Sudan University of Science and
Technology College of Graduate Studies

Student Performance Prediction Using Classification based on their social factors case study Ahfad University for Women

التنبؤ بأداء الطالب باستخدام التصنيف على أساس عواملهم الاجتماعية
(دراسة حالة جامعة الأحفاد للبنات)

A dissertation Submitted in Partial Fulfillment of the Requirements
For MSc Degree in Information Technology

Prepared by:

Fatima Sayed Alsheikh

Supervised by:

Dr. Shaza Mergani

May 2021

Acknowledgement

Primary, I would thank God for being able to complete this research then I would like to thank my Supervisor dr.Shaza Mergani whose valuable guidance has been the once that helped me patch this project, her suggestions and her instructions has save as major contributor toward the completion of this research.

I have taken efforts in this research. However, it would not have been possible without the kind support and help of my friends (Rasha, Elham, and Omnia). I would like to express my special thanks of gratitude to my parents, brothers and sisters who in one way or another shared their support either psychologically or physically in completing my research. Thank you

Abstract

Student's performance is an essential part in learning institutions. Predicting student's performance becomes more challenging due to the large volume of data in educational databases. The adoption of the educational data mining by higher education as an analytical and decision making tool is offering new opportunities to predict student performance. The university management would like to know which features in the currently available data are the strongest predictors of university performance. In order to help the academic advisor to monitor the students' performance in a systematic way by identifies those students which needed special attention to reduce failing ration and taking appropriate action for the next semester at a right time. To meet these objectives the researcher used CRISP-DM Methodology which governs by a series of stages. Starting by business understanding followed by data understanding, data preparation, modeling evaluation and deployment. Many experiments conducted to find out a model that could be useful for predicting students' performance based on their social factors using decision tree (j48, random forest) and Bayesian classifiers (naïve Bayes, Bayes net) as classification techniques. The experimental results showed that J48 is the best algorithm for classification of data. It also showed that social factors have got significant influence over students' performance

الخلاصه

أداء الطالب هو جزء أساسي في المؤسسات التعليمية.اصبح توقع أداء الطالب أكثر صعوبة بسبب الحجم الكبير للبيانات في قواعد البيانات التعليمية. تطبيق التنقيب عن البيانات التعليمية من قبل التعليم العالي كأداة تحليلية وصنع القرار اتاح فرصًا جديدة للتنبؤ بأداء الطلاب.. تود الجامعة معرفة الميزات الموجودة في البيانات المتاحة حاليا واثرها على اداء الطلاب. من أجل مساعدة المرشد الأكاديمي على مراقبة أداء الطلاب بطريقة منهجية من خلال تحديد هؤلاء الطلاب الذين يحتاجون إلى عناية خاصة لتقليل نسبة الفشل واتخاذ الإجراءات المناسبة للفصل الدراسي التالي في الوقت المناسب. لتحقيق هذه الأهداف استخدم الباحث منهجية CRISP-DM التي تحكمها سلسلة من المراحل. البدء بفهم الأعمال متبوعًا بفهم البيانات وإعداد البيانات وتقييم النمذجة والنشر. تم إجراء العديد من التجارب لاكتشاف نموذج يمكن أن يكون مفيدًا للتنبؤ بأداء الطلاب بناءً على عواملهم الاجتماعية باستخدام decision tree (j48, random forest) and Bayesian classifiers (naïve Bayes, Bayes net) كتقنيات تصنيف. أظهرت النتائج التجريبية أن خوارزمية J48 هي أفضل خوارزمية لتصنيف البيانات. كما أظهر أن العوامل الاجتماعية كان لها تأثير كبير على أداء الطلاب

Table of Content

List of Table

List of Figure:

List of Abbreviations

Abbreviations	**Meaning**
AUW	Ahfad University for Women
IHL	Institutions of Higher Learning
DM	Data Mining
EDM	Educational Data Mining
KDD	Knowledge Discovery in Databases
FCM	Fuzzy C Means
EM	Expectation Maximization
GPA	Grade point average
MCA	Master of Computer Applications
TPR	True Positive Rate
FPR	False Positive Rate

CHAPTER I

Introduction

1.1 Introduction

Nowadays, the Institutions of Higher Learning (IHL) database contains so much information about their students. The information is kept increasing by times, but there is no action taken to gain knowledge from it. Data Mining (DM) is the suitable techniques in managing the IHL data to discover new information and knowledge about students. (Ahmad, F., Ismail, N., & Aziz, A. A. (2015). Educational data mining is a new emerging technique of data mining that can be applied on the data related to the field of education. **(**Kumar, S. A. (2011). Educational Data Mining (EDM) is a new trend in the data mining and Knowledge Discovery in Databases (KDD) field which focuses in mining useful patterns and discovering useful knowledge from the educational information systems, such as, admissions systems, registration systems, course management systems (moodle, blackboard, etc…), and any other systems dealing with students at different levels of education, from schools, to colleges and universities. Saa, A. A. (2016). . Educational Data Mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor, and many others.

Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for prediction regarding enrolment of students in a particular course, detection of abnormal values in the result sheets of the students, prediction about students‟ performance and so on Yadav, S. K., & Pal, S. (2012).

1.2 Statement of Problem

The new discovery about students' learning behaviors and the factors contribute to the students' success should be exposed for the community benefits. To discover hidden information and knowledge from the students' data, a few elements such as parameters, methods, and tools need to be identified and considered in order to produce the best model prediction.

1.3 Objective

1- To collect quantitative data that represent social factors of students

2- To build the classification model that classifies a students' performance.

3- To evaluate the performance of different classification techniques.

4- Is to find out if there are any patterns in the available data that could be useful for predicting students' performance at the university based on their social factors. The university management would like to know which features in the currently available data are the strongest predictors of university performance.

1.4 Research Scope

This study focuses on predicting student performance, this prediction will modify the educational process, depending on student achievements. This prediction will motivate the student to increase their marks. The prediction will made on AUW School of management studies freshmen year 2018-2019.

1.5 Significant of The research

.This research is important for higher education institutions, given that the strategic planning of study programs implies expanding or reducing the scope or depth of the curriculum as well as modifying the educational process, depending on student achievements.

1.6 Research organization

This research consist of five chapters organized as follows: Chapter one contains introduction. Chapter two discusses the literature review and related work. Chapter three describes the methodology used by researcher. Chapter four presents the results analysis and their discussion. Lastly, Chapter five contain conclusion and recommendation of future works.

CHAPTER II

Literature review

2.1 Introduction

This chapter consists of two parts, part one demonstrates the data mining and educational data mining concepts, and applications of data mining used in educational sector. The other part shows the related works in the same field of study and contains a summary table that mentions the advantages and limitations in the researcher point of view.

2.2 Data Mining

The amount of data stored in databases is increasing growing in a tremendous speed. This growing need for new techniques and tools to intelligently analyzing huge data sets to gather useful information. This growing need gives birth to a new research field called data mining, also KDD, (Minaei-Bidgoli, 2004) DM is the field of discovering novel and potentially useful information from large amounts of data. DM has been used in areas such as database systems, data warehousing, statistics, machine learning, data visualization, and information retrieval.(Suhirman, Zain and Herawan, 2014).

One of the definitions of DM "is the process of discovering interesting knowledge from large amount of data stored in database, data warehouse, World Wide Web, or other information repository .(Romero and Ventura, 2007) Another, definition "Data Mining is a process that consists of applying data analysis and discovery algorithms that, un-der acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data" (Feyyad, U. M. (1996), it is the task being performed, to generate the desired result"(Choudhury *et al.*, 2017). DM techniques are used to build a model according to which the unknown data identify the new information (Review, 2012).

2.3 Educational data mining (EDM)

The implementation of data mining in the educational sector, recently defined as educational data mining (EDM). EDM is an emerging trend, concerned with developing methods for exploring the huge data that come from the educational system (Barahate, 2012). The EDM research community is constantly growing, starting by organizing workshops since 2004 (Kabakchieva, 2012)**,** They increasing research interests in using data mining in education. EDM is a new emerging technique of data mining that can be applied on the data related to the field of education, currently the huge amount of data stored in educational database contain useful information for predict of students' performance..

According to the international consortium on educational data mining EDM is defined as “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 they learn in “(Srivastava, Jaideep; Cooley, R; Deshpande, M; Tan, 2000) . EDM is the process of transforming raw data compiled by education systems in useful information that could be used to take informed decisions and answer research questions and they have a greater impact on educational research and practice .(Kabakchieva, 2012) Educational Data Mining can be seen as an iterative cycle of hypothesis formation, testing and refinement as shown in figure 2.1.

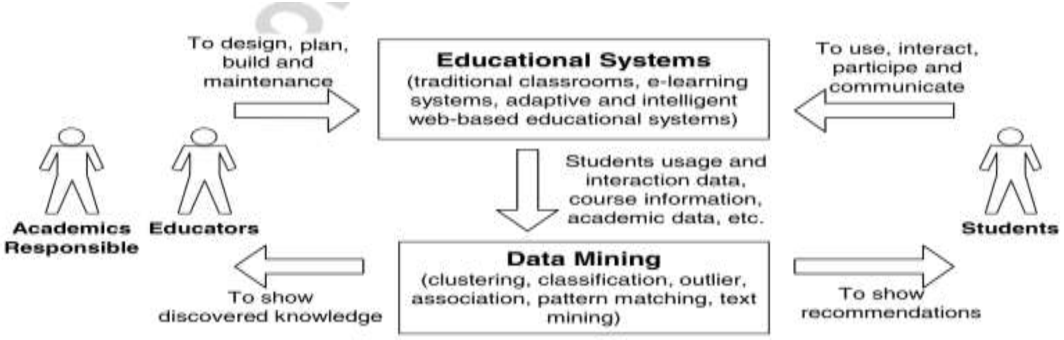

To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Students usage and interaction data, course information, academic data, etc.
Academics Responsible
Educators
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Figure 2.1: The cycle of applying data mining in educational systems

Figure 2.1 shows academic educator section responsibility and are in charge of planning, designing, building and maintaining the educational systems. Students use and interact with them. Starting from all the available information about courses, students, usage and interaction data course information, academic data, different data mining technique can be applied in order to discover useful knowledge that helps to improve the learning process. The discovered knowledge can be used not only by providers (lecturers) but also by own users (students). So, the application of data mining in educational systems can be oriented to different actors with each particular point of view .so in order to get required data & to find the hidden relationship, different data mining techniques are developed & used. Traditionally researchers have applied data mining methods like clustering, classification, prediction to educational context.

2.3.1 Data mining used techniques in educational systems

The main functionality of data mining techniques is applying various methods and algorithms in order to discover hidden patterns and relationships helpful in decision making.(Suhirman, Zain and Herawan, 2014) These interesting patterns are presented to the user and may be stored as new knowledge in knowledge base.(Baradwaj and Pal, 2012) there are many type of data mining techniques used in educational system such as clustering, classification, prediction and many others techniques (Romero and Ventura, 2007). In this research the researcher used classification techniques it is the most powerful techniques in EDM.

1- Classification

Classification is the most commonly applied data mining technique. Classification is the process of data management model building that identifies in-group data to illustrate the differences between groups of data and to predict the data that should be in any class (Cheewaprakobkit, 2015). The data classification process involves the model used to classify data into determined groups is based on an analysis of the data set. This data set would lead the system to classify data by learning and testing. In Learning the training data are analyzed by classification algorithm. In classification test, the rest of the data, as the actual data, will be drawn to test and compare with those acquired from the model, classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples (Suhirman, Zain and Herawan, 2014) The model will be updated and tested to have a satisfactory level. Later, when new data comes and is plugged into the model, the data can predict grouping by the model. The Classification rule used in this research, Decision Tree (J48, Random forest) Bayesian (naïve Bayes, Bayes net).

A- Decision Tree

A decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. (Yadav, S. K., & Pal, S. (2012). The advantages of decision trees are that they represent rules which could easily be understood and interpreted by users,(Kabakchieva, 2013) the tow decision tree algorisms filters applied on the dataset are the j48 and the random forest.

- **J48**

The J48 algorithm is WEKA's implementation of the C4.5 decision tree learner. The algorithm uses a greedy technique to induce decision trees for classification and uses reduced-error pruning. The C4.5 algorithm was proposed in 1992, by Ross Quinlan, to overcome the limitation of the ID3 algorithm (unavailable values, continuous attribute value ranges, pruning of decision trees, etc.). C4.5 uses a divide-and-conquer approach to growing decision trees. The default splitting criterion used by C4.5 is gain ratio, an information-based measure that takes into account different number of test outcomes (Quinlan, R. J. (1996).

- **Random Forest**

The random forest classifier. Which is a tree-based classifier consist of a combination of tree classifiers where each classifiers generated using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class to classify an input vector (Pal, 2005) .The random forest classifier used for this study consists of using randomly combination of features at each node to grow a tree. Bagging. In random forest each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node(Wiener, 2003).

B- Bayesian

One of the major statistical methods in data mining is Bayesian inference**.**

- **Naïve Bayes**

The naïve Bayesian classifier provides a simple and effective approach to classifier learning (Minaei-Bidgoli, 2004). Naive Bayes is a famous classifier which consists on conditional probabilities. Indeed, it employs Bayes' theorem and it supposes that features have a solid independence between each other. However, it has many advantages such as simplicity of use, quick convergence, and high scalability. Finally, Naïve Bayes needs less training data for building a model (Alaoui, Farhaoui and Aksasse, 2018)**.**

2- Prediction

It is used to predict unknown or missing values most commonly used prediction technique is regression analysis. It consists of one or more than one predictor variables (Srivastava and Srivastava, 2013). Prediction is based on the relationship between a thing that is known and a thing need to be predicted In educational data mining prediction can be used to detect student behavior, predicting or understanding student educational outcomes (Choudhury *et al.*, 2017).

3- Clustering

Clustering is referred to as unsupervised learning, it is similar to classification except that the groups are not predefined.(Srivastava and Srivastava, 2013) Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group, called clusters such that “similar” objects fall into the same groups In educational data mining, clustering has been used to group the students according to their behavior(Choudhury *et al.*, 2017).

2.3.2 Applications of Data Mining in Education Sector

There are varieties of popular data mining application within the educational data mining. The most important of them explain as the following:

1- Predicting Students’ Admission in Higher Education

Each student after complete certain course they are taking admission in new course only after screening various factors that are considered important for their overall growth (Srivastava and Srivastava, 2013).Data mining can help management to identify the demographic, geographic and psychographic characteristics of students based on information provided by the students at the time of admission (Kaur, 2015).

2- Predicting Students’ Profiling

EDM can also be used as an effective tool in profiling students based on their skills. Such as academic background, grades and achievements , communication, behavior, attitude, hobbies(Srivastava and Srivastava, 2013).

3- Predicting Students’ Performance

Data Mining is most popularly used to predict performance of students. In educational sector, the mostly predicted values are student’s performance, their marks, knowledge or score(Kaur, 2015).Different techniques and models are applied for prediction of student’s performance particularly classification techniques such as Naïve Bayes and Decision tree based on students ID and marks scored in course (Suhirman, Zain and Herawan, 2014)can be used to predict student’s success in a course also to predict student’s final grade on the basis of features taken from logged data.

4- Teachers’ teaching performance

Teachers' performance evaluation have recently attracted considerable attention and support among researchers and policy makers, There can be various measures to judge teacher’s teaching performance(Kaur, 2015) Student feedback is a popular measure but often it gives skewed results. Because there is high correlation found between marks of the student and feedback of the teacher.

5- Students’ Targeting

Targeting and positioning students is about choose Right student for right kind of course to achieve student satisfaction. It is organized with three dimensions of students targeting: students Value (usage and behavior), student’s characteristics (demographic and Psychographic), and student’s needs (complaints and satisfaction). (Srivastava and Srivastava, 2013).

6- Predicting Students’ course selection

Selection of course by a student depends on various factors such as characteristics, students’ workload, course grades, course type, course duration, and number of time conflicts, final examination time and students’ demand. These factors are used as input of the model Furthermore, by analyze and predict student course satisfaction using this input the researchers found that number of students enrolled to a course and high distinction rate in final grading are the two most influential factors to student course satisfaction (Srivastava and Srivastava, 2013).

7- Predicting Students’ Placement opportunities

Another big challenge in higher education is providing placement to students. Most of the institutions are struggling in this domain. With students becoming more and more demanding, quality placement of students is not only crucial but also very important in creating brand for institutes (Srivastava and Srivastava, 2013).

8- Planning and scheduling

Planning for future courses and course scheduling help the in admission and counseling processes. Different data mining techniques used for this task to analyze enrollee’s course preferences and course completion rates in extension education courses (Kaur, 2015).

2.4 Relative Works

In this research (Hooshyar, Pedaste and Yang, 2020) used student's assignment submission behavior to predict the performance of students with learning difficulties through procrastination behavior (called PPP). Unlike many existing works. PPP firstly builds feature vectors representing the submission behavior of students for each assignment, then applies a clustering method to the feature vectors for labelling students as a procrastinator, procrastination candidate, or non-procrastinator, and finally employs and compares several classification methods to best classify students. To evaluate the effectiveness of PPP, the researcher used a course including 242 students from the University of Tartu in Estonia. The results reveal that PPP could successfully predict students' performance through their procrastination behaviors with an accuracy of 96%. Linear support vector machine appears to be the best classifier among others in terms of continuous features, and neural network in categorical features, where categorical features tend to perform slightly better than continuous, Finally, the researcher found that the predictive power of all classification methods is lowered by an increment in class numbers formed by clustering According to findings, regarding the average of all performance metrics at different k-fold for all classification methods, among all classification methods, L-SVM and R-SVM are the best in two-class, at different k-fold.

(Tampakas *et al.*, 2019) In this study the researcher presented a model for predicting the years taken for a student to complete a bachelor's degree study on School of Health & Social Welfare of Technological Institute of Western Greece. The primary goals of the present research are the accurate and early identification of the students who are at-risk of not completing their studies within six years and the accurate classification of students who have successfully graduated .The researcher collected 282 student records over four years (2010-2013) .. The researcher adopted a two-stage methodology, where the first stage concerns data collection and data preparation, while the second one deploys the proposed two-level classification algorithm. The researcher used (Naive Bayes, Back-Propagation, Sequential Minimal Optimization (SMO), C4.5, JRip, 10NN) classification algorisms, 10NN illustrated the best performance as A-level classifier, since it exhibits the highest accuracy of correctly identifying students who managed to graduate (or not), 98.99% and C4.5 reports the best performance as B- level classifier, illustrating the highest percentage of correctly classified students who have successfully graduated 78.73%.

(Govindasamy, 2018) This study is designed to study and compare four clustering algorithms k-Means, k-Medoids, Fuzzy C Means (FCM) and Expectation Maximization (EM). Data is collected from private Arts and Science Colleges from various departments. More than 1531 student's details are collected with their performance in the Seminar and Assignment. The data are mainly used for evaluating the performance of various clustering algorithms to predict the academic performance of the students in their end of the semester examinations. The cluster quality is evaluated using the number of clusters, execution time, purity, and NMI. Distribution of requirements data set among the clusters: The total number of clusters is three (Average, Good, Excellent). The result showed that FCM and EM algorithm performs well compared with other two clustering algorithms.

(Bashar, 2018) This study proposed a recommendation system to identify weak academic students as soon as possible to help them in a suitable time, encourage students to study hard when they know that they are at risk and to plan their workload carefully. The academic data have been collected from the University of KORDOFAN, faculty of computer studies and statistics. The data contains 1620 records corresponding to students enrolled through the year's study 2008 to 2014. The researcher applied hybrid recommendation techniques which are clustered based on collaborative filtering algorithm and Association rule algorithms Association rule algorithms on each cluster, of which leads to generate strong rules in each year study. The intersection is applied to strong rules which help in recommending students to care which courses are positive and negative effectiveness on GPA. The study also applied Association rule algorithms without using hybrid recommendation techniques. The experiments showed that the applied hybrid recommendation techniques are better.

(Science *et al.*, 2018) proposed an automation for student placement prediction system. The system analyzed the previous year's student's historical data and predict placement chances of "current students" and percentage placement chance of the institution. Students having a better chance of placement are characterized as "good", if not "bad". The system mainly concentrates on student knowledge, skill and attitudes. The system clusters the students based on the characteristics, skills and attitude of the students. The system makes use of previous data to predict the future. The prediction helps to increase the placement rates by helping teachers and placement cell in an institution to coach for the students.

(Gadhavi and Patel, 2017) this study was made to help the student to know his/her performance in advance by using univariate linear regression model. The researcher collected the marks of internal exam components of one subject to predict the final grade in that subject. The internal marks are normalized to 100 (percentage) to have accurate results. The model provides predicted grade of final examinations in particular subjects. It also helps students to know how many marks in the internal examination are required to get particular grade. The model is tested on the same data set of 181 students. Internal exam marks out of 30 are taken into consideration. Then the marks are converted into 100 (percentage) to have uniformity benchmark. These data is used to train the linear regression model to calculate the appropriate value of $\Theta 0$ and $\Theta 1$.

(Mobasher, Shawish and Ibrahim, 2017) This study proposes a complete EDM framework in a form of a rule based recommender system that is not developed to analyze and predict the student's performance only, but also to exhibit the reasons behind it. The framework analyzes the students' demographic data, study related and psychological characteristics to extract all possible knowledge from students, teachers and parents. Seeking the highest possible accuracy in academic performance prediction using a set of powerful data mining techniques. The framework succeeds to highlight the student's weak points and provide appropriate recommendations. The realistic case study that has been conducted on 200 students proves the outstanding performance of the proposed framework in comparison with the existing ones.

