Global Scientific JOURNALS

GSJ: Volume 7, Issue 12, December 2019, Online: ISSN 2320-9186
www.globalscientificjournal.com

MODELING STUDENT'S DROPOUT PREDICTOR IN KAVUMU TVET SCHOOL BY USING DECISION TREE

MAHEREZO Joseph[1], TUYISHIMIRE Jean Damascène[2], Dr. NIYIGENA Papias[3]

Department of Information Technology

Faculty of computing and information sciences

University of Lay Adventists of Kigali

Rwanda

ABSTRACT

Data mining supports to excavate the original and the valued data from the huge amount of dataset. In reality, in Rwanda students are dropping out of school in the different levels of education. Predicting the student's dropout is unique of the solutions for reducing the rate of the students dropping out. This can be accomplished by using the historical data of those specific students stored in an information management system through data mining. Such data can be extracted from the different stakeholders of the education sector, such as nursery schools, primary schools, secondary schools, universities, and other higher learning institutions(HLI) as well as government databases such as the ministry of education. The purpose of this paper is to design a model that enables the HLI to predict the student's dropout and to analyze the causative factors that lead to the dropout of such students. In this situation, the classification method of data mining will be useful to predict useful information through the historical information by data mining tools. This research analyzed data by using three decision tree algorithms; Iterative Dichotomiser3 (ID3), J48 and classification and regression (CART) decision tree.

Keywords: Modeling, Student dropout, predictor, data mining techniques

1. INTRODUCTION

The students are not finishing their studies is the big challenges in the society. The goal of our country is to have the people with the specific level of the study. The local government minister of Rwanda avoids the attitude of the students who drop out the school from nursery till to University schools and the different people inspire the students who drop out to re-register in order to cover all the modules or courses in their program. In line with aspirations to become a knowledge-based society, Rwanda prioritized access to education partly by offering a fee-free and compulsory primary education to all children. And considering that the cost of education can be relatively high for poor families, the government reduced such problems by ensuring that at least twelve basic years of education are free. However, recent reports claim that school dropout rates in Rwanda remain high despite measures undertaken by all relevant parties to combat the problem which is progressively becoming chronic. The dropout rates in schools are tremendously high at 10.8 to 11.9 percentage between boys and 9.6 to 10.4 percentage for girls respectively. The main purpose of this study is to forecasting student Drop out as well as identifying the factors causing it. [1]. EDM refers to techniques and tools designed for automatically mining meaning from large repositories of data created by peoples learning activities in educational settings at a high level, the field seeks to develop and improve methods for exploring this data [2].HLI needs the tools that allow them to analyze the dropout of the student in different schools. ID3, J48, and CART are three types of decision tree algorithm used in this study. WEKA is intelligent learning tools that allow us to analyze data and to build a student's dropout model.

2. Related works (Literature review)

Literature in educational data mining is on the rise in the publishing sector. A number of scholars have published on this subject, though most of their works have not been so conclusive on the matter.

Alaa el-Halees [3], Conducted a study on how data mining is valuable to ameliorate the performance of the students in higher learning education. In this study association rule and classification rule by utilizing decision trees were highly used for data analysis.

Bharadwaj and Pal [4], using the decision tree method for classification to evaluate the performance of students. The objective of this study is to see the knowledge that defines students' performance at the completion of the semester examination. This study was reasonably useful for recognizing the dropout's student in earlier stage and students who want special attention and allow the teacher to provide appropriate advising. In this study, Bharadwaj and Pal[6] conducted a study on learner performance based on choosing 300 students from 5 dissimilar degree university conducting BCA (Bachelor of Computer Application) course of Dr. R. M. L. Awadh University, Faizabad, India. By using Bayesian classification

methods on 17 elements, it was really found that the reasons like grades students in secondary exam, location of living, means of teaching, qualification of mother, students other habits, family annual income, and student's family status were highly correlated with the student academic performance.

Boero, Laureti & Naylor [7], they found that gender (males have a higher probability of dropping out relative to the reference group of females) is one of the principal determinants of the chance of dropping out and age has a major positive effect. the new one will predict student dropout based on units of Kavumu TVET School where students acquiring the knowledge and skills in the academic year 2019 and analyze different data from each units then after the model will be designed based on data collected with the purpose of predicting student's dropout using decision tree in order to overcome problems encountered in previous works as seriously mentioned above. The institution will benefit from it in knowing the main causes why the students in different units drop out each year and also will know how to predict it before occurring.

3. Proposed Method

The methodology begins from the problem definition, data gathering, then preprocessing and the data set and preprocessing units, and 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 [8]. Often known as KDD, data mining refers to excavating knowledge from an immense amount of data. Unseen patterns and relationships helpful for decision making are usually discovered using data mining techniques on massive amounts of data. Sometimes, scholars refer to data mining and KDD as synonyms [9].

