Criteria for student selection	54
4.2.	Understanding of the Data.....	57
4.2.1.	Collect initial data	57
4.2.2.	Description of the collected data.....	57
4.3.	Preparation of data.....	59
4.3.1.	Data Extracted from Database.....	59
4.3.2.	Data Cleaning.....	62

4.3.2.1. Missing Values	62
4.3.3. Data Transformation	63
4.3.4. Setting the Attribute Class.....	64
4.3.5. Data Preparation for WEKA Software.....	65
CHAPTER FIVE	66
EXPERIMENTATION AND EVALUATION.....	66
5.1. Model Building.....	66
5.1.1. Selecting the modeling technique	66
5.1.1.1. WEKA Interface	67
5.1.1.2. Balancing the Dataset	67
5.1.1.3. Attribute Ordering	68
5.2. Experimental Setup.....	69
5.2.1. Classification Using J48 Decision Tree	70
5.2.2. Classification Using PART Rule Induction	71
5.2.3. Classification Using Naïve Bayes	71
5.2.4. Classification Using Logistic Regression	72
5.2.5. Classification Using Sequential Minimal Optimization.....	72
5.2.6. Classification Using Multilayer Perception.....	73
5.3. Experimental Result	73
5.4. Comparison of Classification Models	87
5.5. Generated Rules from Decision Trees.....	89
5.6. Discussion on Major Findings.....	92
5.7. Use of Knowledge	93
5.7.1. Users Evaluation Result	94
CHAPTER SIX	97
CONCLUSION AND RECOMMENDATION	97
6.1. Conclusion	97
6.2. Recommendations	98
REFERENCE.....	100
ANNEXES	105

LIST OF TABLES

Table 2.1: Comparison of the three open sources Data Mining tools [35]	29
Table 3.1: CRISP-DM phases and tasks [49]	43
Table 3.2: Comparison of DM & KD process models and methodologies [49]	45
Table 3.3: Two dimensional Confusion Matrix [55].	46
Table 3.4: Performance measure of ROC Area	48
Table 4.1: Grading system of undergraduate Students at St Mary’s University [63]	54
Table 4.2: Description of the selected attributes from St Mary’s University dataset	58
Table 4.3: Attributes extracted from database	60
Table 4.5: Attributes with missing values and replaced by mode.....	62
Table 4.6: Discretized age attribute	63
Table 4.7: Discretized English/Mathematics EHEEE result attribute.....	63
Table 4.8: A discretized attribute Status	63
Table 5.1: Some of J48 decision tree algorithm parameters with their values.....	70
Table 5.2: Performance result of J48 Decision tree with 10-fold cross validation	74
Table 5.3: Performance result of J48 Decision tree with percentage split (66%)	74
Table 5.4: Performance results of PART Rule Induction with 10 cross validation	76
Table 5.5: Performance result of PART Rule Induction with percentage split.....	77
Table 5.6: Performance result of Naïve Bayes with default 10-fold cross validation.....	78
Table 5.7: Performance result of Naïve Bayes algorithm with default percentage split.....	79
Table 5.8: Performance result of Logistic Regression with default 10-fold cross validation	80
Table 5.9: Performance result of Logistic Regression algorithm with default percentage split	81
Table 5.10: Performance result of SMO with default 10-fold cross validation	82
Table 5.11: Performance result of SMO algorithm with default percentage split	83
Table 5.12: Performance result of Multilayer Perception with default 10-fold cross validation	85
Table 5.13: Performance result of Multilayer Perception algorithm with percentage split	86
Table 5.14: Performance Comparison of the selected models	87
Table 5.15: Confusion Matrix for J48 algorithm using resampling with percentage split.....	88
Table 5.16: Summary of users’ response on the prototype	95

LIST OF FIGURES

Figure 2.1: Areas in relation with EDM [15].....	13
Figure 2.2: Predictive and Descriptive Data mining models [10]	15
Figure 2.3: Decision tree diagram [22].....	18
Figure 2.4: Bayesian graphical model	22
Figure 2.5: A neural network with one hidden layer and W_{xy} is weight from node x to node y.	27
Figure 3.1: The six-step KDP model [26].....	36
Figure 3.2: KDD Process model [26]	40
Figure 3.3: CRISP DM knowledge discovery process model [52].....	41
Figure 3.4: ROC (Receiver Operating Characteristic) [26]	48
Figure 4.1: List of selected attributes in WEKA 3.9.2 explorer window	64
Figure 4.2: Sample CSV format data sets prepared for WEKA	65
Figure 5.1: WEKA interface	67
Figure 5.2: Side by side view of the class (left side) Original data; (right side) balanced data....	68
Figure 5.3: output of attribute ranking with information gain	69
Figure 5.4: Sample result of the prototype.....	94

ACRONYMS

ARFF: Attribute-Relation File Format

CGPA: Cumulative Grade Point Average

CRISP-DM: Cross Industry Standard Process for Data Mining

CSV: Comma Separated Value

DM & KD: Data Mining and Knowledge Discovery

EDM: Educational Data Mining

GNU: General Public License

GPA: Grade Point Average

GUI: Graphical User Interface

HERQA: Higher Education and Relevance Quality Agency

HRM: Human Resource Management

KDD: Knowledge Discovery in Databases

ML: Machine Learning

ROC: Receiver Operating Characteristic

SMU: St. Mary's University

SQL: Structured Query Language

WEKA: Waikato Environment for Knowledge Analysis

ABSTRACT

Current advancements in communication technologies and database technologies have made it easy for organizations to collect, store and manipulate massive amounts of data. Identifying students' behavior in university is a great concern to the higher education managements. An appropriate decision can be made by effectively analyzing and managing the growing volume of data. The general objective of this study is to construct a predictive model that determines the higher education students' performance by applying data mining techniques.

The study followed the six step Hybrid methodology of Knowledge Discovery Process model such as understanding of the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge and use of the discovered knowledge used to achieve the goal. The study tries to understand factors affecting higher education student performance based on the data collected from St. Mary's University from the years 2006 up to 2009. After data preparation using data cleaning, classification algorithms such as J48 Decision tree, PART Rule induction, Naïve Bayes, Logistic regression, Support Vector Machines and Multilayer Perception Neural Network were used for all experiments due to their popularity in recent related works. The study used a dataset containing 11550 instances, 21 attributes and one outcome variable to run the experiments. WEKA 3.9.2 open source software was used as a data mining tool to implement the experiments. The study also used a 10-fold cross validation and 66% split test modes for splitting the data into training and test dataset.

The result of the study showed that J48 Decision tree algorithm has registered best classification accuracy of 97.84%. The results obtained in this study are interesting and encouraging to design a model that predicts higher institution students' performance. Previous study field, number of common course per semester, total course per semester, year, financial source, number of supportive course per semester, were identified as the major factors affecting the student performance.

In this study, an attempt was made to show the use of knowledge extracted by data mining. In the future, we recommend an automatic integration of data mining with a knowledge system so as to design an intelligent system.

KEYWORD: Data Mining, Educational Data Mining, Classification algorithm

CHAPTER ONE

INTRODUCTION

1.1. Background

Every individual in the world needs knowledge to overcome obstacles in their day to day life. There are different mechanisms to get knowledge in specific subject. One of the main mechanisms is formal education. This in turn has created a major impact in the world. The educational activity develops widely and becoming one of the most essential in industries. Education is vital for one country social, economic and political development. The development of a country is insured by students who complete their studies and successfully graduate from higher learning institutions [1].

In all higher education environment students are expected to perform well and achieve good results. Accordingly, higher educational institute are trying to enable students to make the most of their higher education by providing an education of the highest quality and develop learning experience. To this effect, the problems which affect students' performance of higher learning should be studied and need an appropriate solution for corrective action [2].

Nowadays the amount of data stored in educational database is increasing rapidly. These databases contain hidden knowledge for improvement of students' performance. Getting the hidden and unused prized information in data is impossible for individuals or groups with limited resources to find a solution for student's performance from large educational data.

Student accessing test in educational area are common activities to evaluate the student performance. Evaluating students' performance is a difficult when more factors involve in determine the student performance. Data mining is one of the approaches, which can provide an effective assistance in revealing complex relationships with performance.

Data mining can be used in educational field to enhance our understanding of learning process and to focus on identifying, extracting and evaluating variables related to the learning process of

students [3]. Data mining contains tools and algorithms for the extraction of hidden knowledge from large amounts of data. There are other terms similar or a little different meaning to data mining, such as knowledge mining from databases, knowledge extraction and data pattern analysis and other popularly used term "Knowledge Discovery in Databases" (KDD) [4].

Educational Data Mining (EDM) is concerned with developing methods for discovering knowledge from data that come from educational area by applying different Data Mining (DM) tools and techniques. It provides basic knowledge of teaching and learning process for effective education planning [5].

The data available in the educational field can be studied using educational data mining so that the hidden knowledge can be extracted from it. This study focused on higher education student's performance prediction with the help of available data in the institution by using data mining technique. Classification methods like decision trees, rule induction, naïve bayes, logistic regression, support vector machine and neural network, were applied on the educational data for predicting the student's performance [6].

The aim of the study is to design a predictive model that can predict students' performance. The result is useful for higher education institutions to improve the quality of education by providing information for teachers to take making academic decisions to produce students with high caliber and motive.

1.2. Statement of the Problem

In present days the volume of data kept in educational environment has been dramatically increasing. Most of the historical data could be collected from the existing database of educational institutions. However, the task to manage a large amount of data and determine the relationships among variables in the data is not an easy task to be done [7].

Student's performance in any higher educational institute is determined by the combination of internal assessment and external mark. An internal assessment is carried out by the teachers upon the student's performance in various evaluation methods such as tests, assignments, and attendance. An external mark is the one that is scored by the student in final examination in the

semester. Each student has to get the minimum pass mark in internal assessment and external examination [2].

In the growing world, the quality of higher education is given great attention by the government so as to produce competent graduates with the necessary skill and knowledge. Predicting students' academic performance at the beginning of the year helps to support students before they reached risk of failure, effective resource utilization and cost minimization, helping and guiding administrative officers to be successful in management and decision making.

The most vital issue in higher education is predicting students' performance. However, currently there are no functional student performance predictors that accurately determine whether a student will be successful, a dropout, or an average performer. In this regard, not much has been done in Ethiopia [8].

In our country, there are limited number of research works that attempted to apply data mining techniques in predicting student success and failure at the University level. A research done by Muluken [9] focused on the application of data mining technology in the Ethiopian higher education context, particularly at Debre Markos University. Constructing a prediction model to identify success and risk of failure of students; developing a predictive model that could help higher education institutions to identify university students at risk of failure so that they can be treated before the condition escalate into students' academic dismissal and wastage of resources.

Muluken [9] used main attributes for determining the failure or success of students in Debre Markos University; such as number of students in a class, number of courses given in a semester, Higher Education Entrance Certificate Examination result of a student, and sex. These attributes might be best for analyzing the general education system. However, the research didn't others factors for analyzing student performance in Ethiopian education system.

Though varies studies have been conducted using data mining technology, all the available knowledge in the area are not enough to solve the entire prediction problem so as to determine student performance. The studies lack country specific behaviors, and entirety of factors contributing to performance swings. In this study, some unique attributes are used other

researcher didn't consider it. These includes: academic schedule, financial sources and background of study.

Therefore, with the need of higher educational institute development and certainty of the data management, it is necessary to build a predictive model as part of decision support system. So, the purpose of this study is to apply Data Mining techniques for constructing a model for predicting the performance of higher education students.

Hence this research attempts to explore and answer the following main guiding questions.

- ❖ What are the major attributes that contributes to students' performance?
- ❖ Which Data Mining classification algorithms are more appropriate to predict students' performance?
- ❖ To what extent the predictive model performs well in determining students' performance?

1.3. Objective of the Study

1.3.1. General Objective

The general objective of this study is to design a predictive model using data mining techniques so as to determine students' performance at higher education.

1.3.2. Specific Objectives

The following specific objectives are identified in order to achieve the general objective of the research:

- ❖ To collect student data and prepare dataset by applying preprocessing tasks like data cleaning, transformation and attribute selection.
- ❖ To select appropriate data mining algorithms for experimentation.
- ❖ To build a predictive model using data mining tool on cleaned student's data, so as to understand the behavior of students' performance.
- ❖ To develop a prototype that demonstrates the use of hidden knowledge acquired using data mining technology.
- ❖ To test the performance of the prototype using system effectiveness and user acceptance testing.

1.4 Scope and Limitation of the Study

The study attempts to design a predictive model for determining the performance of undergraduate students at higher education. Many factors contribute to a student's academic performance. According to Muluken [9] major factors related to student performance are demography of the student, course taken per semester and previous learning background.

The scope of this study was also limited to the data collected from St. Mary's University, a private higher educational institute in Ethiopia and data collected also restricted from this higher educational institute database. The data collected cover the period from 2006 to 2009 E.C (Ethiopian calendar) concerning undergraduate students.

This study restricted to classification based data mining classification techniques such as Decision tree, Rule Induction, Bayesian, Logistic regression, Support Vector Machines and Neural Network to design a predictive model for determining higher education student's performance.

The unavailability of related literature was one of the limitations encountered to undertake the study. Another limitation of the study was we could not get all needed factors to determine higher education students' performance because of unavailability of clear data found in the database, other demographic data such as family size, family background and health-related data are not included under this study.

1.5 Significance of the study

The main contribution of this study is in identifying factors affecting student performance from large volumes of educational data and extracting hidden knowledge to predict the students' performance. The findings of this study has a great contribution to policy makers, curriculum designers, program managers, instructors, university admission officers, counselors, quality control, parents and other concerned bodies in many respects. Moreover, this study is expected to give some ideas for researchers who may wish to conduct studies on related areas of interests in a very detailed manner.

- ***Policy Makers/ Curriculum designers/Program managers*** to bring change in education and monitoring of school size, class size, school choice, school privatization, tracking, teacher education and certification, teacher pay, teaching methods, curricular content, graduation requirements, school infrastructure investment, and the values that schools are expected to uphold and model.
- ***Instructors*** may improve the academic performance of their students through designing different teaching style that could positively affect the variables of this study.
- ***Higher Education Relevance and Quality Agency*** to monitor the system to ensure that it is producing quality
- ***University admission officers*** can use these variables in their prediction model.
- ***Counselors*** who work with students for better academic and personal planning.
- ***Students*** can evaluate their assets and liability with respect to the psychological variables of this study that have an effect on their academic performance.
- ***Parents/ Guardians*** it may give some insights to know about the condition of their children.

The result of this study will also help other concerned bodies in the teaching-learning processes and it can be used by researchers as a base for exploring students' state in other universities.

1.6 Organization of the Study

This study is organized in five chapters. The first chapter introduces the general introductory concept and also covered the statement of the problem Research Question, objective of the studies, scope of the study and significance of the study.

The Second chapter Discusses the review of literature in the area of education data mining it discusses related to data mining, educational data mining, method educational data mining extracting hidden knowledge from educational data mining and also some related works is present in this chapter.

Third chapter presents understanding of the problem domain, data understanding and data preparation, the data collection and analysis tools and techniques, data transformation, and data integration, data reduction, and feature selection processes.

The fourth chapter presents experimental set up on the selected data mining algorithms as well as model development and testing results of classification algorithms for prediction by using WEKA data mining software tool. Finally extracting rule and discussion of the results are presented.

Finally, the fifth chapter presents conclusion of the study and recommendation for further work in the area.

CHAPTER TWO

LITERATURE REVIEW

In this chapter review of literature is presented on the concepts and techniques of data mining processes, techniques, methodologies and its application particularly in educational environment.

2.1. Data Mining and Knowledge Discovery

Data mining refers to techniques for extracting hidden, interesting patterns and knowledge from large amounts of data. Many other terms that carry a similar or slightly different meaning to data mining are knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [10].

Nowadays it has been estimated that the amount of data stored in the world's databases doubles every 20 months [11]. Intelligently analyzed data is a valuable resource. Data mining is about solving problems by analyzing data already present in databases [11].

The emergency of data mining and knowledge discovery in databases (KDD) as a new technology is due to the fast development and wide application of information and database technologies. With the increasing use of databases, the need to be able to digest large volumes of data being generated is now critical [12].

Data mining and KDD is aimed at developing methodologies and tools to automate the data analysis process and create useful information and knowledge from data to help in the decision making process [12].

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [10].

Data mining is an iterative process within which progress is defined by discovery, through either automatic or manual methods [13]. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an “interesting”

outcome. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers [14]. Data mining is an integral part of knowledge discovery in databases (KDD), which is the overall process of converting raw data into useful information [13].

2.1.1. Educational Data Mining

Different definitions have been provided for the term Educational Data Mining (EDM). Educational data mining (EDM) is increasingly recognized as an emerging discipline [15]. EDM focuses on the development of methods for exploring the unique types of data that come from an educational context. The application of data mining techniques to educational systems improves learning which can be viewed as a formative evaluation technique.

EDM is the application of data mining (DM) techniques to specific type of dataset that come from educational environments to address important educational questions and issues [15] [16].

EDM is both a learning science, as well as a rich application area for data mining, due to the growing availability of educational data. It enables data-driven decision making for improving the current educational practice and learning material.

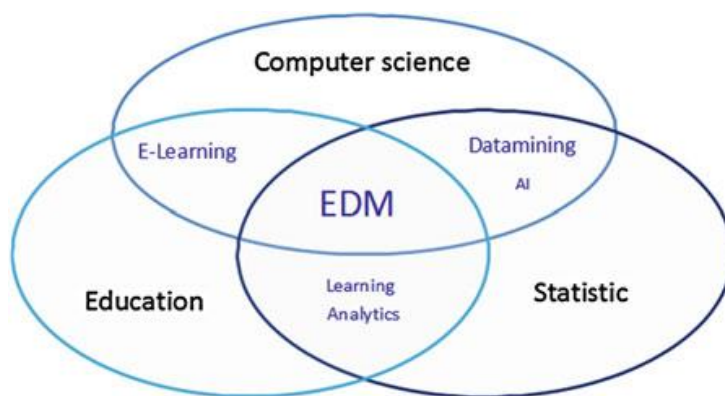


Figure 2.1: Areas in relation with EDM [15]

EDM can be drawn as the combination of three main areas; computer science, education, and statistics [15] . EDM uses methods and applies techniques from statistics, machine learning, data mining, information retrieval, recommender systems, psycho-pedagogy, cognitive psychology and psychometrics (see figure 2.1) [17]

EDM can be applied to assess students' learning performance, to improve the learning process and guide students' learning, to provide feedback and adapt learning recommendations based on students' learning behaviors, to evaluate learning materials and courseware, to detect abnormal learning behaviors and problems, and to achieve a deeper understanding of educational phenomena [18].

Different types of data used in EDM research since the data is collected form historical and operational data of the student's academic and personal data that exist in the database of educational institute. Scholars used many techniques in EDM, such as decision trees, neural networks, k-nearest neighbor, naïve bayes, support vector machines [18]. Using these methods different kinds of knowledge have been discovered using classifications, clustering and association rules.

2.2. Data Mining Tasks

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories; descriptive and predictive [10]. Figure 2.2 Presents DM tasks and functionalities to construct Predictive and Descriptive Data mining models.

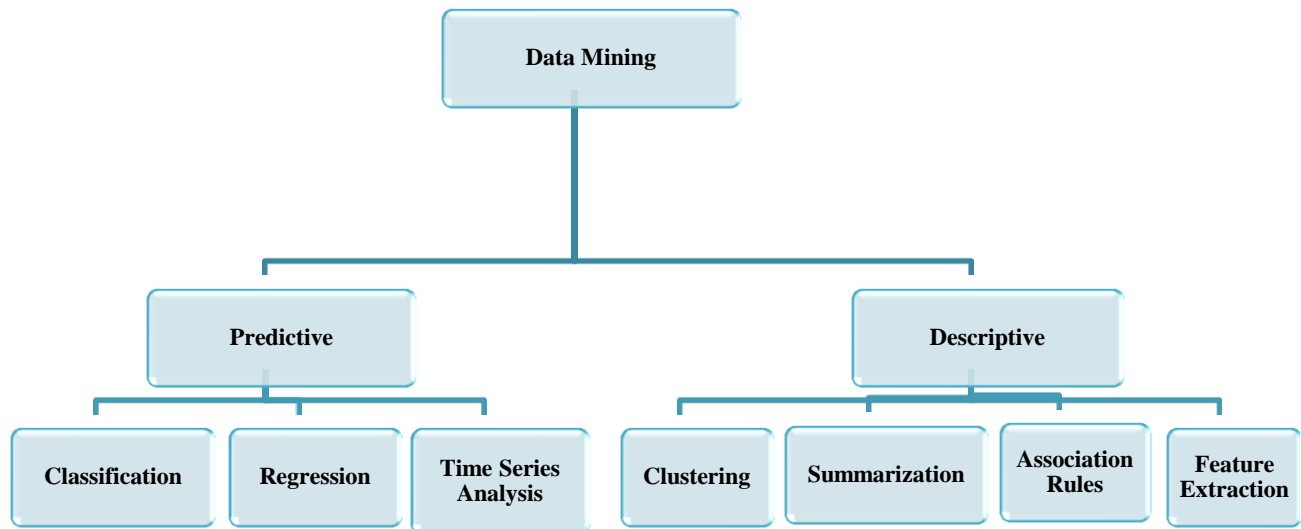


Figure 2.2: Predictive and Descriptive Data mining models [10]

Predictive models are built or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning [19], because it always requires data patterns with known class assignments to train a model which is then used for predicting the class assignment of new data patterns [12]. There are three types of prediction: classification, regression, and prediction [12]. However, Classification is the major task for predictive modeling in educational data mining.

Descriptive techniques are also referred to as unsupervised learning because there is no already-known result to guide the algorithms [19].

Descriptive modeling covers tasks such as Clustering, Association Rules, Summarizations, and Sequence discovery. A descriptive modeling use data mining technique to give a knowledge that identifies patterns found in data and will give image into the past and tells what has happened? Descriptive is used when a need arises to analyze and explain different aspects of the organization, to know, how they stand in the market nowadays and describe the facts and figures about the company. But the Predictive will recognize the future and tells what might happen in future. Reports generated by descriptive are accurate but the reports generated by Predictive are not 100% accurate; it may or may not happen in future [10].

So, since this study aims on predictive data mining one of its types a classification task suitable for this study is discussed below.

2.3. Predictive modeling using classification algorithm

The objective of this task is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables [13].

Classification is the process of finding a model or function that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). Derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks [12].

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large [10].

Some common classification methods include Decision Trees, Naïve Bayes, Neural networks, Rule-Based Classifiers (Rule Induction), K-Nearest Neighbor and support vector machines.

