MODELING AN AUTOMATED STUDENT'S PERFORMANCE PREDICTOR BY USING DECISION TREE

Mukashyaka Charlotte
University of Lay Adventists of Kigali
Email: mukacharlotte@gmail.com

Abstract

The student performance is essential key factors in the educational field as the goal of our country is to promote the students with high quality of education on the market. Nowadays, the students enroll on the modules or courses during the registration period and they pay the school fees according to registered modules. The students who fail the module must retake it in the next years. All High learning institutions use management information system that holds the information about their students, these information's are potential to the different task. Data mining enables them to extract and discover the unknown knowledge from data stored in the database. This field of data mining uses the academic settings called Educational data mining (EDM). Predicting student performance is one of the applications conducted in this field. In this paper, we developed an automated model that will predict student performance, this enables the HLI to analyze and model the performance of their students at the middle of their study. Classification techniques are one technique used for predicting useful knowledge through the historical information by data mining tools. We analyzed data by using three decision tree algorithms named Iterative Dichotomiser3 (ID3), J48 and classification and regression (CART) decision tree. After analyzing data, we find that the ID3 with an accuracy of 70.4% was the best algorithm used because it had higher accuracy compared to J48 and CART

Keywords: Modeling automated, Student performance, predictor, data mining techniques

1. Introduction

Data mining is referred to the process of extracting hidden and useful information in large data repositories. Knowledge Discovery and Data Mining (KDD) is a multidisciplinary area focusing upon methodologies for extracting useful knowledge from data and there are several useful KDD tools to extracting the knowledge. This knowledge can be used to increase the quality of education. Educational Data Mining is concerned with developing new methods to discover knowledge from educational/academic database and can be used for decision making in educational or academic systems (B.Namratha, 2016).

Education is a very important issue regarding the development of any country. Nowadays, the Information Management System used in different services where the organizations store their information in a database in order to manage them in an effective and efficient manner. All higher learning institutions (HLI) have the vast amount of data that are stored in databases. The data contain the identification of students, student marks as well as the modules registered to the student. Educational Data Mining (EDM) is used for mining useful patterns and discovering useful knowledge from the educational information systems, such as, admissions systems, registration systems, course management systems and any other systems dealing with students at different levels of education, from schools, to colleges and universities.

The main objective of higher Learning institutions is to provide quality education to its students. One way to achieve the highest level of quality in higher education system is by discovering knowledge for prediction regarding student performance in order to take measures that enable the student to get a high quality in education (Saa, Educational Data Mining & Students' Performance Prediction, 2016).

All students who start the HLI does not have the same capacity for learning, It is very difficult to know the student who is able for success or not in order to help them to get high performance. The HLI need the tools that allow them to verify the performance of the student from first year to the last one.

Based on (Fadhilah Ahmad*, 2015), (Ahmed Mueen, 2016,) all authors designed a modeling and predicting students' academic performancebased on particular courses in a particular year and semester. These models are semi-automated means that they need the users with special skills that enable them to analyze and predict student performance, the users without that knowledge do not use these systems

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

and do not use them to predict the student from different years or program.

Our study is similar to the previous studies because it analyzes and collects data from registration officer and academic reports that enable us to predict student performance based on relationship between Input variables and predict variables. Almost of the above studies analyze and predict student performance for particular courses, in first year at specific semester while our study analyze and predict data from different level and program.

The contribution of this research is to design an automated model for student performance that is capable to accept the student Independent variables and display the dependent variable based on the student identifications by using data from different years and different program.

ID3, J48 and CART are three types of decision tree algorithm used in this study. WEKA is intelligent learning tools allow us to analyze data and to build a student 'performance model and Java program enable us to design an automated student's performance predictor based on the rules from the decision tree.

2. Methods

2.1. Data mining methodology

Figure 1 depicts the work methodology used in this study, which is based on the framework proposed. The methodology starts from the problem definition, data collection, then preprocessing and the data set and preprocessing sections, then we come to the data mining methods with classification techniques by using decision tree with ID3, J48 and CART followed by the evaluation of results and patterns, finally the knowledge representation process (Mohammed M abutter and Alaa Mustafa EL-Halees, 2012).

2.2. Classification

Classification is the most familiar and the most effective data mining techniques used to classify and predict values. It is the process of finding a set of models that discuss and differentiate data idea and classes, for the purpose of being able to use the model to guess the class whose label is unknown (Eshwari Girish Kulkarni,Raj B. Kulkarni, PhD, 2016).There are multiple different classification methods and techniques used in Knowledge Discovery and data mining. The decision tree is selected to be applied to the students' data. The Classification is one of the most used and studied data mining because it is simple and easy to use. (Amjad Abu Saa, 2016)

