



**Sudan University of Science and Technology**  
**College of Graduate Studies**  
**College of Computer Science and Information Technology**



# **Mining Students' Data to Predict and Evaluate Their Academic Performance**

**( Case study: Faculty of Science, University of Nyala )**

**تنقيب بيانات الطلاب للتنبؤ بأدائهم الأكاديمي وتقييمه**

**( دراسة الحالة: كلية العلوم، جامعة نيالا )**

**A Dissertation Submitted in Partial Fulfillment of the Requirements of Master Degree  
in computer Science, Faculty of Computer Science and Information Technology, Sudan  
University of Science and Technology**

**BY:**

**Ibtihag Fedil Haroun Ali**

**Supervisor:**

**Dr. Khalid Hassan Mohamed Edris**

**March 2019**

## الآية

قال تعالى: ( اقرأ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ (1) خَلَقَ الْإِنْسَانَ مِنْ عَلَقٍ (2) اقرأ وَرَبُّكَ الْأَكْرَمُ (3)  
الَّذِي عَلَّمَ بِالْقَلَمِ (4) عَلَّمَ الْإِنْسَانَ مَا لَمْ يَعْلَمْ (5) )

سورة العلق الآيات (1 - 5)

## DEDICATION

*All praise to Allah, today we fold the day's tiredness and the errand summing up between the cover of this humble work.*

*To the utmost knowledge lighthouse, to our greatest and most honored prophet Mohamed - may peace from Allah be upon him.*

*To those precious to my heart ... mother, father, brother and my husband.*

*To my supervisor and Faculty of Science and Information Technology.*

*I dedicate this work to them all for the support, encouragement, love and prayers that they have always had for us. My Allah blesses them all and grants them happiness all through.*

## **ABSTRACT:**

Recently, institutions of Sudanese higher education testify a technological development in field of computerization management of student information electronically. Many data mining techniques are proposed to extract the hidden knowledge from educational data. Therefore, the educational data mining make it possible to extract these educational data and find information that assistant supporting both teachers and students. Moreover, the extracted knowledge and predicting performance helps the institutions to improve their teaching methods and learning process. These improvements lead to enhance the performance of the students and educational outputs. The aim of this study to design model to predicting the student's performance based on data mining techniques. The data mining tool used to evaluated performance of student's called Weka. The data consists of socio-economic, demographic and academic information included six hundred undergraduate students with eleven attributes .Classification task method was used the classifier tree j48, to predict the final academic results and grades of students in first year, Apriori algorithm was also applied to find the association rule mining among all the attributes and the best rules were also displayed. The results showed that classification process succeeded in training set. Thus, the predicted instance is similar to the training set. This proves the suggested classification model. The algorithm efficiency and effectiveness of j48 algorithm in predicting the academic results, grades, was very good. Furthermore recommend Test other algorithms to design a predictive model ,other than j48 and Increase the size of student data to create another model and compare it with the designer model.

## المستخلص :

في الآونة الأخيرة ، شهدت مؤسسات التعليم العالي السودانية تطوراً تكنولوجياً في مجال إدارة و حوسبة معلومات الطلاب. هنالك العديد من تقنيات التنقيب عن البيانات التي تستخدم لاستخلاص المعلومات التي تساعد متخذي القرار . بالتالي فإن التنقيب عن البيانات والتنبؤ بأداء الطلاب يساعدان مؤسسات التعليم العالي على تقييم أساليب التدريس مما يسهم في العملية التعليمية،فذلك يساعد علي تحسين أداء الطلاب والمخرجات التعليمية. هدفت هذه الدراسة الي تصميم نموذج للتنبؤ بأداء الطلاب اعتمادا على تقنيات تنقيب البيانات. أداة التنقيب عن البيانات التي استخدمت لتقييم أداء الطالب تسمى Weka . شملت الدراسة بيانات ستمائة طالب بمرحلة البكالوريوس، احتوت هذه البيانات علي معلومات اجتماعية واقتصادية وأكاديمية لأحد عشر سنة . تم استخدام تقنية التصنيف تطبيقاً علي خوارزمية 48للتنبؤ بالنتائج الأكاديمية لدرجات الطلاب في السنة الأولى ، بالإضافة الي ذلك تم تطبيق خوارزمية Apriori للحصول علي قواعد منطقية بين السمات المختارة ومدى ارتباطها لمعرفة أسباب الرسوب، ومن ثم تم اختيار أفضل القواعد. أظهرت نتائج الدراسة أن عملية التصنيف نجحت في مجموعة التدريب. وبالتالي ، فإن الحالات المتوقعة مماثلة لمجموعة التدريب . وهذا يثبت كفاءة فعالية الخوارزمية 48j المختبرة في التنبؤ بالنتائج الأكاديمية. توصي الدراسة بمزيد من البحث في هذا المجال والعمل علي خوارزميات أخرى ومقارنتها.

## List of contents

Title	Page No.
الآية	I
DEDICATION	II
Abstract	III
المستخلص	IV
List contents	V
List of Tables	VII
List of Figures	VIII
<b>CHAPTER ONE: INTRODUCTION</b>	1
1.1 Introduction	1
1.2 Problem statement	2
1.3 Important of research	2
1.4 Hypotheses	3
1.5 Objectives of the study	3
1.6 Bounders of research	3
1.7 Contributions	4
1.8 Layout of the research	5
<b>CHAPTER TWO: LITERATURE REVIW</b>	6
2.1 Background of data mining	6
2.1.1 Classification	8
2.1.2 Classification techniques used in Data Mining	8
2.1.3 Classification Algorithms	9
2.1.4 Clustering	11
2.1.5 Association rule	12

2.1.6 Regression	12
2.2 Background of educational data mining(EDM)	13
2.2.1 Goals of educational data mining in higher education	14
2.2.2 Educational data mining methods	15
2.3 Application of data mining	16
2.4 Overview of tools for data mining	17
2.5 WEKA tool	18
2.6 Related work	19
<b>CHAPTER THREE: MATERIALS AND METHODS</b>	21
3.1 Methodology	21
3.1.1 Data collection step	22
3.1.2 Selection Step	22
3.1.3 Pre-processing step	23
3.1.4 Transformation step	23
3.1.5 Data Mining Step	23
3.2 Data Mining Method	23
<b>CHAPTER FOUR: RESULTS AND DISCUSSION</b>	25
4.1 Implementation in Weka tool	25
4.1.1 Classification	25
4.1.2 Classification model implementation	33
4.2 Classification without filtering the attributes	35
4.3 Source code of java	39
4.4 Association rule	39
4.5 Results	42

4.6 Discussion of results	43
<b>CHAPTER FIVE: CONCLUSION AND FUTURE WORK</b>	45
Conclusion and future work	45
Conclusion	45
Future work	45
<b>References</b>	46
<b>Appendix</b>	49

### List of Tables

Title	Page No
Table 1.1 Research scheduling	4
Table 4.1 Performance comparison between different algorithms	33
Table 4.2 Summary of Results Obtained	41



## List of Figures

<b>Title</b>	<b>Page No</b>
Figure 1.1 University of Nyala, South Darfur State.	4
Figure 2.1 Knowledge discovery process in data mining.	7
Figure 3.1 Work methodology	22
Figure 4.1 Front view of Weka tools	25
Figure 4.2 Weka explorer	26
Figure 4.3 Load data set in to the Weka	27
Figure 4.4 Visualize all attributes	28
Figure 4.5 Discretizes filter	29
Figure 4.6 Filter the dataset	30
Figure 4.7 J48 Classifier by default parameter	31
Figure 4.8 J48 Classifier after change min number of obj	32
Figure 4.9 Classifier tree visualize	32
Figure 4.10 Test dataset	34
Figure 4.11 Result of test dataset on the model	35
Figure 4.12 Dataset uploaded without filtering	36
Figure 4.13 Visualize all Attributes	36

Figure 4.14 Apply J48 classifier	37
Figure 4.15 Confusion matrix of classifier	37
Figure 4.16 Visualize J48 classifier	38
Figure 4.17 J48 Classifier tree	38
Figure 4.18 Explorer interface in Weka	40
Figure 4.19 Format of the results after applying Apriori algorithm	40
Figure 4.20 Some of result	41

**CHAPTER ONE**  
**INTRODUCTION**

# CHAPTER ONE

## INTRODUCTION

### 1.1 Introduction:

The Sudanese public higher education institutions comprise of 30 Universities and the private higher education institutions comprise of 13 Universities and 60 colleges. University of Nyala is one of Sudanese public University, it located in South Darfur State. It was established on March 6th 1994. Recently, it had 7 faculties. One of them faculty of Science and Information Technology, its established on 2014 to 2015 include 8 departments; chemistry, physics, zoology, botany, computer science, information technology, mathematics and geology (Ille, 2017). In last decade, the number of higher education Universities and Institutions have proliferated manifolds. Large numbers of graduates and post graduates are graduated by them every year. Universities and Institutes may follow best of the pedagogies; but still they face the problem of dropout students, low achievers and unemployed students, understanding and analyzing the factors for poor performance is a complex and incessant process hidden in past and present information congregated from academic performance and student's behavior. Powerful tools are required to analyze and predict the performance of students scientifically. Although, Universities and Institutions collect an enormous amount of students data, but this data remains unutilized and does not help in any decisions or policy making to improve the performance of students.

In recent years, the goal of any educational institution is to graduate qualified students from different disciplines and to achieve this goal must focus on the performance of the student during the school stage, information technology facilitated the storage of large volumes of data in different forms, which led to an increase in the size of the educational database. With the plenty of existing data stored in so-called databases, has become objectify of Many researchers have questioned the use of data warehousing, and it has become necessary to find techniques, methods and means to extract information and knowledge from such as this data stacked and exploited in problem solving and decision-making, using modern computer applications ,which is a smart modern technology that blindly makes the computer "think as human thought and do man" and so on what is known as artificial

intelligence, the idea of uncovering and excavating this data in intelligent ways to help in problem solving and decision-making. Data mining is global processes that blends artificial intelligence and statistics and generalize the machine and databases, and is a step of exploring knowledge from the databases.

The University of Nyala has a lot of data for students, and this data is increasing day by day. There is no good utilization of these data so that the university can extract knowledge that helps to know the factors that affect students' performance. The student is evaluated at the university through tests, seminars and final exam, So with this huge amount of data to be used well, so there is a tool to help analyze this huge data to benefit from the future called Weka , through mining of data.

Data mining is a process which finds useful patterns from large amount of data. The process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions (Han et al., 2011b). Data Mining is a multidisciplinary field, encompassing areas like information technology, machine learning, statistics, pattern recognition, data retrieval, neural networks, information based systems, artificial intelligence and data visualization(Luan, 2002). Data mining includes four essential methods: classification, association, clustering, and outlier detection.

### **1.2 Problem statement:**

In recent years, the students ration admit in Sudanese Universities are higher compared to low academic performance in university. This issue play potential role in Sudanese Universities; therefore it is very difficult to understand reasons show to these problems. In particular, this problem was found at University of Nyala, Faculty of Science and Information Technology. That accepted students with high rates in contrast to high fail on the first year, so this research is going to look on the reasons that why students' get high marks in Sudanese Certificate but acquire weak marks in the first year exams in University.

### **1.3 Important of research:**

The present study, present application study in the field of Data Mining, to support decision maker in University of Nyala, through extract some patterns which can be participate in processing educational development via technical applications Data Mining to improve the academic performance in University.

#### **1.4 Hypotheses :**

- If using classification technique, it helps to identify the student's academic quality to study the track.
- Does the use of classification; helps to predict the result that a student can obtain based on available data.
- If using association technique help to access to logical rules that can be helpful for detect the reasons of fail.

#### **1.5 Objectives of the study:**

1- Looking on the reasons that why students get high marks in Sudanese Certificate but acquire weak marks in the first year exams in university.

2- Design model to extract knowledge that gives a vision about the students rate.

#### **1.6 Bounders of research:**

- **Place scope:**

The research was conducted in University of Nyala, Faculty of Science and Information Technology, in South Darfur State.



**Figure (1.1):** Map of University of Nyala, South Darfur State (2017).

- **Time scope:**

The research was running from the period October 2018 to February 2019.

**Table (1.1):** Research scheduling.

<b>Subject</b>	<b>Period</b>
Data collection ,cleaning ,selection .	2 month
Data mining , building system.	1 month
Classification ,association.	1 month
Documentation.	1 month

### **1.7 Contributions:**

applying this methodology , it can help decision-makers in Sudanese University, to predict students' levels and potentials based on a model that has been designed in advance, and whether the student is qualified to study the relevant discipline or not. In addition to

reaching the rules of close relationship between students, including the discovery of rules that help in making decisions and the work of future plans to improve performance and raise competencies.

### **1.8 Layout of the research:**

The remaining part of the thesis is organized as follows:

Chapter one gives introduction about the research, defining the problem, objectives, methodology and scope. Chapter 2 is literature review and related work, which contains two parts. Part one represents general background about data mining and educational data mining, part two is the related studies and techniques that used in educational data mining. Chapter 3 explains the methodology, which contains two parts, part one explains tools and techniques used, part two data analysis for the research. Chapters 4 contain results, discussion and recommendations.



**CHAPTER TWO**  
**LITERATURE REVIEW**

# **CHAPTER TWO**

## **LITERATURE REVIEW**

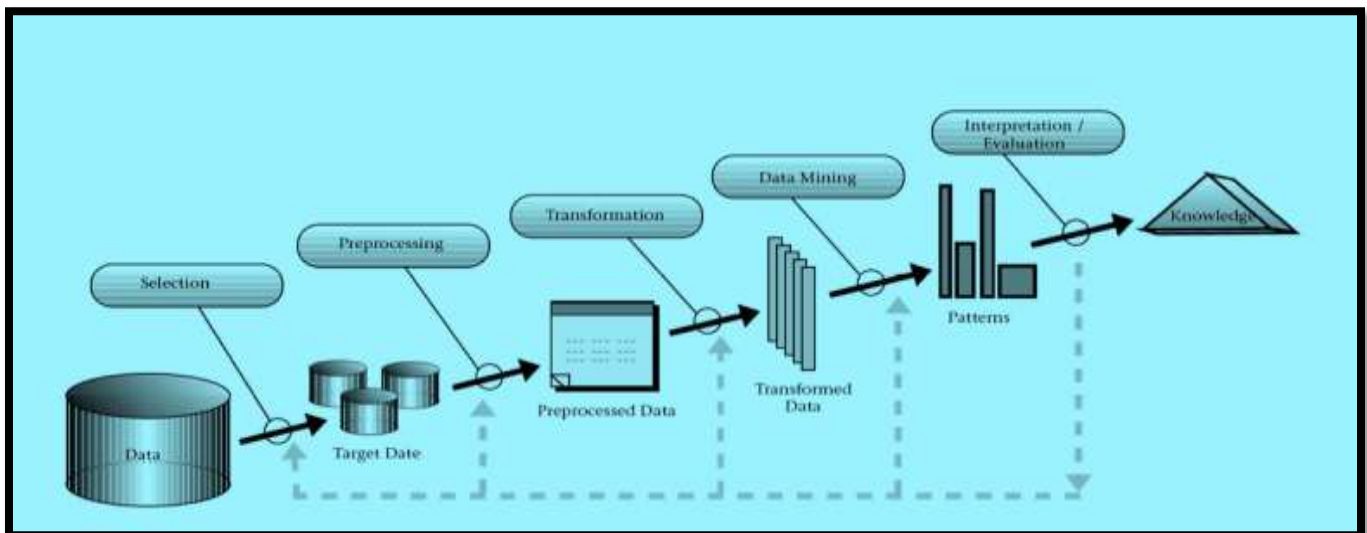
### **2.1 Background of data mining:**

The rapid development of storage technology has greatly increased the volume of data in different enterprises, these data are useful information. This demographic inflation in the volume of data in the institutions did not is keeping pace with effective ways to invest in this massive volume of data, so a new challenge has recently emerged , how to bypass the idea of rules, traditional data that store and search for information only through questions directed by the researcher to techniques used in To infer knowledge by exploring patterns of data in order to take decisions, plan and the formation of a future vision for institutions (Park and Baik, 2006, Linoff and Berry, 2011).