So, the techniques are appropriate for classification tasks seems to be strongly in need of the application, the data mining techniques that are applied in practical applications as a solution to classification problems, among other methods, are decision trees and rule inductions [10]. This section, discusses the concepts and principles of these techniques namely Decision Trees, Rule Inductions, Bayesian classification, Logistic regression, Support Vector Machines and Neural Network.

2.3.1. Decision Trees

Decision trees are a classification methodology, where the classification process is modeled with the use of a set of hierarchical decisions on the feature variables, arranged in a tree-like structure [20]. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called “root” that has no incoming edges (see figure 2.3). All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves also known as terminal or decision nodes [21].

In a decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute’s value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value [21].

Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path [21].

Only nodes that contain a mixture of different classes need to be split further. Eventually, the decision tree algorithm stops the growth of the tree based on a stopping criterion. The simplest stopping criterion is one where all training examples in the leaf node belong to the same class [20].

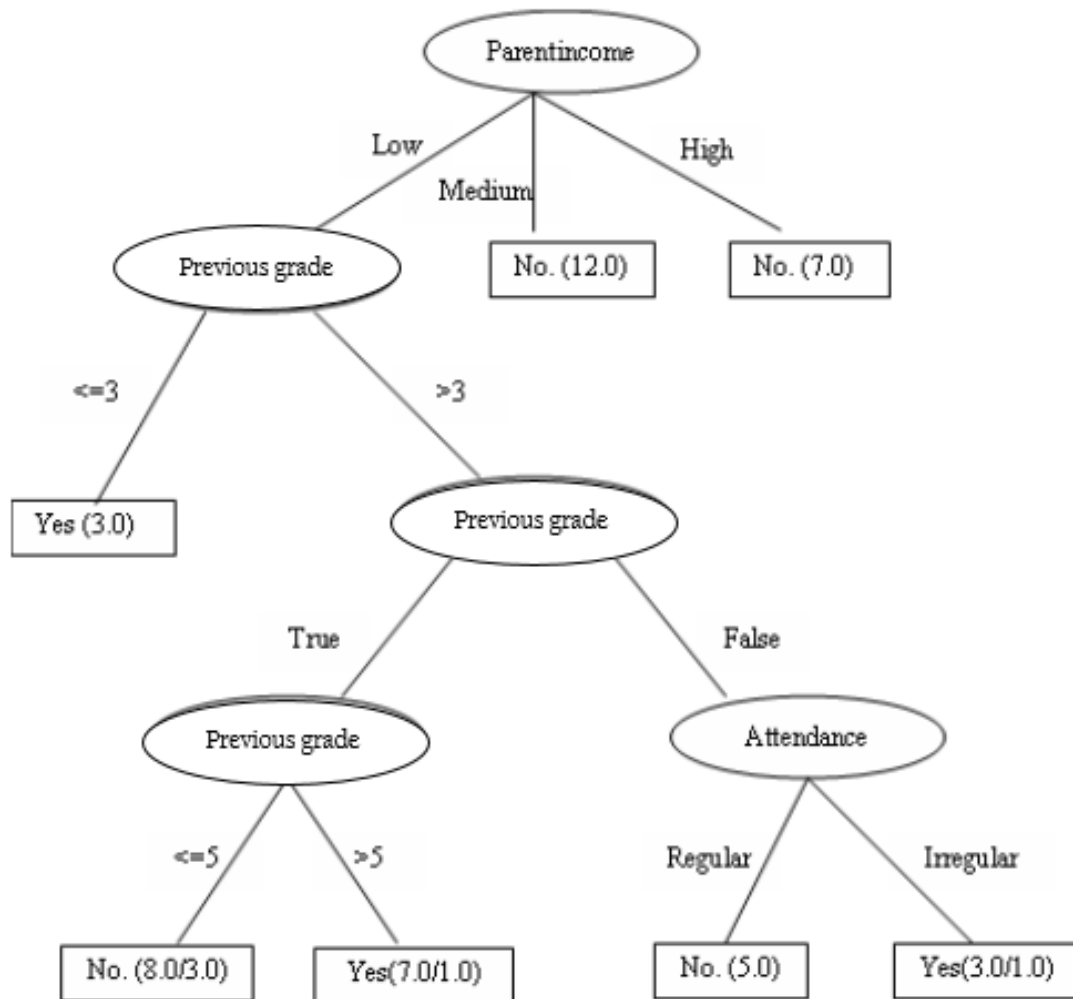


Figure 2.3: Decision tree diagram [22]

Internal nodes are represented as circles, whereas leaves are denoted as triangles. Two or more branches may grow out from each internal node. Each node corresponds with a certain characteristic and the branches correspond with a range of values. These ranges of values must be mutually exclusive and complete. These two properties of disjoint-ness and completeness are important since they ensure that each data instance is mapped to one instance (see Figure 2.3) as described in [23].

Instances are classified by navigating them from the root of the tree down to a leaf according to the outcome of the tests along the path. Start with a root of a tree; consider the characteristic that corresponds to the root and define to which branch the observed value of the given characteristic corresponds. Then, consider the node in which the given branch appears. Repeat the same

operations for this node until reach a leaf. Note that this decision tree incorporates both nominal and numeric attributes. Given this classifier the analyst can predict the response of a potential customer (by sorting it down the tree) and understand the behavioral characteristics of the entire population of potential customers regarding direct mailing. Each node is labeled with the attribute it tests, and its branches are labeled with its corresponding values [23].

Tree Pruning Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex [24]. There are two approaches to prune a tree [24].

- **Pre-pruning** - The tree is pruned by halting its construction early.
- **Post-pruning** - This approach removes a sub-tree from a fully-grown tree.

Decision tree algorithms, such as ID3, C4.5, and CART, were originally intended for classification.

WEKA implements several decision tree induction algorithms. The most frequently used algorithm is the J48 which is a variation of the well-known C4.5 algorithm [23].

2.3.1.1. J48 Decision Tree Algorithm

J48 decision tree handles both categorical and continuous attributes to build a decision tree. J48 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biases of information gain when there are many outcome values of an attribute [25].

Entropy provides an information theoretic approach to measure the goodness of a split. It measures the amount of information in an attribute [26].

$$\text{Entropy}(S) = \sum_{I=1}^c (-p(I) \log_2 p(I))$$

Information Gain that measures expected reduction in entropy caused by knowing the value of a feature F_j , is used [26].

$$\text{Information Gain}(S, F) = \text{Entropy}(S) - \sum_{vi \in V_{fj}} \frac{S_{vi}}{S} * \text{Entropy}(S_{vi})$$

To compensate for the bias of the information gain for cases with many outcomes, a measure called the Gain Ratio is used [26]

$$\text{Gain Ratio (S, F)} = \frac{\text{Information Gain (S, F)}}{\text{Split information (S, F)}}$$

The process of the J48 algorithm to build a decision tree is as follows [27].

1. Choose an attribute that best differentiates the output attribute values using information gain or gain ratio.
2. Create a separate tree branch for each value of the chosen attribute.
3. Divide the instances into subgroups so as to reflect the attribute values of the chosen node.
4. For each subgroup, terminate the attribute selection process if:
 - A. All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.
 - B. The subgroup contains a single node or no further distinguishing attributes can be determined. As in (a), label the branch with the output value seen by the majority of remaining instances.
5. For each subgroup created in (3) that has not been labeled as terminal, repeat the above steps 1-4.

Decision tree induction is closely related to rule induction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part and taking the leaf's class prediction as the class value [21].

2.3.2. Rule-Based Classifiers

A rule is typically expressed in the following form: *IF Condition THEN Conclusion* [20]. The condition on the left-hand side of the rule, also referred to as the antecedent, may contain a variety of logical operators, such as $<$, \leq , $>$, $=$, \subseteq , or \in , which are applied to the feature variables. The right-hand side of the rule is referred to as the consequent, and it contains the class variable. Therefore, a rule R_i is of the form $Q_i \Rightarrow c$ where Q_i is the antecedent, and c is the class variable. The “ \Rightarrow ” symbol denotes the “THEN” condition. The rules are generated from the training data during the training phase. The notation Q_i represents a precondition on the feature set [20].

Rules are expressed in the form of

IF (attribute 1; value 1) and (attribute 2; value 2) and (Attribute n; value n)

THEN (decision; value)

In WEKA there exist many rule induction algorithms OneR, PART, ZeroR and JRip (RIPPER) implements. In this research PART rule induction algorithm is experimented to build the predictive model.

2.3.2.1. PART Rule induction

The rule induction classifier used in this study is PART, which is one of the most commonly applied rule based classifications method. Rules are a good way of representing information or bits of knowledge [28]. PART generates a set of rules based on the divide and conquers strategy, and then it removes all instances from the training collection that are covered by this rule. Finally it precedes recursively until no instance remains. In other words, it combines the divide-and-conquer strategy with separate-and-conquer strategy of rule learning. Such algorithms have been used as the basis of many systems that generate rules. The algorithm generates sets of rules called decision lists ‘which are ordered set of rules. PART obtains rules from partial decision trees using J48 and builds a partial C4.5 decision tree and converts the "best" leaf into a rule [29].

A rule induction system constructs a set of rules. Assuming a set of classified example rules an if-then rule has the form:

IF Sex = male AND Age > 46 AND Number of Course > 3 AND Division = Extension THEN
Status = Success

2.3.3. Bayesian Classification

Bayesian classifiers are statistical classifiers based on Bayes' theorem. Studies [28] comparing classification algorithms have found a simple Bayesian classifier known as the Naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers.

Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases [28]. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes, that is, given the class label of a tuple, the values of the attributes are assumed to be conditionally independent of one another. This assumption is called class conditional independence. It is made to simplify the computations involved. When the assumption holds true, and then the Naïve Bayesian classifier is the most accurate in comparison with all other classifiers [28]. In practice, however, dependencies can exist between variables. Bayesian belief network specifies joint conditional probability distributions. It allows class conditional independencies to be defined between subsets of variables. Bayesian classification provides a graphical model of causal relationships, on which learning can be performed. Trained Bayesian belief networks can be used for classification [28]. Bayesian classification graphical model shown in figure 2.4.

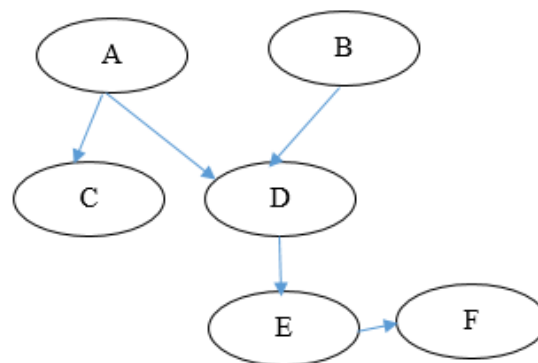


Figure 2.4: Bayesian graphical model

Bayes' Theorem is named after Thomas Bayes, an 18th British mathematician and minister who did early work in probability and decision theory. Let X be a data tuple. It is usually described by measurements made on a set of n attributes. Let H be a hypothesis. $P(H/X)$ is defined as a probability that the hypothesis H holds given the data tuple X . It is the posterior probability, or a posteriori probability, of H conditioned on X . In contrast, $P(H)$ is the prior probability, or a priori probability, of H . Similarly, $P(X/H)$ is the posterior probability of X conditioned on H and $P(X)$ is the prior probability of X . Bayes' Theorem provides a way of calculating the posterior probability, $P(H/X)$, from $P(H)$, $P(X/H)$, and $P(X)$. Bayes' Theorem is

$$P(H/X) = \frac{P(X/H)P(H)}{P(X)}$$

For classification problems, let X be an observed data tuple, and suppose that H is the hypothesis that X belongs to a specified class C . We want to determine the probability $P(H/X)$ that tuple X belongs to class C , based on the attribute description of X [28].

2.3.3.1. Naïve Bayesian

Naïve Bayesian classifier uses the Bayes' rule to compute the probability of each possible value of the target attribute given the instance, assuming the input attributes are conditionally independent given the target attribute i.e. class conditional independence. Due to the fact that this method is based on the simplistic, and rather unrealistic assumption that the causes are conditionally independent given the effect, this method is well known as Naïve Bayes [28] [30].

The naïve Bayesian classifier works as follows [28]:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n -dimensional attribute vector, $\mathbf{X} = (x_1, x_2, \dots, x_n)$, depicting n measurements made on the tuple from n attributes, respectively, A_1, A_2, \dots, A_n .
2. Suppose that there are m classes, C_1, C_2, \dots, C_m . Given a tuple, \mathbf{X} , the classifier will predict that \mathbf{X} belongs to the class having the highest posterior probability, conditioned on \mathbf{X} . That is, the naïve Bayesian classifier predicts that tuple \mathbf{X} belongs to the class C_i if and only if

$$P(C_i/\mathbf{X}) > P(C_j/\mathbf{X}) \text{ for } 1 \leq j \leq m; j \neq i.$$

Thus we maximize $P(C_i | \mathbf{X})$. The class C_i for which $P(C_i | \mathbf{X})$ is maximized is called the *maximum posteriori hypothesis*.

$$P(C_i | \mathbf{X}) = \frac{P(\mathbf{X} | C_i)P(C_i)}{P(\mathbf{X})}$$

3. As $P(\mathbf{X})$ is constant for all classes, only $P(\mathbf{X} | C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \dots = P(C_m)$, and we would therefore maximize $P(\mathbf{X} | C_i)$. Otherwise, we maximize $P(\mathbf{X} | C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i) = \frac{|C_{i,D}|}{|D|}$, where $|C_{i,D}|$ is the number of training tuples of class C_i in D .
4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(\mathbf{X} | C_i)$. In order to reduce computation in evaluating $P(\mathbf{X} | C_i)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$\begin{aligned} P(\mathbf{X} | C_i) &= \left(\prod_{k=1}^n P(x_k | C_i) \right) P(C_i) \\ &= P(x_1 | C_i) * P(x_2 | C_i) * \dots * P(x_n | C_i) \end{aligned}$$

We can easily estimate the probabilities $P(x_1 | C_i)$, $P(x_2 | C_i)$, \dots , $P(x_n | C_i)$ from the training tuples. Recall that here x_k refers to the value of attribute A_k for tuple \mathbf{X} . For each attribute, we look at whether the attribute is categorical or continuous-valued.

5. In order to predict the class label of \mathbf{X} , $P(\mathbf{X} | C_i)P(C_i)$ is evaluated for each class C_i . The classifier predicts that the class label of tuple \mathbf{X} is the class C_i if and only if $P(\mathbf{X} | C_i)P(C_i) > P(\mathbf{X} | C_j)P(C_j)$ for $1 \leq j \leq m, j \neq i$.
In other words, the predicted class label is the class C_i for which $P(\mathbf{X} | C_i)P(C_i)$ is the maximum.

2.3.4. Regression

Regression is predictive modeling and analysis is used to make predictions based on existing data by applying formulas. It is a statistical method of data mining. Regression actually used to model between the one or more dependent variables and independent variables. It can be used for building model or classifiers which can analyses the historical data to predict the future trends using linear or logistic regression techniques from statistics, a function is learned from the existing data. The new data is then mapped to the function in order to make predictions it is uses existing values to forecast what other values will be [19].

2.3.4.1. Logistic Regression

Logistic regression is used to develop a regression model when the dependent variable is categorical. It was developed in 1958 by David Cox. There are three types of logistic regression: (1) binary, for a binary response variable, (2) multinomial - where the dependent variable has more than two non-ordered categories, and (3) ordinal - when the categories are ordered [31].

Logistic Regression (LR) is one of the most important statistical and data mining techniques employed by statisticians and researchers for the analysis and classification of binary and proportional response data sets. The main advantages of LR are that it can naturally provide probabilities and extend to multi-class classification problems [32].

2.3.5. Support Vector Machines

Support vector machine (SVM) were first suggested by Vapnik in the 1960s for classification and have recently become an area of intense research owing to developments in the techniques and theory coupled with extensions to regression and density estimation. Support vector machine (SVM) is a training algorithm for learning classification and regression rules from data, for example the SVM can be used to learn polynomial, radial basis function (RBF) and multi-layer perceptron (MLP) classifiers. SVMs arose from statistical learning theory; the aim being to solve only the problem of interest without solving a more difficult problem as an intermediate step [23].

SVM is a supervised learning algorithm creating learning functions from a set of labeled training data. It has a sound theoretical foundation and requires relatively small number of samples for training; experiments showed that it is insensitive to the number of samples' dimensions. Initially, the algorithm addresses the general problem of learning to discriminate between members of two classes represented as n – dimensional vectors [14].

2.3.5.1. Sequential Minimal Optimization

Sequential Minimal Optimization algorithm, due to John Platt's, gives an efficient way of solving the dual problem arising from the derivation of the SVM. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. In that case the coefficients in the output are based on the normalized data, not the original data, this is important for interpreting the classifier [30].

2.3.6. Neural Networks

An artificial neural network (ANN) often called as a neural network (NN), is a computational model based on the biological neural networks, in other words, is a representation and emulation of human neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation.

A neural network, as the name indicates, is a network structure consisting of a number of nodes connected through directional links. Each node represents a processing unit, and the links between nodes specify the causal relationship between connected nodes. All nodes are adaptive, which means that the outputs of these nodes depend on modifiable parameters pertaining to these nodes [33].

A neural network starts with an input layer, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer. Each input node is connected to every node in the hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables [19].

After the input layer, each node takes in a set of inputs, multiplies them by a connection weight W_{xy} (e.g., the weight from node 1 to 3 is W_{13}), adds them together, applies a function (called the activation or squashing function) to them, and passes the output to the node(s) in the next layer. For example, the value passed from node 4 to node 6 is:

Activation function applied to $([W_{14} * \text{value of node 1}] + [W_{24} * \text{value of node 2}])$

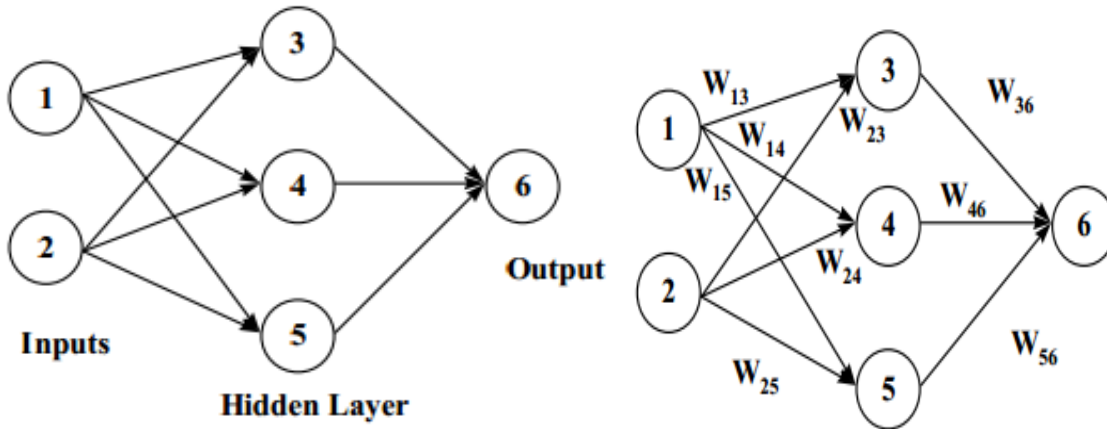


Figure 2.5: A neural network with one hidden layer and W_{xy} is weight from node x to node y

2.3.6.1. Multilayer Perception

Multilayer perceptron is a neural network which is the most used architecture for predictive data mining. It is a feed-forward network with possibly several hidden layers, one input layer and one output layer, totally interconnected. It can be considered as a highly nonlinear generalization of the linear regression model when the output variables are quantitative, or of the logistic regression model when the output variables are qualitative.

The network is feed-forward if the processing propagates from the input side to the output side unanimously, without any loops or feedbacks. In a layered representation of the feed-forward neural network, there are no links between nodes in the same layer; outputs of nodes in a specific layer are always connected as inputs to nodes in succeeding layers. This representation is preferred because of its modularity, i.e., nodes in the same layer have the same functionality or generate the same level of abstraction about input vectors [14].

2.4. Data Mining Tool Selection

There are different data mining software's that can be used to build and test predictive models. Choosing of a data mining tool is not an easy task. The important thing is identifying the tool suite that is easy to use, provides acceptable accuracy and able to perform all the common tasks in a data mining. The most widely known and an open source data mining tool include WEKA, Orange and Rapid Miner [34].

WEKA

WEKA (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code. The WEKA (pronounced Weh-Kuh) workbench contains a collection of several tools for visualization and algorithms for data analytics together with graphical user interfaces for easy access. It includes a set of algorithms for data preparation, predictive modeling, and descriptive modeling and attributes selection.

RAPIDMINER

A software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process. Rapid Miner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures.

ORANGE

Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. Table 2.1 shows the comparison of the data mining tools, specifically WEKA, Rapid Miner and ORANGE.

No	Tool Name	Release Date	Language	Advantage	Disadvantage
1	WEKA	1993	Java	Ease of use, can be extended in RM , visualization, attribute selection, preprocessing	Poor documentation, weak classical statistics, poor parameter optimization,
2	RAPID MINER	2006	Language Independent	Visualization, Statistical, Attribute Selection, Outlier detection, parameter optimization	Requires prominent knowledge of database handling
3	ORANGE	2009	Python C++,C	Better debugger, Shortest scripts, poor statistics, suitable for no voice Experts	Big installation, Limited reporting capabilities

Table 2.1: Comparison of the three open sources Data Mining tools [35]

In this study the researcher chooses WEKA data mining tool. The reason why this tool is specially selected is that it is the only toolkit that has gained widespread adoption and survived for an extended period of time and it is freely available [36]. Other reason is familiarity of the researcher with the tool.

2.5. Related works

Now, there are similar studies that were applied to study the application of data mining techniques on Educational data mining and related issues. The following studies are reviewed to clarify the significance of the research problem and to show the gap in the local and foreign studies.

Baradwaj and Pal [37] conducted a research on a group of 50 students from VBS Purvanchal University, Jaunpur (Uttar Pradesh India). For the study MCA (Master of Computer Applications) program was considered from session 2007 to 2010, with multiple performance indicators, including previous semester mark, class test grades, seminar performance, assignments, general proficiency, attendance, lab work, and end semester marks. They use classification task to evaluate student's performance, ID3 decision tree algorithm to finally construct a decision tree, and if-then rules which will eventually help the instructors as well as the students to better understand and predict students' performance at the end of the semester.

Bhardwaj and Pal [38] conducted a significant data mining research on the student performance based by selecting 300 students' data from 5 different degree colleges using the Naïve Bayes

classification method, on a group of BCA (Bachelor of Computer Applications) students in Dr. R. M. L. Awadh University, Faizabad, India, who appeared for the final examination in 2010. A questionnaire was distributed for collecting data from each student before the final examination, which had multiple personal, social, and psychological questions that was used in the study to identify relations between these factors and the student's performance and grades. They found that the most influencing factor for student's performance is their grade in senior secondary school, which further tells us, that those students, who performed well in their secondary school, were definitely perform well in their Bachelors study. Furthermore, it was found that the living location, medium of teaching, mother's qualification, student other habits, family annual income, and student family status, all of which, highly contribute in the students' educational performance, thus, it can predict a student's grade or generally their performance if basic personal and social knowledge was collected. Limitation of this study is the experiment was conducted with a less rich dataset; the total dataset used in this experiment was only 300.

