Nigerian Journal of Technology (NIJOTECH)
Vol. 37, No. 4, October 2018, pp. **1087 – 1091**
Copyright© Faculty of Engineering, University of Nigeria, Nsukka,
Print ISSN: 0331-8443, Electronic ISSN: 2467-8821
www.nijotech.com
http://dx.doi.org/10.4314/njt.v37i4.31

STUDENT PERFORMANCE PREDICTION BASED ON DATA MINING CLASSIFICATION TECHNIQUES

Y. K. Saheed[1,*], T. O. Oladele[2], A. O. Akanni[3] and W. M. Ibrahim[4]
[1, 3,] DEPARTMENT OF COMPUTER SCIENCE, AL-HIKMAH UNIVERSITY, ILORIN, KWARA STATE, NIGERIA
[2,] DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF ILORIN, ILORIN, KWARA STATE, NIGERIA
[4,] DEPARTMENT OF STATISTICS, AL-HIKMAH UNIVERSITY, ILORIN, KWARA STATE, NIGERIA
E-mail addresses: [1] *yksaheed@alhikmah.edu.ng,* [2] *seyeakanni@alhikmah.edu.ng,*
[3] *wmibrahim@alhikmah.edu.ng,* [4] *oladele.to@unilorin.edu.ng*

ABSTRACT

The process of predicting student performance has become a crucial factor in academic environment and plays significant role in producing quality graduates. Several statistical and machine learning algorithms have been proposed for analyzing, predicting and classifying student performance. However, these classification algorithms still posed issue in terms of the performance classification. This paper presents a method to predict student performance using Iterative dichotomiser 3 (ID3), C4.5 and Classification and Regression tree (CART). The experiment was performed on Waikato Environment for Knowledge Analysis (Weka). The experimental results showed that an ID3 accuracy of 95.9% , specificity of 95.9%, precision of 95.9%, recall of 95.9%, f-measure of 95.9% and incorrectly classified instance of 3.83. The C4.5 gave an accuracy of 98.3%, specificity of 98.3%, precision of 98.4%, recall of 98.3%, f-measure of 98.3% and incorrectly classified instance of 1.70. The CART results showed an accuracy of 98.3%, specificity of 98.3%, precision of 98.4%, recall of 98.3%, f-measure of 98.3% and incorrectly classified instance of 1.70. The time taken to build the model of ID3 is 0.05 seconds, C4.5 is 0.03 seconds and CART of 0.58 seconds. Experimental results revealed that C4.5 outperforms other classifiers and requires reasonable amount of time to build the model.

Keywords*: Student performance, ID3, C4.5, CART, classification, Education data mining.*

1. INTRODUCTION

Universities generate large volumes of data with reference to their students in traditional paper and electronic form. In recent time, the advances in the data mining field make it possible to mine these Universities educational data and generate information that provide and assist teachers and students in decision making.

According to [1], Educational Data Mining (EDM) is concerns with developing and modelling methods that discover knowledge from data originating from educational environment. The performance of students in Universities should be a great concern to Government and parents and not only to the administrators, academicians and educators. Academic feat and achievement is one of the primary factors considered by the employer in recruiting fresh graduates. Thus, students have to place the greatest effort in their study to obtain a good grade in order to fulfil the employer's demand.

In addition, every higher institution must have a database where all relevant data to the academic activities are kept and managed. A students' result repository is part of an educational database. It is a large data bank which stores students' raw scores and grades in different courses [2].

In the past decade much research efforts has focused on predicting students' performance [3] in using data mining approach in higher education institutions to improving learning, going from investigating students' enrolment data to prevent drop-off and improve retention of students [4–5], to predict student retention at an early stage from Portfolios features [6] to analyzing the usage of learning materials uploaded in a E-Learning platform [7] or analyzing mistakes that students make together in a tutoring system [8]. The handbook of educational data mining [9] describes a simplified overview of representative works in the educational data mining area.

