The Use of Classification Algorithm for Students Academic Performance Forecasting

Waidor. K. Tamaramiebi & Akpojaro Jackson
Department of Mathematical Sciences,
University of Africa
Toru – Orua, Bayelsa State
E-mail: zalimaxxx@gmail.com, tamara.waidor@uat.edu.ng; jacksonakpojaro@gmail.com

ABSTRACT

During the last few years, the application of Data Mining has grown exponentially, spurred by the fact that it allows us to discover new, interesting and useful knowledge about data in almost every facet of discipline. Its application in education is also gaining a lot of attention across the globe. The purpose of the research work is an attempt to forecast student academic performance using the Classification (Decision Tree) data mining algorithm which is a learning algorithm. A learning algorithm uses previously established facts or figures to predict future occurrences. The methodology we adopted in this work is the Cross-Industry Standard Process for Data Mining (CRISP-DM) which is a cyclic approach that includes six principal phases. CRISP-DM was preferred over other approaches because it is a well established and generally accepted data mining methodology. The data set used in this experiments is the student academic data of Computer Science Department, University of Africa, Toru – Orua (UAT), Bayelsa State, Nigeria that includes their entry scores into the university and the 2017/2018 academic results. WEKA software which is a very effective data mining program is used as a tool for the data classification and analysis.

Keywords: Data Mining, Classification Algorithm, Learning Algorithm

iSTEAMS Conference Proceedings Paper Citation Format
Waidor. K. Tamaramiebi & Akpojaro Jackson (2018): The Use of Classification Algorithm for Students Academic Performance Forecasting. Proceedings of the 14th iSTEAMS International Multidisciplinary Conference, AlHikmah University, Ilorin, Nigeria, Vol. 14, Pp 153-158

1. INTRODUCTION

The University of Africa, Toru – Orua, located in the Crude Oil rich state of Bayelsa, Nigeria is a nascent citadel of higher learning with aspirations of producing some of the best students in the world. And one way this can be achieved is the through the early determination of academically 'weak' and 'strong' students. This would pave the way to pay necessary attention to the 'weak' students in time. Hence, we look up to Students' Data Mining to predict student academic performance.

Data mining also known as Knowledge Discovery from Data (KDD) (Mythili et al, 2014) is a fast growing branch of Computer science and Statistics. It is the discovery of Knowledge using patterns from a large data repository. The knowledge discovered could be used for predictive purposes or others. The world is data rich but information poor (Jiawei et al 2012). Although some authors argue that the term data mining does not really present all the major components of the iterative data mining process, stressing that the mining of gold from rocks is referred as gold mining in stead of rock mining, hence, analogously, data mining should have been more appropriately named "knowledge mining from data," (Jiawei et al, 2012).

The iterative process of Data mining according to Osmar, 1999 consists to of the following steps:

- Data cleaning/cleansing: Which is the removal of irrelevant data and noise data removed from the data collection.
- Data integration: at this stage, multiple data sources, often heterogeneous (diverse in kind), may be combined in a common source.
- Data selection: Relevant data to the analysis is decided on and retrieved from the data collection.

- Data transformation/Consolidation: Here, selected data is transformed into forms appropriate for the mining procedure.
- • Data mining: This stage is the application of clever techniques to extract patterns potentially useful.
- • Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user, using visualization techniques to help users understand and interpret the mined results.

Data Mining is applicable in different fields of endeavours such as industries, oil and gas, education etc. The application of Data mining in the educational field is known as Educational Data Mining (EDM). There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational environments . Educational Data Mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K-Nearest Neighbor, and many others. (Bridesh et al 2011).

1.1 Types of Data Mining Algorithm

Classification algorithm
Classification is one of the most frequently studied problems by Data Mining and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes, such as the predicting attributes (Cristobal et al, 2008).
Some classification methods are briefly explained below.

Decision Tree is a set of conditions organized in a hierarchical structure (Cristobal et al, 2008). It is a predictive model in which an instance is classified by following the path of satisfied conditions from the root of the tree until reaching a leaf, which will correspond to a class label. A decision tree can easily be converted to a set of classification rules. Some of the most well-known decision tree algorithms are C4.5 (J48) and CART.

