Iree Journal of Science, Technical and Management Studies (*IJOSTEM*)

Iree Journal of Science, Technical and Management Studies x(y): pp-pp, 2015
ISSN: xxxx-yyyy

Modeling and Predicting Students' Academic Performance in Tertiary Institutions

[1]Ayantola, J. A., [2]Ojo, D.A.; [3]Olaleye, G. O. &[4]Adetunji, M. I.

[1,2,3 & 4]Department of Computer Science,
Osun State Polytechnic, Iree
ayantoso2007@yahoo.com

Abstract:This paper addressed the various factors and their influence on student performance in education and predicts their performance. Various factors were put into consideration, factors like previous year results, attendance, financial status, father's income, mother's income, parent educational qualification, internet use of study material, extra classes, extra-curricular activities etc. all these factors plays vital role in students' education. So many literatures were reviewed thoroughly, and this assisted in correcting the arisen issues on the performance of the students in the tertiary institutions. This discusses the most common measurable factors among students. Data mining algorithms were applied in this paper to predict the performance of the students based on some instances collected via a survey. It's presents the result based on the factors i.e. instances used as input into the data mining software by applying the algorithms and the generated output will be used by the educational administrator to make and take effective decision. Rapid 'Miner software is used to get the final outcome which helps in predicting the final result of students performance.

Keywords:Data Mining, Algorithms, Prediction, Educational System, Naive Bayes Classifier, Decision Free Student Performance.

Introduction

There is a growing interest among researchers that use data mining in educationaltechnologies, or educational data mining (EDM). There are so many fields from whichEducational Data Mining methods are derived such as data mining and machine learning,

psychometrics and other areas of statistics and information visualization.

Machine learning has proven to be of great value in data mining problems especially where large databases contain valuable implicit regularities that can only be discovered automatically; in poorly understood domains where humans might not have the knowledge needed to develop effective algorithms such as face recognition from images and in domains where the program must, dynamically adapt to changing conditions (Schaffer, 1994). It is generally taken to encompass automatic computing procedure based on logical or binary operations that learn a task from a series of examples. It involves searching a very large space of possible hypotheses to determine one that best fits the observed data and any prior knowledge held by the learner (Michie, Spiegelhalter & Taylor, 1994).

Machine learning approach to solving problems involves a number of choices such as choosing the type of training experience, the target function to be learned, a representation for this target function, and an algorithm for learning the target function from training examples. The most commonly used machine learning algorithms are Artificial Neural Network, Decision trees, Genetic Algorithms (GA), Rule Inductions Regression Methods (RIRM), and so on.

Basically, this work focus on Artificial Neural Networks (ANN) and Decision trees algorithms.

In addition the introduction of students' evaluation of staff, though not novel, has helped to further shove up educational standard in one of Nigerian universities (Afemikhe, 2007).

Data mining techniques have been applied in many application domains such as Banking, Fraud detection, Instruction detection and Communication. Recently the data mining techniques were used to improve and evaluate the education tasks. There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining (EDM), concerns with developing methods that discover knowledge from data that comes from educational environments. Data can be collected form historical and operational data reside in the databases of educational institutes. The student data can be personal or academic. Also it can be collected from e-learning systems which have a vast amount of information used by most institutes.

The process of the formation of significant models and assessment within KnowledgeDiscovery in Databases- KDD and is referred to as Data mining (Erdogan and M.Timor). The use of data mining techniques may improve the efficiency of educational institutions. If data mining techniques such as decision tree, clustering, association be applied to education processes, it can help improve student performance, selection of course, their retention rate etc.

Nowadays, that the data in the educational institutes has increased into many folds. These databases contain a lot of hidden information that can be used for the improvement of student performance.

Kuyoro (2010) used datasets of students admitted into the department of Agriculture and Industrial Technology (AIT) of Babcock University between 2002 and 2009. This work extends the previous study with a larger dataset of students admitted into Babcock University, Nigeria between 2001 and 2010. The institution was purposively selected and seven thousand, live hundred (7.500) records of students representing 68% of the eleven thousand, and twenty-seven (11,027) admitted into Babcock University, Nigeria between 2001 and 2010 were randomly selected. The students' first year academic performance was measured by Cumulative Grade Point Average (CGPA) at the end of the first session and the previous academic achievement was measured by Senior Secondary Certificate Examination (SSCE) grade score and University Matriculation Examination (UME) score.

Objectives of the study

- Identify students' family background factors and previous academic achievements that influence academic performance
- Generate the classification models (decision trees and rule induction) and a multilayer perception -an artificial neural network function.
- Apply the classifiers to predict the students' final grade.