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

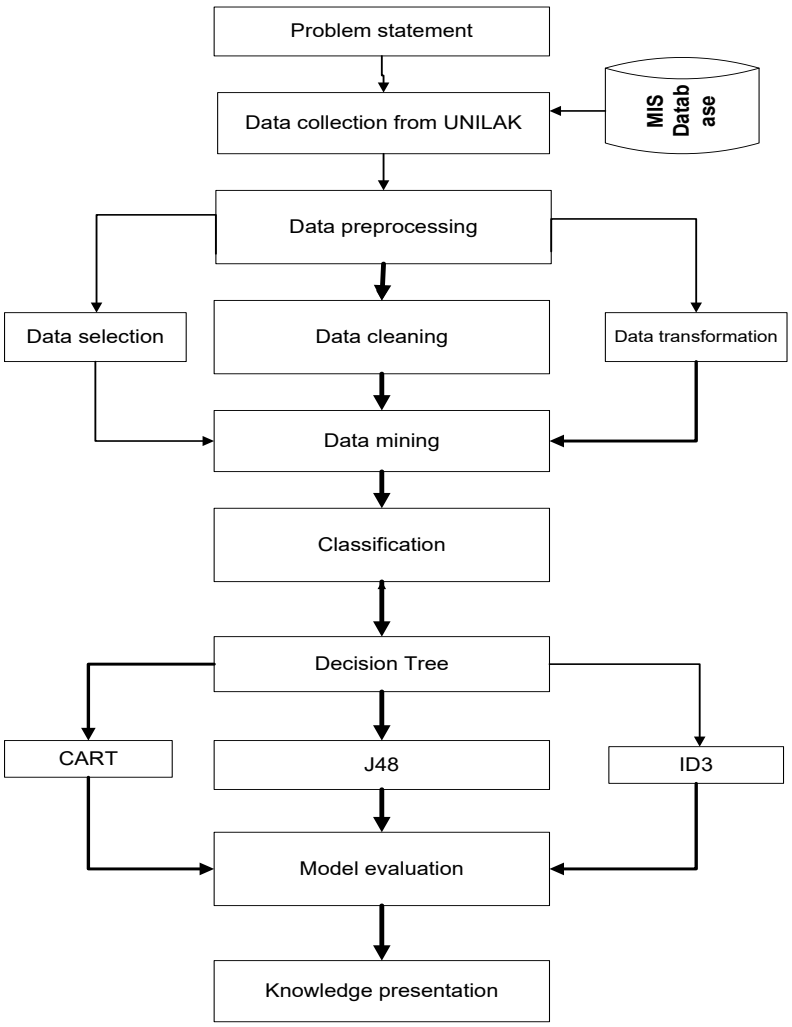


Figure 1: Data mining work methodology

2.2.1 Decision tree

A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily be converted to classification rules. (Jiawei Han, 2012).

ID3, J48, CART are decision tree algorithms to be used in the project.

1. Id3

The ID3 (Iterative Dichotomiser) Algorithm is introduced in 1986 by Quinlan Ross. The ID3 is a Greedy approached learning decision tree algorithm. This algorithm is recursively selected the best attribute as the current node using top down induction. Then the child nodes are generated for the selected Attribute.

It uses on information as entropy based measure to select the best splitting attribute and the

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

attribute with the highest information gain is select as best splitting attribute. ID3 algorithm generates an unpruned full decision tree from a dataset.

2. J48

J48 algorithm is called as optimized implementation of the C4.5 or improved version of the C4.5. Output given by J48 is the Decision tree. Decision tree divides the input space of a dataset into mutually exclusive areas, where each area having a label, a value to describe or elaborate its data point. The splitting criterion is used in decision tree to calculate which attribute is the best to split that portion tree of the training data that reaches a particular node (Tina R. Patil, 2013).

3. The Classification And Regression Tree

The algorithm CART stands for Classification and Regression Tree introduced by Brieman. It's also based on Hunt's Algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. (S.Nagaparameshwara chary, 2017)

3. Result and discussion

This section contains Dataset, data presentation, decision tree, generated rules, and classifier comparison, id3 model and an automated student 'performance predictor

3.1 Dataset

The data made available for analysis came from the UNILAK Management System. These datasets contained 1339 records for the student registered in academic year 2015-2016 in different years and different program in accounting department.Data regarding the student registration have been stored in a Microsoft EXCEL spreadsheet file, then in CVS and covert into an ARFF suitable by WEKA tools.

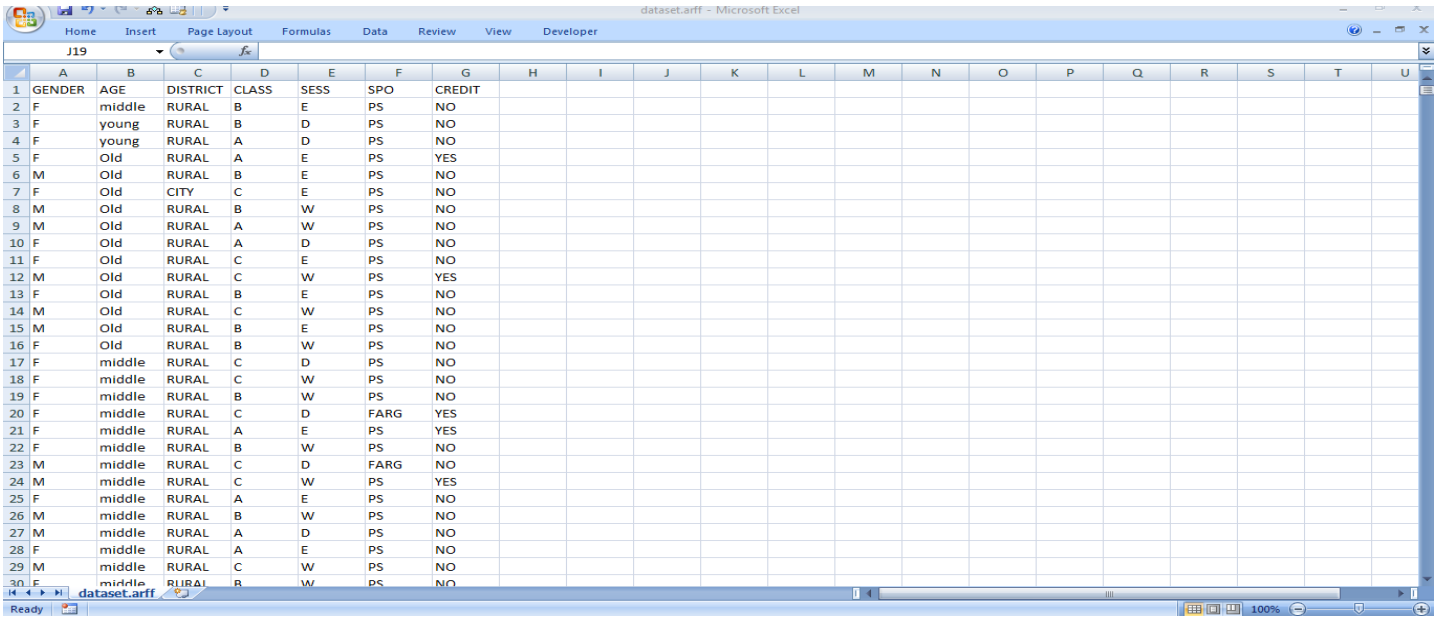
Table 1 describes attributes of a dataset that is divided into two categories independent variables and the dependent variable. First six attributes are independent variables and the last attribute is dependent variable.