The main steps in this method are:

1. Data gathering: Data may be gotten from several unlike and various data sources. This step comprises gathering all available information on students. The set of factors that can affect the students' drop out is first identified and collected from various sources of data available. This is then integrated into a single data set.

2. Data pre-processing: At this phase, the preparation of data set to apply the data mining techniques is done. Traditional pre-processing techniques like data cleaning, data partitioning and data transformation of variables have to be applied. Due to problems of great dimensionality and imbalanced data, here we have also applied the selection of attributes and re-balancing of data[11].

3. Data mining: DM algorithms are applied to examine the factors affecting dropout like a classification problem where a model is constructed. We propose the use of various classification algorithms and techniques that easily produces interpretable models like decision trees and induction rules. Finally, these algorithms have been executed, studied, evaluated and compared in order to determine which one obtains the best outcome with high accuracy[11].

4. Interpretation: The obtained models are analyzed to detect the problem of drop out in this stage. To accomplish this, we interpret the causes that contribute more to the problem and how they are related are considered and assessed [12].

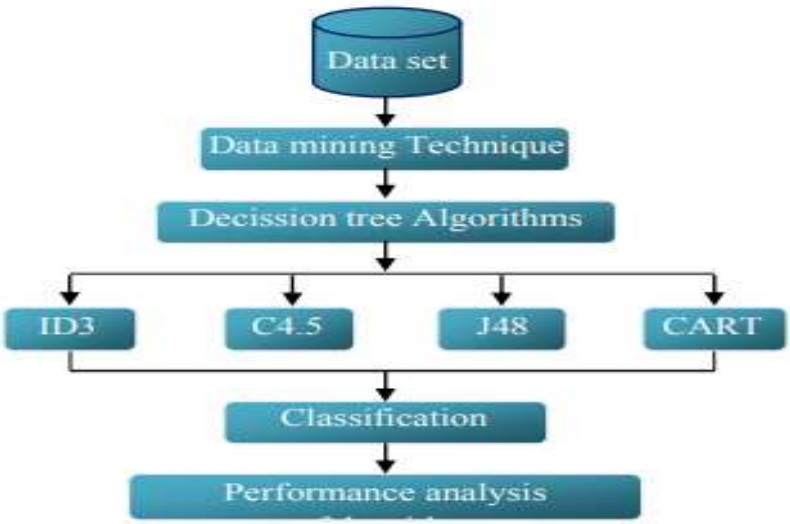

Figure: Steps in Data Mining

4. Presentation of results and discussion

This part contains Dataset, data presentation, decision tree, produced rules and classifier comparison, id3 model and student dropout predictor.Success rate of any TVET School institute can be analyzed by tracking the Main reasons for dropout students. In this research, student information on diverse parameters was gathered through machine learning repository by Forecasting the students' dropout status whether they can carry on their study or not, it needs lots of considerations such as personal gender, district, sponsorship and tuition fees, etc. variables are crucial information for the effective prediction of attributes. Since the current study is in relation to classify & regressing the various quantitative and qualitative factors to identify the causes of dropout from the viewpoint of knowledge discovery and data mining.

4.1 Preparation of data

The dataset used for this study was prepared from the machine learning repository. The data attributes include personal gender, district, and sponsorship and tuition fees) and it was collected by using Kavumu TVET school reports of 2019 and real questionnaires. One dataset is provided regarding student dropout in the academic year 2019 in different sections in the school. After the collection of various data, the

dataset was prepared to apply the data mining techniques. Before the application of the arranged model, data preprocessing was applied to measure the quality and suitability of data. In this stage, only those elements were selected which were needed for data mining. For this, eliminate missing values, smooth out noisy data, selection of related attributes from database or removing irrelevant attributes, identifying or remove outlier values from data set, and resolving inconsistencies of data. Some of the irrelevant parameters were removed from database such as Names Certified, mother tongue, the gender field containing two values female and male because KAVUMU TVET School has both male and female. Data regarding the student registration have been stored in a Microsoft Excel spreadsheets file, and saved in CSV format which is suitable by WEKA tools. The table below describes the attributes of a dataset.

Table of variables

Attribute	**Description**	**Possible values**
Age	Age of student	Middle, young, and old
Sex	Sex of student	Male and female
Dropout	Dropout status	Yes or No
District	Resident district of student	Rural and city
Sponsorship	Sponsorship type	Parent, FARG, NGOS

Table: variables used.