One of these techniques is data mining technology, which is one of the most prominent modern sciences (Han et al., 2011b), specialized in dealing with data and information. It can be said that the age of exploration in the data is just over ten years, where it actually began with the beginning of the third millennium (Bay et al., 2000, Ziegel, 2001). There are several definitions of this concept (data mining): is a multidisciplinary field, encompassing areas like information technology, machine learning, statistics, pattern recognition, data retrieval, neural networks, and information based systems, artificial intelligence and data visualization (Jayaprakash and Jaiganesh, 2018). Data mining is a process which finds useful patterns from large amount of data, the process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions. Data are generally stored in various formats like text, images, audio, video, animated scripts etc in the repository. Han and Kamber define data mining as the process of discovering hidden images, patterns and knowledge within large amount of data and making predictions for outcomes or behaviors (Han et al., 2011b). Also Data Mining (DM), or Knowledge Discovery in Databases (KDD), is an approach to detect or discover useful information from large amount of data (Klüver, 2016), the knowledge discovery in database is systematized in various stages, where as the first stage is selection of data , in which data is gathered from different sources, the second stage is pre-

processing the selected data, the third stage is transforming the data into suitable format so that it can be processed further, the fourth stage consist of Data Mining where suitable Data Mining technique is applied on the transformed data for extracting valuable information and evaluation is the last stage as shown in the Figure ( 2.1 ) below .

**Figure (2.1):** Knowledge Discovery Process in data mining Knowledge Discovery in databases is the process of retrieving high-level knowledge from low level data. It is an



iterative process that comprises steps like Selection of Data, Pre-processing the selected data, Transformation of data into appropriate form, Data mining to extract necessary information and Interpretation/Evaluation of data (Mahindrakar and Hanumanthappa, 2013).

- Selection step collects the heterogeneous data from varied sources for processing. Real life medical data may be incomplete, complex, noisy, inconsistent, and/or irrelevant which requires a selection process that gathers the important data from which knowledge is to be extracted.
- Pre-processing step performs basic operations of eliminating the noisy data, try to find the missing data or to develop a strategy for handling missing data, detect or remove outliers and resolve inconsistencies among the data.
- Transformation step transforms the data into forms which is suitable for mining by performing task like aggregation, smoothing, normalization, generalization, and discretization.

Data mining is a main component in KDD process. Data mining includes choosing the data mining algorithm(s) and using the algorithms to generate previously unknown and hypothetically beneficial information from the data stored in the database. This comprises deciding which models/algorithms and parameters may be suitable and matching a specific data mining method with the general standards of the KDD process. Data mining methods includes classification ,summarization, clustering, regression (Mahindrakar and Hanumanthappa, 2013). Data Mining Techniques are used to manage large amounts of data to discover hidden patterns and relationships. These patterns are helpful in decision making. Data mining techniques includes algorithms like classification, regression, clustering prediction and association. These techniques are used for knowledge discovery from database.

### **2.1.1 Classification:**

Classification is a classic data mining technique based on machine learning. Classification technique maps data into a set of predefined classes to describe a model, classification uses classification rule (IF - Then), decision tree, neural network and etc. For example can apply the classification rule on the past record of the student who left for University and evaluate them (Jindal and Borah, 2013).

#### **2.1.1.1 Classification techniques used in Data Mining:**

It could be some time but not necessarily advisable predictive modeling is seen as a "black box" that makes predictions about the future based on information from the past, and present. Some designs are better than others in terms of accuracy. Some designs are better than others in terms of understanding. For example, models from better understanding of the incomprehensible decision trees, rule induction, and regression models, neural networks. The classification is a type of predictive models. More specifically, the ranking is the appointment process of new objects or predefined categories: given a set of marked files, build a model such as the decision tree, and predicting future records labels is called for classes such of them:

- **Decision Tree :**

Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and

each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. An example of a decision tree for the training set. Decision tree is a predictive model which, as its name suggests, can be seen as a tree. Specifically each branch of the tree is a classification of matter and leaves of trees and data sections with classified, and this technique which is applied in this research.

- **Bayesian Networks :**

A Bayesian Network (BN) is a graphical model for probability relationships among a set of variables features. The Bayesian network structure  $S$  is a directed acyclic graph (DAG) and the nodes in  $S$  are in one-to-one correspondence with the features  $X$ . The arcs represent casual influences among the features while the lack of possible arcs in  $S$  encodes conditional independencies. Moreover, a feature (node) is conditionally independent from its non-descendants given its parents ( $X_1$  is conditionally independent from  $X_2$  given  $X_3$  if  $P(X_1|X_2, X_3) = P(X_1|X_3)$  for all possible values of  $(X_1, X_2, X_3)$ ).

- **K-nearest neighbor classifiers :**

Nearest neighbor classifiers are based on learning by analogy. The training samples are described by  $n$  dimensional numeric attributes. Each sample represents a point in an  $n$ -dimensional space. In this way, all of the training samples are stored in an  $n$ -dimensional pattern space. When given an unknown sample, a  $k$ -nearest neighbor classifier searches the pattern space for the  $k$  training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points.

$$X=(x_1, x_2, \dots, x_n) \text{ and}$$

$$Y=(y_1, y_2, \dots, y_n) \text{ is } (x+a)^n = \sqrt{\sum(x_i - y_i)^2}$$

**2.1.1.2 Classification Algorithms:**

- **Decision tree classifiers :**

A decision tree can be a flow chart resembling a tree structure, where every internal node is denoted by rectangles and the leaf nodes are denoted by ovals. This is often used algorithm because of easy implementation and easier to understand compared to different classification algorithms. Decision tree starts with a root node that helps the users to take required actions. From this node, users split every node recursively according to decision

tree learning algorithm. The ultimate result is a decision tree in which each branch represents an outcome (Pant, 2017).

- **C4.5 (J48) :**

It is the algorithm chosen, this algorithm can be a successor to ID3 developed by Quinlan Ross. It is additionally supported the Hunt's algorithm. C4.5 handles each categorical and continuous attributes to create a decision tree, so as to handle continuous attributes. C4.5 splits the attribute values into 2 partitions based on the chosen threshold. It additionally handles missing attribute values. C4.5 has the concept of Gain Ratio as an attribute selection measure to create a decision tree. It prunes the biasness of information gain once there are many outcome values of an attribute. At first, calculate the gain ratio of every attribute. The root nodes are the attribute whose gain ratio is a maximum. C4.5 uses pessimistic pruning to get rid of unessential branches with in the decision tree to enhance the accuracy of classification (Jayaprakash and Jaiganesh, 2018).

- **Random Forest :**

Random Forests is a bagging tool that leverages the ability of multiple varied analyses, organization strategies, and ensemble learning to supply correct models, perceptive variable importance ranking, and laser-sharp coverage on a record-by-record basis for deep data understanding. Its strengths are recognizing outliers and anomalies in knowledgeable data, displaying proximity clusters, predicting future outcomes, characteristic necessary predictions, discovering data patterns, exchange missing values with imputations, and providing perceptive graphics (Jayaprakash and Jaiganesh, 2018).

- **Neural Network :**

Multilayer Perceptron (MLP) algorithm is one of the most widely used and common neural networks. Multilayer Perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a collection of acceptable output. An MLP consists of multiple layers of nodes in an exceedingly directed graph, with every layer totally connected to the consequent one. Their current output depends solely on the present input instance. It trains victimization back propagation (Jayaprakash and Jaiganesh, 2018).

- IBI :

IBI is nearest neighbor classifier. It uses normalized Euclidean distance to search out the training instance nearest to the given test instance, and predicts the identical category as this training instance. If many instances have the smallest distance to the test instance, the primary one obtained is employed. Nearest neighbor methodology is one of the effortless and uncomplicated learning/classification algorithms, and has been effectively applied to a broad variety of issues (Jayaprakash and Jaiganesh, 2018).

- Decision Table:

Decision Tables are classification models elicited by machine learning algorithms and are used for creating predictions. A decision table consists of a hierarchical table within which entry in a higher level table gets broken down by the values of a pair of additional attributes to make another table.

### **2.1.2 Clustering:**

Clustering is a collection of similar data object. Dissimilar object is another cluster. It is way finding similarities between data according to their characteristic. This technique based on the unsupervised learning. For example, image processing, pattern recognition and city planning (Patil, 2015). Clustering and classification both, is a basic function in data mining. Classification is used mainly as a way of learning under the supervision block to learn uncensored. The objective of the meeting is descriptive, this classification is predictive. Because the goal of the meeting is to find a new set of new categories and groups are important in themselves, and their evaluation is essential. As we mentioned before, classification can be taken as supervised learning process, clustering is another mining technique similar to classification. However clustering is a unsupervised learning process. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects, so that objects within the same cluster must be similar to some extent, also they should be dissimilar to those objects in other clusters. In classification which record belongs which class is predefined, while in clustering there is no predefined classes. In clustering, objects are grouped together based on their similarities. Similarity between objects is defined by similarity functions; usually similarities are quantitatively specified as distance or other measures by corresponding domain experts. Most clustering applications

are used in market segmentation. By clustering their customers are divided into different groups, business organizations can provide different personalized services to different group of markets. For example, based on the expense, deposit and draw patterns of the customers, a bank can clustering the market into different groups of people. For different groups of market, the bank can provide different kinds of loans for houses or cars with different budget plans.

### **Types of clustering methods :**

- Partitioning Methods .
- Hierarchical Agglomerative (divisive) methods .
- Density based methods .
- Grid-based methods .

### **2.1.3 Association rule :**

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

### **2.1.4 Regression:**

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous



response variables). Neural networks too can create both classification and regression models (Han et al., 2011a). Regression is used to map a data item to a real valued prediction variable. In other words, regression can be adapted for prediction. In the regression techniques target value are known. For example, you can predict the child behavior based on family history (Han et al., 2011a).

## **2.2 Background of educational data mining (EDM):**

Data Mining is playing an important role in educational systems where education is considered as one of the key inputs for social development (Jindal and Borah, 2013). In recent years, there has been increasing interest in the use of DM to investigate educational field. Educational Data Mining (EDM) is concerned with developing methods and analyzing educational content to enable better understanding of students' performance (Han et al., 2011a). It is also important to enhance teaching and learning process. The data can be collected from historical and operational data reside in the databases of educational institutes. The student data can be personal or academic. Also it can be collected from e-learning systems which have a large amount of information used by most institutes. Prediction models that include all personal, social, psychological and other environmental variables are necessitated for the effective prediction of the performance of the students (Han et al., 2011a). The prediction of student performance with high accuracy is beneficial for identify the students with low academic achievements initially. It is required that the identified students can be assisted more by the teacher so that their performance is improved in future. So the Educational Data Mining (EDM) have different ways that is defined , most notably:(defined educational data mining as “Educational Data Mining 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 is emerging disciplinewith a suite of computational and psychological methods and research approaches to understanding how students learn and setting which they learn in it (Bassil, 2012). Educational data mining (EDM) is an emerging discipline which focuses applying data mining tools and techniques to educationally related data (Buniyamin et al., 2015). An educational data mining is a broader term that focuses on nearly any type of data in educational institutional, while academic analytics is specific to

data related institutional effectiveness and student retention issues. The scope of educational data mining includes areas that directly impact students. Other areas within EDM include analysis of educational processes including admissions, alumni relations and course selections (Buniyamin et al., 2015). Also defined as academic analytics as the use of statistical techniques and data mining in ways that will help faculty and advisors become more proactive in identifying the position of student (Campbell et al., 2007), as well as the Educational data mining (EDM) deals with the developing and applying the computerized methods to detect patterns in large amount of educational data ,that would be impossible to analyze .The main objective of any higher educational system is to improve the quality of education. To accomplish this goal, the data mining techniques can be used. Educational data mining have some advantages over the higher educational system such as decreasing student's drop-out rate, increasing student's promotion rate, increasing student's retention rate, increasing student's transition rate, increasing educational improvement ratio, increasing student's learning outcome, maximizing educational system efficiency and reducing the cost of system processes. To achieve these goals the data mining system will be helpful to put insights for decision makers in the higher educational system.

### **2.2.1 Goals of educational data mining in higher education:**

- Predicting student's future learning behavior: With the use of student modeling, this goal can be achieved by creating student models that incorporate the learner's characteristics, including detailed information such as their knowledge, behaviors and motivation to learn (Jindal and Borah, 2013).
- Discovering or improving domain models: Through the various methods and applications of EDM, discovery of new and improvements to existing models is possible (Jindal and Borah, 2013).
- Studying the effects of educational support: It can be achieved through learning systems (Jindal and Borah, 2013).
- Advancing scientific knowledge about learning and learners :By building and incorporating student models, the field of EDM research and the technology and software used (Jindal and Borah, 2013).

### 2.2.2 Educational data mining methods:

There are so many promoted methods of educational data mining but all kind of methods Falls within one of the following specified categories:

1. Prediction: (Siemens and d Baker, 2012) has given a detail explanation of prediction in his paper. He mentioned that “In prediction, the goal is to develop a model which can infer a single aspect of data from some combination of other aspects of data. If we study prediction extensively then we get three types of prediction: classification, regression and density estimation. In any category of prediction the input variables will be either categorical or continuous. In case of classification, the categorical or binary variables are used, but in regression continuous input variable s are used. Density estimation can be done with the help of various kernel functions (Siemens and d Baker, 2012).
2. Clustering: In clustering technique, the data set is divided in various groups, known as clusters. When data set is already specified, then the clustering is more useful. As per clustering phenomenon, the data point of one cluster and should be more similar to other data points of same cluster and more dissimilar to data points of another cluster. There are two ways of initiation of clustering algorithm. Firstly, start the clustering algorithm with no prior assumption and second is to start clustering algorithm with a prior postulate.
3. Relationship Mining: Relationship mining generally refers to contrive new relationships between variables. It can be done on a large data set, having a no of variables. Relationship mining is an attempt to discover the variable which is most closely associated with the specified variable. There are four types of relationship mining: association rule mining, correlation mining, and sequential pattern mining and causal data mining. Association data mining is based on if- then rules that is if some particular set of variable value appears then it generally has a specified value. In correlation mining, the linear correlations are discovered between variables. The aim of sequential pattern mining is to extract temporal relationships between variables.

4. Discovery with Models: it includes the designing of model based on some concepts like prediction, clustering and knowledge engineering etc. This newly created models predictions are used to discover a new predicted variable.

### **2.3 Application of data mining**

Now a day's data mining are used in lots of areas but in this section , here mainly listed some application areas for data mining, the diversification of such data has led to the application of data mining in the following sectors:

#### **1. Business Sector:**

In business world data mining is basically used for analyzing performance, profitability index, and customer feedback evaluation and analysis of the stock values of existing organizations and their market trends to aid in future business decisions.

#### **2. Marketing and retailing sector:**

Data mining provide accurate information regarding customer purchase trends, top selling products, so that the retail-store managers are able to identify their loyal customers, and providing discounts and arranging shelves according to customer requirements (Delmater and Hancock, 2001).

#### **3. Bio-informatics:**

Accumulation of medical records of the patient to develop a relationship between the disease and the effectiveness of treatment, assessment of genomic and proteomic data in bio medical field (Bhatnagar et al., 2012).

#### **4. Climatology:**

Assessment of weather conditions over a period of time so as to predict future meteorological patterns for determining natural calamities like cyclone and also weather forecasting.

#### **5. Banking and finance:**

Assessment of individual banking records to generate different marketing strategies for a target customer segment, loan approval, stock forecasting, checking different kinds of fraud and money laundering (Delmater and Hancock, 2001).