(Hamoud and Hashim, 2017) This study provided a model of student's successful prediction based on Bayes algorithms and suggests the best algorithm based on performance details. Two built Bayes Algorithms (naïve Bayes and Bayes network) were used in this model with students' questionnaire answers. The questionnaire consists of 62 questions that cover the fields affecting students' performance the most. The questions refer to health, social activity, relationships and academic performance. The questionnaire is constructed based on a Google form and open-source applications (LimeSurvey); the total number of student answers is 161. To build this model, the tool Weka 3.8 is used. The overall model design process can be divided into two stages. The first stage is finding the most correlated questions to the final class, and the second is applying algorithms and finding the optimal algorithm. A comparison is made between these two Bayes algorithms based on performance details. Bayes Net Classifier has TP Rate 0.655, FP Rate 0.432, Precision 0.643 and Recall 0.655 while the Naïve Bayes classifier has TP Rate 0.667, FP Rate 0.297, Precision 0.706 and

Recall 0.667. Finally, the naïve Bayes algorithm is selected as an optimal choice for students' successful prediction.

(Kaur and Singh, 2016) The researchers conduct this study to maintain the educational quality of the institute by minimizing the diverse effect of three factors (first one Parameters which affect the student performance, Data mining methods and the third one is data mining tool) on student's performance. In this Paper, Prediction of student Performance is done by applying Naïve byes' and J48 decision tree classification techniques WEKA tool. By applying data mining techniques to student data the researcher obtains knowledge which describes the student performance. This knowledge will help to improve the education quality, a student's performance and to decrease the failure rate. All these will help to improve the quality of the institute. This research is made on dummy data to obtain the results of classification. This data set consists 52 instances and each instance consists of 9 attributes. Naïve bays provide 63.59 % accuracy and j48 Provide 61.53% accuracy.

(Al-barrak and Al-razgan, 2016) In this study, the researcher used educational data mining to predict students' final GPA based on their grades in previous courses, the researcher collected students' transcript data for female students who graduated from the Computer Sciences, College at King Saud University in the year 2012 were collected from the database management system and the total number of student's was236 student's .The data included their final GPA and their grades in all courses. After preprocessing the data, the researcher applied the J48 decision tree algorithm to discover classification rules the researcher discovered classification rules to predict students final GPA based on their grades in mandatory courses, also evaluate the most important courses in the study plan that have a big impact on the students' final GPA.

(Mueen, 2016) The main objective of this study is to apply data mining techniques to predict and analyze students' academic performance based on their academic record and forum participation. The researcher has collected students' data from two undergraduate courses. The researcher compared, tested, and analyzed dataset with three classifiers. Those classifiers are Naïve Bayes, Multilayer Perception and C4.5 (J48). All three classifiers were tested on all 38 available attributes. The researcher used tenfold cross validation that means dataset was randomly divided into 10 subsets of the same size. It was observed that Naïve Bayes performs better than other two. Naïve Bayes is also the winner in precision which shows the predictive power. According to recall which represents the sensitivity. Naïve Bayes achieving overall prediction accuracy of 86%. This research assists teachers to early detect student who is expected to fail the course. The instructor can provide special attention to those students and help them to enhance their academic performance. There are a number of studies conducted in this regards identifying different factors such as student personal factor, family factor, or instructor factor. Many factors or combination of different factors affects student performance.

(Sumitha and Vinothkumar, 2016) The main objective of this research is to explore if it is possible to predict the performance of the student (output) based on the various explanatory (input). Data sets about 300 students were collected, dataset around 250 are being used as training dataset and 50 datasets as test data to design student model. Data are collected from I, II, III, IV year B.E CSE of KLN College of Information Technology (Affiliated To Anna University). In this process, a questionnaire form is used to collect the real data from the students that describe the relationship between learning behavior and their academic performance. The variables for judging the learning and academic behavior of students used in the questionnaire are student demographic details, School details, Attendance, CGPA and Final grade in last semester. The algorithm used for classification is Naive Bayes, Multilayer Perception (MLP), REP tree and J48. Each classifier is applied for two testing options - use Training sets and supplied test set. In the designing of student model J48 algorithm provide a maximum accuracy in classifying the instances in an efficient way by 97%. The student model is created in Net Beans using Java coding.

(Cheewaprakobkit, 2015) In this study, the researcher used WEKA open source data mining tool to analyze the attributes for predicting undergraduate students' academic performance in an international program. The data set comprised of 1,600 student records with 22 attributes of students registered between year 2001 and 2011 at a university in Thailand. Preprocessing included attribute importance analysis. The researcher applied the data set to differentiate classifiers (Decision Tree, Neural Network). Results show that the decision tree classifier achieves high accuracy of 85.188%, which is higher than that of a neural network classifier by 1.313%.

(Ahmad, Ismail and Aziz, 2015)A framework for predicting students' academic performance of first year bachelor students in Computer Science course. The data were collected from 8 year period intakes from July 2006/2007 until July 2013/2014 that contains the students' demographics, previous academic records, and family background information. The data set contains 497 record Decision Tree, Naïve Bayes, and Rule Based classification techniques are applied to the students' data in order to produce the best students' academic performance prediction model. The experiment result shows the Rule Based is a best model among the other techniques by receiving the highest accuracy value of 71.3%.

(Baradwaj and Pal, 2012)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, Information's like Attendance, Class test, Seminar and Assignment marks were collected from the student's management system, to predict the performance at the end of the semester. The data set used in this study was obtained from the VBS Purvanchal University, Jaunpur (Uttar Pradesh) on the sampling method of computer Applications department, of course MCA (Master of Computer Applications) from session 2007 to 2010. Initially size of the data is 50. This study help to the students and the teachers to improve the division of the student. This study also works to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination.

(Badr, Din and Elaraby, 2014) In this study, the classification task is used to predict the final grade of students using decision tree (ID3). The data set used in this study was obtained from a student's database used in one of the educational institutions, on the sampling method of the information system department from session 2005 to 2010. Initially size of the data is 1547 records.

(Al-radaideh, 2014) applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods, namely ID3, C4.5, and the NaïveBayes were used. The outcome of their results indicated that the Decision Tree model had a better prediction than other models.

(Borkar and Rajeswari, 2013) In this study, a student’s performance is evaluated using an association rule mining algorithm. Research has been done on assessing the student’s performance based on various attributes. The dataset of 60 students from Master of Computer Application (MCA) course was obtained from M.C.A department of the Pimpri Chinchwad College of Engineering, Pune University. The analysis revealed that student’s university performance is dependent on Unit test, Assignment, Attendance and graduation percentage. In this paper, we find various association rules between attributes like students graduation percentage, Attendance, Assignment work, Unit test Performance and how these attributes affect the student’s university result. The results reveal that the student’s performance level can be improved with university result by identifying students who are poor unit Test, Attendance, Assignment and graduation and giving them additional guidance to improve the university result.

(Kabakchieva, 2013) Predict the student university performance based on the collection of attributes providing information about the student pre-university (place and profile of the secondary school, the final secondary education score, the successful admission exam, the score achieved on that exam, and the total admission score) characteristics, a categorical target variable is constructed based on the original numeric parameter university average score. It has five distinct values (categories) − “excellent”, “very good”, “good”, “average” and “bad”, are determined by the total university score achieved by the students. The dataset used for the project implementation contains 10330 instances. Popular WEKA classifiers are used in the experimental study, including a common decision tree algorithm, C4.5 (J48), two Bayesian classifiers (Naive Bayes and BayesNet), a Nearest Neighbor algorithm (IBk) and to rule learners (OneR and JRip), each classifier is applied for two testing options – cross validation (using 10 folds and applying the algorithm 10 times, each time 9 of the folds are used for training and 1 fold is used for testing) and percentage split (2/3) of the dataset used for training and 1/3 for testing. As the result the decision tree classifier has, the more accuracy than other algorithms 66 -67 % and it considers as the best one.

(Baradwaj and Pal, 2012) In this research, the researcher, aimed to extract knowledge that describes students 'performance in the end semester examination. The classification task is used to evaluate student’s performance using decision tree method. The data set used in this study was obtained from the VBS Purvanchal University, Jaunpur (Uttar Pradesh) on the sampling method of computer Applications department, of course MCA (Master of Computer Applications) from session 2007 to 2010. Initially size of the data is 50.Information‟ s like Attendance, Class test, Seminar and Assignment marks were collected from the student's previous database.

(Yadav and Pal, 2012) The C4.5, ID3 and CART decision tree algorithms are applied to engineering student’s data to predict their performance in the final exam. The outcome of the decision tree predicted the number of students who are likely to pass, fail or promoted to next year. The data set used in this study was obtained from the VBS Purvanchal University, Jaunpur (Uttar Pradesh) on the sampling method for Institute of Engineering and Technology for session 2010. Initially size of the data is 90. The result shows that a C4.5 technique has the highest accuracy of 67.7778% compared to other methods. ID3 and CART algorithms. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for prediction of student’s performance in First Year of engineering exam. From the classifier accuracy it is clear that the true positive rate of the model for the

FAIL class is 0.786 for ID3 and C4.5 decision. Trees that means model is successfully identifying the students who are likely to fail. These students can be considered for proper counseling so as to improve their results.

(Yadav, Bharadwaj and Pal, 2012) Decision tree algorithms are applied to students' past performance data to generate the model and this model can be used to predict the students' performance. It helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising/counseling. The data set used in this study was obtained from the VBS Purvanchal University, Jaunpur (Uttar Pradesh), India on the sampling method of computer Applications department, of course MCA (Master of Computer Applications) from session 2008 to 2011. Initially size of the data is 48. The experimental results show that CART is the best algorithm for classification of data, has higher accuracy of 56.25% compared to other methods. ID3 algorithm.

(J. Kovacic, 2010) this paper aimed to examine to what extent the social-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course programme and course block), that may influence persists or drop out of students. I.e. enrolment data help the researcher in pre-identifying successful and unsuccessful students. The dataset contains the data stored in the Open Polytechnic student management system from 2006 to 2009, covering over 453 students who enrolled to 71150 Information Systems course was used to perform a quantitative analysis of study outcome. Based on a data mining technique (CHAID, CART), among classification tree growing methods Classification and Regression Tree (CART) was the most successful in growing the tree with an overall percentage of correct classification of 60.5%.

2.4.1 Comparative table of the related works

Table 2.1 Comparative table of the related work

NO	***Author***	***Data set size***	***Methods ussed***	***Best methods***	***Advantages***	***limitation***
1	(Hooshyar, Pedaste and Yang, 2020)	242	descriptive statistics, L-SVM and R-SVM, GP, DT, RF, NN ADB ,NB	L-SVM and R-SVM		
2	(Tampakas *et al.*, 2019)	282	Naive Bayes, Back-Propagation, Sequential Minimal Optimization (SMO), C4.5, JRip, 10NN	10NN. C4.5	This work could provide valuable hints and insights for better educational support by offering customized assistance according to students' predicted performance. It can be used as a reference for decision making in the graduate program .	
3	(Bashar, 2018)	1620	hybrid recommendation		Applied Association	-

			techniques: clustering Association rules		rules algorithms on each cluster, of which leads to generated strong rules in each year study. The intersection is applied on strong rules which help for recommending students to care which courses are positive and negative effectiveness on GPA.	
4	(Govindasamy, 2018)	1531	k-Means, k-Medoids, Fuzzy C Means (FCM) , Expectation (EM))	FCM and EM	The main advantage in this research is using clustering algorisms because interesting patterns and structures can be found directly from very large data sets with little	**-**

					or none of the background knowledge.	
5	(Science *et al.*, 2018)	Not assigned	Fuzzy C Means -Naïve Bayes Algorithm	Not assigned	This model helps the placement cell of the organization to identify the weaker students and provide extra care towards them so that they improve their performance henceforth	The research made on dummy data
6	(Gadhavi and Patel, 2017)	181 students	linear regression	linear regression	The model is tested on same data set of 181 students	This model it takes only one variable but it can be extended as multivariate model by adding more parameters to get more accurate results
7	(Mobasher, Shawish and Ibrahim,	200	rule based recommender system		Analyze the student's academic performance to	A larger data set needed to be used in different

	2017)				point out the student’s weak points and provide appropriate recommendations for treatment. The generated rules also showed a perfect matching with the scientific proved facts.	academic stages. Moreover, new psychological characteristics need to be added with the supervision of professional psychologist to better discover new patterns and enhance the prediction results.
8	(Hamoud and Hashim, 2017)	161	Bayes Algorithms (naïve Bayes and Bayes network)	naïve Bayes	The model can be depended on by both students and academic staff to decide the questions/answers that will enhance academic performance and improve institutional success	The size of data set , the number of attributes and the clean dataset affect the accuracy
9	(Kaur and Singh, 2016)	52	Naïve bayes , J48 decision tree	Naïve bayes	prediction of student performance is helpful to	Used small dataset

					identify the abilities of students, their interests and weaknesses and also helpful to cluster the students according to their performance	
10	(Al-barrak and Al-razgan, 2016)	236	J48 decision tree	decision tree	Discovered classification rules to predict students final GPA based on their grades in mandatory courses. evaluate the most important courses in the study plan that have a big impact on the students' final GP	Do not add the elective and general courses to get more accurate results. Not extend the experiment using other data mining techniques,
11	(Mueen, 2016)	Not assigned	Navïe Bayes, Neural Network, Decision Tree	Navïe Bayes	This research assist teachers to early detect student who is expected to fail	Taking just one course. Should include different courses and

					the course. Instructor can provide special attention to those student and help them to enhance their academic performance.	different educational levels
12	(Sumitha and Vinothkumar, 2016)	300	-Naïve bayes -Multilayer Perception -SMO -J48 -REP Tre	J48	This model can be useful in the educational system like Universities and Colleges. By this model we can know the academic status of the students in advance and can concentrate on students to improve their academic results and placements.	The prediction rates are not uniform among the algorithms. The range of prediction varies from (80-98%).Thereby by comparative analysis of classification algorithms (such as Naïve bayes, MLP ,SMO ,Decision Table, REP tree, J48) using WEKA tool
13	Cheewaprakobkit, P.	1,600 student	Decision Tree,	Decision Tree	The researcher applied the	Should reveal the factors that

	(2015)	records with 22 attributes	Neural Network		data set to differentiate classifiers (Decision Tree, Neural Network). A cross-validation with 10 folds was used to evaluate the prediction accuracy. An experimental comparison of the performance of the classifiers was also conducted.	affect academic achievement of students are as follows: 1. The number of hours worked per semester; 2. An additional English course; 3. The number of credits enrolled per semester; 4. Status of students such as single, married, or divorced
14	Ahmad, F., Ismail, N. H. and Aziz, A. A. (2015)	497 record	Decision Tree Naïve Bayes Rule Based	Rule Based	The model will allow the lecturers to take early actions to help and assist the poor and average category students to improve their results.	The limitation of this study is the small size of data due to incomplete and missing value in the collected data.

15	Badr, A., Din, E. and Elaraby, I. S. (2014)	1548	Decision Tree	Decision Tree	This study will help the student's to improve the student's performance, to identify those students which needed special attention to reduce failing ration and taking appropriate action at right time.	Using one algorism
16	Al-radaideh, Q. A. (2014)	Not assign	decision tree (ID3 , C4.5 NaïveBayes	ID3	The higher managements can use such classification model to enhance the courses outcome according to the extracted knowledge. to give a deeper understanding of student's enrollment pattern in the	Should collect a real and large data set from the university student database and apply the model using such data.

					course under study,	
17	Borkar, S. and Rajeswari, K. (2013	60 students	association rule mining algorithm	association rule mining algorithm	find various association rules between attributes like students graduation percentage, Attendance, Assignment work, Unit test Performance and how these attributes affect the student’s university result	Its need further analysis because the associations that the researcher get it from a priori algorithm are not identical with the correlation values of the attributes
18	Kabakchieva, D. (2013)	10330 instances	C4.5 ,J48 NaiveBayes , BayesNet, a Nearest Neighbour ,algorithm (IBk), OneR ,JRip	decision tree	Revealing the high potential of data mining applications for university management.	All tested classifiers are performing with an overall accuracy below 70 % which means that the error rate is high and the predictions are not very reliable.
19	Baradwaj, B. and Pal,	50	Decision Tree	Decision Tree	This study help to the students	Should use more algorisms

	S. (2012)				and the teachers to improve the division of the student. This study also work to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination.	
20	Yadav, S. K. and Pal, S. (2012)	90	Decision tree (ID3 , C4.5), CART	C4.5	The comparative analysis of the results states that the prediction has helped the weaker students to improve and brought out betterment in the result. Its provide steps to	Used small data set Used more algorism

					improve the performance of the students who were predicted to fail or promoted.	
21	Yadav, S., Bharadwaj, B. and Pal, S. (2012)	size of the data is 48	Decision tree (ID3 , C4.5), CART	CART	This study work to identify those students which needed special attention and also work to reduce fail ratio and taking appropriate action for the next semester examination.	Used small data set Used more algorism
22	J. Kovacic, Z. (2010	453	CHAID CART	CART	explores the socio-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course	The risk estimated by the cross-validation and the gain diagram suggests that all trees, based only on enrolment data are not quite good in

					programme and course block), that may influence persistence or dropout of students	separating successful from unsuccessful students.

CHAPTER III

Methodology

3.1 Introduction

This chapter describes the methodology followed to fulfill the objectives of the research. The research methodology followed CRISP-DM Methodology which governs by a series of stages. Starting by business understanding followed by data understanding .data preparation, modeling Evaluation and Deployment as shown in Fig 3.1 WEKA Explorer application has use for data analysis, WEKA stands for Waikato Environment for Knowledge Analysis. (Sumitha and Vinothkumar, 2016).

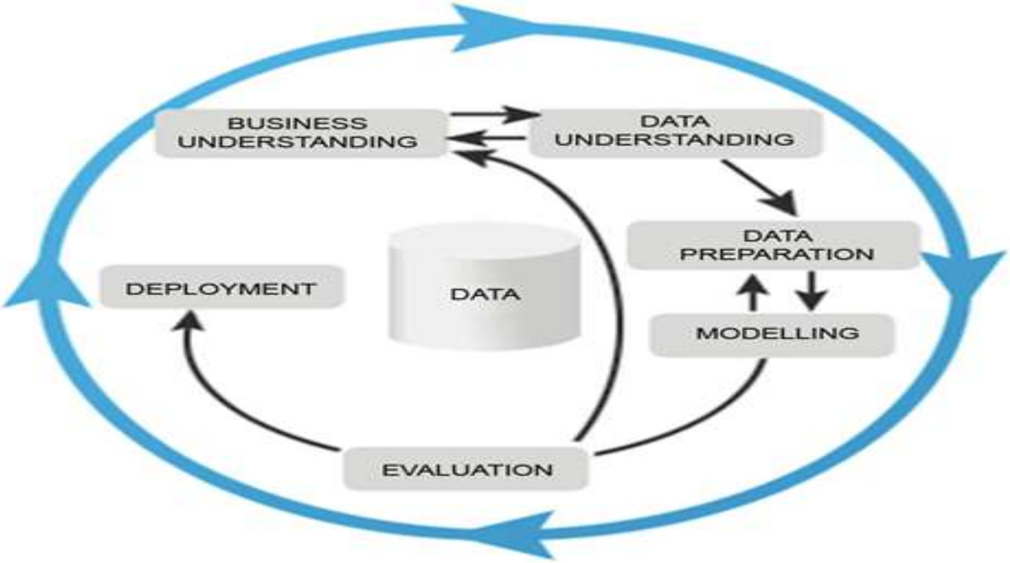


Figure 3.1 the CRISP-DM Methodology

3.1.1 Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective. Converting this knowledge into a problem definition and plan designed to achieve the research objectives.. In AUW, each year consists of two semesters in which the number of courses in each semester is 5 courses and each semester has final grades. In this research, the researcher will use the students final grade in semester one to predict their grade in semester two.

3.1.2 Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information. The primary data was collected using a questionnaire which includes questions related to several personal, socioeconomic and psychological variables as explain in (appendix A). the Creation of the questionnaire allowing the researcher to collect a large amount of data expected to affect student performance on a certain number of students, the researcher distributed this questionnaire to students in freshman year 2018 -2019 at School of Management in AUW.

Secondly, the completed questionnaires led to the construction of the database in which each student is described according to a certain number of criteria or attributes. It is necessary to extract information from the database that allows the researchers to identify their profile using data mining techniques and statistical methods.

3.1.3 Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data. In this phase, the data are put into a suitable form for the modeling phase. Students data collected from two sources, (questionnaire, university database) the data extracted and organized in a new excel sheet using the common field (university number) and converted to. ARFF (Attribute Relation File Format**).**

The original dataset contain data about 260 records with 17 attributes as described in the Table 3.1 .The provided data is subjected to many transformations. Some of unnecessary attributes are removed during cleaning stage, e.g. student index number, student name, Student accumulative average. The selected target variable to be learned by data mining algorithm, is the “Result" A categorical target variable is constructed based on the original numeric parameter university average score. It has five distinct values (categories) − “Distinction”, “Very Good”, “Good”, “Pass” and "Fail". as they shown in Table 3.2.The experiment conducted using two different resampling rate for training and testing dataset ; the training dataset, which used by the classifier to build up the model by learning from the given data, and the testing set (also known as validation dataset); which aims at estimating the performance of the predictive model.

In this study the researcher tested the original dataset two times. In the first time, the original dataset was resampled by dividing it 60%, 40%. 60% as a training set while the 40% used as test set and removed all duplicated data. After applying the filter, the training dataset will contain 156 instances then saved the dataset as training .arrf to use in analysis. While the testing dataset will contain 105 instance then saved the dataset as testing.arrf to use in analysis. as explain in (appendix B)

In the second time, the original data set was resampled by 70%, 30%. 70% as a training set while the 30% used for test set and removed all duplicated data. After applying the filter, the training dataset will contain 182 instances then saved the dataset as training .arrf to use in analysis. While the testing dataset will contain 79 instance then saved the dataset as testing.arrf to use in analysis.