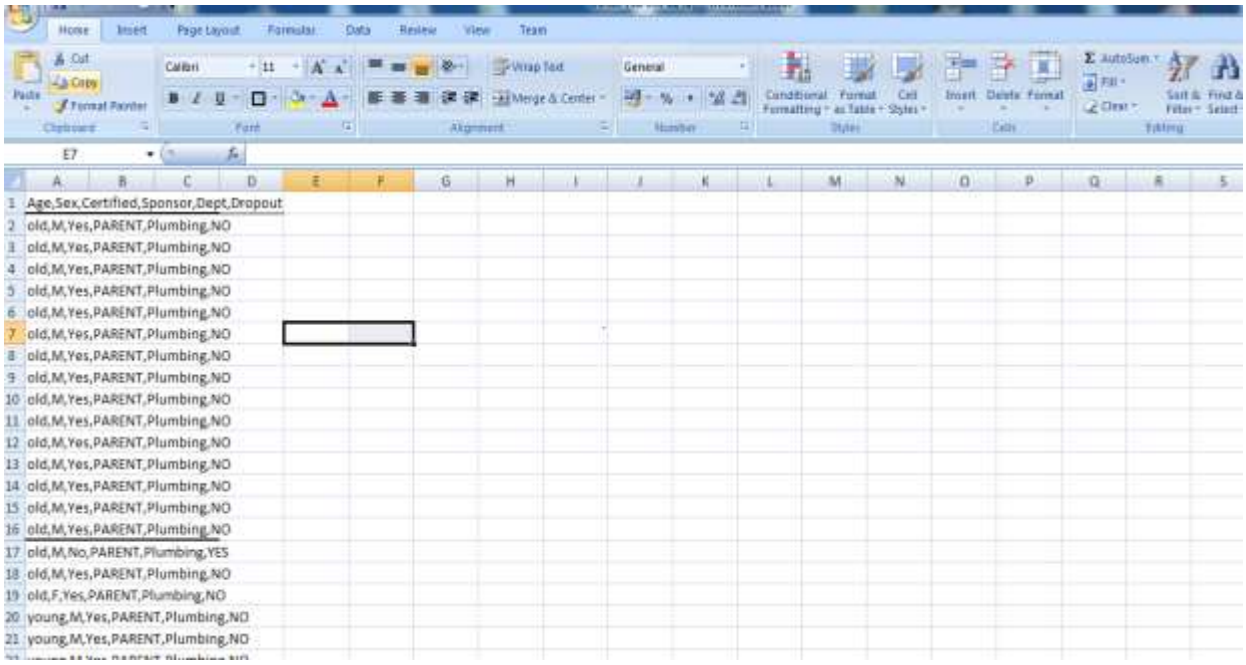

Figure: Dataset in MS Excel

A dataset from the database of Kavumu TVET school are collected into an excel sheet and some information are really combined together.

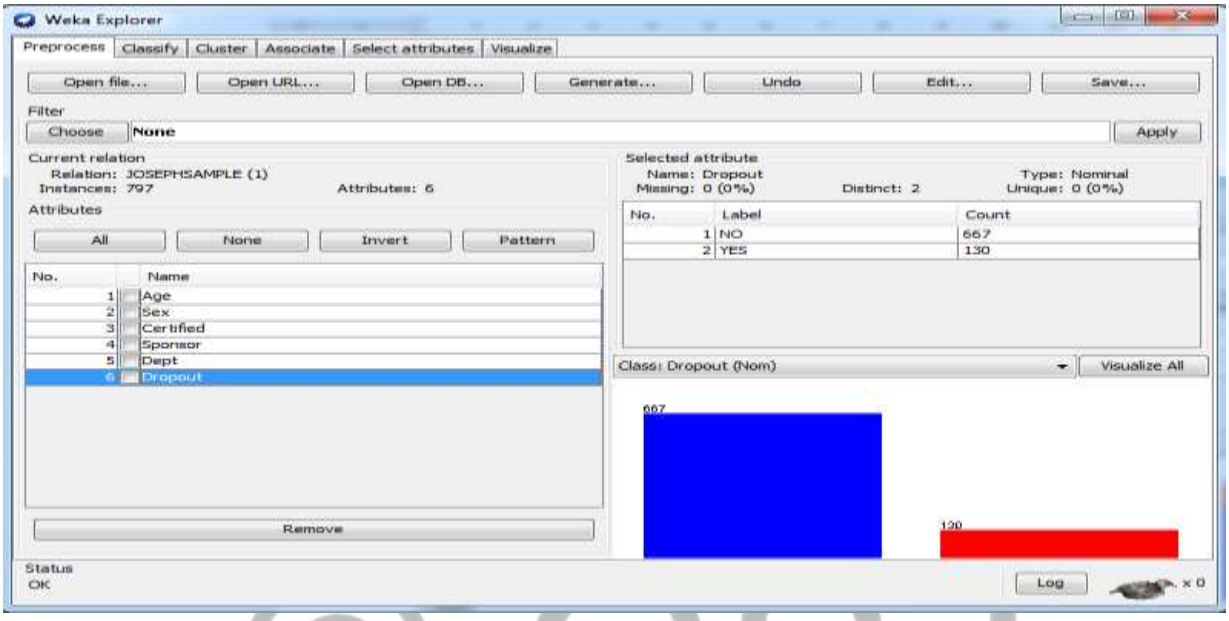

Figure: data preparation in weka

After changing MS excel to the CSV format, data was extracted from CSV into WEKA tool. The above-shown figure the attributes of the dataset and the chart that shows that the Selected attribute as drop out attribute as observed on the figure. The red color indicates the number of students who drop out while blue color indicates the number of students who are not dropped out in academic year 2019

