Addis Ababa University
College of Natural Science
School of Information Science

Students' Placement Prediction Model:

A Data Mining Approach

By
Meseret Getachew
June 2017

Addis Ababa University
College of Natural Science
School of Information Science

Students' Placement Prediction Model:

A Data mining Approach

A Thesis Submitted to the school of school of Information Science of Addis Ababa University in partial Fulfillment of the Requirements for the Degree of Master of Science in information Science

By
Meseret Getachew

June 2017

Addis Ababa University
College of Natural Science
School of Information Science

Students' Placement Prediction Model:

A Data mining Approach

By
Meseret Getachew
June 2017

Name and signature of members of the Examining Board

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

ACKNOWLEDGMENT

First and foremost extraordinary thanks for my Almighty God and His Mother Saint-marry.

There are many People that need to thank for making this long journey so memorable. First I would like to thank you my research advisor. Dr. Wondwossen Mulugeta for his crucial guidance and encouragement in the conception of this research work, My grate full thanks to my instructors and all staff member of the school of information science for their Contribution in success of my study

I would like to extend thank you Ato Mesele Birhanu Dupty resgistrar and all the staff member of registrar office of Addis Ababa University for Kindly Support my study.

Am grate fully thank You Mr. Tariku Teklu and his office staff of (NEAEA) for kindly support. Last but not least I owe my sincere gratitude to my family specially my W/o Adanech Desalegn and My father Getachew weyecha for moral support and consistent

List of Acronyms

ARFF	Attribute relation file format
CART	Classification and Regression Tree
CHAID	Chi-square Automatic Interaction Detector
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	comma separated values
EDM	Educational Data mining
EHEEE	Ethiopian Higher Education Entrance Certificate Examination
FNR	False Negative Rate
FPR	False Positive Rate
GUI	Graphical User Interface
SPSS	Statistical Package for the Social Science
ID3	Cross Industry Standard Process for Data Mining
JDBC	Java Database Connectivity
KDD	Knowledge Discovery in Data Base
KDP	Knowledge Discovery Process
KNN	K nearest neighbor
NEAEA	National Education Assessment and Examination Agency
TNR	True Negative Rate
TPR	True Positive Rate

Contents

List of Table

List of Figure

ABSTRACT

The main objective of Higher Education institutions is to provide quality education to students. To achieve highest level of quality of education institutions may apply discovery of knowledge for prediction placement of the students', consider resource capability and performance of the students. Thus researcher initiated to undertake study on students' placement into different available departments using data mining technique and to propose a predictive model.

The study attempted to build a predictive model for student placement prediction and identify interesting rules from the generated model by applying data mining techniques. The study been carried out using hybrid data mining methodology processing. The targeted data set used for the study was the students' placement and high school score of students', who joined Addis Ababa University during 2015/16 and 2016/17 Academics years. The data is acquired from AAU registrar office and NEAEA. The original dataset consist 34 attributes and very 11320 instances. Thus, to make the initial data appropriate and manageable for the data mining exercise, data preparation task was undertaken. Decision tree and rule induction classification techniques using J48, REPTree and PART algorithms were applied for the experimentation.

The experiments were conducted using six scenarios for each algorithm and the outputs of the experiments were used for comparison of the models based on the set evaluation criteria. After the test design was defined and the dataset separated into training and test dataset, the model was built the training set and its quality was estimated on the separate test set. As a result, the test run using PERTree algorithm has registered best accuracy which is 82.045%. The generated rules were interpreted, analyzed and the discovered knowledge was evaluated against the existing knowledge base and domain expert's validation. Based on the findings of this research work, we can conclude that improved students' placement in to various departments can be done using data driven predictive model with acceptable.

Chapter One
Introduction

1.1. Background

Today we need more information than we can handle from huge collection of data that get from business transactions and scientific data, the collection can get from satellite, pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Hence we created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of patterns from huge collection of data. [1]

In higher educational institution contains a large number of student records hence a variety of information has increased. Therefore they need find use full patterns or characteristics in this large amount of data that provides professional knowledge about students. They can make their stand in corporate sector. [1]

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. Data mining has been applied in a great number of fields, including retail sales, bioinformatics, and counter-terrorism. And also it has been very effective in many business venues. The key is to find actionable information, or information that can be utilized in a concrete way to improve profitability. Some of the earliest applications were in retailing, especially in the form of market basket analysis. [2]

Data mining in education field is a recent area of research and it earning at higher rates because of its developments and progressive in educational institutes' presently. The amount of data stored in Education system or other field of databases is increasing at a tremendous speed. This gives rise to a need for new techniques and tools to aid humans in automatically and intelligently analyzing huge data sets to gather useful information. This growing need Gives birth to a new research field called Data Mining, which has attracted attention from researchers in many different fields including database design, statistics, pattern recognition, machine learning, and data visualization, optimization and High-performance computing. Data Mining can be used in educational field to enhance the understanding of learning process to focus on finding, extracting and validating variables related to the student learning process. Mining is concerned with

developing method that extract Knowledge from data came educational context is called Educational Data Mining (EDM) [3]

Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. EDM can be considered one of the learning sciences, as well as an area of data mining it is recommend improvements to current educational practice. EDM have different key use which likes student behavior model, mining enrollment data, predicting student performance, and curriculum design. EDM provides enable educational institutions to better allocate resources and staff, proactively manage student outcomes, and improve the effectiveness of alumni development ability to uncover hidden patterns in large databases. [4]

The main applications of EDM one of the learning sciences, and considering key use like Statistics and visualization of data, Predicting Student Performance, Outlier Analysis, Grouping Students, Planning and scheduling and enrollment Management are keys frequently used in higher education to describe well-planned strategies and tactics to shape the enrollment of an institution and meet established goals. [5]

Application Educational data mining can be applied by higher education institutions or universities in determining student Performance Employability rate that managing students' enrollment at the beginning of the year, assist students and effective resource utilization and cost minimization, helping and guiding administrative officers to be successful in management and decision making. [6]

Higher Educational institute perform various activities but one of Activities is students, conducting placement which is arranging or positioning in a specified available fields/departments' according institute design or plan. Placements should be performing a valuable role in enhancing students' discipline-relevant skill, improving confidence and making students increase employable. If placement process with good criteria, the students would have good performance and employment opportunities in leading organizations/Industry. Predicting the placement model of students should to give the above beneficiary to students', so models includes all personal, social, and psychological and other variables are required for the effective prediction of the placement of the student. [7]

The number of students that join higher education in Ethiopia is growing rapidly from time to time. These students choose a field of study of their interest and universities they want to learn after they took the entrance exam, Students are placed based on different criteria that are nationally prepared for all universities. The students' choice departments and give numbers according to their order of preference. The students who has high score in EHEEE exam, females and developing regions are given a priority to be placed based on their first choice up to the department in take capacity, the other students also placed by their second, third choices [8].

1.2 Statement of the problem

An educational institute with efficient Data Warehousing and Data Mining approach can find out novel way of improving student's behavior, success rate and course popularity.[8] All these effort may finally improve the quality of education, use better intake capacity, better career counseling and overall practices of education system.

Higher education institute placed Students by their interest (without influence of external force) to available different field of studies/Department in the consideration resource utilizations of institute. The number students' interested field/Department use competitive basis like academic consideration while placing students to colleges/universities

According to [9] Students usually take their ability into consideration when they choose higher institutions. They do such choices in considerations the students' wants they can successfully complete courses and secure good jobs after graduation. However, students' placement to different departments in Ethiopia higher institutions does not solely depend on the interest (choice) of students. Thus, the placement mechanism may have an impact on their performance.

Students are placed based on different criteria that are nationally prepared for all universities. The students choose the departments and give numbers according to their order of preference. The students 'Join the university after take the national exam result so no more enough knowledge about the order of reference choice available field. The students admit higher institute under various courses from different locations, educational background and with varying academic scores in entrance examinations. Moreover, High school may be different subjects in their capability and also different level of depths in their subjects and analyzing the past performance of admitted students would provide a better perspective of the probable academic performance of students in the future. [10]

The placement of department in different academic programs is one of the decision problems that an administrator in universities regularly deals. The applicants' choices for desired department are not matched with capacities of available department. Higher institute have used one or more criteria along with applicants' order of preferences, and finding an effort to best satisfy applicants' interests, while meeting other academic standards. In essence, the placement problem allocation of limited intake capacities of academic programs or departments among applicants while trying to best satisfy the interest of applicants.

Application of data mining technology is steadily growing fast in Higher education institutions are interested in predicting the paths of students and alumni. Thus identifying which students will join particular course programs and which students will require a large number of debates. This is became educational data mining address how to classify data, find associations between data items and perform time and resource analysis. The numerous data mining techniques to analyze large data quantities data [8]

However, there exist a limited number of research works that attempted to apply data mining techniques in Placement process in higher Educational institute. These research work done by [8] whose main objective has focused on to build a predictive model for placement of students. He reported use classification to get the objective and recommended data mining application. The final output shows that most of the students were placed based on their first choice and only some of them were placed without their first choice. The researcher make a model use as information students' preference order of choice departments, but the students not enough information about the fields because they join the university at first time. In this regard, this research work use the sore of each subject in the national exam as input to determine build a model for the prediction of students' placement.

This study presents model based on classification approach to find an enhanced evaluation method for predicting the placement for students. The study attempts to answer the following question

- What are the most interesting patterns or rules generated regarding the placement process

- What attributes are determinant in deciding students placement?

1.3 Objective of the study

1.3.1 General objective

The general objective of this study is to investigate and build a predictive model for student placement.

1.3.2 Specific objectives

In order to achieve the general objective stated above, the following specific objectives have been formulated.

- To explore determinant attributes students' placement of department for existing dataset
- To investigate, identify and select an appropriate data mining techniques, methods and tools,
- To build a predictive model that support in predicting students' placement in to different departments.
- To determine important and interesting rules or patterns from the generated knowledge by having domain expert opinion,

1.4 Scope and limitation of the research

The main aim of this research is to find out the applicability of DM for determining and predicting department of students, when then they join the university and after they had to select the area of the study and university. The research is delimited one educational institute which is AAU and in a two year intake capacity of the students' choice of Department orders, score of final secondary school results and first choice of department. The students with an assumption database representative students of the university.

Application of educational data mining to the AAU is limited to develop the model instead of the deployment the preference department of the student in to department.

1.5 Significance of the study

Data mining is a powerful tool for academic intervention. Higher education institutions can use classification, for example, for a comprehensive analysis of student characteristics, or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success.

They need to analyze and predicting the preference and academicals History can determine in the future specialty of the students. The information content use good decision making for university resource utilization and predict the performance and skill of the students. Educational Data Mining recent field and it provide hidden knowledge make good pattern the model that utilized use to administration and monitoring educational system.

1.6 Organization of the Thesis

This thesis is organized to six chapters the first chapter introduced general introductory concept and also covered the statement of the problem Research Question, objective of the studies, scope of the study and significance of the study

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

The third chapter methodology of research covers the detail description about data mining methodology.

The fourth chapter's presents data understanding and preprocess is covers to prepare data for experimentation

The fifth chapter presents experimentation on the selected datamining algorithm and results and discussion and Finally the Six chapter presents conclusion of the research and recommendation for further work in the area

Chapter Two

Literature Review

In this chapter an attempt has been to review literature on the concepts and techniques of data mining and its application in various section particularly educational environments.

2.1 Data Mining and Knowledge Discovery

Now we live in information age which is ever-increasing progress of network-distributed computing, particularly rapid expansion of the web, advert Modern computer systems or information technology in various fields. It has led to accumulating data at an almost unimaginable rate and from a very wide variety of data storage in various formats. Which is a broad impact on society in a relatively short period of time .Education is on the brim of a new era based on these change. [12]

The data collected from different applications require proper set of activities, all of which involve extracting meaningful new information or it extracting knowledge from large repositories for better decision making. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data is called knowledge discovery in databases (KDD). [13]

We need to realize that we live in information rich but too often knowledge environment. Which need that the problem of knowledge extraction from large databases involves many steps, ranging from data manipulation and retrieval to fundamental mathematical and statistical inference, search, and reasoning. Researchers and practitioners interested in these problems have been meeting since the first Knowledge Discovery for database (KDD) Workshop in 1989 [12]. To bridge the gap of analyzing large volume of data and extracting valuable information and knowledge for decision making using new computerization technologies, DM and KDD has emerged since recent years.

The area of data mining and knowledge Discovery is inherently associated with database. Which means they are inter changeable each other Data mining and knowledge discovery in databases

are related to each other and to other related fields such as machine learning, statistics, and databases The gap from the stored data to the knowledge that could constructed from the data where the DM came into the picture and KDD is over all process Discovery the pattern from the data. [14]

Different scholars have provided different definitions for DM. Data mining is the core process of knowledge discovery in databases. It is the process identifying valid, nontrivial, novel, potentially useful and ultimately understandable pattern from the large database. In order to analyze large amount of information, the area of Knowledge Discovery in Databases (KDD) provides techniques according the above definition provides the term pattern is an expression in some language describe subset of data Finding structure from the data or generally making high level of description of set of data and the pattern should novel and potentially that some benefit of user and the task [1] [15]

2.2 Data Mining Models

The model helps 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. [16]

In order to conduct DM analysis a general process is usually followed. The goal of design a DM process model is to came up with set of processing steps that be followed by practitioners when execute DM projects. The DM process model should provide a complete description of all the steps from problem specification of all the steps from problem to deployment specification of the result. [16]

The systematical approach for successful approach for data mining vendors and organization have specified process models to guide user though a sequence of steps which lead to good result and formalize the knowledge discovery processes (KDPs) within a common framework, we use concept of a process model. [16]

2.2.1 KDD Process Methodology

The knowledge discovery process (KDP), also called knowledge discovery in databases, seeks new knowledge in some application domain. It consists of many steps it attempting to complete a particular discovery task and each accomplished by the application of a discovery method. The model describes procedures that are performed in each of its steps. It is primarily used to plan, work through, and reduce the cost of given Project. [16]

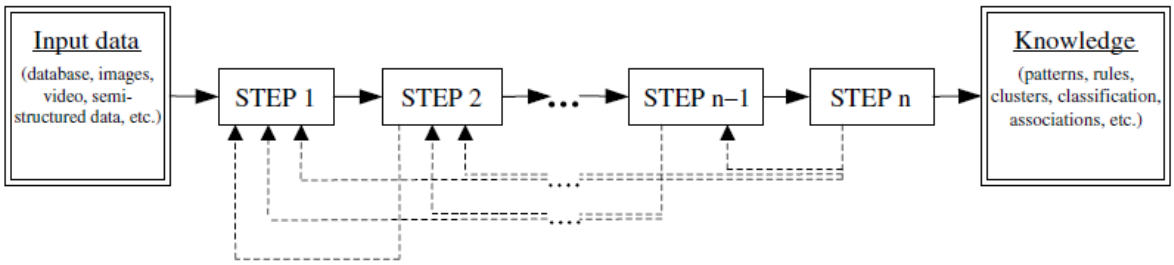

Figure 2.1 Sequential structure of KDP model. [16]

Since 1990s, several different KDPs have been developed. The initial efforts were led by academic research but were quickly followed by industry. The first basic structure of the model was proposed by Fayyad et al. and later improved/modified by others. The process consists of multiple steps that are executed in a sequence. Each subsequent step is initiated upon successful Completion of the previous step, and requires the result generated by the previous step as its input. Another common feature of the proposed models is the range of activities covered, which stretches from the task of understanding the project domain and data, through data preparation and analysis, to evaluation, understanding, and application of the generated results. All the proposed models also emphasize the iterative nature of the model, in terms of many feedback loops that are triggered by a revision process. [16] [12]

Emphasize both the interactive and iterative nature of KDD, that humans make many decisions and that the various steps and methods within them are repeated frequently as knowledge is being refined. Generally there are five step in the KDP Process. [12]

❖ Data selection:

this stage consisting of creating dataset or subset of variable data samples on which on discovery is to be performed and developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the process from the customer's viewpoint.

❖ Data Pre-processing:

This stage data cleaning and preprocessing task in order to obtain consistence data a cleansed source dataset provides better opportunities for successful data mining. The types of operations here include detection and correction of erroneous data (often referred to as data scrubbing), deletion of redundant data, and compensation for missing data. Such operations are usually specific to the data being processed are useful for implementation.

❖ Data Transformation:

This phase covers transformation of selected variable by applying dimension reduction and transformation method different sources into common new format. Constructive data preparation such as production of derived attribute, creating new records value to existing attribute consolidate and summarized fields.in the phase some other activities is done like Discretization it is easily to understand and manageable data analysis for the data mining project

❖ Data Mining:

This is crucial step in which clever techniques are applied to extracting potential useful pattern. I t discover pattern interest in the particular presentation form depending on the DM objectives. In these stage the cleaned and preprocess data sent in to intelligent algorithm for classification, clustering ,reaeration similarity search with in the data and so on here we chose the algorithms that are suitable for discovery pattern in the data. Some of the algorithms provide better accuracy in terms of knowledge discovery than other. Thus selecting the right algorithms can crucial at this point.

❖ Interpretation/evaluation

Interpreting the patterns into knowledge by removing redundant or irrelevant patterns; translating the useful patterns into terms that human understandable.

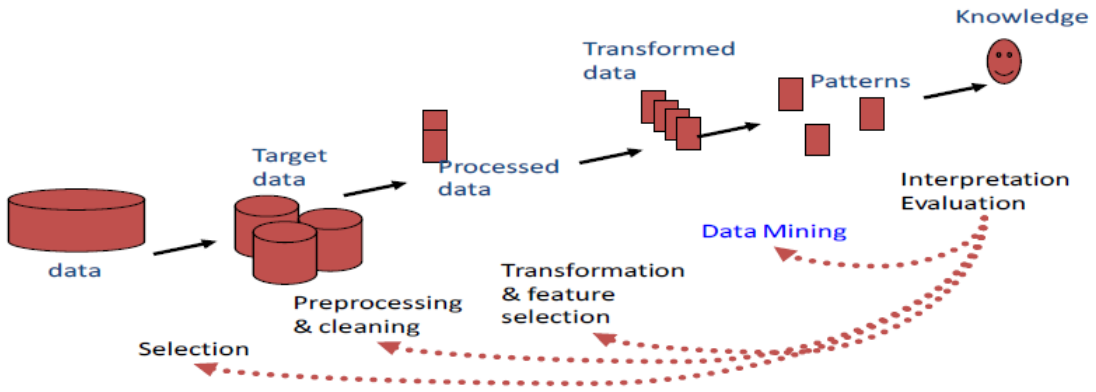

Figure 2.2 KDD process: "From Knowledge Discovery to Data Mining" [12]

2.2.2 The CRISP-DM process model

Cross Industry Standard Process for Data Mining (CRISP-DM) is Comprehensive data mining methodology and process model that provide any one from novice data mining experts with the complete blue print for conducting DM projects .it is standard process model in the industries which consisting of the sequence step that are usually involved data mining in to six phase.