Amirah Mohamed et al. [39] provided an overview on the data mining techniques that have been used for predicting students' performance. A current prediction method is not adequate to classify the most suitable methods for predicting the performance of students in Malaysian institutions. Another one is lack of surveys on the factors affecting students' achievements in particular courses. From their review, they point out systematical literature review on predicting student performance by using data mining techniques is proposed to improve student's achievements and how the prediction algorithm can be used to identify the most important attributes in a student's data. Also, it focused on which prediction methods should be used for student performance. The result shows that Neural Network has the highest prediction accuracy of 98% followed by Decision Tree by 91%; and from several algorithms under the classification techniques, Neural Network and Decision Tree are the two methods highly used by the researchers for predicting students' performance. Results shows that Neural Network method gave the highest prediction accuracy because of the influence from internal and external assessments attributes and least impact on student performance is psychometric factors.

Superby et. al. [40] predicted students at risk of drop-out, determining factors influencing the achievement of the first-year university students. At the beginning of the academic year 2003-2004 researchers distributed a questionnaire to 533 first-year university students during the month of November in three Belgian universities. classifying students into three classes – low-risk, with a high probability of succeeding; medium-risk who may succeed thanks to the measures taken by the university; and high-risk, who have a high probability of failing or dropping out, using decision trees, Random forest method, Neural networks and linear discriminate analysis. Researchers noted that 20% of variables showed significant correlation with academic success, and results obtained by the methods of prediction, discriminant analysis, and to a lesser extent neural networks and random forests, on the condition able to lead to interesting results.

Dorina Kabakchieva [6] studied on education and quality of education in the Bulgarian university. 10330 student's data that have been enrolled as university students during the period between 2007 and 2009, described by 20 parameters, for predicting student performance, based on their personal, pre-university and university-performance characteristics. In general, the main goal is how to give quality information for decision making in effective time, determining significant patterns in large quantities of data. In the presented paper, different data mining classification algorithms such as Neural network, Decision tree algorithm, Rule Learner, and K-Nearest Neighbor method were used. The procedures categorize the students in to Weak and Strong, according to student pre-university data, as mentioned in above description WEKA software was used for classification using different data mining algorithms such as Decision Tree, OneR Rule Learner, Neural Network and K-Nearest Neighbor, The highest accuracy is achieved for the Neural Network model (73.59%), followed by the Decision Tree model (72.74%) and the k-NN model (70.49%). So the Neural Network model is recommended for predictions.

According to Quadri and Kalyankar [22] Predicting the academic outcome of a student needs lots of parameters to be considered. Data pertaining to student's background knowledge about the subject, the proficiency in attending a question, the ability to complete the examination in time will also play a role in predicting performance. Uses decision trees to make important design decisions about the interdependencies among the properties of dropout students. The

scholars proved that student's performance can be predicted using the CGPA grade system where the data set comprised of the student's gender, parental education details, and financial background.

Amjad Abu Saa [41] explored factors affecting Student's academic performance, like personal, environmental, social, and economical variables. What the challenging area in evaluating performance of student's is large volume of data in educational databases. The paper explained that prediction of academic performance is broadly researched. Main objective of the paper is factors that affect courses success rate and student success to get useful information by using different data mining techniques such as classification, prediction, clustering, association, decision tree and sequential patterns. According to paper objective is achieving the main goal and classify clustering techniques and classification to expand academic performance. They found that the student's performance is not totally dependent on their academic efforts; also there are other factors such as personal and social factors that have equal to greater influences as well.

Oyelade et.al. [42] Presents k-means clustering algorithm as a simple and efficient tool to monitor the progression of students' performance in higher institution. They applied the model on the data set academic result of one semester and nine courses of a private university in Nigeria. Demonstrated their technique using k-means clustering algorithm and combined with the deterministic model. Nine courses offered for that semester for each student for total number of 79 students. The overall performance is evaluated by applying deterministic model where the group assessment in each of the cluster size is evaluated by summing the average of the individual scores in each cluster. This analysis showed that cluster size 24 is 50.08% while the overall performance for cluster size 16 is 65.00%. Cluster size 30 has the overall performance of 58.89%, while cluster size 09 is 43.65%. The trends in this analysis indicated that, 24 students fall in the region of "Good" performance (50.08%), while 16 students have performance in the region of "Very Good" performance (65.00%). 30 students have a "Good" performance (58.89%) and 9 students had performance of "Fair" result (43.65%).

Fiseha and Addisalem [43] present the application of educational data mining for predicting students' performance based on their academic record, using a decision tree algorithm. The data was collected from the college of Agriculture, Department of Horticulture Dilla University. The data includes five years period from 2009-2014; datasets are integrated from different datasets. The preprocessing, processing and experimenting was conducted using RapidMiner tool. During processing among a total of 49 various attributes which will help to improve the student's academic performance 27 important rules generated. The work addressed the result of decision tree model reveals that specific courses, student academic status in 1st and 2nd year and sex are attributes that determine the performance of student. Finally, the decision tree algorithm was tested and it provides a promising result of accuracy of useful data mining can be used in higher education particularly to predict students' performance using their academic record.

Alemu and Tamir [44] used data mining approaches to develop a Data Mining model to predict the academic performance of students at the end of the first year degree program. In this study, the data 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 Decision tree learning algorithm and generates five classification rules set classifiers in an experiment. The important six 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). The experiment using a test dataset of 5729 students' record gets 81.4% accuracy.

Tariku [45] also applied data mining for identifying the determinant factors for the student success in the preparatory schools to join higher education. In this work, the researcher used the hybrid data and the study only focused on Addis Ababa 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 and 40328 instances and 15 attributes are selected for analysis. 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. In the researcher finding 84% of students who fail in English subject also fail in Mathematics subject examination in EHEEE.

Different researchers studied on different variables, all those previous studies were conducted by using a very small proportion of the datasets and variables. There for a limited capacity to discover new and unexpected patterns and relationships. This research attempts to apply data mining for predicting student performance using classification algorithm by increase the dataset and variables in order to fill the gaps of the previous research works. Hence this study has its own contribution to get useful information that can help students to know their weakness, as well as instructors and administration to develop a better policy for students using data mining techniques.

CHAPTER THREE

Methodology

Methodology is a way that deals with data collection, analysis and interpretation in order to help the investigators achieve the objective of the research. Hence, the following methods and processes followed in this research work.

3.1. Research Design

This study follows experimental research. Experimental research designs are selected because the primary approach used to investigate causal (cause/effect) relationships and to study the relationship between one variable and another. Researchers use experimental research to compare two or more groups on one or more measures [46].

To conduct an extensive experiment the study uses hybrid data mining process model. This process model is selected because it provides more general, research-oriented description of the steps, Introducing a data mining step instead of the modeling step, Introducing several new explicit feedback mechanisms. The model has six steps (see figure 3.1); understanding of the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and finally use of the discovered knowledge.

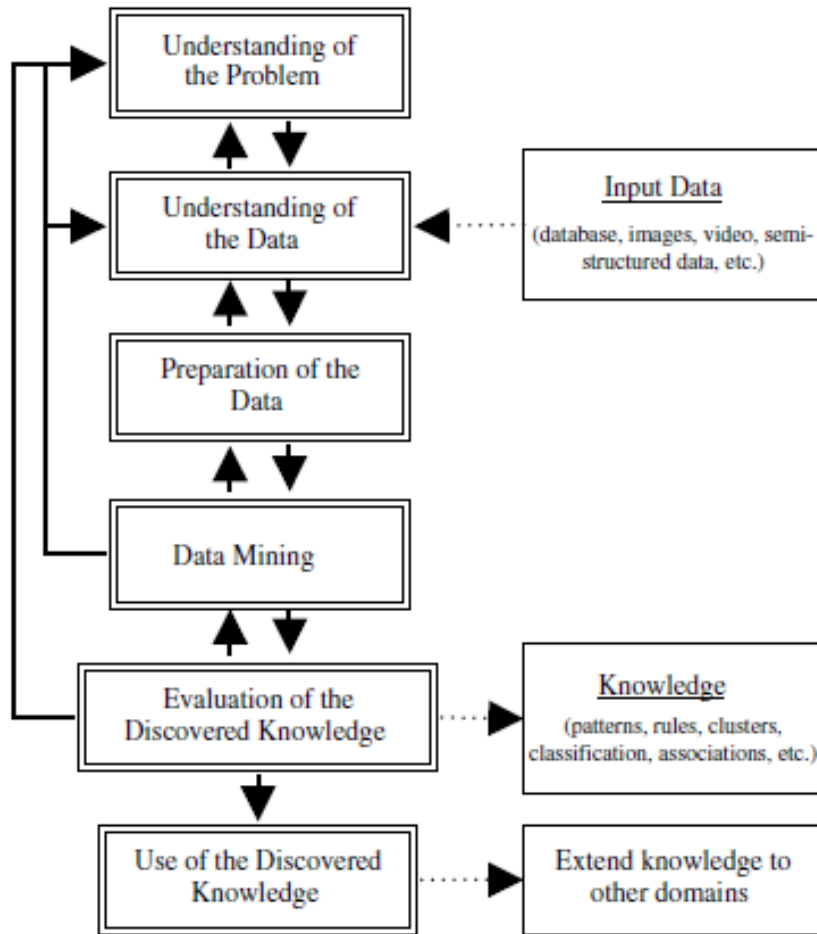


Figure 3.1: The six-step KDP model [26]

Therefore, the overall research design was to build a model that can be used to predict the students' performance on higher instruction dataset. One of the most important aspects of this model is iterative and interactive feature. The feedback loops are necessary because any changes and decisions made in one of the steps can result in changes in following steps. The researcher tries to discuss tasks done and methods used at each step in detail below.

3.1.1. Understanding of the problem domain

This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology [26].

In this study discussion was made with domain experts to understand the problem domain and to have a good understanding about the initial datasets and domain specific terminology. Relevant documents such as journal articles, conference papers and the internet were also reviewed to understand and identify algorithms and methods used for data preparation as well as data mining performance evaluation.

3.1.2. Understanding of the data

This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data was visualize using excel format to check 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. This step need for additional or more specific information about the data in order to guide the choice of specific data preprocessing algorithms [26].

The end result of knowledge discovery process depends on the quality of available data. As a result; careful analysis of the data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the DM tasks to be performed. We used Microsoft Excel 2019 to visualize the extracted data from student record management information system to check for missing values, data redundancy, data integrity, and data completeness.

3.1.3. Preparation of the data

This step is concerned with deciding which data are used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection

and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data granularization) [26]. The end results are data that meet the specific input requirements for the DM tools selected in Step 3.1.4.

The major tasks undertaken in this phase include: description of data sources, carrying out statistical summary measure, filling missing with mode values, and data transformation/reduction activities. Additionally, converting data from one format to another format was also made.

3.1.4. Data mining

One of the major tasks of data mining research is modeling; here the data miner uses various DM methods to derive knowledge from preprocessed data.

In this phase, various data mining techniques were selected and applied and their parameters calibrated to optimal values. Typically, there are several techniques and tool for the same DM problem type. In this study WEKA version 3.9.2 (Waikato Environment for Knowledge Analysis) is selected for DM. It is free software available under the GNU (General Public License). The WEKA work bench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [14]. It is an open source; powerful tool for data mining algorithm will be used for this study. We experimented different DM classification algorithms such decision tree, rule induction, Bayesian, regression, Support Vector Machines and Neural Network. These algorithms are selected because they are usually used for educational data mining and understandable rules can be extracted.

3.1.5. 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. The experimental output of the classifications models are analyzed and evaluated the performances accuracy using confusion matrix.

Furthermore, we can also compute the effectiveness and efficiency of the model in terms of recall and precision, as well as F-measure for good balance between precision and recall. The performance of the classifiers with different parameters is also compared by examining their ROC (Receiver Operating Characteristic) curve [14].

3.1.6. Use of the discovered knowledge

This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

In this research the discovered knowledge is used by integrating the user interface which is designed by java programming language with a WEKA system in order to show the prediction of higher education students' performance. Java is selected because it is easy to build applications. Java has a well-developed GUI /Graphical user Interface/ or windowing environment. Also, the researcher implementers personally prefer an object-oriented approach.

Java is natural choices for application development [47]. With Java's powerful language concepts and distributed application framework, Java offers a major application development framework as used in research and industry. Tactically, Java is a natural choice of language and provides a common framework for exploring and developing applications using data mining [47].

3.2. Data Mining Process Models

A process model is the set of tasks to be performed to develop a particular element, as well as the elements that are produced in each task (outputs) and the elements that are necessary to do a task (inputs) [48]. The goal of a process model is to make the process repeatable, manageable and measurable (to be able to get metrics) [49]. Methodology can be defined as the instance of a process model that lists tasks, inputs and outputs and specifies how to do the tasks [48]. Tasks are performed using techniques that stipulate how they should be done. After selecting a technique to do the specified tasks, tools can be used to improve task performance [49].

3.2.1. KDD Process model

The history of DM and KDD is not much different. In the early 1990s, when the term KDD (Knowledge Discovery in Databases) was first coined [50] there was a rush to develop DM algorithms that were capable of solving all the problems of searching for knowledge in data. Apart from developing algorithms, tools were also developed to simplify the application of DM algorithms. From the viewpoint of DM and KDD process models, the year 2000 marked the most important milestone [49].

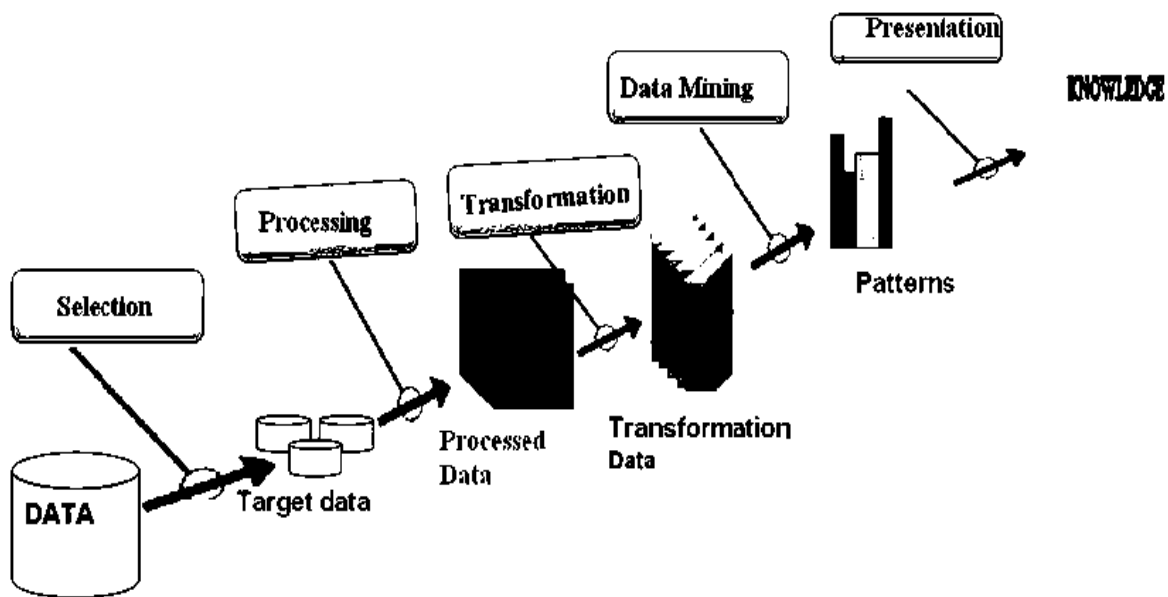


Figure 3.2: KDD Process model [26]

The KDD process model consists of five steps, that are described below [51].

1. **Select a target data set:** The initial step is based on data needed for the DM process may be obtained from many different and heterogeneous data sources.
2. **Data preprocessing:** In this step the data to be used by the process may have incorrect or missing data. There may be abnormal data from multiple sources involving different data types and metrics.
3. **Data transformation:** Attributes and instances are added and/or eliminated from the target data. Data from different sources must be converted into a common format for processing.

4. **Data mining:** A best model for representing the data is created by applying one or more DM algorithms.
5. **Interpretation/evaluation:** The final step the researcher examines the output from step 4 to determine if what has been discovered is both useful and interesting.

3.2.2. The CRISP-DM process model

CRISP-DM (Cross-Industry Standard Process for DM) is the most used methodology and application-neutral standard for developing DM & KD projects. It is actually a “de facto” standard [49]. It is a cyclic approach (see Figure 3.3), including six main phases – Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment [52].

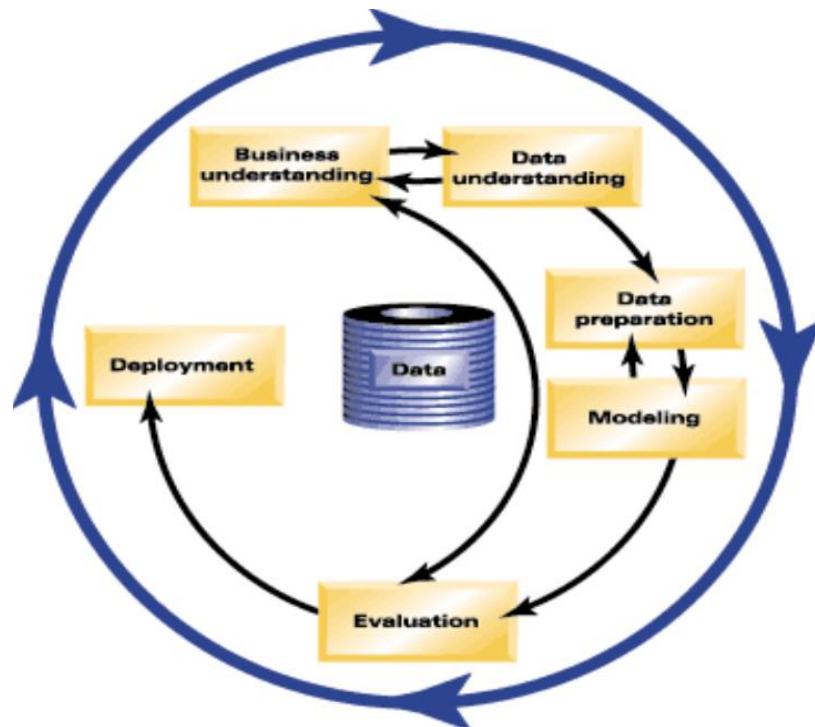


Figure 3.3: CRISP DM knowledge discovery process model [52]

A description of the six steps of CRSP-DM presented below [49]

- 1. Business understanding:** This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.
- 2. Data understanding:** The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
- 3. Data preparation:** The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order.
- 4. Modeling:** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, it is often necessary to step back to the data preparation phase.
- 5. Evaluation:** What are, from a data analysis perspective, seemingly high-quality models will have been built by this stage of the project. Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves the business objectives. At the end of this phase, a decision should be reached on how to use of the DM results.
- 6. Deployment:** Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Table 3.1 presents the six phases of CRISP-DM process model.

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine business objective	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review process	Plan monitoring & Maintenance
Determine DM objective	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

Table 3.1: CRISP-DM phases and tasks [49]

3.2.3. Hybrid Models

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of both CRISP-DM and KDD. One such model is a six-step KDP model developed by Cios et al [50]. It was developed based on the CRISP-DM model by adopting it to academic research [50].

A description of the six steps of KDD process model is given below [26]

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

- 2. Understanding of the data.** This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.
- 3. Preparation of the data.** This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data granularization). The end results are data that meet the specific input requirements for the DM tools selected in Step 1.
- 4. Data mining.** Here the data miner uses various DM algorithms for classification, clustering and association rule discovery to derive knowledge from preprocessed data. In this step, data mining tool is selected for experimenting data mining algorithms and constructing predictive or descriptive models.
- 5. 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.
- 6. Use of the discovered knowledge.** This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented. Finally, the discovered knowledge is deployed.

3.2.4. Comparison of Data Mining process Models

Comparing the KDD stages with the CRISP-DM, first steps 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. Concerning the remaining stages, we can say that: 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 DM; The evaluation phase can be identified with Interpretation/Evaluation [53]. Table 3.2 presents a summary of comparison of Data Mining process models.

In Summary Comparison of DM and KD process models is given in table 3.2 below

Model	Fayyad et al.	Cabena et al.	Anand & Buchner	CRISP-DM	Cios et al.
No of steps	9	5	8	6	6
Steps	Developing and Understanding of the Application Domain	Business Objectives Determination	Human Resource Identification	Business Understanding	Understanding the Data
	Creating a Target Data Set	Data Preparation	Problem Specification		
	Data Cleaning and Pre-processing		Data Prospecting	Data Understanding	Understanding the Data
	Data Reduction and Projection		Domain Knowledge Elicitation		
	Choosing the DM Task		Methodology Identification	Data Preparation	Preparation of the Data
	Choosing the DM Algorithm		Data Pre-processing		
	DM	DM	Pattern Discovery	Modeling	DM
	Interpreting Mined Patterns	Domain Knowledge Elicitation	Knowledge Post-processing	Evaluation	Evaluation of the Discovered knowledge
	Discovered Knowledge	Assimilation of Knowledge		Deployment	Using the Discovered Knowledge

Table 3.2: Comparison of DM & KD process models and methodologies [49]

3.3. Model Evaluation

Evaluation method is the yard stick to examine the efficiency and performance of any model. The evaluation is important for understanding the quality of the model or technique for refining parameters in the iterative process of learning and for selecting the most acceptable model or technique from a given set of models or techniques [54] .

Confusion Matrix

One of the methods to evaluate the performance of a classifier is using confusion matrix. A Confusion matrix that summarizes the number of instances predicted correctly or incorrectly by a classification model. In confusion matrix, there are classifier evaluation metrics like Accuracy, Precision, Recall, and F-measure [55] .

In the two-class case with classes yes and no, a single prediction has four different possible outcomes as shown in table 3.3. The true positives (TP) and true negatives (TN) are correct classifications. A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive.

		PREDICTED CLASS	
		Class = Yes	Class = No
ACTUAL CLASS	Class = Yes	a (TP)	b (FN)
	Class = No	c (FP)	d (TN)

Table 3.3: Two dimensional Confusion Matrix [55].

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

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

The precision for a class is the number of **true positives** (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of **true positives and false positives**, which are items incorrectly labeled as belonging to the class).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and **false negatives**, which are items which were not labeled as belonging to the positive class but should have been).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

In the same way, precision and recall are defused for the negative class. The ratio of the negative over total instance classified as negative will give use precision. On the other hand, the ratio of true negative over total instances of negative class will give uses recall.