* **Corresponding author, tel: +234 – 814 – 268 – 3364**

Several machine learning algorithms have been reported in the literature for analyzing, predicting and classifying student performance. However, these classification algorithms still posed issue in terms of the classifier accuracy of student's performance.

The objective of this paper is to propose data mining technique to predict student performance based on ID3, C4.5 and CART algorithms.

This paper is organized as follows. Section 2 presents the related work. Section 3 entails the methodology and section 4 presents the results and discussion. The paper is concluded in section 5.

2. RELATED WORK

The researchers in [10], proposed Association rule mining to extracts useful information from a large set of data. Likewise, this technique is applied to student's data, their techniques mentioned above are used for matching the organization with the students. This process is very demanding and involves a number of steps. The authors in [3], presented a case study on predicting performance of students. The data of four academic cohorts comprising 347 undergraduate students have been mined with different classifier. The results obtained showed a reasonable accuracy. Recently, the authors in [1], investigated a method based on Decision tree algorithms BfTree, J48 and CART to predict student performance. The results of their study showed that BFtree is the best algorithm for classification with correctly classified instance of 67.07% and incorrectly classified instance of 32.93%.This study focused on three decision tree algorithms ID3, C4.5 and CART for predicting student performance.

3. MATERIALS AND METHOD

3.1 Methodology

The proposed methodology used in this study for predicting student academic performance is based on decision tree algorithms which belong to the process of Data Mining. The stages in the process include the following:

3.2 Data Mining Process

In present day's educational system, a student's performance is determined by the internal assessment and end semester examination. The internal assessment is carried out by considering the parents qualification, living location, economical status, friends and family support, resources and the attendance of the student in lecture room. The end semester examination is the mark obtained by the student at the end of semester examination. Each student has to get minimum marks to pass a semester course from both internal and end semester examination.

3.3 Data Preparation and Summarization

The data set used in this research was obtained from private University in Northern part of Nigeria, on the sampling method of Faculty of Natural Science and Department of Computer Sciences for year 2013 and year 2014 respectively. Initially size of the data is 234. In this step, data stored in different tables was joined in a single table, after the joining process, errors were removed.

3.4 Data Selection and Transformation

This stage involves dataset preparation before applying DM techniques. At this stage, traditional pre-processing methods such as data cleaning, transformation of variables and data partitioning were applied. Also, other techniques such as attributes selection and re-balancing of data were employed in order to solve the problems of high dimensionality and imbalanced data that may be present in the dataset.

3.5 The Data Mining Tools

The experimental tool used was Waikato Environment for Knowledge Analysis (WEKA). WEKA is one of the popular suites of machine learning software developed at the University of Waikato. It is open source software available under the Not Unix General Public License (GNU). The WEKA work bench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality.

3.6 Decision Tree Algorithms

Decision trees are some of the most popular machine learning algorithms used in industry and they are very useful just in general. Decision trees are interpretable, intuitive, mimic the way people like to reason and also powerful non-linear models. Decision tree is a flow-chart tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more children node and the internal nodes contain splits, which test the value of an expression of the attributes.

Table 1: Student Qualitative Data and Its Variables

Variables	Description	Possible Values
A	Age	{Below 18, Above 18}
MS	Marital Status	{Married, Single}
R	Religion	{Muslim, Christian}
S	Sex	{Male, Female}
N	Nationality	{Nigerian, Foreigner}
GENO	Genotype	{ AA, AS, SS}
FOCC	Fathers occupation	{Civil Servant, Trader, Business Man, Teacher, Pilot}
MOCC	Mothers Occupation	{Civil servant, Trader, Business Woman, Teacher, House wife}
CAFA	Course Applied for Admission	B.Sc
CAF	Course Admitted for	Computer
L	Level	100 and 200 Level
MOE	Mode of entry	UTME, IJMB, TRANSFER, DE
YOE	Year of entry	2012, 2013

3.7 ID3

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan in 1986. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node, in order to select the attribute which is most useful for classifying a given sets. A statistical property called information gain is defined to measure the worth of the attribute. From the result of the calculations, the attribute ParQua is used to expand the tree. Then delete the attribute ParQua of the samples in these sub-nodes and compute the Entropy and the Information Gain to expand the tree using the attribute with the highest gain value. Repeat this process until the Entropy of the node equals null. At that moment, the node cannot be expanded anymore because the samples in this node belong to the same class.