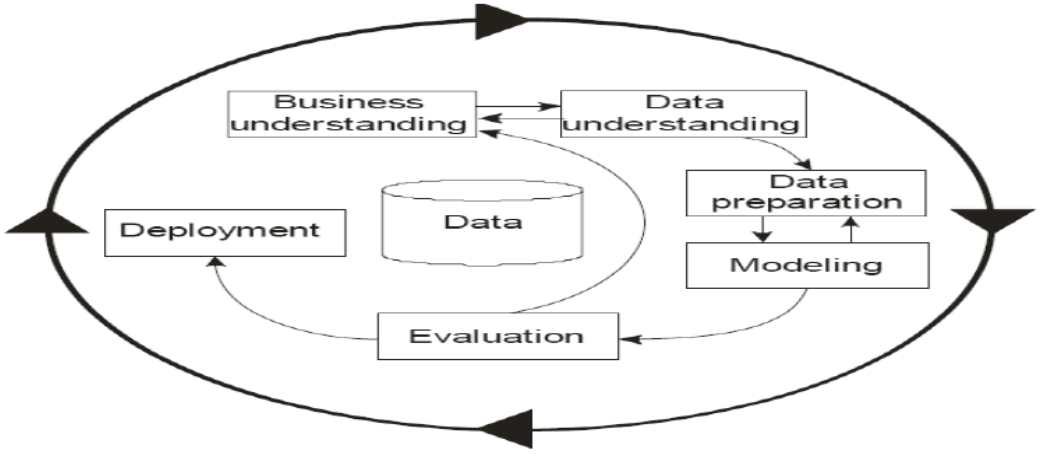

Figure 2.3 CRISP-DM process modeling (source: http://www.crisp-dm.org/).

❖ 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.it is critical of the data mining projects due to the fact that is identified business objective

❖ Data Understanding

The data understanding phase starts with an initial data collection and 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 for hidden information.

❖ 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.

❖ Modeling

In this phase, various modeling 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 on 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 data analysis perspective. Before proceeding to 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**

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. [16]

The above description of each phase of CRISP-DM model showed in as below the table

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

Table 2.1 CRISP –DM phases and tasks [17]

2.2.3 Hybrid Models

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of both. One such model is a six-step KDP model developed by 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 modeling step,

- introducing several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and
- Modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains. [16]

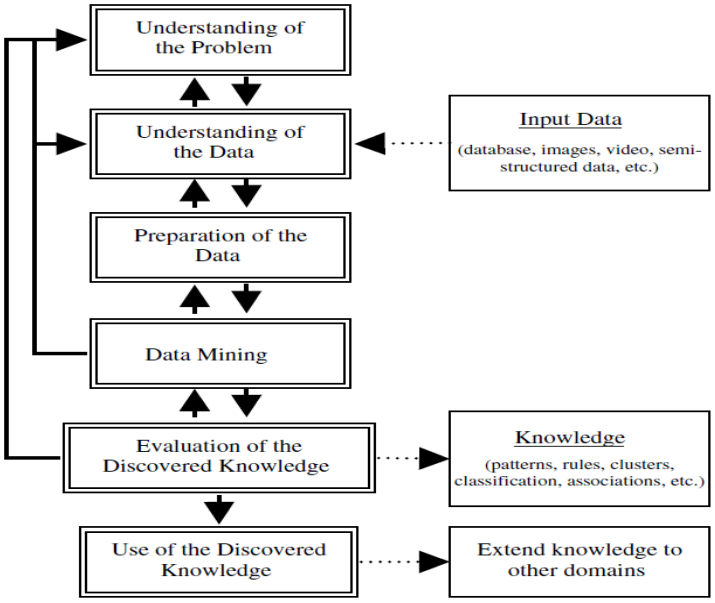

Figure 2.4 the Six-step KDP model. Source [18]

As show the figure above iterative steps for six steps KDP model that followed in hybrid datamining model of the process, drawing from the experience of users of previous models. It identifies and describes several explicit feedback loops: [16]

- Understanding of the data to understanding of the problem domain. This loop is based on the need for additional domain knowledge to better understand the data.
- From preparation of the data to understanding of the data. This loop is caused by the need for additional or more specific information about the data in order to guide the choice of specific data preprocessing algorithms.
- From data mining to understanding of the problem domain, data understanding and preprocess this loop is caused by the need to improve problem domain ,understanding data and preparation, which often results from the specific requirements of the DM method to get the objective of the research.

- Evaluation of the discovered knowledge to data mining. This loop is executed when the discovered knowledge is not novel, interesting, or useful. The least expensive solution is to choose a different DM tool and repeat the DM step.
- Use of the discovered knowledge. This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other. [16]

2.3 Data Mining Tasks

Data mining satisfy its main goal by identifying valid, potentially useful, and easily understandable correlations and patterns present in existing data. Datamining functionalizes in two high-level primary goals of data mining in practice tend to be prediction and description. [12]

A Predictive model makes a prediction about values of data using known results found from different data while the Descriptive model identifies patterns or relationships in data. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties. [19] [2]

In other Approach data-mining task is divided into the two approaches Supervised or directed and in unsupervised or undirected the task for direct datamining uses available data. The goal of supervised or directed data mining is to use the available data as like predictive data-mining to build a model that describes one particular variable of interest in terms of the rest of the available. The user selects the target field and directs the computer to determine estimate, classify or predict its value, In the case indirect data Ming task goal is rather to establish some relationship among all the variables in the data. The user asks the computer to identify patterns in the data that may be significant. Undirected modeling is used to explain those patters and relationships one they have been found [20]
The above Descriptions the datamining task as we can see below the figure

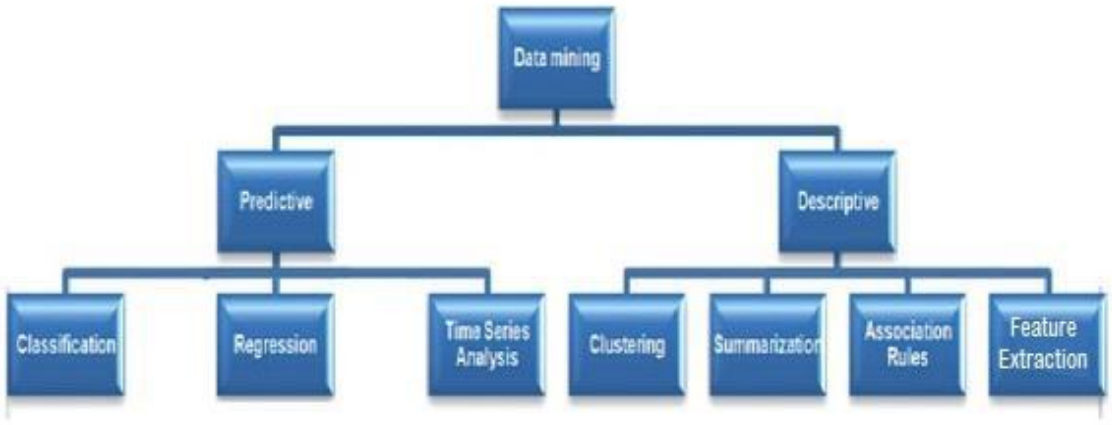

Figure 2.5 Predictive and Descriptive Data mining models [2]

2.3.1 Predictive model

The purpose of Predictive mining model is mainly to predict the future outcome than current behavior. The prediction output can be numeric value or in categorized form. it goals to develop a model which can infer a single aspect of the data (predicted Variable) from some combination of other aspects of the data (predictor variables). [21]

Broadly, there are three types of prediction: classification, regression, and prediction. In Classification, the predicted variable is a binary or categorical variable. In Some popular classification methods include decision trees, logistic regression (for binary predictions), and support vector machines. In regression, the predicted variable is a continuous or numerical. Which like Variable. Neural networks and support vector machine regression.

2.3.2 Classification

Classification is one of the data mining tasks it is composed of two step Supervised learning of training set of data to create a model. Then classify data according to the model or in other way map the data in to predefined classes. It is the process of finding a model which describes and distinguishes data class's or concept for the purpose of being able to use the model to predict the class object whose class label unknown. [19]

Classification is the most commonly applied data mining technique, which employs a set of pre-classified attributes 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 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 sets. The classifier-training algorithm uses these pre-classified attributes to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. [22]

2.3.3 Naïve Bayesian

A Naive Bayesian classifier is a simple method particularly suited when the dimensionality of the inputs is high and classification based on the theory of probability i.e. Bayesian theorem (from Bayesian statistics). It is called naïve because it simplifies problems relying on two important assumptions: it assumes that the prognostic attributes are conditionally independent with familiar classification, and it supposes that there are no hidden attributes that could affect the process of prediction. This classifier represents the promising approach to the probabilistic discovery of knowledge, and it provides a very efficient algorithm for data classification. [23]

Bayesian network model is constructed by explicitly determining all the direct dependencies between the features of the problem domain. And also there has been much interest in learning Bayesian networks from data. Learning such models is desirable simply because there is a wide array of off-the-shelf tools that can apply he learned models as expert systems, diagnosis engines, and decision support systems. Learners also claim that adaptive Bayesian networks have advantages in their own right as a non-parametric method for density estimation, data analysis, pattern classification, and modeling. [24]

2.3.4 Neural Networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the Learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. [25]

2.3.5 Decision Trees

Decision tree are a popular and tree-shape structure that represent the set decision using top-down approach. To classification that divides the data into leaf and node divisions until the entire set has been analyzed. The set of Decision where each internal node denoted a test on an attribute each branch represent an out comes of the test and the leaf nodes represent classes or class distributed

Decision tree is way representing a series rule that lead to a class or value .in simple Decision tree that solves the problem while illustrating all the basic component of the decision tree the first one root node it is the top most node in tree which specifies a test to be carried out the result of this cause the tree to split in to branches. The second component leaves node which lead more than one branches it depends the algorithms. By navigating the decision tree you can assign the value or class to a case by deciding which branch to take starting the root node and moving to each subsequent node until a leaf node is reached .Each node uses the data from the case to choose the appropriate. [26]

The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle multidimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. [19]

A decision tree is made up of a hierarchical structure of decision nodes which are linked with branches. To determine the root node and after that the next nodes, for each attribute we calculated which one will most exactly classify objects according to the decision variable values. From each decision node, branches going out to each node correspond to the various possible answers to the test. For a continuous variable, the test will be in this form: ‘if the value on the variable is higher than a threshold value then take the first connect, otherwise take the second’. For the categorical variables, one of the outgoing branches of the node is associated with each category of the variable. The final nodes are called leaf nodes and are linked during the training phase to one of the categories of the decision variable. To realize a prediction on a new

individual, it is enough to make him traverse the graph to a leaf. The leaf which the new individual reaches will determine the predicted value of the decision variable. [11]

Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. A number of different algorithms may be used for building decision trees including ID3 algorithm it is primarily used for decision making.ID3 (Iterative Dichotomies 3) algorithm invented by Ross Quinlan is used to generate a decision tree makes use a fixed set of decision tree. In order to depict the Dependency of various attributes the resulting tree is used. To decide which attribute goes into the decision node ID3, C.4.5 Decision tree algorithms Developed professor Ross Quinlan in 1993 it has additional features such as handling missing values, categorization of continuous attributes, pruning of decision trees, rule derivation, and others. Basic construction of C4.5, CART stands for Classification and Regression Trees. It is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. The splits are selected using the towing criteria and the obtained tree is pruned by cost complexity Pruning. When provided, CART can consider Misclassification costs in the tree induction. It also enables users to provide prior probability distribution and other one). CHAID can be used for prediction as well as classification, and for detection of interaction between variables. CHAID is often used in the context of direct marketing to select groups of consumers and predict how their responses to some variables affect other variables, although other early applications were in the field of medical and psychiatric research. [27]

2.3.6 Regression

Regression is predictive modeling and analysis is used to make predictions based on existing data by applying formulas. It is a statistical method of data mining. Regression actually used to model between the one or more dependent variables and independent variables. It can be used for building model or classifier which can analyses the historical data to predict the future trends using linear or logistic regression techniques from statistics, a function is learned from the existing data. The new data is then mapped to the function in order to make predictions it is uses existing values to forecast what other values will be. [26]

2.3.7 Prediction

In prediction the goal is to develop a model which can infer Single aspect of data with combination of some aspect of the data [28] like Similar to classification. The difference is that in prediction, the class is not a qualitative discrete attribute but a continuous one. The goal of prediction is to find the numerical value of the target attribute for unseen objects; this problem type is also known as regression, and if the prediction deals with time series data, then it is often called forecasting. Regression analysis, decision trees, and neural nets generally apply [29]

2.3.8 Time-series Analysis

The time-series analysis is a prediction application with one or more time-dependent attributes that usually involves prediction of numeric outcomes such as the future price of individual stock. This time-series represents a collection of values obtained from sequential measurements. Time-series data mining gives natural ability to visualize the shape of data. Time series are very long, considered smooth, as subsequent values are within predictable ranges of one another. Time-series are popular in many applications, such as stock market analysis, economic and sales forecasting. It can be useful for observation of natural phenomena like atmosphere, temperature, wind, earthquake, scientific and engineering experiments, and medical treatments [29]

2.4 Descriptive Modeling

The Descriptive model is the data mining model mostly identifies patterns or relationships in data Sets. It serves as easy way to explore the properties of the data examined earlier and not to predict new properties. It focuses on finding human interpretable patterns describing the data.

The Model are used for Unsupervised or undirected data mining however variable is singled out as the target as like the descriptive mining technique. The goal is rather to establish some relationship among all the variables in the data. The user asks the computer to identify patterns in the data that may be significant. Undirected modeling is used to explain those patters and relationships one they have been found. [2]

The Descriptive model encompasses task to perform as Clustering, Association Rules, Summarizations, and Sequence

2.4.1 Clustering

Clustering is a well-studied and well-known data analysis technique in statistics. Various definitions about clustering method of descriptive modeling exist in literature Clustering is the process of grouping a set of data objects into multiple groups or clusters. So that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures. [19]

In similar way Clustering is the most important descriptive data mining technique, in which a set of data items is partitioned into a set of classes also called as groups. Clustering is the process of finding natural groups, called as clusters, in a database. The set of data items in a group have similar characteristics. Hence, it is mainly used for finding groups of similar items. It is the identification process for classes for a set of objects whose classes are not known. These objects ware so clustered that their intra class similarities can be maximized and minimized on the criteria defined by attributes of objects. The object is label with their corresponding clusters once their particular class or clusters are decided. [2]

Clustering is identify a finite set of categories or clusters to describe the data the categories can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or overlapping categories. Examples of clustering applications in a knowledge discovery context include discovering homogeneous subpopulations for consumers in marketing databases and identifying subcategories of spectra from infrared sky measurements [12]

2.4.2 Association

The Associations or Link Analysis technique are used to discover relationships between variable and object. In these techniques, the presence of one pattern implies the presence of another pattern i.e. item is related to another in terms of cause-and-effect. This is common in establishing a form of statistical relationships among different interdependent variables of a model. For market basket analysis, Commodity management advertising Association Rules is popular and well researched method for discovering interesting relations between the variable in large data

base it is intended a technique that all possible combinations of potentially interesting product groupings can be explored and to identify strung rule discovered in database. They also play an important role in shopping basket data analysis, product clustering, and catalog design and store layout. This association rules build by programmers and use to build capable of machine learning. [2]

2.4.3 Summarization

Summarization is the generalization or abstraction of data .Which is technique maps data into subsets with simple descriptions. The summarized d ate set gives general overview of the data with aggregated information. Summarization can scale up to different levels of abstraction and can be viewed from different angles. It is a key data-mining concept involving techniques for finding a compact description of dataset. Summarization approaches are Basically Mean, Standard Deviation, Variance, tabulating, mode, and median. These approaches are often applied for interactive exploratory data analysis, data visualization and automated and automated report generation. [2]

2.4.4 Sequential pattern mining

Sequential pattern mining is the machine learning task that addresses the problem of discovering the existing frequent sequence in a given database. This mining task is able to discovering the structured behaviors specifically sequential patterns .this patterns exist when the data which is already to be mined has some sequential nature.

2.5 Application of Data Mining

Data mining application is increasingly popular because with growth of data increase in every application, data mining meets the imminent need for effective, scalable, and flexible data analysis in our society. Data mining can be considered as a natural evolution of information technology and a confluence of several related disciplines and application domains.

Data mining applicable in many field is called interdisciplinary field Because of the diversity of disciplines contributing to datamining, datamining research is expected to generate a large variety of data mining systems. Therefore, it is necessary to provide a clear classification of data mining systems, which may help potential users distinguish between such systems and identify those that best match their needs.

According to data mining system, it is applicable to like area of Commercial business, customer analytics, Agriculture, banking, engineering, Security, medicine, Telecommunication, insurance and etc. use of data mining techniques to get discovery of the knowledge or pattern for different field. The main concerned area is about data mining applications in educational systems [26]

2.5.1 Educational Data mining

The application of data mining in education field known as to Educational data mining (EDM) is an emerging interdisciplinary researches and it defined as the area of scientific Inquiry. It concerned with development of methods for exploring the unique type of the data that came from educational environments. Its goal is to better understand how students learn and identify the settings in which they learn to improve educational outcomes and to gain insights into and explain researcher on the area of EDM Techniques what impact the techniques reported on will have on students what the performance expected under the real world model training Constraint [30] [31]

EDM develops and adapts statistical, machine-learning and data-mining methods to study educational data generated basically by students and instructors concerned with developing, researching, and applying computerized methods to explore unique type of the data in large collections of educational system and the tools this is done by developing method to support students and educational setting they use basic reason for the rapidly growth of EDM research is The increase of e-learning resources, instrumental educational software, the use of the Internet in education, and the establishment of state databases of student information has created large repositories of data it necessary [32]

EDM process converts raw educational data into useful information that could be used to improve the quality of administrative decisions of educational institutes. The EDM process is not very different from other application areas of data mining like business, genetics, medicine, etc. Because it uses the same steps as the general data mining process like pre-processing, data mining and post-processing [33]

EDM is an application of data mining, which offer necessary information to educational organization, which hides in educational data of students. It provides various methods for

prediction of student`s performance. Which improve the future result of students which is Prediction of student performance with high accuracy. It is useful in many contexts in all educational institutions for Identifying slow learners and distinguishing students with low academic achievement or weak students who are likely to have low academic achievements. The end product of models would be beneficial to the teachers, Parents and educational planners not only for informing the students during their study, whether their current Behavior could be associated with positive and negative outcomes of the past, but also for providing advice to Rectify problems. As the end products of the models would be presented regularly to students in a comprehensive form, these end products would facilitate reflection and self-regulation during their study. [34]

There are different type of the data used in EDM researcher due to the data is collected form Historical and operational data of the student's academicals and personal data reside in the database of educational institute. It used many techniques such as decision trees, neural networks, k-nearest Neighbor, Naïve Bayes, support vector machines and many others. Using these methods many kinds of knowledge can be discovered such as association rules, classifications and clustering. The is covered knowledge can be used to better understand students' behavior, to assist instructors, to improve teaching, to evaluate and improve e-learning systems, to improve curriculums and many other benefits [35]

2.6 Related research work

Currently, there are some researches that were applied to investigate the application of data mining tools and techniques on Educational data mining and related issues. The following works are reviewed by the researcher to help clarify the significance of the research problem and to show the gap in the current local and international studies.