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

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC Curves

ROC (Receiver Operating Characteristic) curves are two dimensional graphs that visually depict the performance and performance trade off of a classification model. In order to construct ROC curve two performance metrics are to be used. They are true positive rate (TPR) and false positive rate (FPR) [26] [56].

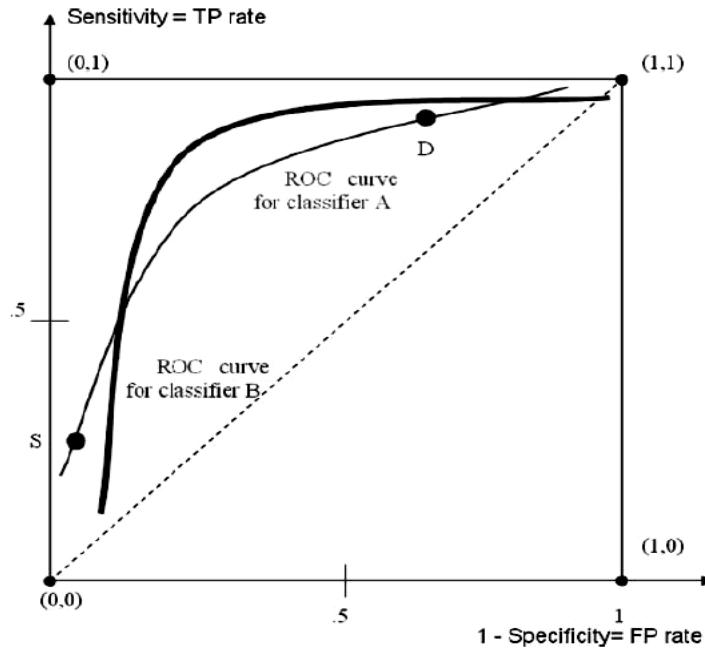


Figure 3.4: ROC (Receiver Operating Characteristic) [26]

The model with perfect accuracy will have an area of 1.0; i.e. the larger the area, the better performance of the model or the larger values of the test result variable indicate the stronger evidence for a positive actual state (1.00). The closer the ROC curve of a model is to the diagonal line, the less accurate the model is closer to the area of 0.5 [26].

ROC Area	Performance
0.9 – 1.0	Excellent (A)
0.8 – 0.9	Good (B)
0.7 – 0.8	Fair (C)
0.6 – 0.7	Poor (D)
0.5 – 0.6	Fail (F)

Table 3.4: Performance measure of ROC Area

Therefore, for this particular study Accuracy, Precision, Recall, F-measure and ROC area are taken in to account when the classifier performance is evaluated to select the best model.

3.4. User Acceptance Testing

User acceptance testing is a testing methodology where the clients or end users involved in testing the product to validate the product against their requirements. The five E's (effectiveness, efficiency, engaging and easy to learn) provide the expert with a set of characteristics that can be used to organize and analyze information from users [57]. They offer traceability from initial information-gathering through requirements setting and finally in evaluation.

3.4.1. Effectiveness

Effectiveness is the completeness and accuracy with which users achieve specified goals. Effectiveness is considered as the accuracy of the prototype to predict student performance. In this study domain experts are asked to evaluate and confirm the effectiveness of the prototype with a comment to improve the performance.

3.4.2. Efficiency

Efficiency can be described as the speed (with accuracy) in which users can complete the tasks for which they use the product. Efficiency metrics include the number of clicks or keystrokes required or the total 'time on task'.

3.4.3. Engaging

An interface is engaging if it is nice and satisfying to use. The graphic design is the clearest element, even within the same class of interfaces; different users may have widely different needs. What is important is that the design meets the expectations and needs of the people who must use the interface.

3.4.4. Easy to Learn

Allow users to build on not only their prior knowledge of computer systems, but also any interaction patterns they have learned through use in a predictable way. Predictability is complementary to interface consistency. A consistent interface ensures that terminology does not change, that design elements and controls are placed in familiar locations and that similar functions behave similarly.

CHAPTER FOUR

PROBLEM UNDERSTANDING AND DATA PREPARATION

4.1. Understanding of the problem domain

Steps in the hybrid knowledge discovery process model include understanding of the problem domain is a stage where the researcher familiarizes with the overall business process or working conditions of the identified work process. To this end, an investigation has been undertaken to understand about problem domain.

The business objective of Higher education institutions (colleges and universities) are faced with a growing number of concerns attracting students, complete a degree program and increasing graduate's donations. Prediction helps to identify students with weak performance and help them to score better marks for improving their performance.

Domain experts were consulted; direct observation of the University existing system was done and reviewing of the different literatures were performed to have brief understanding on the factor affecting student academic performance. Investigation on issues was studied to see a gap where the data mining can be used to fill with the use of Data Mining techniques.

Determinants of students' performance have been identified by many studies [58]. The result of their findings pointed out that hard work; previous schooling, parents' education, family income and self-motivation are factors that have a significant effect on the student's GPA. Cultural differences may play a role in influencing the factors that affect students' performance; it is therefore very important to examine those relevant factors to the Ethiopian culture.

- **GPA:** The use of GPA has a tangible value for future educational and career mobility. It can also be considered as an indication of realized academic potential.
- **Assessment:** It is classified as assignment mark, quizzes, lab work, class test, class attendance, class participation and final exam.
- **Demographic:** Includes gender, age, family background, family size and disability.

- **Gender** is important because they have different styles of female and male students in their learning process. According to Ministry of Education Gender Office, various strategies are employed to increase females' enrollment in all levels of education. Besides, the affirmative action is duly being implemented to increase females' academic performance and to minimize females' attrition rates [59].
- **Age:** The differences in the experiences and maturation of older students involve a relatively better performance in academic.
- **Disability:** students with different types of disabilities academic performance and use of disability support services.
- **High school background:** expected that students with good high school background do better than those less performer students in the background.
- **Instructors:** teaching style has an effect on student's performance. Make sure that effectual teaching learning methods are implemented, provide notes to students and also give private consultation time to help them understand better.
- **Social interaction network:** Students spend more time on social media than focus on learning, which affects greatly their performance on learning.
- **Study Habits:** Study habits of students may be relevant to the prediction of grades because it is possible that student's grades may be related to their study habits. Study skills and learning approaches include, time management, using information resources, taking class notes, communicating with teachers, preparing for and taking examination, and several other learning strategies.
- **Psychometric factor:** is identified as student interest, study behavior, engage time, academic environment and family support.
 - **Physical factors:** When a student is healthy, then he will be able to contribute an active role towards learning. Under this Student's factors group are included such factors as included physical health/illness, physical training/sports activities.
 - **Mental factors:** Attitude falls under mental factors. Students when feel stress and nervousness they can have an effect on a student's academic performance. Alcohol and drug abuse is similarly correlated with poor grades.
 - **Environmental factor:** Physical conditions needed for learning is under environmental factor. One of the factors that affect the efficiency of learning is the

- condition in which learning takes place. This includes the classrooms, textbooks, equipment, school supplies, and other instructional materials.
- **Home Environment:** The importance of home environment or family on students' academic performance. The home has a great influence on the students' psychological, emotional, social and economic state. The family financial support, encouragement and following up have positive impact on students' performance as measured by their GPA.

4.1.1. Overview of the organization, St. Mary's University

St. Mary's University established in 1998 is a private university with four campuses in Addis Ababa, 13 Distance Education Regional Centers, and 160 Coordination Offices throughout Ethiopia, which strive to meet the growing demand for trained manpower [60].

It caters to the needs of six thousand undergraduate students in programs, including Tourism and Hospitality Management, Management, Marketing Management, Accounting, Information Technology, Computer Science, Rural Development, Economics and Sociology. According to the report published by the university [60], twenty thousand students enrolled in distance education programs accessible to the larger society through reasonable tuition focusing on quality and standards in teaching, research and outreach services. There are also two thousand students in graduate programs in six areas of studies: MBA (Master in Business Administration, HRM (Human Resource Management), Accounting and Finance, Agri-business, Agro-economics, and Rural Development and Computer Science [61].

It is one of the leading higher education institutions in Ethiopia that provides quality education in various fields of studies with magnificent dynamism to meet up the rapidly growing demands of students, the industry and the critical need for entrepreneurship in the country.

For quality education, the Ethiopian government established the Higher Education and Relevance Quality Agency (HERQA) to monitor the quality of education provided in higher education institutions. HERQA accredits private institutions but only conducts an institutional audit for public institutions [62].

St. Mary's University mission delivery of quality teaching for its students, research works, academic material production, as well as professional consultancy for the growing needs of Ethiopia and its citizens.

In order to meet and exceed the quality and standard requirements of students and stakeholders, the goals of SMU are the following [60].

- Offer relevant, diverse, learner-centered, and research-led programmed of study;
- Prepare graduates with the requisite knowledge, skills and attitudes embodied in the graduate profile of academic programmed;
- Strengthen assessment methods that validly, reliably and fairly evaluate measurable learning outcomes;
- Promote technology-based, innovative and inter- disciplinary learning environment;
- Augment student support, staff development, facilities and resources;
- Undertake demand-driven research on local, national and international issues and problems;
- Produce and disseminate research outcomes, teaching materials and other publications;
- Ensure the provision of need-based services to the community at large;
- Initiate, sustain and enhance a close network with local and international stakeholders;

The existing system of St. Mary's University to measure their students' performance based on a continuous assessment basis in the form of tests, assignments, class attendance, presentations. To determine the final letter grade, Continuous assessment shall account for 50% of the total course grade. The remaining 50% shall be allotted for a final examination conducted at the end of course delivery. Letter grades are assigned to the marks got out of 100% on a fixed scale (criteria referenced grading system) [63]. The raw marks out of 100% and their equivalent letter grades are indicated in the following table 4.1.

Raw Mark Interval [100 %]	Corresponding fixed Number Grade	Corresponding Letter Grade	Status Description	Class Description
[90,100]	4.0	A+	Excellent	First class with Great distinction
[83, 90)	4.0	A		
[80, 83)	3.75	A–		
[75, 80)	3.5	B+	Very Good	First class with Distinction
[68, 75)	3.0	B		
[65, 68)	2.75	B–	Good	First class
[60, 65)	2.5	C+		Second class
[50, 60)	2.0	C	Satisfactory	Second class
[45, 50)	1.75	C–	Unsatisfactory	Lower class
[40,45)	1.0	D	Very Poor	Lower class
[30,40)	0	Fx	Fail, but possibility for improvement	Lowest class
[<30)	0	F	Fail	Lowest class

Table 4.1: Grading system of undergraduate Students at St Mary’s University [63]

4.1.2. Criteria for student selection

The study determines students’ performance prediction in different department which join higher institute focusing on one university in Ethiopia, St. Mary’s University (SMU). The first step for the students joins in the university from different region of the country is to fill registration form indicating their favorite or interested department. In the institution there are Departments such as Accounting, Management, Marketing Management, Tourism and hospitality management, Computer science and Information Technology in regular and extension division for under graduate school, But the decision-making process was not supported with data mining tools and techniques to enhance the system functionalities by increasing its effectiveness and efficiency.

Specifically, predicting student's performance improves the institution to identify the weakness of their students and provide special attention for their performance improvement.

Based on the investigation on existing problems and suggestion of domain experts, the following attributes are considered to be useful for predicting student performance.

1. **Gender:** female and male students have different learning styles in their learning process.
2. **Age:** Student age when attended in university determines their focus for learning and achieving their objective.
3. **Division:** Extension and regular program students registered to study.
4. **Year:** Academic year of the student attended. Some students are leave their homes and their families for the first time and prepare to face new experiences.
5. **Number of Courses in the Semester:** Total number of credit hours that the students has been taking in the semester whether given courses are minimum, maximum and normal.
6. **Number of Major courses:** Number of major courses that the students taken in the semester.
7. **Number of Supportive courses:** Number of supportive courses that the students taken in the semester.
8. **Number of Common Courses:** Number of common courses that the students taken in the semester.
9. **Marital status:** If a student will be married after filing the Application form, the marital status is single, not married. Married students have much different responsibilities than the unmarried students.
10. **Employment:** Student who has work is didn't attend the class properly because they fill tiredness while the class is conducted so Work has its own impact on student performance.
11. **Financial source:** The self-sponsored students are more a higher level of motivation for attending. Most of the regular program students are sponsored by their family. Student performance affect when any problem happens to their family.
12. **Region:** Region of students where they attended preparatory school. Most of students attended preparatory in Addis Ababa some of from different region of the country.so most of region students are not live with their parents.

- 13. Type of high school:** Expected also that private schools provide better quality education to their students compared to public schools.
- 14. Academic Year:** Academic year student taken EHEEE (Ethiopian higher education entrance examination) exam.
- 15. EHEEE Result:** total result of student gets EHEEE (Ethiopian higher education entrance examination).
- 16. English:** English result in the EHEEE (Ethiopian higher education entrance examination) could have an impact on student.
- 17. Mathematics:** Mathematics result in the EHEEE (Ethiopian higher education entrance examination) could have an impact on student.
- 18. Previous University:** Expected that government University better quality education to their students as compared to private University.
- 19. Previous study Field:** prior knowledge and experience of the student might have impact on student.
- 20. Previous study Program:** prior knowledge and experience might be deferent when student take Degree program and Diploma.
- 21. Previous study GPA:** prior knowledge experience with high performance they get might have impact on student.
- 22. Status:** Current status of students; this attribute is a class.

The main purpose of the study is to apply Data Mining techniques on a student's dataset to predict student's performance using attributes of demographic, internal and external result information. Also identified predicting variables and determined those having a better prediction performance.

In this study WEKA 3.9.2 tool was selected for data mining because the development and application of data mining algorithms require the use of powerful tools. WEKA is selected because it contains tools for data preprocessing, classification, clustering, association rules and visualization. It also uses data file formats like ARFF (attribute relation file format) and CSV (comma separated values) that are accepted by the most applications used for data organization and storage.

4.2. Understanding of the Data

When understanding the problem domain and building a simple plan for solving the problem, the researcher proceeded to the next stage which is data understanding for applying the selected DM techniques. This includes listing out attributes with their respective values and evaluation of their importance for this research and analysis of the data and its structure is done together with domain experts, in order to use the dataset for the Data mining task at hand.

4.2.1. Collect initial data

The initial data for this study were collected after getting permission from the administrations of St Mary's University database. The data collected was about undergraduate students, the table was exported to MS Excel file from Microsoft SQL database.

The study used data collected from St Mary's University undergraduate students database that covers from 2006 - 2009 E.C. this time period is selected because a new admission form was designed with expanded data entry fields and implemented starting from 2006 and required attributes for building the model are found from this database. To understand the data, discussions were made with domain experts from registrar office and university registrar reports and manuals.

4.2.2. Description of the collected data

The decision on the data that was used for the analysis is based on several criteria, including its relevance to the data mining goals. The attributes are selected based on understanding the problem domain with the help of domain expert and literature reviews. Table 4.2 presents the description of the collected attributes with their data type. The final selected data was prepared and preprocessed before developing the model.

No	Attribute Name	Data Type	Description	Value	Missing Value
1	Gender	Nominal	Sex of the student	Male = M, Female = F	0(0.0%)
2	Age	Numeric	Age of the student	[19 – 50]	0(0.0%)
3	Division	Nominal	Student study program	Extension, Regular	0(0.0%)
4	Marital status	Nominal	Student status of marriage	Single, Married, Divorce	0(0.0%)
5	Employment	Nominal	Student employment status	Yes, No	73(1%)
6	Financial source	Nominal	Student Financial source	Self-sponsor, Parent sponsor, organization sponsor, scholarship	80(1%)
7	Year	Nominal	Academic Year	[1, 2, 3, 4]	0(0.0%)
8	Courses Per Semester	Nominal	Total courses given in the semester	[2, 3, 4,...,7]	0(0.0%)
9	Major Courses Per Semester	Nominal	Total Major courses given in the semester	[2, 3, 4,...,6]	0(0.0%)
10	Supportive courses Per Semester	Nominal	Total supportive courses given in the semester	[0,1,2, 3]	0(0.0%)
11	No of Common Courses	Nominal	Total common courses given in the semester	[0,1,2, 3]	0(0.0%)
12	Type of Institution Attended	numeric	Students previously attended Institution	[125-572]	0(0.0%)
13	Region	Nominal	Name of Region	Tigray, Afar, Amhara, Oromiya, Somalia, SNNP, Gambela, Harari, Dire-Dawa, Addis Ababa	316(3%)
14	Attended Years	Date	Students previously attended year	[1984,1985,...2005]	0(0.0%)
15	Attended Result	Numeric	Total Result in EHEEE	Range from 125- 572	0(0.0%)
16	English	Nominal	EHEEE Result student get in English	A,B,C,D	0(0.0%)
17	Mathematics	Nominal	EHEEE Result student get in Mathematics	A,B,C,D,F	0(0.0%)
18	Previous University	Nominal	Student Previously attended University	Addis Ababa, Harmaya, Bahir Dar, Admas, Unity, St Mary's, Tegbared	0(0.0%)
19	Previous study Field	Nominal	Student Previously study Field	Accounting, Management, Marketing Management, Computer science, Information Technology, Engineering, Health, Teaching	0(0.0%)
20	Previous study Program	Nominal	Student Previously study Program	Degree, Diploma	65(1%)
21	Previous study GPA	Numeric	Student Previous GPA	[2 - 3.52]	
22	GPA	Numeric	Total Result students get in the semester	[0.08 – 4]	0(0.0%)

Table 4.2: Description of the selected attributes from St Mary's University dataset

4.3. Preparation of data

The purpose of this data preprocessing/data preparation step is to clean selected data for better quality. Data quality is a multifaceted issue that represents one of the biggest challenges for data mining [19]. It refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed. The presence of duplicate records, the lack of data standards, the timeliness of updates and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to understated differences that may exist in the data [19].

Therefore, each algorithm requires data to be submitted in a discrete format. Data preprocessing/data preparation tasks that include attribute selection, data cleaning mainly handling missing values and attribute reduction are undertaken. It is also concerned with converting from one format to another format to make the dataset understandable and to use it for the Data mining task at hand.

4.3.1. Data Extracted from Database

The data selection method is the first step which includes raw data extraction and selecting the target dataset that are used for the study problem. Selection is the process of choosing the right data from the database on which the tools in data mining can be used to extract useful information, knowledge and pattern from the provided raw data.

This dataset was organized in rows and columns where each column represents an attribute and each row stands for a single record of an individual. The dataset has a total of 22 attributes (columns) and 11550 records (rows) identified; the dataset is presented in Annex 1 and the summary of each of the selected attributes used for model building with their frequency of occurrence are statistically described in detail in Annex 2.

The lists of attributes extracted from database that have been used in this research are shown in the table 4.3 below.

No	Attribute Name	Data Type	Description	Data Values
1	Gender	Nominal	Sex of the students	[M, F]
2	Division	Nominal	Student study time	[Extension, Regular]
3	Year	Numeric	Academic Year	[1, 2, 3, 4]
4	Marital status	Nominal	Student status of marriage	[Single, Married, Divorce]
5	Employment	Nominal	Student employment status	[Yes, No]
6	Financial source	Nominal	Student Financial source	[Self-sponsor, Parent sponsor, organization sponsor, scholarship]
7	Type of high school	Nominal	Students previously attended Institution	[Addis Ababa Government School, Government School, Addis Ababa Private School]
8	Region	Nominal	Name of Region	[Addis Ababa, Amhara, Oromia, SNNP, Harari, Tigray, Dire Dawa, Gambella, Somalia, Afar]
9	Attended Years	Date	Students previously attended year	[Five Subject, Seven Subject, Old Twelve]
10	EHEEE Result	Numeric	Total Result in EHEEE	[Excellent, Very Good, Good, Satisfactory]
11	English	Nominal	EHEEE Result student get in English	[Excellent, Very Good, Good, Satisfactory]
12	Mathematics	Nominal	EHEEE Result student get in Mathematics	[Excellent, Very Good, Good, Satisfactory, Fail]
13	Previous University	Nominal	Student Previously attended institution	[No, Private University, Government TVET College, Government University]
14	Previous study Field	Nominal	Student Previously study Field	[No, Business, Teaching, Health, IT Engineering]
15	Previous study Program	Nominal	Student Previously study Program	[No, Diploma, Degree]
16	Previous study GPA	Numeric	Student Previously GPA	[No, Excellent, Very good, Good Satisfactory]
17	Status (class attribute)	Numeric	Shows student performance	Success, Average, Weak

Table 4.3: Attributes extracted from database

This study derived one attributes from the existing original dataset and four attributes not registered in the database are added based on the domain expert’s advice and values for this research objective. Those attributes for the data mining task of this research were: age, number of courses per semester, number of major courses, number of common courses, and number of supportive courses as shown in table 4.4.

- ❖ **Age:** Age attribute was derived from date of birth filled with application form. We derived this attribute because it helps us to determine students’ status with respect to their age and convert into age category to simplify analysis.
- ❖ **Number of Courses per Semester:** it is counted from academic year and from semesters in the academic year. As per the discussion with experts and with their academic schedule, we added this attribute because it helps us to determine students’ status with respect to the number of courses students taken per Semester.
- ❖ **Number of Major Courses:** it is counted from students taken number of Courses per Semester. We added this attribute because it helps us to determine students’ status with respect to major courses taken by students in a given semester.
- ❖ **Number of Common Courses:** it is counted from students taken number of Courses per Semester. We added this attribute because it helps us to determine students’ status with respect to number of Common courses in the semester the student taken.
- ❖ **Number of Supportive Courses:** it is counted from students taken number of Courses per Semester. We added this attribute because it helps us to determine students’ status with respect to supportive courses per semester students taken.

No	Attribute Name	Data Type	Description	Data Values
1	Age (derived)	continues	Age of the students	[19 – 50]
2	Number of Courses Per Semester	Nominal	Number of courses given in the semester	[Two, Three, Four, Five, Six, Seven]
3	Number of Major Courses	Nominal	Number of major courses given to in the semester	[Two, Three, Four, Five, Six]
4	Number of Common Courses	Nominal	Number of common courses given to in the semester	[Zero ,One, Two, Three]
5	Number of Supportive Courses	Nominal	Number of supportive courses given to in the semester	[Zero ,One, Two, Three]

Table 4.4: Selected attributes with their description

4.3.2. Data Cleaning

Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning procedures fill the missing values and correct inconsistencies in the data.

4.3.2.1. Missing Values

To analyze student's data, there are many attributes with no recorded values. Hence there is a need to fill the missing values for attribute [11] using either of the following ways.

