

**CPU BUSINESS AND INFORMATION TECHNOLOGY COLLEGE**

**DEPARTMENT OF SOFTWARE ENGINEERING**

**GRADUATE PROGRAM IN SOFTWARE ENGINEERING**

**A PREDICTIVE MODEL FOR IDENTIFYING DROP OUT STUDENTS:**

**IN THE CASE OF ST. MARY UNIVERSITY**

BY

HIWOT TESHOME

JUNE, 2021

ADDIS ABEBA, ETHIOPIA

**CPU BUSINESS AND INFORMATION TECHNOLOGY COLLEGE**

**DEPARTMENT OF SOFTWARE ENGINEERING**

**GRADUATE PROGRAM IN SOFTWARE ENGINEERING**

**A PREDICTIVE MODEL FOR IDENTIFYING DROP OUT STUDENTS**

**IN THE CASE OF ST. MARY UNIVERSITY**

BY

HIWOT TESHOME

JUNE 2021

ADDIS ABEBA, ETHIOPIA

**CPU BUSINESS AND INFORMATION TECHNOLOGY COLLEGE**

**DEPARTMENT OF SOFTWARE ENGINEERING**

**GRADUATE PROGRAM IN SOFTWARE ENGINEERING**

**A PREDICTIVE MODEL FOR IDENTIFYING DROP OUT STUDENTS:**

**IN THE CASE OF ST. MARY UNIVERSITY**

BY

HIWOT TESHOME

Name and signature of members of the examining board,

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Title | Signature | Date |
| Sileshi. Yalew | Doctor | \_\_\_\_\_\_\_\_\_\_\_\_ (Advisor) |  |
|  | Doctor | \_\_\_\_\_\_\_\_\_\_\_\_\_ ( Examiner) |  |
|  | Doctor | \_\_\_\_\_\_\_\_\_\_\_\_\_ ( Examiner) |  |

**Declaration**

I hereby declare that Application of Data Mining Techniques to develop a classification model that predicts the dropout of student at St. marry University Result is my own work. All the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

Hiwot Teshome

June, 2021

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Sileshi Yalew(PHD)

June, 2021

# 

**ACKNOWLEDGEMENTS**

First and foremost, My Deepest Thank goes to the Almighty GOD for the wisdom and courage that he has been conferred upon me in all journeys and during this thesis work. And I would like to thank Virgin Mary, mother of Jesus Christ for her Intercession all the time.

I am deeply indebted and very glad to express my sincere gratitude and appreciation to my Advisor Dr.Sileshi Yalew for his important and constructive comments, encouragement, criticisms and professional advice

I am heartily thankful to my instructor who gave me the course datamining and research Dr. Tibebe Beshah for his encouragement, Commitment, support and for help that enabled me to develop an understanding of the subject.

I am deeply grateful to HERQA for collecting and making the data available for this study. In addition, my great thank goes to the Ato Shemelis St, marry University manager and all Registrar office employer for your real cooperation and help for collecting data via Email, Phone and during our meeting.

My special gratitude is to my mother, who always encourage me to be educated woman, My beloved Husband and children who inspire me and remind me that nothing is impossible, and I can do anything, thanks a lot.

Friends Sisters, and Brothers and to all of those who inspired, encouraged, and fully supported me in any respect during the completion of the thesis.

Lastly, I thank my classmates and friends for the stimulating discussions and debates, Support and for all the fun we have had in the last three years.

Table of Contents

[List of Figures 10](#_Toc77966599)

[List of Tables 11](#_Toc77966600)

[List of Acronyms and Abbreviations 12](#_Toc77966601)

[Abstract 13](#_Toc77966602)

[CHAPTER ONE 14](#_Toc77966603)

[1.1 INTRODUCTION 14](#_Toc77966604)

[1.2. PROBLEM STATEMENT 15](#_Toc77966605)

[1.3. SIGNIFICANCE/IMPLICATIONS 16](#_Toc77966606)

[1.4. OBJECTIVES 17](#_Toc77966607)

[General objective 17](#_Toc77966608)

[Specific objective 17](#_Toc77966609)

[1.5. SCOPE AND LIMITATION OF THE RESEARCH 17](#_Toc77966610)

[1.6. ORGANIZATION OF THE THESIS 17](#_Toc77966611)

[CHAPTER TWO 19](#_Toc77966612)

[LITERATURE REVIEW AND RELATED WORK 19](#_Toc77966613)

[2.1 LITERATURE REVIEW 19](#_Toc77966614)

[2.1.1 OVERVIEW OF DATA MINING 19](#_Toc77966615)

[2.1.2 EDUCATIONAL DATA MINING (EDM*)* 19](#_Toc77966616)

[2.1.3. DATA MINING MODELS 20](#_Toc77966617)

[2.1.3.1. Knowledge discovery Process 20](#_Toc77966618)

[2.1.3.2 CRISP-DM process model 22](#_Toc77966619)

[2.1.4. DATA MINING ALGORITHMS AND TECHNIQUES 26](#_Toc77966620)

[2.2 RELATED WORKS 29](#_Toc77966621)

[CHAPTER THREE 33](#_Toc77966622)

[3.1 METHODOLOGY 33](#_Toc77966623)

[3.1.1 RESEARCH DESIGN 33](#_Toc77966624)

[3.1.2. DATA COLLECTION 33](#_Toc77966625)

[3.1.3. DATA PRE-PROCESSING 35](#_Toc77966626)

[*3.1.3.1. Data cleaning* 35](#_Toc77966627)

[*3.1.3.2. Handling the incorrect or inconsistent values*. 35](#_Toc77966628)

[3.1.4. DATA TRANSFORMATION 37](#_Toc77966629)

[3.1.5 DATA FORMAT 39](#_Toc77966630)

[CHAPTER FOUR 41](#_Toc77966631)

[4.1 EXPERIMENTATIONS AND ANALYSIS 41](#_Toc77966632)

[4.1.1 BUILDING MODELING 41](#_Toc77966633)

[4.1.2. EXPERIMENTAL SETUP 44](#_Toc77966634)

[4.2. Experimental Result 47](#_Toc77966635)

[4.2.1. Experimenting with J48 decision tree 47](#_Toc77966636)

[4.2.2. Experimenting PART Rule Induction 49](#_Toc77966637)

[4.2.3. Experimenting Naïve Bayes classification algorithm 51](#_Toc77966673)

[4.2.4. Experimenting Logistic Regression classification algorithm 53](#_Toc77966675)

[4.3. COMPARISON OF CLASSIFICATION MODELS 55](#_Toc77966676)

[4.4. GENERATED RULES FROM RULE PART 57](#_Toc77966677)

[4.5. Discussion on Major Findings 58](#_Toc77966678)

[CHAPTER FIVE 59](#_Toc77966679)

[CONCLUSION AND RECOMMENDATION 59](#_Toc77966680)

[5.1. CONCLUSION 59](#_Toc77966681)

[5.2. RECOMMENDATION AND FUTURE WORKS 60](#_Toc77966682)

[REFERENCE 61](#_Toc77966683)

[ANNEXES 64](#_Toc77966684)

[72](#_Toc77966685)

# List of Figures

[Figure 2.1 knowledge discovery in database (KDD) Process 19](#_Toc77959629)

[Figure 2.2 CRISP-DM Process modeling 21](#_Toc77959630)

[Figure 2.3 The hybrid process model 23](#_Toc77959631)

[Figure 2.4 The SEMMA Analysis Cycle 23](#_Toc77959632)

[Figure 3.1 Preprocess data, sample dataset saved in CSV format 39](#_Toc77959633)

[Figure 4.1 Weka Interface 41](#_Toc77959634)

[Figure 4.2 loaded data view with Weka interface 42](#_Toc77959635)

[Figure 4.3 balanced data 43](#_Toc77959636)

[Figure 4.4 output of attribute ranking with information gain 44](#_Toc77959637)

# 

# List of Tables

[Table 3.1 Selected attributes with their description from SRMIS dataset. 32](#_Toc77960247)

[Table 3.2 summary of incorrect or inconsistent values and method for handling them. 34](#_Toc77960248)

[Table 3.3 summary of missing values and method for filing them. 35](#_Toc77960249)

[Table 3.4 discretized attributes and their values 36](#_Toc77960250)

[Table 3.5 discretized grade 12 result attributes and their values. 37](#_Toc77960251)

[Table 3.6 discretized CGP attributes and their values 38](#_Toc77960252)

[Table 4.1 Some of J48 decision tree algorithm parameters with their value 45](#_Toc77960253)

[Table 4.2 Performance result of J48 Decision tree with 10-fold cross validation 47](#_Toc77960254)

[Table 4.3. Performance result of J48 Decision tree with percentage split (66%) 48](#_Toc77960255)

[Table 4.4 Performance results of PART Rule Induction with 10 cross validations 49](#_Toc77960256)

[Table 4.5 Performance result of PART Rule Induction with percentage split 50](#_Toc77960257)

[Table 4.6. Performance result of Naïve Bayes with default 10-fold cross validation 51](#_Toc77960258)

[Table 4.7. Performance result of Naïve Bayes algorithm with default percentage split. 52](#_Toc77960259)

[Table 4.8. Performance result of Logistic Regression with default 10-fold cross validation 53](#_Toc77960260)

[Table 4.9. Performance result of Logistic Regression algorithm with default percentage split 54](#_Toc77960261)

[Table 4.10 Performance Comparison of the selected models 55](#_Toc77960262)

[Table 4.11 Confusion Matrix for PART algorithm using resampling with 10 fold cross validation 56](#_Toc77960263)

# List of Acronyms and Abbreviations

**CGPA** \_\_\_\_\_\_Cumulative Grade Point Average

**DBMS \_\_\_\_\_\_\_** Data Base Management System

**EC \_\_\_\_\_\_\_**Ethiopian calendar

**ECLCE \_\_\_\_\_**\_\_Ethiopian School Leaving Certificate Examination

**GC** **\_\_\_\_\_\_\_**Gregorian Calendar

**EDM** **\_\_\_\_\_\_\_**Educational Data Mining

**CART**  \_\_\_\_\_\_\_Classification and Regression Tree

**CLI**  **\_\_\_\_\_\_\_**Command Line Interface

**CRISP-DM** \_\_\_\_\_\_\_Cross Industry Standard Process for Data Mining

**CSV \_\_\_\_\_\_**Comma Separated Values

**DM** Data \_\_\_\_\_\_ Mining

**EC \_\_\_\_\_\_** Ethiopian Calendar

**EDM** \_\_\_\_\_\_\_Educational Data Mining

**GC** \_\_\_\_\_\_\_Gregorian Calendar

**GPA \_\_\_\_\_\_\_**Grade Point Average

**IEEE**  \_\_\_\_\_\_\_Institute of Electrical and Electronics Engineers

**KDD** \_\_\_\_\_\_\_\_Knowledge Discovery in Database

**dd/mm** \_\_\_\_\_\_\_ date/month

**SE** \_\_\_\_\_\_\_System Engineering

**SEMMA** \_\_\_\_\_\_\_Sample Explore Modify Model Assess

**SMOTE** \_\_\_\_\_\_\_Synthetic minority oversampling technique

**SRIMIS \_\_\_\_\_\_\_\_**Student Record Management Information System

**SMU \_\_\_\_\_\_\_\_\_**St Marry University

**SVM \_\_\_\_\_\_\_\_\_** Support Vector Machine

**UAT** \_\_\_\_\_\_\_\_\_University Admit ion Test

**WEKA \_\_\_\_\_\_\_\_**Waikato Environment for Knowledge Analysis

**WWW** \_\_\_\_\_\_\_\_World Wide Web

# Abstract

Nowadays student dropout became a universal problem in higher education. To improve student dropout, one should understand the reason behind student dropout. Student dropout in higher education institution can be affected by a wide variety of factors. These factors include, demographic, social, economic, academic and institution aspects, which are the major contributing aspects that leads to dropout of students in higher education.