Literature Review

Recently, focus is on application of machine learning to educational datasets to proffer solutions to education challenges especially in relation to predicting students' academic performance. (Thai-Nghe, et al, 2007). The most commonly studied model on students' academic performance, persistence in school and students dropouts was Tinto's theoretical model. This model was followed by multiple studies that used different methods to test it (Alkhasawneh, 2011). In the recent time, other models have been developed either as a result of using Tinto's model as building block, reference to it or independent of it. This section discusses literature review of machine learning application to students' academic performance; starting from the theoretical concept of Tinto's model.

It has been proposed that educational data mining methods are often different from standard data mining methods, due to the need to explicitly account for (and the opportunities to exploit) the multi-level hierarchy and non-independence in educational data. For this reason, it is increasingly common to see the use of models drawn from the psychometrics literature in educational data mining publications (Barnes, 2005).

Tinto's Model

Tinto's model is the basic building block of studies on students' performance, retention or dropout. This model presents the predominant theoretical framework for

considering factors in academic success. Tinto (1,975) noted that integration into college system, academically and socially; inform decision on whether to stay on or to dropout. Tinto (1982) considers the process of student attrition as a socio-psychological interplay between the characteristics of the student entering university and the experience at the institute. This interaction between the student's past and the academic environment leads to a degree of integration of the student into his/her new environment as depicted in Figure 1. Later studies of this author tried to make this model operational by identifying factors such as peer group interactions, interactions with faculty, faculty concern for student development and teaching, academic and intellectual development and institutional and goal commitments that affect the student's integration. These factors proved to have a predictive capacity across different institutions.

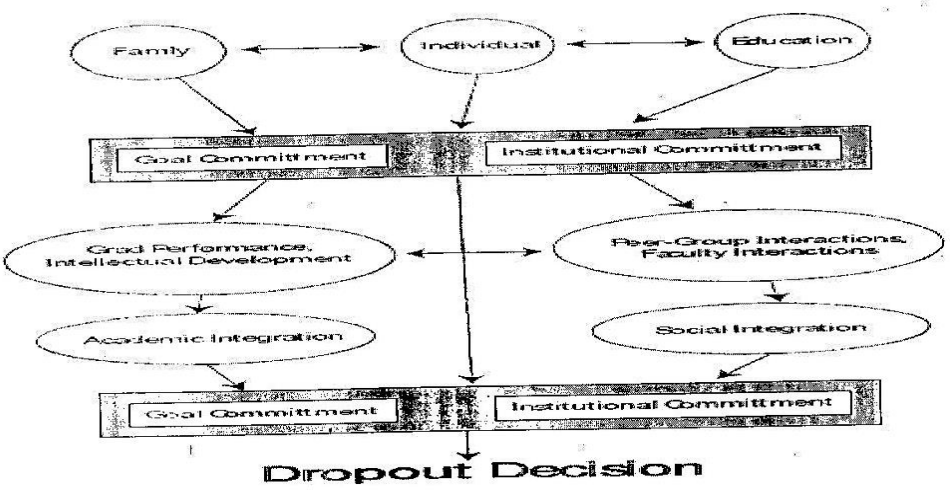

Figure 2.1: Tinto's model for graduation. Source: (Tinto, 1982)

Tinto's postulation in 1982 pointed out that though other factors contribute to students' integration to tertiary institution; students' past is also very significant in determining academic success.

Factors affecting students' performance

The problem relating to students' academic performance has been referred to in literature as one-hundred factor problem (Marquez-Vera et al, 2011). Studies have shown that though many factors contribute to the outcome of student performance family influences are more strongly related to students' academic outcomes than other factors. It has been proposed that for students to achieve academic success what parents do at home is very significant. Kellaghan, (1993) and Redding (1999)) emphasized that it is important to recognize the nature of interrelationships between family background characteristics and other refined family influences. There are so many factories that can be considered as family background factors, poverty level or financial status, family type, socioeconomic status among others (Hodge & Mellin, 2010).

Poverty is an important factor accounting for differences in students' performance and achievement across rural, sub-urban and urban districts. The United States Department of Education (2000) found in a study that the relationship between poverty and students performance is relatively simple and direct. The study, however, concluded that poverty alone does not account for all the differences in the performance of the students. Financial status of parents has elastic effects on their children academic performance as they lack enough resources and funds to sponsor their education. (Johnson, 1996) Parents' poverty made education and learning difficult for the children and can lead to low performance and psychological problems (Mba, 1991)

Hijazi and Naqvi (2006) conducted a study on the student performance by selecting a sample of 300 students (225 males, 75 females). The hypothesis was stated as "Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance." It was found that the factors like

mother's education and student's family income were highly correlated with the student academic performance.