- Ignore the tuple usually done when the class label is missing
- Fill in the missing value manually this approach is time consuming
- Use a global constant to fill in the missing value. Label like “Unknown” or $-\infty$.
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class as the given tuple
- Use the most probable value to fill in the missing value

In the original dataset that some attributes contain missing values, as presented in table 4.5 in this dataset missing values were observed in nominal variables. The missing values were replaced by the most common value of an attribute using mode. Microsoft Excel version 2019 was used for handling the missing values.

No	Attribute	Frequency of missing value	Percentage of missing values	Fill missing value
1	Employment	73	1%	the most common values using mode
2	Financial source	80	1%	the most common values using mode
3	Region	316	3%	the most common values using mode
4	Previous study Program	65	1%	the most common values using mode

Table 4.5: Attributes with missing values and replaced by mode

4.3.3. Data Transformation

Data transformation is the process of discretizing continuous value. In this page, the continuous valued age, Ethiopian higher education entrance examination English/Mathematics result and status were discretized as shown in table 4.6, 4.7 and table 4.8 respectively.

In consultation with domain experts the age ranges of students joined the University are divided into N intervals. After completing the discretization process distinct values of the age attribute were reduced to 4 from 28 distinct values.

Age	Represented value
19 - 26	Teenager
27 - 34	Young
35 - 42	Middle age
43 - 50	Adult

Table 4.6: Discretized age attribute

English/Mathematics	Represented value
A	Excellent
B	Very Good
C	Good
D	Satisfactory
F	Fail

Table 4.7: Discretized English/Mathematics EHEEE result attribute

The same was done with status score to convert into the stated success, average and weak.

Status	Represented value
[2.75 – 4]	Success
[2.0 - 2.75)	Average
[0.08 – 2.0)	Weak

Table 4.8: A discretized attribute Status

4.3.4. Setting the Attribute Class

In supervised classification technique predefined classes are required in order to train and test classification models. The setting of a predefined class is done intentionally because the technique for this study is J48 Decision Tree, PART Rule induction algorithm, Naïve Bayes Bayesian, Logistic regression, (SMO) Support Vector Machines and Multilayer Perception Neural Network classifications. In order to classify students' performance, the target attribute selected is status of students, which contains three outcome instances; Success, Average and Weak.

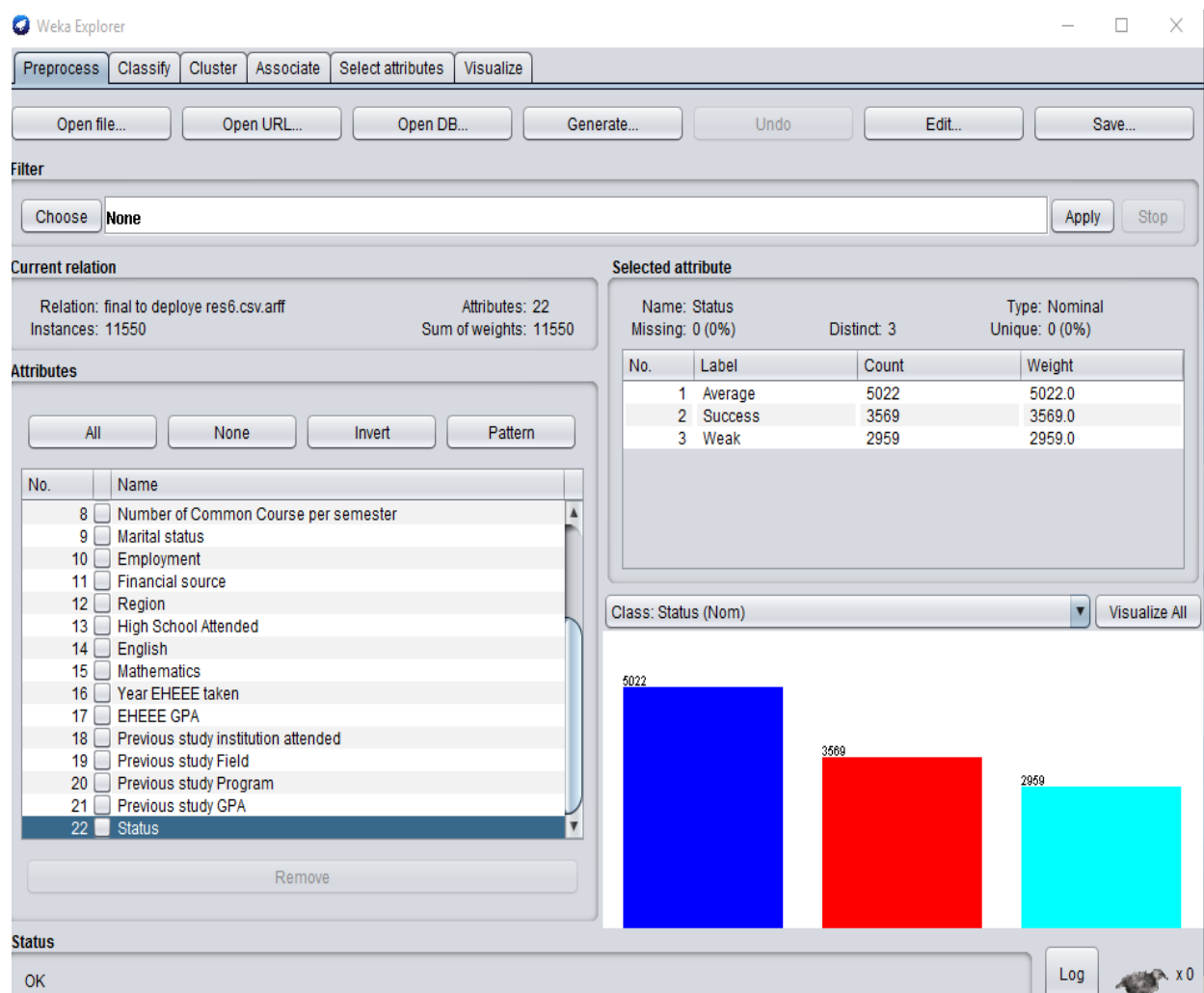


Figure 4.1: List of selected attributes in WEKA 3.9.2 explorer window

4.3.5. Data Preparation for WEKA Software

The raw data initially in SQL database extracted to Microsoft Excel 2019 format which is not understandable by the WEKA tool. To make such data format understandable by WEKA saved the exported Microsoft Excel file is saved as a CSV (Comma Separate Values) file format (see figure 4.2) to make ready for an input to WEKA 3.9.2 data mining tool. The last task done to make the data format suitable for WEKA tool was converting the CSV file into ARFF (Attribute-Relation File Format) file format.

```
Division,Gender,Age,Year,Total course per semester,Number of
Major Course per semester,Number of Supportive Course per
semester,Number of Common Course per semester,Marital
status,Employment,Financial source,Region,High School
Attended,English,Mathematics,Year EHEEE taken,EHEEE GPA,Previous
study institution attended,Previous study Field,Previous study
Program,Previous study GPA,Status
Extension,F,Age Three,Three,Three,Three,Zero,Zero,Single,Yes,Self
Sponsored,Addis Ababa,Addis Ababa Government,Good,Good,Old
Twelve,Old Twelve,Private University,No,Diploma,No,Average
Regular,F,Age One,One,Six,Two,Two,Two,Single,No,Self
Sponsored,Addis Ababa,Addis Ababa
Government,Excellent,Satisfactory,Seven
Subject,Good,No,No,No,No,Average
Regular,M,Age Two,Two,Three,Three,Zero,Zero,Single,No,Parent
Sponsored,Addis Ababa,Government,Good,Good,Five
Subject,Good,No,No,No,No,Average
Extension,F,Age Two,Three,Four,Three,One,Zero,Single,Yes,Parent
Sponsored,Tigray,Addis Ababa
Government,Satisfactory,Satisfactory,Five Subject,Grade
Ten,Private University,No,Diploma,No,Average
Regular,F,Age Two,Two,Six,Three,Two,One,Single,No,Parent
Sponsored,Oromia,Addis Ababa Government,Good,Good,Seven
Subject,Good,No,No,No,No,Average
Regular,M,Age Two,One,Seven,Three,One,Three,Single,No,Parent
Sponsored,Addis Ababa,Addis Ababa Private,Good,Good,Five
Subject,Good,No,Teaching,No,Satisfactory,Average
Extension,M,Age Two,Two,Four,Three,Zero,One,Single,Yes,Parent
Sponsored,Addis Ababa,Addis Ababa Government,Good,Good,Old
Twelve,Old Twelve,Government TVET College,No,Diploma,No,Average
```

Figure 4.2: Sample CSV format data sets prepared for WEKA

CHAPTER FIVE

EXPERIMENTATION AND EVALUATION

This chapter discusses the experiments, experimental results and model building shown in the study. As mentioned earlier the main objective of this study is to predict higher education student's performance in data mining classification techniques were applied to develop predictive models. Therefore, it is important to conduct different experiments to find the best model for solving the problem. Also, different experiments were evaluated their performance with the output from the algorithms.

5.1. Model Building

Modeling is one of the tasks undertaken under the phase of hybrid data mining process model. In this phase, different techniques can be used for the data mining problems. The tasks include: selecting the modeling technique, experimental setup, building a model and evaluating the model.

5.1.1. Selecting the modeling technique

In this work an attempt was done to build a model using selected algorithms for classification of higher education student's performance. To achieve the objectives of this study, six classification techniques have been selected for model building. The analysis was performed using WEKA environment. Among the different available classification algorithms in WEKA, J48 algorithms from decision tree, PART algorithms from rule induction, Naïve Bayes algorithm from Bayesian, Logistic regression algorithm from regression, Sequential Minimal Optimization (SMO) algorithm from Support Vector Machines and Multilayer Perception algorithm from Neural Network are selected for experimentation.

The reason for selecting the above algorithms is its popularities in recently published papers, easy to understand and interpret the result of the model for the studies. In this study for creating a predictive model, total size of 11550 records and 22 attributes including class are used for training and testing.

5.1.1.1. WEKA Interface

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tool [64].



Figure 5.1: WEKA interface

5.1.1.2. Balancing the Dataset

In class imbalance problems, the number of examples of one class (minority class) is much smaller than the number of examples of the other classes, with the minority class being the class of greatest interest and that with the biggest error cost from the point of view of learning [65]. Therefore, to solve the problem of class imbalance we applied supervised resampling technique by changing the default values of (biasToUniformClass (0.0) into (1.0)). As shown in figure 5.2 after we applied to resample the three dependent classes become equal with each class value success, average and weak having 3850 instances.

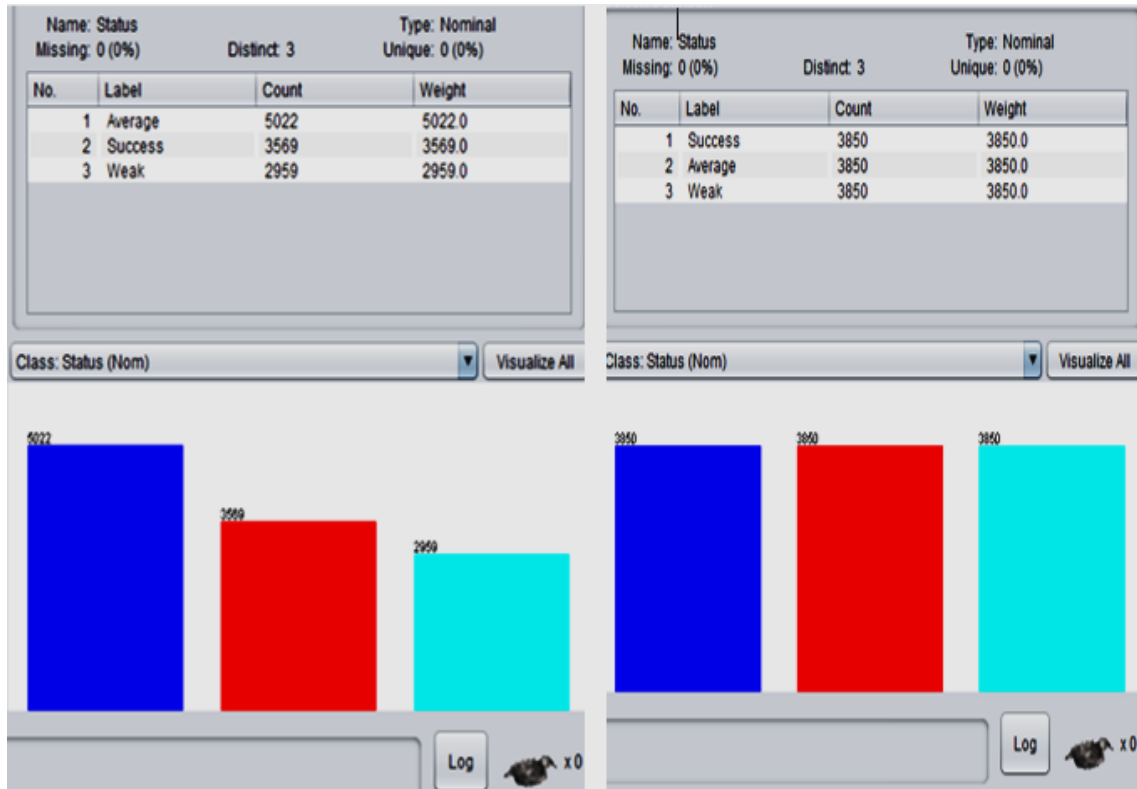


Figure 5.2: Side by side view of the class (left side) Original data; (right side) balanced data.

5.1.1.3. Attribute Ordering

Since attribute selection is important because in most cases all attributes are not equally useful for predicting the target so we need to first evaluate the usefulness of each attribute before conducting the experiment of classification in order of this the tried to rank the attribute based on information gain. It was calculated based on entropy value of the attribute.

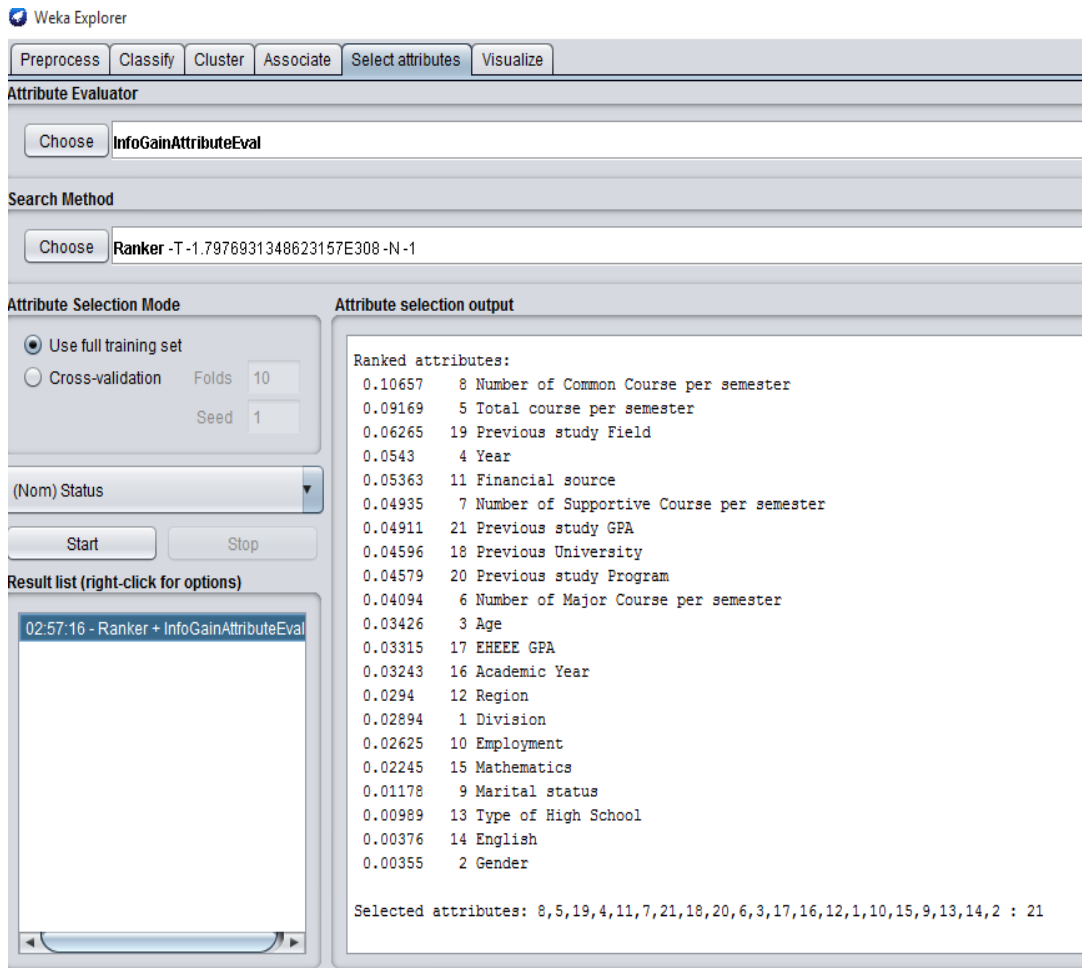


Figure 5.3: output of attribute ranking with information gain

As show the figure above depict ranked order of attribute based on their relevance for the reason that such attributes are very important for later experimentations.

5.2. Experimental Setup

The experiment is conducted using the default 10-fold cross validation and 66% percentage split. In 10-fold cross validation, the initial data are randomly partitioned into 10 mutually exclusive subsets or folds, 1,2,3,4 ...10, each approximately with equal size. The training and testing is performed 10 times. In the first iteration, the first fold is reserved as a test set, and the remaining 9 folds are collectively used to train the classifier. In percentage split the default ratio is 66% for training and 34% for testing.

5.2.1. Classification Using J48 Decision Tree

One of the classification techniques applied for building the classification model in this study is the J48 algorithm with the default parameter. J48 algorithm contains some parameters such as confidence factor, pruning and un-pruning, hanging the generalized binary split decision classification that can be changed to further improve classification accuracy as shown in table 5.1 and in annex 3.

Parameters	Default value	Description	Types
binarySplits	False	Whether to use binary splits on nominal attributes when building trees	Boolean
confidenceFactor	0.25	The confidence factor used for pruning (smaller values incur more pruning)	Numeric
minNumObj	2	The minimum number of instances per leaf	Numeric
unpruned	False	Whether pruning is performed	Boolean

Table 5.1: Some of J48 decision tree algorithm parameters with their values

For J48 decision tree four test modes were considered:

- **Experiment 1:** Pruned J48 algorithm with 10 fold cross-validation test option.
- **Experiment 2:** Pruned J48 algorithm using resampling with a 10-fold cross-validation test option.
- **Experiment 3:** Pruned J48 algorithm with Percentage (%) split test which is train a model and then supply the unseen remaining part of the record for testing the performance of the model.
- **Experiment 4:** Pruned J48 algorithm using resampling with Percentage (66%) split test option.

5.2.2. Classification Using PART Rule Induction

The second data mining technique used in this study is PART Rule induction algorithm. PART algorithm extracts rules. Due to this reason the algorithm is categorized under classification by rule induction. The algorithm builds partial decision trees and reads a path from the root of the tree to the leaf to read a rule. The rules are added together to give a complete set of rules. PART has almost a similar set of parameters with J48 algorithm that can be adjusted to build better model from datasets. Four experiments were conducted using all variables. In the experiment the 10-fold cross validation and percentage split was used for all experiments.

- **Experiment 1:** PART algorithm with 10-fold cross-validation test mode
- **Experiment 2:** PART algorithm using resampling with a 10-fold cross-validation test mode.
- **Experiment 3:** PART algorithm with Percentage (%) split test mode
- **Experiment 4:** PART algorithm using resampling with Percentage (%) split test mode.

5.2.3. Classification Using Naïve Bayes

The third type of classification technique applied in this study is the Naïve Bayes algorithm. Four experiments were conducted using all variables. In the experiment the 10-fold cross validation and percentage split was used for all experiments. The Naïve Bayes experiment was designed to build the model for predicting student performance and to compare the performance with J48 algorithm and PART algorithm.

- **Experiment 1:** Naïve Bayes algorithm with 10-fold cross-validation test mode
- **Experiment 2:** Naïve Bayes algorithm using resampling with a 10-fold cross-validation test mode.
- **Experiment 3:** Naïve Bayes algorithm with Percentage (%) split test mode
- **Experiment 4:** Naïve Bayes algorithm using resampling with Percentage (%) split test mode.

5.2.4. Classification Using Logistic Regression

The fourth type of classification technique applied in this study is the linear regression algorithm. Four experiments were conducted using all variables. In the experiment the 10-fold cross validation and percentage split was used for all experiments. The linear regression experiment was designed to build the model for predicting student performance and to compare the performance with others selected algorithms.

- **Experiment 1:** linear regression algorithm with 10-fold cross-validation test mode
- **Experiment 2:** linear regression algorithm using resampling with a 10-fold cross-validation test mode.
- **Experiment 3:** linear regression algorithm with Percentage (%) split test mode
- **Experiment 4:** linear regression algorithm using resampling with Percentage (%) split test mode.

5.2.5. Classification Using Sequential Minimal Optimization

The fifth type of classification technique applied in this study is the Sequential Minimal Optimization algorithm. Four experiments were conducted using all variables. In the experiment the 10-fold cross validation and percentage split was used for all experiments. The Sequential Minimal Optimization experiment was designed to build the model for predicting student performance and to compare the performance with others selected algorithms.

- **Experiment 1:** Sequential Minimal Optimization algorithm with 10-fold cross-validation test mode
- **Experiment 2:** Sequential Minimal Optimization algorithm using resampling with a 10-fold cross-validation test mode.
- **Experiment 3:** Sequential Minimal Optimization algorithm with Percentage (%) split test mode
- **Experiment 4:** Sequential Minimal Optimization algorithm using resampling with Percentage (%) split test mode.

5.2.6. Classification Using Multilayer Perception

The sixth type of classification technique applied in this study is the Multilayer Perception algorithm. Four experiments were conducted using all variables. In the experiment the 10-fold cross validation and percentage split was used for all experiments. The Multilayer Perception experiment was designed to build the model for predicting student performance and to compare the performance with others selected algorithms.

- **Experiment 1:** Multilayer Perception algorithm with 10-fold cross-validation test mode
- **Experiment 2:** Multilayer Perception algorithm using resampling with a 10-fold cross-validation test mode.
- **Experiment 3:** Multilayer Perception algorithm with Percentage (%) split test mode
- **Experiment 4:** Multilayer Perception algorithm using resampling with Percentage (%) split test mode.

5.3. Experimental Result

5.1.1. Experimenting with J48 decision tree

Four experiments are conducted using J48 decision tree by changing test mode and by applying resampling techniques.