## **6. Security and data integrity:**

Data mining can be used to monitor different systems and raises alarm when ever any kind of security breach or intrusion is detected. It can help in identifying the reason for security problems in firewall.

## **7. Electronic commerce:**

Data mining techniques are used in ecommerce to analyze customer search patterns to promote up sale and cross sale.

## **8. Forensic and criminal investigation:**

Data mining technique is used in forensic and criminal department to assess previous criminal records in order to identify the criminal as well as to determine the crime pattern ,sentiments of the accused and the accuse (Pal, 2011).

## **9. Government records:**

Data mining technique is used in government record. For generating citizen specific data which can include anything starts form employment record till medical history, law enforcement, profiling tax cheaters (Pal, 2011).

## **10. Cloud computing:**

Today cloud computing can be considered as one of the major source of every kind of data. Moreover cloud servers are fast, reliable, efficient and secure and reduce the cost of infrastructure of individual (Pal, 2011). Furthermore using of KDD techniques and different kinds of data mining algorithms, one can even create different search patterns and applications for finding any information which remains hidden in unstructured data.

### **2.4 Overview of tools for data mining :**

- Gephi : is an interactive visualization and exploration platform. It is mainly focused on networks and dynamic and hierarchical graphs.
- Graphviz : is a graph visualization software specialized for representing structural information through diagrams of networks.
- KNIME : is a modular data exploration platform that enables users to depict data flows and examine results using interactive views on data and models.
- Pajek : is a visualization tool mainly focused on the analysis of complex network data. It is usually utilized for social network analysis.

- R: is both the programming language and the environment facilitating data manipulation, calculation, visualization and DM.
- RapidMiner : is also a tool for machine learning (ML) and DM tasks which allows users to create data flows. They are similar to pipelines offered by Weka.
- VisuaLinks : is a graphical analysis tool designed for discovering patterns and hidden networks in sources of heterogeneous data. It addresses the analytical process from access and integration to presentation and reporting.
- VizTree :is aVDM system with the specialization on the time series analysis using tree visualization and interaction.
- Weka : is a collection of ML and DM algorithms for analytical tasks that allows users to create pipelines in order to visualize and control the analytical process.
- XmdvTool : is an interactive tool for visual exploration of multivariate data sets.

Additionally, tools mainly providing statistical and mathematical visualization involve Mathematic ,or Matlab Matla is a language and an environment enabling users to perform computationally intensive tasks. The octave is a language similar to Matlab. It is mainly intended for numerical computations. It also provides capabilities for data visualization and manipulation.

## **2.5 WEKA tool:**

The Weka or woodhen (*Gallirall usaustralis*) is an endemic bird of New Zealand. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand (Hall et al., 2009). The Weka suite contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. It provides many different algorithms for data mining and machine learning. Weka is open source and freely available. It is also platform-independent. As shown in fig3 the GUI Chooser consists of four buttons: Explorer: An environment for exploring data with WEKA. Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes. Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag and drop interface.

One advantage is that it supports incremental learning. Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface. This Java-based version (Weka 3) is used in many different application areas, the version (weka 3.6) is used in this research, in particular for educational purposes and research.

#### **Various advantages of Weka:**

- It is freely available under the GNU General Public License.
- It is portable, since it is fully implemented in the Java programming language and thus runs on almost any architecture.
- It is a huge collection of data preprocessing and modeling techniques.
- It is easy to use due to its graphical user interface. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

#### **2.6 Related work:**

Data mining has been widely applied in the higher education field as universities provide huge amount of data. Some of the application is to study factors that affect student retention through monitoring the academic behavior and providing powerful strategies to intervene as proposed (Yu et al., 2010). (Abu Tair and El-Halees, 2012), showed the concept of Educational data mining (EDM) and used it to improve students' performance, the data are collected from graduate students information collected from the college of Science and Technology – Khanyounis ,each one of these method discovered rule knowledge , can be used to improve the performance of graduate student (Abu Tair and El-Halees, 2012).

(Patil, 2015), the data collected from students of polytechnic institute, the study shows more attention should be given to improve basic fundamentals in mathematics and language based on some academic attributes of students. Comparing results of DT and NB algorithms is observed that decision tree (J48) gives better result than Naive Bayesian algorithm in terms of accuracy in classifying the data (Velmurugan and Anuradha, 2016). Data collected from Sri Sai University Palampur, used the clustering, decision tree and neural networks techniques to evaluate student's performance. The result was that , by using above techniques ,Teachers can easily evaluate the performance of the students (Anoopkumar and

Rahman, 2018). Data mining will be considered most useful in educational field. Predicting student's academic performance, by applying data mining techniques and tools, in prediction of student performance is helpful to identify the abilities of students, their interests and weaknesses performance. Prediction of student Performance is done by applying Naive Bayesian and J48 decision tree classification techniques WEKA tool, Naive Bayesian provide 63.59 % accuracy and j48 provide 61.53% accuracy (Sivasakthi, 2017). (Bassil, 2012) proposed a model for typical university information system that based on transforming an operational database whose data are extracted from an already existing operational database. The purpose of the proposed design is to help decision makers and university principles. (Romero and Ventura, 2007) introduced a survey of the specific application of data mining in learning management systems and a case study tutorial with the Model system. (Ahmed and Elaraby, 2014), used the classification task to predict the final grade, by presenting a study, that can help the student's instructors to improve the student's performance, by identifying those students who needed special attention to reduce failing and taking appropriate action at right time.



**CHAPTER THREE**  
**MATERIALS AND METHODS**

## **CHAPTER THREE**

### **MATERIALS AND METHODS**

At the beginning of this chapter, I will explain all the practical steps that have been taken, to implement the methodology that has been followed to achieve the required goals step by step.

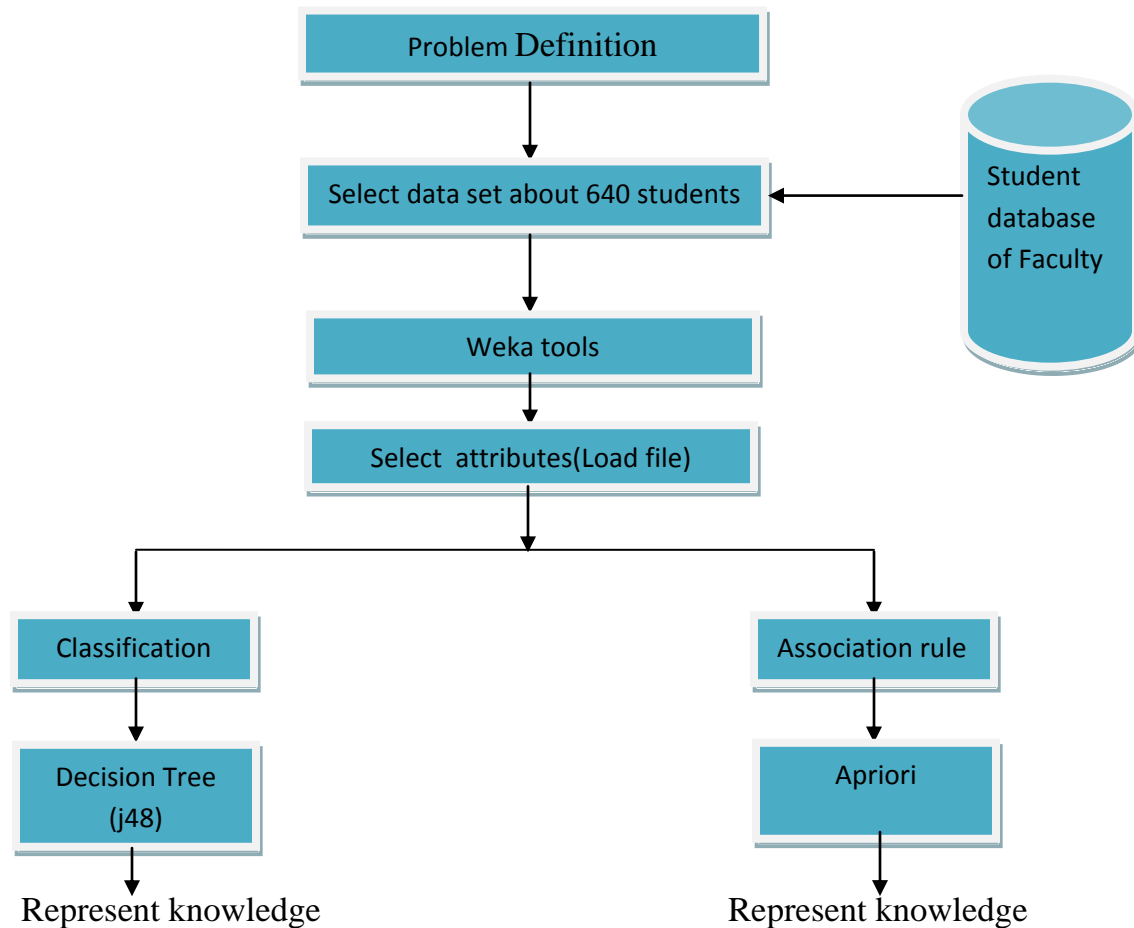
#### **3.1 Methodology:**

The data set will be used in this study is about 300 students, contains personal and academic information collected from the faculty of Science and Information Technology, University of Nyala.

Before applying the data mining techniques on the data collected above, there should be a methodology that governs my work. Figure 1 depicts the work methodology used in this research, which is based on the framework proposed according to the protocol described previously (Hall et al., 2009). The methodology starts from the problem definition, then Select data set about 640 students which contains personal and academic information, then come to the data mining tool (WEKA) to apply techniques which are Classification, Association Rule and Clustering.

Then the data will be in file , Attribute Relation File Format (.Arff) , this file is one of the file format that WEEKA tool deal with it, the file will contain all the attribute that I will be used to process the data collected above , then apply data mining techniques classification ,clustering ,association rules, by choosing decision tree (j48) algorithm for classification and Aproiri algorithm for association rule .

The preparation and preprocessing of the data set will be done according to the flowing steps:



**Figure(3.1): Work Methodology.**

### **3.1.1 Data collection step:**

The students data of the Faculty of Science and Information Technology were collected in the first batch of eight sections, Computer Science, Information Technology, Geology, Chemistry, Physics, Mathematics, Zoology and Botany. The data was taken from students files , which is contains personal data and Sudanese secondary certificate, in addition to academic data from the result, all data was written manually.

### **3.1.2 Selection step:**

The heterogeneous data was selected from data collected includes personal data (age, family economic status, student residence, student gender), secondary data (success rate, additional subject ,number times to sit for the Sudanese certificate exam), and academic data(student results in the first year, attendance, type of admission).

### **3.1.3 Pre-processing step:**

Performs basic operations of eliminating the noisy data, try to find the missing data or to develop a strategy for handling missing data.

### **3.1.4 Transformation step:**

Transforms the data into forms which is suitable for mining by performing task like aggregation, smoothing, normalization, generalization, and discretization. All students data were written in an Excel file and then converted to a file in a format ARFF (Attribute-Relation File Format ) to be ready to work in the WEKA tool.

### **3.1.5 Data mining step:**

Classification and Association rules data Mining technique is applied on the transformed data for extracting valuable information, for classification technique , various algorithms was applied and compare between them , which is better , and present better result. In addition to association rules which present rules, help to understand the relation between the students.

## **3.2 Data mining method:**

Data mining is a computational method of processing data which is successfully applied in many areas that aim to obtain useful knowledge from the data (Klösgen and Zytkow, 2002). Data mining techniques are used to build a model according to which the unknown data will try to identify the new information. Regardless of origin, all data mining techniques show one common feature: automated discovery of new relationships and dependencies of attributes in the observed data. If the goal of the analysis is the categorization of data by class, then that is the new information on classes to which data belongs. In doing so, the algorithms are divided into two basic groups unsupervised algorithms and supervised algorithms. When the mining is "unsupervised" or "undirected", the output conditions are not explicitly represented in the data set, the task of unsupervised algorithms to discover automatically inherent patterns in the data without the prior information about which class the data could belong, and it does not involve any supervision (Cios et al., 2007). Conversely, in supervised learning, no target variable to be learned is identified as such. Instead, the supervised learning algorithm searches for patterns and

structure among all the variables. The goal of such model is to uncover data patterns in the set of input fields.

**CHAPTER FOUR**  
**RESULTS AND DISCUSSION**

# CHAPTER FOUR

## RESULTS AND DISCUSSION

### 4.1 Implementation in Weka tool:

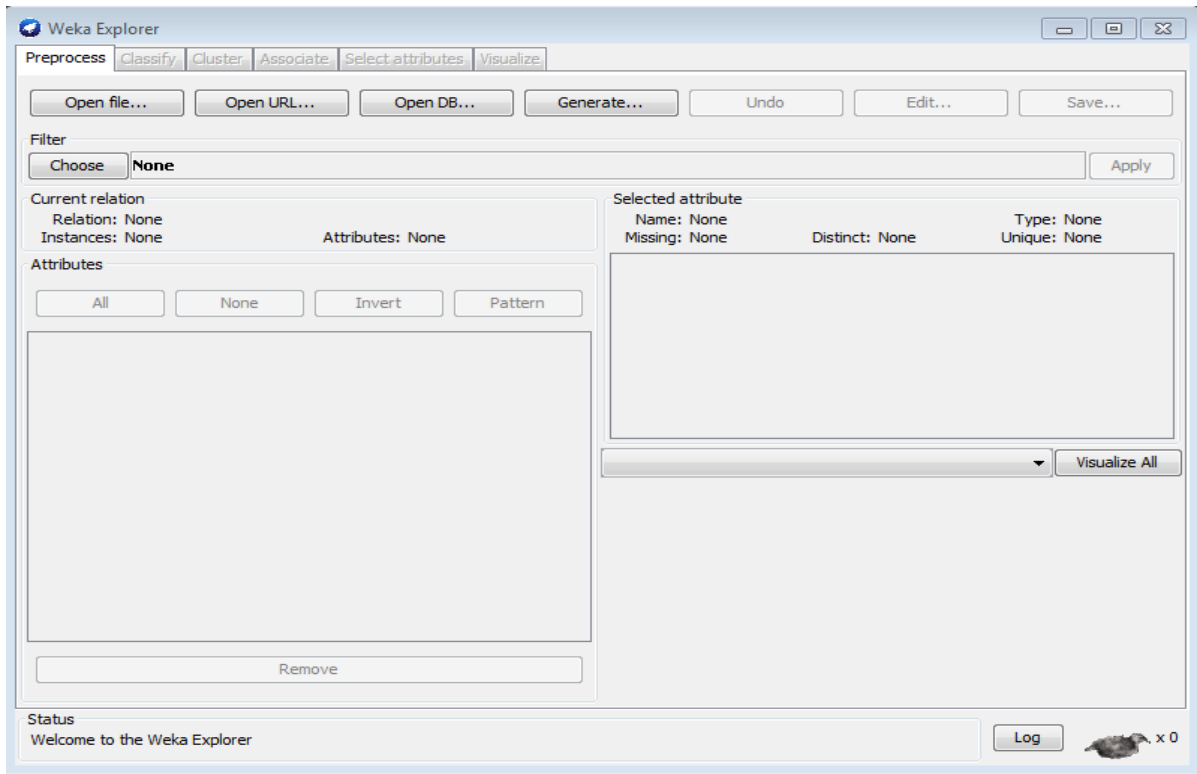
Download and install WEKA tool, Start Weka get the Weka GUI chooser window , the initial window of the program will appear in this format, which contains four options, three main Graphical User Interfaces(GUI) the Explorer (exploratory data analysis),the Experimenter” (experimental environment) and the KnowledgeFlow (new process model inspired interface) , see (Figure: 4.1) below:



**Figure (4.1) :** Front view of weka tools.

#### 4.1.1 Classification:

Click on the Explorer button and get the Weka Knowledge Explorer window, Click on the “Open File..” button and load an ARFF file” our dataset” to start working on the dataset , after pressing , the window will appear in the following Figure :