The main objective of this study is to develop a predictive model using data mining technology to identifying drop out of students. In this study, the KDD data mining process model is followed. The KDD data mining process model has five steps select a target data set, data pre-processing, data transformation, data mining, and interpretation/evaluation. In this study based on the problem understanding, nine attributes are selected, and 2872 instances are used to experiment with designing a predictive model that can be determine students’ status. The study tries to understand factors affecting higher education student dropout based on the data collected from St. Mary ‘s University from the years 2011 entry. The data obtained from SRMIS (student record management information system) was in two table formats. So, merging the two tables into one table format was the major challenge of this study. It is also difficult to get well organized, correct, enough and quality data for the mining tasks. Data has been prepared using data cleaning and data transformation. Classification algorithms such as J48 Decision tree, PART Rule induction, Naïve Bayes, Logistic regression are used in the model building process. And 10-fold cross-validation and 66% split test option are used to train and test the classifier model. WEKA 3.8.4 open-source software was used as a data mining tool to implement the experiments. Among the four algorithms tested, the PART Rule induction algorithm scored the highest accuracy of 74.5% followed by the J48 Decision tree, Naïve Bayes, Logistic regression algorithms, respectively. Depending on the extracted hidden pattern using the PART algorithm marital status (single, marred, divorced), Grade 12 result, Age, GPA, and nationality were identified as the major contributing factors behind student dropout.

From the experiment the PART algorithm found that has better accuracy, so the study use PART to extract rule and develop a model. To get a best model additional attribute should be study for that, we suggest educational institutions to maintain their data symmetrically for data analyses.

Keyword: - Data Mining, Educational Data Mining, Classification algorithm, dropout,

# CHAPTER ONE

## 1.1 INTRODUCTION

Data mining can discover hidden information to inform decision-making in various domains. Data mining has a wide range of applications in different areas, including marketing, telecommunications, scientific discovery, surveillance, banking, fraud detection, and educational research [14]. The education system is one of these domains.

Educational data mining (EDM) is “an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings with the aim of developing models to improve the learning experience and institutional effectiveness” [14].

Data mining in higher education is a recent and popular research field and this area of research is gaining popularity because of its potentials for educational institutes.

There is currently an increasing interest in researching the topic of university dropout around the world, [24]. Dropout negatively affects institutions in reducing enrolment and the non- achievement of institutional objectives like number of graduate students in the university [20]. Therefore, students, universities and governments are affected in both economic and social terms. Furthermore, dropout becomes a critical topic when university administrators do not possess the tools necessary to identify students who are at risk of leaving the institution. [24].

Dropout has been a common education issue that needed attention. A high rate of dropout can have bad impact on the university [15]. The increasing number of drop-out makes many researchers attempt to find the best influencing factors. They use the data mining method to model the dropout, the dropout can be influenced by many factors from academic [15]. The academic factors that mostly used are course scores, the number of students present in the course, and achievement. [12].

Dropout cases cannot be determined only using the academic factors always. There are non-academic factors that can influence the student to dropout [15]. The dropout can be influenced by the demographics, the finances, the motivation, the social interaction, and the personality. [12].

Students leave the University not necessarily because of performance, but also other factors external to academics. Like, psychosocial factors, environmental factors, and socialization factors. Using data mining techniques which are the most important variables that affect a student decides to drop out of higher education, can be determined and the importance of the variables can be established, both the academic results and the socioeconomic situation influence a student's decision to stay in his or her respective career. [2].

The dropout can be influenced by many factors’ such as academics [15]. The academic factors that are mostly used are course scores, the number of students present in the course, and achievement. [12]. Researchers use data mining methods to model the dropout.

Dropout negatively affects institutions in the reduction of enrolment and the non-achievement of institutional objectives. Therefore, students, universities and governments are affected in both economic and social terms. Furthermore, dropout becomes a critical topic when university administrators do not possess the tools necessary to identify students who are at risk of leaving the institution.

In turn, potential corrective measures are reduced, [16]. which might have enabled student retention at higher education institutions. In the same way, the early prediction of student dropout has become a major challenge and identifies the factors that contribute to this increasingly occurring phenomenon. [16]. One possible reason that there are still high university dropout rates may be associated with the fact that most of the prediction models applied to solve this problem are difficult to interpret. [24]. A significant effort has been made to close the university dropout gap and thus reduce dropout rates. Nonetheless, this effort has been insufficient. [24].

In this study, we want to understand the key determinants of dropout, to accurately identify students with high probability of dropping out, analyzing the best factors that can influence the dropout of university student, and make a model to predict the student dropout with high accuracy using well-known classification methods. To do this, we model student dropout using data gathered from academic databases. The correlation between the results which will be obtained in this work will be presented.

## 1.2. PROBLEM STATEMENT

University Student dropout is a problem that affects universities around the world, with consequences such as reduced enrolment, reduced profits for the university and the country as all, and financial losses for the State which funds the studies, and constitutes a social problem for students, their families, and society in general,

The early identification of vulnerable students who are prone to drop out from the university is crucial for the success of any university withholding a good strategy. And to try to reduce the problem, it is necessary to detect students who are at risk as early as possible and thus provide some care to prevent these students from quitting their studies. This would allow educational institutions to undertake timely and proactive measures [24]. Hence it is very important, predicting university dropout or identify dropout students in advance to design a model for tackle this problem.

Though various studies have been conducted using data mining technology, all the available knowledge is not enough to solve the entire prediction problem to determine student dropout. The studies lack country specific behaviors, and entirety of factors contributing to dropout. In this study, some unique attributes are used other researcher did not consider it. These includes marital status, nationality, and age. 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.

This research is also different from the other works since it focused on student dropout for five semester which is since 2011. And Student dropout include academic dismissal, academic suspension, and withdrawal (occurs each semester until graduation.) So, this study aims to apply Data Mining techniques for constructing a model for predicting the dropout of higher education students. Hence this research attempts to explore and answer the following main guiding questions.

* Which set of data attributes can be a factor to identify at risk student’s dropout?
* Which DM algorithms is best used to develop a model for predicting student dropout?
* To what extent the predictive model determines the risk status of students?

## 1.3. SIGNIFICANCE/IMPLICATIONS

It is imperative to study predicting university dropout and develop a model giving that such models have been significance factor in defining and examine the student behavior, that leads to dropout. Identifying at early state can prevent the dropout by taking necessary action towards the dropout reason. The findings of this study have a significant contribution to students, family, school, and the country as a whole. 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.

## 1.4. OBJECTIVES

The prime objective of the proposed model is to provide a solution for dropping out students from Higher Education and tried to show where the Institutions can get information that provides an insight into the respective effect which is Universities and other external stakeholders as to gain the optimal advantage.

### General objective

The main objective of this study is to develop a predictive model to identifying the drop out of students at St. Marry University.

### Specific objective

To achieve the general objective, the following specific objectives have been formulated.

* To identify a set of data attributes that can be a factor for student’s dropout.
* To build a model for the student dropout.
* To identify which algorithms are best to develop a model for predicting student dropout.
* To determine to what extent the predictive model determines the risk status of students

## 1.5. SCOPE AND LIMITATION OF THE RESEARCH

The main aim of this research is to find out the applicability of DM for determining and predicting dropout of students The research is delimited one educational institute and dropout of undergraduate regular and extension student information enrolled in a year of 2011. The distance and the postgraduate program students are not included in this study.

The study used limited attribute because could not get all needed attribute to use as a factor to determine higher education student dropout because of unavailability of clear data found in the database, other demographic data such as distance from home, family background, native language, and place of birth location (urban or rural), and health-related data are not included under this study.

## 1.6. ORGANIZATION OF THE THESIS

The rest of this study is organized as follows:

Chapter two covered Literature review in education data mining it discusses related to data mining and related work on the basic concept of the predictive model on drop out students and others: different publications and related works done on Predicting model of dropout of the student.

Chapter three describes the Research Methodology and strategy that aims to provide a model for identifying drop out students at St. marry University. It covers the detailed description of data mining methodology, how the data is prepared for the data mining task, and the transformations performed on the data are presented.

Chapter four presents an experimental setup on the selected data mining algorithms and model development and testing results of classification algorithms for prediction by using WEKA data mining software tool. Extracting rules and discussion of the results are presented.

Chapter Five presents the conclusion of the study and recommendation for further work in the area.

# 

# CHAPTER TWO

## LITERATURE REVIEW AND RELATED WORK

## 2.1 LITERATURE REVIEW

This chapter reviews journals, books, the internet, and research paper related to data mining to have detailed knowledge about the problem domain, a detailed conceptual understanding of data mining, data mining process model, the data mining tasks, overview of educational data mining and the tasks are presented first, this is followed by a review of different related works concerning student dropout.

Many recent studies have focused on the problem of drop out of the student. While there has been much research on the area

### 2.1.1 OVERVIEW OF DATA MINING

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically [14]

Data mining is a process of extraction of useful information and patterns from huge data. It is also called a knowledge discovery process, knowledge mining from data, knowledge extraction or data pattern analysis. [11]

Data Mining (DM) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining (DM) and knowledge discovery are intelligent tools that help to accumulate and process data and make use of it. DM bridges many technical areas, including databases, statistics, machine learning, and human computer interaction. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. [4].

Data mining, also called knowledge discovery from data (KDD), is the automation of extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, the web, other massive information repositories, or data streams [14].

According to Jiawei et al [14], Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, in contrast, others view data mining as merely an essential step in the process of knowledge discovery.

### 2.1.2 EDUCATIONAL DATA MINING (EDM)

It is used to study available data in the educational context and extract value from the hidden information. This information can be used in several educational processes such as predicting course enrollment, estimating student dropout rate, detecting atypical values in students’ transcripts, and performance prediction. [31]

Educational data mining also referred to as EDM uses mining strategies and techniques to answer few important educational questions. It is based on computational approaches that analyze data collected from academic institutions. This can also help improve e-learning experience for students. EDM is dependent on the distribution of data collected; hence slight variations can cause a change in results. [7]

Educational data mining is new growing research for knowledge discovery from a large amount of educational data. Different countries researchers are putting in their efforts in finding out patterns and factors that can be helpful in the progress of education. Data mining can be used in the education filed to enhance our understanding of the learning process to focus on identifying extracting and evaluating variables related to students’ learning process. The essences of data mining concepts are used in the educational fields for the purpose of extracting useful information on student behavior in the learning process [23].

### 2.1.3. DATA MINING MODELS

The model helps an organization to better understand the process model and provides a roadmap to follow while planning and executing the project. This in turn results in cost and time savings, better understanding, and acceptance of the results of such projects. Before one attempts to extract useful knowledge from data, it is important to understand the Overall approach. Simply knowing many algorithms used for data analysis is not sufficient for a Successful DM project.

KDD, SEMMA and CRISP-DM are popular data mining process models that are used in data mining projects. [11]

### 2.1.3.1. Knowledge discovery Process

The knowledge discovery process is an iterative sequence of the following steps [19] which is shown below in Figure.2.1

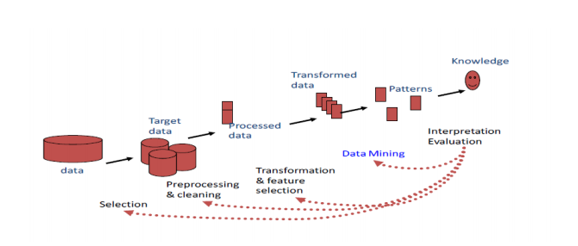


Figure 2.1 knowledge discovery in database (KDD) Process

The goal of a process model is to make the process repeatable, manageable, and measurable. This section discusses the most popular datamining process model knowledge Discovery Database (KDD) process model. The data mining and researchers mostly practice this model.

According to [18] the process of extracting knowledge from big data consists of five sequenced steps namely selection of dataset, data preprocessing, data transformation, data mining, interpretation/ evaluation, and brief descriptions of each of the steps in each broadly used methodologies in data mining are as follows:

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

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

**Step 3. Data Transformation**: During the transformation phase, data are combined into forms appropriate for mining to reduce the data size by dividing the range of attribute value into intervals each containing approximately same number of samples or to scale attribute data to fall within a 19 specified range. Therefore, values of attributes are changed to a new set of replacement values to ease data mining. For example, discretization of variables or production of derived variables is transformation of data.

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

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

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

### 2.1.3.2 CRISP-DM process model

Cross Industry Standard Process for Data Mining (CRISP-DM) is a comprehensive data mining methodology and process model that provides anyone—from novices to data mining experts—with a complete blueprint for conducting a data mining project

It breaks down the life cycle of a data mining project into six phases.

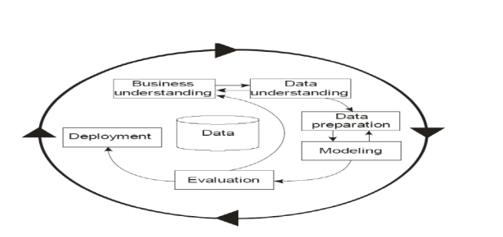


Figure 2.2 CRISP-DM Process modeling

Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective; and then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Data Understanding

The data understanding phase starts with an initial data collection. It proceeds with activities to become familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses.

Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

Modelling

In this phase, various modelling techniques are selected and applied; their Parameters are calibrated to optimal values. Typically, there are several techniques for the same Data mining problem type. Some techniques have specific requirements in the form of data. Therefore, stepping back to the data preparation phase is often needed.

Evaluation

At this stage in the project the model (or models) built appears to have high quality from a data analysis perspective. Before proceeding to the final deployment of the model, it is important to evaluate the model more thoroughly, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been considered sufficiently. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

The creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the client can use. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the client, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the client to understand up front what actions will need to be carried out to make use of the models created. [11]

2.1.3.3 Hybrid Models

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

The main differences and extensions include.

Providing more general, research-oriented description of the steps,

introducing a data mining step instead of the modelling step,14

introducing several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and

Modifying the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains. [11]

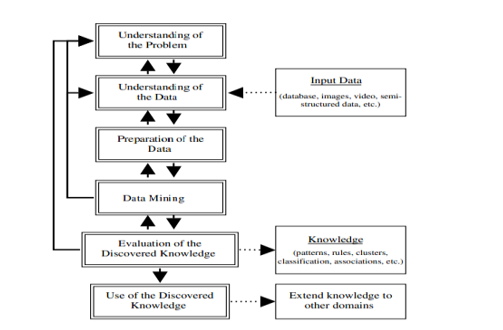


Figure 2.3 The hybrid process model

**2.1.3.4** The SEMMA Process Model

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

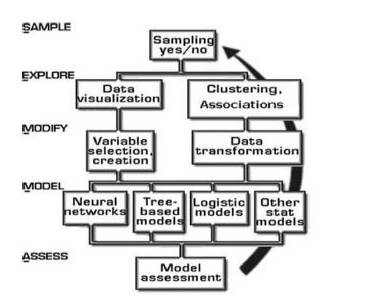


Figure 2.4 The SEMMA Analysis Cycle

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

**Explore**: this stage consists of exploring the data by searching for unanticipated trends and anomalies to gain understanding and ideas.

**Modify:** This stage consists of modifying of the data by creating, selecting, and transforming the variables to focus the model selection process.

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

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

### 2.1.4. DATA MINING ALGORITHMS AND TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Artificial

Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithms, the Nearest Neighbor method etc., are used for knowledge discovery from databases. [11]

#### 2.1.4.1. Classification

Classification is the most applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. The data classification process involves learning and classification. In Learning, the training data are analyzed by a classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples [11]

We are going to take a tour of 5 top classification algorithms in Weka.

Types of classification models which is the 5 top classification algorithms in Weka are.

1. Regression
2. Bayesian Classification
3. Decision Tree
4. Rule-Based Classifiers
5. Support Vector Machines (SVM1)

**A. Regression**

Regression is predictive modelling and analysis is used to make predictions based on existing data by applying formulas. It is a statistical method of data mining. Regression used to model between one or more dependent variables and independent variables. It can be used for building models 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 using existing values to forecast what other values will be [32].

**B. Logistic regression**

Logistic regression is a binary classification algorithm. It assumes the input variables are numeric and have a Gaussian (bell curve) distribution. This last point does not have to be true, as logistic regression can still achieve good results if your data is not Gaussian. In the case of the Ionosphere dataset, some input attributes have a Gaussian-like distribution, but many do not.

The algorithm learns a coefficient for each input value, which are linearly combined into a regression function and transformed using a logistic (s-shaped) function. Logistic regression is a fast and simple technique but can be very effective on some problems. The logistic regression only supports binary classification problems, although the Weka implementation has been adapted to support multi-class classification problems. [32]

**C. Bayesian Classification**

Bayesian classifiers are statistical classifiers based on Bayes’ theorem. Studies [32] 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.

**D. Naive Bayes**

Naïve Bayes is a classification algorithm. Traditionally it assumes that the input values are nominal, although it numerical inputs are supported by assuming a distribution.

Naive Bayes uses a simple implementation of Bayes Theorem (hence naive) where the prior probability for each class is calculated from the training data and assumed to be independent of each other (technically called conditionally independent).

This is an unrealistic assumption because we expect the variables to interact and be dependent, although this assumption makes the probabilities fast and easy to calculate. Even under this unrealistic assumption, Naive Bayes has been shown to be a very effective classification algorithm.

Naive Bayes calculates the posterior probability for each class and makes a prediction for the class with the highest probability. As such, it supports both binary classification and multi-class classification problems.[32]

**E. Decision Tree**

Decision trees can support classification and regression problems. Decision trees are more recently referred to as Classification and Regression Trees (CART). They work by creating a tree to evaluate an instance of data, start at the root of the tree and moving town to the leaves (roots) until a prediction can be made. The process of creating a decision tree works by greedily selecting the best split point in order to make predictions and repeating the process until the tree is a fixed depth.

After the tree is constructed, it is pruned in order to improve the model’s ability to generalize to new data.[32] [10]

**J48 Decision Tree Algorithm**

J48 decision tree handles both categorical and continuous attributes to build a decision tree. J48 uses the 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 [10]. Entropy provides an information theoretic approach to measure the goodness of a split. It measures the amount of information in an attribute [10].

**Rule-Based Classifiers**

A rule is typically expressed in the following form: IF Condition THEN Conclusion [10]. 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 , =, ⊆, or ∈, 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 Ri is of the form Qi ⇒ c where Qi is the antecedent, and c is the class variable.

The “⇒” symbol denotes the “THEN” condition. The rules are generated from the training data during the training phase. The notation Qi represents a precondition on the feature set [10].

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 study PART rule induction algorithm is used to build the predictive model.

**PART Rule induction**

“The rule induction classifier used in this study is PART, which is one of the most applied rule-based classifications methods. Rules are a good way of representing information or bits of knowledge” 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 the 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 builds a partial C4.5 decision tree and converts the "best" leaf into a rule [19].

## 2.2 RELATED WORKS

[2] In recent years there has been an increase in rate of dropouts in correspondence courses and part-time courses. Many factors could be behind this increase in rate of dropout. Researchers are working towards this field to identify the factors so that teachers and management could be informed. “If the root cause of the problem is known, it will be easier to find solutions for problems.” Researchers can identify factors and predict which factors are dominating the dropout rate. This could be achieved by using data mining. Data can be taken from databases of universities and colleges and by application of various techniques we can perform educational data mining to predict useful knowledge. Data mining methods are considered better then statistical methods as the data may be huge in size and it is difficult to process large datasets using statistical methods. Data mining process involves many steps like data gathering, data preprocessing, applying mining techniques, result interpretation.

[8] The study Explain Dropouts can be due to various risk factors like financial conditions, parental education, marital status, etc. Relationship between dropout performance and these risk factors must be understood before devising any strategy. For mining of data, the tools used is one of popular tools which is Weka and suggest that educational institutions can use the outcome of analysis in strategic planning to help students improve performance.

[7]. A case study on the detection of the dropout of System Engineering (SE) showed preliminary results for predicting student attrition from a large, heterogeneous dataset of student demographics and transcript records. In the findings system engineering courses performance are correlated to physics and mathematics courses performances has been discovered. The irregularity (standard deviation of term’s averages) is positively correlated to drop out, so this accuracy could be confident enough to help in early dropping out early detection. Courses related to SE have the greatest impact in dropout prediction.

[2]. This research establishes the importance of the variables that affect the university student´s decision to drop out by using data mining techniques, comparing them with the variables identified in the literature review. The results obtained from the data provided by the Engineering departments of the University of Mumbai, in India, determine that the variables that best explain a student's dropout are the socioeconomic factors and the income score provided by the University Admission Test (UAT). According to the decision tree technique, it is concluded that the retention is 78.3%. The quality of the classifiers allows to ensure that their predictions are correct, with statistical levels of the ROC curve are 76%, 75%, and 83% successful for Bayesian network classifiers, decision tree, and neural network, respectively. The results found in this analysis show that the academic results and the socioeconomic situation influence a student's decision to stay in his or her respective career. Managing these variables helps reduce the dropout rates in the university system.

[3]. A data mining study that evaluates the number of factors that may be responsible for raising the dropout ratio of girl children in the Theni district, the data have been collected through a survey of individual people’ opinions and treated by a data mining tool called XL miner. The result shows that parent’s attitude towards a girl child’s education is positively related to the raising of school dropouts.

[30]. Introduces a methodology to predict the student dropout using Naive-Bayes Classification Algorithm in R language. And examine the reason for student drop out at an early stage and predict whether the student will drop or not. It also mentions the factors that affect a student to commit dropout. The method that the paper proposed is a combined approach that takes into consideration factors such as demographics, academic performance, health issues, place of residence etc. and argue that the existing method is very time consuming and not very accurate and focuses on only specific factors but the proposed one is which increases the accuracy and implements methods that reduce the time taken for prediction.

[7]. This paper understands the key determinants of dropout, to accurately identify students with a high probability of dropping out, and find the important characteristics associated with the graduation level of students in a computer science program. Data was gathered from academics and used two approaches to identify the key determinants of dropout. The first one following the CRISP–DM methodology and applying Decision Tree, Logistic Regression and Naive Bayes models. The second one, using Watson Analytics to automatically establish these determinants. They do not consider data related to the enrollment process like demographic information. Main results are presented to decrease the dropout rate by identifying potential causes. In addition, they present some findings related to data quality to improve the student’s data collection process.

[24]. Made a systematic review of literature on the prediction of university student dropout through data mining techniques. The study was developed as a systematic review of the literature of empirical research results regarding the prediction of university dropout. In this phase, the review protocol, the selection requirements for potential studies and the method for analyzing the content of the selected studies were provided. To address the problem of dropout, they try to identify highly accurate techniques that are being developed, however they cannot identify one technique that is clearly superior, but they suggest for prediction accuracy depends mainly on the context, data, and technique characteristics; any potential alternative must consider these factors.

[4]. Analyses the factors affecting students’ academic performance that contributes to the prediction of their failure and dropout using educational data mining. This paper suggests the use of various data mining techniques to identify the weak students who are likely to perform poorly in their academics. WEKA, an open-source tool for data mining was used to evaluate the attributes predicting school failure. Various classification techniques like induction rules and decision tree have been applied to the data. The results of each of these approaches have been compared to select the one that achieves high accuracy.

[12]. The researchers raise hypothesis that the potential dropout students can be determined from non- academic factors. Five non-academic factors criteria that can be used as determinants of dropout, demography, social interaction, finance, motivation, and personal. These criteria give rise to 37 factors that are considered influential in determining the potential dropout. The factors processed into three phases are collecting data, preprocessing data, and modelling.

Based on the result of correlation test there are two factors had correlation, so the modelling done with two combination factors. They suggest that the best model is using combination of factor.

[25]. Presents the most recent investigation of student dropout at Mae Fah Luang University, Thailand, and the novel reuse of link-based cluster ensemble as a data transformation framework for more accurate prediction. The empirical study on mixed-type data collection related to students’ demographic detail, academic performance, and enrollment record, suggests that the proposed approach is usually more effective than several benchmark transformation techniques, across different classifiers.

Asmerom et.al [18] examined the magnitude of students’ dropout by faculties and subject. According to the study major cause of students’ dropout can be categorized as academic and non-academic. However, "dropping out" from higher institutions is basically dependent on academic performance.

Getahun [27] explored the possibility of applying data mining techniques for predicting the likelihood of a student’s dropout. In this work, the data of the students is extracted from the student information management system of the St. Mary’s University College, most of the required attribute for building the model were generated form 2007/08. J48, Random Forest, Multilayer and Perceptron are selected to conduct this experiment. the strongest predictor of dropout was found to be CGPA, as a result, higher performance is demonstrated more by (J48 and random forest), the findings of the study indicate that students’ dropout is more related to performance than other predictors in the study considered.

# 

# CHAPTER THREE

## 3.1 METHODOLOGY

The 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.

### 3.1.1 RESEARCH DESIGN

This study follows experimental research. Experimental research designs are selected because the approach used to investigate causal (cause/effect) relationships which study the relationship between one variable and another. Researchers use experimental research to compare two or more groups on one or more measures [21].

To conduct an experiment. the study uses the KDD data mining process model. This process model is selected because it is considered as a programmed, exploratory analysis and modelling of vast data repositories.

KDD data mining process model has five steps: Selection, Preprocessing, Transformation, Data Mining, and Interpretation/Evaluation

### 3.1.2. DATA COLLECTION

Originally the study was planned to conduct with data from Government University, but it was difficult to get data. And after try to look at CPU College and Unity University but still couldn’t get enough data for all selected attribute Comparing to St. Marry data so, we chose St. Marry University to conduct the study.