Table 1: attribute description

Attribute	Description	Possible values
Gender	Sex with student	F and M
Age	Student's age	Middle, young and old
District	Resident district of student	Rural and city
Class	Class of student	A, B, C and D

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

Session	Programs	D, E and W
Sponsorship	Sponsorship (private or FARG	PS and FARG
Credit	Credit failed	Yes and No



	A	B	C	D	E	F	G
1	GENDER	AGE	DISTRICT	CLASS	SESS	SPO	CREDIT
2	F	middle	RURAL	B	E	PS	NO
3	F	young	RURAL	B	D	PS	NO
4	F	young	RURAL	A	D	PS	NO
5	F	Old	RURAL	A	E	PS	YES
6	M	Old	RURAL	B	E	PS	NO
7	F	Old	CITY	C	E	PS	NO
8	M	Old	RURAL	B	W	PS	NO
9	M	Old	RURAL	A	W	PS	NO
10	F	Old	RURAL	A	D	PS	NO
11	F	Old	RURAL	C	E	PS	NO
12	M	Old	RURAL	C	W	PS	YES
13	F	Old	RURAL	B	E	PS	NO
14	M	Old	RURAL	C	W	PS	NO
15	M	Old	RURAL	B	E	PS	NO
16	F	Old	RURAL	B	W	PS	NO
17	F	middle	RURAL	C	D	PS	NO
18	F	middle	RURAL	C	W	PS	NO
19	F	middle	RURAL	B	W	PS	NO
20	F	middle	RURAL	C	D	FARG	YES
21	F	middle	RURAL	A	E	PS	YES
22	F	middle	RURAL	B	W	PS	NO
23	M	middle	RURAL	C	D	FARG	NO
24	M	middle	RURAL	C	W	PS	YES
25	F	middle	RURAL	A	E	PS	NO
26	M	middle	RURAL	B	W	PS	NO
27	M	middle	RURAL	A	D	PS	NO
28	F	middle	RURAL	A	E	PS	NO
29	M	middle	RURAL	C	W	PS	NO

Figure 2: Dataset in excel

A dataset from the database are collected into an excel sheet and some information are combined together.

The WEKA cannot accept any data than ARFF, CSV, C4.5 and binary

We save the excel file into CSV Format (Office button-→Save as →save as types →choose CSV then Save

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

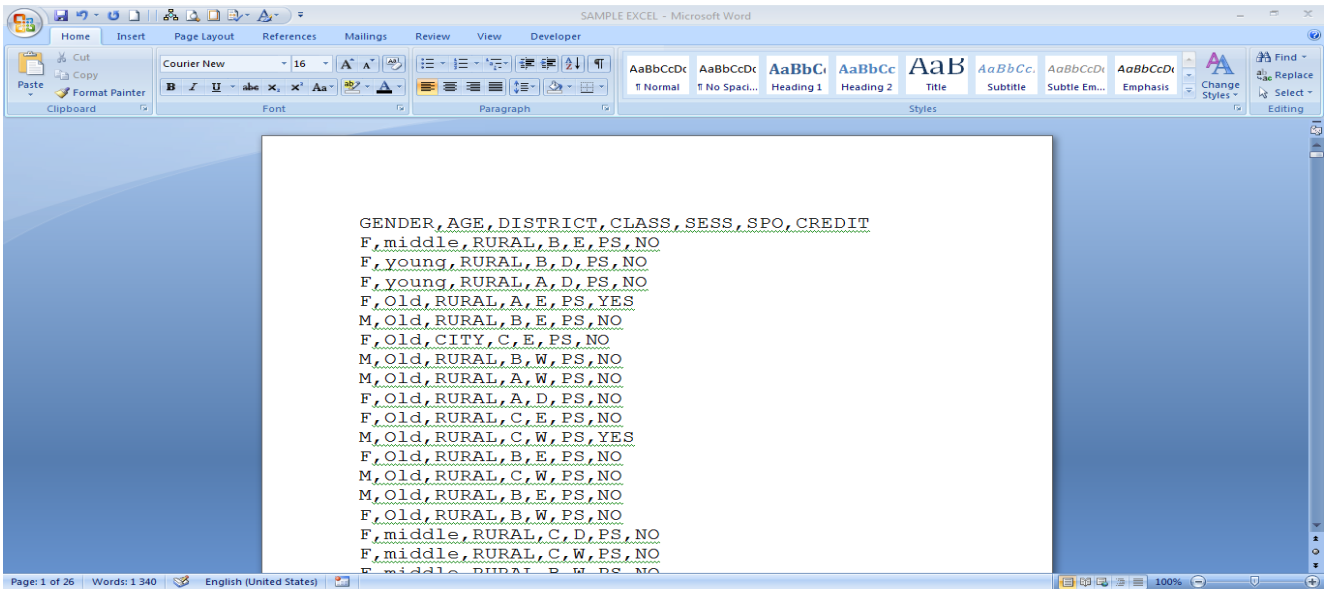


Figure 3: Dataset in MS Word

When MS EXCEL dataset figure2 is opened in MS word, we find the comma that separate element to another see figure3.