In the above-shown dataset, there is 797 students taken from Kavumu TVET departments, the percentage of female is 1, 88.8% and the percentage of male is 98, 11 % of male in dataset. This meant that 667 out of 797 equivalent to 83,68 percentage got their certificates means they completed well their studies contrary to this 130 students out of 797 equivalent to 16.31 percentage dropped out in learning this meant that those students did not finish their studies in academic year 2019.

4.2 Comparison of classifier Performance

Decision tree algorithm	Accuracy	Error rate
ID3	80%	0.120
CART	79,5%	0.059
J48	78.8%	0.0522

Figure:

The ID3 is the best model used in this work due to higher accuracy of 80% than other algorithms. The second algorithm is J48 with 78.8 % and the third one is the CART. After studying the above-mentioned algorithms, the result shows that the best algorithm that can use to build the model is an ID3 because the model can classify 797 instances correctly are 782 with an instances are 15 with 1,11% shows that the model can accurately predict upcoming unidentified values.

Factors leading students to drop out in Kavumu TVET School

No	Reasons or factors	Percentages (%)
1.	Family issues	**9.31%**
2.	Human sickness	**4%**
3.	Change of objectives/goals	**1%**
4.	School environment	**1%**
5.	High schools fees	**1%**

In reality few students likely to drop out due to, homesickness, illness, peer problems, high school fees, and adjustment problems, etc.

Several of the association rule mining technique was used to discover the relationship between unrelated variables in a large database of KAVUMU TVET School.

5. CONCLUSION

The main purpose of the study was to examine the key factors causing the dropout of students in Kavumu TVET School and designing a model to predict students drop out. Based on the review of the related literature done by previous researchers. The study was taking797 students as samples from different department of Kavumu TVET School. Data were well collected from the Kavumu management system and analyzed using Data Mining techniques.

The result indicated that ID3 is the best algorithm with an accuracy of 80 % compared to J48 with accuracy 79.5 % and CART with 78.5. This model is important to Kavumu TVET School

For the prediction of students' Drop out. This model will help the institution to prevent students' drop out before occurring based on factors leading students to drop out in the school. With this model, the culture of dropping out in school will be perfectly eradicated.

REFERENCES

1. [1] Herzog, S. Measuring determinants of student return vs. dropout/stop out vs. transfer:
A first-to-second year analysis of new freshmen. In Proc. of 44th Annual Forum of the
Association for Institutional Research (AIR), 2004.
2. [2,3]Eshwari Girish Kulkarni, Raj B. Kulkarni, Ph.D. (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
3. [4,6]Alaa E-Halees, Ghadeer S. Abu-Oda, "Data Mining in Higher Education: University Student **Dropout Case Study**" International Journal of Data Mining & Knowledge Management Brijesh Kumar Baradwaj and Saurabh Pal "Mining Educational Data to Analyze Students" Performance" International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011
4. [7]Boero, Gianna et.al "An econometric analysis of student withdrawal and progression in postreform Italian Universities" 2005.
5. [8,5], S.Nagaparameshwara chary, Dr.B.Rama," A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining", International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X), Volume.3,Special Issue.1,March.2017
6. [9]Romero, Cristobal and Ventura, Sebastian. Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1):12–27, 2013.
7. [10] Anjana Pradeep, Jeena Thomas," Predicting College Students Dropout using EDM Techniques", International Journal of Computer Applications (0975 – 8887) Volume 123 – No.5, August 2015
8. [11]Anjana Pradeep, Jeena Thomas," Predicting College Students Dropout using EDM Techniques", International Journal of Computer Applications (0975 – 8887) Volume 123 – No.5, August 2015
9. [12]Luan, J. Data mining and its applications in higher education. New Directions ForInstitutional Research, p. 17-36, Spring 2002.

Authors:

1. **MAHEREZO Joseph,** email: mahejose2020@gmail.com
   Student in Master of Science in Information Technology, University of Lay Adventists of Kigali
2. **TUYISHIMIRE Jean Damascène**, email: tjeandamascne@gmail.com

Student in Master of Science in Information Technology, University of Lay Adventists of Kigali

Correspondence author:

1. **Dr. NIYIGENA Papias**, email: papiasni@yahoo.fr

Lecturer in Master of Science in Information Technology, University of Lay Adventists of Kigali