Table 3.1 Dataset description:

No	Attribute Description	Possible values
1	Student grade in semester 1	Numerical percentage
2	Student grade in semester 2	Numerical percentage
3	Student accumulative average	Numerical percentage
4	Student Family Status	Still married, devours
5	Student live in a dormitory	Yes, No
6	Father qualification	no-education , elementary, secondary, graduate, Post- graduated, Doctorate
7	Mother qualification	no-education ,elementary, secondary, graduate, Post- graduated, Doctorate
8	Family annual income	Poor, Medium, High
9	Group of study friends	Alone, 2-3, 4-5, 6-more
10	Father Occupation	Worker, retired, not-applicable
11	Mother Occupation	House wife, Worker, retired, not-applicable
12	Reason to choose this school	My choice, My parent choice
13	Family living Country	Inside country, outside country
14	Do you have any of your parent passed away	Father, Mother, Both, None of them
15	Result	Distinction, Very Good , Good , Pass , Fail

Table 3.2 University Average Score

Percentage Range	Value	Class
100% to 80%,	Distinction	**A**
79% to 75%,	Very Good	**B**
74% to 65%,	Good	**C**
64% to 50%,	Pass	**D**
Range below 50%.	Fail	**F**

3.1.4 Modeling

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. The main objective is to explore if it is possible to predict the performance of the student (output) based on the various explanatory (input) variables which are retained in the model. The implementation of the model is made using a data mining tool WEKA, WEKA classifier is an open source software that implements a large collection of machine leaning algorithms used in the experimental study (Yadav, S. K., & Pal, S. (2012). The algorithms used for classification are Naive Bayes, J48, Random forest, BayesNet. Each classifier is applied for two testing options - use Training set and supplied test set. The classify panel enables the user to apply classification to the resulting dataset, to estimate the accuracy of the resulting predictive mode. This predictive model provides way to predict the student's future learning outcome.

3.1.5 Evaluation and Deployment

At this stage, the model obtained are more thoroughly evaluated and the steps executed to construct the model that are reviewed to be certain it properly achieves the business objectives, the criteria of comparison between algorithms can be made by measuring accuracy, speed, scalability, interpretability and robustness .in this research the model was evaluated by measuring the accuracy of algorithms which can be defined as the capacity of a classifier to predict the class label correctly, Confusion matrix is an approach which presents the predicted and actual classification based on multiple standers such as True Positive Rate (TPR "are examples correctly labeled as positives"(Staeheli and Mitchell, 2010). and False Positive Rate **(**FPR**)** "is refer to negative examples incorrectly labeled as positive"(Staeheli and Mitchell, 2010). A high TP Rate indicates that algorithm returns more relevant results than irrelevant and high FP Rate means that most of the results retuned by the algorithms are relevant.

The knowledge gained will need to be organized and presented in a way that the customer can use it. The final models from the previous phase, are then applied on a testing assessing dataset predictive accuracy and consistency.

CHAPTER IV

Results analysis and Discussion

4.1 Introduction

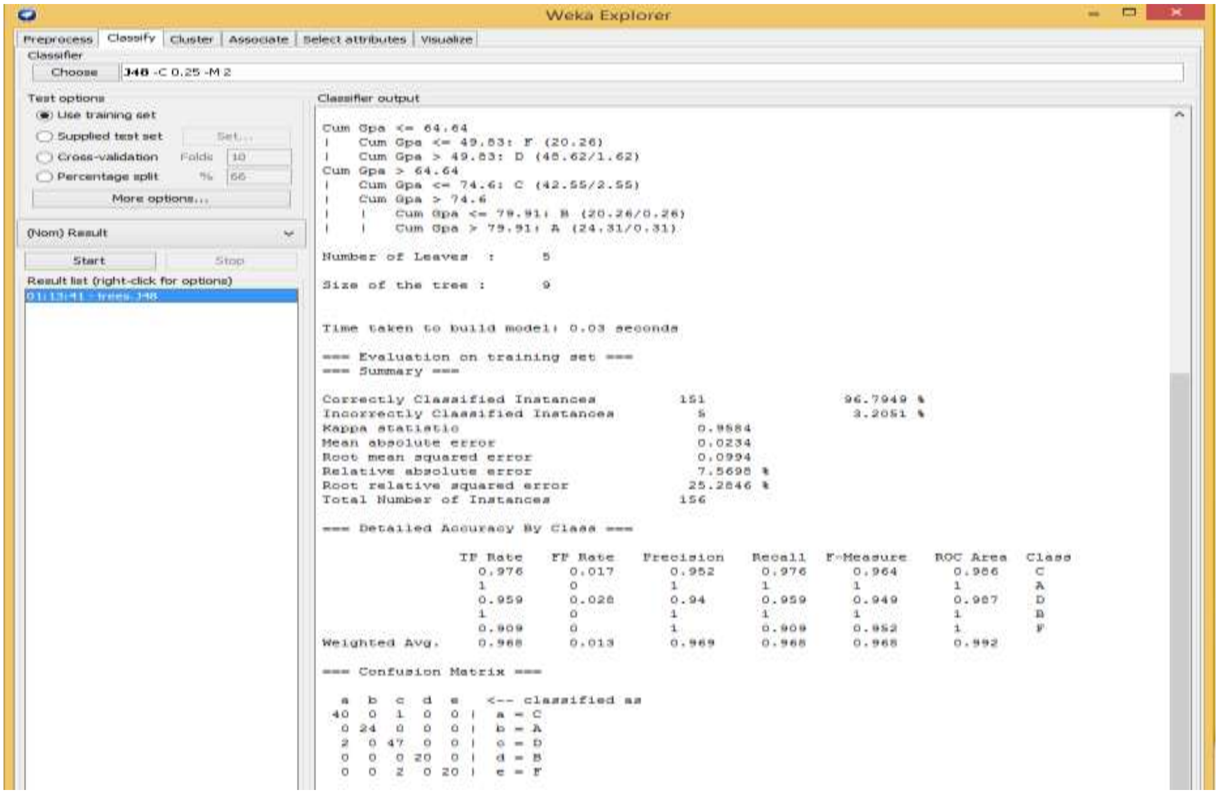
This chapter discusses all the findings and results. Firstly, results of different classification decision tree classifiers (j48, Random forest) and Bayesian classifiers **(**naïve Bayes, Bayes net), are applied for building the classification model. The WEKA Explorer application is used at this stage. Each classifier is applied for two testing options .The first one by splitting the original data set by 60% - 40%, 60% used to train the data while 40% used for the test. The second one by splitting the original dataset by 70% -30%, 70% used for training and 30% used for the testing. Secondly, the performance of different classifier is evaluated on the basis of TP Rate and FP Rate. Finally the results showed which factor is most affected student performance.

4.2 Result of Decision tree classifier

Decision tree algorisms filters applied on the dataset are the j48 and the random forest. Both of them are trained and tested two time, by splitting the data to 60%, 40% firstly and 70 -30 % secondly.

4.2.1 The result when using j48 classifier

On evaluating the data set under J48 classifier the result generated is as shown in Figure 4.1 .which shows the correctly classified instance by 97.4% in both 60% and 70% splitting the percentage split that produces a classification tree with a size of 9 nodes and 5 leaves.



Classifier: J48 -C 0.25 -M 2

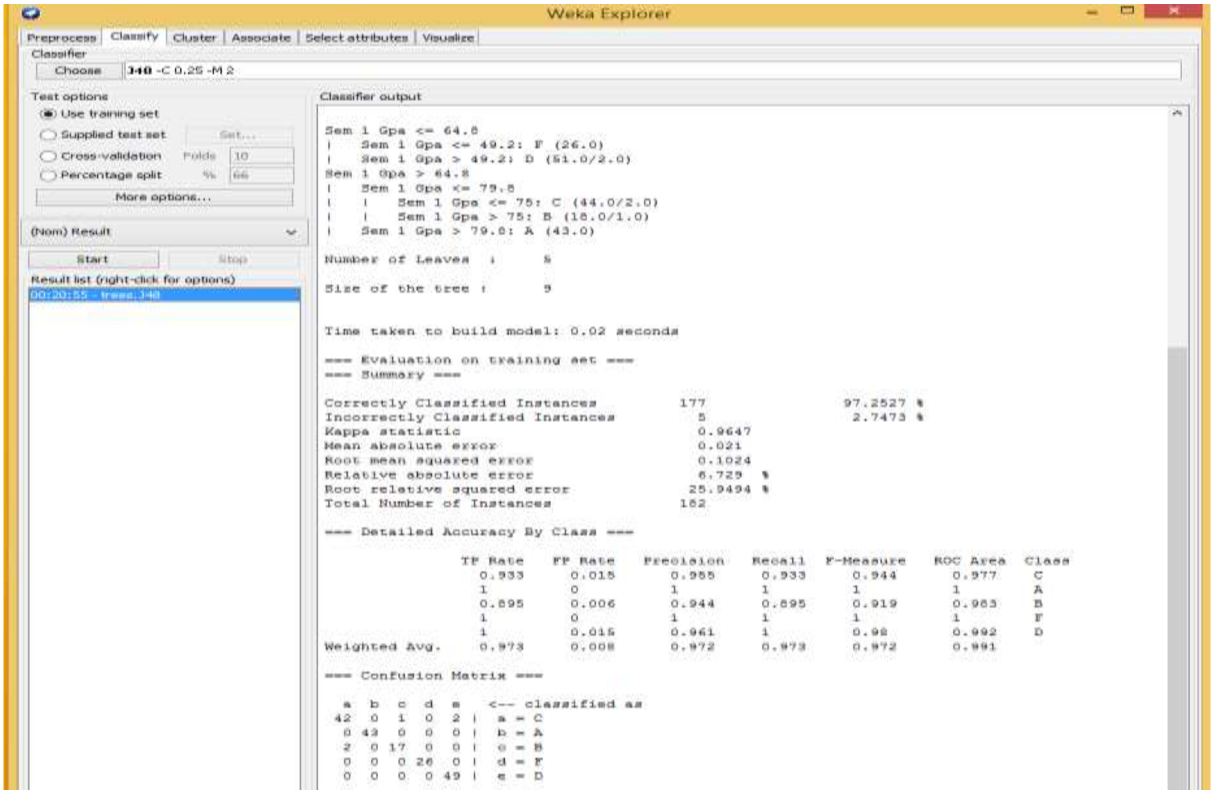
Test options:
- ● Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) Result

01:13:41 - trees.J48

```
Cum Gpa <= 64.64
|   Cum Gpa <= 49.83: F (20.26)
|   Cum Gpa > 49.83: D (48.62/1.62)
Cum Gpa > 64.64
|   Cum Gpa <= 74.6: C (42.55/2.55)
|   Cum Gpa > 74.6
|   |   Cum Gpa <= 79.91: B (20.26/0.26)
|   |   Cum Gpa > 79.91: A (24.31/0.31)

Number of Leaves  :      5

Size of the tree :       9


Time taken to build model: 0.03 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         151               96.7949 %
Incorrectly Classified Instances         5                3.2051 %
Kappa statistic                          0.9584
Mean absolute error                      0.0234
Root mean squared error                  0.0994
Relative absolute error                  7.5698 %
Root relative squared error             25.2846 %
Total Number of Instances              156

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.976     0.017      0.952     0.976     0.964      0.986    C
                 1         0          1         1         1          1        A
                 0.959     0.028      0.94      0.959     0.949      0.987    D
                 1         0          1         1         1          1        B
                 0.909     0          1         0.909     0.952      1        F
Weighted Avg.    0.968     0.013      0.969     0.968     0.968      0.992

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 40  0  1  0  0 |  a = C
  0 24  0  0  0 |  b = A
  2  0 47  0  0 |  c = D
  0  0  0 20  0 |  d = B
  0  0  2  0 20 |  e = F
```

Figure 4.1: The result of j48 classifier when using 60% of data set for training



Classifier

J48 -C 0.25 -M 2

Test options

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

(Nom) Result

Result list (right-click for options)

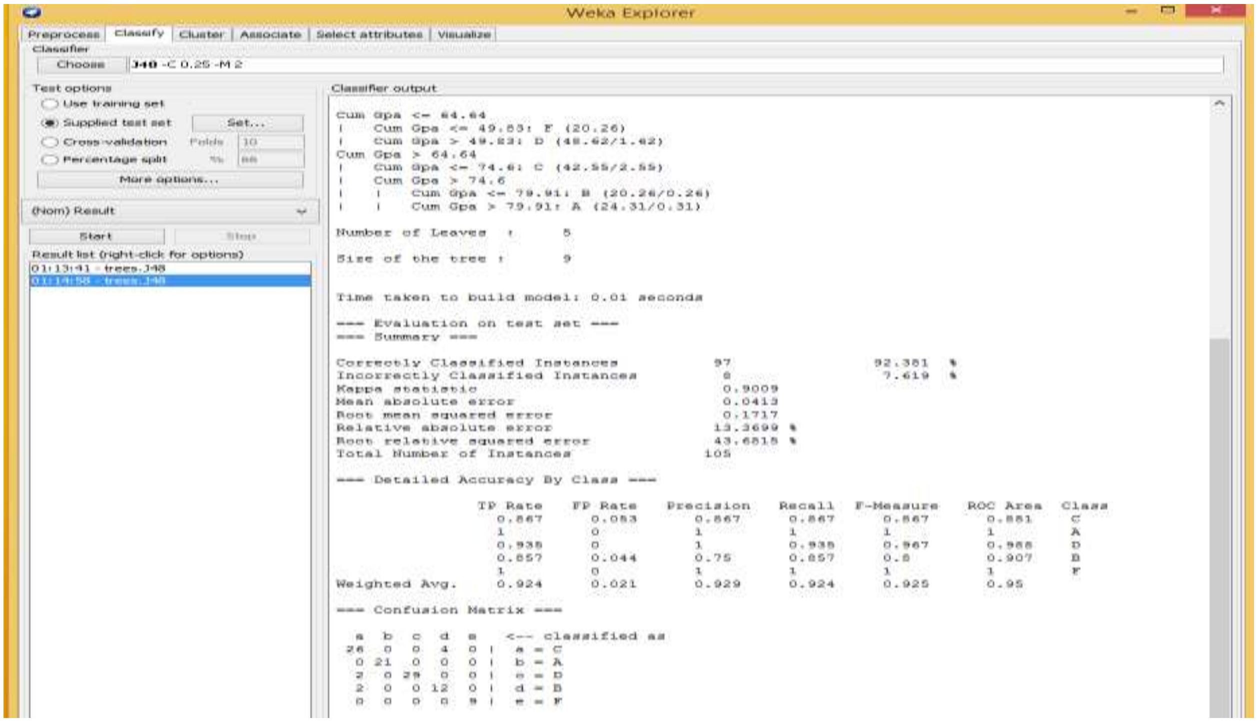
00:20:55 - trees.J48

Classifier output

```
Sem 1 Gpa <= 64.8
|   Sem 1 Gpa <= 49.2: F (26.0)
|   Sem 1 Gpa > 49.2: D (51.0/2.0)
Sem 1 Gpa > 64.8
|   Sem 1 Gpa <= 79.8
|   |   Sem 1 Gpa <= 75: C (44.0/2.0)
|   |   Sem 1 Gpa > 75: B (18.0/1.0)
|   Sem 1 Gpa > 79.8: A (43.0)

Number of Leaves  :      5

Size of the tree :      9


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         177               97.2527 %
Incorrectly Classified Instances         5                2.7473 %
Kappa statistic                          0.9647
Mean absolute error                      0.021
Root mean squared error                  0.1024
Relative absolute error                  6.729  %
Root relative squared error             25.9494 %
Total Number of Instances              182

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.933     0.015     0.955      0.933     0.944      0.977    C
                 1         0         1          1         1          1        A
                 0.895     0.006     0.944      0.895     0.919      0.963    B
                 1         0         1          1         1          1        F
                 1         0.015     0.961      1         0.98       0.992    D
Weighted Avg.    0.973     0.008     0.972      0.973     0.972      0.991

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 42  0  1  0  2 |  a = C
  0 43  0  0  0 |  b = A
  2  0 17  0  0 |  c = B
  0  0  0 26  0 |  d = F
  0  0  0  0 49 |  e = D
```

Figure 4.2: The result of j48 classifier when using 70% of data set for training

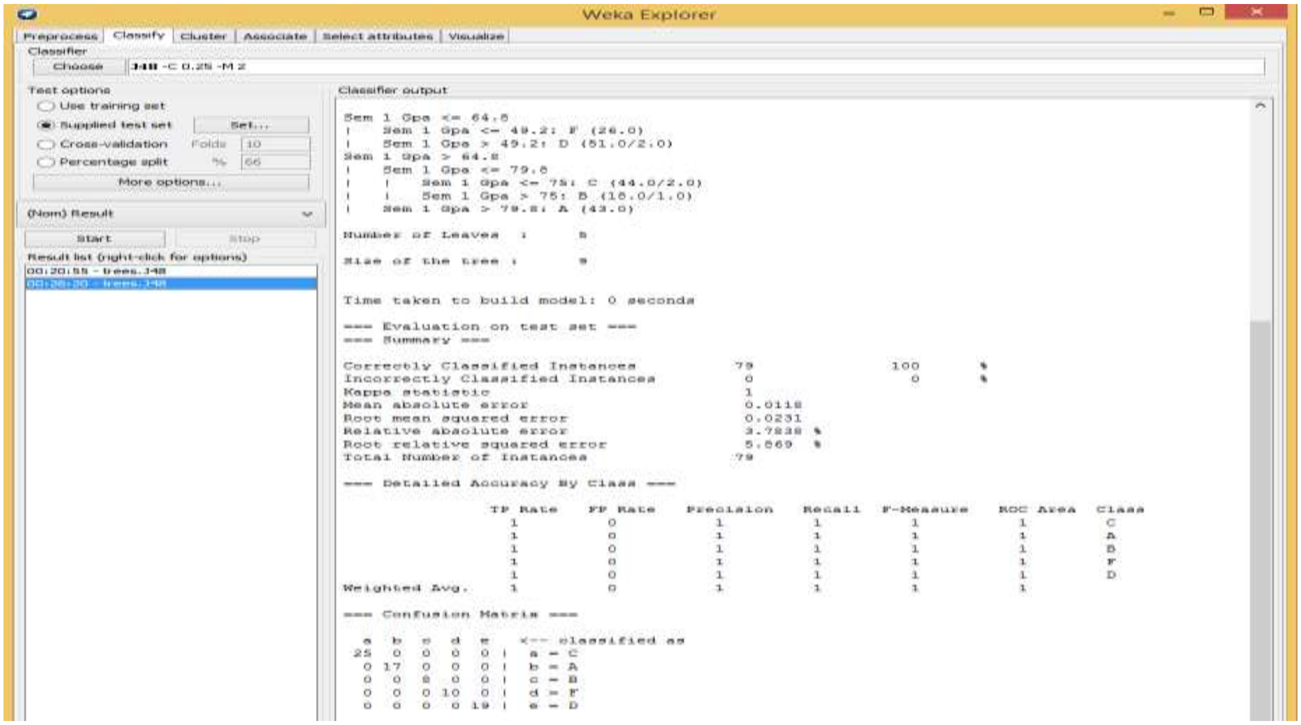
To analyze the performance of students, the researcher used test data to predict performance based on training data. The Figures below 4.2, 4.3 represent the j48 algorithm implementation for test data 40% and 30% and tree visualization of J48 algorithm. The first figure shows the correctly classified instance by 95% for 40% test dataset, while the second one 30% has very high accuracy than the first one with percentage of 100% and 0 error rate which is very reliable.



J48 -C 0.25 -M 2

```
Cum Gpa <= 64.64
|   Cum Gpa <= 49.83: F (20.26)
|   Cum Gpa > 49.83: D (48.62/1.62)
Cum Gpa > 64.64
|   Cum Gpa <= 74.6: C (42.55/2.55)
|   Cum Gpa > 74.6
|   |   Cum Gpa <= 79.91: B (20.26/0.26)
|   |   Cum Gpa > 79.91: A (24.31/0.31)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0.01 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          97               92.381  %
Incorrectly Classified Instances         8                7.619  %
Kappa statistic                          0.9009
Mean absolute error                      0.0413
Root mean squared error                  0.1717
Relative absolute error                 13.3699 %
Root relative squared error             43.6818 %
Total Number of Instances              105

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.867     0.053     0.867      0.867     0.867      0.881    C
                 1         0         1          1         1          1        A
                 0.935     0         1          0.935     0.967      0.988    D
                 0.857     0.044     0.75       0.857     0.8        0.907    B
                 1         0         1          1         1          1        F
Weighted Avg.    0.924     0.021     0.929      0.924     0.925      0.95

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 26  0  0  4  0 |  a = C
  0 21  0  0  0 |  b = A
  2  0 29  0  0 |  c = D
  2  0  0 12  0 |  d = B
  0  0  0  0  9 |  e = F
```

Figure 4.3: The result of j48 classifier when using 40% of data set for testing



J48 -C 0.25 -M 2

```
Sem 1 Gpa <= 64.8
|   Sem 1 Gpa <= 49.2: F (26.0)
|   Sem 1 Gpa > 49.2: D (51.0/2.0)
Sem 1 Gpa > 64.8
|   Sem 1 Gpa <= 79.8
|   |   Sem 1 Gpa <= 75: C (44.0/2.0)
|   |   Sem 1 Gpa > 75: B (18.0/1.0)
|   Sem 1 Gpa > 79.8: A (43.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          79              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0118
Root mean squared error                  0.0231
Relative absolute error                  3.7838 %
Root relative squared error              5.869  %
Total Number of Instances               79

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0         1          1         1          1        C
                 1         0         1          1         1          1        A
                 1         0         1          1         1          1        B
                 1         0         1          1         1          1        F
                 1         0         1          1         1          1        D
Weighted Avg.    1         0         1          1         1          1

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 25  0  0  0  0 |  a = C
  0 17  0  0  0 |  b = A
  0  0  8  0  0 |  c = B
  0  0  0 10  0 |  d = F
  0  0  0  0 19 |  e = D
```

Figure 4.4: The result of j48 classifier when using 30% of data set for testing

Table 4.1: TP Rate FP Rate using j48 algorism

J48								
60% , 40%					Used of 70% , 30%			
Class	Training Set		Testing Set		Training Set		Testing Set	
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate
A	1	0	1	0.048	1	0	1	0
B	0.833	0.007	0.733	0	0.895	0.006	1	0
C	0.95	0.017	0.967	0	0.933	0.015	1	0
D	1	0.009	1	0.013	1	0.015	1	0
F	1	0	1	0	1	0	1	0
Weighted Average	0.974	0.007	0.952	0.013	0.973	0.008	1	0

4.2.1.1 Discussion of j48 classifier result

A- Result of 60% - 40% splitting

The results for the detailed accuracy by classes in J48, including the True Positive (TP) rate and the False Positive (FP) rate .The results reveal that the True Positive Rate (TP)is excellent for three of the classes and have the same percentage in training and testing dataset A,D,F by (100%) and high for other two classes, class B in training dataset (83%) while in testing dataset is decreased by (73%) and class C in training dataset (95%) while in testing dataset increased by (96%).