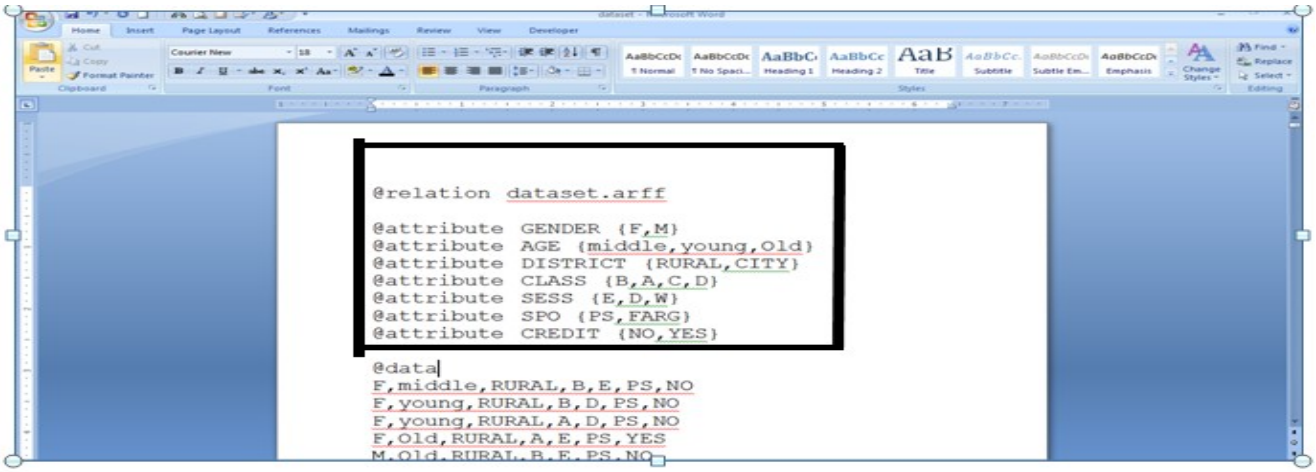


Figure 4: Dataset in ARFF

The highlighted lines are used to convert MS word into ARFF format which is suitable for WEKA tool, replace the first line "GENDER, AGE, DISTRICT, CLASS, SESS, SPO, CREDIT" in Figure 3 with the highlighted lines.

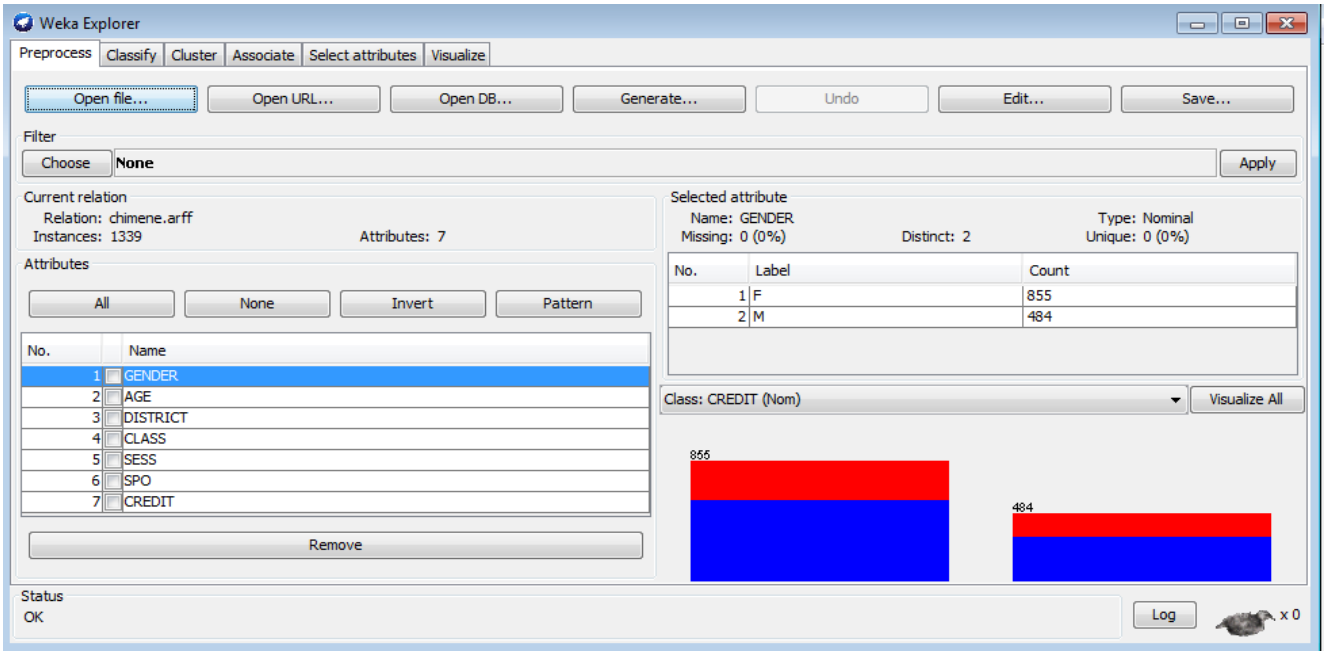


Figure 5: data preparation

After converting MS word into the ARFF format we extracted data from ARFF into WEKA tool. The figure 5 shows the attributes of the dataset and the graph that shows the selected attribute as gender attribute. The red color shows fail target while blue color is success target. In our dataset, we have 855 of female (63.8%) with579 success and 276 fail and 484 of male (36.2%) with 314 successes and 170 fail. Based on the number of dataset table2 shows value of fail and success for each attribute.Table3 describe the percentage for each value

Table 2: attribute name, the number of failing and success

Attribute	Element	Success	Fail	Total
Gender	Female	579	276	855
	Male	314	170	484
Age	Young	417	212	629
	Middle	359	186	545
	Old	117	48	165
District	City	177	111	288
	Rural	716	335	1051
	A	305	157	462

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

Class	B	162	121	283
	C	379	156	535
	D	47	12	59
SESSION	Day	386	158	544
	Evening	154	71	225
	Weekend	353	217	570
SPONSOR	FARG	155	68	223

Table 3: Percentage of success and fail

ATTRIBUTE VALUE	Success (%)	Fail (%)
Female	67.7	32.2
Male	64.8	35.2
Young	66.2	33.8
Middle	65.8	34.2
Old	70.9	29.1
City	61.4	38.5
Rural	68.1	31.8
1st	66	34
2nd	57.2	42.7
3rd	70.8	29.2
4th	79.6	20.4
Day	70.9	29.1
Evening	68.4	31.5
Weekend	61.9	38.1
FARG	69.5	30.5
Private	66.1	33.9

The table3 shows that the Female are more successful than male. The old are more successful than middle and youngest.