Artificial Neural Networks (ANNs) are parallel computational models comprised of densely interconnected, adaptive processing units, characterized by an inherent propensity for learning from experience and also discovering new knowledge (Loannis et el 2012) . It can also be viewed as a computing paradigm that is modeled after cortical structures in the brain. It consists of interconnected processing elements called nodes or neurons that work together to produce an output function (Cristobal et al 2008).

2. RELATED WORKS

Brijesh and Saurabh, (2011) set out to justify the capabilities of data mining techniques in context of higher education by offering a data mining model for higher education system in the university. They used the Decision Tree classification method (collecting Information like Attendance, Class test, Seminar and Assignment marks) to extract knowledge that describes students' performance in end semester examination to enhance early identification of the dropouts and students who need special attention in order to give appropriate advice or counsel to help reduce fail ratio and take appropriate action for the next semester examination. Dorina K, (2012) in her paper aimed to reveal the high potential of data mining applications for university management and to boost the university enrolment campaigns so as to attract the most desirable students. The paper focused on the development of data mining models for predicting student performance, based on their personal, pre-university and university-performance characteristics.

Classification algorithms such as OneR rule learner, decision tree, neural network and Nearest Neighbour, were applied on the dataset. It was observed that OneR Rule Learner, Decision Tree, Neural Network and K-Nearest Neighbour (k-NN), had classification accuracy between 67.46% and 73.59%, with the highest accuracy achieved for the Neural Network model (73.59%), followed by the Decision Tree model (72.74%) and the k-NN model (70.49%). The Neural Network model predicts with higher accuracy the "Strong Student" class, while the other three models perform better for the "Weak Student" class. Cristóbal, Sebastián, Pedro and César (2008) in their paper compared different data mining methods and techniques to classify students based their respective courses using the moodle mining tool.

They developed a specific mining tool for making the configuration and execution of data mining techniques easier and also applied discretization and rebalance preprocessing techniques on the original numerical data in order to verify if better classifier models are obtained. They were able to show that some algorithms improve their classification performance when such preprocessing tasks as discretization and rebalancing data where applied, but others did not.

Qasem, Emad and Mustafa used the classification data mining processes to evaluate student data to study the main attributes that may affect the student performance in courses. This they did by building a system of generated rules which allows students to predict the final grade in a course under study. They were able to prove that attributes such as High school grades, Teacher's grade and funding etc. could affect the student academic performance.

3. METHODOLOGY

The method adopted in the paper is the Cross-Industry Standard Process for Data Mining (CRISP-DM). This is a cyclic approach that includes six principal phases – Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment, with a number of internal feedback loops between the phases, resulting from the very complex non-linear nature of the data mining process and ensuring the achievement of consistent and reliable results (Dorina, 2012).

Business Understanding
At this phase, understanding business objectives is very vital and what actions to be taken on the likely outcomes. For this research work, our major concern is the to detect early the academically weak students with the intention of rendering them the needed attention to improve their performance.

Data Understanding
At this phase, data sources and fields are identified that may have an impact on the Business objectives. In our study the critical data collected/ or fields (numerical data) are the students entry scores, the first and second semester Grade Point Average. While the others are (nominal data) students' registration numbers, department and standing – either Pass (P) or Fail (F).

Modelling
Here, we used WEKA, which is a data mining software developed by Waikato University, New Zealand. WEKA is an acronym for Waikato Environment for Knowledge Analysis. It provides a collection of data mining and machine learning algorithms which includes Classification, Clustering and Regression etc.
The modelling was done using the classification algorithm (decision tree algorithm C4.5 (J48)), reason being that classification are a popular choice for researchers and have the propensity to yield an acceptable results.

Evaluation
At this stage, data are partitioned into 2 sets: Training or Modelling Set and Test or Hold out Set (John M, 2013). Here data are analysed and decisions made based on the business objectives. In our case, we are looking at using our findings to assist the students having difficulties in their studies in order to boost their academic performance.

Deployment
At this phase, how the results gotten are utilized and who use them are considered. In our case, recommendations would be made to the Computer Science department.