The data source of this study is SRMIS (student record management information system) of St. Mary’s University. The database was filled up with different attributes which are interrelated to students. Among those attributes, depending on problem understanding and available information the selected attributes are related to student status (drop out and Active). The dataset comprised 2870 undergraduate degree students who registered in 2011 and may graduate from the university in 2013 and 2014. The collected data was organized in a Microsoft Excel sheet. Each student record had the following attributes:

Age, Sex, Program, 12th result, Department, , Nationality, Marital status, CGPA (cumulative GPA) and Students’ status( Dropout or Active) are selected to build the predictive model. For the data analysis student’s data covering from 2011EC are considered. The original dataset obtained from SRMIS (student record management information system) has eleven attributes and more than 2870 instances. Before obtained the final dataset, incomplete and missed data were removed from each record. In this work, attributes such as Name, and student ID are removed because they are identifiers of students in this study.

#### 3.1.2.1. Data description

During this phase, the data analyst examines the “gross” or “surface” properties of the acquired data and report the results, examining issues such as the format of the data, the quality of the data, the number of records and fields in each table, the identity of the field, and any other surface features of the data [28]. The obtained data from SRMIS was in two different table form which the first table includes ID, Student Full Name, Date of birth, Sex, Program, Admission Date, Nationality, CGPA, Department and grade 12 result and the second table includes Student Id, Student full name, Address, Marital status, and Student status (Drop or Active). The two tables are merged into one table by cross-matching based on students ID numbers. Finally, nine attributes are selected. The descriptions of the selected attributes are listed below in table 3.1.

Table 3.1 Selected attributes with their description from SRMIS dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Variables | Descriptions | Data types | Possible values | Remark |
| 1 | Age | Age of student | Nominal | One (19-29), two (30-40), three (>40) | Original |
| 2 | SEX | Sex of student | Nominal | Male. Female | Original |
| 3 | Department | A stream the students intended to learn | Nominal | marketing, Computer science accounting and finance and management | Original |
| 4 | Nationality | Citizen of student’s particular nation | Nominal | Ethiopian, foreigner | Original |
| 5 | Program | Modality | Nominal | Extension, Regular | Original |
| 6 | CGPA | Current cumulative result or the result before dropout | Nominal | <1.5 (low)  1.5 -2.0 (Satisfactory)  2.0 -2.5 (good)  2.5 - 3.5 (v good)  3.5 – 4 (Excellent) | Original |
| 7 | Marital Status | Student’s relationship | Nominal | Single, Marred, Divorce | Derived |
| 8 | Grade 12 result | Preparatory, ECLCE results and Diploma grade report | Nominal | Satisfactory, fair Good, and very good | Original |
| 9 | Dropout | The academic status | Nominal | Yes. No | Derived |

### 3.1.3. Data Pre-processing

This step consists of the core of data mining. It takes much of the research time and effort of the entire knowledge discovery process but, is the most necessary activity in the data mining process. Data preprocessing involves all the action taken before the actual data analysis process starts. [14]

## 3.1.3.1. Data cleaning

Data cleaning is routine work to “clean” the data by filling in missing values, smoothing noisy data, identifying, or removing outliers, and resolving inconsistencies [14]. Without clean data, the result of a data mining analysis is in question. In this study data cleaning has been applied to handle incorrect, inconsistent, and missing entry values which emerged from the attended result.

The dataset collected from the university had some common mistakes such as inaccuracies, missing score, and inconsistent data. Therefore, to achieve the study goal, attributes such as attribute age is obtained by converting date of birth in this manner, they were cleaned from the dataset.

## 3.1.3.2. Handling the incorrect or inconsistent values.

The incorrect and inconsistent values have been noticed in different attributes. The key methods that are used for correcting the incorrect and inconsistent entries are as follows [1]

* **Inconsistency detection**: this typically done when the data is available from a different source in different format.
* **Domain knowledge:** A significant amount of domain knowledge is often available in terms of the ranges of the attributes or rules that specify the relationships across different attributes.
* **Data-centric methods**: In these cases, the statistical behaviour of the data is used to detect outliers. Data-centric methods for cleaning can sometimes be dangerous because they can result in the removal of useful knowledge from the underlying system.

Attribute: The grade 12 result is one of the attributes in which data is registered in different format. There are student’s data score from range 150 to 600 out of 700 and another format of grade 12 result is like from 2.0 to 2.8 out of 4 for a student who took ECLEC exam in old curriculum and a diploma student. The number of data which have grade 12 result in the second format is 23 students out of all.

The researcher converts the ECLCE and Diploma result in to the first format which is out of 700 to keep consistency.

Attribute: Age, In the data base Date of birth is available so It convert to age

Table 3.2 summary of incorrect or inconsistent values and method for handling them.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Attribute | Total Instance | No of inconsistent Value | In % | Handling Mechanism |
| 1 | Department | 2872 | 8 | 0.28% | Domain knowledge |
| 2 | Age () | 2872 | 10 | 0.34% | Domain knowledge |
| 3 | Age | 2872 | 2 | 0.07% | Data-centric methods |
| 5 | Grade 12 result | 2872 | 23 | 0.8% | Inconsistency detection |

#### 3.1.3.2. Handling missing values

As Jiawei et.al [14] suggested, some of the methods to handle the missing values are presented as follows.

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

Table 3.3 summary of missing values and method for filing them.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Attribute | Total Instance | No of inconsistent Value | In % | Handling Mechanism |
| 1 | Age | 2872 | 6 | 0.2% | mean for all samples belonging to the same class as the given tuple |
| 2 | Grade 12 result | 2872 | 2 | 0.07% | The most probable values using mode |
| 3 | Department | 2872 | 5 | 0.17% | The most probable values using mode |
| 4 | Nationality | 2872 | 12 | 0.42% | The most probable values using referring name |

### 3.1.4. DATA TRANSFORMATION

Data transformation is about transforming or consolidating the data to make it appropriate for mining.

In this stage, the creation of appropriate data for Data Mining is prepared and developed. Techniques here incorporate dimension reduction (for example, feature selection and extraction and record sampling), also attribute transformation (for example, discretization of numerical attributes and functional transformation). [14].

#### 3.1.4.1. Data discretization

Jiawei et. al [14] noted that Data discretization could be used to replace a numeric attribute by interval labels or conceptual labels. The labels, in turn, can be recursively organized into high-level concepts, resulting in concept hierarchy for the numeric attributes. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data. This leads to a concise, easy to use, knowledge-level representation of mining results. In this study, five attributes are discretized into their respective represented values these are:

Age: Age attribute contains numeric instances of students’ age from 19 to 50.

Age attribute discretized into the following.

* one: Students are joined the University at the age of 19-29.
* two: students are joined the University in between age 30 and 39.
* three: students are joined the University at the age of 40 or 50.

Table 3.4 discretized attributes and their values

## 

|  |  |  |
| --- | --- | --- |
| Attribute name | Values | Represented values. |
| Age | 19-29 | One |
| 30 and 39 | Two |
| 40-50 | Three |

Grade 12 Result: grade 12 result attribute contains numeric instances of students’ grade 12 result from 161 to585 out of 700.

Grade 12 result attribute discretized into the following.

* satisfactory: Students they got on grade 12 result between 160 to 200.
* fair: Students they got on grade 12 result between 201 to 299.
* good: Students they got on grade 12 result between 300 to 399.
* very good: Students they got on grade 12 result greater than 400.

Table 3.5 discretized grade 12 result attributes and their values.

|  |  |  |
| --- | --- | --- |
| Attribute name | Values | Represented values. |
| Grade 12 result | 160-200 | satisfactory |
| 201-299 | fair |
| 300-399 | good |
| >400 | Vgood |

CGPA: CGPA (Cumulative Grade point average) attribute contains numeric instances of students’ last GPA before dropping out for dropout student and current GPA for active students from 1 to 4

GPA attribute discretized into the following.

* low: Students with a GPA between 1.00 to 1.49
* satisfactory: Students with GPA between 1.50 to 2.00
* good: Students with GPA between 2.01 to 2.49.
* very good: Students with GPA greater than 2.51 to 3.50
* Excellent: Students with GPA greater than 3.51 to 4.00

Table 3.6 discretized CGP attributes and their values

|  |  |  |
| --- | --- | --- |
| Attribute name | Values | Represented values. |
| CGPA | 1.00 to 1.49 | low |
| 1.50 to 2.00 | satisfactory |
| 2.01 to 2.49. | good |
| 2.50 to 3.50 | vgood |
| 3.51 to 4.00 | Excellent |

### 3.1.5 DATA FORMAT

In this study, the original data extracted from the student record management information system My-SQL database of SMU. We preprocessed the original data using a different preprocessing technique using Microsoft Excel.

The attribute of Age is converted from the date of birth from the original data, including the dd/mm/year format.

To get the age first the year only discard from the date and month. Some of the student’s year was recorded in G.C (Gregorian calendar) so that it changed to E.C (Ethiopian calendar) and then taking only the year and subtracted from the year 2011.

Before applying the data mining algorithm, the data converted into CSV (Comma Delimited) file format which is Weka acceptable format. Figure 3.1 shows sample dataset saved in CSV format for the data mining tasks using classification algorithms.

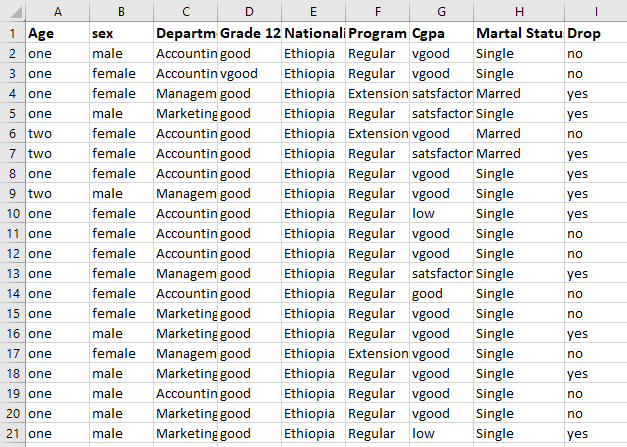


Figure 3.1 Preprocess data, sample dataset saved in CSV format

# CHAPTER FOUR

## 4.1 EXPERIMENTATIONS AND ANALYSIS

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 dropout in data mining classification techniques that were applied to develop predictive models. Therefore, it is important to conduct different experiments to find the best model for solving the problem.

### 4.1.1 BUILDING MODELING

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

#### 4.1.1.1 Selecting the modeling technique

Selecting an appropriate model depends on data mining goals. Consequently, to attain the objectives of this research four classification algorithms has been selected for model building.

In this study, a total of 16 experiments are conducted using the four-classification algorithm stated above using 66% percentage split test and 10-fold cross-validation test options. The researcher selected the above algorithms, because its popularities in recently published papers, ease of understanding and interpretation of the Result of the model.

#### 4.1.1.2 Weka interface

Weka version 3.8 software. According to Priyanga et. Al. [9], Weka is open-source software developed at the University of Waikato and the programming language is based on Java. Weka has four different applications, Explorer, Experimenter, Knowledge Flow, and Simple CLI. Knowledge Flow is a node and linked based interface and Simple CLI is the command line prompt version where each algorithm is run by hand. In this study, Explorer applications of the Weka has been used. WEKA contains many inbuilt algorithms for data mining and machine learning. Weka implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes visualization tools.