Students of Rural district are more successful than city district. The students of third and fourth are more successful than students of first and second year. The students studied at day and evening is more successful than the students of the weekend. The students funded by FARG are more successful than private's students

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

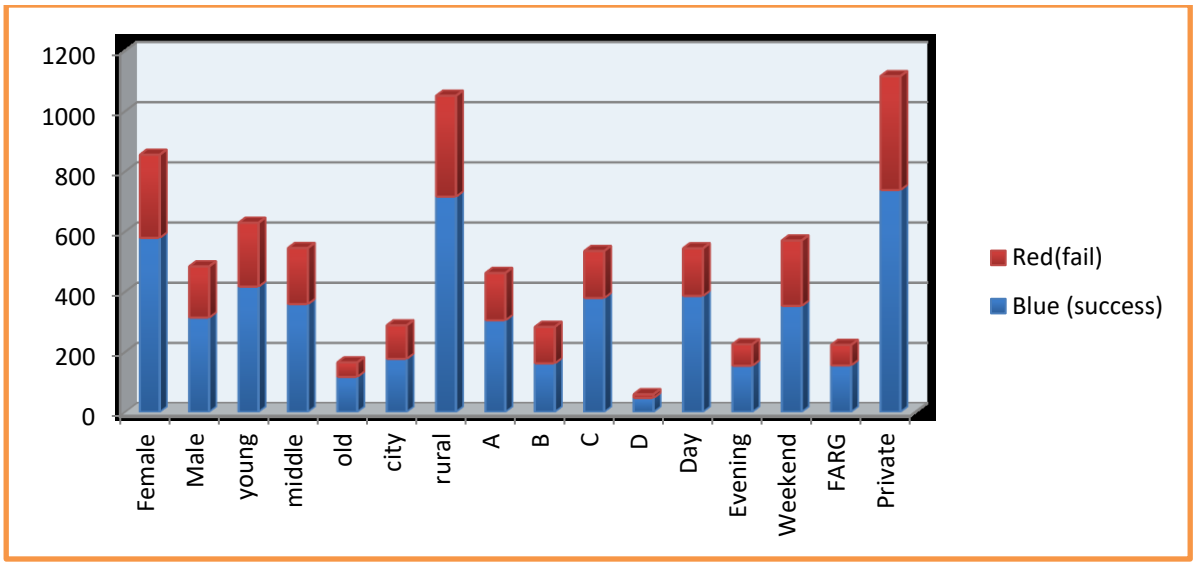


Figure 6 :success and fail of each attribute

3.2. Decision tree Visualization

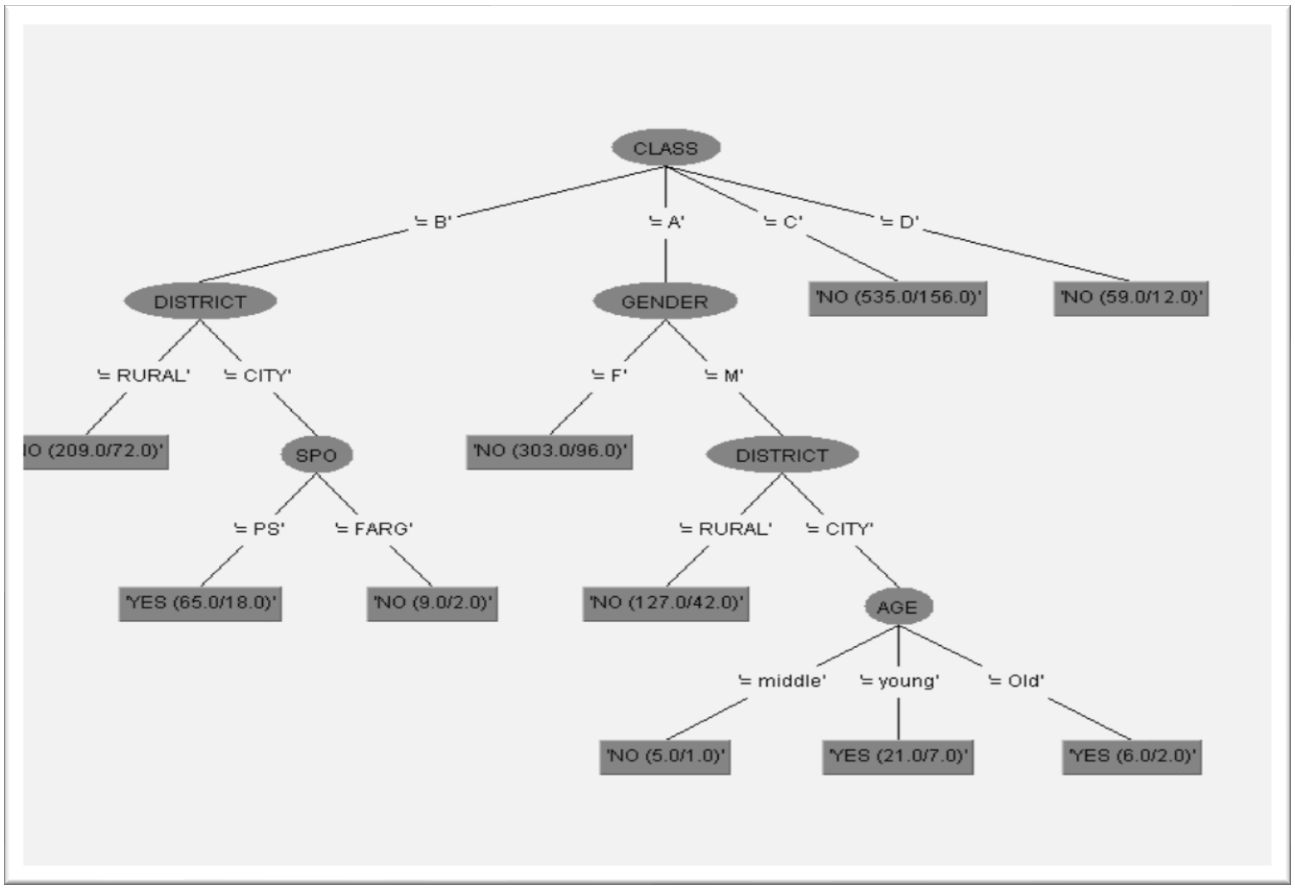


Figure 7: Decision tree visualization

The above decision tree was shown that the class attribute is the root of decision tree because it contains the highest information gain than the others.

The result of a decision tree shows the following: the students from class "C" (class 3) and "D" (class 4) are performing well than the classes. The students from second year (class B) their Performance depends on the distinct means that the students from rural district perform well than the student from city district. The success of students from the city based on the sponsorship means the students funded by FARG are well performed. First year student (Class "A"), the gender is one factor that affects the student of this year, females are well performed while the success of male depend on the district means the males from Rural are well performed. The success of the male from the city districtdepends on the age means the student with young and old age are more failed than students of middle age.

3.3 Generated rules from decision tree

1. IF class ='C' and Class= 'D'then Credit='NO'
2. IF class ='B' and district ='Rural' then Credit ='NO'
3. IF class ='B' and district ='CITY' and sponsor ='PS' then Credit= 'YES'
4. IF class ='B' and district ='CITY' and sponsor ='FARG' then Credit= 'NO'