The (FP) performed excellent for class F and A (0%) and D (1%) for both train and test data. For A class also performed excellent in train data (0%) but the percentage increased in test data by (5%) with a high percentage. Class B and C in train dataset perform well (1%), (2%) respectively, but they have an excellent present in test dataset.

A- Result of 70% - 30% splitting

The results of 70% - 30% splitting reveal that the True Positive Rate (TP) is excellent for three of the classes and in test dataset A, D, F by (100%) and high for other two classes. Class B in training dataset (89%) and class C in training dataset (95%) while in test dataset all classes have excellent performance by (100%).

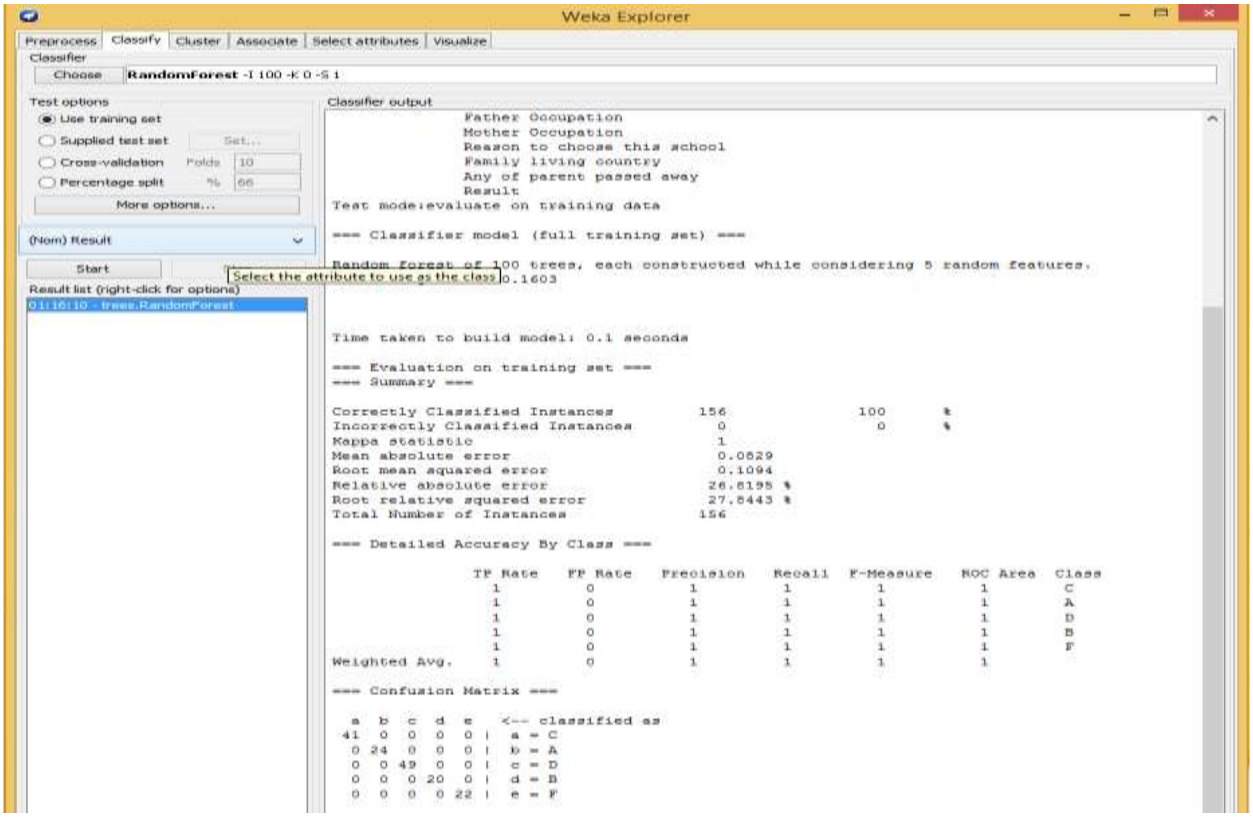
The (FP) performed excellent for class F and A (0%) and B (1%) for train data. Class C and D in train dataset perform well by (2%), while in test dataset all classes have excellent performance (0%).

B- Comparison between the results

From the above result for 60% - 40% splitting mean classes F and A which falls between 0 and 1 with a higher number that indicate better classification performance. And other classes B, C, D its performance in test data set better than its performance in train dataset. In both FP rate and TP rate, while 70% - 30 % splitting performed excellent for all classes which fall between 0 and 1 indicating better classification performance which is very reliable.

4.2.2 Result of Random forest Classifier

On evaluating the data set under J48 classifier the result generated as shown in Figure 4.5 which shows the correctly classified instance by 100%. In both trained data.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Choose
Set...
More options...
Start
Select the attribute to use as the class
Result list (right-click for options)
01:16:10 - trees.RandomForest

Classifier
RandomForest -I 100 -K 0 -S 1

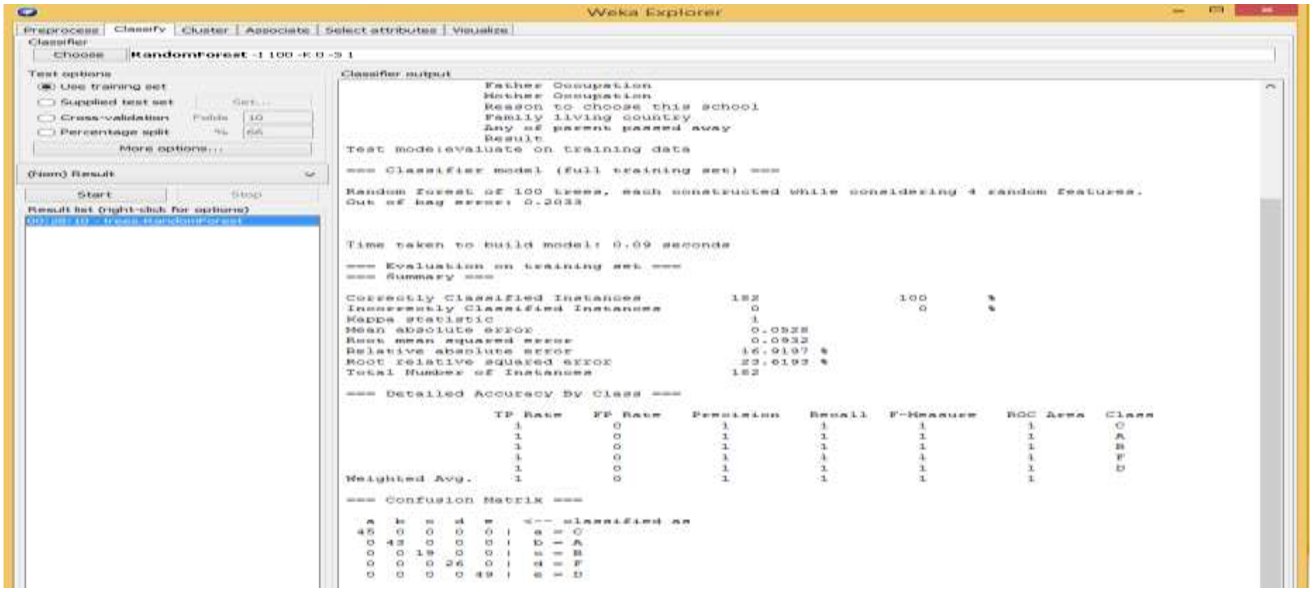
Test options
● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Result

Classifier output

```
                    Father Occupation
                    Mother Occupation
                    Reason to choose this school
                    Family living country
                    Any of parent passed away
                    Result
Test mode:evaluate on training data

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

Random forest of 100 trees, each constructed while considering 5 random features.
                      0.1603



Time taken to build model: 0.1 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         156              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0829
Root mean squared error                  0.1094
Relative absolute error                 26.8195 %
Root relative squared error             27.8443 %
Total Number of Instances              156

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1          1        1           1        C
                 1         0          1          1        1           1        A
                 1         0          1          1        1           1        D
                 1         0          1          1        1           1        B
                 1         0          1          1        1           1        F
Weighted Avg.    1         0          1          1        1           1

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 41  0  0  0  0 |  a = C
  0 24  0  0  0 |  b = A
  0  0 49  0  0 |  c = D
  0  0  0 20  0 |  d = B
  0  0  0  0 22 |  e = F
```

Figure 4.5: The result of Random forest classifier when using 60% of data set for training

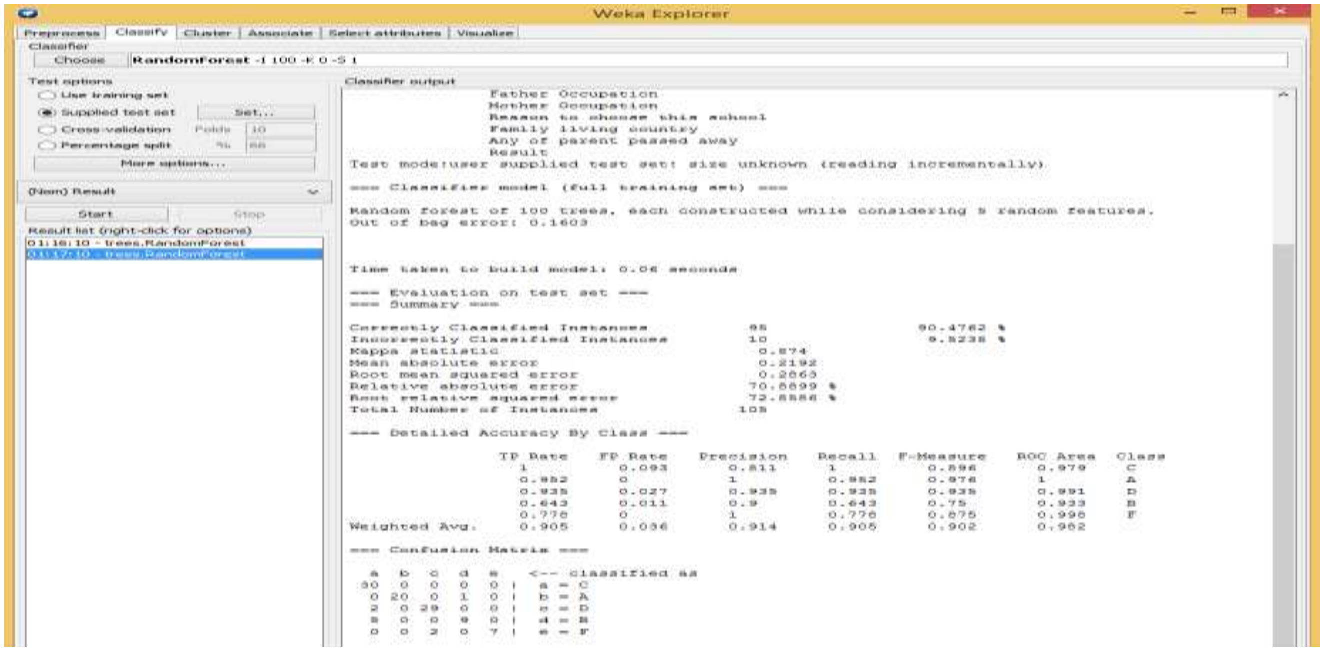


Classifier: RandomForest -I 100 -K 0 -S 1

Test options: ● Use training set ○ Supplied test set ○ Cross-validation Folds 10 ○ Percentage split % 66

```
                    Father Occupation
                    Mother Occupation
                    Reason to choose this school
                    Family living country
                    Any of parent passed away
                    Result
Test mode:evaluate on training data

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

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.2033


Time taken to build model: 0.09 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         182              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0528
Root mean squared error                  0.0932
Relative absolute error                 16.9197 %
Root relative squared error             23.0193 %
Total Number of Instances              182

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   1         0          1         1        1          1        C
                   1         0          1         1        1          1        A
                   1         0          1         1        1          1        B
                   1         0          1         1        1          1        F
                   1         0          1         1        1          1        D
Weighted Avg.      1         0          1         1        1          1

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 45  0  0  0  0 |  a = C
  0 43  0  0  0 |  b = A
  0  0 19  0  0 |  c = B
  0  0  0 26  0 |  d = F
  0  0  0  0 49 |  e = D
```

Figure 4.6: The result of Random forest classifier when using 70% of data set for training

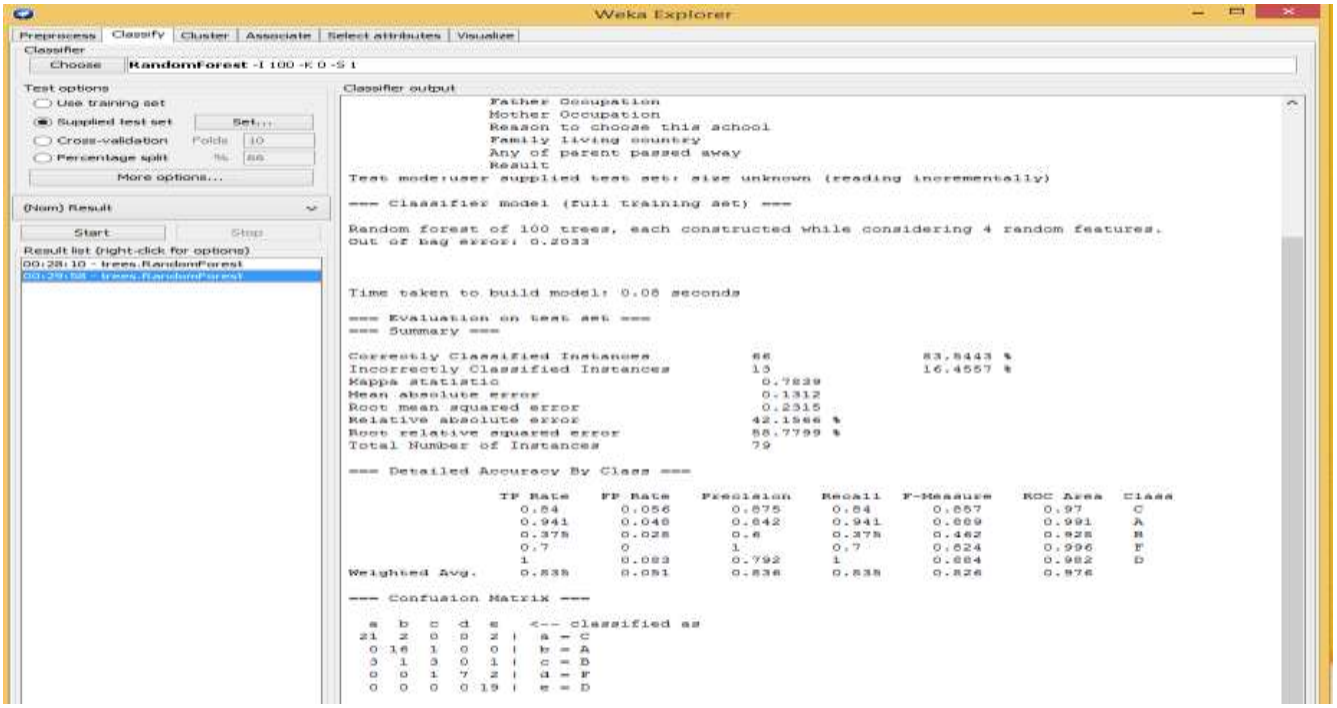


Classifier: RandomForest -I 100 -K 0 -S 1

Test options: ○ Use training set ● Supplied test set ○ Cross-validation Folds 10 ○ Percentage split % 66

```
                    Father Occupation
                    Mother Occupation
                    Reason to choose this school
                    Family living country
                    Any of parent passed away
                    Result
Test mode:user supplied test set: size unknown (reading incrementally)

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

Random forest of 100 trees, each constructed while considering 5 random features.
Out of bag error: 0.1603


Time taken to build model: 0.06 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          95               90.4762 %
Incorrectly Classified Instances        10                9.5238 %
Kappa statistic                          0.874
Mean absolute error                      0.2192
Root mean squared error                  0.2863
Relative absolute error                 70.6899 %
Root relative squared error             72.8886 %
Total Number of Instances              105

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   1         0.093      0.811     1        0.896      0.979    C
                   0.952     0          1         0.952    0.976      1        A
                   0.935     0.027      0.935     0.935    0.935      0.991    D
                   0.643     0.011      0.9       0.643    0.75       0.933    B
                   0.778     0          1         0.778    0.875      0.998    F
Weighted Avg.      0.905     0.036      0.914     0.905    0.902      0.982

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 30  0  0  0  0 |  a = C
  0 20  0  1  0 |  b = A
  2  0 29  0  0 |  c = D
  5  0  0  9  0 |  d = B
  0  0  2  0  7 |  e = F
```

Figure 4.7: The result of Random forest classifier when using 40% of data set for testing



RandomForest -I 100 -K 0 -S 1

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

(Nom) Result

00:28:10 - trees.RandomForest
00:29:58 - trees.RandomForest

```
                   Father Occupation
                   Mother Occupation
                   Reason to choose this school
                   Family living country
                   Any of parent passed away
                   Result
Test mode:user supplied test set: size unknown (reading incrementally)

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

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.2033



Time taken to build model: 0.08 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          66               83.5443 %
Incorrectly Classified Instances        13               16.4557 %
Kappa statistic                          0.7829
Mean absolute error                      0.1312
Root mean squared error                  0.2315
Relative absolute error                 42.1566 %
Root relative squared error             58.7799 %
Total Number of Instances               79

=== Detailed Accuracy By Class ===

                   TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
                     0.84      0.056     0.875      0.84      0.857      0.97      C
                     0.941     0.048     0.842      0.941     0.889      0.991     A
                     0.375     0.028     0.6        0.375     0.462      0.828     B
                     0.7       0         1          0.7       0.824      0.996     F
                     1         0.083     0.792      1         0.884      0.982     D
Weighted Avg.        0.835     0.051     0.836      0.835     0.826      0.976

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 21  2  0  0  2 |  a = C
  0 16  1  0  0 |  b = A
  3  1  3  0  1 |  c = B
  0  0  1  7  2 |  d = F
  0  0  0  0 19 |  e = D
```

Figure 4.8: The result of Random forest classifier when using 30% of data set for testing

Table 4.2: TP Rate FP Rate using random forest algorism

Random Forest								
60% , 40%					70% , 30%			
Class	Training Set		Testing Set		Training Set		Testing Set	
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate
A	1	0	0.955	0.084	1	0	0.941	0.048
B	1	0	0.2	0.022	1	0	0.375	0.028
C	1	0	0.867	0.093	1	0	0.84	0.056
D	1	0	0.955	0.084	1	0	1	0.083
F	1	0	0.2	0.022	1	0	0.7	0
Weighted Average	1	0	0.781	0.068	1	0	0.835	0.051

4.2.2.1 Discussion of Random Forest classifier result

A- Result of 60% - 40% splitting

The results indicate that the (TP) is excellent for all five classes and have the same percentage in training dataset A, B, C, D and F by (100%), while in the testing dataset the TP is excellent for class D (96%) and high for two classes, A (95%) and C (87%) while it’s very low in B and F (2%).

The (FP) is also performed excellent for all five classes and have the same percentage in training dataset A, B, C, D and F by (0%) but the percentage increased in testing data by (8%) for class A, (2%) for class B and F , (9%) for class C and (8%)for class D .

B- Result of 70% - 30% splitting

The results indicate that the (TP) is excellent for all five classes and have the same percentage in training dataset A, B, C, D and F by (100%), while in the testing dataset the TP is excellent for class D (100%) and high for two classes, A (94%) and C (84%), while it is low in class B (38%) and very low in F (7%).

The (FP) also performed excellent for all five classes and have the same percentage in training dataset A, B, C, D and F by (0%), but the percentage increased in testing data by (5%) for class A, (3%) for class B , (6%) for class C and (9%) for D, while it has excellent performance for class F (0%).

