African Scientist Vol. 20, No. 2 June 30, 2019 1595-6881/2019 $10.00 + 0.00

Printed in Nigeria

http://www.niseb.org/afs

AFS2019064/20205

The Use of Classification Algorithm for Forecasting the Academic Performance of Students of Biological Sciences, University of Africa, Toru-Orua

T.K. Waidor* and J. Akpojaro

Department of Mathematical Sciences, University of Africa, Toru-Orua, Bayelsa State, Nigeria

*Corresponding author; Email: tamara.waidor@uat.edu.ng, Tel: +2347067392967

(Received June 4, 2019; Accepted in revised form June 16, 2019)

ABSTRACT: In recent years, the application of Data Mining has grown exponentially, spurred by its ability to allows us 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. In this research, a data mining technique known as classification algorithm (Decision Tree) was used to forecast students' academic performance. The methodology 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 experiment is the student academic data of Biological Sciences, University of Africa, Toru-Orua (UAT), Bayelsa State, Nigeria.From our findings, the performance of the students was predicted with very high accuracy of 95.24% using WEKA Datamining tool.

Keywords: Data Mining, Classification Algorithm, Learning Algorithm

Introduction

The University of Africa, Toru-Orua, is located in the crude oil rich state of Bayelsa, Nigeria. It is a nascent citadel of higher learning with aspirations of producing some of the best students in the world. To achieve this, there is need to determine early the 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) (Brijesh, *et al.*, 2011; Mythili *et al*., 2014), is the discovery of knowledge using patterns from a large data repository. The knowledge discovered could be used for predictive or prescriptive purposes. Mining data helps to gain value from it, as the world is data rich but information poor (Jiawei *et al*., 2012).

Data mining which is an iterative process; according to Osmar (1999) it consists of the following steps:

- **Data cleaning/cleansing**: This 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**: This 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.

However, authors such as Jiawei *et al*. (2012) opined 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 instead of rock mining, hence, analogously, data mining should have been more appropriately named "knowledge mining from data".

Data Mining is applicable in different fields of endeavours such as Biology, Industries, Oil and Gas, Education and Agriculture etc. The application of data mining in the educational field is known as Educational Data Mining (EDM) (Bridesh *et al*., 2011). 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 Neighbour, and many others (Bridesh *et al*., 2011).

There are many types of data mining algorithms, some of which are:

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***:** This 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)***:** These 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 al*., 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).

Related Works

To justify the capabilities of data mining techniques in the context of higher education by offering a data mining model for higher education system in the university (Brijesh *et al*., 2011), 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. This was 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.

Data mining application has very high potential for university management to boost enrolment campaigns to attract the most desirable students. Dorina (2012) 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.

In comparing different data mining methods and techniques to classify students based their respective courses using the moodle mining tool, Cristóbal *et al.* (2008) 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 (such as GAP, ADLinear and Corcoran etc.) improved their classification performance when such preprocessing tasks as discretization and rebalancing data were applied, but others (such as Kernel, KNN, AprioriC and Decision Tree etc.) did not.

Qasem *et al.* (2006) used the classification data mining processes to evaluate student data to study the main attributes that may affect the student performance in their undertaken courses. This they did by building a system of generated rules which allows prediction of students' 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.

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 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).

Modeling: 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 modeling was done using the classification algorithm (decision tree algorithm C4.5 (J48)), reason being that classification area 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 Modeling Set and Test or Hold out Set. 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.

Discussion of Data Analysis and Results Obtained from the Decision Tree Data Mining Algorithm

Our dataset was collected in Excel format but because WEKA works mostly with dataset in Attribute Related File Format (Arff) and Comma Separated Values (CSV) etc., hence the dataset was converted to CSV file format before be loading it on the WEKA platform. The table below represents the Attributes of our Dataset.