5. IF class ='A' and Gender ='F' then Credit='NO'
6. IF class ='A' and Gender ='M' and District ='Rural' then Credit='NO'
7. IF class ='A' and Gender ='M' and District ='Rural' then Credit='NO'
8. IF class ='A' and Gender ='M' and District ='CITY' and age =Young then Credit='Yes'
9. IF class ='A' and Gender ='M' and District ='CITY' and age = old then Credit='Yes'
10. IF class ='A' and Gender ='M' and District ='CITY' and age = middle then credit='NO'

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

3.4. Evaluation matrix

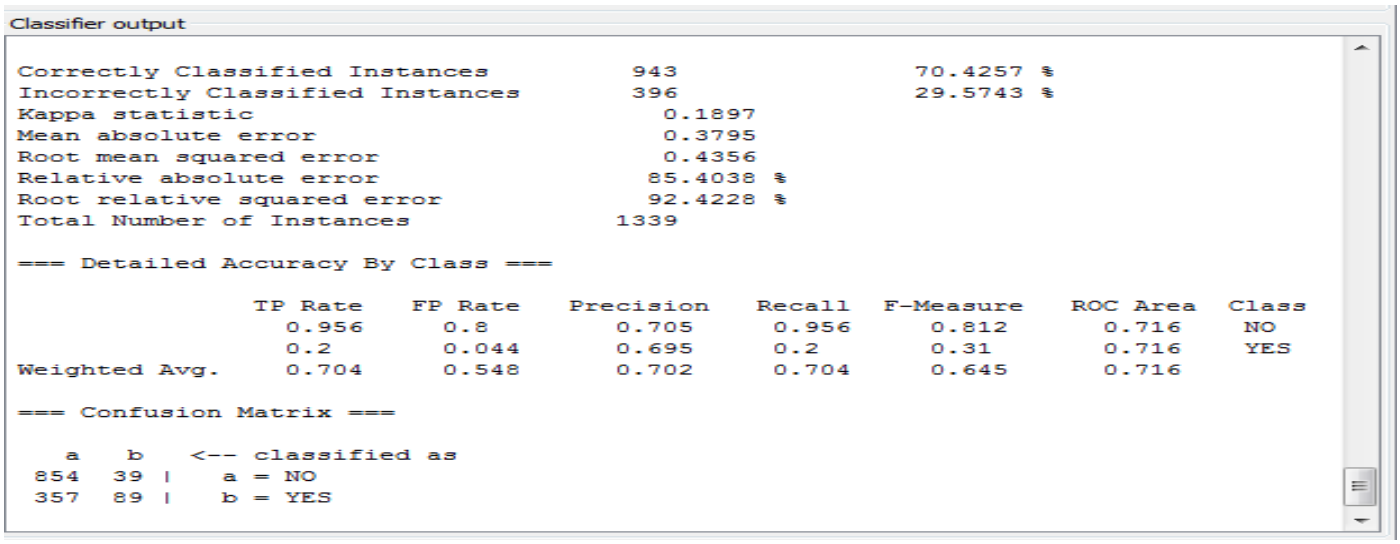


Figure 8:ID3 classifier output

The figure 8 the classifier output from Classification techniques with ID3 Algorithm by using full training .The result shows that number of attributes, the summary of evaluation training set ,detailed accuracy by class and the confusion matrices' through the above figure we have correct classified instances of 943(70.42%) and incorrect classified instance of 396(29.57%) both numbers are useful in calculation of accuracy and sub elements of accuracy

Accuracy of ID3= (Total number of corrected prediction)/Total number ofInstances =943/1339=**70.4%**

$$\text{Error rate}=\frac{FP+FN}{P+N}=\frac{89+39}{893+446}=0.095$$

$$\text{Sensitivity(recall)}==\frac{TP}{FN+TP}=\frac{372}{461}=0.806$$

$$\text{Specificity (true negative)}=\frac{TN}{TN+FP}=\frac{804}{877}=0.916$$

Precision

$$\text{Precision}=\frac{TP}{TP+FP}=\frac{372}{372+73}=\frac{372}{445}=0.835$$

Score

$$\text{Score}=\frac{2\ (precisionxrecall}{precisionxrecall}=\frac{2\ x0.835x0.806}{0.835+0.806}=0.820$$

We can use the same formula on J48 and CART and calculate sub elements of Accuracy in all

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

classifiers after calculating them we find the value of Table3.