Using data mining techniques improve efficiency of higher educational institute. Where techniques such as Clustering, decision tree and association are applied to higher education processes, these are useful in student improve students' performance, their life cycle management, selection of courses, to measure their retention rate and the grant fund management of an institution. Therefore, standardization of input data and output model are needed. Data mining techniques are useful in student marketing, Selection revenue analysis, predicting student

performance, planning of courses and result analysis. It is has a wide array of applications for the higher education it has a wide array of applications for the higher education sector. [5]

Prediction of student's academic performance at higher secondary level examination remains a difficult task Examinations serve a unique position as a measure of assessing the academic performance of a student. Measuring of academic performance of students is challenging tasks since student performance is a product of the factor namely demographic, academic-environment socio-economic factors. The analysis on the data obtained from the students of the higher secondary schools containing 35 attributes with 5650 data that proposing an efficient prediction models that use data mining techniques use Bayesian network BN model provides a robust mechanism to predict the academic performance of higher secondary students with better predictive accuracy [24]

Educational Data Mining is concerned with developing new methods to discover knowledge from educational database and can used for decision making in educational system. Researcher collected the student's data that have different information about their previous and current academics records and presents a model based on classification approach to find an enhanced evaluation method for predicting the placement for students data mining techniques using Decision tree and Naïve Bayes classifier to interpret potential and useful knowledge and make model can determine the relations between academic achievement of students and their placement in campus. The Researcher suggest that placement of student he three selected classification algorithms based on Weka. The best algorithm based on the placement data Naïve Bayes classifier has the potential to significantly improve the conventional classification methods for use in placement [22]

The knowledge is hidden among the educational data set and it is extractable through data mining techniques. Applicable for discovering knowledge for prediction regarding enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students' performance and so on. The researcher is designed to justify the capabilities of data mining techniques in context of higher education by offering a data mining model for higher education system in the university. In this research, the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here. Information's like Attendance, Class

test, Seminar and Assignment marks were collected from the student's previous database, to predict the performance at the end of the semester. This study will help to the students and the teachers to improve the division of the student. This study will also work to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination. [25]

EDM offers Academia attributes of students such as internal marks, lab marks, age or such information to educational organization from educational data. EDM provides such attribute and various methods for prediction of student`s performance about placement of students which improve the future result of students. The researcher use method for sum different method which used to find pattern from the given dataset with statistical analysis the model efficiently predict the placement of the students. EDM has been used for predicting the performance about placement of final year students [34]

Predicting students' performance has been an issue studied in educational data mining research in the context of success or failure of a student in a course or program is a problem that focus on two problems prediction of approval/ failure and prediction of grade that the researcher address the problem using different data mining Techniques. The classification task while regression task. Separate models are trained for each course. The experiments were carried out using administrate data from the University of Porto, concerning approximately 700 courses. The algorithms with best results overall in classification were decision trees and SVM while in regression they were SVM, Random Forest, and AdaBoost.R2. However, in the classification setting, the algorithms are however, in the classification setting, the algorithms are finding useful patterns, while, in regression, the models obtained are not able to beat a simple baseline [36].

In academic institutions like universities and colleges the students' placement in to different departments is one of the activity that data mining can be applied to predict the departments. Which is the students will be placed based on the order of their preference. In this study, we collected the student's data that have different information about their entrance exam result and then apply different classification algorithm using Data Mining tools (WEKA). The study used three algorithms J48, Naïve Bayes and Random Forest to build a prediction model for placement of students. The analysis result shows that Random Forest algorithms has performed well and can be used as a best predicting model algorithm than the other two algorithms. Results from the

study have shown that the problem of predicting departments of students in the University could be supported by the use of data mining technique. [8]

Educational institute seeks more efficient technology to better manage and support decision making procedures or assist them to set new strategies and plan for a better management of the current processes One way to effectively address the challenges for improving the quality is to provide new knowledge related to the educational processes and entities to the managerial system using the techniques of data mining technology. Data mining techniques are analytical tools that can be used to extract meaningful knowledge from large data sets. Various data mining Techniques can be applied to the set of educational data and what new explicit knowledge or models are discovered. The models are classified based on the type of techniques used, including predictive and descriptive. The capabilities of data mining in the Context of higher educational system by enhance their current decision processes, and applying data mining techniques to discover new explicit knowledge which could be useful for the decision making processes [37]

Education data mining is improve educational system one of is improvement of activities placement. Placement is a very important issue for any educational organization. The researcher actually deals with the application of data mining to the educational data to improve placement. It takes education data i.e student data as input and predict the status of placement for the student. After the collection of desired data researcher use WEKA data mining tool for the development of prediction model using random tree algorithm ID3, Bayes Net, J48, network, Random tree So Random tree algorithm is best suited for prediction of placement status for a student. Therefore by using Random tree algorithm we generate a decision tree which can be used for the building the model for predicting placement of students accurately. [38]

Data mining in educational environment is a powerful tool for academic intervention. Tree Classifiers have main role to developing new methods to discover knowledge from educational database and can used for decision making in educational system. The tree classifier are J48, NBTree, REPTree, Decision Stump, Random Forest and Random Tree algorithm using Data Mining tools (WEKA) for analysis the student's academics performances. The researcher deals with a comparative study of various Tree Classifiers from these he suggest that the higher management's executives for training & placement department of engineering colleges or Company Executives can use such classification model to measures or visualized the students' performance according to the extracted knowledge. [39]

The main objectives of higher education institutions are to provide quality education to its students for their better placement opportunity. The researcher use Decision tree algorithms to predict student selection in placement was implemented by using Rapid Miner tool. From the results obtained above it is proved that Decision tree algorithms are applied on student's data for analyzing the student's placement selection. The efficiency of different decision tree algorithms was analyzed based on the accuracy ID3 algorithm is appropriate for predicting student placement. ID3 than other decision tree [40]

Based on the data mining techniques the most important predictors for student success were the students' high school GPA and gender the researcher to explore the socio-demographic variables like gender, region lived and studied, nationality and high school degree that may influence success of students. Use data mining techniques classification in algorithms of decision tree We examine to what extent these factors help us to predict students' academic achievement although all the variables except the region individually significant predictors as described in attribute selection trees displayed only two variables Gender and secondary school significant impact on students success [41]

Educational data mining develop student model use their academic record, using a decision tree algorithm. The preprocessing, processing and experimenting was conducted using Rapid Miner tool. During processing among a total of 49 various attributes which will help to improve the student's academic performance 27 important rules were generated. From the generated model specific courses, sex, academic status in 1st and 2nd year of the students determines the performance of student. Finally, the decision tree algorithm was tested and it provides a promising result of accuracy of useful data mining can be used in higher education particularly to predict students' performance using their academic record. [42]

Data Mining Algorithm use Fuzzy logic and K nearest neighbor (KNN) predict and analyses whether a student (he/she) will be recruited or not in the campus placement. The campus placement activity is very much important as institution point of view as well as student point of view. The algorithms are applied on the data set and attributes used to build the model. The accuracy obtained after analysis for KNN is 97.33% and for the Fuzzy logic is 92.67%. Hence,

from the above said analysis and prediction it would be better if the KNN is used to predict the placement results [43]

Measuring student's performance including academic performance, technical skills, soft skills, training and projects are considered to capture the desirability and ability of a student for placement the researched Identification of range of students according to their academic performance via the data mining technique this study uses simple K-Mean clustering and j-48 classification algorithm for student's segregation. This system will validate the Results and classify that which student should come under which cluster. [7]

Nowadays Educational Data Mining Approaches such as Predicting student performance, Analysis and visualization of data, providing feedback for supporting instructors, Recommendations for students, Social network analysis and so comparative analysis of classification algorithms (such as Naïve bayes, MLP ,SMO ,Decision Table, REP tree, J48) using WEKA tool, it is proven that the attributes chosen from the original dataset have high influence using J48 with an accuracy of 97% under comparative study of data mining classification algorithms and students to improve their academic results and placements. [44]

Chapter Three

Methodology and Techniques

In data mining many methodologies have been used according to systematically conduct data mining analysis which is related objective of data mining project. Use datamining process model (KDP) process should have the in-depth and complete knowledge of the problem. This could decide the technique and algorithm to be applied on the dataset. [1]

The KDP design consists of the various strategies see in the literature review and used to provide a good coverage starting with from problem domain, data preprocessing, analysis, inferences drawn, and implementation.

According to section 2.2.3 use Hybrid data mining modeling more research oriented and it's developed by Cios et al and based on the CRISP-DM model by adopting[33].Hence the researcher choose Hybrid model and it provides

- It more general research-oriented,
- The description of the steps use datamining steps
- The model combines aspects of both which is industrial and academicals model.

The model steps are understanding of the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and finally use of the discovered knowledge.

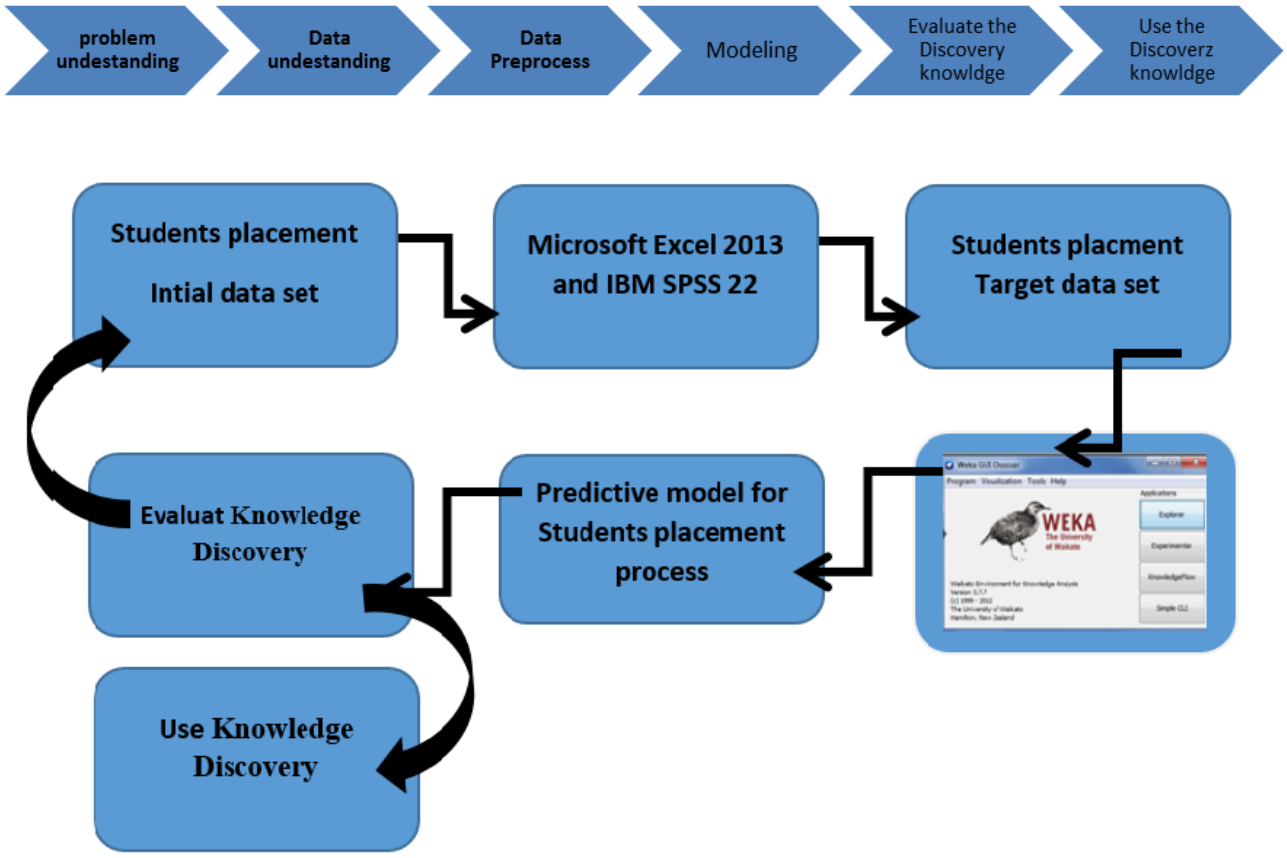

Figure 3.1: Diagrammatic Overall Research Design

Overall purpose of this research work is to explore the predictive model for placement of students in different departments. Which use the standard logical steps in Hybrid Modeling of data mining methodology has been followed in order to come up with reliable and efficient predictive model use in educational dataset?

In this research data mining Tools is not the only tools that we use require instead review of literature (conceptual and theoretical) be done. Moreover, we need to understand the system that contain the initial data and other Intermediate data processing tools and technique Such as MS Excel and SQL transfer and Pre-process the data for actual application of data mining task using Hybrid data mining model. The researcher tries to discuss each step in details below

3.1.1 Understanding Problem Domain

This initial phase to involve working closely with domain expert to identify, and formulate the problem domain with respect datamining project. The requirements a converting this knowledge into a data mining problem definition and designing a preliminary plan to achieve the data mining project.

In this study, a thorough discussion was made with domain experts to understand the problem domain and to have a good understanding about the initial datasets. The domain experts are those placement expert, data analysts, computer programmer and statisticians who processed the placement process to provide statistical information for information of placement students. In addition to discussion made with domain experts, relevant documents were deeply reviewed to gain insight about the nature of the data items and what is contained in the dataset.

3.1.2 Data Understanding

Once the problem to be addressed is understood, the next step is analyzing understanding Data. Because, the end result of knowledge discovery process heavily depends on the quality of available data. The original data collected for this research should be described briefly. Its description includes listing out attributes with their respective values, missing and outlier and evaluation of their importance to the research goal. Additionally, careful analysis of the data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the DM tasks to be performed

3.1.3 Preparation of the Data

This is the most crucial phases in which the success of the entire knowledge discovery process Depends. It usually consumes much of the research effort. In this phase, the final dataset is also constructed from the initial raw data. Data preparation tasks were performed in iterative Ways. The major tasks include: description of data sources, carrying out statistical summary measure, Finding out distinct value, filling missing values, outlier and noisy data and data transformation/reduction activities were also undertaken in this phase. Additionally, feature /attribute construction was also made too.

Data cleaning routines was applied to fill in missing values, smooth out noise, and detect outliers in the data. To support the preprocessing tasks, the researcher used WEKA 3.6.13 to normalize and to fill missing value with mode value. It also helps to train the selected algorithms with training sample and testing case.

3.1.4 Modeling

One of the major tasks of data mining research is modeling. Here the data miner uses various DM methods and techniques to derive knowledge from preprocessed data set

Since, we were focused on building models which can predict Department, we more emphasizes on application of predictive modeling methods/techniques. There are many predictive model in DM methods one of this are classification. It is most powerful predictive modeling techniques. In classification DM analysis Decision Tree is one and the other is the Rule Induction algorithms Therefore, after exploring the underlined tasks with the nature of the research problem, the researcher selected two data mining techniques that are decision tree and rule induction in order to be used and to accomplish this research work.

3.1.5 Evaluation of the Discovered Knowledge

Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. At this stage, before proceeding to use the final discovered knowledge, it is important to evaluate the model and thoroughly review the steps executed to construct the model that properly achieves the business objectives. Although, there are several model evaluation methods, in this study, evaluation of the models were done by comparing the accuracy of the models in terms of the confusion matrices such as True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR), Relative Operating Characteristics (ROC), F-Measure, and the number of correctly classified instances which are available on Weka tool. [11]

3.1.6 Use of the Discovered Knowledge

One of the basic tasks of exploratory data analysis is to present the salient features of a dataset in a format that can be understood by humans. Even though, the purpose of modeling is to increase the knowledge about the data, the discovered knowledge need to be organized and presented in a way that can be used in an organization's decision making processes.

3.2 Data mining Techniques

Data mining techniques are used to find patterns in large datasets. Algorithms of data mining are capable of handling large and growing of datasets. Consequently to attain the objectives of research and selecting appropriate techniques. In data mining have various techniques like Classification, Clustering, Regression, and Association Rules. Among those algorithms According to data mining goals use statistical analysis and classification analysis.

There are many modeling techniques available to solve different types of data mining problems. Nonetheless, not all tools and techniques are applicable to each and every task. For certain problem types, only few data mining techniques are appropriate. Besides, among these tools and techniques, there are other requirements and constraints which further limit the choice available to the data miner, and sometimes there may be only technique is available to solve the problem at hand Many modeling techniques use specific built-in assumptions about the data, data quality or the data format which requires comparing these assumptions with the actual data characteristics described in the Data Description Report for deciding on the selection of appropriate modeling techniques to be used for the experiment [47]

The classification task consists in assigning instances from a given domain, described by a Set of discrete- or continuous-valued attributes. Hence classification algorithms are selectee in this study. The data mining algorithms for experimentation in this study are decision tree, rule based classification analysis are used.

3.3 Data mining tool Selection

There are varieties of tool available for data mining like WEKA, Shogun, Orange, Scikit-learn, R and Rapid Miner and those tools comprise on the basis of operating system and file formats supported, general features and language bindings. This is useful for various users to select the tool best suitable for their application. All the tools do not support all the data mining operations. However WEKA and Shogun supports all the three so in this research the tools selected WEKA stands for Waikato Environment open source tools available for data mining from some of tools work is for Knowledge Analysis. It is developed in Java programming language. It contains tools for data preprocessing, classification, clustering, association rules and visualization. Data file can be used in any format like ARFF (attribute relation file format), CSV (comma separated values), C4.5 and binary and can be read form a URL or from SQL database as well by using JDBC. One additional feature is that data sources, classifiers etc. are called as beans and these can be connected graphically consulting studies conducted in the area of DM tools comparison,

Chapter Four

Data Understanding and Pre-processing

In this section showed that how data is prepared for the purpose of the experiment. As showed in chapter one, under the methodology section of this study thesis, the DM process model selected Hybrid Models.it was initially developed, by adapting Cross-Industry Standard Process for Data Mining (CRISP-DM) model to the needs of academic research community [12]

Hybrid Modeling involves six iterative processes or steps that were adopted by Cios et al. [16] these steps include understanding of the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and finally use of the discovered knowledge.

4.1 Understanding of the problem Domain

This initial step that focus on whole process of data mining is to understand the need to do data mining. Which is understand the problem we need to solve this is because the goal of data mining process depending on the type of the problem to be solved. Using data mining technology in order to start actual datamining task we should able to clearly defined the problem and good understanding of the data to be used in data mining task.

The research attempt to discover to placement process of students in different department which join in the higher institute that focus one university in Ethiopia, we call it Addis Ababa University (AAU) .AAU is oldest higher learning and research institution which started its operation in 1950,it was renamed as Haile Selassie I University in 1962 and Addis Ababa university in 1975,the university started with a few teaching programmers by enrolling 71 students ,but according to AAU registrar office report add new programs and increase the number of student's currently enrolling 53861 students

As mentioned in general objective the research analyze placement process of the students this mainly find the interested model which help process predicting Department with respect of interested department of students'.

After defined the problem and goal of our datamining task the next step will be selecting an appropriate data mining tool which can perform the expected tasks in the data mining process

In our case vision use WEKA USE 3.6.13 was chosen for clustering and classification and association rule mining analysis can select easily and efficiently according to their work [45]

.

4.2 Data understanding

After defining problem to be address the next step data understanding considers data requirements it can consider analyze and understand the available data there are several things to be learned about the data before actual application of data mining techniques. In this phase can include sub tasks like initial data collection, data description, data exploration, and the verification of data quality. Specific aim of data understanding is other than identification of data quality problems, initial insights into the data, and detection of interesting or usefulness of data mining goal.