Table 1: Dataset information

Run Information and Summary (Dataset for Student Result Prediction)

=== Run information ===

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     DATASET FOR STUDENT RESULT PREDICTION
Instances:    84
Attributes:   7
                 MAT_No {UAT18100…UAT18530}
                 Gender {M,F}
                 Programme {Biology, Biochemistry, Microbiology}
                 FirstSemester_GPA
                 SecondSemester_GPA
                 Cummulative_GPA
                 Result_CurrentStanding {probation, Poor, Good, V.Good and
```

```
                Excellent}
Test mode:    evaluate on training data

=== Summary ===

Correctly Classified Instances          80        95.2381 %
Incorrectly Classified Instances        4         4.7619 %
Kappa statistic                         0.9365
Mean absolute error                     0.0292
Root mean squared error                 0.1208
Relative absolute error                 11.5856 %
Root relative squared error             34.1555 %
Total Number of Instances               84
```

Table 2 represents finding from the Decision Tree 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 2: Detailed accuracy by classification

Data Mining Algorithm	**Decision Tree (J48)**						
	Probation	**Poor**	**Poor**	**Good**	**V.Good**	**Excellence**	**Weighted Average**
Evaluation Parameters							
Correctly Classified Instance	95.2381%						
Incorrectly Classified Instance	4.7619 %						
Kappa Statistic	0.9365						
TP Rate	1.000	1.000	0.000	0.897	1.000	1.000	0.952
FP Rate	1.000	0.049	0.000	0.000	0.015	1.000	0.016
Precision	1.000	0.885	?	1.000	0.944	1.000	?
Recall	1.000	1.000	0.000	0.897	1.000	1.000	0.952
F-Measure	1.000	0.939	?	0.945	0.971	1.000	?
MCC	1.000	0.917	?	0.922	0.965	1.000	?
ROC Area	1.000	0.975	0.849	0.965	0.993	1.000	0.978
PRC Area	1.000	0.885	0.038	0.948	0.944	1.000	0.928

The major concern here is the accuracy of prediction, and the Decision Tree classification (J48) model gave a high accuracy of prediction – 95.2381%. 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, 2012).