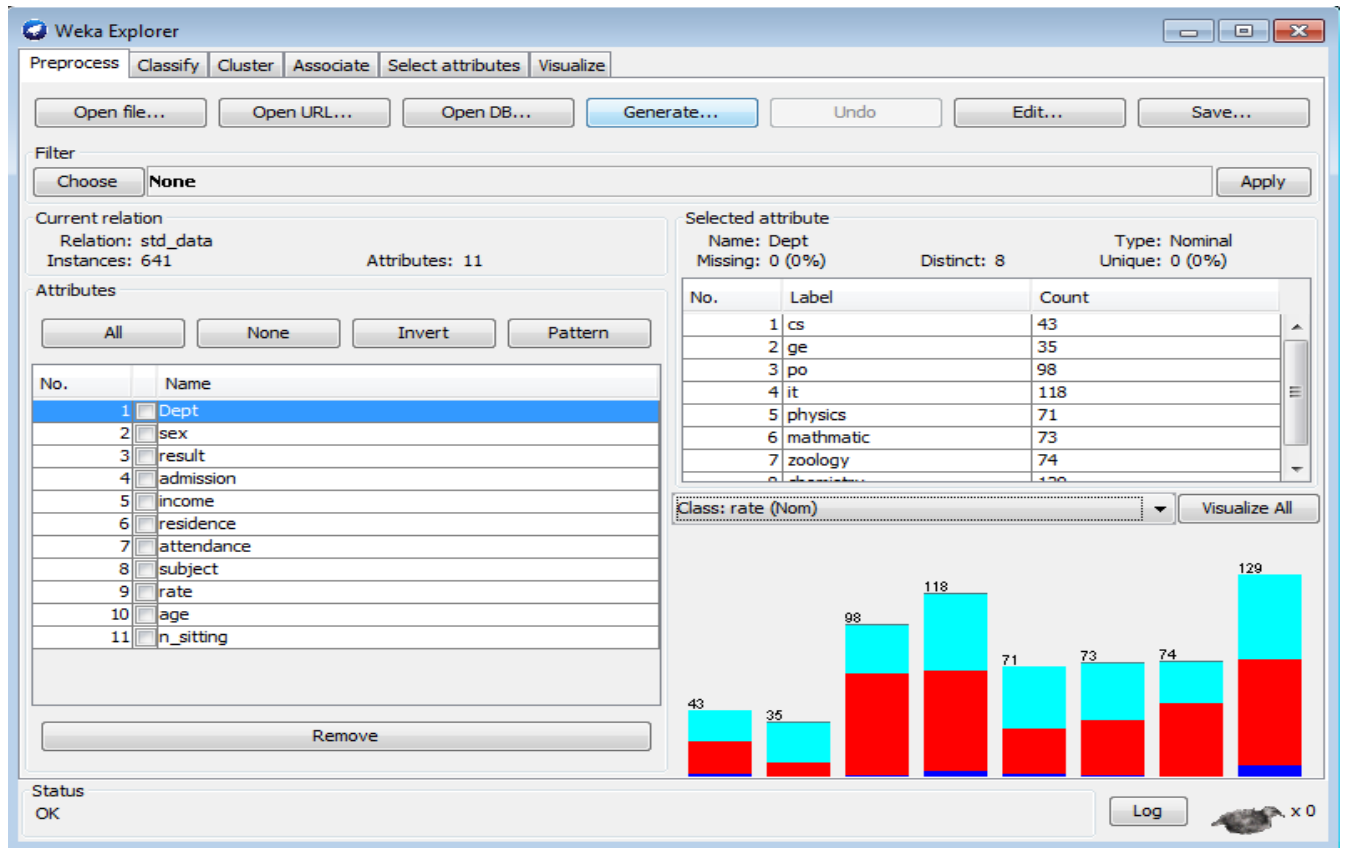
**Figure (4.2) :Weka Explorer.**

**The first phase:**

- pre-processing the data , Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary or can also be read from a URL , SQL database (using JDBC) , here data imported from file format ARFF, see Figure(5) .
- after data imported , you see current relation which contains :  
 Relation :std\_data (the name of the relation in ARFF file dataset).  
 Attribute: 11 attributes ( the number of column in the dataset which columns refer to the attributes ).

Instances: 641 instance (the number of rows in the dataset which rows refer to the instances).





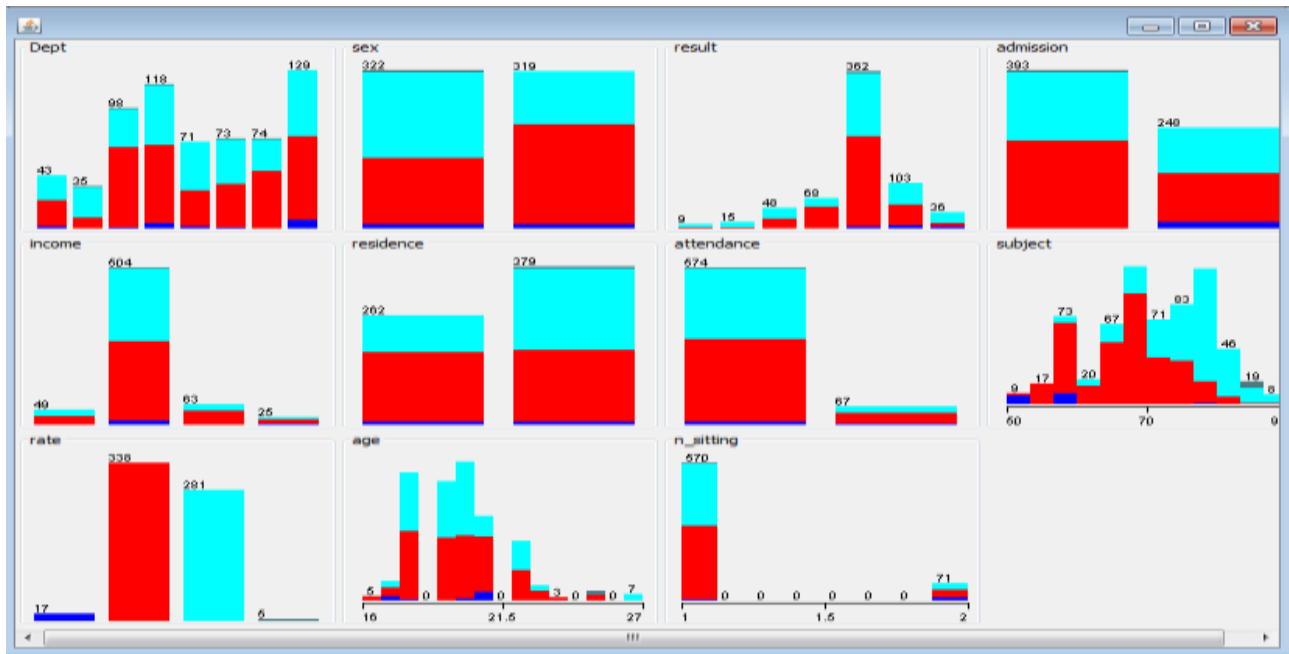
**Figure (4.3):** Load data set in to the weka.

Visualize all attributes in the dataset , based on the success rate of the Sudanese certificate exam , in Figure (4.3) .

Attributes names:

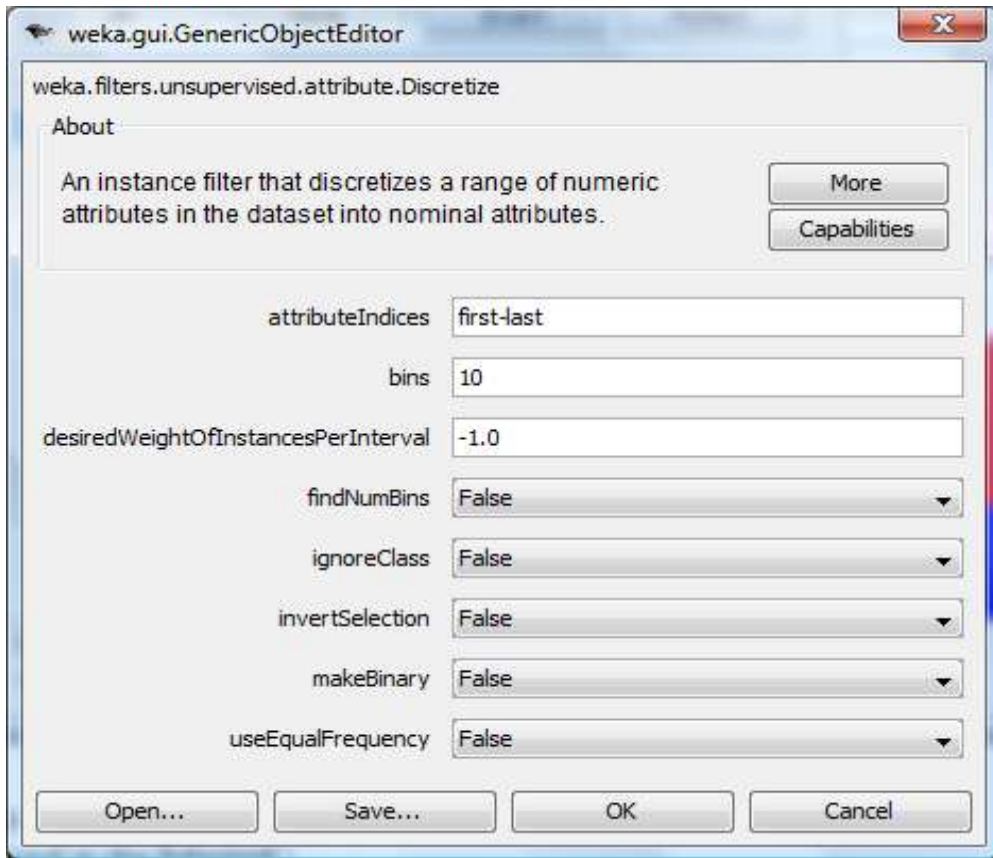
- Dept: the Department of the student in the faculty ( cs (computer science) ,it (information technology) ,zoology, potany, mathematic ,chemistry, physics and geology.
  - Sex: the gender of the student is either male (m) or female (f).
  - Result: as a result of the student at the end of the first year Excellent(E), very good|(vg),good(g), failed(F),dismiss(d) Acceptable(A),retreating(R) .
  - Admission: Type of admission either public or special.
  - Income: family economic status (high , medium , low and Atel).
  - Residence: the place where the student exams high Sudanese certificate either (in) Nyala city, or (out) Nyala city.
  - Attendance: either (f) full or (uf).
  - Subject: the degree of the subject was awarded in Sudanese certificate.

- Rate : the degree which the student chives in Sudanese certificate .
- Age : age of student when accepted to the college.
- N\_sitting : number of sitting student to Sudanese certificate .



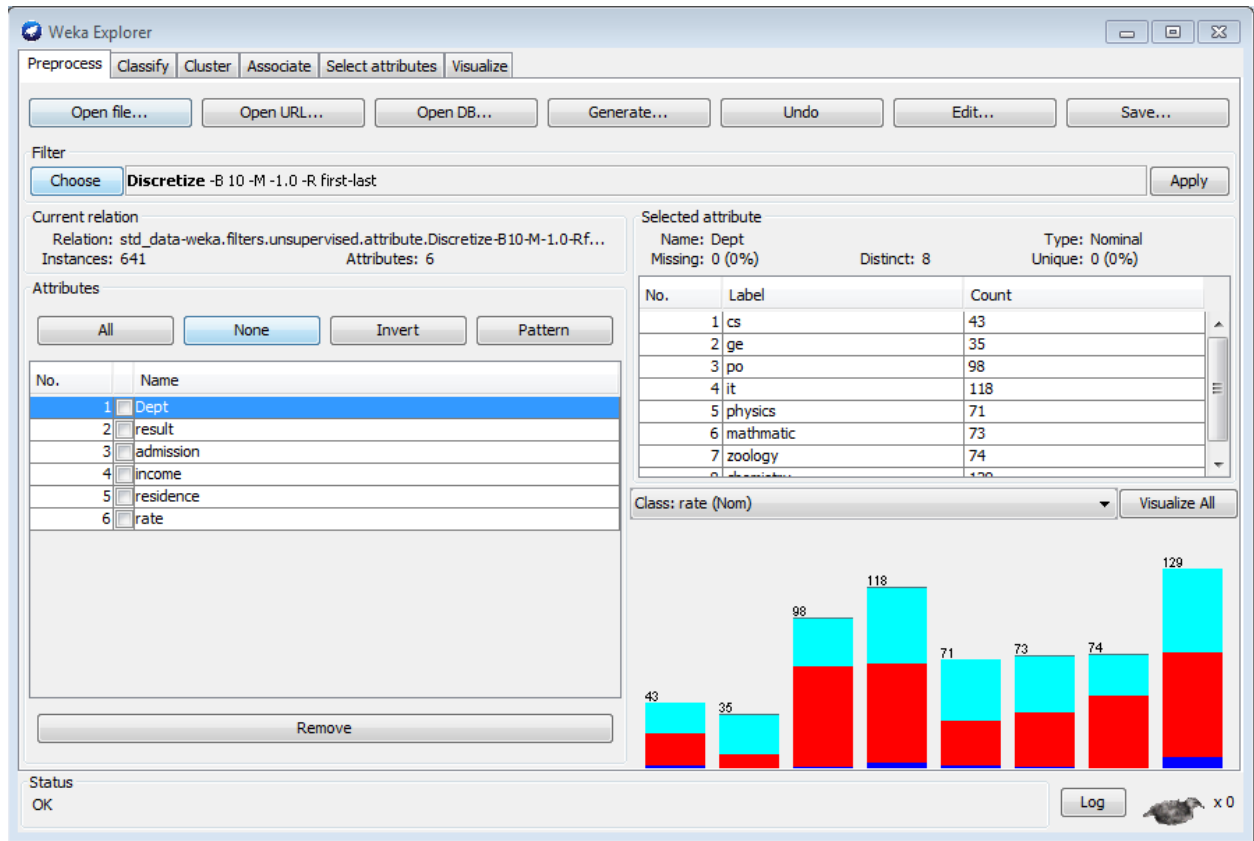
**Figure (4.4) :** Visualize all attributes.

Pre-processing tools in WEKA are called filters, now on this dataset applied filter, there are a lot of different filters, Allfilter and MultiFilter are ways of combining filters. There are supervised and unsupervised filters. Supervised filters are ones that use a class value for their operation; they aren't so common as unsupervised filters, which don't use the class value. There are attribute filters and instance filters here want to remove an attribute. There are so many filters under attribute, I chose Discetize which is An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes. So click on choose and select filters/unsupervised/attribute/Discetize. Then click on the area right of the Choose button. You get the following:



**Figure (4.5) :** Discretizes filter.

You see in fig (4.5) the default parameters of this filter. Click on More to get more information about these parameters, click on the Apply button to do the discretization. Then select one of the original numeric attributes (e.g. Dept) and see how it is discretized in the Selected attribute window.