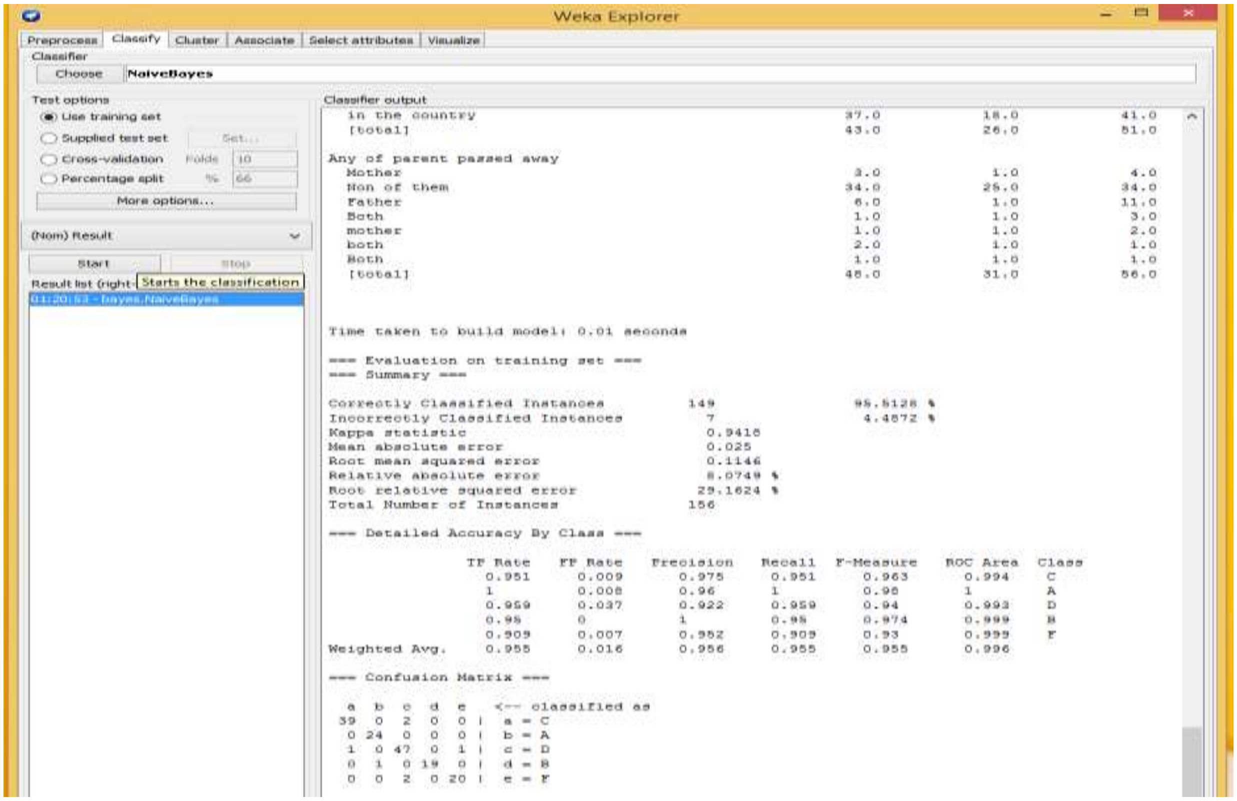
C- Comparison between the results

The training dataset which falls between 0 and 1 with a higher number indicate better classification performance than the testing dataset in all classes for the both splitting option , while in testing datasets the 70% - 30% perform better than 60% - 40% splitting by (83.5%) accuracy.

4.3 Result of Bayesian classifiers

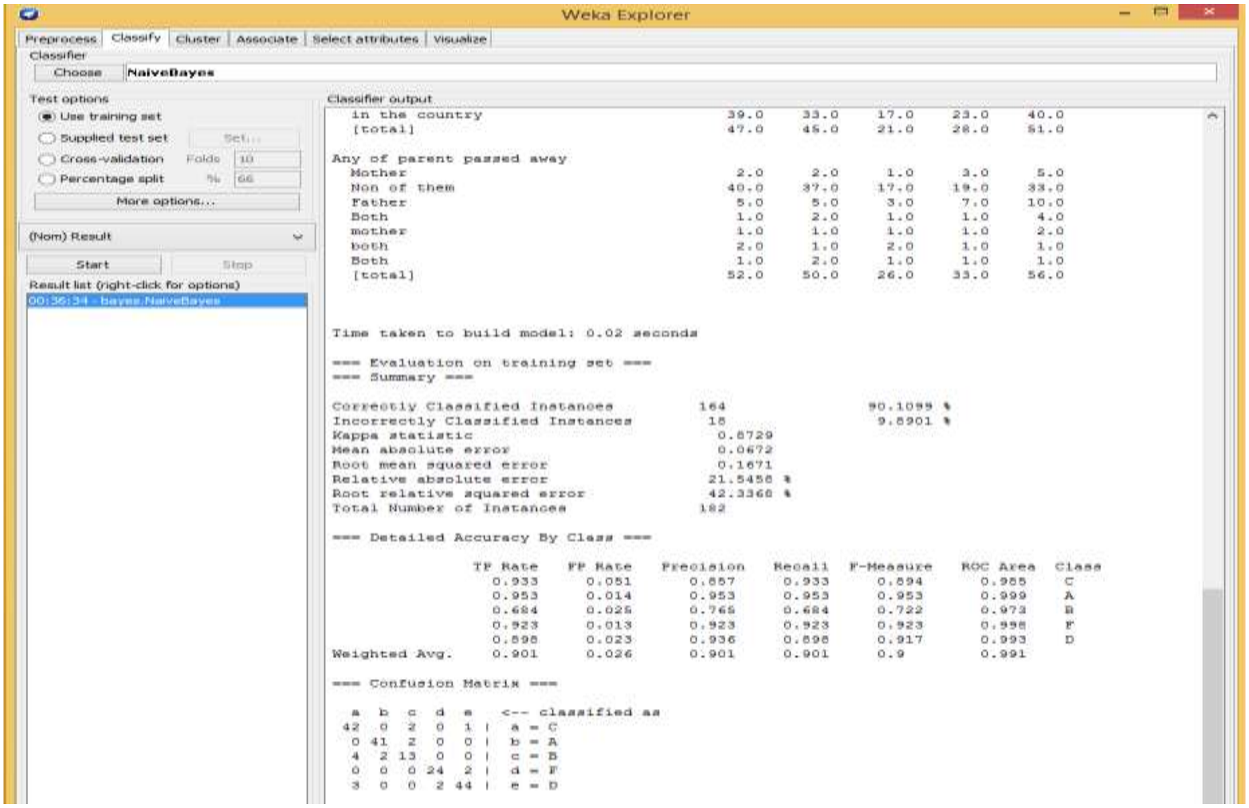
The two WEKA classification filters applied on the dataset are the NaiveBayes and the BayesNet. Both of them are trained and tested two time, the first time by split percentage 60%, 40% respectively. The Second time by 70 -30 %, the achieved results are presented in figures 4.6, 4.7 and Table 4.3.

4.3.1 Result of Naive Bayes Classifier



```
   in the country                                         37.0          18.0          41.0
   [total]                                                43.0          26.0          51.0

Any of parent passed away
   Mother                                                  3.0           1.0           4.0
   Non of them                                            34.0          25.0          34.0
   Father                                                  6.0           1.0          11.0
   Both                                                    1.0           1.0           3.0
   mother                                                  1.0           1.0           2.0
   both                                                    2.0           1.0           1.0
   Both                                                    1.0           1.0           1.0
   [total]                                                48.0          31.0          56.0


Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         149               95.5128 %
Incorrectly Classified Instances         7                4.4872 %
Kappa statistic                          0.9418
Mean absolute error                      0.025
Root mean squared error                  0.1146
Relative absolute error                  8.0749 %
Root relative squared error             29.1624 %
Total Number of Instances              156

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.951     0.009     0.975     0.951     0.963      0.994     C
                 1         0.008     0.96      1         0.98       1         A
                 0.959     0.037     0.922     0.959     0.94       0.993     D
                 0.95      0         1         0.95      0.974      0.999     B
                 0.909     0.007     0.952     0.909     0.93       0.999     F
Weighted Avg.    0.955     0.016     0.956     0.955     0.955      0.996

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 39  0  2  0  0 |  a = C
  0 24  0  0  0 |  b = A
  1  0 47  0  1 |  c = D
  0  1  0 19  0 |  d = B
  0  0  2  0 20 |  e = F
```

Figure 4.9: The result of Naive Bayes classifier when using 60% of data set for training



Classifier
NaiveBayes

Test options
● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

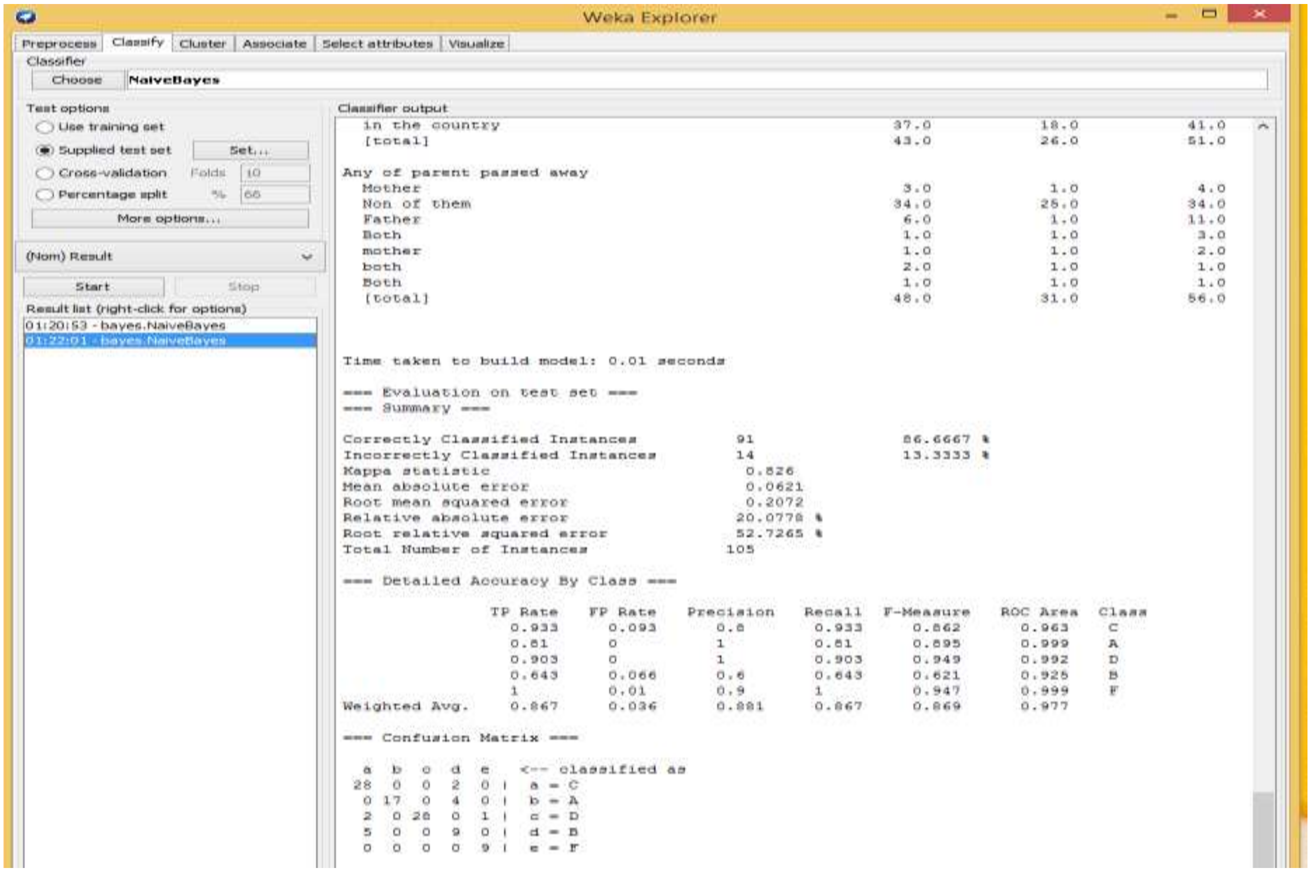
(Nom) Result

Result list (right-click for options)
00:36:34 - bayes.NaiveBayes

Classifier output

```
   in the country                          39.0      33.0      17.0      23.0      40.0
   [total]                                 47.0      45.0      21.0      28.0      51.0

Any of parent passed away
   Mother                                   2.0       2.0       1.0       3.0       5.0
   Non of them                             40.0      37.0      17.0      19.0      33.0
   Father                                   5.0       5.0       3.0       7.0      10.0
   Both                                     1.0       2.0       1.0       1.0       4.0
   mother                                   1.0       1.0       1.0       1.0       2.0
   both                                     2.0       1.0       2.0       1.0       1.0
   Both                                     1.0       2.0       1.0       1.0       1.0
   [total]                                 52.0      50.0      26.0      33.0      56.0



Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         164               90.1099 %
Incorrectly Classified Instances        18                9.8901 %
Kappa statistic                          0.8729
Mean absolute error                      0.0672
Root mean squared error                  0.1671
Relative absolute error                 21.5458 %
Root relative squared error             42.3368 %
Total Number of Instances              182

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.933     0.051     0.857     0.933     0.894      0.985    C
                 0.953     0.014     0.953     0.953     0.953      0.999    A
                 0.684     0.025     0.765     0.684     0.722      0.973    B
                 0.923     0.013     0.923     0.923     0.923      0.998    F
                 0.898     0.023     0.936     0.898     0.917      0.993    D
Weighted Avg.    0.901     0.026     0.901     0.901     0.9        0.991

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 42  0  2  0  1 |  a = C
  0 41  2  0  0 |  b = A
  4  2 13  0  0 |  c = B
  0  0  0 24  2 |  d = F
  3  0  0  2 44 |  e = D
```

Figure 4.10: The result of Naive Bayes classifier when using 70% of data set for training



Classifier: NaiveBayes

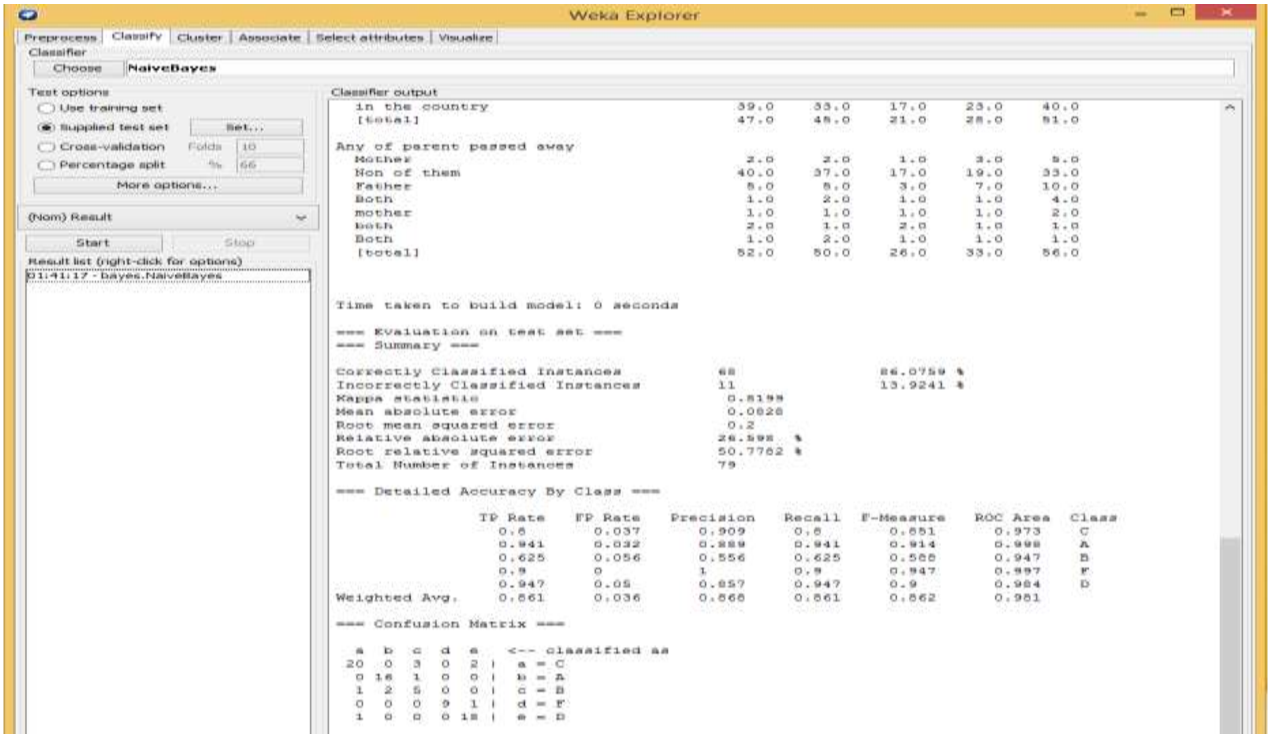
○ Use training set
● Supplied test set   Set...
○ Cross-validation   Folds 10
○ Percentage split   % 66
More options...

(Nom) Result

01:20:53 - bayes.NaiveBayes
01:22:01 - bayes.NaiveBayes

```
  in the country                                          37.0         18.0         41.0
  [total]                                                 43.0         26.0         51.0

Any of parent passed away
  Mother                                                   3.0          1.0          4.0
  Non of them                                             34.0         25.0         34.0
  Father                                                   6.0          1.0         11.0
  Both                                                     1.0          1.0          3.0
  mother                                                   1.0          1.0          2.0
  both                                                     2.0          1.0          1.0
  Both                                                     1.0          1.0          1.0
  [total]                                                 48.0         31.0         56.0



Time taken to build model: 0.01 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          91               86.6667 %
Incorrectly Classified Instances        14               13.3333 %
Kappa statistic                          0.826
Mean absolute error                      0.0621
Root mean squared error                  0.2072
Relative absolute error                 20.0778 %
Root relative squared error             52.7265 %
Total Number of Instances              105

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.933     0.093      0.8       0.933     0.862      0.963    C
                 0.81      0          1         0.81      0.895      0.999    A
                 0.903     0          1         0.903     0.949      0.992    D
                 0.643     0.066      0.6       0.643     0.621      0.925    B
                 1         0.01       0.9       1         0.947      0.999    F
Weighted Avg.    0.867     0.036      0.881     0.867     0.869      0.977

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 28  0  0  2  0 |  a = C
  0 17  0  4  0 |  b = A
  2  0 28  0  1 |  c = D
  5  0  0  9  0 |  d = B
  0  0  0  0  9 |  e = F
```

Figure 4.11: The result of Naive Bayes classifier when using 40% of data set for testing



Classifier

Choose NaiveBayes

Test options

○ Use training set
● Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Result

Result list (right-click for options)

01:41:17 - bayes.NaiveBayes

Classifier output

```
   in the country                         39.0    33.0    17.0    23.0    40.0
   [total]                                47.0    48.0    21.0    28.0    51.0

 Any of parent passed away
   Mother                                  2.0     2.0     1.0     3.0     5.0
   Non of them                            40.0    37.0    17.0    19.0    33.0
   Father                                  8.0     8.0     3.0     7.0    10.0
   Both                                    1.0     2.0     1.0     1.0     4.0
   mother                                  1.0     1.0     1.0     1.0     2.0
   both                                    2.0     1.0     2.0     1.0     1.0
   Both                                    1.0     2.0     1.0     1.0     1.0
   [total]                                52.0    50.0    26.0    33.0    56.0


Time taken to build model: 0 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          68               86.0759 %
Incorrectly Classified Instances        11               13.9241 %
Kappa statistic                          0.8198
Mean absolute error                      0.0628
Root mean squared error                  0.2
Relative absolute error                 26.598  %
Root relative squared error             50.7762 %
Total Number of Instances               79

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.8       0.037      0.909     0.8       0.851      0.973    C
                 0.941     0.032      0.889     0.941     0.914      0.998    A
                 0.625     0.056      0.556     0.625     0.588      0.947    B
                 0.9       0          1         0.9       0.947      0.997    F
                 0.947     0.05       0.857     0.947     0.9        0.984    D
Weighted Avg.    0.861     0.036      0.868     0.861     0.862      0.981

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 20  0  3  0  2 |  a = C
  0 16  1  0  0 |  b = A
  1  2  5  0  0 |  c = B
  0  0  0  9  1 |  d = F
  1  0  0  0 18 |  e = D
```

Figure 4.12: The result of Naive Bayes classifier when using 30% of data set for testing

Table 4.3: TP Rate FP Rate using Naive Bayes algorism

Naive Bayes								
	60% , 40%				70% , 30%			
Class	Training Set		Testing Set		Training Set		Testing Set	
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate
A	1	0.008	0.909	0.133	0.953	0.014	0.941	0.032
B	0.667	0	0.067	0.033	0.684	0.025	0.625	0.056
C	0.95	0.043	0.833	0.08	0.933	0.051	0.8	0.037
D	0.93	0.044	0.88	0.05	0.898	0.023	0.947	0.05
F	0.87	0.008	0.923	0.011	0.923	0.013	0.9	0
Weighted Average	0.923	0.026	0.762	0.069	0.901	0.026	0.861	0.036

4.3.1.1. Discussion of Naive Bayes classifier result

A- Result of 60% - 40% splitting

The detailed accuracy results for the Naive Bayes classifiers in training dataset revealed that the (TP) is excellent for the four classes. A (100%), C (95%), D (93%) and high for class F (87%) while it is good for class B (67%). (TP) for the testing dataset is excellent for two classes A (90%), F (92%). And also very high for two classes C (83%) , D (88%) and good for class B (67%) which it is the same present of the trained dataset.

The (FP) the training dataset is excellent for three classes B (0%), A (1%) and F (1%). Good for class C and D (4%) but in test dataset the percentage for class B and A increased by (3%), (13%) respectively while class F has the same percentage (1%) and its decreased in other two classes C by (3%), D by (1%).