4. DISCUSSION OF DATA ANALYSIS AND RESULTS OBTAINED FROM THE DECISION TREE DATA MINING ALGORITHM

Our dataset was collected in Excel format. But WEKA work mostly with dataset in Attribute Related File Format (Arff) so there are arose a need to first convert the excel format to Comma Separated Values (CSV) file format before preceding to convert the CSV file format to Arff on the WEKA platform. However this proved a little difficult for us so we found a way of directly converting the excel file format to Arff using the an open source software called Excel2ArffConverter. However to achieve this, the excel file would first have to be saved in Excel 97–2003 Workbook.

Find Below our Dataset in Attribute Related File Format (Arff). Please note that the Reg_No were altered for the sake of confidentiality.

@relation Stud_Performance

@attribute Reg_No
{uat18000,uat18001,uat18032,uat18036,uat18061,uat18071,uat18089,uat18093,uat180153,uat18010,uat18015,uat18016,uat18014,uat18015,uat18021,uat18020,uat18023,uat18232,uat18247,uat18270,uat18006,uat18015,uat18032,uat18039,uat18063,uat18073}
@attribute Dept {CSC}
@attribute Gender {M,F}
@attribute EntryScore numeric
@attribute FirstSemester_GPA numeric
@attribute SecondSemester_GPA numeric
@attribute Cummulative_GPA numeric
@attribute 'Results _Current Standing' {pass,fail}

@data
uat18000,CSC,M,400,4.48,4,4.24,pass
uat18001,CSC,M,204,1.39,1.91,1.65,fail
uat18032,CSC,M,234,1.52,2.39,1.96,fail
uat18036,CSC,F,301,3.26,2.52,2.89,pass
uat18061,CSC,M,312,3.3,3.52,3.41,pass
uat18071,CSC,M,340,3.26,3.7,3.48,pass
uat18089,CSC,M,230,1.78,1.87,1.83,fail
uat18093,CSC,F,390,3.91,3.83,3.87,pass
uat18013,CSC,M,319,2.96,2.78,2.87,pass
uat18010,CSC,M,309,2.91,2.83,2.87,pass
uat18015,CSC,M,354,3.43,2.3,2.87,pass
uat18016,CSC,M,267,2.3,1,1.65,fail
uat18014,CSC,M,329,3.04,2.61,2.83,pass
uat18015,CSC,M,222,1.83,1.26,1.55,fail
uat18021,CSC,M,262,2.09,1.43,1.76,fail
uat18020,CSC,M,345,3.52,2.61,3.07,pass
uat18023,CSC,M,250,1.35,1.09,1.22,fail
uat18232,CSC,M,319,3.3,2.39,2.85,pass
uat18247,CSC,F,359,3.39,1.13,2.26,pass
uat18270,CSC,M,212,1.91,0.57,1.24,fail
uat18006,CSC,M,326,2.26,0.83,1.55,fail
uat18015,CSC,M,310,2.35,3.78,3.07,pass
uat18032,CSC,M,239,2.04,0.87,1.46,fail
uat18039,CSC,M,216,1.78,0.87,1.33,fail
uat18063,CSC,M,299,3,0.61,1.81,fail
uat18073,CSC,M,329,2.57,2.04,2.31,pass

Table 1 depicts our findings from the Decision Table data mining algorithm having the following evaluation measures: Percentage (%) of correctly classified instances, Percentage of incorrectly classified instances, Kappa Statistic, True Positive (TP) and False Positive (FP) Rates, Precision, Recall, F-Measure and ROC Area.

Table 1: Findings from the Decision Table Data Mining Algorithm

Data Mining Algorithm	**DECISION TREE (J48)**		
	Pass	Fail	Weighted Average
Evaluation parameters			
Correctly Classified Instance	100%		
Incorrectly Classified Instance	0%		
Kappa Statistic	1		
TP Rate	1.0	1.0	1.0
FP Rate	1.0	1.0	1.0
Precision	1.0	1.0	1.0
Recall	1.0	1.0	1.0
F-Measure	1.0	1.0	1.0
ROC Area	1.0	1.0	1.0

The Decision Tree classification (J48) model gave a very high accuracy of prediction – 100%. This high accuracy we believe is due to the little quantity of data analysed which in turn was due to small amount of data repository available from our data source. Kindly note that the University of Africa, Toru – Orua is a nascent university, hence, don't yet have a large repository of data. But this would improved greatly in the coming years. This model is easily interpretable because it produces a set of easy to understand rules, and works well with both numeric variable and nominal variables (Dorina K, 2012).