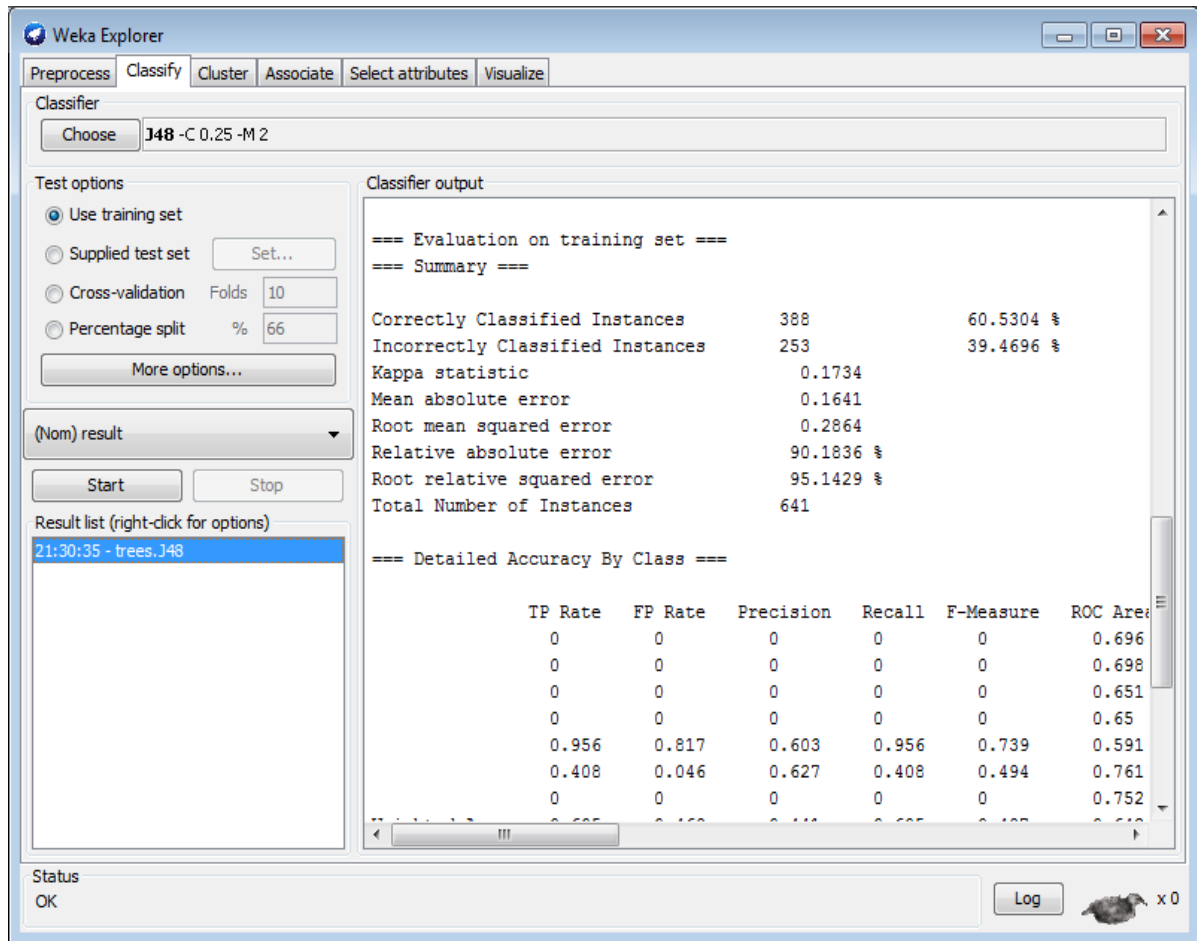


**Figure (4.6):** Filter the dataset.

**The second phase :**

(Han et al., 2012) mentioned Data Mining techniques; here will start working on classification technique.

- Select the "classify" tab and click the "choose" button to select the classifier.
- Classifiers, like filters, are organized in a hierarchy: here will working on J48 which has the full name weka.classifiers.trees.J48.



**Figure (4.7):** j48 classifier by default parameter.

The classifier is shown before in Figure (4.7) in the text choose button: It reads J48 -C 0.25 -M 2. This text gives the default parameter settings for this classifier ,only shows-C Confidence value (default 25%), lower values incur heavier pruning and --Mie. Minimum number of instances in the two most popular branches (default 2).

- Open the weka.gui.GenericObjectEditor and start changing the default values, you will see that more switches will appear in the box, as shown in the figures (9) and (10).

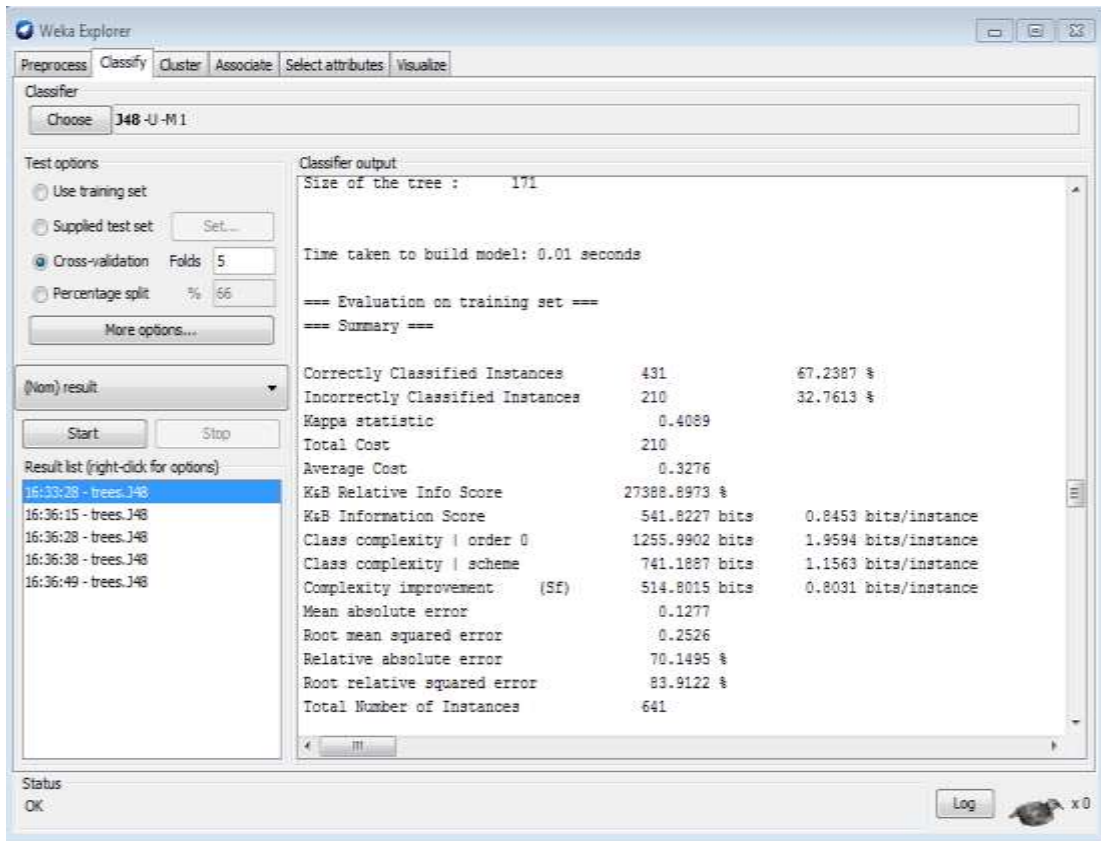


Figure (4.8): j48 classifier after change min number of obj.

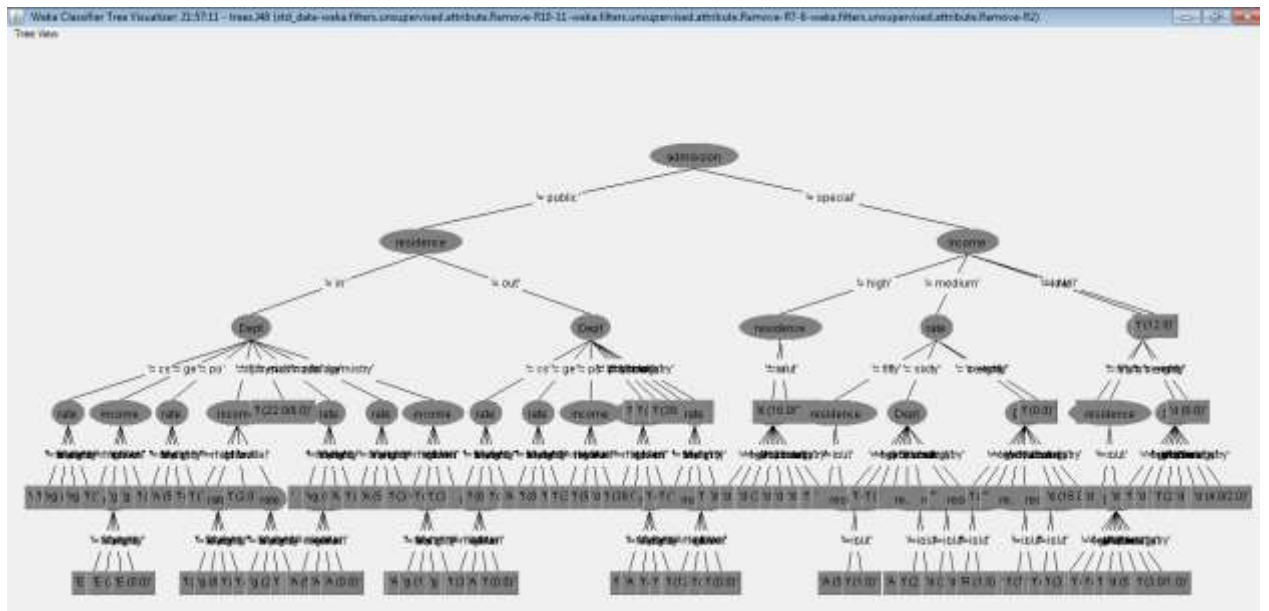


Figure (4.9): Classifier tree visualize.

**Table (4.1):** Performance comparison between different algorithms.

<b>Algorithms</b>	<b>Accuracy</b>	<b>Error</b>
ZeroR	56.4743 %	43.5257 %
LADtree	68.4867 %	31.5133 %
UserClassifier	56.4743 %	43.5257 %
Naivebayes	65.5226 %	34.4774 %
J48(filter)	67.2387%	32.77 %
J48(no filter)	87.9875 %	12.0125 %

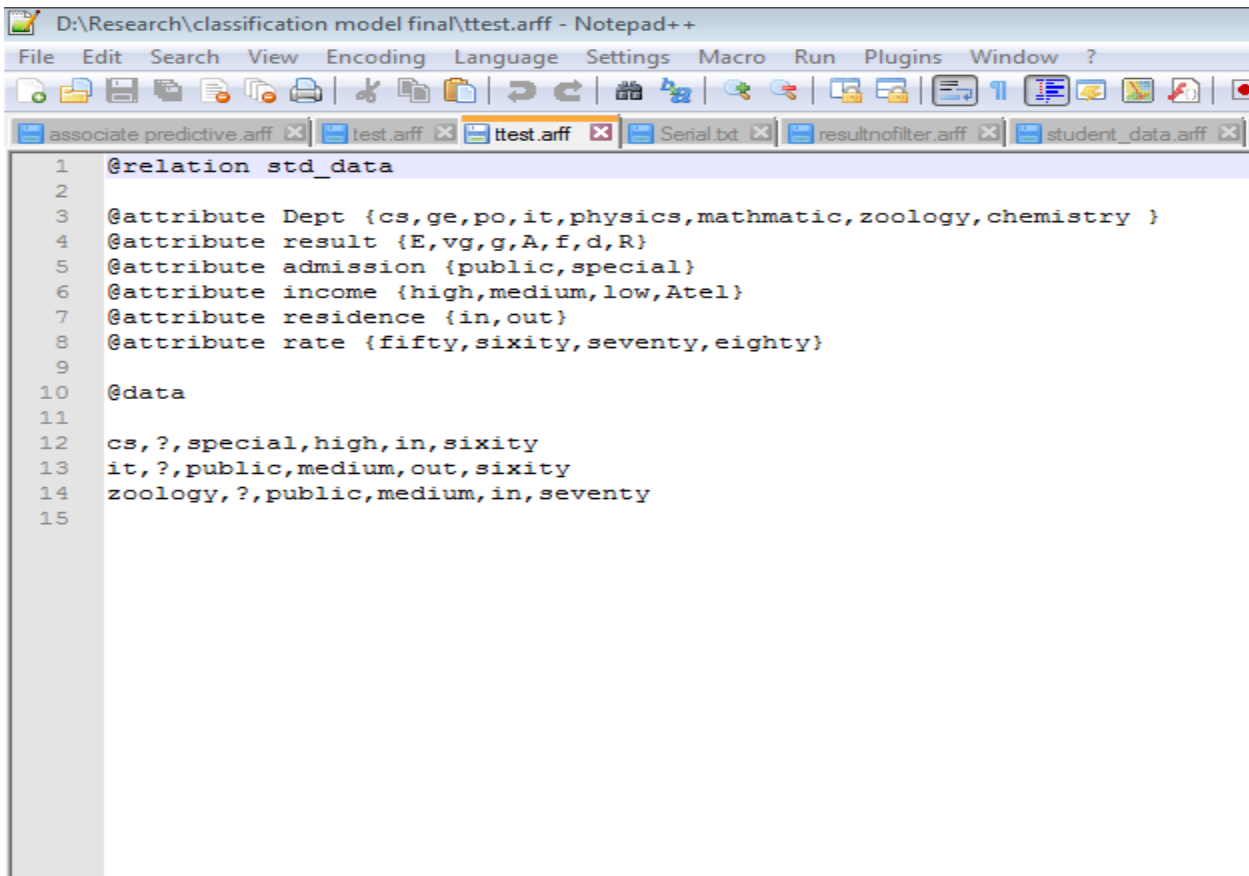
The model is designed and tested to predict the student's level or level at the start of admission according to available information,

#### **4.1.2 Classification model implementation:**

The model has been performed, where the student's academic results, data collected over a period of time at the Faculty of Science and Information Technology, Nyala University. The main aim of the study was to find how our model managed to classify the new instances(tested set) on supplied instances(training set), I use the final results of students departments for the year 2016 to 2018), Faculty of Science and Information Technology.

The model to classify the instances of the figure (4.10) (sample of the train set), the file ARFF(tttest.arff),WEKA source file which include final grade of graduated students , is depicted in figure ( 4.10) . The predicted instances (tested set), the file ARFF (result.arff) see figure (4.11), note that the attribute section is identical to the training data, the file includes the tested instances of the attribute ("result "). The values of "result "attribute is left as "?", thus WEKA has no actual values to which it can compare the predicted values of new instances.