B- Result of 70% - 30% splitting

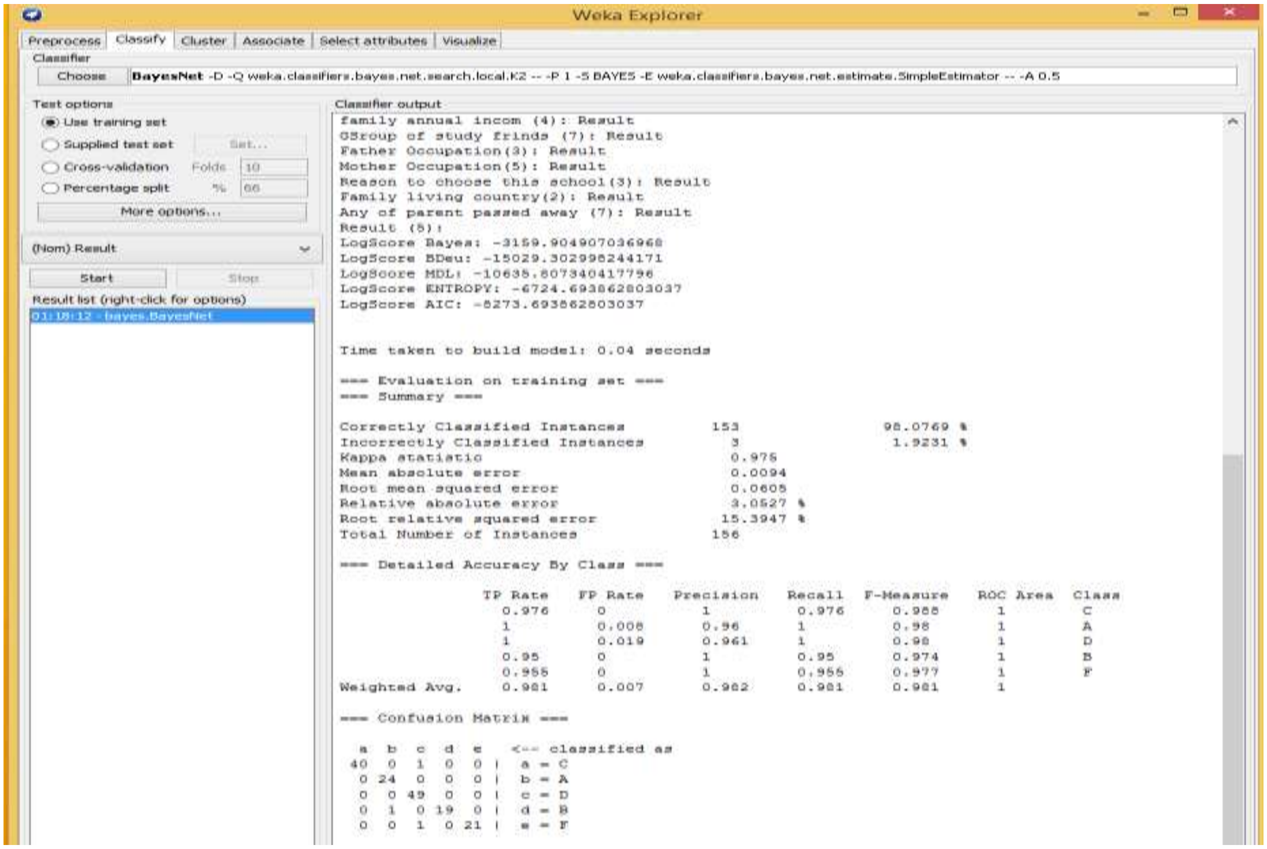
The detailed accuracy results for the Naive Bayes classifiers in the training dataset reveal that the (TP) is excellent for the four classes. A (95%), C (93%), F (92%) and D (90%) while it is good for class B (68%). (TP) for the testing dataset is excellent for two classes A (94%), D (95%), good for class B (63%) and very low for two classes C (8%), F (9%).

(FP) in training dataset is excellent for two classes A and F (1%), good for class B (3%), D (2%) and C (5%) but in the testing dataset the percentage for class B and A increased to (6%), (3%) respectively while class F has excellent percentage (0%), and decreased in other two classes C by (3%), D by (5%).

C- Comparison between results

The results show that the training dataset in 60%- 40% splitting to indicate better classification performance than 70% -30% splitting. While in testing datasets the 70% - 30% the perform better than 60% - 40% splitting has (86%) accuracy.

4.3.2 Result of BayesNet Classifier



Classifier

BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Test options

● Use training set

○ Supplied test set

○ Cross-validation Folds 10

○ Percentage split % 66

(Nom) Result

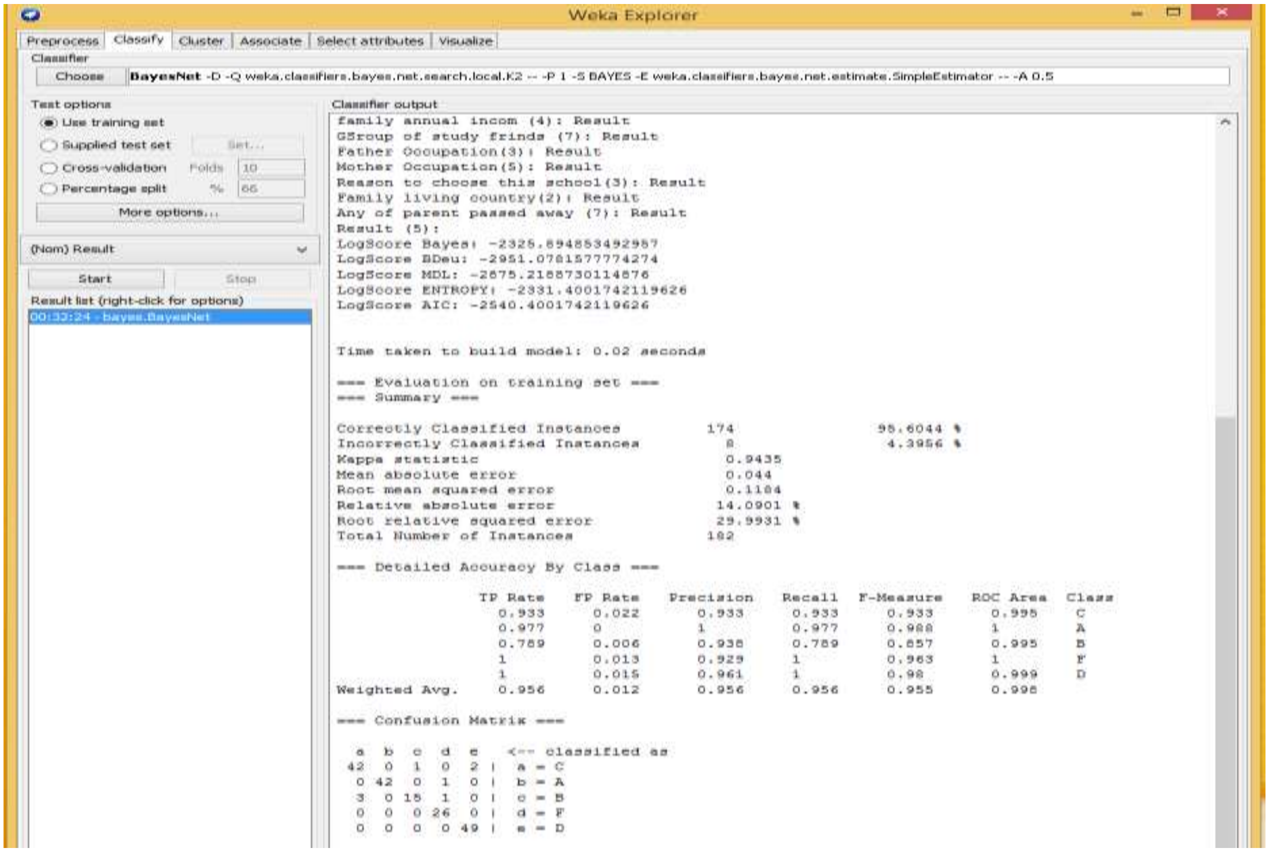
Result list (right-click for options)

01:18:12 - bayes.BayesNet

Classifier output

```
family annual incom (4): Result
GGroup of study frinds (7): Result
Father Occupation(3): Result
Mother Occupation(5): Result
Reason to choose this school(3): Result
Family living country(2): Result
Any of parent passed away (7): Result
Result (5):
LogScore Bayes: -3159.904907036968
LogScore BDeu: -15029.302998244171
LogScore MDL: -10635.807340417796
LogScore ENTROPY: -6724.693862803037
LogScore AIC: -8273.693862803037


Time taken to build model: 0.04 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         153               98.0769 %
Incorrectly Classified Instances         3                1.9231 %
Kappa statistic                          0.975
Mean absolute error                      0.0094
Root mean squared error                  0.0605
Relative absolute error                  3.0527 %
Root relative squared error             15.3947 %
Total Number of Instances              156

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.976     0         1         0.976     0.988      1        C
                 1         0.008     0.96      1         0.98       1        A
                 1         0.019     0.961     1         0.98       1        D
                 0.95      0         1         0.95      0.974      1        B
                 0.955     0         1         0.955     0.977      1        F
Weighted Avg.    0.981     0.007     0.982     0.981     0.981      1

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 40  0  1  0  0 |  a = C
  0 24  0  0  0 |  b = A
  0  0 49  0  0 |  c = D
  0  1  0 19  0 |  d = B
  0  0  1  0 21 |  e = F
```

Figure 4.13: The result of BayesNet classifier when using 60% of data set for training



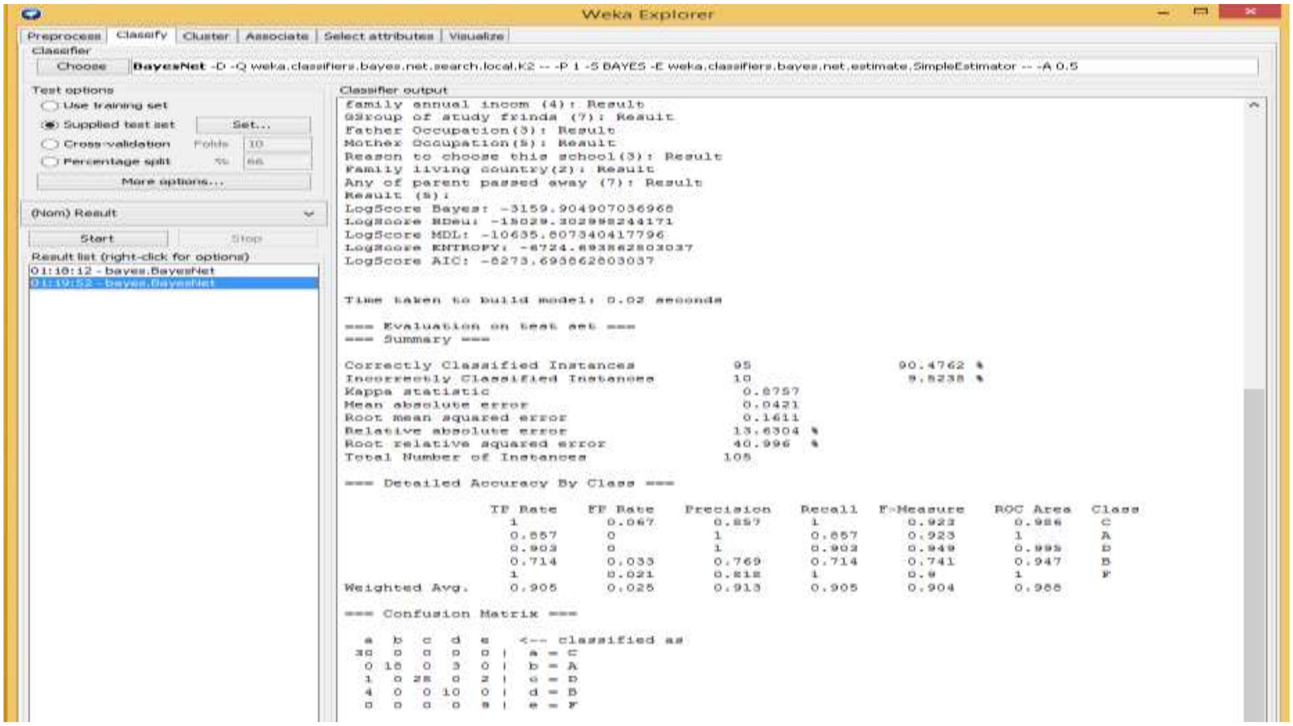
BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

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

(Nom) Result

```
family annual incom (4): Result
GSroup of study frinds (7): Result
Father Occupation(3): Result
Mother Occupation(5): Result
Reason to choose this school(3): Result
Family living country(2): Result
Any of parent passed away (7): Result
Result (5):
LogScore Bayes: -2325.894853492957
LogScore BDeu: -2951.0781577774274
LogScore MDL: -2675.2188730114876
LogScore ENTROPY: -2331.4001742119626
LogScore AIC: -2540.4001742119626


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         174               95.6044 %
Incorrectly Classified Instances         8                4.3956 %
Kappa statistic                          0.9435
Mean absolute error                      0.044
Root mean squared error                  0.1184
Relative absolute error                 14.0901 %
Root relative squared error             29.9931 %
Total Number of Instances              182

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.933     0.022     0.933     0.933     0.933      0.995    C
                 0.977     0         1         0.977     0.988      1        A
                 0.789     0.006     0.938     0.789     0.857      0.995    B
                 1         0.013     0.929     1         0.963      1        F
                 1         0.015     0.961     1         0.98       0.999    D
Weighted Avg.    0.956     0.012     0.956     0.956     0.955      0.998

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 42  0  1  0  2 |  a = C
  0 42  0  1  0 |  b = A
  3  0 15  1  0 |  c = B
  0  0  0 26  0 |  d = F
  0  0  0  0 49 |  e = D
```

Figure 4.14 : The result of BayesNet classifier when using 70% of data set for training



Classifier: BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Test options
○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

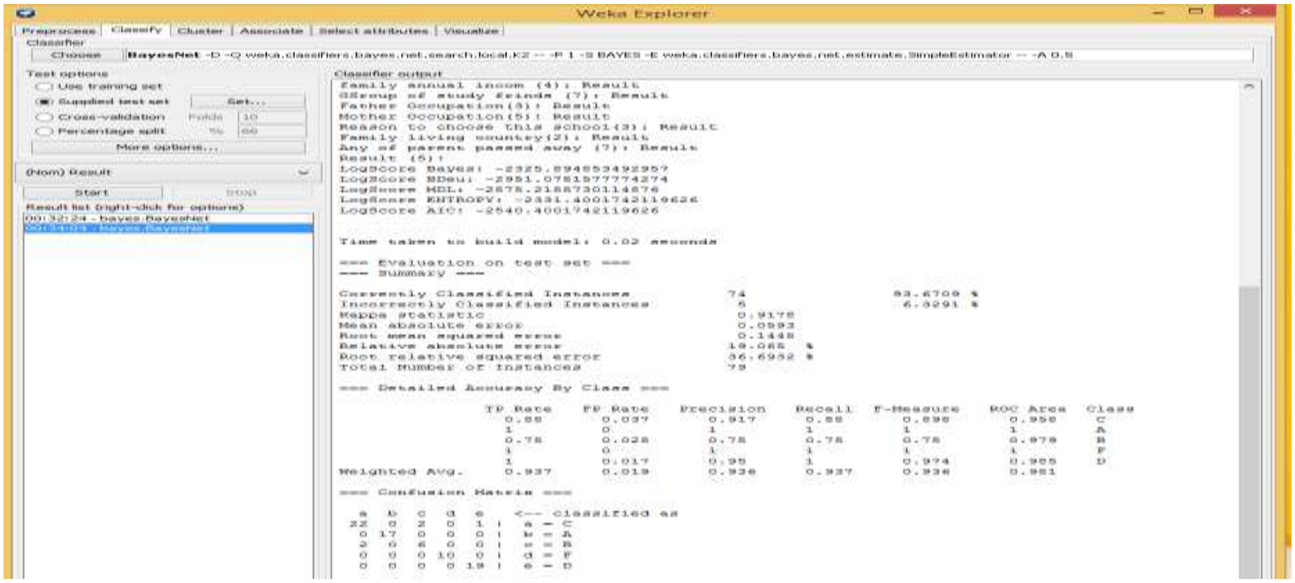
(Nom) Result

Result list (right-click for options)
01:18:12 - bayes.BayesNet
01:19:52 - bayes.BayesNet

Classifier output

```
family annual incom (4): Result
GGroup of study frinds (7): Result
Father Occupation(3): Result
Mother Occupation(5): Result
Reason to choose this school(3): Result
Family living country(2): Result
Any of parent passed away (7): Result
Result (5):
LogScore Bayes: -3159.904907036968
LogScore BDeu: -15029.302898244171
LogScore MDL: -10635.607340417796
LogScore ENTROPY: -6724.693862803037
LogScore AIC: -8273.693862803037


Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          95               90.4762 %
Incorrectly Classified Instances        10                9.5238 %
Kappa statistic                          0.8757
Mean absolute error                      0.0421
Root mean squared error                  0.1611
Relative absolute error                 13.6304 %
Root relative squared error             40.996  %
Total Number of Instances              105

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.067     0.857      1         0.923      0.986    C
                 0.857     0         1          0.857     0.923      1        A
                 0.903     0         1          0.903     0.949      0.995    D
                 0.714     0.033     0.769      0.714     0.741      0.947    B
                 1         0.021     0.818      1         0.9        1        F
Weighted Avg.    0.905     0.025     0.913      0.905     0.904      0.988

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 30  0  0  0  0 |  a = C
  0 18  0  3  0 |  b = A
  1  0 28  0  2 |  c = D
  4  0  0 10  0 |  d = B
  0  0  0  0  9 |  e = F
```

Figure 4.15 : The result of BayesNet classifier when using 40% of data set for testing



Classifier: BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Test options
○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Result

Result list (right-click for options)
00:32:24 - bayes.BayesNet
00:34:04 - bayes.BayesNet

Classifier output

```
family annual incom (4): Result
GGroup of study frinds (7): Result
Father Occupation(3): Result
Mother Occupation(5): Result
Reason to choose this school(3): Result
Family living country(2): Result
Any of parent passed away (7): Result
Result (5):
LogScore Bayes: -2325.894853492957
LogScore BDeu: -2951.0781577774274
LogScore MDL: -2678.2188730114876
LogScore ENTROPY: -2331.4001742119626
LogScore AIC: -2540.4001742119626


Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          74               93.6709 %
Incorrectly Classified Instances         5                6.3291 %
Kappa statistic                          0.9178
Mean absolute error                      0.0593
Root mean squared error                  0.1448
Relative absolute error                 19.085  %
Root relative squared error             36.6952 %
Total Number of Instances               79

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.88      0.037     0.917      0.88      0.898      0.958    C
                 1         0         1          1         1          1        A
                 0.75      0.028     0.75       0.75      0.75       0.978    B
                 1         0         1          1         1          1        F
                 1         0.017     0.95       1         0.974      0.985    D
Weighted Avg.    0.937     0.019     0.938      0.937     0.936      0.981

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 22  0  2  0  1 |  a = C
  0 17  0  0  0 |  b = A
  2  0  6  0  0 |  c = B
  0  0  0 10  0 |  d = F
  0  0  0  0 19 |  e = D
```

Figure 4.16 : The result of BayesNet classifier when using 30% of data set for testing

Table 4.4: TP Rates and FP Rates when using BayesNet algorism

BayesNet								
	60% , 40%				70% , 30%			
Class	Training Set		Testing Set		Training Set		Testing Set	
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate
A	1	0	0.909	0.048	0.977	0	1	0
B	0.833	0	0.467	0.033	0.789	0.006	0.75	0.028
C	0.975	0.017	0.9	0.067	0.933	0.022	0.88	0.037
D	1	0.009	1	0.025	1	0.015	1	0.017
F	1	0	0.769	0.022	1	0.013	1	0
Weighted Average	0.981	0.007	0.848	0.043	0.956	0.012	0.937	0.019

4.3.2.1 Discussion of BayesNet classifier result

A- Result of 60% - 40% splitting

The results of training dataset show that the (TP) is excellent for three classes A, B and C (100%) and high for other two classes B (83%), C (97%). In testing dataset the percentage decreased in four classes A (91%), B (47%), (1%) and F (77%). While class D has the same percentage (100%).

The (FP)in training dataset also performs excellent for three classes A, B and F (0%) , while its high for other two classes C (2%) and D (1%) while in testing dataset the percentage increased in all classes class B and D have the same present (3%) , class F (2%) , class A (5%) and class C (7%)

B- Result of 70% - 30% splitting

The results of training dataset shows that the (TP) is excellent for three classes in both training and testing dataset D and F (100%) and high for other two classes A (98%), C (93%),quite high for class B (79%). In testing dataset the percentage increased for class A (100%) and decreased for the other two classes, B (75%) and C (88%).

The (FP) in training dataset also performs excellent for three classes A (0%) and B, F (1%), while its high for other two classes C and D (2%).While in testing dataset the class A and F have excellent performance (0%) and the percentage increased for other three classes, B (3%), C(4%) and D (2%)

C- Comparison between the results

The result show that the training dataset in 60%- 40% splitting performed better classification (TP,FP)than 70% -30% splitting. While in test datasets the 70% - 30% performed better than 60% - 40% splitting.

4.4 Performance analysis for all applied classifiers

Table 4.5: Summarize the performance of different classifier

60%, 40% splitting								
	Training				Test			
	J48	Random Forest	Naive Bayes	BayesNet	J48	Random Forest	Naive Bayes	Bayes Net
Correctly Classified Instances	97%	100%	92%	98%	95.2%	78 %	76.1%	85%
Incorrectly Classified Instances	3%	0%	7%	1.9%	4.7 %	21.9 %	23.8 %	15%
70%, 30% splitting								
Correctly Classified Instances	97%	100%	90%	96%	100%	84%	86%	94%
Incorrectly Classified Instances	3%	0%	10%	4%	0%	16%	14%	6%

Designing the student model to predict students' performance by analyzing training data 60% and testing data 40%, random forest classifier has maximum accuracy in the training set 100% followed by BayesNet 98%, while in testing set J48 has maximum accuracy by 95%. But when using training data 70% and test data 30%. Random forest has maximum accuracy in training dataset by 100%. Followed by J48 97%, while in testing set J48 has maximum accuracy by 100% which mean that the J48 is more reliable algorisms used to predict student performance.