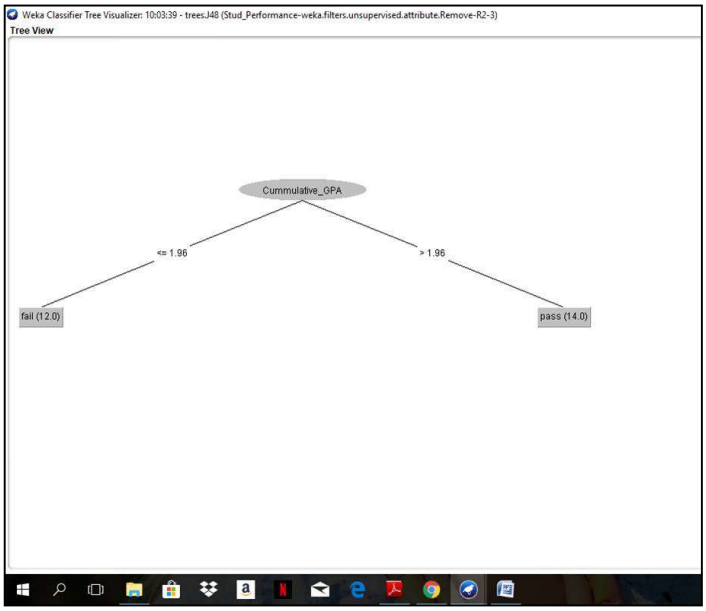

Fig. 1. Decision Tree.

Figure 1 above is a Snapshot of the Decision Tree from our analysis.

5. CONCLUSION

We got 100% prediction accuracy from the dataset we collected. Meaning we can obtain a very high accuracy of student academic performance as of now. However, in the future as more data are available from our data source, we intend looking at a different format of predicted target variable (a nominal variable with 4 distinct values – Poor, Good, Very Good and Excellence), and using more than one classifier algorithm in order to do a comparative analysis. The results we got also opened our eyes to the dimension to take in future student academic performance research in the University of Africa, Toru – Orua, which include adding more departments, adding new data, and of course, more attributes.

REFERENCES

1. Brijesh Kumar Baradwaj and Saurabh Pal (2011). *Mining Educational Data to Analyze Students" Performance.* International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6. P 63.
2. Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás (2008). *Data Mining Algorithms to classify Students.* Educational Data Mining 2008. The 1st International Conference on Educational Data Mining Montréal, Québec, Canada, June 20-21, 2008 Proceedings 1 Educational Data Mining. Pp 6-7.
3. Dorina Kabakchieva (2012): *Student Performance Prediction by using Data Mining Classification Algorithms.* International Journal of Computer Science and Management Research Vol 1 Issue 4 November 2012 ISSN 2278-733X Pp 686 – 689
4. Jiawei Han, Micheline Kamber and Jian Pei (2012). *Data Mining Concepts and Techniques.* Morgan Kaufmann Publishers. P 6.
5. John McConnell (2013). *An Introduction to CRISP-DM.* Smart Vision Europe Ltd. Pp 14-17
6. Ioannis E. Livieris, Konstantina Drakopoulou and Panagiotis Pintelas. *Predicting Students' Performance Using Artificial Neural Network.* P 2.
7. Mythili M. S. and Shanavas Mohamed A.R (2014). *An Analysis of students' performance using classification algorithms.* IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. III (Jan. 2014) p.63
8. Osmar R. Zaiane (1999). *Introduction to Data Mining.* CMPUT690 Principles of Knowledge Discovery in Databases. Pp 1-5
9. Qasem A. Al-Radaideh, Emad Al-Shawakfa and Mustafa I. Al-Najjar (2006). *Mining Student Data Using Decision Trees.* The 2006 International Arab Conference on Information Technology (ACIT'2006), Nov. 2006, Jordan. P 1.