4.2.1 Initial Data Collection

The process to identify the relevant data for the primary key tasks of successful data mining goal The outcome of data mining and knowledge discovery heavily depends on the quality and quantity of available data .Thus, the initial data source for this research work, retrieved from integrated students information system for Addis Ababa university which is collecting the last two years placements of the students in different department in the university. Which contain the students the end preparatory Academic result and preference of the departments, the table were exported to MS Excel file from Microsoft SQL 2012 data base and NAEAE data base.

4.2.2 Description of the data Collected

In the process placement of departments in higher institute use the student score in EHEE and the interest of students. As described above, the initial students record that were taken from two different data base AAU Data base and NEAEA data bases database contains many attributes together with their Instances. The file organized in row and columns represent records of students whereas the columns represent attributes of score of course and preference of the interest. Summary of the data sources and general description of the collected data are illustrated in the table below.

Source Data	Year	Number of attribute	Size
AAU Data base	2009 E.C	28	1.9MB
	2008 E.C	28	1.9MB
NAEAE Data base	2009 E.C	20	90.6MB
	2008 E.C	20	87.25

Table 4.1 List of attributes names, their description and data type

4.2.3 Data Quality Verification

Several fundamental issues related to the data have a significant impact on the quality of the outcome of a knowledge discovery process. Which one is the problems related to data quality, such as imprecision, incompleteness, noise, missing values, and redundancy this issues must be addressed before we perform any further work on the data; otherwise, we may encounter serious problems later in the knowledge discovery process

In this regard, the initial dataset has been statistically described and visualized using Microsoft Excel to examine the properties of the whole dataset records and to obtain high level information regarding the data mining questions. Simple statistical analysis has been performed to verify the quality of the dataset, addressing questions such as: Does the data cover all cases required? Is the data correct or does it contains errors? Are there missing values in the data? See more about the statistical descriptions of all attributes in the initial data set.

4.2.4 Exploration of the data

This task addresses to visualize through tables and graphics such as graphs and histograms as well as through more specific techniques that help data mining questions viewing summary statistics, visualization, and reporting techniques. These include distribution of key attributes (for example, the target attribute of a prediction task) relationships between pairs or small numbers of attributes, results of simple aggregations, properties of significant sub-populations, and simple statistical analyses. These analyses may directly address the data mining Goals; they may also

contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed for further analysis.

As a result, the researcher has performed basic statistical data analysis on the initial dataset to clarify the data mining goals or to make them more precise. In this task, basic statistical exercise like frequency distribution Descriptive statistics' explore cross tab and so on are used to conducted identify the characteristics of interesting sub-populations using SPSS and MS Excel software. Then, we have analyzed the properties of major attributes that indicate the data characteristics or lead to interesting data subsets for further examination. Accordingly, in the following section we describe the first findings of the primary data analysis and evaluate this information regarding their impact on the remainder of the study.

The Ministry of Education administers entrance examination at national level for college preparatory II (grade 12) students every year then the national examination center corrects, grades, and distributes the results to students every year. Based on the criteria set by the Ministry of Education those students who satisfy the entrance requirement will fill in a form about their choice of field of study and higher education institution they would like to join.

In Ethiopia higher education institutions, Which Accept number of students that full requirement based of ministry of education and resource utilization of higher institute. The same process is done by university to manage student placement process based on choice preference of interested Department based on available field and total Entrance Exam.

4.2.4.1 Choice of Preferences

Department Placement process done in the university one attribute is of choice order preference had as placed band (Field of studies) it was already placed by institute of NEAEA. The placement process in different department is responsibility for university. As we know AAU have 72 under graduate regular program according to ministry of education those program have placed more than 6 bands each bands have their own number of possible choice programs prepare to choose by as below table

Band no	Band name	No Choice For preference	No of students
1A	Engineering and technology	4	3394
2	Computer science	2	419
	Natural science and computational science	7	1024
3	Medicine	1	487
	Other Health science	6	46
	Dental Medicine	1	941
4	Veterinary medicine	2	127
	Other agricultural science	5	
5	Business and economics	4	777
	Business and Economics (Commerce)	8	541
6	Other social science and humanities	20	1868
	Law	1	53
	Music Theatrical and fine arts	3	292

Table 4.2 Band of field access of choice

The above table shows that number of the students prepares for process of placement in the last two academics' years and their possible number of choice with respected band. Each band independent of field of Choice and the students' expected to put order of for each possible number of choices according to their interest with each respect bands.

4.2.4.2 Result of the subject

According the order of preference placed their department have a few criteria, one of the criteria those students 'get their first choice using high score of exam in EHEEE. Hence from the above data the population of the students was taken placement with result about 78% of the population. The others were taken female and disability of the students placement criteria where the students to get the choice of preference. Where students' take as criteria the sum each subject that performing in the educational entrance examination (EHEEE). It result compline the subject of in social and natural stream of the students. of the results from this the last two years students were placed in Addis Ababa university 11320 students where 62.95% Natural students and the other 37.05% where social students

Natural stream	Social stream	In common
Natural Mathematics	Social mathematics	English
Physics	Geography	Aptitude
Chemistry	History	Civics
Biology	Economics	

Table 4.3 Type of the subject for EHEEE

The above table were see the subject of the students were taken as Ethiopian higher Educational entrance Exam where taken the result out of 100%

The below the Score the Score the subject that classified with natural and social science on average the score

Descriptive Statistics

	Minimum	Maximum	Mean	Std. Deviation
Eng	19	93	65.75	13.839
NMa	14	95	56.72	13.686
Apt	25	97	66.87	12.604
Phy	16	92	52.46	13.906
Che	23	99	70.06	11.940
Bio	24	99	79.42	11.564
Civ	17	95	74.08	10.986
Valid N (listwise)				

Table 4.4 Descriptive statistics ‘for natural science Score EHEEE

Descriptive Statistics

	Minimum	Maximum	Mean	Std. Deviation
Eng	16	90	55.72	14.586
Apt	23	92	56.53	12.842
Geo	20	97	68.94	16.030
His	18	89	50.24	13.870
Eco	19	92	59.86	14.825
SMa	0	80	41.25	13.062
Civ	18	93	70.44	11.673
Valid N (listwise)				

Table 4.5 Descriptive statistics ‘for Social science score EHEEE

4.2.4.3 Region

The students came from differ region to university where are placed according to EHEEE Criteria the same thing use the attribute region use the students placement process students one that region came from developing as came some quota so it is one of the criteria for the placement process there is 9 region and 2 subsite administration and one region out of Ethiopian area. The population see below the graph.

Region

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Addis Ababa	4518	39.9	39.9	39.9
	Afar	24	.2	.2	40.1
	Amhara	1618	14.3	14.3	54.4
	Benishangul Gumuzi	24	.2	.2	54.6
	Dire-Dawa	134	1.2	1.2	55.8
	Gamble	48	.4	.4	56.2
	Harare	50	.4	.4	56.7
	Oromia	1898	16.8	16.8	73.4
	Riyadh and Jeddah	10	.1	.1	73.5
	SNNPRS	1998	17.7	17.7	91.2
	Somalia	258	2.3	2.3	93.5
	Tigray	740	6.5	6.5	100.0
	Total	11320	100.0	100.0	

Table 4.6 Students 'Statistics in region

4.2.4.4 Choice of preference with band

According to in the two year data few number of the students in the Band/Fields which is Prepared to Students placement process

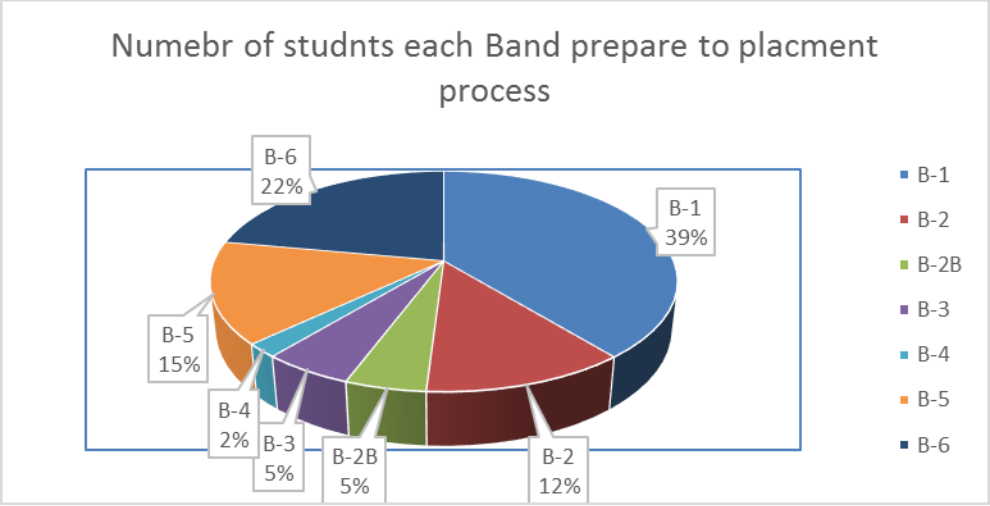

Figure 4.1 number of students with their placed Band

Pre-engineering B-1		% Selected Department	Pharmacy		% Selected Department
Order_pref	3031	**87.50%**	Order_pref	394	**48.22%**
1	2834		1	190	
2	172		2	78	
Biology			Accounting and Finance		**25.15%**
Order_pref	945	**37.78%**	B-5	1956	
1	357		1	492	
2	291		2	419	
Computer Science			Social work		**34.90%**
Order_pref	367	**85.56%**	B-6	1639	
1	314		1	572	
2	53		2	467	

Table 4.7 order of the first choose and sample Departments

AS we have seen the above some Department with respect Band some department like in Band1 per-engineering department out of the total student's in this band 87.5% of the interested with Department of Engineering. In the second Band -2 Natural and Computational Science are six more choose but from this Biology 37.78% of students are interested in the next, Band-3 other Health science pharmacy 48.22% are selected, in Band-5 Business and Economics and the Band -6 other Social Science 25.15%and 34.90 where inserted with Accounting and finance and Social work respectively choose the students

4.2.4.5 Profile

In this Attribute that students explore other than academicals description but support and use the placement process like school type, age , Sex ,nationality , Signing use as deceptive statistics as

SPSS IBM 22 AS Follows but in the case nationality almost all are Ethiopian but 0.1% are non-Ethiopian.

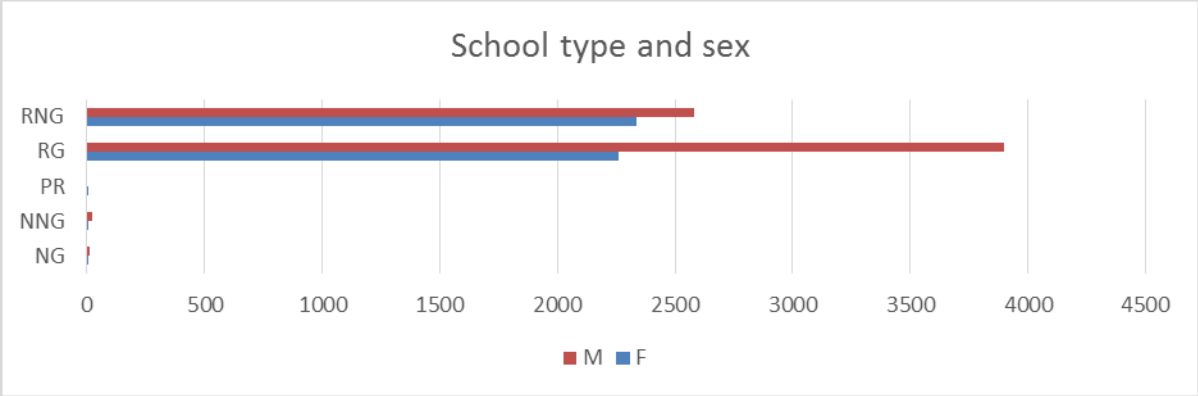

Figure 4.2 school type with sex

The number students in great in Regular Government and regular non-Governmental students' the other is insignificant in number as we have seen the above chart the other age population grouped in to tree which is use as group below 17 and above 22 are frequency small in number

Age Category		Age and sex		Total
		F	M	
Age	Below 17	52	56	304
	17	1164	1292	2456
	18	2312	2566	4878
	19	792	1312	2104
	20	232	862	1094
	21	36	178	214
	22	8	104	112
	Above 22	14	144	158
Total		4610	6514	11320

Table 4.8 Descriptive statistic's for age and sex

After exploit all data using prepare to preprocess data use cleaning detected outlier

4.3 Data Preparation

Data preparation is the start of the data mining process. Many of the in the raw data highly susceptible to irrelevant and redundant information present or noisy and unreliable data, which is affecting the result in the quality of data mining result to improve quality of data or consequently

data mining result data preprocess key issue to improve efficiency and ease of the mining process. Data preprocess key stage in the knowledge discovery process which involves many different task are routine tedious and time consuming activity. The time spent 60-80% for data preprocessing. The phase Basic operations include removing noise cleaning if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes. After we integrated the data into one file, to increase interpretation and comprehensibility, we discretized the numerical attributes to categorical ones to get final dataset found. [46]

In this study, the major activities done during data preparation phase included data cleaning, data selection, attribute or feature selection, data transformation and aggregation, data integration and formatting of the dataset. The purpose of these activities was to produce best model that can predict placement of department university students.

4.3.1 Data selection

The whole data set may not be used for data mining task. Irrelevant or unneeded or not related to the data mining object data are usually eliminated from the data mining data set. Before starting the actual data mining function. which belongs to the most fundamental steps in data processing We can define data selection as a process of finding a subset of data set, from the original set of features forming patterns in a given data set, according to the defined criterion of selection Here, we consider feature selection as a process of finding the best feature subset from the original set of pattern features, according to the defined feature goodness criterion which objective the problem data mining without additional feature transformation or construction. Other criteria for excluding data may include resource constraints, cost restriction on data use quality problem [16]

As we have seen the above initial data collection the target file exported in to MS-Excel file contains 11320 records and 34 attributes. Not all of collected dataset for experimentation. Firstly to check the relevance of each record to overall research goal and missing attribute.

Also from the total records of selected fill two criteria one students' placed order of preference from one up to three and no missing value and the score of students' in exam NEAEA in the data set.

4.3.2 Data Cleaning

Data cleaning routine work of the phase which refer to the preprocess of data in order remove and reduce which data by filling in missing values, smoothing noisy data, identifying outliers, and resolving inconsistencies. It is the process of ensuring that all values in a data set are consistent and correctly recorded. This may involve selection of clean subsets of the insertion of suitable defaults or more ambitious techniques such as the estimation of missing data by modeling. Generally, data cleaning is a process which fills missing values, removes noise (invalid data), and corrects data inconsistency the researcher used the MS-Excel application for cleaning data to familiarity with application.

Data cleaning task deals with all the data quality issues until the targeted dataset reaches the level required by the selected analysis techniques. Accordingly, the researcher took actions such as Missing Value Handling, and Outlier Detection and Removal on the selected dataset to improve the quality of the targeted dataset .The data cleaning report describes the decisions and actions that were taken to address the data quality problems of the initial

4.3.3 Missing Value Handling

Missing value refers to the values of one or more attributes in a data that do not exist. Sometimes missing value may be significant by itself and has to be properly handled. Theoretically, there are several methods suggested in the data mining literature to handle missing values, such as: calculating the average of continuous attribute values and filling this mean value for missing attribute values, removing the tuple, using the global constant value, and filling in the missing value manually [16]

In the initial dataset shows that some attributes in the original dataset contain missing values, ranging from 0.1% to as high as 100%, which are difficult to predict, replace or fill. Thus, the researcher has decided to ignore those attributes that have large amount of missing values from the dataset to assure the quality of the data. This is because, the researcher believe that trying to fill these missing values using any accepted method will result in changing the original input data with artificial data

4.3.4 Outlier Detection and Removal

The data stored in a database may reflect outlier – Noise is a random error or variance in a measured variable. Given a numerical attribute noise, exceptional case, incomplete data object and random error in a measure of attribute values. These incorrect attribute values may occur due to data entry problems, faulty data collection, inconsistency in naming convention or technology limitation. According to there are four basic noise handling methods for a given dataset such as: binning, clustering, regression and combined computer and human inspection methods [16]

In the study first to identified and detected some noise or outlier value from the target dataset, through analysis of statistical measure of variables use available in SPSS use the method applied for handling outliers from the dataset for those attributes containing continuous numeric values such age of the students, Then, high and low extreme values were replaced with mean value for those data instances and In addition, some of the identified outlier values were corrected manually and others were removed from the dataset based on a combined computer, researcher and domain expert

4.4 Data Integration

Data Integration refers combines data from different sources into a single mining Database and requires reconciling differences in data values from the various sources. It in include merging of data from multiple data Stores by which information is combined from multiple tables or other information sources to create new records through joining two or more tables together that have different information about the same objects.

Accordingly, to integrate the students placement data integrated tree different tables which the first table have score in the exam subject (EHEEE), the Second Table have the order of interested Department or preference of the students and the last table placed department . This at of the tribute was included to make aggregation of the placement of the students department data simple and suitable for using MS-Excel after completing basic data pre-processing tasks separately on each file, the data files (i.e. from 2015/16 to 2016/17 G.C.)

In this study datasets are integrated from different datasets. During this step attributes that has the same name is merged as one attribute and attribute that has the same value is adjusted to a

common value. In addition, the same attribute that have different name representation is adjusted to common attribute name. Generally, attribute that have the same instance but have different Express in different attribute can merge using use registration number and use Id number of the universe merge and in this step data inconsistence is resolved and one integrated dataset is created in this step; list of student record was found. Sample record of the dataset are have different information about the selected variable were integrated or merged into one file for further analysis. [42]

4.4.1 Attribute or Feature Selection

EDM Feature Selection is made of candidate Variable .It influence the predictive accuracy of any performance model and also elaborate the model is connected to Feature Selection techniques. Data sets for analysis may contain hundreds of attributes, many of which may be Irrelevant full of noise to the mining task or redundant which affect the success of data mining algorithms on a given tasks and the quality of the data. Attribute selection is the process of selecting a subset of relevant features. Attribute selection is a particularly important step in analyzing the data from many Experimental Techniques for use in model construction. There are several reasons for attribute selection mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand [46]

Consequently, after consultation with the domain experts the meaning of the attributes, the researcher decided to eliminate some attributes from the target dataset, since their instances are irrelevant for the analysis of this data mining problem domain and it is not related for the object of the paper. Thus 8 Attributes are irrelevant or redundant or already represented by other attributes in the database were excluded from the target dataset. Besides, those attributes with no variation in their value throughout the dataset and attributes which serve to assign sequence number for the records were also eliminated.