## Tested model :

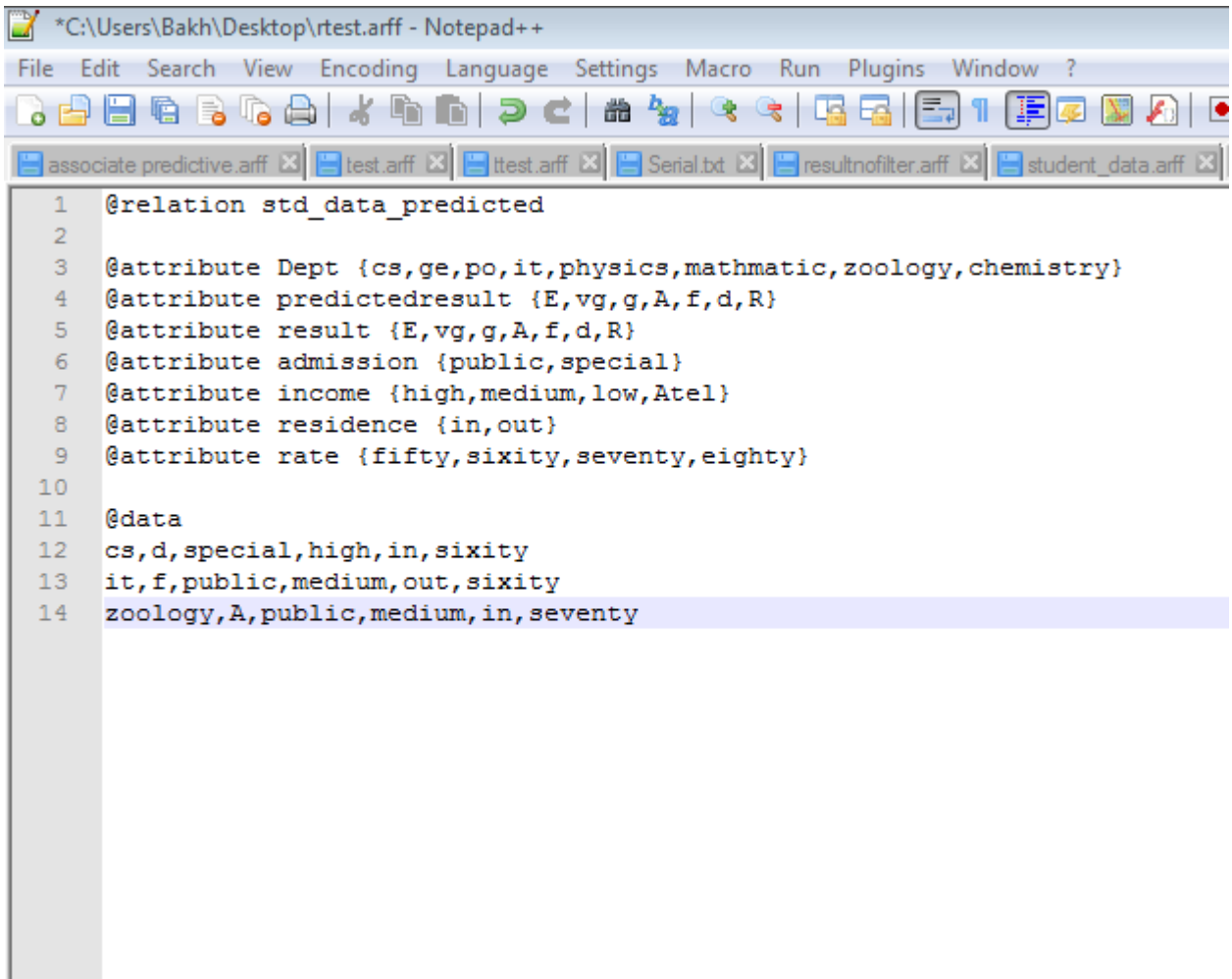


The image shows a Notepad++ window with the following content:

```
D:\Research\classification model final\ttest.arff - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
associate predictive.arff x test.arff x ttest.arff x Serial.txt x resultnofilter.arff x student_data.arff x
1 @relation std_data
2
3 @attribute Dept {cs,ge,po,it,physics,mathmatic,zoology,chemistry }
4 @attribute result {E,vg,g,A,f,d,R}
5 @attribute admission {public,special}
6 @attribute income {high,medium,low,Atel}
7 @attribute residence {in,out}
8 @attribute rate {fifty,sixity,seventy,eighty}
9
10 @data
11
12 cs,?,special,high,in,sixity
13 it,?,public,medium,out,sixity
14 zoology,?,public,medium,in,seventy
15
```

Figure (4.10): Test dataset





```
*C:\Users\Bakh\Desktop\rttest.arff - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
associate_predictive.arff test.arff ttest.arff Serial.txt resultnofilter.arff student_data.arff
1 @relation std_data_predicted
2
3 @attribute Dept {cs,ge,po,it,physics,mathmatic,zoology,chemistry}
4 @attribute predictedresult {E,vg,g,A,f,d,R}
5 @attribute result {E,vg,g,A,f,d,R}
6 @attribute admission {public,special}
7 @attribute income {high,medium,low,Atel}
8 @attribute residence {in,out}
9 @attribute rate {fifty,sixity,seventy,eighty}
10
11 @data
12 cs,d,special,high,in,sixity
13 it,f,public,medium,out,sixity
14 zoology,A,public,medium,in,seventy
```

**Figure (4.11):** Result of test dataset on the model.

#### **4.2 Classification without filtering the attributes:**

click on the Explorer button and get the Weka Knowledge Explorer window, Click on the “Open File..” button and load an ARFF file” our dataset” to start working on the dataset after pressing , the window will appear in the following Figure below:

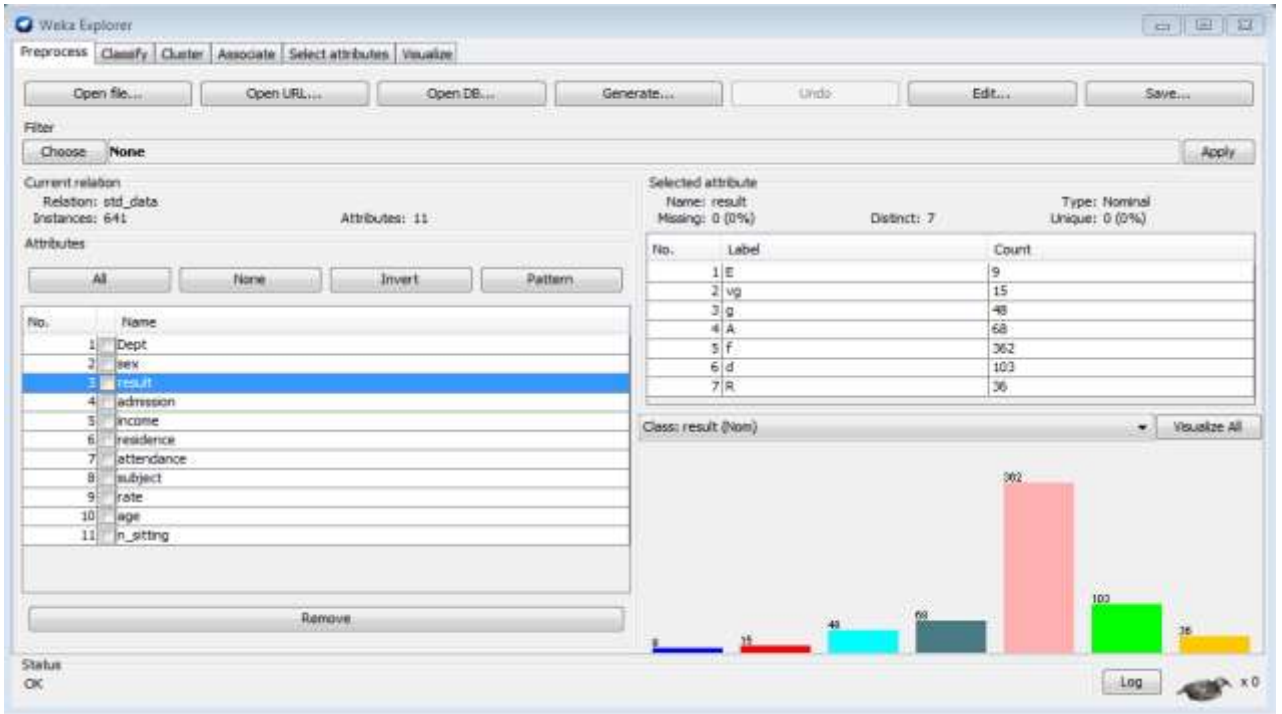


Figure (4.12): Data set uploaded without filtering.



Figure (4.13): Visualize all attributes.

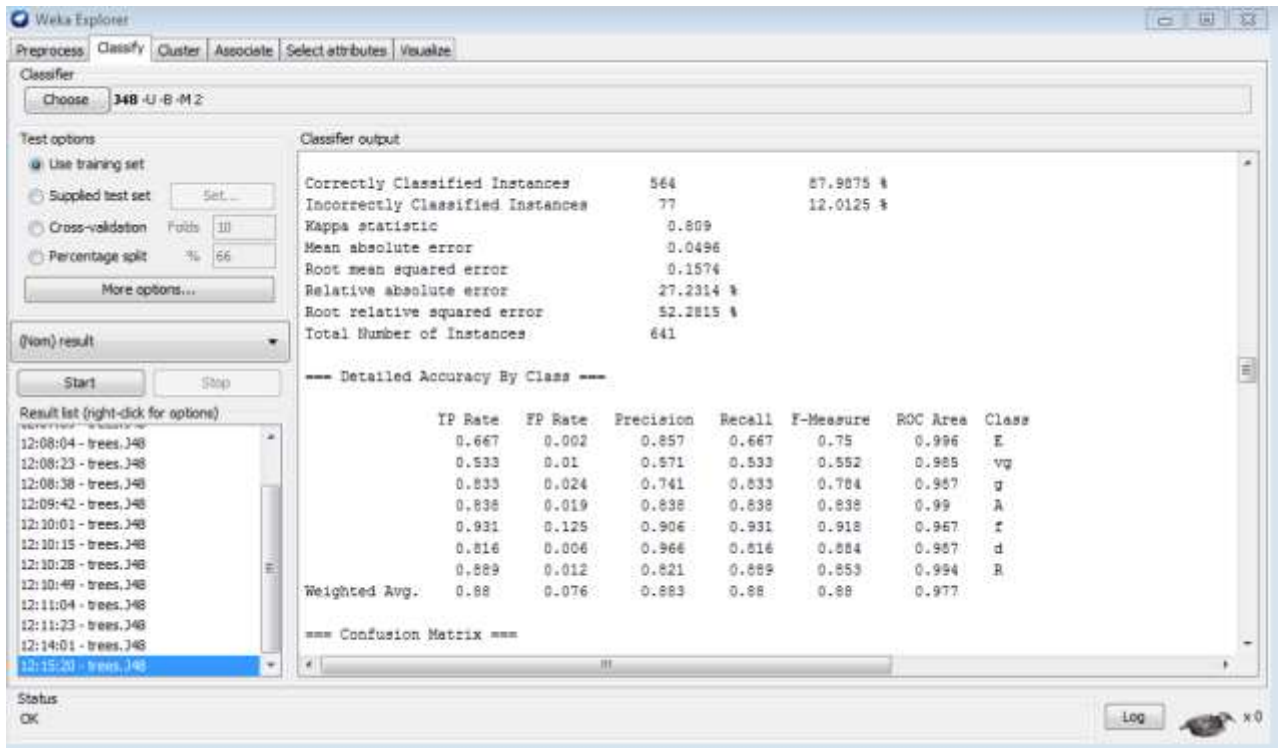


Figure (4.14): Apply j48 classifier.

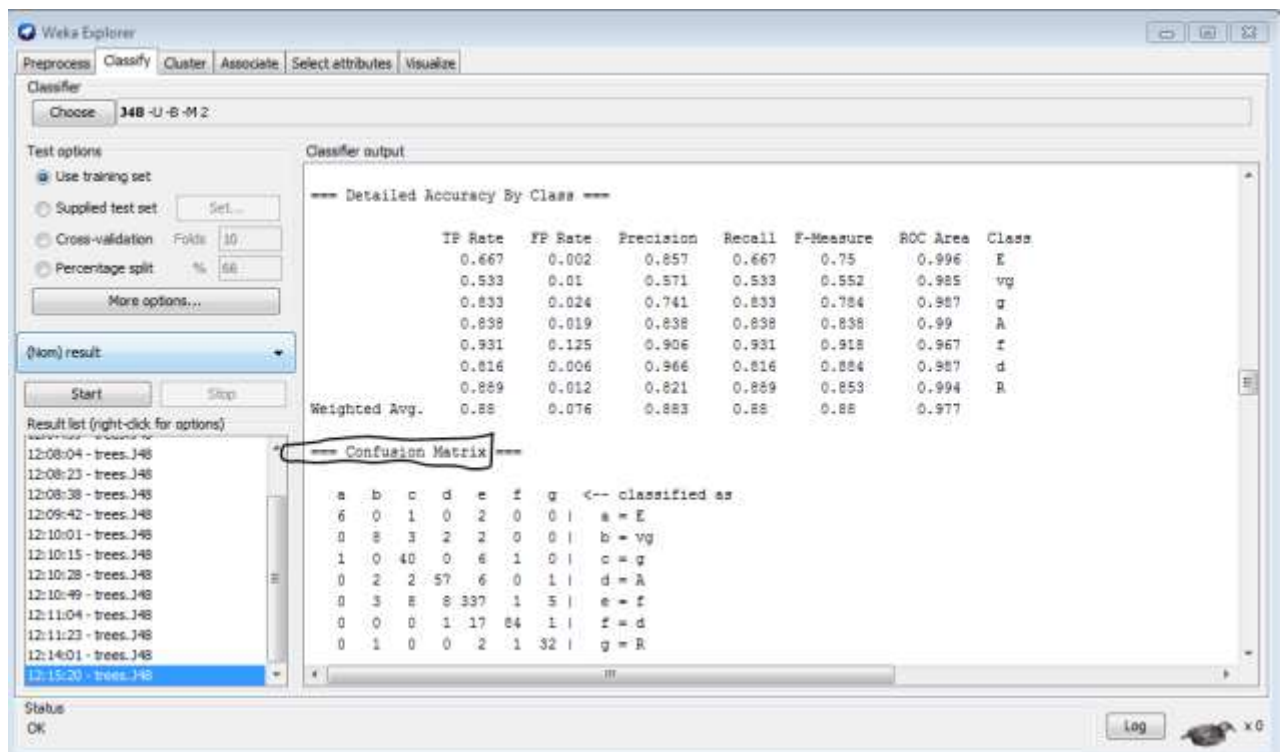


Figure (4.15): Confusion matrix of classifier.

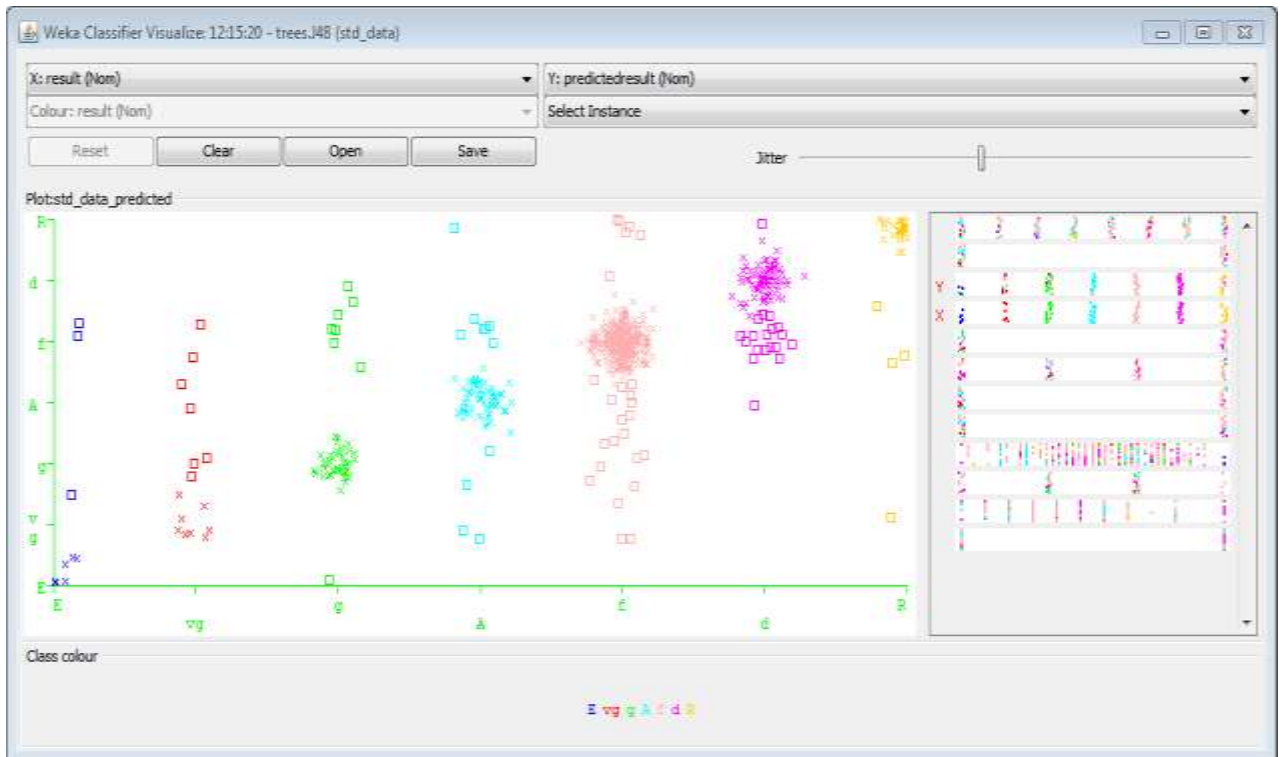


Figure (4.16): Visualize j48 classifier.