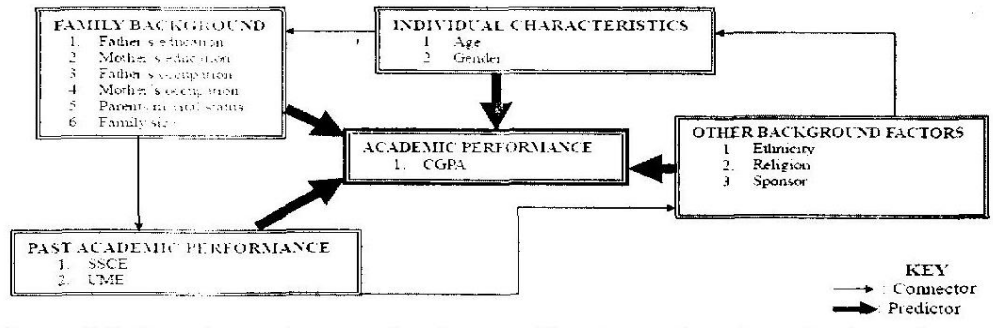

Figure 2.2: Causal model presenting factors affecting students' academic performance

Machine Learning Algorithms and Students Academic Performance

Machine Learning (ML), a subfield of Artificial Intelligence (AI) is concerned with the development, understanding and evaluation of algorithms and techniques that enables computers to learn. It intertwines with other disciplines such as statistics, human psychology and brain modeling (Taiwo, 2009). Machine Learning techniques embody some of the facets of the human mind that allow us to solve hugely complex problems at speeds which outperform even the fastest computers (Schank. 1982).

Machine learning is considered the most suitable technology in giving additional insight into educational entities such as student, lecturer, staff, alumni and managerial behavior. It acts as an active automated assistant facilitating improved decision making processes in higher learning institutions. The improvement include; increasing student's promotion rate, retention rate, transition rate, increasing educational improvement ratio, increasing student's success, increasing student's learning outcome, maximizing educational system efficiency, decreasing student's drop-out rate, and reducing the cost of system processes (Delavari et al, 2005).

Application of Decision trees algorithms to students' performance

Decision trees arc tree-shaped structures that represent a series of rules that lead to sets of decisions. These decisions generate rules for the classification of a dataset. Decision trees present a clear, logical model that can be easily understood by people who are not mathematically inclined. They show how the value of a target variable can be predicted by using the values of a set of predictor variables. Each node represents a set of records (rows) from the original dataset. Nodes that have child nodes are called interior nodes while nodes that do not have child nodes are called terminal or leaf nodes. The topmost node is called the root node. Unlike a real tree, decision trees are drawn with their root at the top. The root node represents all the rows in the dataset. A typical decision tree has three types of nodes: (a) decision node (b) chance node, and (c) leaf node (Kuyoro. 2001). Hence, since decision trees score so well on so many application to different fields; this significant machine learning algorithms are being used in a variety of academic problems for Null exploration and prediction (Ranjan and Ranjan, 2010).

Application of Neural Networks algorithms to students' performance

Neural Network (NN) is a class of systems modeled after the human brain. As the human brain consists of millions of neurons that are interconnected by synapses; neural networks are also formed from hire numbers of simulated neurons, connected to each other in a manner similar to brain neurons. Like in the human brain, the strength of neuron interconnections may change (or be changed by the learning algorithm) in response to a presented stimulus or an obtained output, which enables the network to 'teach'" (Akinola et al, 2012). The most common type of artificial neural network consists of three groups, or layers of units: input, hidden and output layer. A layer of input units connected to a layer of hidden units, which is connected to a layer of output units.

Research is a systematic, controlled, empirical, and critical investigation of natural phenomena guided by theory and hypotheses about the presumed relations among suchphenomena (Kerlinger 1986). It is the systematic, objective analysis and recording of controlled observations that may lead to the development of generalizations, principles, or theories, resulting in prediction and possibly ultimate control of events (Best 1985). Research design is the plan and structure of the investigation so conceived as to obtain answers to research questions. The plan is the overall scheme or programme of the research. The plan includes an outline of what the investigator do join the formulation of hypotheses and their operational implications to the final analysis of data (Kerlinger 1986).

Methodology

Data gathering is very crucial in research, as the data is meant to contribute to a better understanding of theoretical framework (Bernard 2002, Manly 1994). It then becomes imperative that selecting the manner of obtaining data and from whom the data will be acquired be done with sound judgment, especially since no amount of analysis can make up for improperly collected data (Bernard et al. 1986).