At this point, the number of attributes has diminished considerably, and out of the 28 attributes of the original dataset the researcher selected 17 attributes that are found to be more relevant and necessary to meet the research goals. Descriptions of the selected attributes that have been taken from the original dataset for experimentation are shown in Table

S.No	**Attribute Name**	**Data Type**	**Description**
1	Age	Numerical	Age of the students
2	Sign	Char	Sign of capability
3	Sex	Char	Sex
4	Math	Numerical	Mathematical score in EHEEE In 100
5	Eng	Numerical	English score in EHEEE In 100
6	Nmath	Numerical	Natural Mathematical score in EHEEE In 100
7	Apt	Numerical	Aptitude score in EHEEE In 100
8	Phy	Numerical	Physics score in EHEEE In 100
9	Chem	Numerical	Chemistry score in EHEEE In 100
10	Bio	Numerical	Biology score in EHEEE In 100
11	Civics	Numerical	Civics score in EHEEE In 100
12	Reg	String	Region
13	First	String	First preference of department
14	Second	String	Second preference of department
15	Third	String	Third preference of department
16	D.Placed	String	Placed department
17	School code	Nominal	School code

Table 4.9 List of attribute

4.5 Transformation and Reduction

It important component of data preparation, data transformation is data are transformed or consolidate into forms for appropriate for mining it use simple mathematical formulations or learning curves to convert different measurements of selected, and clean, data into a unified numerical scale for the purpose of data analysis.

Data transformation can involve Smoothing, which works to remove noise from the data. Such techniques include binning, regression, and clustering, Aggregation, where summary or aggregation operations are applied to the data, Generalization of the data, where low-level is replaced by higher-level concepts through the use of concept hierarchies. Normalization the attribute data are scaled so as to fall within a small specified Attribute construction (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process. [16]

The data needed to be reduced in order to make the analysis process manageable, reduce Complex data analysis, cost- effective and mining on huge amounts of data can take a long time. It makes such analysis impractical or infeasible that data reduction techniques include

Dimensionality which discussed in the above reduce the number of attribute according to initial object and Numerosity reduction, where encoding mechanisms are used to reduce the data set size reduction. Where the data are replaced or estimated by alternative, smaller data representations such as sampling use purposively the researcher focus on the placement process department for engineering and science of the students. [16]

Data discretization is one of data transformation methods which is used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values, data cube aggregation, dimensional reduction (irrelevant or redundant attributes are removed), and data compression data is encoded to reduce the size, numerously reduction (models or samples are used instead of the actual data). [16]

Data analysis using classification technique can only be performed on definite data (discrete or nominal values). Thus, data reduction on numeric or continuous values of attributes is required in order to make the analysis process manageable. According to [14] data reduction techniques include data discretization, which is one of several data transformation methods used to reduce the number of continuous values of a given attribute by dividing the range of the attribute into intervals. Then, interval labels can be used to replace the actual data values using different methods such as: data cube aggregation, dimensional reduction (irrelevant or redundant attributes are removed), data compression (data is encoded to reduce the size), and numerous reduction (models or samples are used instead of the actual data). Generally, discretization reduces significantly the number of possible values of continuous feature since large numbers of possible values contribute to slow and ineffective process of inductive machine learning.

In this research some attributes in the targeted dataset like the score the subject like English, natural mathematics Aptitude, Biology, Chemistry, civics have numeric data type or it have continues value. Hence, these continuous values have been grouped into ranges to have a combined smaller number of group values or intervals, and their data type converted from numeric to nominal Were discretized to reduce the unlike values of the attribute to obtain knowledge (pattern) and to make the dataset suitable for data mining tools.

4.5.1 Data Discretization

It divides numerical data in to categorical classes that are easier to understand the data. Discretization and concept hierarchy use raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of Abstraction, and are a powerful tool for data mining. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results

Concept hierarchies for numerical attributes can be constructed automatically based on data discretization it has some method to use .In this research use Bin method. Binning is a top-down splitting technique based on a specified number of bins. Attribute values can be discretized by applying equal-width, or equal-frequency binning, and then replacing each bin value by the bin mean or median, as in smoothing by bin means or smoothing by bin medians, respectively.

In our case all subject of score have numerical value use to in categorical class that are easily understand where discretization using the equal-width method it is method divides the data into a fixed number of interval of Equal length so all data were categorized in to five interval which labeled (Excellent ,Good ,Average, poor and fail. The interval were set by taking the maximum and minimum values of activities in computing width of interval in equal-width of interval binning method. Each subject their own minimum and maximum value so we calculate their value in to different interval

The following tables illustrate the transformed attribute values that were created by grouping their continuous values into smaller number of categorical data to reduce complexity of the results of each subject of the score

Type score	Type of category contain value				
	Excellent	Good	Average	Poor	Fair
English	95	79	64	49	34
Mathematics	100	85	68	51	34
Aptitude	97	70	56	42	28
Physics	98	80	64	48	32
Biology	99	84	69	54	39
Chemistry	98	85	70	55	40
Civics	98	81	65	49	33

Table 4.10 categorical value of the score the subject

4.5.2 Discretizing the Age attribute value

The other numerical data in to categorical classes that are easier to understand the data. Use to discretize equal frequency of the age value. The value grouped in to three group use ascending order the first to and then grouped in to categorical data

The end result of the data preparation phase is to come up with a prepared dataset for best enhanced modeling. In the targeted input dataset there are 12 selected attributes having 4303 records and the following table presents the descriptions of these attributes with their possible values.

No	Attribute name	Data type	Description	Possible value
1	S.type	Numerical		Regular Government school students (RG)=01 Regular non-government school type (RNG)=02 Night Government school students (NG)=03 Night non-government school students (NNG) =04 Private students Who sat only for examination (PR)=05
2	Sex	Char	Sex	Male =M Female F
3	Region	Char	Region	Tigray=1 , Afar,=2 , Amhara=3,Oromiya=4 , Somalia =5 Benishangul Gumuzi =6 , , SNNP=11, Gambela=12, Harari=13 Dire-Dawa = 15, Addis Ababa =14
4	F.choice S,Choice Third choice	String	First choice Second choice Third choice	Pre-engineering =pre Bachelor of Science in Architecture =Arch Bachelor of Science in Biology=Bio Bachelor of Science in Chemistry=che Bachelor of Science in Mathematics=math Bachelor of Science in Construction Technology & Management=Cotm Bachelor of Science in Computer Science=comp Bachelor of Science in Information Systems=INF Bachelor of Science in Physics=phy Bachelor of Science in Geology=Geol Bachelor of Science in Sport Science=spo Bachelor of Science in Urban and Regional Planning=url Bachelor of Science in Statistics=stat
5				
6				

Table 4.10 List of attribute and possible value

After the process of Discretizing data formatting task come next which primarily refers to syntactic modifications or changes made to the data without changing its meaning, but might be required to satisfy the requirements of the specific modeling too. Most data mining tools in use can handle only certain types of file format and data types. For instance, WEKA software requires the target dataset to be prepared in the Comma Separated Value (CSV) file format to run experiments. This software either process the CSV file format itself or a file in the form of Attribute Relation File Format (ARFF). WEKA software allows the conversion of CSV file format into an ARFF file format which is acceptable for running experiments on WEKA software.

In our case all data pre-processing was performed in MS –Excel format however selected WEKA 3.6.13.It does not accept MS-Excel format so the which MS-Excel format first convert to comma separated value text format. (csv format) then to ARFF(Attribute relation file format) comma delimited is applied for the list of records where the item are separated by commas where ARFF is an extension of file format that is convenient for the data mining tool WEKA. In convert the data into ARFF format declaration such as @relation<relation name >, @Attribute <attribute name > together with the data type and @data are added in the CSV files tasks have been completed, the target dataset were imported and saved into CSV file format using SPSS package, and then, the targeted data was converted successfully into Attribute Relation File Format (ARFF).

1	AdmissionNumber	Year	Scho	School code	Age group	Age	Sex	Region	Sig	Nation	Eng_Al	Nma_Al	Apt_Al	phy_Al	Chem_Al	Bio_al
2	545896	2009	304	RG	G-1	15	M	3	N	E	Good	Good	Good	Average	Average	Good
3	793024	2008	1405	RG	G-1	16	F	14	N	E	Average	Fail	Poor	Fail	Average	Average
4	820863	2008	1406	RNG	G-1	16	F	14	N	E	Excellet	Average	Good	Average	Average	Good
5	752081	2009	420	RNG	G-1	16	F	4	N	E	Poor	Average	Average	Poor	Average	Average
6	820864	2008	1406	RNG	G-1	16	M	14	N	E	Excellet	Average	Good	Average	Average	Good
7	808420	2008	420	RNG	G-1	16	F	4	N	E	Poor	Poor	Poor	Poor	Fail	Good
8	823446	2008	1409	RNG	G-1	16	M	14	N	E	Good	Poor	Good	Good	Poor	Good
9	817834	2008	1404	RNG	G-1	16	M	14	N	E	Good	Poor	Good	Good	Good	Excellent
10	749567	2008	1103	RG	G-1	16	F	11	N	E	Average	Poor	Poor	Fail	Average	Poor
11	818940	2008	1405	RNG	G-1	16	F	14	N	E	Good	Average	Average	Poor	Average	Good
12	826319	2008	1410	RNG	G-1	16	M	14	N	E	Good	Poor	Good	Poor	Good	Excellent
13	668666	2008	311	RG	G-1	16	F	3	N	E	Good	Average	Average	Average	Good	Excellent
14	810308	2008	503	RNG	G-1	16	M	5	N	E	Average	Average	Good	Good	Average	Good
15	811302	2008	1106	RNG	G-1	16	M	11	N	E	Average	Average	Poor	Average	Good	Good
16	809670	2008	427	RNG	G-1	16	M	4	N	E	Good	Average	Good	Average	Good	Excellent
17	660177	2008	308	RG	G-1	16	M	3	N	E	Average	Poor	Average	Average	Average	Good
18	813559	2008	1119	RNG	G-1	16	F	11	N	E	Average	Average	Good	Average	Good	Average

Figure 4.3 preprocess data using Excel

Chapter Five

Experimentation and Discussion on Results

In this chapter, the researcher describes the techniques that have been used in developing a model to predict placed department of the students. In addition to incorporated typical stages that characterize a data mining process. This study has been organized according to hybrid datamining process model, which is described and discussed section methodology in chapter one. Here the researcher discuss the experimentation process by relating the steps followed, the choice made , the task accomplished , the result obtained, evaluation of the model and results , and it present a way that the organization can easily understand and use it.

5.1 Selection of Modeling Technique

Modeling is one of the major tasks which is undertaken under the phase of data mining in hybrid methodology. In this phase several data mining techniques various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, different techniques can be employed for similar data mining problems. Some of the tasks include, selecting the modeling technique, experimental setup or design, building a model and evaluating the model.

Selecting appropriate model depends on data mining goals. Consequently, to attain the objectives of these research three classification techniques has been selected for model building. The analysis was performed using WEKA environment. Among the different available classification algorithms in WEKA, PERTree, J48 and are PART tree re used for experimentation of this study. The researcher selected the above algorithms, easy of understanding and interpretation of the Result of the model.

As we have seen, decision tree algorithms are based on a divide-and-conquer approach to the classification problem. They work from the top down, seeking at each stage an attribute to split on that best separates the classes; then recursively processing the sub problems that result from the split. This strategy generates a decision tree, which can if necessary be converted into a set of classification rules although if it is to produce effective rules, the conversion is not trivial [11]

A PART algorithm is common types of rule induction technique which generate a model as a set of rules. In Other case the J48 algorithms of decision tree generate a model by constructing decision tree where each internal node is a feature or attribute. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specific call in decision analysis, to help identify a strategy most likely to reach a goal. [11]

5.2 Experimental Design

Prior to building a model, it is necessary to define a procedure to test the model's quality and validity, Which is means the model prepare for experimentation plan for training, testing and evaluating is required [33]. Thus, in this experiment 17 attributes were selected during the data preparation phase to train and test the selected classifiers .A total of 4284 records, extracted from the original dataset using a proportional purposive sampling technique
[34].It is also important to build models by intelligently adjusting these parameters. Often, only repeated experiments and familiarity with the data will test out the best set of options as shown in Table 4.1 below which is taken from WEKA Manual

Name	Possible values	Data Type	Default value	Description
Binary Splits	True/False	Boolean	False	Whether to use binary splits on nominal attributes when building the trees
confidence Factor	10..-.05	Numerical	0.25	The confidence factor used for pruning (smaller values incur more pruning
Min-Num Obj		Numerical	2	The minimum number of instances per leaf.
Sub-tree Raising	Yes/No	Boolean	Yes	Whether to consider the subtree raising operation when a leaf during pruning
Unpruned		Boolean	False	Whether pruning is performed.

Table 5.1: Summary of Classifier Parameters for the Experimentation

As shown above in Table default classifiers present smaller value of confidence factor help to enforce more pruning force the algorithm to use the unpruned tree instead of the pruned one; or use Laplace smoothing for predicted probabilities. The default confidence factor is 0.25 Confidence factor works only when the unpruned parameter is set to "False". The*unfolds* parameter (default 3) determines the size of the pruning set: the data is divided equally into that number of parts and the last one used for pruning. One experiment is done by reducing the confidence factor. The minimum number of instances by default is 2 and another experiment is carried out by increasing this value. The sub tree raising parameter is by default set to "True" to replace nodes in a decision tree with a leaf during pruning. Binary Splits parameter by default is set to "False" and it is used as it is to generate generalized decision tree instead of binary decision tree, [11]

Commonly used parameter of classifier like J48 algorithms was illustrated with various options related to tree pruning. Hence, the process ensures the maximum accuracy on the training data, but creates excessive rules there is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that subtree rising can be somewhat computationally complex. [11]

Reduced-error pruning option: Error rates are used to make actual decisions about which parts of the tree to replace. There are multiple ways to do so. The simplest method is to reserve a portion of the training data to test on the decision tree. The reserved portion can then be used as test data for the decision tree, helping to overcome potential over fitting. This approach is known as reduced-error pruning. This method reduces the overall amount of data available for training the model.

Confidence factor option: Confidence factor is the approach that seeks to forecast the natural variance of the data, and to account for that variance in the decision tree. This approach requires a confidence threshold, which by default is set to 25 percent. This option is important for determining how specific or general the model should be. If the value of the confidence factor is lowered on the selected model, the training data is expected to fit in closely to the data we would like to test. As a result of this a more pruned or more generalized tree will be produced.

Minimum number of instances per leaf option: this is the lowest number of instances that can constitute a leaf. The higher the number the more general the tree will be Lowering the number will produce more specific trees, as the leaves become more granular.

Binary split option: The binary split option is used with numerical data. If turned on, this option will take any numeric attribute and split it into two ranges using inequality. This greatly limits the number of possible decision points. Rather than allowing for multiple splits based on numeric ranges, this option effectively treats the data as a nominal value. Turning this encourages more generalized trees.

Laplace smoothing option: This option is used to prevent probabilities from ever being calculated as zero. This is mainly to avoid possible complications that can arise from zero probabilities

Besides, other standard measure including precision, recall, sensitivity and specificity are available. Therefore, the test design specifies that the dataset should be separated into training and test set, and builds the model on the training set and estimate its quality on the separate test set. Process of building predictive models requires a well-defined training and validation protocol in order to insure that most accurate and robust prediction [26]

As above suggestion as in this Research use the data set as training and testing In WEKA Environment has used to set up and measure the quality, validity and test of the selected model. For purpose of this study k-fold (10-folds) cross validation percentage than 30-70 and 66-33 split test options are used because of its relatively low bias and variations

5.2.1 Attribute ordering

Since attribute selection is important in decision tree models, because in most case all attribute are not equally useful for predicting the target so we need to first evaluate the usefulness of each attribute before conducting the experiment of classification in order of this the tried to rank the attribute based on information gain. It was calculated based on entropy value of the attribute. As explained information gain is calculated from sum of entropy for every attribute. The formula for calculating intermediate values is:

$$\text{Info (D)} = -\Sigma^{m}_{i=1} \text{Pi} \log_2 {}^{\text{Pi}}$$

Where, Pi is the probability that an arbitrary tuples in sets of training data D belongs to certain class. Info (D) is also known as the entropy of *D*. After you calculate information gain for each attribute, select the one with the highest information gain as the root node, and continue the calculation recursively until the data is completely classified by J48 algorithms in this [25]

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 17 D.placed):
        Information Gain Ranking Filter

Ranked attributes:
 1.71419  14 First _Pref
 1.20308  15 Second_Pref
 0.77749  16 Third_Pref
 0.3679    7 Eng_Al
 0.33073   9 Apt_Al
 0.26316  12 Bio_al
 0.20947   8 Nma_Al
 0.20927  10 phy_Al
 0.20212  11 Chem_Al
 0.18936  13 Civic_AL
 0.18689   1 School code
 0.09035   4 Region
 0.08061   3 Sex
 0.05775   2 Age group
 0.00856   5 Sig
 0.00245   6 Nation