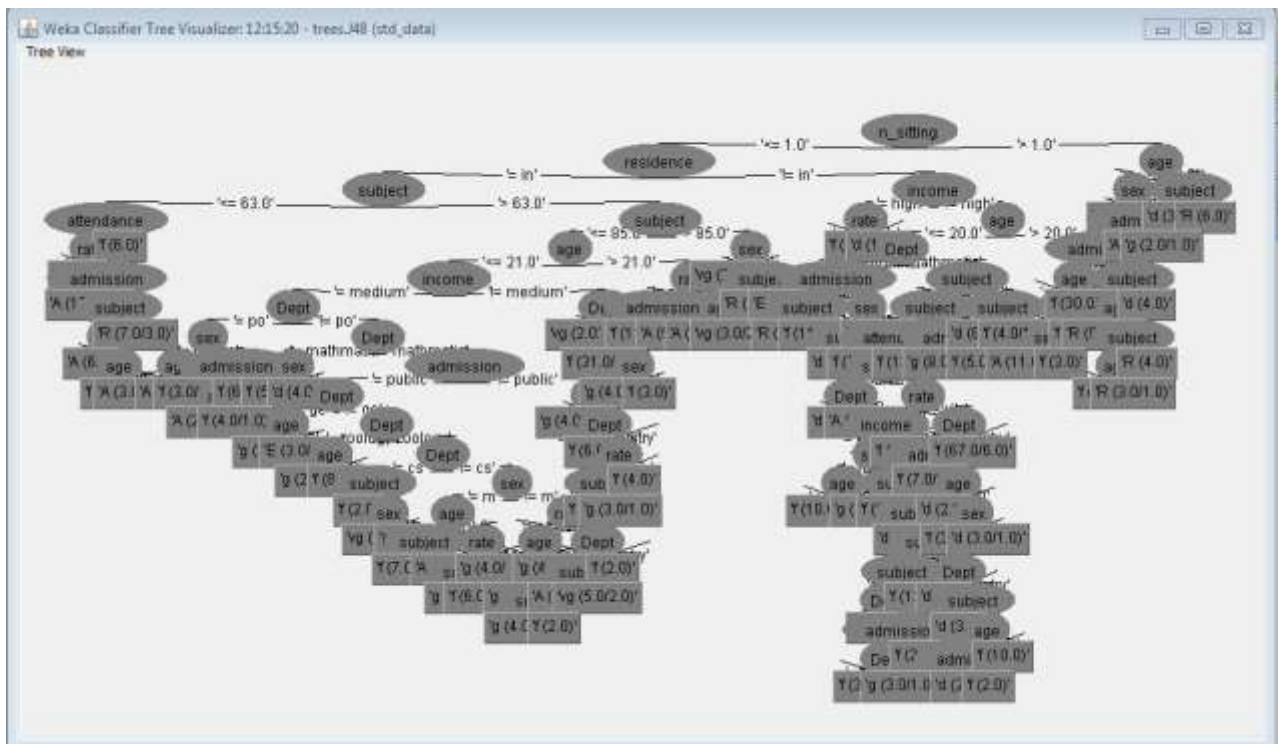


Figure (4.17): j48 classifier tree.

Although, this technique gives better accuracy about (87.22%) compared with classification with filter, but in testing classification with filter gives better result and was near to the real live. so the model of classification with filter was chosen and applied tests on this model.

#### **4.3 Source code of java:**

The model has source code of java, to implement j48 algorithm or want to understand the steps deeply, the code is available in appendix .

#### **4.4 Association rule:**

In the second part of the above methodology we will work on Association rules, which means if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases and Apriori algorithm was applied.

- Upload data in Explorer:

Let's start loading the data in the explorer and start loading it, we start the (Weka program) to get the graphical selection interface as shown in the previous figure (3).

- Then select the explorer from the four available diagrams on the right side of the figure (3).
- In fact, the format shows what is displayed after loading the data basic operations supported by Navigator, currently in the initial processing platform, press the "Open" button you can select the file you want to format. To load the data we specify the attributes that we want to perform the compilation process in general, and also exist an optional option for all assembly requirements. And then by reference to the first lists we choose associate then start, after the program generates the results and outputs as shown in the figure (10).

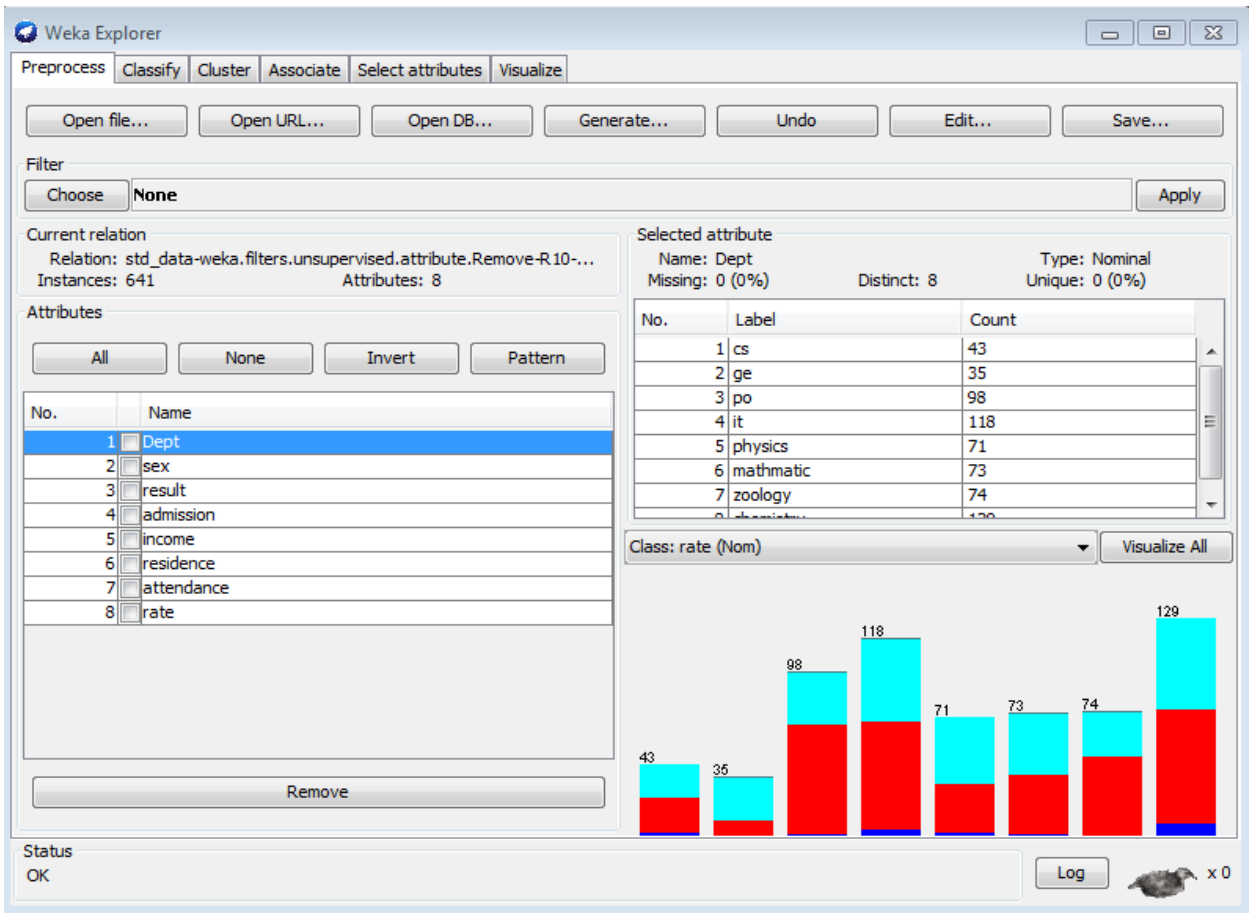


Figure (4.18): Explorer interface in weka.

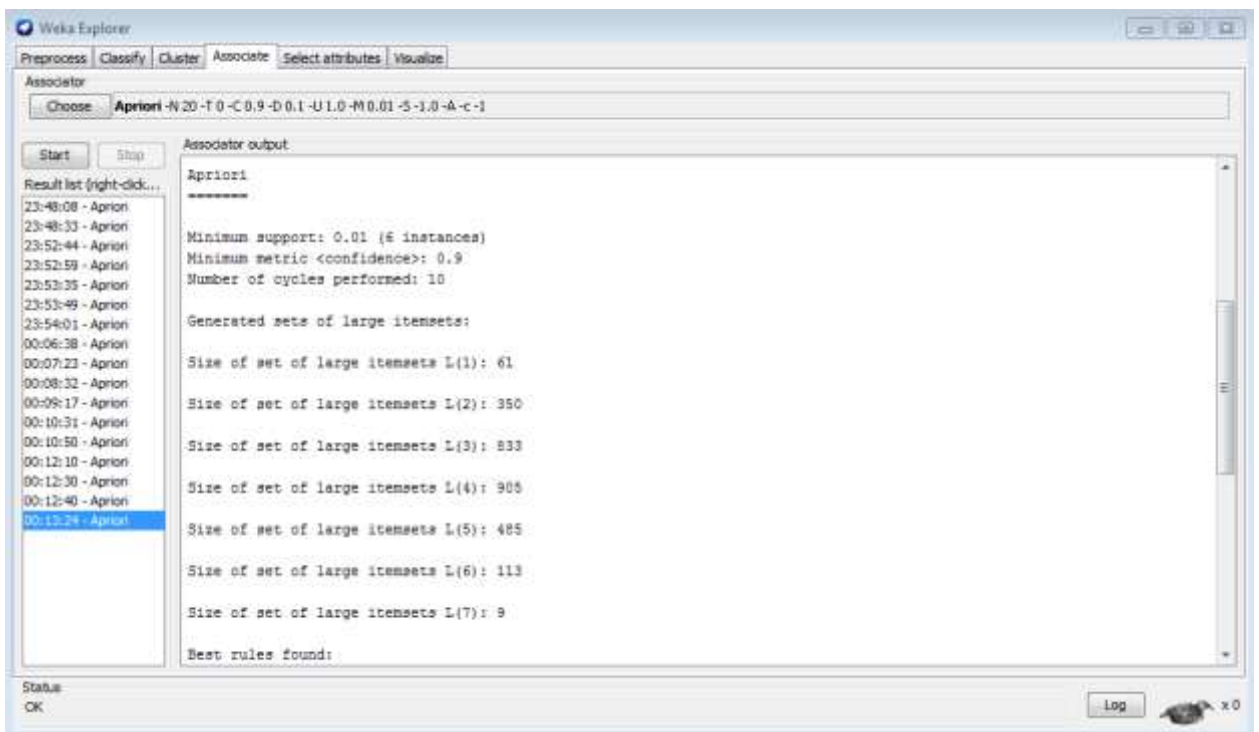


Figure (4.19) : Format of the results after applying Apriori algorithm.

After many experiments were conducted on the Weka program, by changing values(Confidence) and (Support), and change the attributes selected in the configuration phase of the data(Preprocessing), A large amount of rules have been obtained. After loading those rules it is found that, some of them are classified as trivial and some of them have a logical degree of value (Confidence).

**Table (4.2):** Summary of results obtained.

Confidence	Support	n.Rules	Large item set
0.9	0.01	151	7
0.2	0.1	159	5
0.9	0.1	119	6

As a result, we notice that the greater the number of (item set) and the smaller the number of items produced, the more the relations between them are more logical, the change in both the values of "confidence" and " (increase or decrease) causes different rules with varying degrees of logic to appear. There was also a large amount of intuitive results and illogical results when determining all the attributes to find The relations between them, which made me identify certain qualities and conduct operations on them to discover more relationships logical between results.

```

3. Dept=ge admission=special 13 ==> result=f 13    acc:(0.99076)
4. admission=special income=Atel 12 ==> result=f 12    acc:(0.98969)
5. Dept=zooology admission=special rate=seventy 10 ==> result=f 10    acc:(0.9864)
6. Dept=ge sex=m residence=out rate=seventy 9 ==> result=f 9    acc:(0.98383)
7. Dept=zooology sex=m admission=special 8 ==> result=f 8    acc:(0.98027)
8. admission=public income=low rate=seventy 8 ==> result=f 8    acc:(0.98027)
9. Dept=zooology income=low 7 ==> result=f 7    acc:(0.97524)
10. sex=f admission=special rate=fifty 7 ==> result=d 7    acc:(0.97524)
11. sex=m admission=special residence=in rate=seventy 7 ==> result=d 7    acc:(0.97524)
12. Dept=zooology sex=m residence=out 23 ==> result=f 22    acc:(0.97417)
13. income=high residence=out rate=sixty 6 ==> result=f 6    acc:(0.968)
14. rate=eighty 5 ==> result=f 5    acc:(0.95733)
15. Dept=zooology residence=out rate=seventy 19 ==> result=f 18    acc:(0.95466)
16. Dept=physics income=Atel 4 ==> result=f 4    acc:(0.94122)
17. Dept=zooology admission=special residence=in 4 ==> result=f 4    acc:(0.94122)
18. Dept=ge residence=out rate=seventy 16 ==> result=f 15    acc:(0.92856)
19. Dept=zooology sex=m rate=seventy 15 ==> result=f 14    acc:(0.91675)
20. Dept=ge sex=f rate=sixty 3 ==> result=f 3    acc:(0.91616)
21. Dept=zooology sex=m residence=in rate=sixty 3 ==> result=f 3    acc:(0.91616)
22. sex=m admission=public income=Atel residence=in 3 ==> result=g 3    acc:(0.91616)
23. sex=m income=Atel residence=in rate=seventy 3 ==> result=g 3    acc:(0.91616)
24. Dept=zooology sex=m 27 ==> result=f 25    acc:(0.87985)

```

**Figure (4.20):** some of result.

### **Best rules found:**

1. sex=f result=f income=low 22 ==> rate=sixty 22 conf:(1)
2. admission=special income=high residence=out 16 ==> rate=seventy 16 conf:(1)
3. sex=f result=f income=low residence=out 15 ==> rate=sixty 15 conf:(1)
4. sex=m income=high residence=out 14 ==> rate=seventy 14 conf:(1)
5. sex=f admission=public income=low 13 ==> rate=sixty 13 conf:(1)
6. result=A admission=special 12 ==> rate=sixty 12 conf:(1)
7. sex=m result=A residence=out 12 ==> rate=sixty 12 conf:(1)
8. result=d income=high residence=out 12 ==> rate=seventy 12 conf:(1).
9. result=A admission=special income=medium 12 ==> rate=sixty 12 conf:(1)
10. result=A admission=special residence=in 12 ==> rate=sixty 12 conf:(1)
11. sex=m result=A admission=public residence=out 12 ==> rate=sixty 12 conf:(1)
12. sex=m result=A income=medium residence=out 12 ==> rate=sixty 12 conf:(1)
13. sex=m result=E income=high residence=out 12 ==> rate=seventy 12 conf:(1).
14. result=f admission=special income=medium residence=out42 ==> rate=eighty42conf:(1).
15. result=E admission=public income=medium attendance=f 9 ==> rate=sixty 9conf:(1)
16. result=R admission=special residence=out attendance=f 42 ==> rate=seventy42conf:(1).

### **4.5 Results:**

1. Building model that helped to identify if the student's academic qualify to study the track and predict the result that a student can obtain based on available data .
2. The results of the Sudanese certificate are totally untrue or there is jugglery.
3. using association technique help to access to logical rules that can be helpful for evaluation students' performance based on available data, one of the strongest rules has been reached :
  - Most of student's which accepted by high scores, were coming from outside the city or from localities, and achieved bad rates.
  - The results of Sudanese certificate are not really or transparency.
  - The students who were achieved pass rates; their ages are between 17 and 19.