Table 4 :Classifiers and their evaluation matrix

Decision tree algorithm	Accuracy	Error rate	Sensitivity	Specificity	Precision	F-score
ID3	70.4	0.120	0.806	0.916	0.835	0.820
J48	69.5	0.059	0.937	0.941	0.861	0.897
CART	68.8	0.0522	0.937	0.940	0.899	0.960

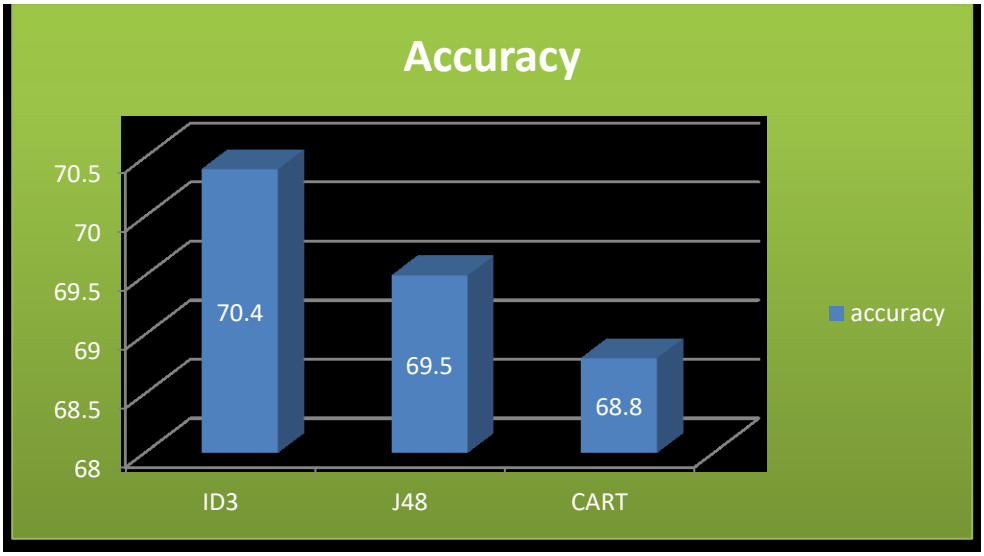


Figure9:ID3, J48 and CART and their accuracy

The ID3 is the best model that we used in this study because it has higher accuracy of 70.4% than the other algorithm. The second algorithm is J48 with 69.5% and the last one is CART. After analyzing the above algorithms we find that the best algorithm that we use to build our model is an ID3 because the model classifies 1339 instances correctly are 943 with an accurate rate of 70.42.3 % and incorrect instances are 396 with 29.57% indicates that our model will accurately predict future unknown values

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

Save and Building model

Classification is used to find a model that segregates data into predefined classes. Classification is based on the features present in the data. The result is a description of the present data and a better understanding of each class in the database. This classification provides a model for describing future data. Prediction helps users make a decision. Predictive modeling for knowledge discovery in databases predicts unknown or future values of some attributes of interest based on the values of other attributes in a database

Table 5: Testing Model

NO	GENDER	AGE	DISTRICT	CLASS	SESS	SPO	CREDIT
1	F	Middle	RURAL	B	E	PS	?
2	F	Young	RURAL	B	D	PS	?
3	F	Young	RURAL	A	D	PS	?
4	F	Old	RURAL	A	E	PS	?
5	M	Old	RURAL	B	E	PS	?
6	F	Middle	RURAL	C	D	FARG	?
7	F	Middle	RURAL	A	E	PS	?
8	F	Middle	RURAL	B	W	PS	?
9	M	Middle	RURAL	C	D	FARG	?
10	M	Middle	RURAL	C	W	PS	?
11	F	Middle	RURAL	A	E	PS	?
12	M	Middle	RURAL	B	W	PS	?
13	M	Middle	RURAL	A	D	PS	?
14	M	Young	CITY	B	W	PS	?
15	M	Young	RURAL	B	W	PS	?
16	F	Young	CITY	B	W	PS	?
17	M	Old	CITY	D	W	PS	?
18	M	Old	CITY	A	D	PS	?
19	M	Young	CITY	A	D	PS	?
20	M	Old	RURAL	B	W	PS	?

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

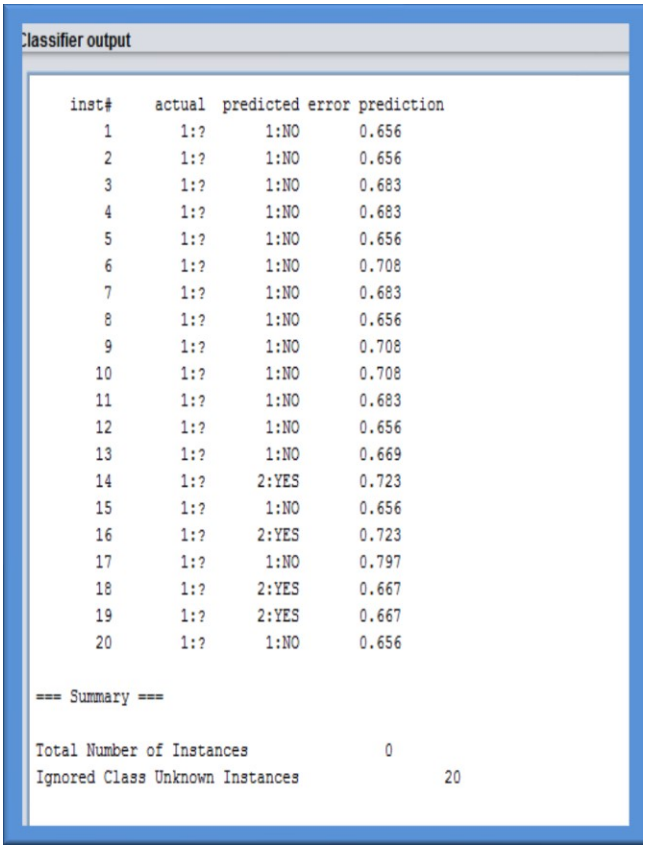


Figure 10: Testing Output

In the testing dataset, 16 instances have NO credit predicted and 4 instances predicted YES credit the figure 10 model test the new instance correctly. It is applicable for predicting student performance from different year.