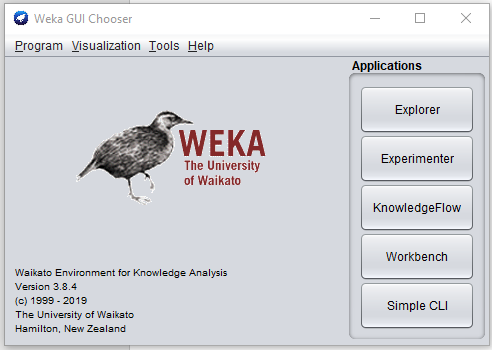


Figure 4.1 Weka Interface

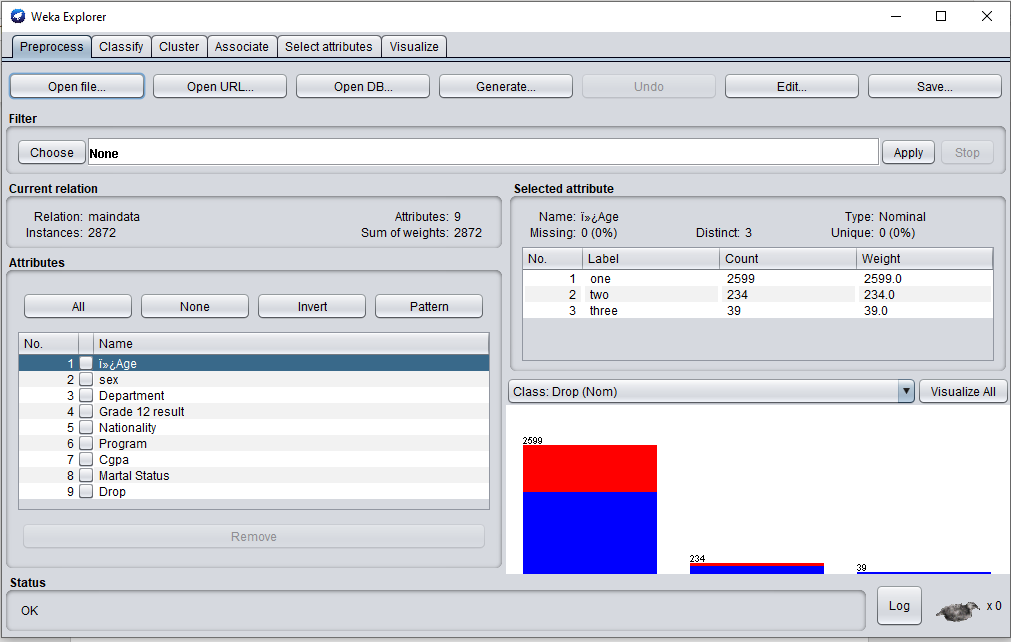


Figure 4.2 loaded data view with Weka interface

#### 4.1.1.3. 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 [26]. Therefore, to solve the problem of class imbalance we applied supervised class balancer technique by changing the default values of (biasToUniformClass (0.0) into (1.0)). As shown in figure 4.2 after we applied to resample the three dependent classes become equal with each class value success, average and weak having 2872 instances

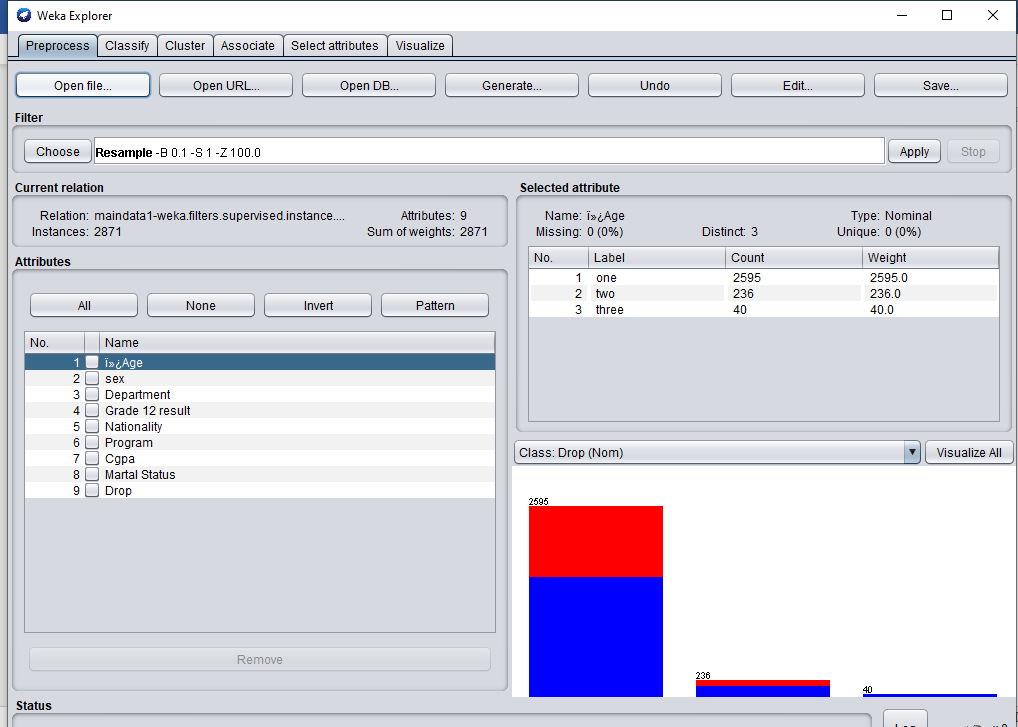


Figure 4.3 balanced data

#### 4.1.1.4. Attribute Ordering

Since attribute selection is important because in most cases all attributes are not equally useful for predicting the target, so better 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.

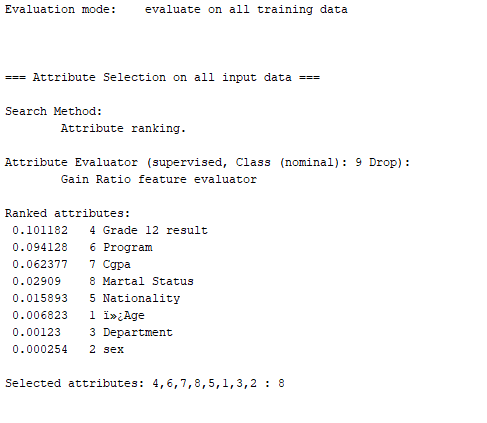


Figure  4.4 output of attribute ranking with information gain

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

### 4.1.2. 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 ten mutually exclusive subsets or folds, 1,2,3,4 ...10, each approximately with equal size. The training and testing are performed ten 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.

#### 4.1.2.1. 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 4.1 and annexe 3.

Table 4.1Some of J48 decision tree algorithm parameters with their value

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

For J48 decision tree four test modes were considered:

* Experiment 1: Pruned J48 algorithm with 10-fold cross-validation test mode
* Experiment 2: Pruned J48 algorithm using resampling with a 10-fold cross-validation test option.
* Experiment 3: Pruned J48 algorithm with Percentage (66%) split test mode.
* Experiment 4: Pruned J48 algorithm using resampling with Percentage (66%) split test option.

#### 4.1.2.2. 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 a better model from datasets. Four experiments were conducted. In the experiment the 10-fold cross validation and percentage split were 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 (66%) split test mode.

• Experiment 4: PART algorithm using resampling with Percentage (66%) split test mode.

#### 4.1.2.3. Classification Using Naïve Bayes

The second 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 were used. The Naïve Bayes experiment was designed to build the model for predicting student drop out and to compare the dropout 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 (66%) split test mode.
* Experiment 4: Naïve Bayes algorithm using resampling with Percentage (66%) split test.

#### 4.1.2.4. Classification Using Logistic

The fourth type of classification technique applied in this study is the linear regression algorithm. Four experiments were conducted. In the experiment the 10-fold cross validation and percentage split were used. The linear regression experiment was designed to build the model for predicting student dropout 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 (66%) split test mode.
* Experiment 4: linear regression algorithm using resampling with Percentage (66%) split test mode.

## 4.2. Experimental Result

### 4.2.1. Experimenting with J48 decision tree

Four experiments are conducted using the 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 52 and 37 leaves. The experiment has generated a model with an accuracy of 70.64%, weighted precision of 72.6%, weighted Recall of 70.6%, weighted F-Measure of 71.2% and weighted ROC area of 80.8%. From the total instances, 70.64 % are correctly classified and 29.35% are incorrectly classified. The performance of the two experiments is summarized and presented in the table 4.2 below.

Table 4.2 Performance result of J48 Decision tree with 10-fold cross validation

| Experiment | Model | Accuracy | Leaf size | Tree Size Weighted | Time Taken | Weighted TP Rate | FP Rate Weighted | Precision Weighted | Recall Weighted | F-Measure Weighted | ROC Area |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | J48 Pruned with 10 fold cross-validation | 70.6% | 37 | 52 | 0 sec | 70.6 | 29.7 | 72.6 | 70.6 | 71.2 | 81.6 |
| 2 | J48 Pruned using Resampling with 10 fold cross-validation | 73.9 % | 66 | 95 | 0.01s | 73.9 | 23.7 | 76.6 | 73.9 | 74.4 | 83.2 |

In the second experiment as shown in the above table 4.2, the same number of attributes and records are used but the data is balanced using resampling technique, the model developed with this percentage results in 73.9 % correctly classified instances and 26% incorrectly classified instances.

Accordingly, there is an improvement of performance with accuracy, weighted precision, weighted F-Measure and weighted ROC area of 73.9%, 76.6%, 74.4 and 83.2 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 dropout 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 4.3.

Table 4.3. Performance result of J48 Decision tree with percentage split (66%)

| Experiment | Model | Accuracy | Leaf size | Tree Size Weighted | Time Taken | Weighted TP Rate | FP Rate Weighted | Precision Weighted | Recall Weighted | F-Measure Weighted | ROC Area |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | J48 pruned Percentage (%) split | 69.2 | 37 | 52 | 0 sec | 69.2 | 45.5 | 67.5 | 69.2 | 67.1 | 78.9 |
| 2 | J48 pruned using Resampling Percentage (%) split | 72.10 % | 66 | 95 | 0.01 sec | 72.1 | 32.3 | 72.1 | 72.1 | 72.1 | 81.8 |

# 

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 4.3, the model developed with this percentage results in 72.1% correctly classified instances and 27.8% incorrectly classified instances. 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 better classification accuracy of 72.1 % predicting. 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.

### 4.2.2. Experimenting PART Rule Induction

Four experiments are conducted 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, 2872 dataset was used as an input. The PART Rule induction algorithm with 10-fold cross validation scored an accuracy of 70.3. This result shows that out of the total training datasets 70.3 % records are correctly classified instances, and the remaining 29.7 % 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 4.4 below.

Table 4.4 Performance results of PART Rule Induction with 10 cross validations

| Experiment | Model | Accuracy | No of Rule | Time Taken | Weighted TP Rate | Weighted FP Rate | Weighted  Precision | Weighted  Recall | Weighted F-Measure | ROC Area |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | PART Pruned with 10 fold cross-validation | 70.3% | 28 | 0.02s | 70.3 | 29.8 | 72.4 | 70.3 | 70.9 | 80.5 |
| 2 | PART Pruned using Resampling with 10 fold cross-validation | 74.5% | 38 | 0.03s | 74.5 | 23.7 | 76.8 | 74.5 | 75 | 84.1 |

In the second experiment as shown in the above table 4.4, the same number of attributes and records are used but the data is balanced using resampling technique. In this case, correctly classified instances are 74.5% and the resampling 25.46% are incorrectly classified.

Based on the above experiment, PART algorithm using resampling technique with 10 cross validations has scored a better accuracy, weighted precision, weighted F-Measure and weighted ROC area of the model were74.5 % ,76.8%, 75% and 84.1% respectively than PART algorithm without using resampling technique. Running information of PART algorithm using resampling technique with 10 cross validations (74.5%). technique is provided on annex-6

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 4.5.

Table 4.5 Performance result of PART Rule Induction with percentage split

| Experiment | Model | Accuracy | No of Rule | Time Taken | Weighted TP Rate | Weighted FP Rate | WeightedPrecision | WeightedRecall | Weighted F-Measure | ROC Area |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | PART pruned Percentage (%) split | 69.73% | 28 | 0.01 | 69.7 | 43.3 | 68.2 | 69.7 | 68.2 | 79.5 |
| 2 | PART pruned using Resampling Percentage (%) split | 73.6 % | 38 | 0.03s | 73.6 | 32.9 | 73.1 | 73.6 | 73.2 | 82.9 |

As shown in table 4.5, PART rule induction without applying resampling techniques 69.7% correctly classified instances, while 43.3% of the records are incorrectly classified instances. On the other hand, PART rule induction using resampling technique with percentage split scored an accuracy of 73.9 % while 26.4% of the records are incorrectly classified instances. constructed a model with accuracy, weighted precision, weighted F-Measure and weighted ROC area of the model were 73.6%, 73.1%, 73.2% and 85.6% 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 resampling technique with 10 cross validations resampled data registered best classification result of 74.5%. Therefore, among the different PART rule induction models the second model, resampling technique with 10 cross validations, has been chosen due to its better classification accuracy. Therefore, It has been selected for comparing with other classifier results.