3.8 C4.5

C4.5 algorithm can also be called J48 and it is a successor of ID3 that uses gain ratio as splitting criterion to partition the data set. The algorithm applies a kind of normalization to information gain using a “split information” value.C4.5 decision tree algorithm can be designed when we take the original samples as the root of the decision tree. As the result of the calculation, the attribute ParQua is used to expand the tree. Then delete the attribute ParQua of the samples in these sub-nodes and compute split information to split the tree using the attribute with highest gain ratio value. This process continues on until all data are classified perfectly or run out of attributes. Repeat this process until the Entropy of the node equals null. At that moment, the node cannot be expanded anymore because the samples in this node belong to the same class.

3.9 CART

CART stands for Classification and Regression Trees introduced by Brieman. It is also based on Hunt’s algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tree. Unlike ID3 and C4.5 algorithms, CART produces binary splits. Gini Index measure does not use probabilistic assumptions like ID3, C4.5. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

A CART decision tree can be designed when the Gini Index and information gain is calculated for all the nodes. As the result of the calculation, the attribute ParQua is used to expand the tree. Then delete the attribute ParQua of the samples in these sub-nodes and compute the Gini Index and the Information Gain to expand the tree using the attribute with highest gain value. Repeat this process until the Entropy of the node equals null. At that moment, the node cannot be expanded anymore because the samples in this node belong to the same class.

4. RESULTS AND DISCUSSION

The study main objective is to predict the academic performance for the year 2013 and 2014. Several different algorithms are applied for building the classification model, each of them using different classification techniques. The WEKA Explorer application is used at this stage. The classify panel enable us to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, or the model itself. The classification algorithms used for this study are ID3, C4.5 and CART. Under the "Test options", the 10-fold cross-validation is selected as our evaluation approach.

Table 2: Performance of the classifiers

Algorithms	Accuracy	Specificity	Precision	Recall	F measure	Incorrectly classified instance
ID3	95.9	95.9	95.9	95.9	95.9	3.83
C4.5	98.3	98.3	98.4	98.3	98.3	1.70
Simple CART	98.3	98.3	98.4	98.3	98.3	1.70

Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. The model is generated in the form of decision tree.

4.1. Results for Classification Algorithms

The Table 1 shows the accuracy of ID3, C4.5 and CART algorithms for the classification using 10-fold cross validation

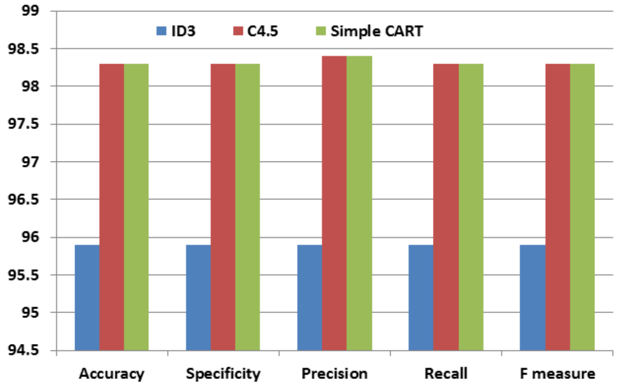

Figure 1: Comparison of ID3, C4.5 and CART classifiers

Table 3:Time taken to build the performance of the model

Algorithms	Time taken
ID3	0.05sec
C4.5	0.03sec
CART	0.58sec

From Table 1 and Table 2, it was obtained that the accuracy of the ID3 decision tree algorithm used gave a prediction accuracy of 95.9%, incorrectly classification instance of 3.83%, recall of 95.9%, specificity of 95.9%, precision is 95.9%, F-measure is 95.9%, and the time taken 0.05sec sequentially. The rank and information gain among attribute, were the course applied for, which is the highest ranked and the information gain was 234.