Selected attributes: 14,15,16,7,9,12,8,10,11,13,1,4,3,2,5,6 : 16
```

Figure 5.1 output of attribute ranking with information gain attribute ranking

As show the figure above depict ranked order of attribute based on their relevance for the reason that such attributes are very important for later experimentations. As you can see from the result of attribute selection using entropy based information gain method of WEKA, from attribute the last tree attribute (age, nationality sign) are not interest and no more relevant to the target class.

Put some scenario in the selected classification method of algorithm model. Then compare the result of one model to other and finally help us to find out the performing model based on criteria of evaluation. Consequently for both of the methods following scenario has been done for each of three selected model with default parameter value of WEKA software. [48]

Accordingly, different scenarios, which were considered by the researcher to run the experiments, with their adjusted parameters' values are listed below:

- **Scenario One:** Decision tree using J48 algorithm with default parameters settings
- **Scenario Two:** Decision tree using J48algorithm with adjusted parameters settings.
- **Scenario Three:** Decision tree using REP Tree algorithm with default parameters settings.

- **Scenario Four:** Decision tree using REP Tree algorithm with adjusted parameters settings.
- **Scenario Five:** Decision rule using PART algorithm with default parameters settings.
- **Scenario Six:** Decision rule using PART algorithm with adjusted parameters settings.

5.2.2 Model Building using J48 Decision Tree

Model building is one of the major tasks that are undertaken under the data mining phase in the hybrid methodology of conducting Data Mining Researches. The following figure shows the attributes which were selected for the experiments [49]

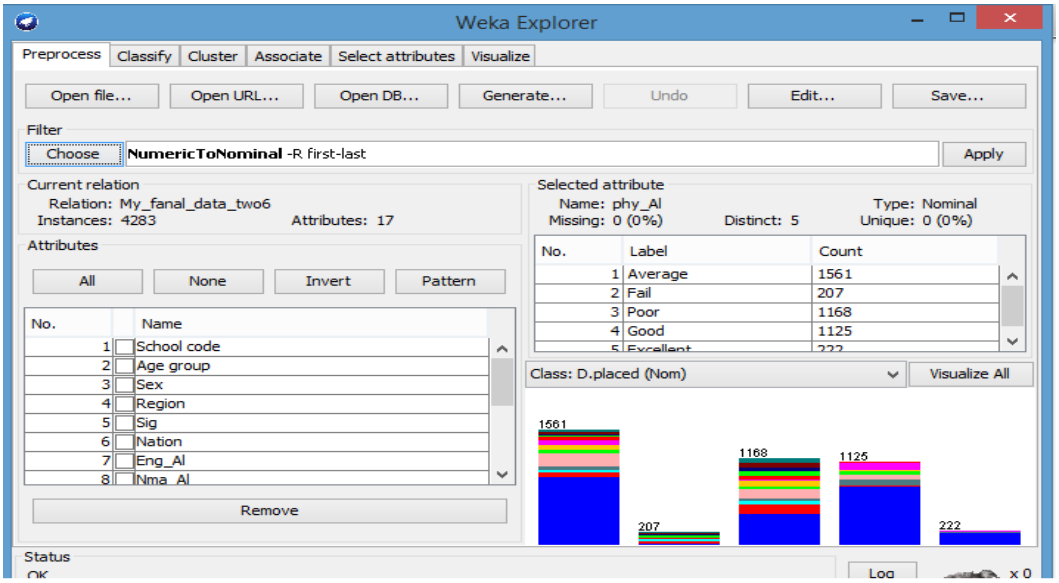

Figure 5.2: A Screenshot that Shows the Attributes for the Experiments.

The most widespread classification algorithms are algorithms for building decision trees. Decision trees are very frequently used for classification, as they are easy to use and understand. There are two criteria for the quality of decision trees one is classification accuracy and the other decision tree size.

The classification accuracy is the proportion of correctly classified instances in a given set of instances. If a decision tree would have been based on all available instances training and testing. Therefore, a decision tree is built on a subset of all available data, called the learning set. The remaining data are used to assess the accuracy of the built classification tree and is called the test set. To avoid dividing sensitive information between the learning set and the test set, the

procedure is repeated n times, where the test set is one of the n parts of the data set and the rest is used for the learning set. This method is called n-fold cross-validation

The Size of decision tree is expressed number of node in which increasing size of the decision tree its incomprehensibility also increases, which is particularly undesirable if the decision tree is used for the interpretation of the discovered relationships in the training data. If a decision tree is used for classification, its size it is not so significant, as the implementation of decision trees in practice is not demanding. The size of decision trees can be controlled during the construction process by pruning. There are two approaches to decision tree pruning: preprinting (terminating the subtree construction during the tree-building process) and post running (pruning the subtrees of an already constructed tree). [11]

According to [11]Model building step, many parameters settings could be adjusted on the modeling tool and can be used as standard measure to enhance the quality of the model. Before running the model building experiments the researcher has tried to adjust different parameter settings of J48 algorithm with its chosen values (i.e., confidence Factor=0.50, minNumObj=5 and Pruned=True) in order to obtain models with better accuracy

Experimets	Parameters setup		
	Pruned	**Confidence Factor**	**(minNumObj)**
Experiment #1	True	.25	2
Experiment #2	True	.50	2
Experiment #3	True	.25	5
Experiment #4	True	.50	5
Experiment #5	False	.25	2
Experiment #6	False	.5	2
Experiment #7	False	.25	5
Experiment #8	False	.50	5

Table 5.3: Adjusted Parameters Values of J48 Algorithm

The above table show according the above scenario one and two default parameter and adjusted parameter in algorithm J48 decision tree .The selected modeling technique with its adjusted parameters values which changing the parameter pruned ,confidence factor and Minnumobj

Check possible experiment.to see the performance of the model

Once the modeling techniques, tool and the evaluation criteria were established, then building models with a number of parameters setup which governs the kind of models generated by the process was followed .Since, WEKA Explorer is sensible to its default values, initially, J48 algorithm with its default parameters values has been run on the targeted input dataset and the experiments continued by adjusting those default parameters values in the object editor of WEKA 6.13 The following table shows the summary of experimental results of J48 Decision Tree.

Evaluation Criteria	**Results of the Experiments for J48**							
	Scenario One – (Pruned)				**Scenario Two – (Unpruned**			
	Exp #1	Exp #2	Exp #3	Exp #4	Exp #5	Exp #6	Exp #7	Exp #8
Numbers of Leaves	1395	1395	584	584	363	749	101	340
Number of tree	1696	1696	716	716	454	925	122	418
Time Taken(in sec)	0.02	0.03	0.03	0.02	0.09	0.04	0.02	0.02
Accuracy (%)	79.43	79.43	80.22	80.22	81.46	80.76	81.63	80.81
AV. TP Rate	0.794	0.794	0.802	0.802	0.815	0.808	0.816	0.808
AV. FP Rate	0.065	0.065	0.084	0.084	0.078	0.065	0.088	0.067
AV. Precision Rate	0.786	0.786	0.79	0.79	0.801	0.795	0.806	0.796
AV. Recall Rate	.0.794	0.794	0.802	0.802	0.815	0.808	0.816	0.808
AV. F-Measure	0.789	0.789	0.794	0.794	0.798	0.799	0.795	0.798
AV. ROC Area	0.905	0.905	0.938	0.938	0.943	0.924	0.949	0.951

Table 5.4 Summary of Experimental Results of J48 Algorithm

KEY: Accuracy: Registered performance of model, **AV**: Average, **TPR**: True Positive Rate. **FPR**: False Positives Rate, **PR**: precision rate, **RR**: Recall rate, **ROC Area**: Relative Optical character curve.

As we have seen the above table, different experimental results were obtained by applying J48 algorithm with its default parameters value and by adjusting its parameters values with user

chosen values. Then, by comparing the obtained results, best performing model has been chosen based on the criteria

The summarized results in Table illustrate the accuracy level of the generated models in percentage that show how much number of instances are classified correctly according to the scenario. The results show that there is significant difference in their accuracy among the eight models experiments. Result obtained from experiment five scenario two that is applyingJ48 decision tree algorithm without pruning and keeping parameters adjusted value default values performed with highest accuracy than other models. That is the obtained experimental result shows that 81.63% of the total records were classified correctly.

Below as we some sample the selected model for j48 decision tree can be presented in the form of easily understandable and an executable code with if-then constructions. Here the transformations of the decision tree representation of the J48 algorithm results to the IF_THEN form are given in below. See more List rule

- **List of Sample Rules Predicted by J48 Model**

Rule1: IF First _Pref = Pre AND Sex = M AND Nma_Al = Good THEN Pre (698.62/27.0)

Rule 2: IF First _Pref = Pre AND Nma_Al = Fail AND Apt_Al = Good phy_Al = Average THEN Pre (7.0/1.0)

Rule 3: IF First _Pref = Pre AND Nma_Al = Fail AND Apt_Al = Good AND phy_Al = Poor THEN COTM (7.0/3.0)

Rule 4: IF First_Pref =Pre ANF Nma_Al = Average: Pre (845.62/190.62)

Rule 5: IF First _Pref = COMP AND Eng_Al = Good AND Civic_AL = Excellent THEN COMP (93.06/16.06)

Rule 6: IF First _Pref = COMP AND Eng_Al = Good AND Civic_AL = Excellent THEN COMP (93.06/16.06)

Rule 7: IF First _Pref = COMP AND Eng_Al = Good AND Civic_AL = Good THEN COMP (75.06/25.06)

Rule 9: IF First _pref = Geol AND Second_Pref = Bio THEN Geol (133.97/73.0)

Rule 10: If second_pref = Geol AND Sex = M Chem_Al = Average: Geol (17.0/2.0)

Consequently, among all the possible lists of J48 decision rules generated, only few of the discovered knowledge that could enable us to predict the department of students want to learn are analyzed. The results are evaluated based on the existing knowledge base as interpreted by the domain experts. Hence, the most interesting rules or discovered knowledge were.

Rule 2: IF First _Pref = Pre AND Nma_Al = Fail AND Apt_Al = Good phy_Al = Average THEN Pre (7.0/1.0)

This rule indicated that Department of Pre engineering associated interested pre engineering field in addition to with Good result in Aptitude even if the mathematics to below poor

Rule 4: IF First_Pref =Pre ANF Nma_Al = Average: Pre (845.62/190.62)

This rule indicated that Department of Pre engineering associated interested pre engineering field in addition to with enough to result to Mathematical result Average

Rule 5: IF First _Pref = COMP AND Eng_Al = Good AND Civic_AL = Excellent THEN COMP (93.06/16.06)

This rule indicated that Department of Computer associated interested Computer science field in addition Result English and Civics Good and Excellent respectively

Rule 7: IF First _Pref = COMP AND Eng_Al = Good AND Civic_AL = Good THEN COMP (75.06/25.06)

This rule indicated that Department of Computer associated interested Computer science field in addition Result English and Civics both have score Good

First _Pref = Math: Math (23.01/0.01)

This rule indicated that Department of Mathematics interested in the field of mathematics that is good

5.2.3 **Model Building Experiment using REPT Tree Decision Tree**

This experiment was performed following the same procedure applied on the previous experiment scenarios which are presented above. That means the experiments were conducted by

applying REPT Tree algorithm, initially with its default parameters settings and then, by partially adjusting its parameters values. The experiments were run on the training dataset to build the model and its quality was estimated on the test dataset. The lists of parameters settings with their chosen values used for the experiment are shown below in table

Experimental	**Parameters setup**		
	Pruned	**Max Depth**	**(minNum)**
Experiment #1	True	-1	2
Experiment #2	True	10	2
Experiment #3	True	10	3
Experiment #4	True	20	5
Experiment #5	False	-1	2
Experiment #6	False	10	2
Experiment #7	False	10	3
Experiment #8	False	20	5

Table 5.5 Adjusted Parameters Values of PER Tree Algorithm

NB: maxDepth -- The maximum tree depth (-1 for no restriction) minNum -- The minimum total weight of the instances in a leaf

Accordingly, the summarized experimental results obtained by applying REPTree algorithm based on the above scenarios are illustrated in the following Table.

Evaluation Criteria	Results of the Experiments							
	Scenario One – (Pruned)				Scenario Two – (Unpruned			
	Exp #1	Exp #2	Exp #3	Exp #4	Exp #5	Exp #6	Exp #7	Exp #8
Numbers of Leaves	2100	1997	1434	1016	205	205	193	136
Time Taken(in sec)	0.06	0.05	0.05	0.04	0.12	0.05	0.03	0.02
Accuracy (%)	79.38	79.417	79.66	80.20	81.53	81.53	82.045	82.045
AV. TP Rate	0.794	0.794	0.797	0.802	0.815	0.815	0.82	0.82
AV. FP Rate	0.069	0.69	0.068	0.0702	0.085	0.085	0.088	0.092
AV. Precision Rate	0.78	0.781	0.783	0.79	0.811	0.811	0.825	0.833
AV. Recall Rate	0.794	0.794	0.797	0.802	0.815	0.815	0.82	0.82
AV. F-Measure	0.785	0.785	0.787	0.791	0.794	0.794	0.798	0.796
AV. ROC Area	0.911	0.913	0.928	0.942	0.955	0.955	0.955	0.953

Table 5.6 Summary of Experimental Results of REPTree Algorithm

The summarized results shown above in Table illustrate the accuracy level of the generated models in percentage, which show the number of instances that were classified correctly based on the given scenarios and the employed algorithm. From the results it can be observed that there is noticeable difference among the performance of the REPTree decision tree models. Result obtained from experiment eight in the scenario four that is by applying REPTree decision tree algorithm without pruning and adjusted parameters values performed with highest accuracy than the other models. That means the obtained experimental result shows that 82.045%of the total records were classified correctly.

- **List of Sample Rules Predicted by PERT tree Model**

Rule1: First _Pref = Pre Nma_Al = Good : Pre (620/22) [293.63/6]

Rule 2: First _Pref = Pre Nma_Al = Fail Sex = M phy_Al = Average : Pre (5/1) [5/2]

Rule 3: Rule 2: First _Pref AND ,Nma_Al AND Average AND Eng_Al AND Good AND phy_Al AND Average AND Civic_AL = Excellent : Pre (52/0) [31/1]

Rule 1: Rule1: First _Pref = Pre Nma_Al = Good : Pre (620/22) [293.63/6]

This rule indicated that Department of Pre engineering associated interested pre engineering score of mathematics is good.

Rule 2: Rule 2: First _Pref AND ,Nma_Al Average AND Eng_Al Good AND phy_Al Average AND Civic_AL = Excellent : Pre (52/0) [31/1]

This rule indicated that Department of Pre engineering associated interested pre engineering score of mathematics is average and English Good and physics and civics excellent.

5.2.4 Model Building Experiment using PART Rule Induction

This experiment was performed on the PART Rule Induction it is an alternative representative of decision tree follows the same procedure applied on the previous experiment scenarios in five and six which are presented above. That means the experiments were conducted by applying RART Rule induction algorithm, initially with its default parameters settings and then, by partially adjusting its parameters values. The experiments were run on the training dataset to build the model and its quality was estimated on the test dataset. The lists of parameters settings with their chosen values used for the experiment are shown below in table

Experiments	Parameters setup-PART Rule induction		
	Pruned	**Confidence Factor**	**(minNumObj)**
Experiment #1	True	.25	2
Experiment #2	True	.50	2
Experiment #3	True	.25	5
Experiment #4	True	.50	5
Experiment #5	False	.25	2
Experiment #6	False	.50	2
Experiment #7	False	.25	5
Experiment #8	False	.50	5

Table 5.7 Adjusted Parameters Values of RART Algorithm

Evaluation Criteria	Results of the Experiments							
	Scenario Five – (Pruned)				Scenario Six – (Unpruned			
	Exp #1	Exp #2	Exp #3	Exp #4	Exp #5	Exp#6	Exp #7	Exp #8
Numbers of Leaves	638	638	320	320	266	257	87	113
Time Taken(in sec)	0.46	0.46	0.27	0.26	0.16	0.17	0.09	0.09
Accuracy (%)	78.38	78.38	79.97	79.97	80.50	80.6	81.06	81.11
AV. TP Rate	0.784	0.784	0.8	0.8	0.805	0.806	0.811	0.811
AV. FP Rate	0.064	0.062	0.067	0.067	0.06	0.058	0.067	0.065
AV. Precision Rate	0.779	0.779	0.788	0.788	0.796	0.797	0.799	0.801
AV. Recall Rate	0.784	0.784	0.8	0.8	0.805	0.806	0.811	0.811
AV. F-Measure	0.78	0.78	0.792	0.792	0.799	0.8	0.799	0.802
AV. ROC Area	0.899	0.899	0.936	0.936	0.934	0.931	0.962	0.956

Table 5.8 Summary of Experimental Results of PART Tree Algorithm

When we compare the performance of the models produced by PART algorithm applying on the targeted input datasets, the generated results obtained from experiment eight of in the scenario six indicated that by applying PART rule induction algorithm without pruning and keeping other parameters as Adjusted value default values performed with highest accuracy than other models which is (81.11%)

The selected PART model generated 113 rules and the possible list of decision rules generated by the algorithm is shown in Annex 5.Among those rules produced by the PART rule induction algorithm, some interesting rules are selected and their interpretations have been illustrated as follows

❖ List of Sample Rules Predicted by PART Model

Rule 1: First _Pref = Pre AND Nma_Al = Good: Pre (698.55/27.0)

Rule 2: First _Pref = COMP AND Eng_Al = Excellet AND Civic_AL = Good: COMP (22.0/2.0)

Rule 3: First _Pref = COMP AND Eng_Al = Good AND Nma_Al = Average AND Chem_Al = Good AND Apt_Al = Good: COMP (20.0)

Rule 4: First _Pref = COMP AND Civic_AL = Excellent AND Apt_Al = Good AND Chem_Al = Average AND Nma_Al = Average: COMP (18.0/3.0)

Rule 5: First _Pref = Pre AND Nma_Al = Average AND Eng_Al = Good AND Bio_al = Excellent: Pre(203.0/11.0)

Rule 6: First _Pref = Pre AND Nma_Al = Excellent: Pre (157.0/4.0)

Rule 7: First _Pref = Pre AND Apt_Al = Excellent: Pre (29.0/4.0)

Rule 8: First _Pref = Pre AND phy_Al = Average AND Eng_Al = Good AND Bio_al = Good AND Apt_Al = Average AND Nma_Al = Average: Pre (18.0/4.0)

Rule 9: First _Pref = Pre AND Civic_AL = Good AND Apt_Al = Good AND phy_Al = Poor: COTM (7.0)

Rule 10: First _Pref = Bio AND Chem_Al = Good: Bio (6.01/2.01)

Rule 11: First _Pref = Bio AND Civic_AL = Good AND Chem_Al = Poor AND Nma_Al = Poor: Bio (11.0/6.0)

Rule12: Third_Pref = chem AND Apt_Al = Fail AND Civic_AL = Average: chem (7.0/4.0)

As can be observed from the above partial list of rules, partial targeted attributes have been used as input to construct rules and the classifier provided a separate class for each rule which is predicted by the model. The numeric values which appeared in the bracket next to the class label indicates the number of correctly and incorrectly classified records, respectively. Hence, sample rules generated by the model are interpreted as follows, and similarly, the remaining rules can be interpreted in the same way.

Rule 1: First _Pref = COMP AND Eng_Al = Good AND Nma_Al = Average AND Chem_Al = Good AND Apt_Al = Good: COMP (20.0)

This rule indicates that department of computer is placed associated with result of English, chemistry and aptitude have Good and natural mathematics' have average result

Rule 2: First _Pref = Pre AND Nma_Al = Average AND Eng_Al = Good AND Bio_al = Excellent: Pre(203.0/11.0)

The rule indicate that the Department pre-engineering was placed see the criteria that get result for, mathematics average, English Good and Biology Excellent.

5.3 Evaluations of the Models

Classifiers and predictive models evaluation is one of the key points in any data mining process. The main and frequently evaluation criteria desired in classification perspective is the criteria of overall accuracy obtained by model validation method [11]

One of the specific objectives of this study is to select the best alternative classification technique for building a model which can be implemented for placed Department of the students. In line with this, the overall performance of the three selected classification models were compared with each other and evaluated based on their respective results using different evaluation techniques as follows

Algorithm	Accuracy (%)	Time taken (Sec)	Number of Leaves
J48	81.63	0.02	101
PER Tree	82.05	0.02	136
PART	81.11	0.09	113

Table 5.9 the Comparison of the Selected Three Models with their Accuracy Results

The test results shown that, decision Tree using PERTree algorithm performs better than others in terms of its accuracy. That means, its ability in correctly classifying records into both 'True' and 'False' classes performs better 82.05%.