Running information of PART algorithm using resampling technique with Resampling Percentage (%) split technique is provided on annex-7

### 4.2.3. Experimenting Naïve Bayes classification algorithm

# In this case also four experiments are conducted 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 70.89 % while only 29.1% instances were classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 69.8%, 69.9%, % and 80.6% respectively. The overall performance of this test has a small difference from what has scored in the previous experiment of J48 decision tree and PART rule induction it exceed by 0.3 from J48 and 0.6 from PART.

Table 4.6: Performance result of Naïve Bayes with default 10-fold cross validation

| Experiment | Model | Accuracy | Time Taken | Weighted TP Rate | Weighted FP Rate | Weighted  Precision | Weighted  Recall | Weighted F-Measure | ROC Area |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Naïve Bayes Pruned with 10 fold cross-validation | 70.9% | 0s | 70.9 | 39.7 | 69.8 | 70.9 | 69.9 | 81.9 |
| 2 | Naïve Bayes Pruned using Resampling with 10 fold cross-validation | 72.0% | 0s | 72.0 | 32.5 | 71.9 | 72.0 | 71.9 | 81.5 |

As shown in the above table 4.6, the second experiment conducted by Naïve Bayes using the resampling technique with 10-fold cross validation scored an accuracy of 72.0 %. This result shows that out of the total training datasets 72 % of records are correctly classified instances and remaining 28% of the records are incorrectly classified instances. Constructed a model with accuracy, weighted precision, weighted F-Measure and weighted ROC area of the model were 72.2%, 71.9%, 71.9% and 81.5% respectively.

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

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 70 % and 30% instances are classified incorrectly. Constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 68.6%, 68.5%, and 80.4% respectively. The outcome of this experiment is presented in table 4.7.

Table 4.7: Performance result of Naïve Bayes algorithm with default percentage split.

| Experiment | Model | Accuracy | Time Taken | Weighted TP Rate | Weighted FP Rate | Weighted  Precision | Weighted  Recall | Weighted F-Measure | ROC Area |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Naïve Bayes pruned Percentage (%) split | 70 % | 0s | 70 | 43 | 68.6 | 70 | 68.5 | 80.4 |
| 2 | Naïve Bayes pruned using Resampling Percentage (%) split | 71.6% | 0s | 71.6 | 36.0 | 70.9 | 71.6 | 71.1 | 81.5 |

As shown in the above table 4.7, the Fourth experiment conducted Naïve Bayes using resampling technique with percentage split scored an accuracy of 71.6 %. This result shows that out of the total training datasets 71.6 % records are correctly classified instances and remaining 28.3% of the records are incorrectly classified instances. constructed a model with accuracy, weighted precision, weighted F-Measure, and weighted ROC area of the model were 71.6, 70.9%, 71.1%, and 81.5% respectively.

Based on the above experiment, Naïve Bayes algorithm using technique Resampling with 10-fold cross-validation. has scored a best result of 72% than Naïve Bayes algorithm using technique with percentage split (66%). Therefore, among the different Naïve Bayes models the second model; with the resampling 10-fold cross-validation has been chosen due to its better classification accuracy.

Running information of Naïve Bayes algorithm using resampling technique 10-fold cross-validation technique is provided on annex-8

### 4.2.4. Experimenting Logistic Regression classification algorithm

In this case also four experiments are conducted using logistic by changing testing modes and by applying resampling techniques. Experiment with Logistic 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 70.4% while only 29.6% instances were classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 70.2%, 70.3%, and 80.5% respectively. The overall performance of this test has lowered from what has scored in the previous experiment of J48 decision tree.

Table 4.8: Performance result of Logistic Regression with default 10-fold cross validation

| Experiment | Model | Accuracy | Time Taken | Weighted TP Rate | Weighted FP Rate | Weighted  Precision | Weighted  Recall | Weighted F-Measure | ROC Area |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Logistic regression Pruned with 10 fold cross-validation | 70.4% | 0.05s | 70.4 | 35.7 | 70.2 | 70.4 | 70.3 | 80.5 |
| 2 | Logistic regression Pruned using Resampling with 10 fold cross-validation | 71.9 | 0.012s | 71.7 | 33.2 | 71.8 | 71.7 | 71.7 | 82.0 |

As shown in the above table 4.8, the second experiment conducted logistic regression using resampling technique with 10-fold cross validation scored an accuracy of 71.9 %. This result shows that out of the total training datasets 71.9% records are correctly classified instances and remaining 28.1% of the records are incorrectly classified instances. constructed a model with accuracy, weighted precision, weighted F-Measure, and weighted ROC area of the model were 71.9%, 71.8%, 71.7% and 82.0 respectively. Based on the above experiment, logistic regression algorithm using resampling technique with 10 cross validations. has scored a better accuracy than logistic regression algorithm with 10 cross validations.

**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 69.7% and 33.3% instances are classified incorrectly. constructed a model with weighted precision, weighted F-Measure and weighted ROC area of the model were 68.3%, 68.4% and 80.6% respectively. The outcome of this experiment is presented in table 4.9.

| Experiment | Model | Accuracy | Time Taken | Weighted TP Rate | Weighted FP Rate | Weighted  Precision | Weighted  Recall | Weighted F-Measure | ROC Area |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Logistic regression pruned Percentage (%) split | 69.7% | 0.05s | 69.7 | 42.5 | 68.3 | 69.7 | 68.4 | 80.6 |
| 2 | Logistic regression pruned using Resampling Percentage (%) split | 70.1% | 0.15s | 70.1 | 34.4 | 70.2 | 70.1 | 70.1 | 82.2 |

Table 4.9: Performance result of Logistic Regression algorithm with default percentage split

As shown in the above table 4.9, the Fourth experiment conducted logistic regression using resampling technique with a percentage split with an accuracy of 70.1 %. This result shows that out of the total training datasets 70.1 % records are correctly classified instances and the remaining 29.9% of the records are incorrectly classified instances. Constructed a model with accuracy, weighted precision, weighted F-Measure, and weighted ROC area of the model were 70.1%, 70.2%, 70.1% and 82.2% respectively.

Based on the above experiment, the logistic regression algorithm using the resampling technique with percentage split (66%) has scored the best result of 70.1% than logistic regression algorithm with percentage split (66%) which is 69.7%. Therefore, among the different logistic regression models the second model, 10-fold cross-validation algorithm using resampling technique, which score (71.9%) has been chosen due to its better classification accuracy. Running information of logistic regression algorithm with percentage split (71.9%) technique is provided on annex-9

## 

## 4.3. COMPARISON OF CLASSIFICATION MODELS

Selecting a better classification technique for building a model, which performs best in predicting student dropout, is one of the aims of this study. For that reason, four classification techniques i.e. Decision Tree, Bayesian, Rule function and Regression, were applied. Four algorithms were selected implement of classification modelling, namely, J48, Naïve Bayes, PART and Logistic. Then, four experiments were conducted for each algorithm, and the obtained results were compared. For each algorithm, the best model with the highest accuracy is selected and presented in table 4.14 below.

Table 4.10: Performance Comparison of the selected models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type of algorithm | Test mode | Accuracy | Recall | Precision | F-measure | ROC |
| J48 decision tree | 10 fold cross validation using resampling Technique | 73.9 % | 73.9% | 76.6% | 74.4% | 83.2% |
| PART rule induction | 10 fold cross validation using resampling Technique | 74.5 % | 74.5% | 76.8% | 75% | 84.4% |
| Naïve Bayes | 10 fold cross validation using resampling Technique | 72% | 72.% | 71.9% | 71.9% | 82.5% |
| Logistic regression | 10 fold cross validation using resampling Technique | 71.9% | 71.9% | 73.3% | 72.3% | 82.7% |

As it is shown in Table 4.14, among the four algorithms PART algorithm using resampling technique 10 fold cross validation performed the highest accuracy of 74.5 %.

As a result, PART rule induction with 10 fold cross validation was selected and using resampling technique as a final model for the study. The confusion matrix of the selected model of PART tree algorithm using resample technique is shown in table 4.15 below.

Table 4.11: Confusion Matrix for PART algorithm using resampling with 10 fold cross validation

|  |  |  |
| --- | --- | --- |
| === Confusion Matrix === | | |
| A | B | Classified as |
| 1311 | 506 | a = no |
| 225 | 829 | b = yes |

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 values 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 4.15 above. Accordingly, it shows that out of 2872 instances, 1311 instances are correctly classified as Active student, 829 instances are correctly classified a dropout.

This classifier incorrectly classified 506 instances as Active student and 225 instances incorrectly classified as dropout, The reason for the misclassification of the classes was if Active status occurs, there is also a possibility that drop status to have occurred

## 4.4. GENERATED RULES FROM RULE PART

In this study from the model developed in the above-mentioned experiments, PART classifier 10 with fold cross validation using resampling Technique have achieved relatively the highest accuracy in most of performance evaluation criteria compared to J48, Naïve Bayes and Logistic Regression Therefore, the model generated by PART classifier with all attributes was selected as the model that can predict student dropout. PART rules generate 38 rules for predicting student dropout; the following nine rules were found interesting rules extracted. Therefore, those rules selected based on 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**: CGPA = satisfactory: yes (86.0)

*This rule indicates if the student current semester GPA bellow 1.5 the student will dropout.*

**Rule 2:** Grade 12 result = good AND Nationality = Ethiopia AND CGPA = good AND

Department = Computer Science: yes (50.0/17.0)

*This rule indicates if the student previous Grade 12 result is between 300and 399 out of 700 and if the department is computer science the student will dropout.*

**Rule 3:**Grade 12 result = good AND Nationality = Ethiopia AND CGPA = low: yes (31.0)

*This rule indicates if the student previous Grade 12 result is between 300and 399 out of 700 and nationality is Ethiopian and semester GPA is less than 1,5 the student will dropout.*

**Rule 4:** Nationality = Ethiopia AND Grade 12 result = good AND Marital Status = Single AND Age = one AND Department = Computer Science AND sex = male: yes (81.0/25.0)

*This rule indicates if the student previous Grade 12 result is between 300and 399 out of 700 and nationality is Ethiopian and semester GPA is less than 1,5 the student will dropout.*

**Rule 5:** Nationality = Ethiopia AND Grade 12 result = good AND Marital Status = Single AND Age = one AND CGPA = good AND Department = Accounting and Finance: yes (193.0/67.0)

*This rule indicates If the student nationality is Ethiopian and previous grade 12 result is good (between 300 to 399) and marital status is single and age between 19 to 29 and semester CGPA is good (between 2 and 2.5) and the department is accounting and finance the student will dropout*

**Rule 6:** Marital Status = Single AND Age = one AND Cgpa = good AND Department = Management: yes (43.0/17.0)

This rule indicate Student marital status is single and age between 19 to 29 and semester GPA is between 2 and 2.5and department is management the student will dropout.

**Rule 7:** Marital Status = Single AND Age = one AND Grade 12 result = good AND sex = male: yes (454.0/210.0)

This rule indicates marital status is single and age is between 19 and 29 and grade 12 result is good (between 300and 399) and sex is male the student will dropout

**Rule 8:** Marital Status = Single AND Age = one AND Grade 12 result = good AND Department = Accounting and Finance: yes (313.0/149.0)

This rule indicates It student marital status is single, and age is between 19 to 29 and grade 12 result is good (between 300 and 399), and the department is accounting and finance the student will dropout

**Rule 9:** Marital Status = Single AND Age = one AND Grade 12 result = good AND Department = Management: yes (71.0/31.0)

This rule indicates It student marital status is single, and age is between 10 to 29 and grade 12 result is good (between 300 and 399), and the department is management the student will dropout

## 4.5. Discussion on Major Findings

From the generated rules it is observed that the most determinant factors are previous study field, Marital status, Nationality, Grade 12 result, CGPA and Age. The first factor identified in this study is Marital Status; Students come to the classroom with a broad range of prior knowledge, background, experience, skills, beliefs, and attitudes, which influence their status throughout they stay in the school. Other factors identified in this study are, sex and department.

Attribute CGPA is also identified as the major factor that determine student drop out. Those who gain CGPA less than 1.5 is going to drop because of dismissal from the university.

The study also show Age and department are another attribute which is a factor to drop out.

The findings observed in the interpreted rules shows that attributes like sex and program have less effect on student performance.