Experiment with J48 decision tree using default 10-fold cross validation

The first two experiments are conducted with J48 default 10-fold cross validation. The default 10-fold cross validation test option is employed for training and testing the classification model. The result of the experiment 1 shows that the experiment has generated a model with a tree size of 671 and 907 leaves. The experiment has generated a model with an accuracy of 81.71%, weighted precision of 82.5%, weighted Recall of 81.7%, weighted F-Measure of 81.7% and weighted ROC area of 91.8%. From the total instances, 81.71 % are correctly classified and 18.29% are misclassified. The performance of the two experiments is summarized and presented in the table 5.2 below.

Experiment	Model	Accuracy	Leaf Size	Tree Size	Time Taken	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
1	J48 Pruned with 10 fold cross-validation	81.71 %	671	907	0.09	0.817	0.09	0.825	0.817	0.817	0.918
2	J48 Pruned using Resampling with 10 fold cross-validation	97.77 %	495	672	0.36	0.978	0.011	0.978	0.978	0.978	0.998

Table 5.2: Performance result of J48 Decision tree with 10-fold cross validation

In the second experiment as shown in the above table 5.2, the same number of attributes and records are used but the data is balanced using resampling technique. Accordingly, there is an improvement of performance with accuracy, weighted precision, weighted F-Measure and weighted ROC area of 97.77%, 97.80%, 97.80 and 97.80% respectively. Running information of J48 algorithm with 10-fold validation technique is provided on annex-4.

Experimenting with J48 decision tree using percentage split

This experiment is performed, by changing the 10-fold cross validation to percentage split of 66%. The use of this parameter was to assess the performance of the algorithm by changing the 10-fold cross validation to the default value of the percentage split (66%). The result of experiment three and four presented below in table 5.3.

Experiment	Model	Accuracy	Leaf Size	Tree Size	Time Taken	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
3	J48 pruned Percentage (%) split	80.26%	671	907	0.09	0.803	0.1	0.809	0.803	0.803	0.91
4	J48 pruned using Resampling Percentage (%) split	97.84%	495	672	0.16	0.978	0.011	0.978	0.978	0.978	0.998

Table 5.3: Performance result of J48 Decision tree with percentage split (66%)

In experiment three out of the total records 66% of the records are used for training purpose while 34% of the records are used for testing purpose. As shown in table 5.3, the model developed with this percentage results in 80.26% correctly classified instances and 19.74% incorrectly classified instances.

The fourth experiment is done after resampling the data. As shown in above table 5.3, correctly classified instances are 97.84% and the remaining 2.16% are misclassified. Running information of J48 algorithm with percentage split of 66% is provided on annex-5

Generally, from the four experiments conducted before using J48 decision tree, the model developed with the percentage split (66%) using resampling technique gives a best classification accuracy of 97.84 % predicting performance. Therefore, among the different decision tree models built in the foregoing experimentations, the fourth model, with the percentage split (66%) using resampling, has been chosen due to its better classification accuracy.

In the process of building model and finding the best model measures like adjusting the values of the parameters were also taken. Experiments carried out by varying the parameters of the algorithm, but changing the default parameters was not resulting in a significant change to the performance of the model. Therefore the researcher decided to proceed the experiment by using the default parameters.

5.1.2. Experimenting PART Rule Induction

We conducted four experiments using PART rule induction by applying 10-fold cross validation and percentage split, as well as resampling technique.

Experiment with PART Rule Induction using default 10-fold cross validation

To build the PART Rule induction model, 11550 dataset was used as an input. The PART Rule induction algorithm with 10-fold cross validation scored an accuracy of 82.71. This result shows that out of the total training datasets 82.71 % records are correctly classified instances, and the remaining 17.29 % of the records are incorrectly classified. The result of PART Rule induction

algorithm using resampling technique and without balancing the dataset with 10-fold cross validation is shown in table 5.4 below.

Experiment	Model	Accuracy	Number of Rules	Time Taken	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
1	PART Pruned with 10 fold cross-validation	82.71 %	189	0.53	0.827	0.083	0.836	0.827	0.827	0.922
2	PART Pruned using Resampling with 10 fold cross-validation	97.48 %	226	0.67	0.975	0.013	0.975	0.975	0.975	0.997

Table 5.4: Performance results of PART Rule Induction with 10 cross validation

In the second experiment as shown in the above table 5.4, the same number of attributes and records are used but the data is balanced using resampling technique. In this case, correctly classified instances are 97.48% and the resampling 2.52% are misclassified. Running information of PART algorithm with 10-fold validation technique is provided on annex-6.

Based on the above experiment, PART algorithm using resampling technique with 10 cross validation has scored a better accuracy of 97.48 % with weighted precision, weighted F-Measure and weighted ROC area of the model were 97.48%, 97.5%, 97.5% and 99.7% respectively than PART algorithm without using resampling technique. Therefore, the PART algorithm using resampling technique with 10 cross validation been selected for comparing with other classifier results.

Experiment with PART Rule Induction using percentage split

The third and fourth experiment is performed, by changing the default testing option of 10-fold cross validation to the percentage split (66%). The outcome of these experiments is presented in table 5.5.

Experiment	Model	Accuracy	Number of Rules	Time Taken	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
3	PART pruned Percentage (%) split	81.23 %	189	0.55	0.812	0.093	0.82	0.812	0.812	0.911
4	PART pruned using Resampling Percentage (%) split	97.17 %	226	0.01	0.972	0.014	0.972	0.972	0.972	0.995

Table 5.5: Performance result of PART Rule Induction with percentage split

As shown in table 5.5, PART rule induction without applying resampling techniques 81.23% correctly classified instances, while only 18.77% of the records are incorrectly classified instances. On the other hand, PART rule induction using resampling technique with percentage split scored an accuracy of 97.17 % while only 2.83% of the records are incorrectly classified instances. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 97.2%, 97.2%, 97.2% and 99.5% respectively.

The resulting confusion matrix, of the PART algorithm using resampling technique with default percentage split (66%).

Generally, after conducting the four experiments using PART rule induction the model developed with 10-fold cross validation resampled data registered best classification result of 97.48%. Therefore, among the different PART rule induction models the second model, with the 10-fold cross validation using resampling, has been chosen due to its better classification accuracy.

5.1.3. Experimenting Naïve Bayes classification algorithm

In this case also we conducted four experiments using Naïve Bayes by changing testing modes and by applying resampling techniques.

Experiment with Naïve Bayes using default 10-fold cross validation

This experiment was conducted using all attributes to build the model with default 10-fold cross validation. The Naïve Bayes model built is correctly Classified Instances 53.89 % while only 46.11% instances were classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 54.0%, 53.9%, 53.9% and 72.5% respectively.

The overall performance of this test has lowered from what has scored in the previous experiment of J48 decision tree and PART rule induction.

Experiment	Model	Accuracy	Time Taken to test model	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
1	Naïve Bayes Pruned with 10 fold cross-validation	53.89 %	0	0.539	0.255	0.540	0.539	0.539	0.725
2	Naïve Bayes Pruned using Resampling with 10 fold cross-validation	52.34 %	0.05	0.523	0.238	0.528	0.523	0.513	0.717

Table 5.6: Performance result of Naïve Bayes with default 10-fold cross validation

As shown in the above table 5.6, the second experiment conducted Naïve Bayes using resampling technique with 10-fold cross validation scored an accuracy of 52.34 %. This result shows that out of the total training datasets 52.34 % records are correctly classified instances and remaining 47.66% of the records are incorrectly classified instances. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 52.8%, 52.3%, 52.3% and 71.7% respectively.

Based on the above experiment, Naïve Bayes algorithm with 10 cross validation has scored a better accuracy than Naïve Bayes algorithm using resampling technique with 10 cross validation. Running information of Naïve Bayes algorithm with 10-fold validation technique is provided on annex-7.

Experiment with Naïve Bayes using percentage split

This experiment is performed, by changing the default testing option of 10-fold cross validation to the percentage split (66%). The Naïve Bayes model built is correctly Classified Instances 53.88 % and 46.12% instances are classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 54.1%, 53.9%, 53.9% and 71.6% respectively. The outcome of this experiment is presented in table 5.7.

Experiment	Model	Accuracy	Time Taken to test model	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
3	Naïve Bayes pruned Percentage (%) split	53.88 %	0.05	0.539	0.251	0.541	0.539	0.539	0.716
4	Naïve Bayes pruned using Resampling Percentage (%) split	52.10 %	0.05	0.521	0.238	0.529	0.521	0.511	0.717

Table 5.7: Performance result of Naïve Bayes algorithm with default percentage split

As shown in the above table 5.7, the Fourth experiment conducted Naïve Bayes using resampling technique with percentage split scored an accuracy of 52.10 %. This result shows that out of the total training datasets 52.10 % records are correctly classified instances and remaining 47.89% of the records are incorrectly classified instances. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 52.9%, 52.1%, 51.1% and 71.7% respectively.

Based on the above experiment, Naïve Bayes algorithm with percentage split (66%) has scored a best result of 53.89% than Naïve Bayes algorithm using resampling technique with percentage split (66%). Therefore, among the different Naïve Bayes models the first model; with the 10-fold cross validation has been chosen due to its better classification accuracy.

5.1.4. Experimenting Logistic Regression classification algorithm

In this case also we conducted four experiments using logistic regression by changing testing modes and by applying resampling techniques.

Experiment with Logistic Regression using default 10-fold cross validation

This experiment was conducted using all attributes to build the model with default 10-fold cross validation. The logistic regression model built is correctly Classified Instances 62.23 % while only 37.76% instances were classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 62.5%, 62.2%, 41.0% and 67.2% respectively. The overall performance of this test has lowered from what has scored in the previous experiment of J48 decision tree.

Experiment	Model	Accuracy	Time Taken	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
1	Logistic regression Pruned with 10 fold cross-validation	62.23 %	8 sec	0.622	0.217	0.625	0.622	0.410	0.672
2	Logistic regression Pruned using Resampling with 10 fold cross-validation	64.35 %	9.97	0.644	0.178	0.643	0.644	0.465	0.731

Table 5.8: Performance result of Logistic Regression with default 10-fold cross validation

As shown in the above table 5.8, the second experiment conducted logistic regression using resampling technique with 10-fold cross validation scored an accuracy of 64.35 %. This result shows that out of the total training datasets 64.35 % records are correctly classified instances and remaining 35.64% of the records are incorrectly classified instances. constructed a model with

weighted precision, weighted F-Measure and weighted ROC area of the model were 64.3%, 64.4%, 46.5% and 73.1% respectively.

Based on the above experiment, logistic regression algorithm with 10 cross validation has scored a better accuracy than logistic regression algorithm using resampling technique with 10 cross validation.

Experiment with Logistic Regression using percentage split

This experiment is performed, by changing the default testing option of 10-fold cross validation to the percentage split (66%). The logistic regression model built is correctly Classified Instances 62.10 % and 37.89% instances are classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 62.1%, 62.1%, 40.7% and 67.3% respectively. The outcome of this experiment is presented in table 5.9.

Experiment	Model	Accuracy	Time Taken	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
3	Logistic regression pruned Percentage (%) split	62.10%	0.03	0.621	0.218	0.621	0.621	0.407	0.673
4	Logistic regression pruned using Resampling Percentage (%) split	64.65%	0.02	0.647	0.176	0.646	0.647	0.470	0.738

Table 5.9: Performance result of Logistic Regression algorithm with default percentage split

As shown in the above table 5.9, the Fourth experiment conducted logistic regression using resampling technique with percentage split scored an accuracy of 64.65 %. This result shows that out of the total training datasets 64.65 % records are correctly classified instances and remaining 35.34% of the records are incorrectly classified instances. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 64.6%, 64.7%, 47.0% and 73.8% respectively.

Based on the above experiment, logistic regression algorithm using resampling technique with percentage split (66%) has scored a best result of 64.65% than logistic regression algorithm with percentage split (66%). Therefore, among the different logistic regression models the first model, with the percentage split (66%) using resampling, has been chosen due to its better classification accuracy. Running information of logistic regression algorithm with percentage split (66%) technique is provided on annex-8.

5.1.5. Experimenting Sequential Minimal Optimization classification algorithm

In this case also we conducted four experiments using Sequential Minimal Optimization by changing testing modes and by applying resampling techniques.

Experiment with Sequential Minimal Optimization using default 10-fold cross validation

This experiment was conducted using all attributes to build the model with default 10-fold cross validation. The Sequential Minimal Optimization model built is correctly Classified Instances 62.46 % while only 37.54 % instances were classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 63.1%, 62.5%, 62.4% and 74.6% respectively.

Experiment	Model	Accuracy	Time Taken	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
1	Sequential Minimal Optimization Pruned with 10 fold cross-validation	62.458 9 %	446.8 9 sec	0.625	0.221	0.631	0.625	0.624	0.746
2	Sequential Minimal Optimization Pruned using Resampling with 10 fold cross-validation	67.116 9 %	545.6 1 sec	0.671	0.164	0.671	0.671	0.671	0.792

Table 5.10: Performance result of SMO with default 10-fold cross validation

As shown in the above table 5.10, the second experiment conducted Sequential Minimal Optimization using resampling technique with 10-fold cross validation scored an accuracy of 67.11%. This result shows that out of the total training datasets 67.11 % records are correctly classified instances and remaining 32.88% of the records are incorrectly classified instances. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 67.1%, 67.1%, 67.1% and 79.2% respectively.

Based on the above experiment, Sequential Minimal Optimization algorithm with 10 cross validation has scored a better accuracy than Sequential Minimal Optimization algorithm using resampling technique with 10 cross validation.

Experiment with Sequential Minimal Optimization using percentage split

This experiment is performed, by changing the default testing option of 10-fold cross validation to the percentage split (66%). The Sequential Minimal Optimization model built is correctly Classified Instances 61.49 % and 38.5% instances are classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 62.6%, 61.5%, 61.4% and 74.0% respectively. The outcome of this experiment is presented in table 5.11.

Experiment	Model	Accuracy	Time Taken	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
3	Sequential Minimal Optimization pruned Percentage (%) split	61.4973 %	470.05 sec	0.615	0.231	0.626	0.615	0.614	0.740
4	Sequential Minimal Optimization pruned using Resampling Percentage (%) split	67.125 %	13409.4 sec	0.671	0.165	0.670	0.671	0.669	0.795

Table 5.11: Performance result of SMO algorithm with default percentage split

As shown in the above table 5.11, the Fourth experiment conducted Sequential Minimal Optimization using resampling technique with percentage split scored an accuracy of 67.12 %. This result shows that out of the total training datasets 67.12 % records are correctly classified instances and remaining 32.87% of the records are incorrectly classified instances. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 67.1%, 67.1%, 66.9% and 79.5% respectively.

Based on the above experiment, Sequential Minimal Optimization algorithm using resampling technique with percentage split (66%) has scored a best result of 67.12% than Sequential Minimal Optimization algorithm with percentage split (66%). Therefore, among the different Sequential Minimal Optimization models the first model, with the percentage split (66%) using resampling, has been chosen due to its better classification accuracy. Running information of Sequential Minimal Optimization algorithm with percentage split (66%) technique is provided on annex-9.

5.1.6. Experimenting Multilayer Perception classification algorithm

In this case also we conducted four experiments using Multilayer Perception by changing testing modes and by applying resampling techniques.

Experiment with Multilayer Perception using default 10-fold cross validation

This experiment was conducted using all attributes to build the model with default 10-fold cross validation. The Multilayer Perception model built is correctly Classified Instances 81.79 % while only 18.20 % instances were classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 82.4%, 81.8%, 81.8% and 92.1% respectively.

Experiment	Model	Accuracy	Time Taken	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
1	Multilayer Perception Pruned with 10 fold cross-validation	81.792 2 %	466.3 9 sec	0.818	0.090	0.824	0.818	0.818	0.921
2	Multilayer Perception Pruned using Resampling with 10 fold cross-validation	97.8 %	532.0 8 sec	0.978	0.011	0.978	0.978	0.978	0.996

Table 5.12: Performance result of Multilayer Perception with default 10-fold cross validation

As shown in the above table 5.12, the second experiment conducted Multilayer Perception using resampling technique with 10-fold cross validation scored an accuracy of 97.80 %. This result shows that out of the total training datasets 97.80 % records are correctly classified instances and remaining 2.19% of the records are incorrectly classified instances. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 97.8%, 97.8%, 97.8% and 99.6% respectively.

Based on the above experiment, Multilayer Perception algorithm with 10 cross validation has scored a better accuracy than Multilayer Perception algorithm using resampling technique with 10 cross validation.

Experiment with Multilayer Perception using percentage split

This experiment is performed, by changing the default testing option of 10-fold cross validation to the percentage split (66%). The Multilayer Perception model built is correctly Classified Instances 80.21 % and 19.78 % instances are classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 80.8%, 80.2%, 80.3% and 91.5% respectively. The outcome of this experiment is presented in table 5.13.

Experiment	Model	Accuracy	Time Taken	Weighted TP Rate	Weighted FP Rate	Weighted Precision	Weighted Recall	Weighted F-Measure	Weighted ROC Area
3	Multilayer Perception pruned Percentage (%) split	80.2139 %	470.4 sec	0.802	0.102	0.808	0.802	0.803	0.915
4	Multilayer Perception pruned using Resampling Percentage (%) split	97.5809 %	459.3 sec	0.976	0.012	0.976	0.976	0.976	0.992

Table 5.13: Performance result of Multilayer Perception algorithm with percentage split

As shown in the above table 5.13, the Fourth experiment conducted Multilayer Perception using resampling technique with percentage split scored an accuracy of 97.58 %. This result shows that out of the total training datasets 97.58 % records are correctly classified instances and remaining 2.41% of the records are incorrectly classified instances. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 97.6%, 97.6%, 97.6% and 99.2% respectively.

Based on the above experiment, Multilayer Perception algorithm using resampling technique with percentage split (66%) has scored a best result of 97.58% than Multilayer Perception algorithm with percentage split (66%). Therefore, among the different Multilayer Perception models the first model, with the percentage split (66%) using resampling, has been chosen due to its better classification accuracy. Running information of Multilayer Perception algorithm with 10-fold cross validation technique is provided on annex-10.

5.4. Comparison of Classification Models

Selecting a better classification technique for building a model, which performs best in the prediction of students' performance are one of the aims of this study. For that reason, six classification techniques i.e. Decision Tree, Rule Induction, Bayesian, Regression, Support Vector Machine and Neural Network were applied. Six algorithms were selected for the implementation of classification modeling namely; J48, PART, Naïve Bayes, Logistic Regression, Sequential Minimal Optimization and Multilayer Perception. Then, four experiments were conducted for each algorithm, and the obtained results were compared. For each algorithm, the best model with heights accuracy is selected and presented in table 5.14 below.

Type of algorithm	Test mode	Accuracy	Recall	Precision	F-Measure	ROC
J48 decision tree	Percentage split (66%) using resampling Technique	97.84%	97.80%	97.80%	97.80%	97.80%
PART rule induction	10 fold cross validation using resampling Technique	97.48%	97.50%	97.5%	97.50%	99.7%
Naïve Bayes	10 fold cross validation without using resampling Technique	53.89%	53.9%	54.0%	53.9%	72.5%
Logistic regression	Percentage split (66%) using resampling Technique	64.65%	64.7%	64.6%	47.0%	73.8%
Sequential Minimal Optimization	Percentage split (66%) using resampling Technique	67.12%	67.0%	67.1	66.9%	79.5%
Multilayer Perceptron	10 fold cross validation using resampling Technique	97.80%	97.8 %	97.8 %	97.8 %	99.6%

Table 5.14: Performance Comparison of the selected models

As it is shown in above Table 5.14, among the six algorithms J48 tree algorithm using resampling technique with percentage split (66%) performed the highest accuracy of 97.84%. As a result, we selected the result of J48 decision tree with percentage split and using resampling technique as a final model for the study.

The confusion matrix of the selected model of J48 decision tree algorithm using resample technique is shown in table 5.15 below.

==== Confusion Matrix ====			
A	B	C	classified as
1313	7	18	A = Success
3	1287	16	B = Average
5	36	1242	C = Weak

Table 5.15: Confusion Matrix for J48 algorithm using resampling with percentage split

There is a lot to be learned from closely examining the errors made by a classification model. These errors represent the difference between what the model predicts and what the actual outcome turns out to be in the real world. Whenever a model turns out to be worth considering for application, the next step is to examine why classification errors occur in the test dataset. Sometimes, the predicted and actual value may differ in predicting a record to a certain class label. The classifier mostly predicts the records into a certain class as there are similar attributes that lie in the same class boundary.

The confusion matrix of the final model for the study is shown in table 5.15 above. Accordingly, it shows that out of 3927 instances 1313 instances are correctly classified as Success, 1287 instances are correctly classified as Average and 1242 instances are classified as Weak. This classifier incorrectly classified 7 instances as Average and 18 instances as Weak while in fact, they belong to Success; incorrectly classified 3 instances as Success, 16 instances as Weak while in fact, they belong to Average and incorrectly classified 5 instances as Success, 36 instances as Average while in fact, they belong to Weak. The reason for the misclassification of the three classes was if success status occurs there is also a possibility that average or weak status to have occurred.

5.5. Generated Rules from Decision Trees

In this study from the model developed in the above mentioned experiments, J48 classifier with default percentage split (66%) and resampling have achieved relatively the highest accuracy in most of performance evaluation criteria compared to PART, Naïve Bayes, Logistic Regression, Sequential Minimal Optimization and Multilayer Perception algorithms. Therefore, the model generated by J48 classifier with all attributes was selected as the model that can predict student performance.

J48 decision tree generated 495 rules for predicting student performance; the following 12 rules were found interesting rules extracted. Therefore, those rules selected based on the discussion with the domain expert and have the highest accuracy selected. The numeric values which appeared in the bracket next to the class label indicate the number of correctly and incorrectly classified records, respectively. Hence, rules generated by the model are interpreted as follows.

Rule 1: IF Previous study Field = No and Number of Common Course per semester = Zero and Marital status = Single and Number of Supportive Course per semester = Zero and Type of Institution Attended = Government and Gender = M and Year = Two: Success (129.0)

This rule indicates that if the student doesn't have previous study (doesn't have degree or diploma) and not taken common course and supportive course in the semester and marital status is single and student attended High School in Government school and gender is male and year of attended is two then the student performance is Success.

Rule 2: IF Previous study Field = No and Number of Common Course per semester = Zero and Marital status = Single and Number of Supportive Course per semester = Zero and Type of Institution Attended = Addis Ababa Private and Division = Extension and Total course per semester = Three: Average (116.0)

This rule indicates that if the student doesn't have previous study (doesn't have degree or diploma) and not taken common course and supportive course in the semester and marital status is single and student attended High School in Addis Ababa private school and division is Extension and total number of course per semester three given then the student performance is Average.

Rule 3: IF Previous study Field = No and Number of Common Course per semester = Zero and Marital status = Married and Number of Major Course per semester = Three and Financial source = Self Sponsored: Success (182.0)

This rule indicates that if the student doesn't have previous study (doesn't have degree or diploma) and not taken common course and marital status is married and three major course given per semester and financial source is self-sponsored then the student performance is Success.