4.4.1 Comparison using 60% - 40%

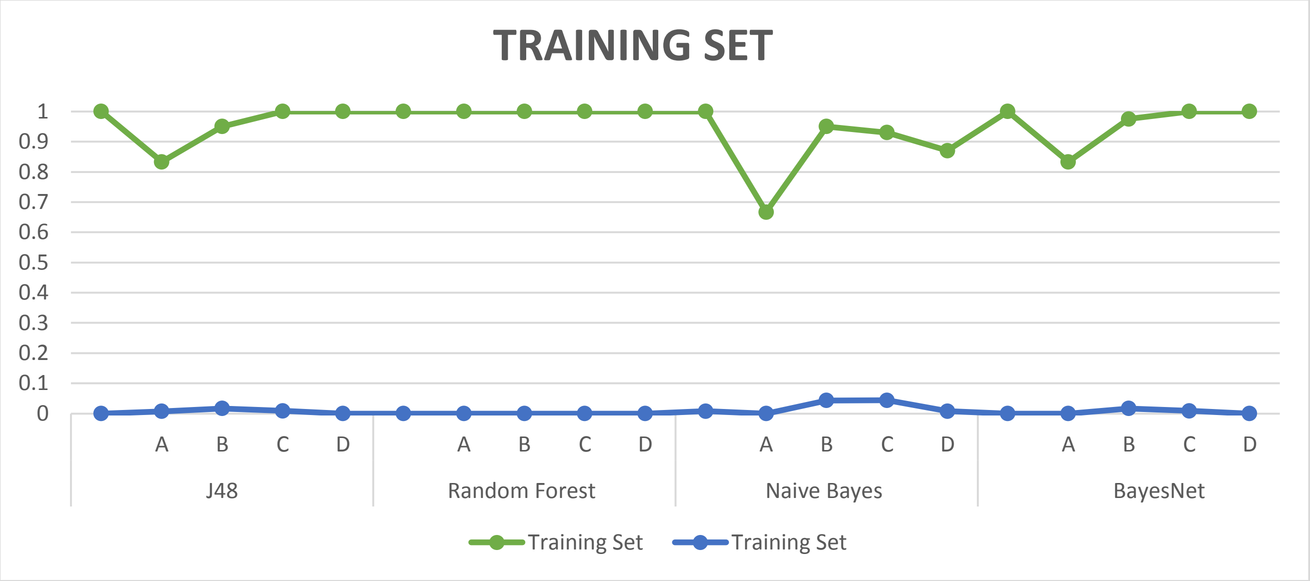


Figure 4.17 the performance charts of different classifier based on TP Rate, FP Rate when using 60% training dataset

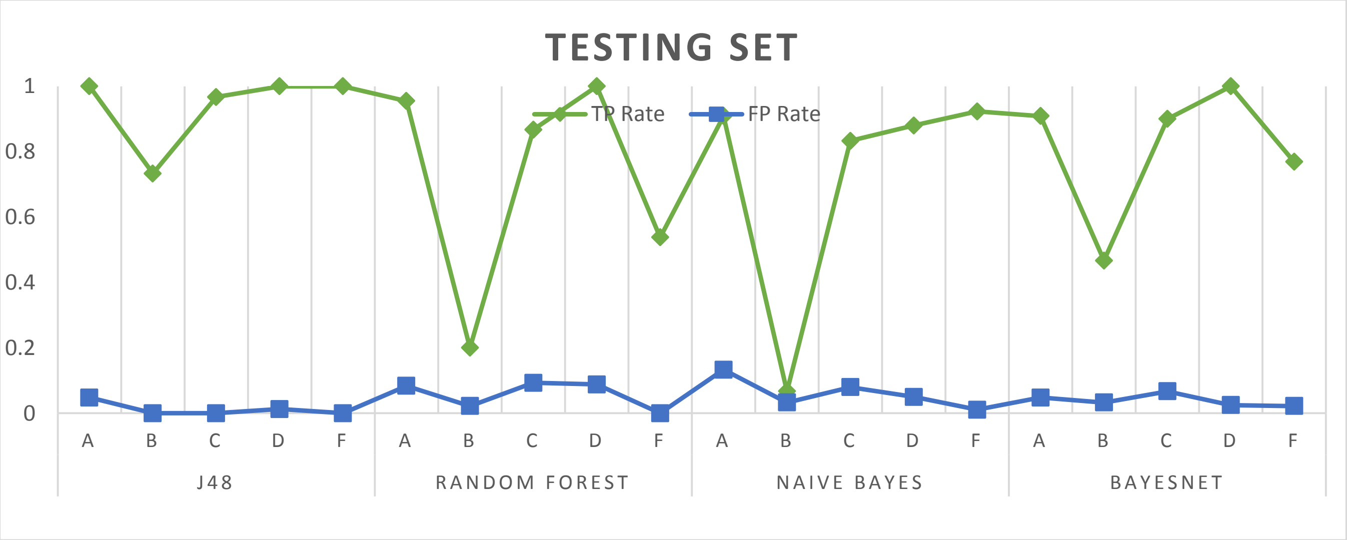


Figure 4.18 the performance charts of different classifier based on TP Rate, FP Rate when using 40% testing dataset

Figures (4.18, 4.19) show the performance charts of different classifier based on TP Rate, FP Rate in both training dataset and test dataset. In training dataset with the Random forest algorithm, all of (TP Rate and FP Rate) obtain high values followed by the J48 algorisms, which have the high value in FP Rate. FP Rate in naïve Bayes and Bayes net has the low value. The J48, Bayes net and naïve Bayes algorithm obtained low values in TP Rate. While in the test dataset J48 algorisms obtain high value in TP Rate and FP Rate compared by others algorisms. Thus, based on this charts, j48 is the better choice for prediction process.

4.4.2 Comparison using 70% - 30%

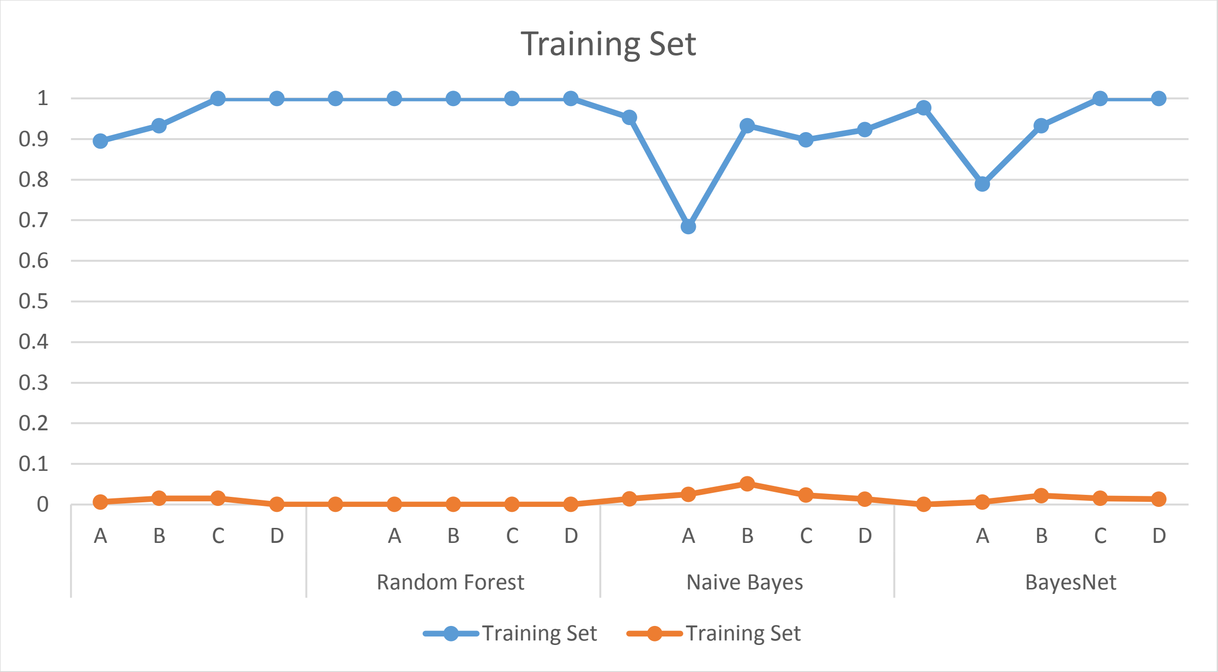


Figure 4.19 the performance charts of different classifier based on TP Rate, FP Rate when using 70% training dataset

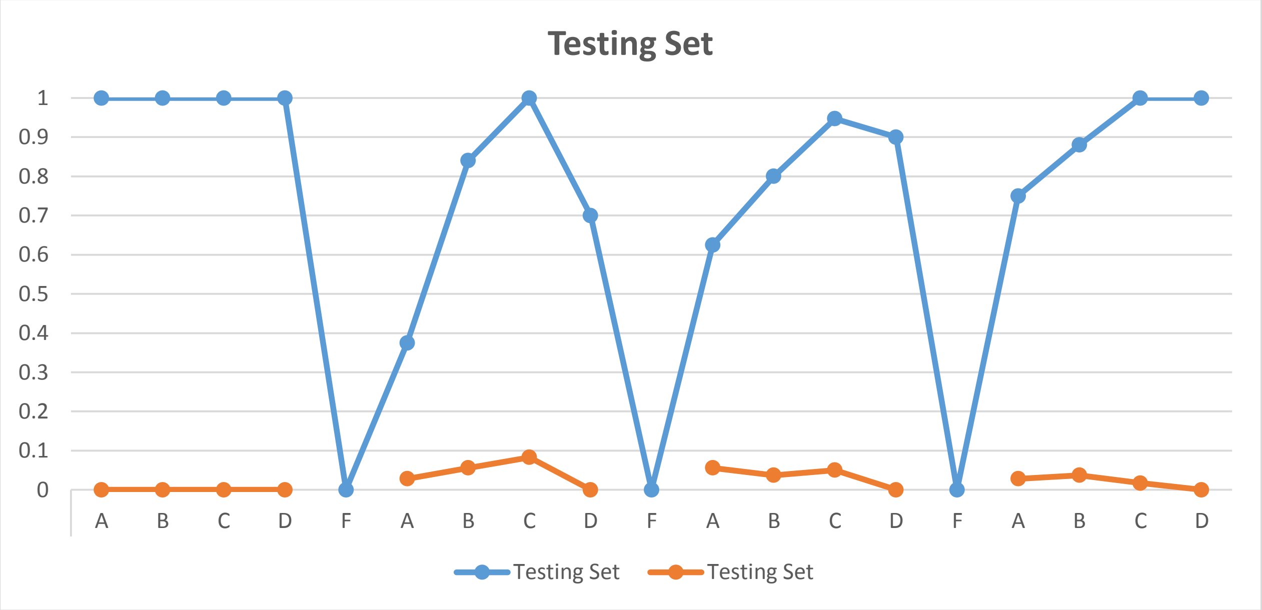


Figure 4.20 the performance charts of different classifier based on TP Rate, FP Rate when using 30% testing dataset

Figure (4.20, 4.21) shows the performance charts of different classifier based on TP Rate, FP Rate in both training dataset (70%) and test dataset (30%). In training dataset with the Random forest algorithm, all of (TP Rate and FP Rate) obtain high values followed by the J48 algorisms, FP Rate in naïve Bayes and Bayes net has the low value. The Bayes net and naïve Bayes algorithm obtains low values in TP Rate. While in the test dataset J48 algorisms obtain highest value in TP Rate and FP Rate compared by others algorisms. Thus, based on this charts j48 is the better choice for prediction process. Because it was the best method for prediction. the researcher make tree Visualization ,the attribute with high ranking in two test was Sem 1 GPA appears at the first level of the tree, as shown in figure 4.21.

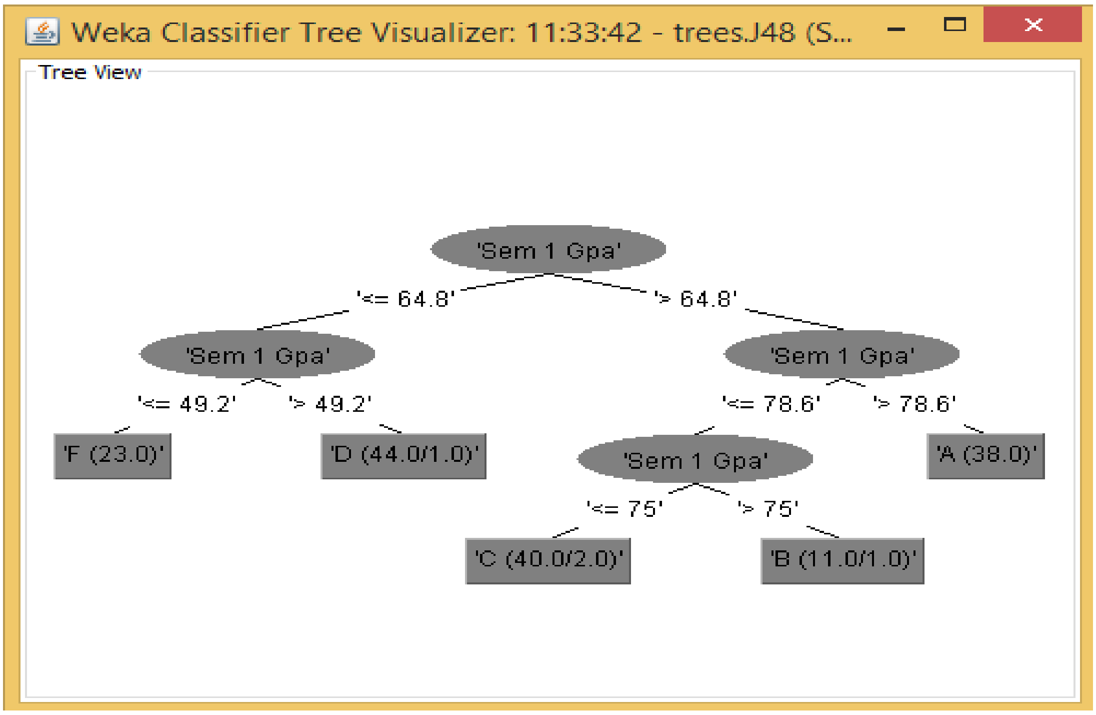


Figure 4.21: J48 Tree Visualization

4.5 The strongest predictor's factors that affect students' performance:

To find out the most affected factor on student performance the researcher used SPSS analysis (cross table, chi-squer), it's found that there are two factor that mostly affected the students' performance (student family status, any of parents pass away) as they shown in the tables and figures below.

Table: 4.6 Family status

Family Status				
Result	**Divorce**		**Still married**	
	Frequency		**Frequency**	
High Performance	22	8%	135	52%
Low Performance	34	13%	69	27%
P.Value	.000			

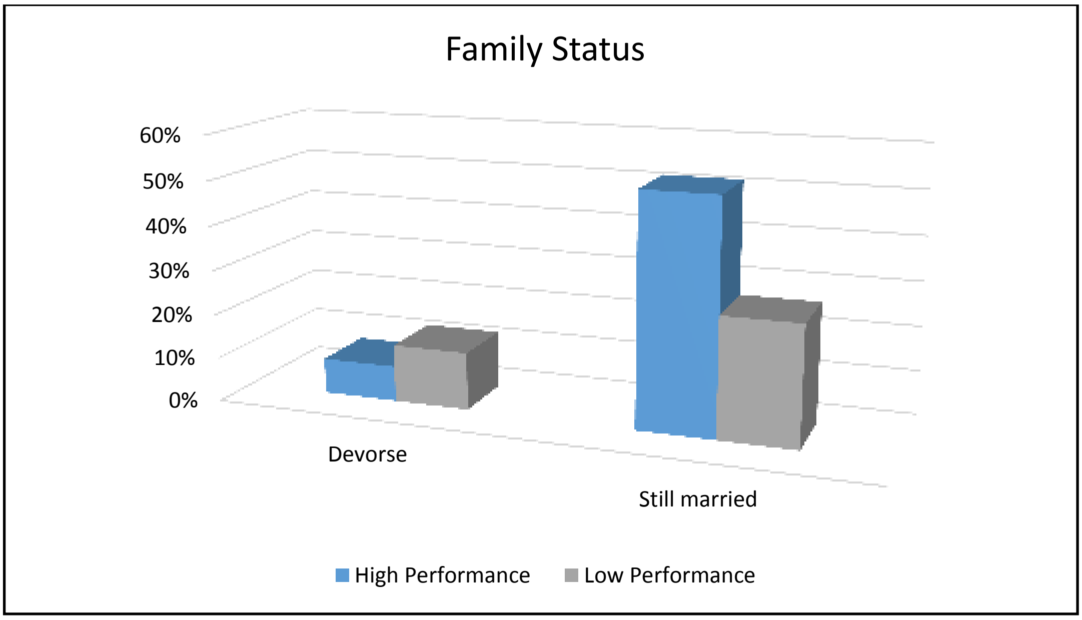


Figure 4.22 Family Status

Table 4.6 and figure 4.22 Show that most of students that their family status divorce have greater percentage in low performance (13%), than the high performance which is (8%). On the other hand to the student family status still married, have high performance (52%) are greater than students that have low performance (27%). Consequently the family status affect student performance with p.value .000.

Table 4.7 Show if any of the parents of students pass away

Result	Any of parents passed away							
	Both		Father		Mother		None of them	
	Frequency		Frequency		Frequency		Frequency	
High Performance	6	2%	13	5%	2	0.7%	133	51%
Low Performance	5	2%	24	9%	13	5%	63	24%
P.Value	.002							

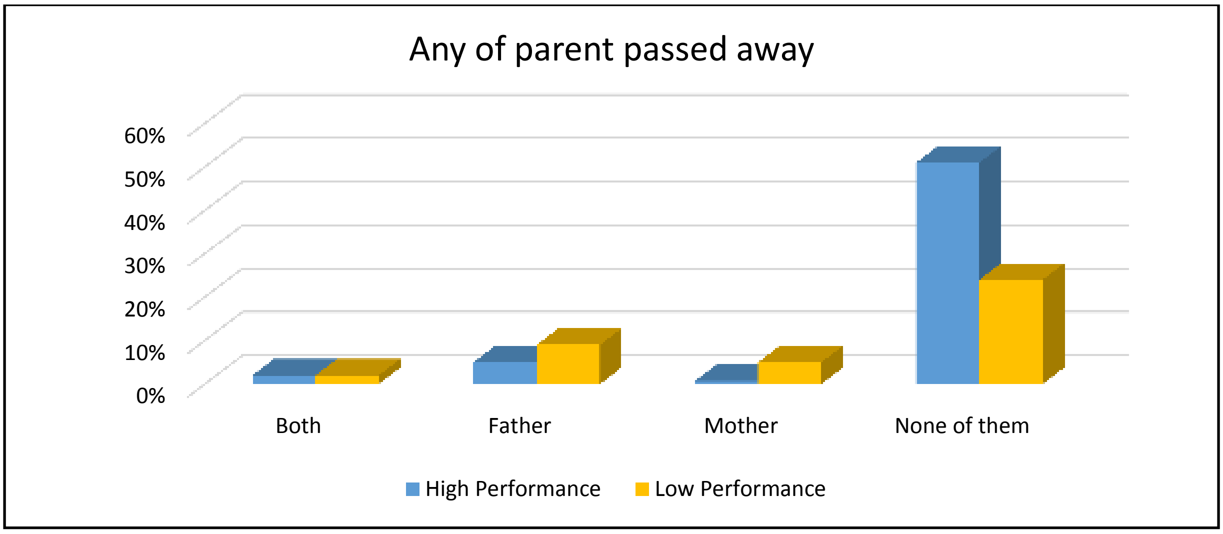


Figure 4.23 Show if any of the parents of students pass away

Table 4.7 And figure 4.23 indicate that student that have one or both of their parents died have greater percentage in low performance than the high performance , while the student that does not have any of their parents died have greater percentage in high performance than low performance. Therefore the existence of the parents influences students' performance with p.value .002.

CHAPTER V

Conclusion and Recommendation

5.1 Conclusion

This study aimed at collecting quantitative data that represent social factors of students, to build the classification model that classifies a students' performance, to find out if there are any patterns in the available data that could be useful for predicting students' performance at the university based on their social factors. The university management would like to know which features in the currently available data are the strongest predictors of university performance and to evaluate the performance of different classification techniques. In order to meet this objective, the researcher start collecting quantitative data, which includes questions related to several personal, socioeconomic and psychological variables the researcher distributed this questionnaire to students in freshman year 2018 -2019 at School of Management in AUW.

And to find out if there are any patterns in the available data that could be useful for predicting students' performance at the university based on students social factors the researcher used a comparative analysis of four classification techniques; decision tree (j48, random forest) and Bayesian classifiers (naïve Bayes, Bayes net) using WEKA tool.

Finally to know which features in the currently available data are the strongest predictors of university performance the researcher used SPSS analysis.

The experimental results showed that J48 is the best algorithm for classification of data, it also showed that academic performance of the students is not always depend on their own efforts. The Study showed that social factors have got significant influence over students' performance, furthermore, the family status and any of parents died they are the most affected factors on the student performance by .000 and .002 Pvalue respectively.

This study will help the academic advisor to monitor the students' performance in a systematic way by identify those students who needed special attention to reduce failing ration and taking appropriate action for the next semester at a right time.

5.2 Recommendation

In the future, this study will be expanded by adding more data from different years and several departments by increase dataset size as following

- Used historical data.to get more information about student social factor
- Applied in the rest of year of study.
- Applied in the rest college of university.
- Applied in different university.

In order to extract useful information about key factors affecting students' performance and increase the accuracy of the prediction.