Automated student's performance

After analyzing data by using WEKA, we developed an automated student performance predictor that enables user to predict individual student based on the Gender, district, class, sponsorship, session and age and predict if a new student success or fail.

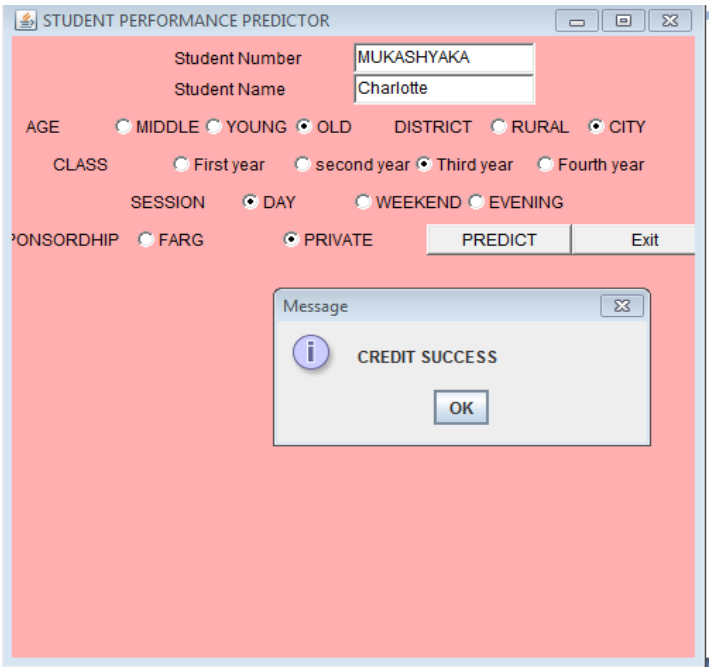


Figure 11: Automated student performance predictor

Conclusion

This study examines the factors associated with a student's performance that will enable the UNILAK to improve their learning process based on the result from this project. We used data which collected from the UNILAK management system in the last academic year 2015-2016.We applied data mining techniques to discover knowledge about student performance. ID3 was the best algorithm with the accuracy of 70.4% compared to J48 with accuracy 68.5 % and CART with 68.5.This model is very important to any academic institution for the prediction of students' performance. This model will also help academic services to put effort into the classes of first years and the second one because the students of these classes are

http://eajournal.unilak.ac.rw/ EAJST (Online Version) ISSN: 2227-1902 Email: eajst_editor@unilak.ac.rw / eajscience@gmail.com

weaker performing than the other classes. The new students need the advice that enables them to get care about the learning materials. The students from cities also need the We recommended UNILAK to apply this model in order to improve the rate of success for their students which lead them to have students with the highest quality in education. We recommended also the future researcher to apply data mining techniques on an expand the data set with most distinctive attributes and using courses. Dataset from different academic year to get more accurate results. As implementation in more than one institution. Experiments could be done using more data mining techniques such as neural nets, genetic algorithms, k-nearest Neighbor, and others.

Acknowledgement

To the Almighty God for the gift of life so as to get this far I have come.

My sincere acknowledgment addresses to the authorities of University of lay Adventist of Kigali (UNILAK), more especially to the Faculty of Computing and information sciences, though I acquired multitudes of knowledge having permitted me to carry out this project

Bibliography

Ahmed Mueen, B. Z. (2016), Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *I.J. Modern Education and Computer Science*, 11, 36-42.

Amjad Abu Saa. (2016). Educational Data Mining & Students' Performance Prediction. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 2012-220.

B.Namratha. (2016). Educational Data Mining – Applications and Techniques. *International Journal of Latest Trends in Engineering and Technology (IJLTET)* , 484.

Eshwari Girish Kulkarni,Raj B. Kulkarni, PhD. (2016). WEKA Powerful Tool in Data Mining. *International Journal of Computer Applications (0975 – 8887)National Seminar on Recent Trends in Data Mining (RTDM 2016)*, 10-15

Fadhilah Ahmad*, N. H. (2015). The Prediction of Students' Academic Performance. *Applied Mathematical Sciences, Vol. 9, 2015, no. 129-13.*

HashmiaHamsa at all. (2016). Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm. *Colloquium on*

Recent Advancements and Effectual Researches in Engineering, Science and Technology (RAEREST2016), 326-332.

Jiawei Han, M. K. (2012). Data Mining Concepts and Techniques, Third Edition. Waltham in USA: Morgan Kaufmann

Lior Rokach, O. M. (2015). *DATA MINING WITH DECISION TREES Theory and Applications, 2nd Edition.* Singapore 596224: World Scientific Publishing Co. Pte. Ltd

Megha Gupta, Naveen Aggarwal. (2010). classification techniques analysis. *National Conference on Computational Instrumentation*, 19-20.

Mohammed M abuteir and Alaa Mustafa EL-Halees. (2012). Mining Educational Data to Improve Students' performance. *International Journal of Information and Communication Technology Research*, 142-154

Tina R. Patil, M. .. (2013). Performance analysis of Naive Bayes.*International journal of computer sciences and application,* 253-259.