Rule 4: IF Previous study Field = No and Number of Common Course per semester = One and Year = Two and EHEEE GPA = Good and Type of Institution Attended = Addis Ababa Government and Mathematics = Satisfactory and Age = Teenager: Success (130.0)

This rule indicates that if the student doesn't have previous study (doesn't have degree or diploma) and one common course taken per semester and year of attended is two and EHEEE GPA is good and student attended high school in Addis Ababa government and mathematics is satisfactory and age of the student is teenager then the student performance is Success.

Rule 5: IF Previous study Field = No and Number of Common Course per semester = One and Year = One and EHEEE Year taken = Seven Subject and Number of Supportive Course per semester = Two and Mathematics = Good and Financial source = Parent Sponsored and Type of Institution Attended = Addis Ababa Government and Gender = F and English = Very good: Average (192.0/45.0)

This rule indicates that if the student doesn't have previous study (doesn't have degree or diploma) and one common course given per semester and two supportive course in the semester given and EHEEE year taken seven subject and mathematics have good result and financial source is parent sponsored and high school attended in Addis Ababa Government School and gender is female and English result is very good then the student performance is Average.

Rule 6: IF Previous study Field = No and Number of Common Course per semester = Two and Number of Major Course per semester = Two and EHEEE GPA = Good and English = Good and Gender = F and Age = Age Two and Financial source = Parent Sponsored: Weak (126.0)

This rule indicates that if the student doesn't have previous study (doesn't have degree or diploma) and two common course given per semester and two major course in the semester given and EHEEE GPA result good and English have good result and gender is female age is age two and financial source is parent sponsored then the student performance is Weak.

Rule 7: IF Previous study Field = No and Number of Common Course per semester = Two and Number of Major Course per semester = Two and EHEEE GPA = Good and English = Good and Gender = F and Age = Age One and Financial source = Parent Sponsored: Average (185.0/43.0)

This rule indicates that if the student doesn't have previous study (doesn't have degree or diploma) and two common course given per semester and two major course in the semester given and EHEEE GPA result good and English have good result and gender is female age is age two and financial source is parent sponsored then the student performance is Average.

Rule 8: IF Previous study Field = No and Number of Common Course per semester = Three and EHEEE GPA = Good: Weak (572.0)

This rule indicates that if the student doesn't have previous study (doesn't have degree or diploma) and three common course given per semester and EHEEE GPA is good then the student performance is Weak.

Rule 9: IF Previous study Field = Teaching and Previous study GPA = Satisfactory: Success (152.0)

This rule indicates that if the students have Teaching in previous study field and previous study GPA is Satisfactory the Success.

Rule 10: IF Previous study Field = Health: Success (85.0)

This rule indicates that if the students have Health in previous study field then the student performance is Success.

Rule 11: IF Previous study Field = IT and Number of Supportive Course per semester = Zero and EHEEE Year taken = Old Twelve: Average (97.0)

This rule indicates that if the students have IT in previous study field and not given supportive course per semester and year EHEEE taken old twelve then the student performance is Average.

Rule 12: IF Previous study Field = Engineering: Average (18.0/1.0)

This rule indicates that if the students have Engineering in previous study field then the student performance is Average.

5.6. Discussion on Major Findings

From the generated rules it is observed that the most determinant factors are previous study field, number of common course per semester, total course per semester, year, financial source, number of supportive course per semester, previous study GPA, previous study institution attended, previous study program and number of major course per semester.

The first factor identified in this study is Previous study Field; Students come to the classroom with a broad range of prior knowledge experience, skills, beliefs, and attitudes, which influence how they attend, interpret and organize incoming information.

Other factors identified in this study are total course per semester, number of major course per semester, number of common course per semester, and number of supportive course per semester; less academic load during in the semester give result in greater student success. Academic load is measured in terms of credit load and course difficulty. The possibilities that weaker students might be more successful with lighter credit loads or those stronger students might be more successful with more difficult courses.

Attribute Year is also identified as the major factor that determine student performance. Those who join universities leave their homes and their families for the first time and prepare to face new experiences. Therefore, they might miss their families and friends. They might also face difficulties to manage their lives and to familiarize themselves with the new atmosphere. Student experience may improve over time.

The study also identified financial sources as factors that affect student performance. The self-sponsored students are more satisfied than those that get their money from their parents. Those who invest their own money should have a higher level of motivation for attending. Parents are not financially stable; students worried about money, this financial worry may affect their academic performance.

The findings observed in the interpreted rules shows that attributes like the use of other than marital status, religion, employment and gender have less effect on student performance.

The related research works concerning the issue of student performance. The study conducted by [9] the research findings indicated that EHEECE (Ethiopian Higher Education Entrance Certificate Examination) result, Sex, Number of students in a class, number of courses given in a semester, and field of study are the major factors affecting the student performances.

The other research work was done by [44], also showed 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 previous study (previous GPA, attended University, study program) number of courses given in per semester (Major courses, common courses and supportive courses given per semester) financial sources and Academic year become a new finding that the other researcher didn't consider it.

5.7. Use of Knowledge

After evaluating the discovered knowledge, the last step is the use of this knowledge for determining student performance. In this step, the knowledge discovered is incorporated into performance system and take this action based on the discovered knowledge.

The development of a graphical user interface in this study was done using Java. This graphical user interface was developed based on the model generated by J48 decision tree classifier with pruned parameter and with all attributes. The rules used by the studies to design the graphical user interface for predicting the student's performance are used. This prototype prediction model can be used for predicting student's performance based on the rules generated by J48 classifier. A sample result is shown in Figure 5.4 shown in below.

St Mary's University
Students' Performance Prediction Model

Division	Regular	Total course per semester	Seven	English	Good
Gender	Female	Number of Major Course per semester	Three	Mathematics	Excellent
Age	Teenager	Number of Supportive Course per semester	Two	Region	Addis Ababa
Year	Two	Number of Common Course per semester	Two	Previous study Field	Health
Marital status	Single	Type of high school	Addis Ababa Government	Previous study Program	Degree
Employment	Yes	Academic Year	Seven Subject	Previous study GPA	Satisfactory
Financial source	Parent Sponsored	EHEEE GPA	Excellent	Previous University	Government University

Student Status
Success

Solome Samson Altaye

Figure 5.4: Sample result of the prototype

5.7.1. Users Evaluation Result

The researcher developed a user interface prototype to predict higher education students' performance, based on the objective to check the validity of the interface. Six domain experts have been asked to give feedback on five point scale ranging from excellent (5) to poor (1) and the researcher discusses with each of the experts about the interface for prediction of higher institution students' performance. Table 5.16 below presents a summary of user acceptance testing results.

Questionnaires	Excellent (5)	Very Good (4)	Good (3)	Satisfactory (2)	Poor (1)
Effectiveness:					
➤ Is the prediction completeness?	50%	25%	25%	-	-
➤ Is the prediction produces a desired result?	50%	25%	25%	-	-
➤ Are you satisfied with the prediction result?	75%	25%	-	-	-
Efficiency:					
➤ Are the prediction taken less times?	75%	-	-	25%	-
➤ Is the prediction saves energy & materials?	75%	25%	-	-	-
Engaging:					
➤ Is the prediction user interface likeable?	75%	25%	-	-	-
Easy to Learn:					
➤ Is the prediction easy to learn?	25%	50%	25%	-	-
➤ Is the prediction system User friendly?	75%	25%	-	-	-

Table 5.16: Summary of users' response on the prototype

As shown in Table 5.16 above; most of the respondents have positive feedback towards the validity of the prototype. In the case of effectiveness, they revealed that this prediction model produces the desired result, but in order to make it perfect other factors could better be included for improvement is suggested by domain expert.

In this research, efficiency is considered as a time taken to predict student performance by taking inputs from the user. Majority of domain experts that account for 75% responded the prototype is much efficient. 25% respondents give satisfactory because they stated that student performance analysis must be perfect because it has a significant impact on the students that might need special attention to get better performance.

As discussed before this experiment registers a better performance of 97.84% accuracy with J48 decision tree algorithm. Making a user interface attractive and easy to use domain experts satisfies with the interface design and we prepare sample screenshots on the document.

During the discussion with expert, they reveal that our prediction model was easy to learn, and it is easier to remember. They also agreed that the prediction model was user-friendly and it is clearer. The newly developed system is a simple and manageable one. There is no need for advanced computer training for managing the user interface.

In this research, we introduced new trends for St. Mary's University to use the historical institutional students' records for determining students' status by applying data mining technique. Respondents suggested that predicting student status is a new technique that the university didn't use before. Therefore St. Mary's University can use this system to effectively identify students' status. To improve the prototype they suggested that, if other factors and another feature also be included it will be better.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1. Conclusion

In higher education institutions learning curriculum or process doesn't have a special concern to renovate student performance by showing students back history. The main focus of teaching-learning is just given learning to acquire knowledge. This type of process has done for a long period of time and not gives the necessary information in resolving students learning problem.

Educational data mining (EDM) is an emerging field for high quality research that mines large data sets in order to answer educational research questions that shed light on the learning process. EDM has been concerned with developing methods for exploring unique and increasingly large scale data that come from educational settings and using those methods to better understand students and the settings which they learn in. The main objective of this research is to identify important and interesting patterns from the academic history of the students' records that can enable students' performance in higher education.

In this research, the methodology employed was Hybrid Data mining process model; it involves six steps and the researcher thoroughly passes through all the steps and iterated as needed. The study was conducted using WEKA software version 3.9.2 and six data mining algorithms for classification techniques. A total of 11550 datasets, 21 attributes and 1 outcome variable were used to build the model. Several experiments were conducted in order to build models that can predict student's performance in higher education.

Different experiments are conducted using J48 decision tree algorithm, PART rule induction, Naïve Bayes, Logistic Regression, Sequential Minimal Optimization and Multilayer Perception algorithm using 10- fold cross validation and percentage split. The experimental result shows that J48 decision tree algorithm outperforms with an accuracy of 97.84%.

Based on, the extracted hidden pattern using J48 algorithm, previous study field, number of common course per semester, total course per semester, year, financial source, number of supportive course per semester, previous study GPA, previous study institution attended, previous study program and number of major course per semester are identified as the major finding factors of student status.

The strength of this study is an achievement of all the stated goals, Data mining goal was to identify major attributes that contributes to students' performance the study has identified and selected 10 attributes that are significant in predicting student performance. To design a predictive model Data mining techniques are more appropriate to predict students' performance. The selected model built with J48 classier was able to answer this question by predicting 97.84% of the cases correctly and develop a prototype of the student performance prediction interface. Understanding of the problem some attributes not considered due to data completeness and missing values found in the dataset are the challenge of this study.

6.2. Recommendations

This research work is conducted mainly for academic achievement. However, the researcher strongly believes that the findings of the study can be used by the concerned organizations to further investigate. Based on the findings obtained from the research, the researcher makes the following recommendation.

- ❖ In this study work, using some set of attributes that were considered more important by domain experts. However, understanding of the problem shows that there are attributes that are missing from the database. Recording important variables might help for decision making variables such as assignment mark, quizzes, lab work, class test, class attendance, partners' occupation, partners' education level, family size, disability, instructors performance, social interaction network, psychometric factor, study habits. Hence, higher institutions should have to make their database and data warehouse with better quality and stronger for educational data mining.

- ❖ In this research, the researcher use only St. Mary's University data; however further investigation is needed by including the other higher institution data.
- ❖ This study has attempted to apply DM techniques on student's data but it could also be applied in other education areas for decision making and problem solving concerning the quality of education, student placement etc.
- ❖ Education planners together with other interested parties should use the proposed potential set of attributes to design good and suitable plans to solve student academic weakness.
- ❖ In order to design an intelligent system integration of the discovered classification rules with knowledge-based system is need for the future.

REFERENCE

- [1] Mohammed I.Al-Twijri a, Amin Y. Noamanb, "A New Data Mining Model Adopted for Higher Institutions," in *ference on Communication, Management and Information Technology (ICCMIT 2015)*, Jeddah, Saudi Arabia, 2233.
- [2] Surjeet Kumar Yadav , Brijesh Bharadwaj and Saurabh Pal, "Data Mining Applications: A comparative Study for Predicting Student's performance," *International journal of innovative technology & creative engineering (ISSN:2045-711)*, vol. 1, no. 12, pp. 13-19, December 2013.
- [3] Jiawei Han and Micheline Kamber, *Data mining: Concepts and techniques*, Burnaby: Morgan Kaufmann , 2000.
- [4] Patil Sameer G. and Barahate Sachin, "Educational Data Mining –A New Approach to the Education Systems," *International Journal On Advanced Computer Theory And Engineering (IJACTE)*, vol. 5, no. 1, pp. 18-20, 2016.
- [5] Rajni Jindal and Malaya Dutta Borah, "A survey on educational data mining and research trends," *International Journal of Database Management Systems (IJDMS)*, vol. 5, no. 3, pp. 53-73, June 2013.
- [6] D. Kabakchieva, "Student Performance Prediction by Using Data Mining Classification Algorithms," *International Journal of Computer Science and Management Research*, vol. 1, no. 4, pp. 668-690, 4 November 2012.
- [7] R. Sumitha, E.S. Vinothkumar, "Prediction of Students Outcome Using Data Mining Techniques," *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, vol. 2, no. 6, pp. 132-139, June 2016.
- [8] Paul Golding and Opal Donaldson, "Predicting Academic Performance," in *36th ASEE/IEEE Frontiers in Education Conference*, Jamaica, 2006.
- [9] M. A. Yehuala, "Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University)," *International journal of scientific & technology research*, vol. 4, no. 04, pp. 91-94, APRIL 2015.
- [10] Jiawei Han and Micheline Kamber, *Data Mining:Concepts and Techniques Second Edition*, San Francisco: Morgan Kaufmann Elsevier's Science, 2006.
- [11] Soumen Chakrabarti et al., *Data Mining Know It All*, United States: Morgan Kaufmann is an imprint of Elsevier, 2009.
- [12] Xue Z. Wang, *Data Mining and Knowledge Discovery for Process Monitoring and Control*,

London Berlin Heidelberg: Springer-Verlag , 1999.

- [13] Pang.Ni Ng Tan, Michael Steinbach, Vi Pi N Ku Mar, Introduction to Data mining, united States of America: Pearson Education, Inc, 2006.
- [14] M. Kantardzic, Data mining : concepts, models, methods, and algorithms, Second edition, Hoboken, New Jersey: John Wiley & Sons, Inc., 2011.
- [15] Cristóbal Romero and Sebastián Ventura, "Data mining in education," *Wiley Interdisc. Rev.: Data Min. Knowl.*, vol. 3, no. 1, pp. 12-27, 2013.
- [16] Cristóbal Romero and Sebastián Ventura, "Educational data mining: a review of the state of the art," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev*, vol. 40, no. 6, pp. 601-618, 2010.
- [17] A. Peña-Ayala, Educational Data Mining Applications and Trends, Switzerland: Springer Cham Heidelberg New York Dordrecht London, 2014.
- [18] W. He, "Examining students' online interaction in a live video streaming environment using data mining and text mining," *Comput. Hum. Behav*, vol. 29, no. 1, pp. 90-102, 2013.
- [19] T. C. Corporation, "Introduction to Data Mining and Knowledge Discovery 3rd Ed," USA, Two Crows Corporation, 2005.
- [20] C. C. Aggarwal, Data Mining: The Textbook, Switzerland: Springer International Publishing, 2015.
- [21] Prof. Oded Maimon and Dr. Lior Rokach, Data Mining and Knowledge Discovery Handbook , Second Edition, New York Dordrecht Heidelberg London: Springer Science+Business Media, LLC 2005., 2010.
- [22] Mr. M. N. Quadri and Dr. N.V.Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques," *Global Journal of Computer Science and Technology*, vol. 10, no. 2, pp. 2-5, April 2010.
- [23] Lior Rokach and Oded Maimon, DATA MINING WITH DECISION TREES Theory and Applications 2nd Edition, Singapore: World Scientific Publishing Co. Pte. Ltd, 2014.
- [24] Mrs.N.Sivanagamani, "Review on Data Mining Techniques," *International Journal of Computational Engineering Research (IJCER)*, vol. 8, no. 2, p. 2250 – 3005, February – 2018.
- [25] K. R. a. D. A. Payal Dhakate, "Analysis of Different Classifiers for Medical Dataset using Various Measures," *International Journal of Computer Applications* , vol. 111, no. 5, pp. 20-24, February 2015.

- [26] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz A. Kurgan, *Data Mining: A Knowledge Discovery Approach*, 2007: Springer Science+Business Media, USA.
- [27] Asha Gowda Karegowda, Punya V, M.A.Jayaram and A.S .Manjunath, "Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5," *International Journal of Computer Applications (0975 – 8887)*, vol. 45, no. 12, pp. 45-50, May 2012.
- [28] J Han and M Kamber, *Data mining: concepts and techniques*, 2nd ed, San Fransisco: CA: Morgan Kaufmann , 2006.
- [29] S Krishnaveni and M Hemalatha, "A perspective analysis of traffic accident using data mining techniques," *International Journal of Computer Applications*, vol. 23, no. 7, 2011.
- [30] H. Witten, "Data Mining: Practical machine learning tools and techniques.," 2005.
- [31] Marijana Zekić-Sušac, Nataša Šarlija, Adela Has and Ana Bilandžić, "Predicting company growth using logistic regression and neural networks," *Croatian Operational Research Review*, no. 7, pp. 229-248, 2016.
- [32] M. Maalouf, "Logistic regression in data analysis: An overview," *International Journal of Data Analysis Techniques and Strategies*, pp. 1-20, July 2011.
- [33] M. Kantardzic, *DATA MINING Concepts, Models, Methods, and Algorithms*, Canada: A JOHN WILEY & SONS, INC, 2011.
- [34] Kalpana Rangra and Dr. K. L. Bansal, "Comparative Study of Data Mining Tools," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 6, pp. 216-223, June 2014.
- [35] Kalpana Rangra and Dr. K. L. Bansal, "Comparative Study of Data Mining Tools," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 6, pp. 216-223, June 2014.
- [36] Zdravko Markov and Ingrid Russell, "An introduction to the WEKA data mining system," in *11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education*, Bologna, Italy, June 2006.
- [37] Brijesh Kumar Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Students' Performance," (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 6, pp. 63-69, 2011.
- [38] Brijesh Kumar Bhardwaj and Saurabh Pal, "Data Mining: A prediction for performance improvement using classification," (*IJCSIS*) *International Journal of Computer Science and Information Security*, vol. 9, no. 4, 2011.

- [39] Amirah Mohamed Shahiria, Wahidah Husaina and Nur'aini Abdul Rashid, "A Review on Predicting Student's Performance using Data Mining Techniques," in *The Third Information Systems International Conference*, Penang, Malaysia, 2015.
- [40] J.F. Superby and J.-P. Vandamme N. Meskens, "Determination of factors influencing the achievement of the first-year university students using data mining methods".
- [41] A. A. Saa, "Educational Data Mining and Students' Performance Prediction," (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, pp. 212-20, 2016.
- [42] Oyelade, O. J , Oladipupo, O. O and Obagbuwa, I. C, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance," (*IJCSIS*) *International Journal of Computer Science and Information Security*, vol. 7, no. 1, pp. 292-295, 2010.
- [43] Fiseha Berhanu and Addisalem Abera, "Students' Performance Prediction based on their Academic Record," *International Journal of Computer Applications (0975 – 8887)*, vol. 131, no. 15, pp. 27-35, December 2015.
- [44] Alemu K. Tegegne and Tamir A. Alemu, "Educational Data Mining for Students' Academic Performance Analysis in Selected Ethiopian Universities," *Information Impact: Journal of Information and Knowledge Management*, vol. 9, no. 2, pp. 1-15, 2018, Vol. 9 (2) Pg 1 - 15.
- [45] T. Teklu, "Identifying Determinant Factors for Students' Success in Preparatory Schools Using Data Mining Techniques," AAU respository, Addis Ababa.
- [46] M. Macin, Center for innovation in research and teaching , [Online]. Available: https://cirt.gcu.edu/research/developmentresources/research_ready/experimental/overview. [Accessed 3 April 2019].
- [47] Mark F. Hornick , Erik Marcadé and Sunil Venkayala, *Java Data Mining: Strategy, standard, and practice : a practical guide for architecture, design, and implementation*, United States of America: Morgan Kaufmann Publishers is an imprint of Elsevier, 2007.
- [48] Pressman, R., "Software Engineering: A Practitioner's Approach. McGraw-Hill," New York, 2005.
- [49] Julio Ponce and Adem Karahoca, *Data Mining and Knowledge Discovery in Real Life Applications*, Vienna, Austria: In-Teh is Croatian branch of I-Tech Education and Publishing KG, 2009.
- [50] Piatetsky-Shapiro, G. and Frawley, W., "Knowledge Discovery in Databases," in *AAAI/ MI Press, MA*, 1991.
- [51] Fayyad, "The KDD Process for Extracting Useful Knowledge," in *Communications of the ACM*, New York, USA., 1996.

- [52] P. Ch a p m a n, " CRISP-DM 1.0: Step-by-Step Data Mining Guide 2000," in *SPSS Inc, CRISPWP-0800*, 2000. http://www.spss.ch/upload/1107356429_CrispDM1.0.pdf, 2000.
- [53] Ana Azevedo and Manuel Filipe Santos, "KDD, SEMMA AND CRISP-DM: A parallel overview," in *ISBN: 978-972-8924-63-8 © 2008 IADIS*, Portugal, 2008.
- [54] B. Kiranmai and Dr. A. Damodaram, "A Review on Evaluation Measures for Data Mining Tasks," *International Journal Of Engineering And Computer Science ISSN:2319-7242*, pp. 7217-7220, July, 2014.
- [55] Weiss, Sholom M. and Zhang, Tong, *The Hand book of data mining. Performance analysis and evaluation*, New Jersey, USA: Lawrence Erlbaum Associates Inc, 2003.
- [56] Daniel T. Larose and Chantal D.Larose, *Data mining and predictive analytics second edition*, Hoboken, New Jersey: John Wiley & Sons, Inc, 2015.
- [57] L. Luo, "Software Testing Techniques Technology Maturation and Research Strategy," 2012.
- [58] N. H. a. A. El-Shaarawi, "Factors Affecting Students' Performance," 16 July 2006.
- [59] F. A, "The Role of Affirmative Action on Empowering Women's, in the Case of L/HaHale," *Journal of Civil & Legal Sciences*, vol. 6, no. 1, pp. 1-7, 2017.
- [60] St. Mary University , graduate student hand book, Addis Abeba: SMU printing press, 2014/2015, p. 2nded.
- [61] Z. Ayalew, "overview of St. Mary university," in *St.Mary university school of graduate*, Addis Ababa, 2018.
- [62] Adamu, Abebaw, Y.; Addamu, A.M, "Quality assurance in Ethiopian higher education: Procedures and," *Pricedia-Social and behavioral sciences.*, vol. 6, no. 4, pp. 838-864, December, 2012.
- [63] St. Mary's University, Student handbook, Addis Ababa, Ethiopia: St. Mary's University Press, October,2017.
- [64] "Weka Machine Learning Project," [Online]. Available: <http://www.cs.waikato.ac.nz/~ml/index.html>.
- [65] Olatz Arbelaitz and Ibai Gurrutxaga and Javier Muguerza and Jesús María Pérez, "Applying Resampling Methods for Imbalanced Datasets," *Springer-Verlag Berlin Heidelberg 2013*, pp. 111 - 120, 2013.