# CHAPTER FIVE

## CONCLUSION AND RECOMMENDATION

### 5.1. CONCLUSION

Student dropout is a universal problem in the academic area. It has both educational and cost implications. St. Mary’s University among others PHEI (private higher educational institution) offer conventional and distances education that is accessible to the large society. Despite this, student dropout also becomes a chronic problem that the university faces daily and losing unexpected cost expenditure every year. The aim of this study is to develop a predictive model using data mining classification techniques to determine undergraduate students’ status (Dropout, Active) in higher education. As indicated in this study, there are various aspects that can have a great potential contribution to student dropout status. A variable considered under this study includes, age, sex, department, Grade 12 result, nationality, program, CGPA, marital Status and dropout (yes or no) were used to build the predictive model. This study was conducted by obtaining the institutional data from St. Mary’s university student record management information system My-SQL database to develop a model that can determine students’ status using data mining technology. In this study, we used one years of data that covers from 2011EC to 2013EC. The SSD data mining process model and the Weka version 3.8.4 data mining tool were employed to undertake the experiment. In this study, 16 experiments have been carried out using four classification algorithms such as decision tree classifier (J48), rule induction (PART), Function (Logistic) and probabilistic classifier (naïve bayes) using 10- fold cross validation and percentage split. and a total of 2872 datasets, 9 attributes were used to build the model. Hence, among the four algorithms tested, the rule induction (PART), classifier algorithm scored the highest accuracy which is 74.5% followed by J48 Decision tree, Naïve Bayes, Logistic regression algorithms. In addition to this, to solve data imbalanced problem (overfitting) and to increase classification accuracy we also applied SMOTE (synthetic minority oversampling technique) and resampling technique. Based on, the extracted hidden pattern using PART algorithm, marital status (single, marred, divorced), Grade 12 result, Age, GPA, and nationality are identified as the major contributing factors of student dropout. Other demographic data such as distance from home, family background, native language, and place of birth location (urban or rural), and health-related data are not included under this study. From the result of accuracy, we understood that there are other factors of student dropout that should be researched by considering other demographic data like listed above. In addition, the data obtained from SRMIS (student record management information system) was in two table formats. So, merging the two tables into one table format was the major challenge of this study.

### 5.2. RECOMMENDATION AND FUTURE WORKS

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

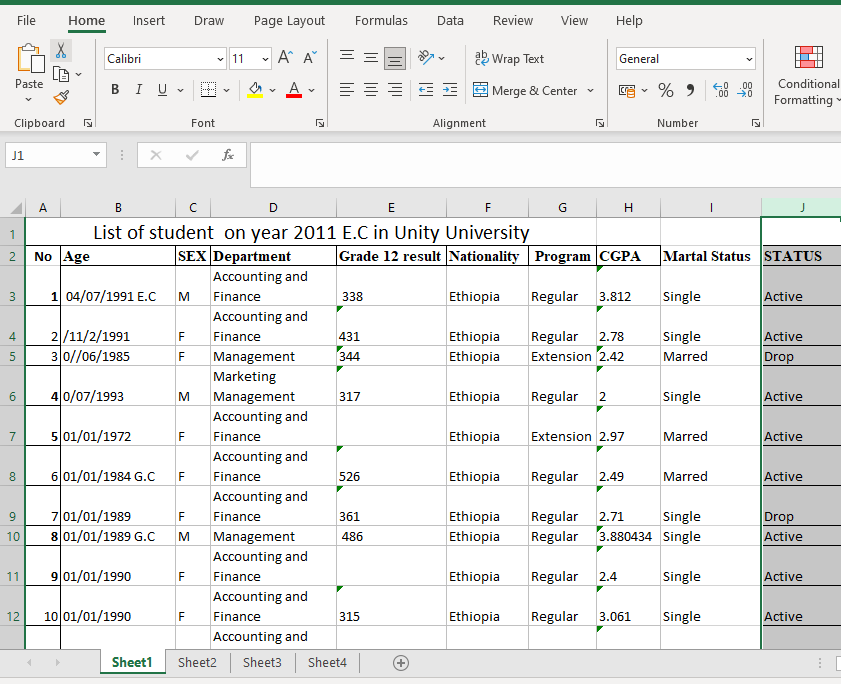
* In this study, we used a decision tree, rule induction, Naïve Bayes and function Logistic data mining classifiers. To improve the performance of the predictive model, further research is needed using other models.
* This research has been carried out to find out the major causes of student dropout at St. Mary’s University. We recommended other researchers to investigate other causes such as family-related problems, health problems, distance related issues from home to school, transferring the program from regular to extension or distance program, university change, late registered, and out of the country. And the Institution should have this information for future use.
* It is difficult to get well organized, correct and quality data for the mining tasks. We suggest educational institutions analyze their data symmetrically for data analyses and have more information about the students so that future researcher will develop a different model that the institution can use it.
* In this study the determination of student status using data mining techniques was only done for St. Mary’s University. Still we proposed other researchers to explore the other private or public university trends of student status, and we also propose private and public universities to organize their data.

# REFERENCE

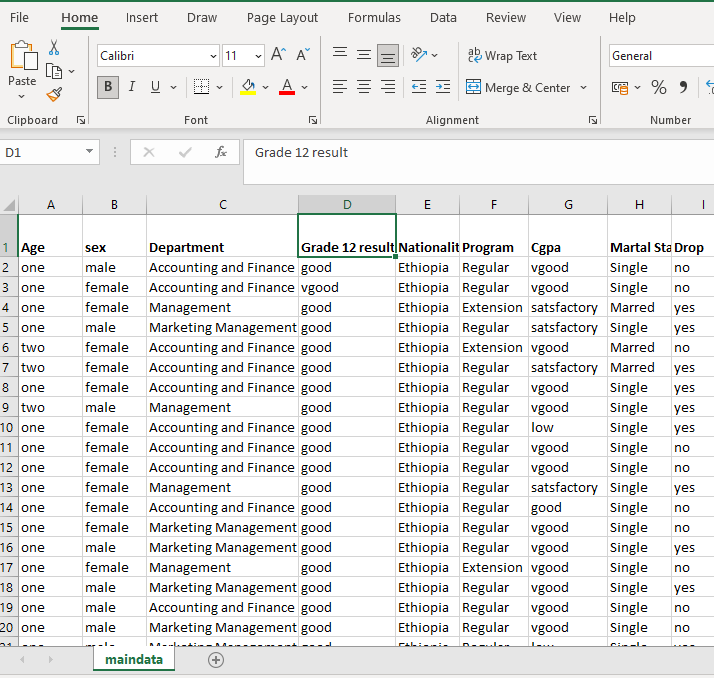
1. Aggrwal, CH.data mining the textbook. Switzerland: springer, p.1,2015
2. Amelec V, etal Integration of Data Technology for Analyzing University Dropout. ScienceDirect 2019; Halifax, Canada
3. Anitha , S.B. Jasmine, B.. Deepalakshmi.S. A Study On School Dropouts In Theni District: A Data Mining Analysis. 2013; Vol. 2. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 [www.ijert.org](http://www.ijert.org/)
4. Anjana P, Smija D, Jubilant J K. Students Dropout Factor Prediction Using EDM Techniques. 2015; International Conference on Soft-Computing and Network Security (ICSNS), Coimbatore, INDIA
5. Azevedo, A., & Santos, M. ”KDD, SEMMA and CRISP-DM: a parallel overview”, IADIS European conference on data mining, pp.184, 2008.
6. Bharati M. Ramageri, Data mining techniques and applications/ Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305
7. Boris P, Camilo C, Dario C. Applying Data Mining Techniques to Predict Student Dropout: A Case Study. 2018: IEEE 978-1-5386-6740-8/18/$31.00 3
8. Carol M, Jeevitha D, Mr. Suman A. A survey on dropout of students using predictive analytics. 2017; pp. 415-41: International Journal of Latest Trends in Engineering and Technology 4
9. Chandrasekar, P. i Qian, K. Shahriar,H. Bhattacharya, P.”Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing”, IEEE 41st Annual Computer Software and Applications Conference,2017
10. Charu C. Aggarwal.Data Mining: The Textbook, Switzerland: Springer International Publishing, 2015.
11. Cios, K.J., Pedrycz, W., Swiniarski, R.W., Kurgan, L.Data Mining A Knowledge Discovery Approach, USA : Springer Science+Business Media, 2007.
12. Dharmawan, T. Ginardi, H. Munif, A. Dropout Detection Using Non-Academic Data. 2018; 4th International Conference Science and Technology (ICST), Yogyakarta, Indonesia
13. Hammond,C. Smink,J and Drew, S. “Dropout Risk Factors And Exampary Programs,” Individual Risk Factors, vol. 19, no. 4, pp. 1– 8, 2008.
14. Han, J.Kamber Mand pei, J data mining concepts and techniques, 3rd ed. Waltham: Elsevier,2012
15. Hidayat, M Purwitasari, D. Ginardi, H. and Informatika, J. T. “analisis prediksi drop out berdasarkan perilaku sosial mahasiswa dalam educational data mining.”
16. Jia P, Malone T. Using predictive modelling to identify students at risk of poor university outcomes. Higher Education. 2015; 70(1):127–49. https://doi.org/10.1007/ s10734-014-9829-7
17. Khanna, L. Singh, D and Aset, D.” educational data mining and its role in determining factors affecting students’ academic performance: A systematic review”, information processing(IICIP), 2016 1st India international conference, 2017.
18. Kidane,A Wolde, L. Makonnen, T. Yusuf Y.and Abdi, O. ”students drop-out in institutions of higher learning in Ethiopia magnitude causes and cures”, the Ethiopian journal of education vol.x, no.2, 1989
19. Krishnaveni, S. and Hemalatha,M "A perspective analysis of traffic accident using data mining techniques," International Journal of Computer Applications, vol. 23, no. 7, 2011.
20. Lin SH. Data mining for student retention management. Journal of Computing Sciences in Colleges. 2012; 27(4): 92–9
21. Macin, M. Center for innovation in research and teaching , [Online]. Available: https://cirt.gcu.edu/research/developmentresources/research\_ready/experimental/overview. [Accessed 3 April 2019].
22. Márquez-Vera C, Morales CR, Soto SV. Predicting school failure and dropout by using data mining techniques. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje 2013; 8(1):7–14.<https://doi.org/10.1109/RITA.2013.2244695> ,7
23. Mayra A, David M. Predicting University Dropout through Data Mining: A Systematic Literature. Indian Journal of Science and Technology. 2019; 12(4), DOI: 10.17485/ijst/2019/v12i4/139729,
24. Mehata, A. and.Buch, N.”depth and breadth of educational data mining: researchers’ point of view”, IEEE access.
25. Natthakan I, Tossapon B. Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. 2015; Springer-Verlag Berlin Heidelberg: DOI 10.1007/s13042-015- 0341-x
26. 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.
27. Semeon, G. “using data mining technique to predict student dropout in St. Mary’s university college: its implication to quality of education”, UNECA Conference Center Addis Abeba, Ethiopia, pp.51- 76.2011
28. Shearer, C. ”the crip-dm model: the new blueprint for data mining”, journal of data warehousing, vol.5, no.4, pp.13-14, 2000
29. Varun K, Anupama C. An Empirical Study of the Applications of Data Mining Techniques in Higher Education. 2011; 2(3): International Journal of Advanced Computer Science and Applications,
30. Vinayak H, Prageeth P. Higher Education Student Dropout Prediction and Analysis through Educational Data Mining.2018; IEEE Xplore Compliant - Part Number: CFP18J06-ART, ISBN:978-1- 5386-0807-4; DVD Part Number: CFP18J06DVD, ISBN:978-1-5386-0806-7
31. Yukseltu,E. Ozekes, S.and Kılıc¸Y. Turel. ¨ Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and E-learning, 17(1):118– 133, 2014
32. <https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/>

# ANNEXES

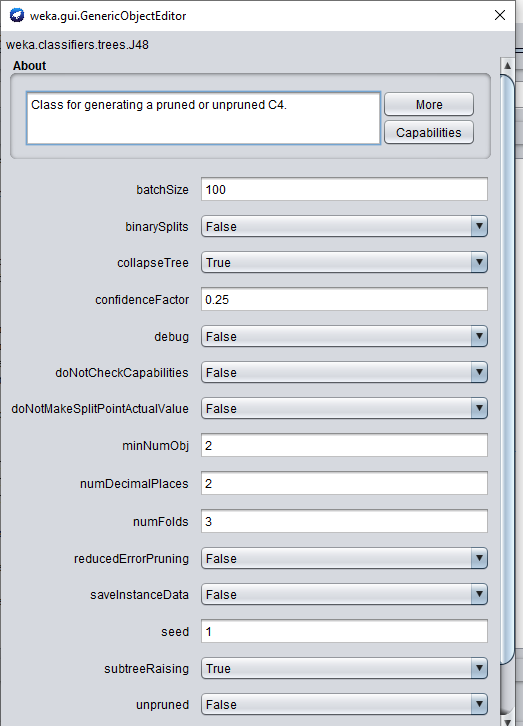
Annex 1: The original sample of collecting data.