- The Students who was retracted from the university, are the students who have been accepted by high degree and coming from localities.
- The Students who admitted to public admission, were achieve better result than special admission.

#### **4.6 Discussion of results:**

1. The female students, who were accepted in the sixties, were of low income and therefore get a failure result.
2. The Students who were admitted to private admission and rate in the seventy and have high economic income, it's student's examine outside the city.
3. The female students who were accepted in the sixties, low income, examine outside the city, therefore get a failure result.
4. The students were accepted in the seventies, those ones who have high-income and coming from outside the city.
5. The female students, who were admitted to public admission and rate in the sixty, have low economic income.
6. The Students who achieve an acceptable percentage were the student who were admitted to a special admission and were accepted was sixty.
7. The Students which are coming from outside the city or from the localities are the same Students in result 6.
8. The Students, who have been dismissed, are students who come from outside the city or from localities, high-income, and accepted by seventy.
9. The Students, who have achieved an acceptable score, are students who have been accepted by seventy, their acceptance was a special admission, and their income is affordable.
10. The Students who have achieved an acceptable score, are students who have been accepted by seventy, their acceptance was a special admission, and coming from inside the city.
11. The male Students who have achieved an acceptable score, are students who have been accepted by sixty and income was medium , their acceptance was public admission, and coming from outside the city or from localities.(result 11,12).

12. The Students who have achieved an Excellent score, are students who have been accepted by sixty, their acceptance was public admission, and coming from inside the city.
13. The Students who was retracted from the university, are the students who have been accepted by seventy, their acceptance was special admission, full attendance and coming from outside the city.

**Summary of this Chapter:**

In this chapter, the data of the available students were examined. The data were categorized using the J48 algorithm; they were able to predict the students' rate according to the model they designed. Comparisons were made with a number of algorithms, but the J48 algorithm overwhelmed them with better test results. Associated rules have been created and the best rules that have achieved better results have been selected and compared with the real reality.

**CHAPTER FIVE**  
**CONCLUSION AND FUTURE WORK**

## **CHAPTER FIVE**

### **CONCLUSION AND FUTURE WORK**

#### **Conclusion:**

In these study, discussed the various data mining techniques which can support education system via generating strategic information. Since the application of data mining brings a lot of advantages in higher learning institution, it is recommended to apply these techniques in the areas like optimization of resources, prediction of academic results of students, look on the reasons that why students' get high marks in Sudanese Certificate but acquire weak marks in university, to solve the problem of failing in academic results for the students in higher institutions. Also, we can use data mining applications in analyzing the students' academic results for long periods, to improve the results. Data mining can be used in the process of choosing applications to the posts by reducing gap between the numbers of candidates applied for the post, number of applicants.

#### **Future work:**

1. Test other algorithms to design a predictive model other than an algorithm j48.
2. Increase the size of student data and create another model and compare it with the designer model.
3. Establish data center at the University of Nyala to facilitate gets the students, staffs and courses information which helps in data collection.
4. The data should be stored well enough, to be accessible, thus helping to complete research in this area.
5. In order to increase the efficiency of the results obtained from this research, add more characteristics, especially the data of the schools to which the Sudanese certificate was taken.
6. Extensions the students before submitting the application according to their desires.

## References:

- ABU TAIR, M. M. & EL-HALEES, A. M. 2012. Mining educational data to improve students' performance: a case study. *Mining educational data to improve students' performance: a case study*, 2.
- AHMED, A. B. E. D. & ELARABY, I. S. 2014. Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*, 2, 43-47.
- ANOOPKUMAR, M. & RAHMAN, A. M. Z. 2018. Model of Tuned J48 Classification and Analysis of Performance Prediction in Educational Data Mining. *International Journal of Applied Engineering Research*, 13, 14717-14727.
- BASSIL, Y. 2012. A Data Warehouse Design for A Typical University Information System. *arXiv preprint arXiv:1212.2071*.
- BAY, S. D., KIBLER, D. F., PAZZANI, M. J. & SMYTH, P. 2000. The UCI KDD archive of large data sets for data mining research and experimentation. *SIGKDD explorations*, 2, 81-85.
- BHATNAGAR, A., JADYE, S. P. & NAGAR, M. M. 2012. Data mining techniques & distinct applications: a literature review. *International Journal of Engineering Research & Technology (IJERT) Vol*, 1.
- BUNIYAMIN, N., BIN MAT, U. & ARSHAD, P. M. Educational data mining for prediction and classification of engineering students achievement. 2015 IEEE 7th International Conference on Engineering Education (ICEED), 2015. IEEE, 49-53.
- CAMPBELL, J. P., DEBLOIS, P. B. & OBLINGER, D. G. 2007. Academic analytics: A new tool for a new era. *EDUCAUSE review*, 42, 40.
- CIOS, K. J., SWINIARSKI, R. W., PEDRYCZ, W. & KURGAN, L. A. The knowledge discovery process. *Data Mining*, 2007. Springer, 9-24.
- DELMATER, R. & HANCOCK, M. 2001. *Data mining explained: a manager's guide to customer-centric business intelligence*, Digital press.
- HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. & WITTEN, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11, 10-18.
- HAN, J., KAMBER, M. & PEI, J. 2011a. *Data mining concepts and techniques third edition*. Morgan Kaufmann.
- HAN, J., KAMBER, M. & PEI, J. 2012. *Data mining: concepts and techniques*, Waltham, MA. Morgan Kaufman Publishers, 10, 978-1.
- HAN, J., PEI, J. & KAMBER, M. 2011b. *Data mining: concepts and techniques*, Elsevier.
- ILLE, E. 2017. Political, Financial and Moral Aspects of Sudan's Private Higher Education. *Rethinking Private Higher Education*. BRILL.
- JAYAPRAKASH, S. & JAIGANESH, V. 2018. A Survey on Academic Progression of Students in Tertiary Education using Classification Algorithms.
- JINDAL, R. & BORAH, M. D. 2013. A survey on educational data mining and research trends. *International Journal of Database Management Systems*, 5, 53.

- KLÖSGEN, W. & ZYTKOW, J. M. 2002. *Handbook of data mining and knowledge discovery*, Oxford University Press, Inc.
- KLÜVER, C. Steering clustering of medical data in a self-enforcing network (SEN) with a cue validity factor. 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 2016. IEEE, 1-8.
- LINOFF, G. S. & BERRY, M. J. 2011. *Data mining techniques: for marketing, sales, and customer relationship management*, John Wiley & Sons.
- LUAN, J. 2002. Data mining and its applications in higher education. *New directions for institutional research*, 2002, 17-36.
- MAHINDRAKAR, P. & HANUMANTHAPPA, M. 2013. Data mining in healthcare: A survey of techniques and algorithms with its limitations and challenges. *International Journal of Engineering Research and Applications*, 3, 937-941.
- PAL, J. K. 2011. Usefulness and applications of data mining in extracting information from different perspectives.
- PANT, A. 2017. *For Partial Fulfillment of the Requirements for the Degree of Master of Computer Information System Awarded*. Pokhara University.
- PARK, H.-S. & BAIK, D.-K. 2006. A study for control of client value using cluster analysis. *Journal of Network and Computer Applications*, 29, 262-276.
- PATIL, P. 2015. A Study of Student's Academic Performance Using Data Mining Techniques. *International Journal of Research in Computer Applications and Robotics*. ISSN, 2320-7345.
- ROMERO, C. & VENTURA, S. 2007. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33, 135-146.
- SIEMENS, G. & D BAKER, R. S. Learning analytics and educational data mining: towards communication and collaboration. Proceedings of the 2nd international conference on learning analytics and knowledge, 2012. ACM, 252-254.
- SIVASAKTHI, M. Classification and prediction based data mining algorithms to predict students' introductory programming performance. 2017 International Conference on Inventive Computing and Informatics (ICICI), 2017. IEEE, 346-350.
- VELMURUGAN, T. & ANURADHA, C. 2016. Performance evaluation of feature selection algorithms in educational data mining. *Performance Evaluation*, 5.
- YU, C. H., DIGANGI, S., JANNASCH-PENNELL, A. & KAPROLET, C. 2010. A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, 8, 307-325.
- ZIEGEL, E. R. 2001. *Mastering Data Mining*. Taylor & Francis.

## **CHAPTER SIX**

### **Appendix**

## Appendix

A	B	C	D	E	F	G	H	I	J	K
cs	m	g	public	medium	out	f	81	71	19	1
cs	m	f	special	medium	in	uf	51	59	17	2
cs	f	d	special	high	in	uf	66	60	21	2
cs	f	d	special	high	in	f	67	61	22	2
cs	f	f	special	medium	in	f	56	63	20	1
cs	m	f	public	medium	in	f	72	66	22	1
cs	m	f	public	high	in	f	69	65	21	1
cs	f	g	public	medium	in	f	74	70	19	1
cs	m	f	special	medium	out	f	68	69	22	1
cs	m	vg	public	high	in	f	70	71	19	1
cs	f	f	public	medium	in	f	70	63	19	1
cs	m	vg	public	medium	in	f	76	73	17	1
cs	f	f	special	low	out	uf	68	64	18	1
cs	f	A	special	medium	in	f	56	64	22	1
cs	f	vg	public	medium	in	f	74	72	23	1
cs	f	A	public	low	out	f	59	65	18	1
cs	f	A	public	medium	out	f	60	71	18	2
cs	m	vg	public	medium	in	f	86	72	19	1
cs	f	d	public	medium	out	uf	59	60	21	2
cs	m	f	public	low	out	f	63	71	20	2
cs	m	d	special	low	out	f	65	61	21	2
cs	f	A	public	medium	in	f	60	65	22	1
cs	m	R	special	medium	in	f	59	61	18	1
cs	m	A	public	medium	in	f	60	65	20	1
cs	m	f	public	medium	in	f	65	70	18	1



ge	m	E	public	medium	in	f	79	72	19	1
ge	f	f	special	medium	out	uf	80	75	20	1
ge	f	g	public	medium	in	f	81	72	18	1
ge	f	f	special	medium	in	f	70	69	18	1
ge	m	g	public	atel	in	uf	75	71	19	1
ge	m	A	public	medium	out	uf	64	64	18	1
ge	f	f	special	medium	in	uf	54	60	21	1
ge	m	f	public	medium	out	uf	84	70	22	1
ge	m	f	special	medium	out	f	81	77	20	1
ge	f	f	public	medium	out	f	79	76	18	1
ge	f	f	special	medium	out	f	80	78	22	1
ge	m	f	public	medium	out	f	78	75	22	1
ge	f	E	public	medium	in	f	90	76	18	1
ge	m	f	special	medium	out	uf	81	79	19	2
ge	m	g	public	medium	in	f	74	71	18	1
ge	m	f	public	medium	out	f	80	75	19	1
ge	m	f	public	medium	out	f	57	61	19	2
ge	f	vg	public	medium	out	f	81	73	18	1
ge	m	A	public	medium	out	f	75	69	21	1
ge	m	f	special	Atel	in	uf	70	63	18	1
ge	m	f	public	low	out	f	79	78	20	1
ge	m	f	special	medium	out	f	80	77	20	1
ge	m	f	public	medium	out	f	85	80	25	1
ge	f	f	public	high	in	f	70	65	18	1
ge	m	f	public	medium	out	f	80	76	20	1

ge	m	E	public	medium	in	f	79	72	19	1
ge	f	f	special	medium	out	uf	80	75	20	1
ge	f	g	public	medium	in	f	81	72	18	1
ge	f	f	special	medium	in	f	70	69	18	1
ge	m	g	public	atel	in	uf	75	71	19	1
ge	m	A	public	medium	out	uf	64	64	18	1
ge	f	f	special	medium	in	uf	54	60	21	1
ge	m	f	public	medium	out	uf	84	70	22	1
ge	m	f	special	medium	out	f	81	77	20	1
ge	f	f	public	medium	out	f	79	76	18	1
ge	f	f	special	medium	out	f	80	78	22	1
ge	m	f	public	medium	out	f	78	75	22	1
ge	f	E	public	medium	in	f	90	76	18	1
ge	m	f	special	medium	out	uf	81	79	19	2
ge	m	g	public	medium	in	f	74	71	18	1
ge	m	f	public	medium	out	f	80	75	19	1
ge	m	f	public	medium	out	f	57	61	19	2
ge	f	vg	public	medium	out	f	81	73	18	1
ge	m	A	public	medium	out	f	75	69	21	1
ge	m	f	special	Atel	in	uf	70	63	18	1
ge	m	f	public	low	out	f	79	78	20	1
ge	m	f	special	medium	out	f	80	77	20	1
ge	m	f	public	medium	out	f	85	80	25	1
ge	f	f	public	high	in	f	70	65	18	1
ge	m	f	public	medium	out	f	80	76	20	1

ge	t	vg	public	medium	in	t	77	74	22	1
ge	f	f	special	medium	in	f	80	70	21	1
it	f	f	public	medium	out	f	65	71	20	1
it	f	f	public	medium	out	f	67	68	19	1
it	m	f	special	medium	out	f	72	70	20	1
it	m	f	special	medium	out	f	69	71	20	1
it	m	f	public	medium	in	f	77	68	21	1
it	f	f	public	medium	out	f	62	65	20	1
it	f	f	public	medium	out	f	72	69	20	1
it	f	f	public	medium	out	f	74	68	23	1
it	m	f	public	medium	out	f	61	61	18	1
it	m	g	public	medium	out	f	64	69	20	1
it	m	A	public	medium	in	f	75	70	19	1
it	m	g	public	medium	in	f	66	60	20	1
it	m	f	public	medium	out	f	75	71	18	1
it	m	f	public	medium	out	f	80	79	19	1
it	f	f	public	medium	out	f	78	68	20	1
it	f	f	special	medium	out	f	75	72	20	1
it	f	f	special	high	in	f	65	65	21	1
it	f	f	public	Atel	in	f	80	69	18	1
it	f	f	special	medium	in	f	66	65	18	1
it	f	f	special	medium	out	f	76	65	18	1

no_stude	Dept	sex	result	admission	income	rate	age	residence	subject	n_sitting	attendanc
1	cs	m	A	public	medium		70	22 in		70	1 f
2	cs	m	g	public	medium		71	19 out		81	1 f
3	cs	m	f	special	medium		59	17 in		51	2 uf
4	cs	f	d	special	high		60	21 in		66	2 uf
5	cs	f	d	special	high		61	22 in		67	2 f
6	cs	f	f	special	medium		63	20 in		56	1 f
7	cs	m	f	public	medium		66	22 in		72	1 f
8	cs	m	f	public	high		65	21 in		69	1 f
9	cs	f	g	public	medium		70	19 in		74	1 f
10	cs	m	f	special	medium		69	22 out		68	1 f
11	cs	m	vg	public	high		71	19 in		70	1 f
12	cs	f	f	public	medium		63	19 in		70	1 f
13	cs	m	vg	public	medium		73	17 in		76	1 f
14	cs	f	f	special	low		64	18 out		68	1 uf
15	cs	f	A	special	medium		64	22 in		56	1 f
16	cs	f	vg	public	medium		72	23 in		74	1 f
17	cs	f	A	public	low		65	18 out		59	1 f
18	cs	f	A	public	medium		71	18 out		60	2 f
19	cs	m	vg	public	medium		72	19 in		86	1 f
20	cs	f	d	public	medium		60	21 out		59	2 uf
21	cs	m	f	public	low		71	20 out		63	2 f
22	cs	m	d	special	low		61	21 out		65	2 f
23	cs	f	A	public	medium		65	22 in		60	1 f
24	cs	m	R	special	medium		61	18 in		59	1 f
25	cs	m	A	public	medium		65	20 in		60	1 f