References

- Ahmad, F., Ismail, N. H. and Aziz, A. A. (2015) ‘The prediction of students’ academic performance using classification data mining techniques’, Applied Mathematical Sciences, 9(129), pp. 6415–6426. doi: 10.12988/ams.2015.53289.
- Al-barrak, M. A. and Al-razgan, M. (2016) ‘Predicting Students Final GPA Using Decision Trees : A Case Study’, 6(7). doi: 10.7763/IJIET.2016.V6.745.
- Al-radaideh, Q. A. (2014) ‘Mining Student Data Using Decision Trees’, (January 2006).
- Alaoui, S. S., Farhaoui, Y. and Aksasse, B. (2018) ‘Classification Algorithms in Data Mining : A Survey’, 3(1), pp. 349–355.
- Alaoui, S. S., Farhaoui, Y. and Aksasse, B. (2018) ‘Classification Algorithms in Data Mining : A Survey’, 3(1), pp. 349–355.
- Badr, A., Din, E. and Elaraby, I. S. (2014) ‘Data Mining : A prediction for Student ’ s Performance Using Classification Method’, World Journal of Computer Application and Technology, 2(2), pp. 43–47. doi: 10.13189/wjcat.2014.020203.
- Baradwaj, B. and Pal, S. (2012) ‘Mining educational data to analyze student’s performance’, Internation Journal od Advamced Computer Science and Applications, 2(6), pp. 63–69. doi: vol.2,No.6.
- Barahate, S. R. (2012) ‘Educational Data Mining as a Trend of Data Mining in Educational System’, Proceedings of IJCA International Conference and Workshop on Emerging Trends in Technology, pp. 11–16.
- Bashar, M. O. (2018) ‘Sudan University of Science and Technology College of Graduate Studies Improving Students Academic Performance Using Hybrid Recommendation Techniques نيسحت ءادأ بلاطلا يميداكلأا مادختساب قرط ةيصوت نيجه’, (November).
- Borkar, S. and Rajeswari, K. (2013) ‘Predicting students academic performance using education data mining’, International Journal of Computer Science and Mobile Computing, 2(7), pp. 273–279.
- Cheewaprakobkit, P. (2015) ‘Predicting Student Academic Achievement by Using the Decision Tree and Neural Network Techniques’, 12(2), pp. 2408–137.
- Choudhury, S. D. et al. (2017) ‘Altered translational repression of an RNA-binding protein, Elav by AOA2-causative Senataxin mutation’, Synapse, 71(5), pp. 20–25. doi: 10.1002/syn.21969.

- Feyyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. IEEE expert, 11(5), 20-25.
- Gadhavi, M. and Patel, C. (2017) ‘STUDENT FINAL GRADE PREDICTION’, 8(3), pp. 274–279.
- Govindasamy, K. (2018) ‘ANALYSIS OF STUDENT ACADEMIC PERFORMANCE USING’, 119(15), pp. 309–323.
- Hamoud, A. K. and Hashim, A. S. (2017) ‘Students ’ Success Prediction based on Bayes Algorithms Students ’ Success Prediction based on Bayes Algorithms’, (November). doi: 10.5120/ijca2017915506.
- Hooshyar, D., Pedaste, M. and Yang, Y. (2020) ‘Mining educational data to predict students’ performance through procrastination behavior’, Entropy, 22(1), p. 12. doi: 10.3390/e22010012.
- J. Kovacic, Z. (2010) ‘Early Prediction of Student Success: Mining Students Enrolment Data’, pp. 647–665. doi: 10.28945/1281.
- Kabakchieva, D. (2012) ‘Student performance prediction by using data mining classification algorithms’, International Journal of Computer Science and Management Research, 1(4), pp. 686–690.
- Kabakchieva, D. (2013) ‘Predicting student performance by using data mining methods for classification’, Cybernetics and Information Technologies, 13(1), pp. 61–72. doi: 10.2478/cait-2013-0006.
- Kaur, G. and Singh, W. (2016) ‘Prediction Of Student Performance Using Weka Tool’, 17(January), pp. 8–16.
- Kaur, H. (2015) ‘EDM: A Review of Applications of Data Mining in the Field of Education’, India, 4(4), pp. 409–412. doi: 10.17148/IJARCCE.2015.4492.
- Kumar, S. A. (2011). Efficiency of decision trees in predicting student’s academic performance.
- Minaei-Bidgoli, B. (2004) ‘Data Mining for a Web-Based Educational System’, Thesis, p. 267.
- Mobasher, G., Shawish, A. and Ibrahim, O. (2017) ‘Educational Data Mining Rule based
- Recommender Systems’, 1(Csedu), pp. 292–299. doi: 10.5220/0006290902920299.
- Mobasher, G., Shawish, A. and Ibrahim, O. (2017) ‘Educational Data Mining Rule based Recommender Systems’, 1(Csedu), pp. 292–299. doi: 10.5220/0006290902920299.

- Mueen, A. (2016) ‘Modeling and Predicting Students ’ Academic Performance Using Data Mining Techniques’, (November), pp. 36–42. doi: 10.5815/ijmecs.2016.11.05.
- Pal, M. (2005) ‘Random forest classifier for remote sensing classification’, International Journal of Remote Sensing, 26(1), pp. 217–222. doi: 10.1080/01431160412331269698.
- Review, E. (2012) ‘Data mining approach for predicting student performance’, X(1), pp. 3–12.
- Romero, C. and Ventura, S. (2007) ‘Educational data mining: A survey from 1995 to 2005’, Expert Systems with Applications, 33(1), pp. 135–146. doi: 10.1016/j.eswa.2006.04.005.
- Science, I. et al. (2018) ‘STUDENT PLACEMENT ANALYZER : A RECOMMENDATION SYSTEM USING MACHINE LEARNING’, (3), pp. 2058–2060.
- Saa, A. A. (2016). Educational Data Mining & Students’ Performance Prediction. International Journal of Advanced Computer Science and Applications, 7(5), 212-220.

- Science, I. et al. (2018) ‘STUDENT PLACEMENT ANALYZER : A RECOMMENDATION SYSTEM USING MACHINE LEARNING’, (3), pp. 2058–2060.

- Srivastava, Jaideep; Cooley, R; Deshpande, M; Tan, P. N. (2000) ‘Web Usage Mining: Discovery and Applications of Usage Patterns fromWeb Data’, SIGKDD Explorations, 1(2), pp. 12–23. doi: 10.1145/846183.846188.
- Srivastava, J. and Srivastava, D. A. K. (2013) ‘Data Mining in Education Sector: A Review’, Special Conference Issue: National Conference on Cloud Computing & Big Data, pp. 184–190.
- Staeheli, L. A. and Mitchell, D. (2010) ‘Relevance’, The SAGE Handbook of Social Geographies, pp. 546–559. doi: 10.4135/9780857021113.n29.
- Suhirman, X. X., Zain, J. M. and Herawan, T. (2014) ‘Data mining for education decision support: A review’, International Journal of Emerging Technologies in Learning, 9(6), pp. 4–19. doi: 10.3991/ijet.v9i6.3950.
- Sumitha, R. and Vinothkumar, E. S. (2016) ‘Prediction of Students Outcome Using Data Mining Techniques’, (6).
- Tampakas, V. et al. (2019) ‘Prediction of students’ Graduation time using a two-level classification algorithm’, Communications in Computer and Information Science,

993, pp. 553–565. doi: 10.1007/978-3-030-20954-4_42.

- Wiener, A. L. and M. (2003) ‘Classification and Regression by randomForest. R News 2’, 3(December 2002), pp. 18–22.
- Yadav, S., Bharadwaj, B. and Pal, S. (2012) ‘Data mining applications: A comparative study for predicting student’s performance’, International Journal of Innovative Technology & Creative Engineering, 1(12), pp. 13–19.
- Yadav, S. K. and Pal, S. (2012) ‘Data Mining : A Prediction for Performance Improvement of Engineering Students using Classification’, World of Computer Science and Information Technology Journal WCSIT, 2(2), pp. 51–56. doi: 10.1142/9789812771728_0012.

Appendix

Appendix A

Sudan University of Science and Technology

College of Graduate Studies

Master of Information Technology

Questionnaire for freshmen students' school of management studies

2018 – 2019

Student Performance Prediction Using Data Mining Techniques

(Ahfad University for Women

Index No _________________________

1- Student Family Status

a- Diverse **b-** Still married

2- Student live in a dormitory

a- Yes **b-** No

c-

3- Father qualification

a- no-education **b-** elementary **c-** secondary **d-** graduate

e- post-graduate **f-** Doctorate **g-** not-applicable

4- Mother qualification

a- no-education **b-** elementary **c-** secondary **d-** graduate

D- post-graduate **f-**Doctorate **g-** not-applicable

5- Family annual income

a- Poor **b-** medium **b-** high

6- Group of study friends

a- Alone **b-** 2-3 **d-** 4-5 **d** - 6 – more

7- Father Occupation

a- Service **b-** retired **e-** not-applicable

8- Mother Occupation

a- House wife **b-** Service **c-**retired **f-** not-applicable

9- Reason to choose this school

a- My choice **b-** my parent choice

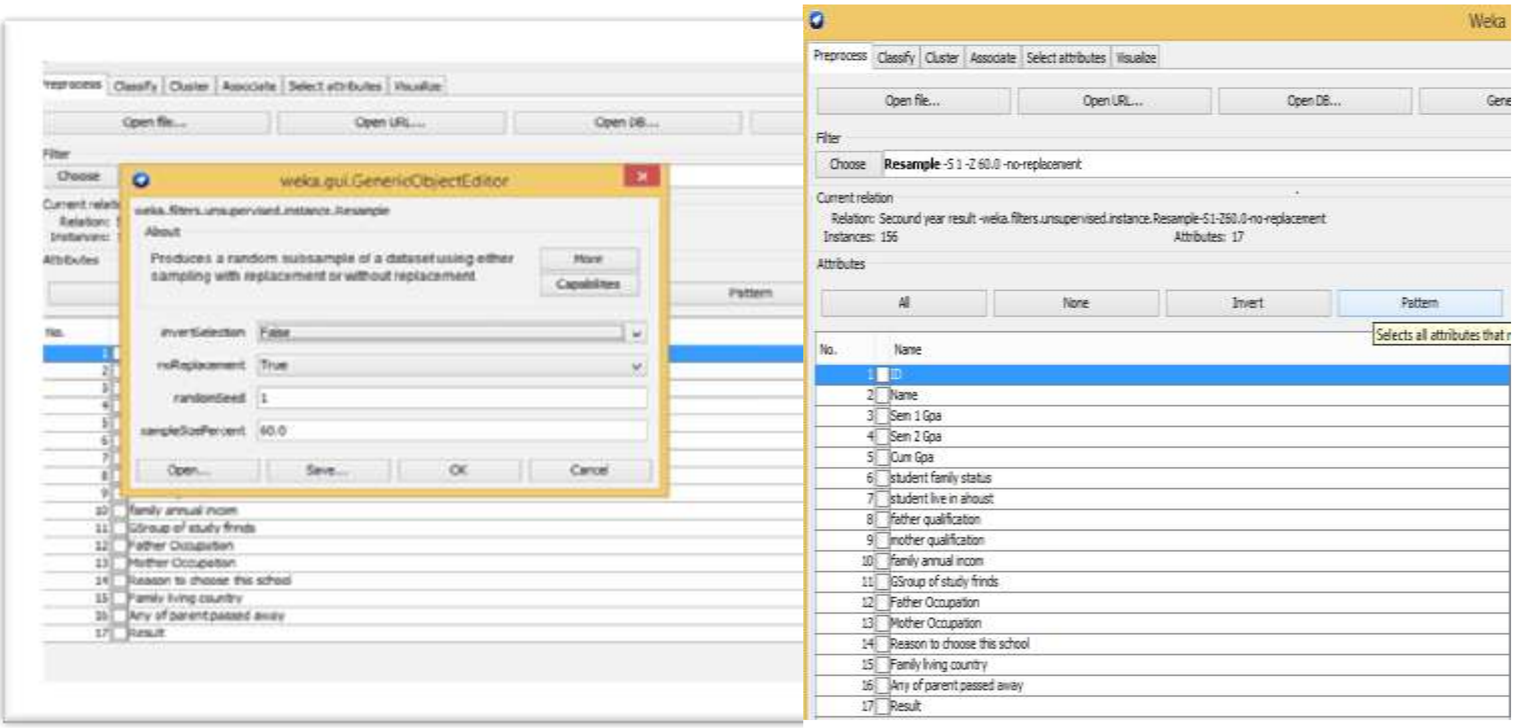
10- Family living Country

a- Outside the country **b-** in the country

11- Do you have any of your parent passed away

a- Father **b-** Mother **c-**Both **c-** None of them

Appendix B



weka.gui.GenericObjectEditor

weka.filters.unsupervised.instance.Resample

About

Produces a random subsample of a dataset using either sampling with replacement or without replacement.

More
Capabilities

invertSelection: False
noReplacement: True
randomSeed: 1
sampleSizePercent: 60.0

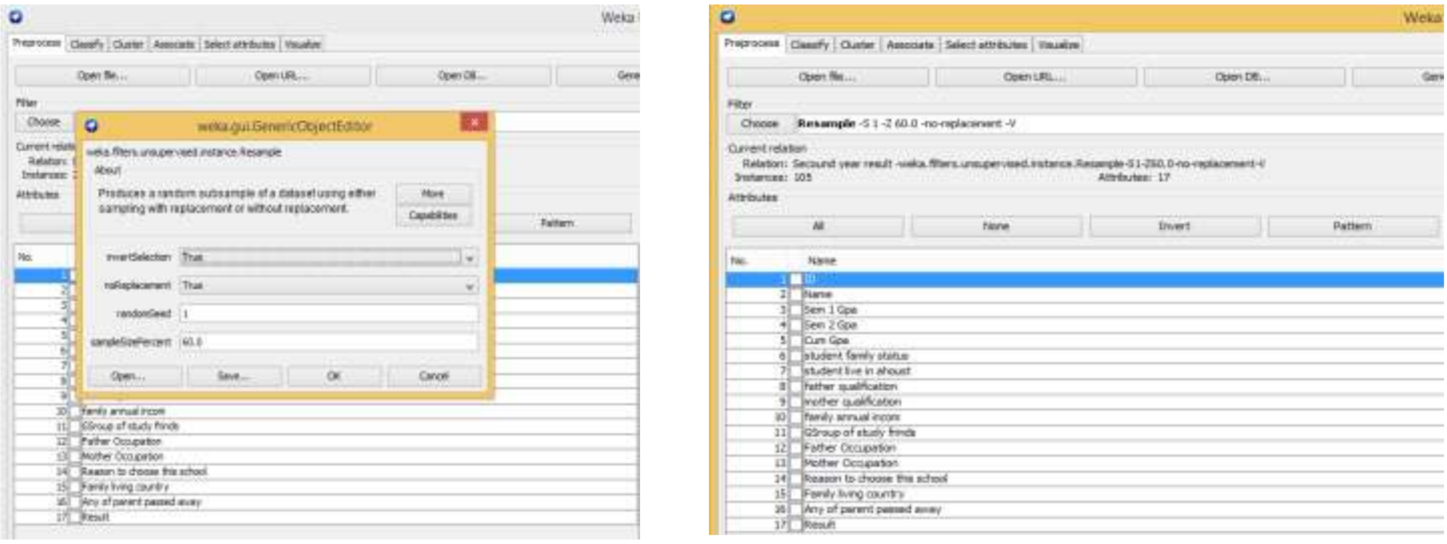
Open... Save... OK Cancel

Filter: Resample -S 1 -Z 60.0 -no-replacement

Relation: Secound year result -weka.filters.unsupervised.instance.Resample-S1-Z60.0-no-replacement
Instances: 156
Attributes: 17

No.	Name
1	ID
2	Name
3	Sem 1 Gpa
4	Sem 2 Gpa
5	Cum Gpa
6	student family status
7	student live in ahoust
8	father qualification
9	mother qualification
10	family annual incom
11	GSroup of study finds
12	Father Occupation
13	Mother Occupation
14	Reason to choose this school
15	Family living country
16	Any of parent passed away
17	Result

Figure B.1 Create Training Set



weka.gui.GenericObjectEditor

weka.filters.unsupervised.instance.Resample

About

Produces a random subsample of a dataset using either sampling with replacement or without replacement.

More
Capabilities

invertSelection: True
noReplacement: True
randomSeed: 1
sampleSizePercent: 60.0

Open... Save... OK Cancel

Filter: Resample -S 1 -Z 60.0 -no-replacement -V

Relation: Secound year result -weka.filters.unsupervised.instance.Resample-S1-Z60.0-no-replacement-V
Instances: 105
Attributes: 17

No.	Name
1	ID
2	Name
3	Sem 1 Gpa
4	Sem 2 Gpa
5	Cum Gpa
6	student family status
7	student live in ahoust
8	father qualification
9	mother qualification
10	family annual incom
11	GSroup of study finds
12	Father Occupation
13	Mother Occupation
14	Reason to choose this school
15	Family living country
16	Any of parent passed away
17	Result

Figure B.2 Create Testing Set

Appendix C

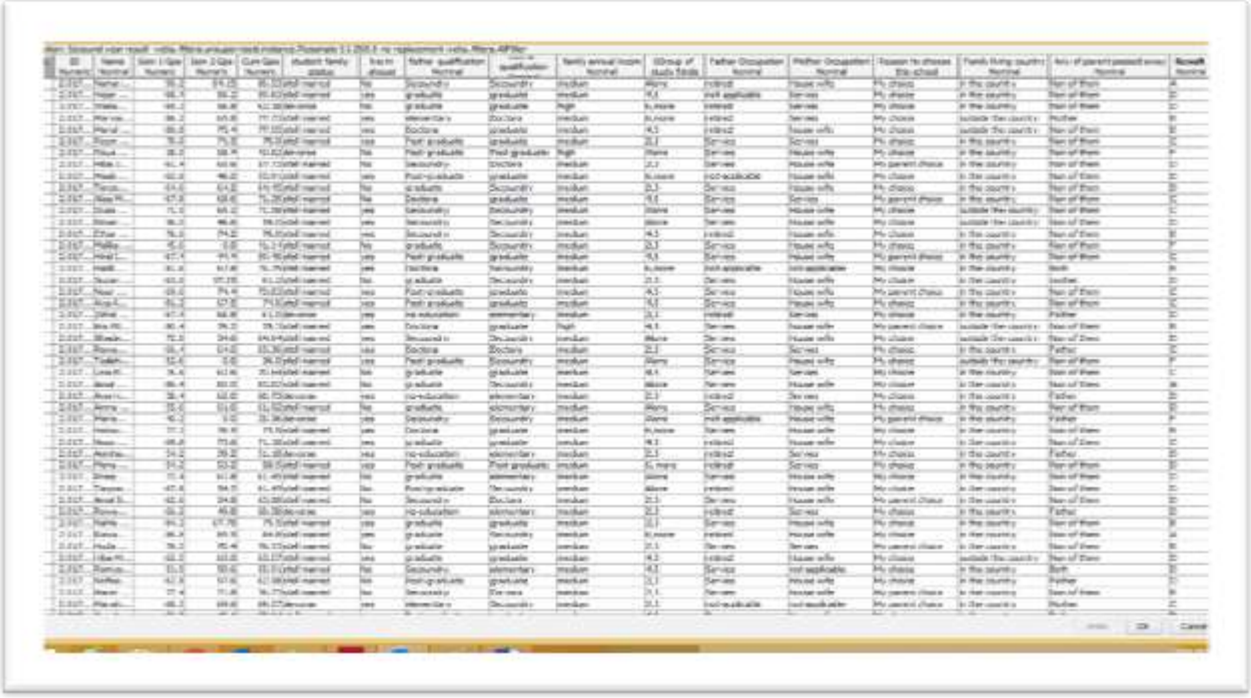


Figure C.1 Example of Training Data set

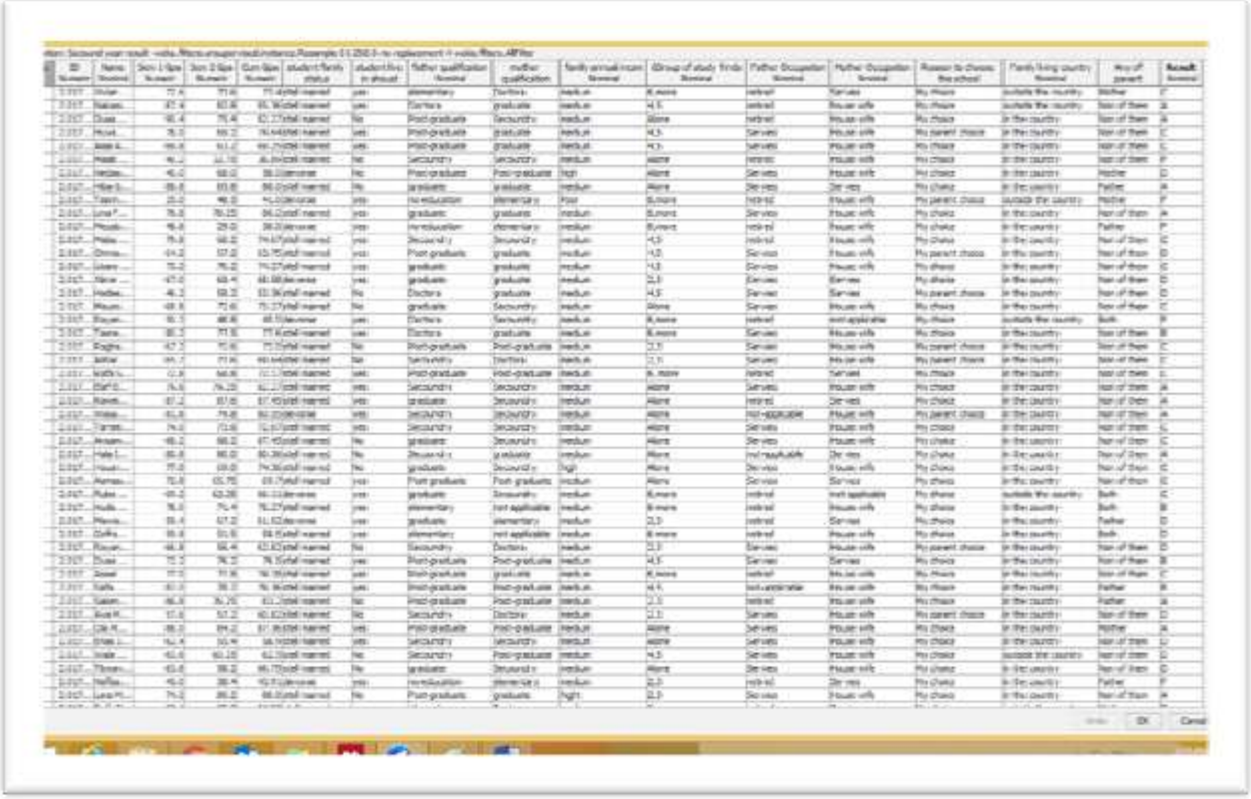


Figure C.2 Example Test Data set