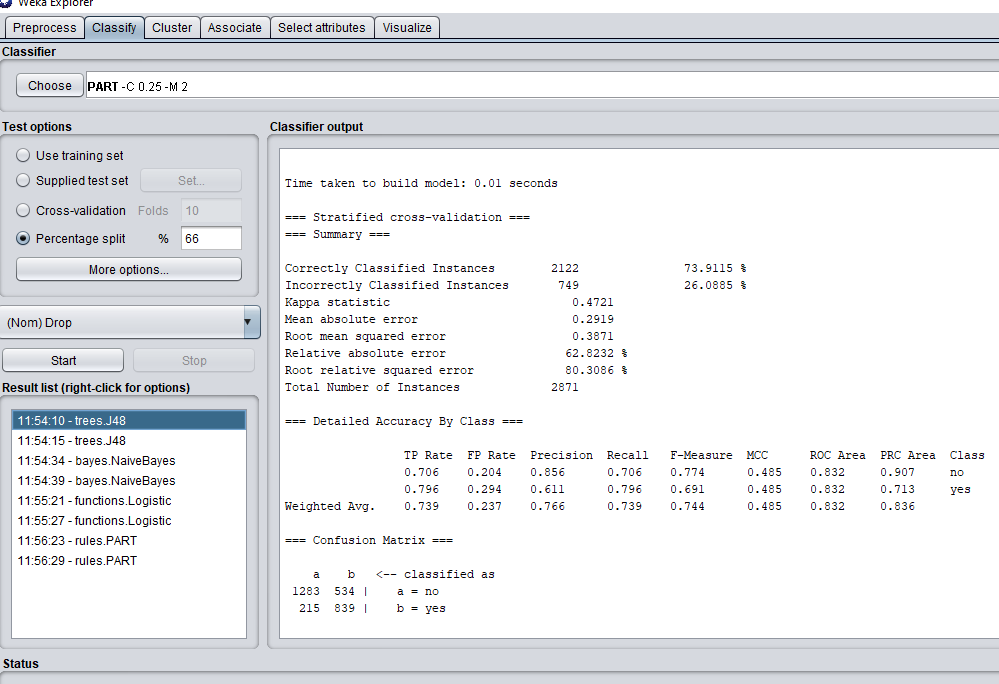
Annex 2: sample statistical summary of data.



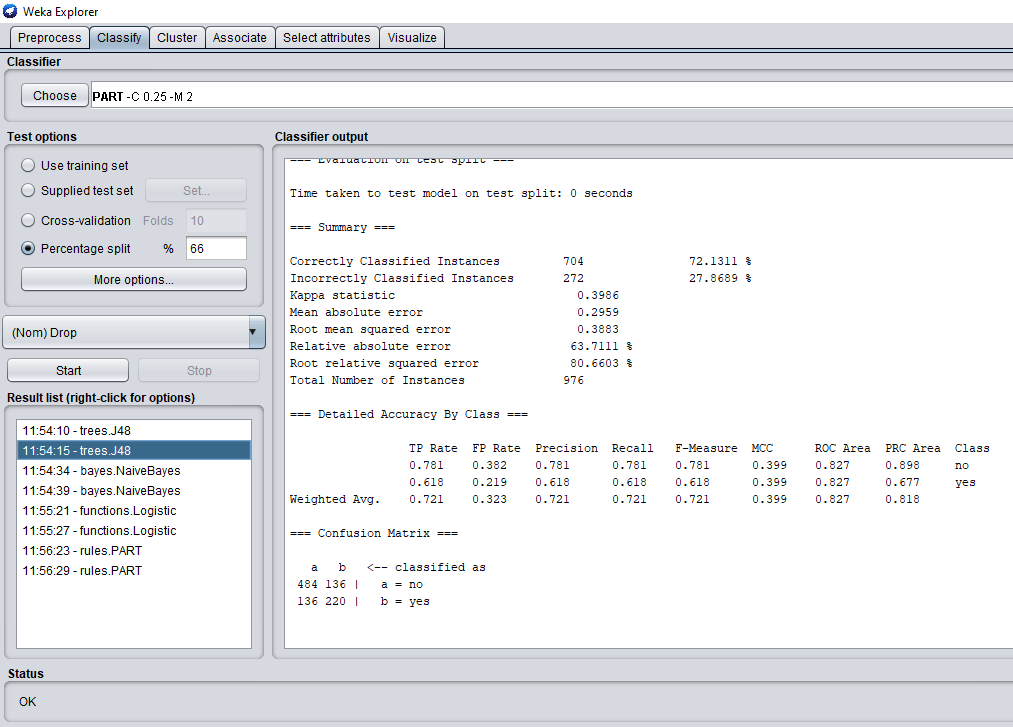
**Annex 3: Parameter settings of the J48 used in conducting the experiments**.

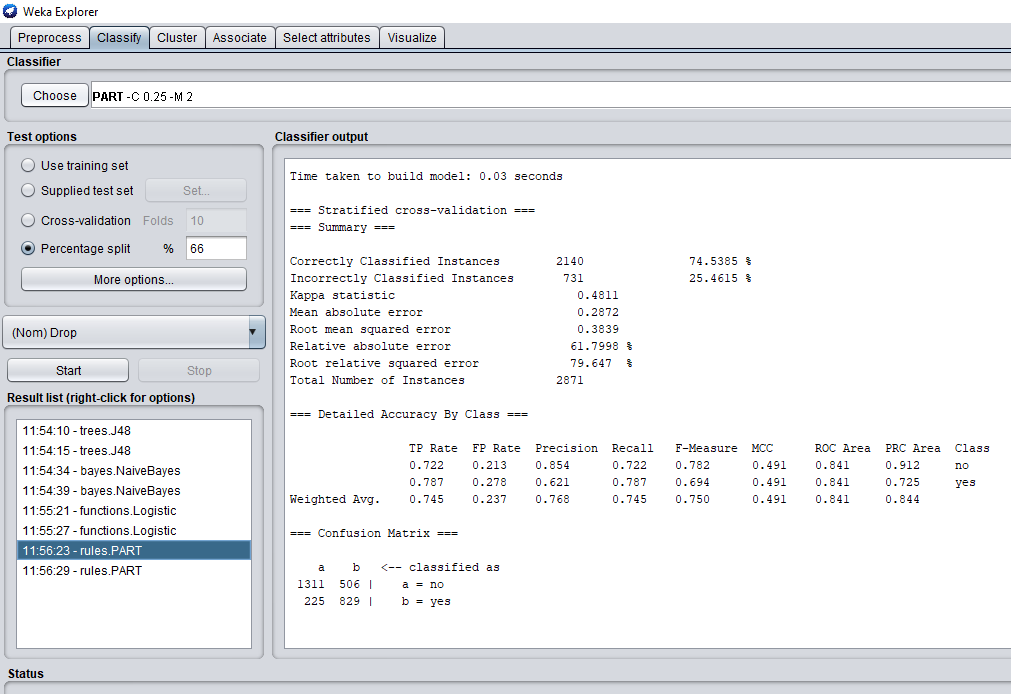


**Annex 4: The snapshot running information of J48 algorithm using resampling technique with 10-fold validation technique.**

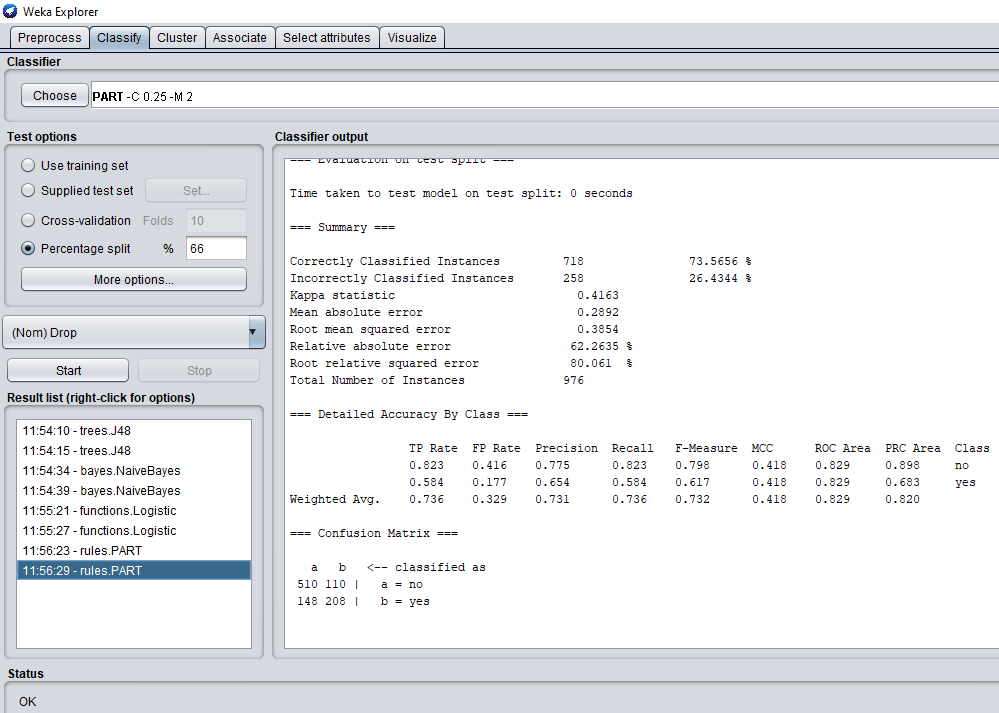


**Annex-5: The snapshot running information of J48 algorithm using resampling technique with percentage split 66% technique.**

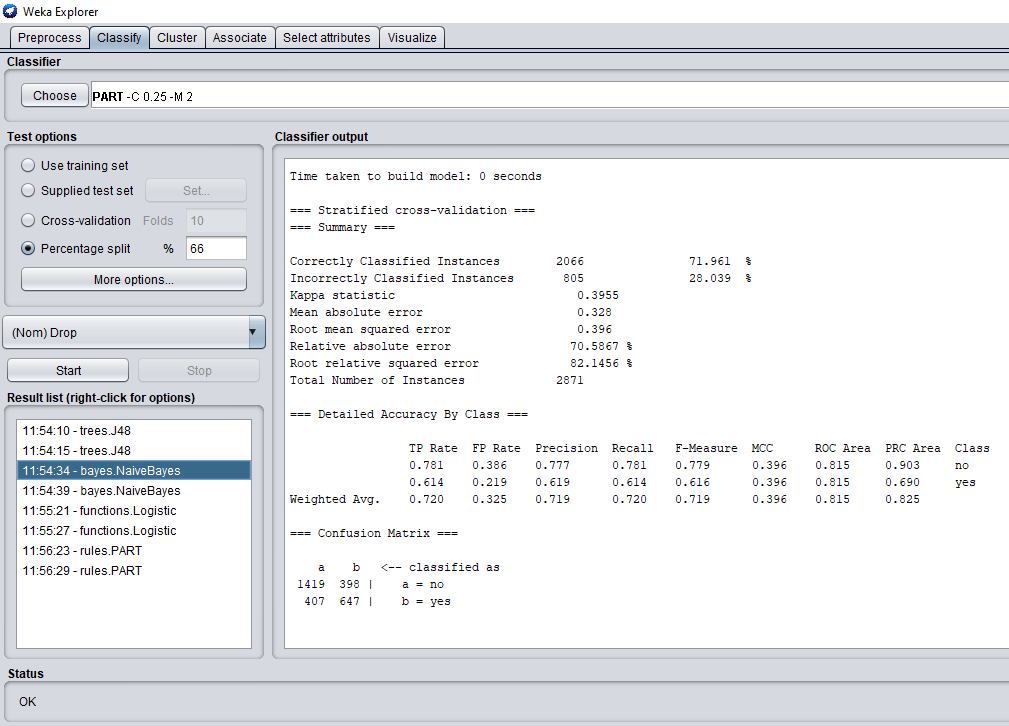


**Annex-6: The snapshot running information of PART algorithm using resampling technique with 10-fold validation technique**

**Annex-7: The snapshot running information of PART algorithm using resampling technique with percentage % split technique.**



**Annex-8: The snapshot running information of Naive Bayes algorithm with 10- fold validation technique**



**Annex-9: The snapshot running information of Logistic Regression technique using resampling technique with percentage split 66% technique.**

# 