The C4.5 decision tree algorithm used gave a prediction accuracy of 98.3%, incorrectly classification instance of 1.70%, recall of 98.3%, specificity of 98.3%, precision of 98.4%, F-measure of 98.3%, and the time taken 0.03sec sequentially. The rank and information gain among attribute, were the course applied for, which is the highest ranked and the information gain was 234.

The Simple CART tree algorithm used gave a prediction accuracy of 98.3%, incorrectly classification instance of 1.70%, recall of 98.3%, specificity of 98.3%, precision of 98.4%, F-measure of 98.3%, and the time taken 0.58sec sequentially. The rank and information gain among attribute, were the course applied for, which is the highest ranked and the information gain was 234.

During analysis of academic students performance, many attribute have been examined and some of them are found effective which includes Educational Factors Course Applied For, Satisfaction Level, Socio-Demographic Factors (Age) and Parental factors (Parents and Friends).

5. CONCLUSION

This paper presents the prediction of student academic performance in a private University in Northern part of Nigeria. EDM has been a research hot spot to academicians and decision makers. The decision tree algorithms ID3, C4.5 and CART were used in this experiment. Results obtained showed that C4.5 performed better than other algorithms. This research would help educational stakeholder in making significant decision. Future work would be to investigate other machine learning algorithms to predict the student performance and also extending the coverage of the dataset used in this paper.

6. REFERENCES

[1] Abdulsalam, S. O., Saheed, Y. K., Hambali, M. A., Salau-Ibrahim, T. T., and Akinbowale, N. B., “Students’ Performance Analysis Using Decision Tree Algorithms”, *Anale. Seria Informatică.* Vol. XV fasc. pp.55-62. 2017.

[2] Abdulsalam, S. O., Hambali, M. A, Salau-Ibrahim, T. T, Saheed, Y. K., and Akinbowale, N., B. “Knowledge Discovery From Educational Database Using Apriori Algorithm”. *GESJ: Computer Science and Telecommunications* No. 1 (51) pp. 41-51. 2017.

[3] Raheela, A., Agathe, M., Mahmood, K. P. “Predicting Student Academic Performance at Degree Level: A Case Study. *I.J. Intelligent Systems and Applications,* 01, pp.49-61 2015.

[4] Delen, D. "A comparative analysis of machine learning techniques for student retention management." *Decision Support Systems*, vol. 49, pp. 498–506, 2010.

[5] Kovačić, Z. J., "Predicting student success by mining enrolment data". *Research in Higher Education Journal*, vol. 15, pp. 1–20, 2012.

[6] Aguiar, E., N. V. Chawla, J. Brockman, G. A. Ambrose, V. Goodrich. "Engagement vs Performance: Using Electronic Portfolios to predict first semester engineering student retention". *International Conference on Learning Analytics and Knowledge,* ACM, 2014.

[7] Valsamidis, S., and S. Kontogiannis. "E-Learning Platform Usage Analysis". *Interdisciplinary Journal of E-Learning and Learning Objects*, vol. 7, pp. 185-204, 2011.

[8] Merceron, A., and Yacef, K. "Measuring Correlation of Strong Symmetric Association Rules in Educational Data". *In: Handbook of Educational Data Mining ,* edited by C. Romero, S. Ventura, M. Pechenizkiy & R.S.J.d. Baker, CRC Press, , pp. 245 -256. 2010.

[9] Romero, C., Ventura, S., M. Pechenizkiy, and R.S.J.d. Baker, "*Handbook of Educational Data Mining*". CRC Press, ISBN: 978-1-4398-0457-5, 2010.

[10] Magdalene D, Delighta A. & Samuel I., Peter, J. "Association Rule Generation using Apriori Mend Algorithm for Student‟s Placement". *International Journal of Emerging Sciences*, Vol. 2, No. 1, Pp. 78–86.