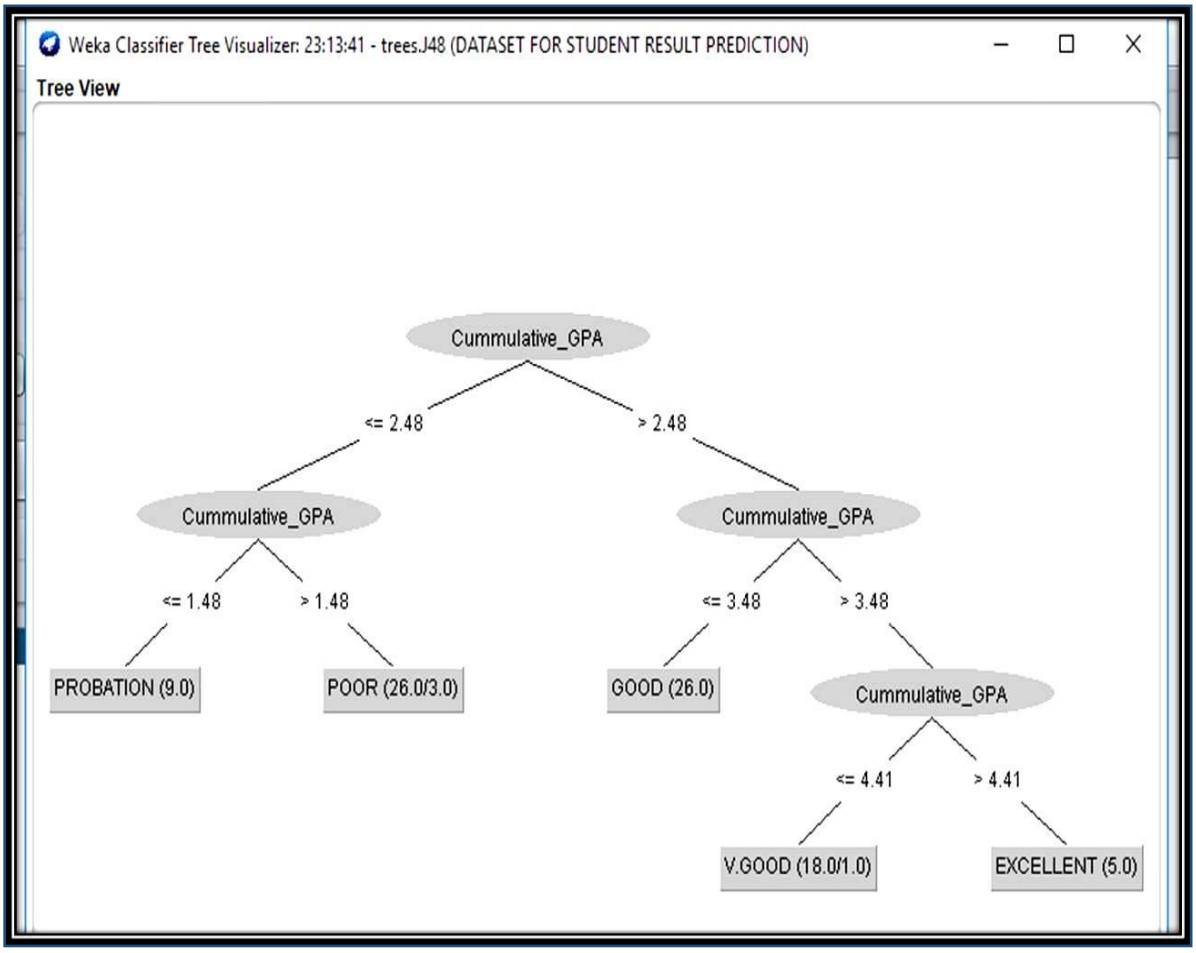

Figure 1: A snapshot of the Decision Tree from our analysis

Conclusion

A 95.2381% prediction accuracy was obtained from the dataset collected in this study, implying a very high accuracy of the Biological Sciences students' 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 and using more than one classifier algorithm to do a comparative analysis.

The results obtained can be used to advice management of the institution on student performance optimization. This research also helped to give an insight into the dimension to take in future student academic performance research in the University of Africa, Toru-Orua, which could include adding more departments, adding new data, and of course, more attributes.

References

Brijesh KB, Saurabh P: Mining educational data to analyze students‟performance. Int J Adv Compute Sci Appl 2(6): 63. 2011.

Cristóbal R, Sebastián V, Pedro GE, César H: *Data Mining Algorithms to Classify Students.* Educational Data Mining. The 1st International Conference on Educational Data Mining. Montréal, Québec, Canada. June 20-21, 2008 Proceedings 1 Educational Data Mining. pp 6-7. 2008.

Dorina K: Student performance prediction by using data mining classification algorithms. Intl J Compute Sci Manage Res 1(4): 686-689. 2012.

Jiawei H, Micheline K, Jian P: *Data Mining Concepts and Techniques.* Morgan Kaufmann Publishers. p6. 2012.

John M: *An Introduction to CRISP-DM*. Smart Vision Europe Ltd. pp 14-17. 2013.

Ioannis EL, Konstantina D, Panagiotis P: *Predicting Students' Performance Using Artificial Neural Network.* p2. 2013

Mythili MS, Shanavas MAR: An analysis of students' performance using classification algorithms. IOSR J Compute Eng 16(1): 63. 2014.

Osmar RZ: *Introduction to Data Mining.* CMPUT690. Principles of Knowledge Discovery in Databases. pp 1-5. 1999.

Qasem A, Al-Radaideh, EA, Mustafa IA: *Mining Student Data Using Decision Trees*. The 2006 International Arab Conference on Information Technology (ACIT) Jordan. p1. 2006.