ANNEXES

Annex-1: The original collected sample data

The screenshot shows a Microsoft Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Division	Gender	Age	Year	Total cour	Number o	Number o	Number o	Marital st	Employment	Financial sc	Region	High Scho	English	Mathema	Year EHEE	EHEEE GP	Previous s	Previous s	Previous s	Previous s	Status
2	Extensior	M	Age Two	Two	Two	Two	Zero	Zero	Single	Yes	Organizati	Addis Ab	Addis Ab	Good	Good	Five Subj	Good	No	No	No	No	Success
3	Regular	M	Age One	Two	Six	Three	Two	One	Single	No	Parent Spc	Addis Ab	Addis Ab	Good	Good	Seven Su	Good	No	No	No	No	Success
4	Regular	F	Age One	Two	Six	Three	Two	One	Single	No	Parent Spc	Addis Ab	Addis Ab	Very good	Satisfact	Seven Su	Good	No	No	No	No	Success
5	Extensior	M	Age Two	One	Five	Two	One	Two	Single	Yes	Self Spons	Addis Ab	Government	Good	Good	Five Subj	Satisfact	No	No	No	No	Success
6	Extensior	M	Age Two	Three	Four	Three	One	Zero	Married	Yes	Self Spons	SNNP	Addis Ab	Very good	Very good	Five Subj	Good	No	No	No	No	Success
7	Extensior	F	Age Threi	Three	Two	Two	Zero	Zero	Single	Yes	Organizati	Addis Ab	Addis Ab	Good	Good	Old Twel	Old Twel	Private U	Business	Diploma	Very good	Success
8	Regular	F	Age One	One	Six	Three	Two	One	Single	No	Parent Spc	Addis Ab	Addis Ab	Very good	Satisfact	Five Subj	Good	No	No	No	No	Success
9	Extensior	F	Age Two	Three	Four	Three	One	Zero	Single	Yes	Self Spons	Amhara	Addis Ab	Very good	Very good	Five Subj	Good	No	No	No	No	Success
10	Extensior	M	Age Two	One	Two	Two	Zero	Zero	Single	Yes	Organizati	Addis Ab	Government	Very good	Very good	Old Twel	Old Twel	No	No	No	No	Success
11	Regular	F	Age Two	Two	Six	Three	Two	One	Single	No	Parent Spc	Addis Ab	Addis Ab	Good	Excellent	Seven Su	Good	No	No	No	No	Success
12	Extensior	M	Age Two	Two	Four	Three	Zero	One	Single	Yes	Self Spons	Addis Ab	Addis Ab	Good	Good	Old Twel	Old Twel	Government	Business	Diploma	Very good	Success
13	Extensior	M	Age Two	Three	Four	Three	One	Zero	Married	Yes	Self Spons	SNNP	Addis Ab	Very good	Very good	Five Subj	Good	No	No	No	No	Success
14	Extensior	F	Age Two	One	Four	Two	One	One	Single	Yes	Parent Spc	Addis Ab	Addis Ab	Good	Good	Seven Su	Good	No	No	No	No	Success
15	Regular	F	Age One	Three	Seven	Six	One	Zero	Single	No	Parent Spc	Addis Ab	Addis Ab	Good	Good	Seven Su	Good	No	No	No	No	Success
16	Extensior	F	Age Two	Three	Two	Two	Zero	Zero	Single	Yes	Self Spons	Addis Ab	Addis Ab	Good	Good	Five Subj	Satisfact	Government	Business	Degree	Satisfact	Success
17	Extensior	M	Age Two	Two	Four	Three	Zero	One	Single	Yes	Self Spons	Amhara	Addis Ab	Good	Good	Five Subj	Good	No	No	No	No	Success
18	Regular	F	Age One	Two	Three	Three	Zero	Zero	Single	No	Self Spons	Amhara	Addis Ab	Very good	Good	Seven Su	Good	No	No	No	No	Success
19	Regular	M	Age Two	One	Six	Three	Two	One	Single	No	Self Spons	Addis Ab	Government	Very good	Very good	Seven Su	Good	No	No	No	No	Success
20	Extensior	M	Age Threi	Four	Five	Five	Zero	Zero	Single	Yes	Self Spons	Addis Ab	Addis Ab	Good	Good	Five Subj	Satisfact	Government	Business	Degree	Satisfact	Success
21	Regular	F	Age One	Two	Six	Three	Two	One	Single	No	Parent Spc	Addis Ab	Addis Ab	Very good	Satisfact	Seven Su	Good	No	No	No	No	Success
22	Extensior	F	Age Threi	Four	Five	Five	Zero	Zero	Single	Yes	Parent Spc	Addis Ab	Addis Ab	Excellent	Good	Seven Su	Good	No	No	No	No	Success
23	Regular	M	Age Two	Four	Four	Four	Zero	Zero	Single	No	Parent Spc	Addis Ab	Addis Ab	Excellent	Very good	Seven Su	Excellent	No	No	No	No	Success
24	Regular	F	Age One	Two	Six	Three	Two	One	Single	No	Parent Spc	Addis Ab	Addis Ab	Very good	Satisfact	Seven Su	Good	No	No	No	No	Success
25	Regular	F	Age Two	Two	Three	Three	Zero	Zero	Married	No	Self Spons	Addis Ab	Addis Ab	Very good	Good	Seven Su	Good	No	No	No	No	Success
26	Extensior	F	Age One	One	Four	Two	One	One	Single	Yes	Self Spons	Addis Ab	Addis Ab	Good	Good	Five Subj	Grade Te	Government	Business	Diploma	Very good	Success
27	Extensior	M	Age Two	Two	Two	Two	Zero	Zero	Single	Yes	Self Spons	Addis Ab	Addis Ab	Excellent	Good	Five Subj	Good	No	No	No	No	Success
28	Extensior	F	Age One	Three	Three	Three	Zero	Zero	Single	Yes	Self Spons	Amhara	Addis Ab	Good	Good	Five Subj	Good	No	No	No	No	Success
29	Regular	F	Age Two	One	Six	Three	Two	One	Single	No	Parent Spc	Addis Ab	Addis Ab	Good	Good	Five Subj	Good	No	No	No	No	Success
30	Regular	M	Age Two	One	Five	Three	One	One	Single	No	Parent Spc	Addis Ab	Addis Ab	Excellent	Very good	Five Subj	Good	No	No	No	No	Success

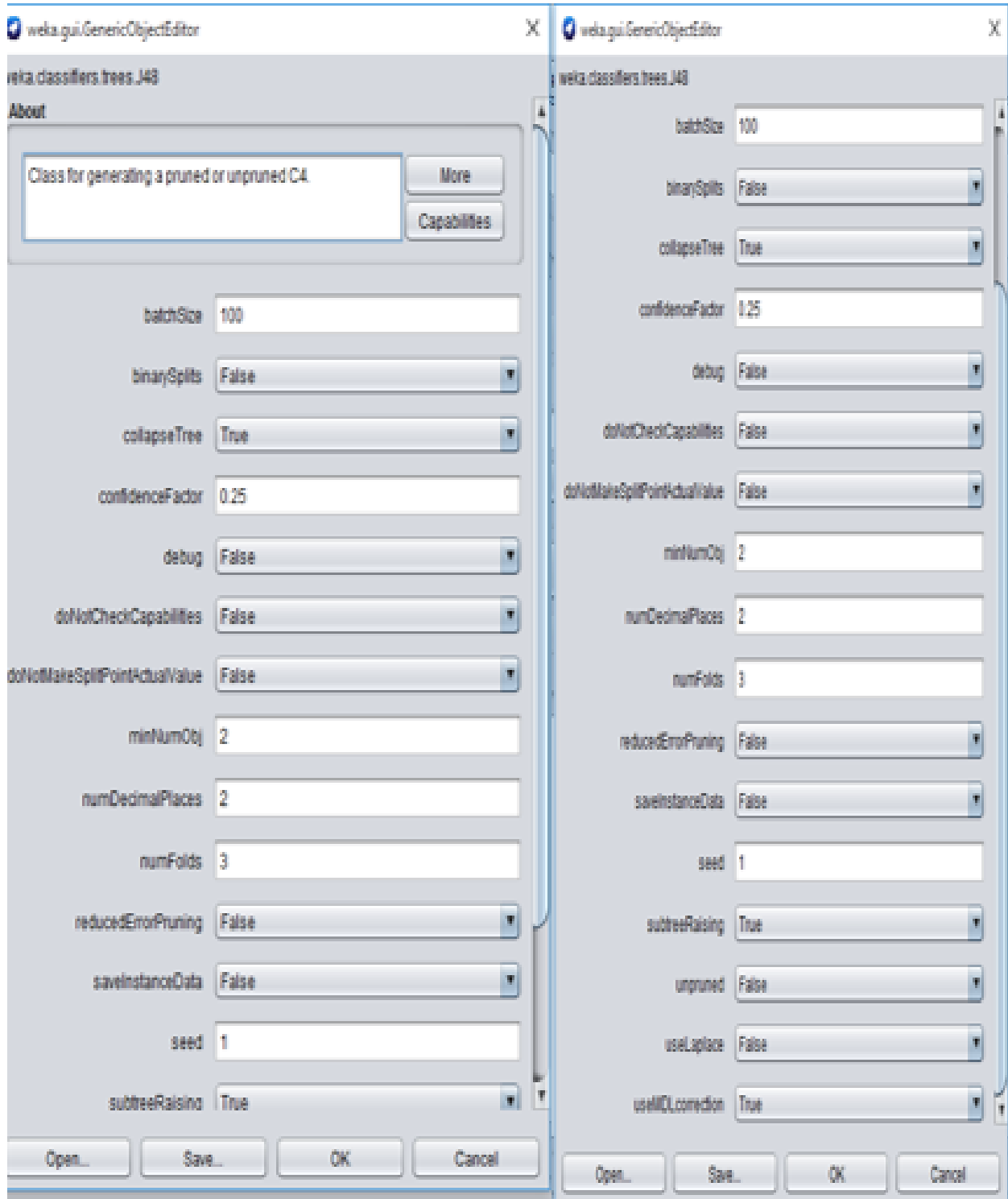
Annex-2: Statistical summary of the selected attributes

No.	Variables	Instances	Frequency	Percent (%)
1	Division	Extension	5526	47.84
		Regular	6024	52.16
2	Gender	M	4409	38.17
		F	7141	61.83
3	Age	Age One	3821	33.08
		Age Two	6202	53.70
		Age Three	1377	11.92
		Age Four	150	1.30
4	Year	One	4405	38.14
		Two	3725	32.25
		Three	2474	21.42
		Four	946	8.19
5	Total course per semester	Two	1185	10.26
		Three	1336	11.57
		Four	3009	26.05
		Five	1456	12.61
		Six	3973	34.40
		Seven	591	5.12
6	Number of Major Course per semester	Two	4027	34.87
		Three	5008	43.36
		Four	797	6.90
		Five	1062	9.19
		Six	656	5.68
7	Number of Supportive Course per semester	Zero	4563	39.51
		One	3992	34.56
		Two	2924	25.32
		Three	71	0.61
8	Number of Common Course per semester	Zero	6135	53.12
		One	3525	30.52
		Two	1164	10.08
		Three	726	6.29
9	Marital status	Single	10625	91.99

		Married	812	7.03
		Divorced	113	0.98
10	Employment	Yes	5453	47.21
		No	6097	52.79
11	Financial source	Organization Sponsor	409	3.54
		Parent Sponsored	6577	56.94
		Self-Sponsored	4557	39.45
		Scholarship	7	0.06
12	Region	Addis Ababa	9800	84.85
		SNNP	318	2.75
		Amhara	568	4.92
		Tigray	164	1.42
		Oromia	601	5.20
		Harari	28	0.24
		Dire Dawa	17	0.15
		Somali	43	0.37
		Afar	2	0.02
		Gambella	9	0.08
13	High School Attended	Addis Ababa Government	7545	65.32
		Government	1714	14.84
		Addis Ababa Private	2291	19.84
14	English	Excellent	1824	15.79
		Very good	3509	30.38
		Good	6078	52.62
		Satisfactory	139	1.20
15	Mathematics	Excellent	586	5.07
		Very good	1831	15.85
		Good	6906	59.79
		Satisfactory	2203	19.07
		Fail	24	0.21
16	Year EHEEE taken	Five Subject	4061	35.16
		Seven Subject	5765	49.91
		Old Twelve	1724	14.93
17	EHEEE GPA	Excellent	188	1.63

		Very good	403	3.49
		Good	7390	63.98
		Satisfactory	1343	11.63
		Old Twelve	1724	14.93
		Grade Ten	502	4.35
18	Previous study institution attended	No	9720	84.16
		Private University	944	8.17
		Government TVET College	316	2.74
		Government University	570	4.94
19	Previous study Field	No	9689	83.89
		Business	1411	12.22
		Teaching	190	1.65
		Health	85	0.74
		IT	157	1.36
		Engineering	18	0.16
20	Previous study Program	No	9681	83.82
		Diploma	1183	10.24
		Degree	686	5.94
21	Previous study GPA	No	9697	83.96
		Excellent	31	0.27
		Very good	717	6.21
		Good	447	3.87
		Satisfactory	658	5.70
22	Status	Success	3850	33.33
		Average	3850	33.33
		Weak	3850	33.33

Annex-3: Parameter settings of the J48 used in conducting the experiments



Annex-4: The snapshot running information of J48 algorithm using resampling technique with 10-fold validation technique

The screenshot shows the Weka Explorer interface with the J48 classifier selected. The 'Test options' section is configured for cross-validation with 10 folds. The 'Classifier output' section displays the results of the stratified cross-validation, including a summary of performance metrics and a detailed accuracy by class table.

Test options

- Use training set:
- Supplied test set: Set...
- Cross-validation: Folds: 10
- Percentage split: %: 66
- More options...:

Classifier output

```

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

Correctly Classified Instances      11293           97.7749 %
Incorrectly Classified Instances     257             2.2251 %
Kappa statistic                     0.9666
Mean absolute error                  0.0219
Root mean squared error              0.1069
Relative absolute error              4.9346 %
Root relative squared error          22.6749 %
Total Number of Instances           11550

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                0.977   0.003   0.994     0.977   0.986     0.979  0.998    0.997    Success
                0.992   0.020   0.961     0.992   0.976     0.965  0.998    0.994    Average
                0.963   0.010   0.979     0.963   0.971     0.957  0.997    0.994    Weak
Weighted Avg.   0.978   0.011   0.978     0.978   0.978     0.967  0.998    0.995

=== Confusion Matrix ===

  a   b   c  <-- classified as
3763  30  57 |  a = Success
  6 3821  23 |  b = Average
 16 125 3709 |  c = Weak
    
```

Result list (right-click for options)

06:46:45 - trees_J48

Status

OK

Annex-5: The snapshot running information of J48 algorithm using resampling technique with percentage split 66% technique

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose J48 -C 0.25 -M 2

Test options:

- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

(Nom) Status

Result list (right-click for options)

11:55:55 - trees.J48

Classifier output

Time taken to test model on test split: 0.01 seconds

=== Summary ===


Correctly Classified Instances	3842	97.8355 %
Incorrectly Classified Instances	85	2.1645 %
Kappa statistic	0.9675	
Mean absolute error	0.0224	
Root mean squared error	0.1073	
Relative absolute error	5.0469 %	
Root relative squared error	22.7554 %	
Total Number of Instances	3927	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.981	0.003	0.994	0.981	0.988	0.981	0.999	0.998	Success
	0.985	0.016	0.968	0.985	0.976	0.965	0.997	0.995	Average
	0.968	0.013	0.973	0.968	0.971	0.957	0.998	0.994	Weak
Weighted Avg.	0.978	0.011	0.978	0.978	0.978	0.968	0.998	0.996	

=== Confusion Matrix ===

a	b	c	<-- classified as
1313	7	18	a = Success
3	1287	16	b = Average
5	36	1242	c = Weak

Status: OK  x0

Annex-6: The snapshot running information of PART algorithm using resampling technique with 10-fold validation technique

The screenshot displays the Weka Explorer interface with the PART classifier selected. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' pane provides a detailed summary of the model's performance, including a stratified cross-validation summary and a detailed accuracy table by class.

Classifier: PART -M 2 -C 0.25 -Q 1

Test options:

- Use training set
- Supplied test set (Set...)
- Cross-validation Folds: 10
- Percentage split %: 66
- More options...

Classifier output:

Time taken to build model: 0.66 seconds

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

Correctly Classified Instances	11259	97.4805 %
Incorrectly Classified Instances	291	2.5195 %
Kappa statistic	0.9622	
Mean absolute error	0.0235	
Root mean squared error	0.1144	
Relative absolute error	5.2832 %	
Root relative squared error	24.2589 %	
Total Number of Instances	11550	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.977	0.006	0.987	0.977	0.982	0.973	0.998	0.995	Success
	0.987	0.020	0.961	0.987	0.974	0.961	0.997	0.995	Average
	0.960	0.011	0.977	0.960	0.969	0.953	0.997	0.994	Weak
Weighted Avg.	0.975	0.013	0.975	0.975	0.975	0.962	0.997	0.994	

=== Confusion Matrix ===

a	b	c	<-- classified as
3763	30	57	a = Success
21	3799	30	b = Average
29	124	3697	c = Weak

Status: OK

Annex-7: The snapshot running information of Naive Bayes algorithm with 10-fold validation technique

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose NaiveBayes

Test options

- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

(Nom) Status

Result list (right-click for options)

12:03:45 - bayes.NaiveBayes

Classifier output

```

=== Summary ===
Correctly Classified Instances      6225      53.8961 %
Incorrectly Classified Instances    5325      46.1039 %
Kappa statistic                    0.2922
Mean absolute error                 0.3397
Root mean squared error             0.4377
Relative absolute error             78.4182 %
Root relative squared error        94.0556 %
Total Number of Instances          11550

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
	0.541	0.343	0.548	0.541	0.545	0.199	0.664	0.615
	0.536	0.193	0.554	0.536	0.545	0.346	0.773	0.625
	0.538	0.180	0.508	0.538	0.523	0.352	0.771	0.533
Weighted Avg.	0.539	0.255	0.540	0.539	0.539	0.283	0.725	0.597

```

=== Confusion Matrix ===
 a  b  c  <-- classified as
2719 1209 1094 |  a = Average
1205 1913  451 |  b = Success
1034  332 1593 |  c = Weak

```

Status: OK x0

Annex-8: The snapshot running information of Logistic Regression technique using resampling technique with percentage split 66% technique

Classifier

Choose **Logistic -R 1.0E-8 -M 1 -num-decimal-places 4**

Test options

- Use training set
- Supplied test set
- Cross-validation Folds
- Percentage split %

(Nom) Status

Result list (right-click for options)

08:46:29 - functions.Logistic

Classifier output

Time taken to test model on test split: 0.17 seconds

=== Summary ===

Correctly Classified Instances	2539	64.655 %
Incorrectly Classified Instances	1388	35.345 %
Kappa statistic	0.4699	
Mean absolute error	0.2999	
Root mean squared error	0.3882	
Relative absolute error	67.4763 %	
Root relative squared error	82.3326 %	
Total Number of Instances	3927	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.706	0.152	0.706	0.706	0.706	0.553	0.860	0.774	Success
	0.549	0.175	0.610	0.549	0.578	0.385	0.806	0.680	Average
	0.684	0.202	0.621	0.684	0.651	0.471	0.843	0.758	Weak
Weighted Avg.	0.647	0.176	0.646	0.647	0.645	0.470	0.836	0.738	

=== Confusion Matrix ===

	a	b	c	<-- classified as
944	214	180		a = Success
234	717	355		b = Average
160	245	878		c = Weak

Status

OK

Activate Windows
Go to Settings to activate Windows.

Log

Annex-9: The snapshot running information of Sequential Minimal Optimization technique using resampling technique with percentage split 66% technique

Classifier

Choose **SMO** -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -

Test options

- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

(Nom) Status

Result list (right-click for options)

09:06:32 - functions.SMO

Classifier output

```

Correctly Classified Instances      2636      67.125 %
Incorrectly Classified Instances    1291      32.875 %
Kappa statistic                    0.5066
Mean absolute error                 0.3142
Root mean squared error             0.4051
Relative absolute error             70.6974 %
Root relative squared error         85.9314 %
Total Number of Instances          3927

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.758    0.183    0.681     0.758   0.717     0.561    0.821    0.642    Success
      0.577    0.159    0.644     0.577   0.609     0.431    0.767    0.550    Average
      0.677    0.151    0.685     0.677   0.681     0.527    0.796    0.593    Weak
Weighted Avg.   0.671    0.165    0.670     0.671   0.669     0.507    0.795    0.595

=== Confusion Matrix ===

      a   b   c  <-- classified as
1014  187  137 |  a = Success
 290   754  262 |  b = Average
 185   230  868 |  c = Weak
    
```

Status

OK

Annex-10: The snapshot running information of Multilayer Perceptron techniques using resampling technique with 10 fold cross validation technique

Weka Explorer

Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier: Choose **MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a**

Test options

Use training set
 Supplied test set (Set...)
 Cross-validation Folds **10**
 Percentage split % **66**
 More options...

(Nom) Status

Start Stop

Result list (right-click for options)

09:12:36 - functions.MultilayerPerceptron

Classifier output

```

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      11297      97.8095 %
Incorrectly Classified Instances     253        2.1905 %
Kappa statistic                    0.9671
Mean absolute error                 0.0196
Root mean squared error             0.1028
Relative absolute error              4.413 %
Root relative squared error         21.8172 %
Total Number of Instances          11550

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
          0.987   0.008   0.983     0.987   0.985     0.978  0.997   0.976   Success
          0.985   0.016   0.968     0.985   0.976     0.965  0.994   0.994   Average
          0.962   0.008   0.984     0.962   0.973     0.959  0.998   0.997   Weak
Weighted Avg.   0.978   0.011   0.978     0.978   0.978     0.967  0.996   0.989

=== Confusion Matrix ===

  a   b   c  <-- classified as
3801  21  28 |  a = Success
  23 3794  33 |  b = Average
  42  106 3702 |  c = Weak
  
```

Status: OK

Log