One of the famous other method of visualizing the metric predictive performance can be obtained by constructing a confusion matrix. Specific matrix layout that allows visualization of the performance of an algorithm shows the predicted and actual classifications. And also The matrix is a valuable tool because it not only shows how frequently the model correctly predicted a value, but also shows which other values the model most frequently predicted incorrectly [11]

In confusion matrix measure the value Precision, Recall, F-Measure and ROC are also a good performance measures for the model built. The ROC is also known as a Relative Operating Characteristic curve, because it is a comparison of two operating characteristics (TPR & FPR) as the criterion changes. Then, the test runs results

Evaluator	Weighted Average Values		
	J-48 test run	PERTree	PART
Precision	0.806	0.833	0.801
Recall	0.816	0.82	0.811
F-measure	0.795	0.796	0.802
ROC	0.949	0.953	0.956

Table 5.10 Evaluate model of classification algorithm

The average of these parameters like precision, recall for individual predictive classes becomes the final values of the entire model. So above from above table it can be observed that the weighted average values of Precision, Recall, F-Measure and ROC values of the selected models are nearly the same. The ROC area of all the three models is also above 0.5, which is the minimum possible acceptable value for ROC curve. ROC area is plotted from True positive Rate (TPR) on the y axis against the False Positive Rate (FPR) on the x-axis. In a case where the ROC is 0.50 it means that there is a 50-50 chance for TPR and FPR, which is quite unacceptable. When we drown the ROC Curves, they appear above the diagonal. Based on these criteria it is possible to evaluate the performance of the three selected models. Hence, among the three models, model generated by using PERTree algorithm showed best performance which is 0.956. Then, it is possible to conclude that the performance of PERTree model is promising and performs with better accuracy when compared with the other two models.

5.3.1 Confusion Matrix

In any branch of science, it is almost a common requirement that performance of various models have to be compared with each other to understand the suitability of a model to a given problem. In data mining also it is a common requirement and in this work "**Confusion matrix**" was used for this purpose.

For classification modeling, a confusion matrix is a very useful tool for understanding the results, because a confusion matrix shows the counts of the actual versus predicted class values. Besides, it shows not only how well the model predicts, but also presents the details needed to identify exactly where things may have gone wrong. The following table is an illustration of the

confusion matrix of the selected model. Whereas, the columns show the actual classes, the rows show the predicted classes and the diagonal shows all the correct predictions as we have seen below result confusion matrix with selected model.

Placed Department	Pre	Bio	Chem	Arch	COTM	INF	Urb	Comp	Spo	Math	Phy	Geol	Stat	Actual Total	Accuracy Rate (%)
pre	2296	0	0	16	0	6	0	0	0	0	0	0	0	2318	99.05%
Bio	1	200	0	0	0	0	0	0	11	1	0	12	1	226	88.50%
Chem	0	35	55	0	0	0	0	0	0	0	0	22	0	112	49.11%
Arch	0	0	0	175	0	0	0	0	0	0	0	0	0	175	100.00%
Cotm	188	0	0	0	182	0	5	0	0	0	0	0	0	375	48.53%
INF	0	0	0	0	0	75	0	54	0	0	0	0	0	129	58.14%
Urb	113	0	0	0	27	0	20	0	0	0	0	0	0	160	12.50%
Comp	0	0	0	0	27	0	0	171	0	0	0	0	0	198	86.36%
Spo	0	10	0	0	0	0	0	0	99	0	0	12	4	125	79.20%
Math	0	37	4	0	0	0	0	0	3	24	0	29	12	109	22.02%
Phy	0	32	4	0	0	0	0	0	3	1	22	17	7	86	25.58%
Geol	0	2	2	0	0	0	0	0	4	4	0	112	4	128	87.50%
Stat	0	16	2	0	0	0	0	0	5	3	0	35	63	124	50.81%

Table 5.11 Confusion Matrix of the Selected PERTreeModel

The above table of confusion matrix shows that the selected PER Tree model were classified, out of the 4283 total records of the testing dataset delivered to the program, 3514 (i.e., the sum of the diagonal results of label Pre=2296, Bio`=220,Chem=55,Arch=175,COTM= 182, INF=75,urb=20, Comp = 171, Spo=99, math =24,phy =22,Geol =112,stat=63 and e=900 **Total = 3514**) records were classified correctly and 769 records were classified incorrectly.

At this stage, it is reasonable to conclude that PER Tree decision rule induction algorithm is more appropriate for this particular type of problem domain. As a result, PERT tree decision algorithm is taken as the final working classification model to predict the Department placed of students

5.4 Evaluation of the Discovered Knowledge

Before proceeding to the final deployment, the discovered knowledge has to be thoroughly evaluated and the steps executed to construct the model reviewed to proof whether it properly achieves the research objectives. Evaluation of the discovered knowledge includes understanding the results, checking whether it is novel and useful, verifying against a given knowledge base, interpretation of the results by the domain experts, and examining the impact of the discovered knowledge on the data mining goals

Accordingly, in this study two classification techniques i.e. decision tree and rule induction were applied. Three algorithms were selected for the implementation of classification modeling namely; J48, REPTree, and PART. Then, eight experiments were conducted for each algorithm, and the obtained results were compared. Among the three algorithms PERT tree algorithm performed with highest accuracy which is 82.045%.As a result, according the result PER Tree choose as a final model for the study. Consequently, the researcher has discussed on the results with the Placement process experts to verify the discovered knowledge against a given knowledge base or what is already known in the real world practice, and then, the impact of the discovered knowledge on the data mining goals examined

Some list of rules if then format that generated some of them

Rule 1: IF First _Pref = Pre AND Nma_Al = Fail AND Sex = F THEN Pre (14/0) [2/0]

Rule 2: IF First _Pref = Pre AND Nma_Al = Good THEN Pre (620/22) [293.63/6]

Rule 3 First _Pref = Pre and Nma_Al = Fail and Sex = M and phy_Al = Poor THEN COTM (5/3) [3/1]

Rule 4 First _Pref = Pre and Nma_Al = Average and Eng_Al = Good and phy_Al = Average then Pre (202/14) [103/8]

Rule 6 IF First _Pref = Pre Nma_Al = Average AND Eng_Al =Good AND phy_Al = Poor AND Sex = M AND Bio_al = Good Chem_Al = Average : Pre (15.63/6.63) [6/4]

Rule 7 :IF First _Pref = Pre Nma_Al = Average Eng_Al = Good , phy_Al = Poor Sex = M Bio_al = Good Chem_Al = Poor : Urb (4/2) [1/0]

Rule 8 :IF First _Pref = Pre Nma_Al = Average Eng_Al = Good phy_Al = Poor Sex = M Bio_al = Good AND Chem_Al = Good : Pre (31/5) [13/2]

Rule 9 :IF First _Pref = Pre Eng_Al = Good phy_Al = Poor Sex = M Bio_al = Average : COTM (3/1) [1/0]

Rule 10 :IF First _Pref = Pre Nma_Al = Average Eng_Al = Good phy_Al = Poor Sex = F : Pre (49/1) [21/0]

Rule 11 IF First _Pref = Pre AND Nma_Al = Fail Nma_AND Al = Average AND Eng_Al = Good phy_Al = Good : Pre (65/0) [43/0]

Rule 12 IF First _Pref = Pre First _Pref = Pre Nma_Al = Average Eng_Al = Good phy_Al = Excellent: Pre (5/0) [4/0]

Rule 13 IF First _Pref = Pre Nma_Al = Average AND Eng_Al = Average Sex = M phy_Al = Average Chem_Al = Average Bio_al = Good : Urb (20/10) [9/6]

Rule 14: IF First _Pref = Pre Nma_Al = Average Eng_Al = Average Sex = M phy_Al = Average Chem_Al = Average AND Bio_al = Excellent : Pre (7/2) [4/0]

Rule 15: IF First _Pref = Pre Nma_Al = Average Eng_Al = Average Sex = M phy_Al = Average Chem_Al = Good Civic_AL = Excellent : Pre (14/0) [5/0]

Rule 15: IF First _Pref = Pre Nma_Al = Average Eng_Al = Average Sex = M phy_Al = Average Chem_Al = Average Civic_AL = Good Region < 7.5 : Pre (26/8) [18/6]

Rule 16: First _Pref = COMP ,phy_Al = Average , Civic_AL = Excellent : COMP (29/5) [15/4]

Rule 17: First _Pref = COMP phy_Al = Average Civic_AL = Average : INF (4/1) [3/0]

Rule 18: First _Pref = COMP phy_Al = Average Civic_AL = Good Eng_Al = Good : COMP (19/6) [11/3]

Rule 19: First _Pref = COMP phy_Al = Average Civic_AL = Good Eng_Al = Average : INF (5/2) [1/0]

- **Some interested run with experiment and**

Rule 2: IF First _Pref = Pre AND Nma_Al = Good THEN Pre (620/22) [293.63/6]

If the students first interested with preengineering with mathematics score Good the placed department pre engineering

Rule 3: First _Pref = Pre and Nma_Al = Average and Eng_Al = Good and phy_Al = Average then Pre (202/14) [103/8]

Students choose firsters engineering and score of mathematics is average and score English Good and physics score average then the placed department will be pre-engineering

Rule 4: IF First _Pref = Pre First _Pref = Pre Nma_Al = Average Eng_Al = Good phy_Al = Excellent: Pre (5/0) [4/0]

Students' interested with pre engineering and mathematics average and score English Good and Physics Excellent then they placed department pre-engineering

Rule 5: IF First _Pref = Pre Nma_Al = Average Eng_Al = Average Sex = M phy_Al = Average Chem_Al = Good Civic_AL = Excellent: Pre (14/0) [5/0]

The students interested in pre engineering and score mathematics Average and English average and physics average Chemistry Good and Civics excellent then the students' placed in pre engineering

Rule 6: IF First _Pref = Pre Nma_Al = Average Eng_Al = Average Sex = M phy_Al = Average Chem_Al = Average Civic_AL = Good Region < 7.5 : Pre (26/8) [18/6]

The students interested in pre engineering and score mathematics Average and English average sex male and physics average Chemistry Average and Civics Good and then region came from 1-7 the students' placed in pre engineerin

Rule 14: First _Pref = COMP ,phy_Al = Average , Civic_AL = Excellent : COMP (29/5) [15/4]

The first choose computer and physics average and Civic Excellent Good then Placed in department computer science

Rule 14: First _Pref = COMP phy_Al = Average Civic_AL = Good Eng_Al = Good : COMP (19/6) [11/3]

The first choose computer and physics average and Civic Good and English Good then Placed in department computer science

5.5 Discussion

The related research works is a study conducted by [39] whose main objectives developing new methods to discover knowledge from educational database and can used for decision making in educational system, it collected enrolled student's data from engineering institute. Use different information about their previous and current academics records like students passing percentage & other information applied method six classification (Tree Classifiers) methods on student data i.e. J48, NBTree, REPTree, Decision Stump, RandomForest and Random Tree Classification method he notice that according to experimental result J48 & Random Tree Classifiers is most suitable method for this type of student dataset.

The other research work was done by [8] whose main objective was to assess the application datamining techniques to predict the departments which the students will be placed based on the order of their preference his Experiment The study used three algorithms J48, Naïve Bayes and Random Forest to build a prediction model for placement of students the result showed the result gained from the experiments show that the best prediction model using Random Forest algorithm is more acceptable

Generally, the research works done in above not address the problem raised by the present researcher that is building a predictive model for students' placement in the side of score of and the type of subject. In this regard, this research has been conducted to fill the gaps of the previous research works with main objective to build a predictive model that the placement students. Thus, the experiments done by the current researcher have been conducted using decision tree and rule induction methods applying three algorithms: namely J48, REPTree and PART algorithms. The results obtained from the experiments showed that REPTree algorithm performed best with accuracy of 82.04% than the other models

5.6 Use of the Discovered Knowledge: The Predictive Model

Once a data mining model is built and validated, it can be used in the first way is for a Domain expert to recommend actions based on simply viewing the model and its results. For example, the Expert may look at the classification the model has identified, the rules that define the model, or ROI charts that depict the effect of the model and the discovered knowledge always needs to

be organized and presented in a way that the end users can utilize it [15]. As a result, the above discovered knowledge or interesting rules could be used for identifying students' placed department. The organization can incorporate these newly generated rules among with those currently used information for the prediction

5.7 Findings

Findings can be defined as anything apart from the model that is important in meeting objectives of the business or important in leading to new questions, or side effects e.g. data quality problems uncovered by the data mining exercise. Although the model is directly connected to the business objectives, the findings need not be related to any questions or objectives, but are important to the initiator of the study.

Consequently, the following important research findings were observed during the course of data mining model building and discussed as follows

- The findings observed in the interpreted rules indicate that the algorithms used for building the models. Assumed that all attributes of the targeted dataset were used for the construction of decision rule sets. However, some attributes like use of other than first preference order, Region , school of type and age have less effect on placed department

- Use of placement of process can easily to use with students'. The information of their source can get them hands to predict the placed department. According to full the criteria for academicals history of the students for placement.

- Use the placement Department of the university placed capable of the students' were placed. It tells detail performance of students'.
- Finally, since no field data is required for the newly developed predictive model, it gives precise solution it alternative way of facilitate the placement process easily by students' themselves. It faced for the problems faced by the higher educational institute optimize the capacity of department collection method every Academic Year. This is because the model works with the predicting Department

Chapter Six

Summary, Conclusions and Recommendations

6.1 Summary

The main objective of this research was to explore the students' placement in to difference Available department/field and identify important and interesting patterns from the academic history of the students 'record.

This investigation has been carried out using hybrid data mining model processing methodology. Which consists of problem understanding, data understanding, data preparation, data mining, evaluation of the discovered knowledge, and finally use of the discovered knowledge. The iterative nature of the process assisted the data miner to go back and forth, at different steps of the study, to fix the problem wherever it is identified.

The targeted data set used for this study was the students' records in the placement process of Addis Ababa university Registrar office in the year 2015/16 and 2016/17. The original dataset consisted of 34 attribute and 11320 cases from this make the data manageable 4283 cases were sampled and 11 attribute selected for the data mining exercise. The major task undertaken was data preparation to make the initial data suitable for the data mining tasks which took much time of the study works.

In the experimentation of the study two classification techniques i.e. decision tree and rule induction. For the implementation of these modeling techniques three classification algorithms (J48, REPTree and PART) were used. The experiments were conducted using six scenarios for each algorithm and the outputs of the experiments were used for comparison of the accuracy of the models. In addition, by changing the training test options and the default parameter values of the algorithms models were tested and evaluated. As a result, the test run on PERTREE algorithm registered better accuracy which is 82.045% without pruning and keeping others algorithm parameter values as default. The generated decision rules were analyzed and interpreted. The discovered knowledge was evaluated based on the existing knowledge base and the domain expert validation.

6.2 Conclusion

The obtained results of this study deal with students' placement in to different department prediction model use academicals history of the students' prediction ,discovering important knowledge and interesting patterns from the existing data, lead to the following conclusions:

- The obtained results show a significant trend or patterns available in the targeted dataset. Moreover, attributes like subject interested field and type of use to predicted placed department is one of the criteria for placement.

- According to the result the subject of score can request performance subject is required like pre engineering request score good result in mathematics and Computer science English performer is needed

- Generally, encouraging results were obtained by employing decision tree and rule induction technique, and the decision rules generated by PERTree algorithm are easily understandable by the domain experts. Thus, the results obtained in this study have proved the applicability of data mining technology to build predictive model for placement process by employing decision tree and rule induction methods.

- More attention has to be given to placement process for the sake of department placement use an attribute performance of each subject in the process it good in the have of the students' motivation and the also department easy to predicate their capacity using visible in order to obtain high quality data.

Generally, the experimentation results obtained from this research work show that the final output of the study could be used to investigate the applicability of data Mining technology in developing a predictive model for such type of dataset

6.3 Recommendations

This research work has been conducted mainly for academic achievement. However, the researcher strongly believes that the findings of the study can be used by the concerned organizations to further investigate their data quality problems.to choose appropriate data analysis methods, techniques and tools that are currently in use for processing Students placement process model hence, based on the findings of this study, the following recommendations are forwarded:

- There were data contain high school results is difficult to find the big problem and affect performance. The other missing value hence, the effects of those omitted variables require further investigation.

- In the process of this study works, it was learnt that more study will be needed to enable the full exploitation of the data mining technology in related placement process of in the higher institute. Hence, the following areas were identified as deserving further research work:

- In this study, an attempt has been made to assess the applicability of data mining technology to predict Department of the Students by using some set of variables that were considered as more important by the domain experts. However, it remains to investigate further the effect of those future performance and psychological of the students' because use and variables, due to their high missed values, to build models with better accuracy and performance than the models built in this study.

- This research work has not be considered factor of interested department and performance students 'though the university.

7 References

[1] D. M. D. Angeline, "Association Rule Generation for Student Performance Analysis using Apriori Algorithm," *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), ,* vol. 1, no. 1, 2013.

[2] P. P. Sondwale, "Overview of Predictive and Descriptive Data Mining Techniques," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 5, no. 4, 2015.

[3] M. M. T. A. M.El-Haees, "Mining Eduactional Data To imrove Students perfomance Acase Studies," *information and Communication Technology Technology ,* vol. 2, no. 2, 2012.

[4] M. M. Ali, "Role of data mining in Education sector," *Computer Science and Mobile Computing,* vol. 2, no. 4, pp. 374-383, 2013.

[5] M. G. a. R. Vohra, "Applications of Data Mining in Higher Education," *Computer Science Issues,* vol. 9, no. 2, 2012.

[6] M. A. Yehuala, "Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University," *SCIENTIFIC & TECHNOLOGY RESEARCH,* vol. 4, no. 4, 2015.

[7] D. R. V. Praveen Rani, "Generating Placement Intelligence in Higher Education Using Data Mining," *Computer Science and Information Technologies,* vol. 6, no. 3, 2016.

[8] V. S. Getaneh Berie Tarekegn, "Application of Data Mining Techniques to Predict Students Placement in to Departments," *Research Studies in Computer Science and Engineering,* vol. 3, no. 2, 2016.

[9] T. Dejenie, "Student's Placement by First Choice and Without at Mekelle University: The Impact on Academic Performance," *Ethiop. J. Educ. & Sc.,* vol. 6, no. 1, 2010.

[10] A. G. A. D. R. J. a. V. H. Kalpesh Adhatrao, "PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS," *Data Mining & Knowledge Management Process,* vol. 3, no. 5, 2013.

[11] E. F. Ian H. Witten, Data mining Practical Machine Learning Tools and Techniques, San Francisco: Elsevier Inc, 2005 .

[12] G. P.-S. a. P. S. Usama Fayyad, " From data minig to knowldge discovery data base," *AI Magazine,* vol. 17, no. 3, 1996.

[13] S. P. ,. R. K. P. Neha Purohit, "Data Mining, Applications and Knowledge Discovery," *Advanced Computer Research,* vol. 2, no. 6, 2012.

[14] M. S. Pratiyush Guleria, "Data mining in Education : Areview on the knowledge DIscery Perspective," *Data Mining & Knowledge Management Process,* vol. 4, no. 5, 2014.

[15] A. Arora, "Data mining Techniques and tools For knowledge Discovery in Agricultural Datasets," IASRI, 2011.

[16] W. P. W. S. A. K. Krzysztof J. Cios, Data Mining A Knowledge Discovery Approach, USA : Springer Science+Business Media, 2007.

[17] T. D. Akal, "Constructing Predictive model for Netework Induction Detection," Addis ababa, 2015.

[18] L. j. Nikhil R.pal, Advanced Techniques in Knowledge Discovery and Data Mining, Springer , 2005.

[19] M. K. Jiawei Han, Data Mining:Concepts and Techniques, Elsevier Inc., 2006.

[20] J. R. Lalit Dole, "A Decision Support System for Predicting Student Performance," *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 2, no. 12, 2014.

[21] R. S.J.d.Baker, "Datamining For Education," *Pittysburgh pennsyvania,* p. 14, 2014.

[22] S. P. Ajay Kumar Pal, "Classification Model of Prediction for Placement of Students," *I.J.Modern Education and Computer Science,* no. 11, p. 8, 2013.

[23] M. S. Edin Osmanegovic, "Data mining Aproach for pridicting Stunents Performance," *Journal of Economical and Business ,* no. 1, p. 10, 2012.

[24] R. R. M. Ramaswami, "Student Performance Prediction Modeling A Bayesian Network Approach," *Computational Intelligence and Informatics,* vol. 4, no. 1, 2012.

[25] S. P. Brijesh Kumar Baradwaj, "Mining Educational Data to Analyze Students‟ Performance," *Journal of Advanced Computer Science and Applications,* vol. 2, no. 6, 2011.

[26] T. C. Corporation, "Introduction to Data Mining and knowldge Discovery," *Two Crows Corporation,* no. 3, 2005.

[27] s. m. L. ,. m. Hitarthi bhatt, "use ID3 Decision Tree Algorithn For placement Prediction," *Computer Science and Information Technology ,* vol. 6, no. 5, 2015.

[28] R. S. Baker, "Data mining of education," *Encyclopedia ,* no. 3.

[29] J. Jackson, "DATA MINING: A CONCEPTUAL OVERVIEW," *Communications of the Association for Information Systems,* vol. 8, 2002.

[30] C. R. a. S. Venture, "Data mining in Education," *Wire Data mining Knowl Discov ,* no. 3, 2013.

[31] Q. Y. W. T. Zachary A. Pardos, "The real world significance of performance prediction," in

International Conference on Educational Data Mining.

[32] S. A. .. A. J. K. A. Kolo David Kolo, "A Decision Tree Approach for Predicting Students Academic Performance," *Education and Management Engineering,* vol. 5, no. 2, 2015.

[33] N. G. H. A. A. Raheela Asif, "Prediction of Undergraduate Student's Performance using Data Mining Methods," *Computer Science and Information Security,* vol. 14, no. 5, 2016.

[34] S. P. D. G. G. Ramanathan.L, "Mining Educational Data for Students' Placement Prediction using Sum of Difference Method," *Computer Applications,* vol. 99, no. 18, 2014.

[35] R. T. a. A. K. Sharma, "A Detailed Analysis of Educational Data Mining," *RESEARCH IN EMERGING SCIENCE AND TECHNOLOGY,* vol. 2, no. 7, 2015.

[36] L. C. S. M.-M. A. Pedro Strecht, "A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance".

[37] S. P.-A. R. B. Naeimeh DELAVARI, "Data Mining Application in Higher Learning Institutions," *Informatics in Education,* vol. 7, no. 1, 2008.

[38] A. K. S. Ravi Tiwari, "A Data Mining Model to Improve Placement," *Computer Applications,* vol. 120, no. 12, 2015.

[39] V. K. Samrat Singh, "Performance Analysis of Tree Classifiers using on Engineering Student's Recruitment Dataset," *Computing and Technology,* vol. 1, no. 2, 2014.

[40] N. A. S. K. T. Jeevalatha, "Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms," *Computer Applications,* vol. 108, no. 15, 2014.

[41] A. U. Nihat Cengiz, "Prediction of Student Success Using Enrolment Data".

[42] A. A. Fiseha Berhanu, "Students' Performance Prediction based on their Academic Record," *Computer Applications,* vol. 131, no. 5, 2015.

[43] M. S. P. S. Bakare, "Prediction of Campus Placement Using Data Mining Algorithm-Fuzzy logic and K nearest neighbor," *Advanced Research in Computer and Communication Engineering,* vol. 5, no. 6, 20016.

[44] E. V. R. Sumitha, "Prediction of Students Outcome Using Data Mining Techniques," *Scientific Engineering and Applied Science,* vol. 2, no. 6, 2016.

[45] S. G. Mansi Gera, "Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity," *Computer Applications,* vol. 113, no. 18, 2015.

[46] Y. Z. Y. Guandong Xu, Applied Data Mining, London: CRC, 2013.

[47] J. C. (. R. K. (. Pete Chapman (NCR), Step-by-step data mining guide, SPSS Inc, 2000.

[48] T. B. Gossa, "BUILDING A PREDICTIVE MODELFOR ANNUAL CEREAL CROPS PRODUCTION USING DATA MINING TECHNIQUES," Gondor Ethiopia , 2014.

[49] J. C. (. R. K. (. K. (. T. R. (. Pete Chapman (NCR), "CRISP-DM 1.0," *SPSS Inc,* 2000.

[50] L. C. L. a. A. A. J. Perez, "Educational Data mining and Learning Analytics Difference Similarities and Time evolution," *Revista De UniveridadSociedad Ddel Conccimiento,* vol. 12, no. 3, 2015.

[51] M. S.-R. Mashael A.Al-Barrak, "Predicting Stunents performace though Classification :Acase Studies," *Theortical and Applied information Technology,* vol. 75, no. 2, 2015.

[52] K. D.Kabakchievad, "Analyzing University Data for Determining Student Profiles and Predicting Performance".

[53] P. P. Sondwale, "Overview of Predictive and Descriptive Data Mining Techniques," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 5, no. 4, 2015.

8 Annexes

Annex 1: List of Attributes Name, Data Type and their Description

Attribute	type	Distribution
Adm.no	Numeric	Admission Nubmer
S.code	Numeric	School code
Sex	char	Sex
Age	Numeric	Age
Sig	char	visualization
Nation	char	nationality
Eng	Numeric	Score in English subject
NMa	Numeric	Score in natural mathematics subject
Apt	Numeric	Score in Aptitude subject
Phy	Numeric	Score in physics subject
Che	Numeric	Score in chemistry subject
Bio	Numeric	Score in Biology subject
Geo	Numeric	Score in Geography subject
His	Numeric	Score in history subject
Eco	Numeric	Score in economics subject
SMa	Numeric	Score in soccial mathematics subject
Civ	Numeric	Score in civis subject
T.Mark	Numeric	Total mark
Reg	String	Region
Benefit Group	String	Reason n for placement
Full Name	String	name the students’
Field Of study	String	field of the studies
Band hoice No	String	numeric choose with selected bad
Choose pref	String	order of the department choose by students
Palced _Dept		placement of department

Annex 2 Descriptive statistics for with Band/field distribution of the students'

Field of studies	Eng	NMa	Apt	Phy	Che	Bio	Geo	His	Eco	SMa	Civ
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
Business and Economics	65	-	64	-	-	-	79	58	70	47	78
Computer Sciences	74	60	74	57	70	83	-	-	-	-	77
Dental Medicine	74	70	76	67	79	90	-	-	-	-	77
Engineering Sciences	70	59	70	54	72	82	-	-	-	-	76
Law	68	-	63	-	-	-	82	64	74	41	81
Medicine	80	75	81	73	85	93	-	-	-	-	86
Natural and Computational Sciences	49	42	52	39	57	65	-	-	-	-	62
Other Agricultural and Natural Resource	51	47	54	44	63	69	-	-	-	-	67
Other Health Sciences	68	58	67	52	72	83	-	-	-	-	76
Other Social Sciences and Humanities	49	-	50	-	-	-	61	43	51	37	65
Theatrical, Fine Arts and Music	47	-	53	-	-	-	63	44	53	39	64
Veterinary Medicine	58	50	60	47	69	77	-	-	-	-	74
Veterinary science	56	46	59	44	65	72	-	-	-	-	69

Annex 3 The first 20th the first presence of the students'

Field of studies	first _pref
Pre-Engineering	2834
Bachelor of Arts in Social Work	572
Bachelor of Arts in Sociology	440
Bachelor of Science in Biology	357
Bachelor of Arts in Accounting and Finance (CoBE)	351
Bachelor of Science in Computer Science	314
Bachelor of Arts in Economics (CoBE)	295
Bachelor of Science in Geology	285
Bachelor in Pharmacy	190
Bachelor of Science in Construction Technology and Management	187
Bachelor of Arts in Management (CoBE)	156
Bachelor of Arts in Accounting and Finance (Commerce)	141
Bachelor of Arts in Political Science and International Relations	125
Bachelor of Science in Sport Science	124
Bachelor of Arts in Geography and Environmental Studies	116
Bachelor of Science in Anesthesia	104
Bachelor of Science in Agricultural Economics	75
Bachelor of Science in Statistics	73
Bachelor of Arts in Business Administration and Information Systems (Commerce)	72
Bachelor of Arts in Afan Oromo and Literature	70
Bachelor of Arts in Marketing Management (Commerce)	70
Bachelor of Arts in Economics (Commerce)	68
Bachelor of Science in Plant Science	66
Bachelor of Science in Chemistry	63
Bachelor of Arts in Journalism and Communication (Broadcast Journalism)	61
Bachelor of Arts in Management (Commerce)	60
Grand Total	**7269**

Annex 4 The first 20th the placed department in two academic year

Field of studies placed department	Frequency
Pre-Engineering	1518
Doctor of Medicine	363
Bachelor of Arts in Theatre Arts	315
Bachelor of Science in Construction Technology and Management	299
Bachelor of Science in Urban and Regional Planning	189
Bachelor of Arts in Accounting and Finance (CoBE)	166
Bachelor of Arts in Management (CoBE)	157
Bachelor of Science in Biology	148
Bachelor of Arts in Economics (CoBE)	135
Bachelor of Laws	116
Bachelor of Science in Architecture	104
Bachelor of Science in Computer Science	103
Bachelor of Arts in Public Administration & Development Management (CoBE)	95
Bachelor of Arts in Social Work	90
Bachelor of Science in Information Systems	90
Bachelor of Science in Sport Science	85
Bachelor of Science in Mathematics	84
Bachelor of Science in Geology	79
Bachelor of Arts in Accounting and Finance (Commerce)	77
Bachelor of Science in Statistics	76
Bachelor of Arts in Special Needs and Inclusive Education	69
Bachelor of Science in Chemistry	69
Bachelor of Arts in Geography and Environmental Studies	67
Bachelor of Arts in Political Science and International Relations	66
Bachelor of Arts in Psychology	66
Bachelor of Arts in Sociology	65
Bachelor of Science in Physics	60
Bachelor of Arts in Ethiopian Language(s) Literature- Amharic	52

Annex 5: Sample Test Run Results generated by J48 Algorithm Model

=== Run information ===

J48 pruned tree

First _Pref = Pre

| Sex = M

| | Nma_Al = Good: Pre (698.62/27.0)

| | Nma_Al = Fail

| | | Apt_Al = Good

| | | | phy_Al = Average: Pre (7.0/1.0)

| | | | phy_Al = Fail: Pre (0.0)

| | | | phy_Al = Poor: COTM (7.0/3.0)

| | | | phy_Al = Good: Pre (4.0/2.0)

| | | | phy_Al = Excellent: Pre (0.0)

| | | Apt_Al = Poor: COTM (1.0)

| | | Apt_Al = Average: Pre (6.0/2.0)

| | | Apt_Al = Excellent: Pre (0.0)

| | | Apt_Al = Fail: Pre (0.0)

| | Nma_Al = Average: Pre (845.62/190.62)

| | Nma_Al = Poor

| | | Third_Pref = urb: COTM (9.0/4.0)

| | | Third_Pref = Math: Pre (0.0)

| | | Third_Pref = Urb

| | | | Civic_AL = Excellent: Pre (54.0/13.0)

| | | | Civic_AL = Average: Urb (17.0/7.0)

| | | | Civic_AL = Good

| | | | | phy_Al = Average

| | | | | | Bio_al = Good: Pre (33.0/11.0)

| | | | | | Bio_al = Average: Urb (4.0/1.0)

| | | | | | Bio_al = Excellent: Pre (28.0/13.0)

| | | | | | Bio_al = Poor: Pre (0.0)

| | | | | | Bio_al = Fail: Pre (0.0)

| | | | | phy_Al = Fail: Urb (2.0)

| | | | | phy_Al = Poor

| | | | | | Apt_Al = Good

| | | | | | | Chem_Al = Average: COTM (16.0/7.0)

| | | | | | | Chem_Al = Fail: Pre (1.0)

| | | | | | | Chem_Al = Poor: Urb (2.0)

| | | | | | | Chem_Al = Good: Pre (9.0/3.0)

| | | | | | | Chem_Al = Excellent: Pre (0.0)

| | | | | | Apt_Al = Poor: COTM (1.0)

| | | | | | Apt_Al = Average: Urb (17.0/7.0)

| | | | | phy_Al = Good: Pre (10.0/3.0)

| | | | | phy_Al = Excellent: Pre (0.0)

| | | Third_Pref = COTM: Urb (15.0/5.0)

| | Nma_Al = Excellent : Pre (157.0/4.0)

| Sex = F: Pre (716.0/5.0)

First _Pref = Bio: Bio (347.16/126.16)

First _Pref = chem: chem (64.03/10.03)

First _Pref = Arch: Arch (175.08

Number of Leaves : 101

Size of the tree : 122

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 3496 81.625 %

Incorrectly Classified Instances 787 18.375 %

Kappa statistic 0.7191

Mean absolute error 0.0417

Root mean squared error 0.1485

Relative absolute error 39.5177 %

Root relative squared error 64.6388 %

Total Number of Instances 4283

== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.989	0.154	0.884	0.989	0.933	0.958	Pre
0.898	0.032	0.633	0.898	0.743	0.968	Bio
0.496	0.005	0.737	0.496	0.593	0.948	chem
1	0	1	1	1	1	Arch
0.48	0.008	0.849	0.48	0.613	0.901	COTM
0.527	0.005	0.756	0.527	0.621	0.98	INF
0.144	0.006	0.479	0.144	0.221	0.814	Urb
0.887	0.015	0.739	0.887	0.807	0.982	COMP
0.8	0.007	0.775	0.8	0.787	0.969	Spo
0.229	0.006	0.5	0.229	0.314	0.895	Math
0.267	0.002	0.742	0.267	0.393	0.903	Phy
0.68	0.02	0.506	0.68	0.58	0.959	Geol
0.589	0.012	0.593	0.589	0.591	0.951	stat
Weighted Avg. 0.816	0.088	0.806	0.816	0.795	0.949	

Annex 6: Sample Test Run Results Generated by the Selected PART Model

=== Classifier model (full training set) ===

PART decision list

First _Pref = Pre AND

Sex = F: Pre (716.0/5.0)

First _Pref = Pre AND

Nma_Al = Good: Pre (698.55/27.0)

First _Pref = COTM: COTM (180.09/10.09)

First _Pref = Arch: Arch (175.09/0.09)

First _Pref = chem: chem (64.03/10.03)

First _Pref = INF: INF (48.02/3.02)

First _Pref = Math: Math (23.01/0.01)

First _Pref = Phy: Phy (22.01/0.01)

First _Pref = COMP AND

Eng_Al = Excellet AND

Civic_AL = Good: COMP (22.0/2.0)

First _Pref = COMP AND

Civic_AL = Excellent AND

Apt_Al = Excellent: COMP (26.0/1.0)

First _Pref = COMP AND

Eng_Al = Good AND

Civic_AL = Excellent AND

Nma_Al = Good: COMP (47.04/8.04)

First _Pref = COMP AND

Eng_Al = Good AND

Nma_Al = Average AND

Chem_Al = Good AND

Apt_Al = Good: COMP (20.0)

First _Pref = COMP AND

Number of Rules : 113

Time taken to build model: 0.1 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	3474	81.1114 %
Incorrectly Classified Instances	809	18.8886 %
Kappa statistic	0.7192	
Mean absolute error	0.0367	
Root mean squared error	0.145	
Relative absolute error	34.7436 %	
Root relative squared error	63.1315 %	
Total Number of Instances	4283	

== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.958	0.108	0.913	0.958	0.935	0.973	Pre
0.837	0.027	0.658	0.837	0.737	0.969	Bio
0.504	0.009	0.613	0.504	0.553	0.906	chem
1	0	1	1	1	1	Arch
0.563	0.022	0.708	0.563	0.627	0.927	COTM
0.674	0.012	0.635	0.674	0.654	0.989	INF
0.35	0.016	0.467	0.35	0.4	0.858	Urb
0.744	0.01	0.775	0.744	0.759	0.988	COMP
0.808	0.008	0.748	0.808	0.777	0.938	Spo
0.266	0.009	0.433	0.266	0.33	0.887	Math
0.291	0.004	0.595	0.291	0.391	0.822	Phy
0.672	0.017	0.551	0.672	0.606	0.96	Geol
0.605	0.013	0.591	0.605	0.598	0.902	stat
Weighted Avg. 0.811	0.065	0.801	0.811	0.802	0.956	

Annex 7: Sample Test Run Results Generated by the Selected PART Model

=== Run information ===

REPTree

First _Pref = Pre

| Nma_Al = Good : Pre (620/22) [293.63/6]

| Nma_Al = Fail

| | Sex = M

| | | phy_Al = Average : Pre (5/1) [5/2]

| | | phy_Al = Fail : Pre (1/0) [1/0]

| | | phy_Al = Poor : COTM (5/3) [3/1]

| | | phy_Al = Good : Pre (4/2) [1/0]

| | | phy_Al = Excellent : Pre (0/0) [0/0]

| | Sex = F : Pre (14/0) [2/0]

| Nma_Al = Average

| | Eng_Al = Good

| | | phy_Al = Average

| | | | Civic_AL = Excellent : Pre (52/0) [31/1]

| | | | Civic_AL = Average

| | | | | Age group = G-1

| | | | | | Apt_Al = Good : Pre (3/0) [4/0]

| | | | | | Apt_Al = Poor : Pre (3/1) [0/0]

| | | | | | Apt_Al = Average : COTM (2/0) [3/1]

| | | | | | Apt_Al = Excellent : Pre (0/0) [0/0]

| | | | | | Apt_Al = Fail : Pre (0/0) [0/0]

| | | | | Age group = G-2 : Pre (7/0) [2/0]

| | | | Civic_AL = Good : Pre (134/10) [63/4]

| | | | Civic_AL = Poor : COTM (1/0) [0/0]

| | | | Civic_AL = Fail : Pre (0/0) [0/0]

| | | phy_Al = Fail : Pre (7/4) [4/1]

| | | phy_Al = Poor

| | | | Sex = M

| | | | | Bio_al = Good

| | | | | | Chem_Al = Average : Pre (15.63/6.63) [6/4]

Size of the tree : 193

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	3514	82.0453 %
Incorrectly Classified Instances	769	17.9547 %
Kappa statistic	0.7253	
Mean absolute error	0.0409	
Root mean squared error	0.1459	
Relative absolute error	38.7363 %	
Root relative squared error	63.5317 %	
Total Number of Instances	4283	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.991	0.154	0.884	0.991	0.934	0.963	Pre
0.894	0.033	0.625	0.894	0.736	0.971	Bio
0.487	0.003	0.821	0.487	0.611	0.956	chem
1	0	1	1	1	1	Arch
0.485	0.011	0.809	0.485	0.607	0.903	COTM
0.581	0.006	0.758	0.581	0.658	0.99	INF
0.125	0.003	0.645	0.125	0.209	0.832	Urb
0.877	0.013	0.76	0.877	0.814	0.993	COMP
0.792	0.006	0.786	0.792	0.789	0.981	Spo
0.22	0.002	0.727	0.22	0.338	0.924	Math
0.256	0	1	0.256	0.407	0.904	Phy
0.875	0.031	0.469	0.875	0.61	0.973	Geol
0.508	0.007	0.692	0.508	0.586	0.953	stat
Weighted Avg. 0.82	0.088	0.825	0.82	0.798	0.955	

DECLARATION

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

_____

Meseret Getachew Weyecha

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

_____

Wondwossen Mulugeta (PhD) Advisor

June 2017