The data set used in this study has been obtained from different students reading in various departments. A questionnaire was prepared containing a lot of multiple choice of questions. Also a semi interview was also conducted with some students. These questionnaires were distributed among the students and finally collected back from them. Bach student filled the questionnaires with independence and according to his/her real data. Those fields that are needed by data mining software and for prediction purpose are selected. The data so collected is then divided into two categories; Training set and testing set.

Table 1.1 Items used

ITEM	DATA	INPUT TYPE	VALID INPUT OPTION
Age on Entry	Integer	Text Field (parsed tointeger)	Integer value (17>= value <=25)
Family Size	Integer	Text Field (parsed tointeger)	Integer value (value > 0)
Gender	String	Combo Box	Male: Female
Sponsor	String	Combo Box	Parent; Scholarship; Self; Father;Mother; Guardian; Other
Mother'sEducation	String	Combo Box	No formal education; Primary;SSCE; first degree; Second degree;PhD
Mother's Occupation	String	Combo Box	Unemployed; Government worker;Private; Self employed
Father's Occupation	String	Combo Box	Unemployed; Government worker;Private; Self employed
Extra Curricula	String	Text Field	Clubbing, Party, Sport, etc.

Source:

Model Building

In the machine learning Field, there are three basic phases: pre-processing, processing and postprocessing in setting up a model.

WEKA is an open source machine learning package developed in Java; it is a collection of machine learning algorithms for data mining tasks. It contains tools for data preprocessing, classification. Regression clustering, association rules, and visualization. There are many machine learning algorithms implemented in WEKA including Bayesian classifiers, Decision Trees, Rules. Functions, Lazv classifiers and miscellaneous classifiers.

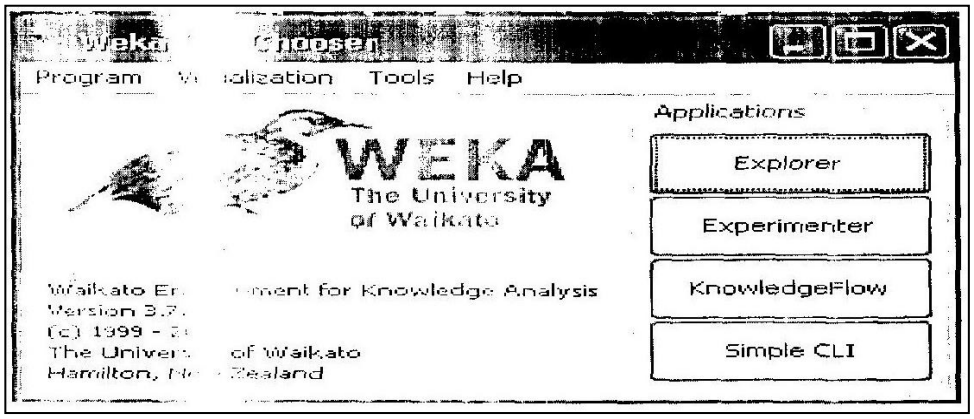

Figure 3: WEKA GUI Chooser Source: Kirkbyet all (2007)

Decision trees

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision-making. Decision tree starts with a root node on which it is possible for user to take actions. From this node, users split each node recursively according to the decision tree learning algorithm

$$\text{Entropy} = \sum_{c} -\left(\frac{n_{bc}}{n_b}\right)\log_2\left(\frac{n_{bc}}{n_b}\right) \quad \text{.....(1)}$$

$$\text{Average Entropy} = \sum_{b}\left(\frac{n_b}{n_t}\right) \times \left[\sum_{c} -\left(\frac{n_{bc}}{n_b}\right)\log_2\left(\frac{n_{bc}}{n_b}\right)\right] \quad \text{.....(2)}$$

- **Gini impurity**

$$I_G(f) = \sum_{i=1}^{m} f_i(1-f_i) = \sum_{i=1}^{m} f_i - f_i^2 = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2 \quad \text{.....(3)}$$

- **Information gain**

$$I_E(f) = -\sum_{i=1}^{m} f_i \log_2 f_i \quad \text{.....(4)}$$

- **Gain ratio**

$$IntrinsicInfo(S,A) \equiv -\sum \frac{|S_i|}{|S|}\log_2\frac{|S_i|}{|S|} \quad \text{.....(5)}$$

$$GainRatio\ (S,A) = \frac{Gain\ (S,A)}{IntrinsicI\ \ nfo\ (S,A)} \quad \text{(6)}$$

Neural Networks

A neuron is a special biological cell with information processing ability. Neural networks try to mimic interconnected; neurons in human brain in order to make the algorithm capable of complex learning for extracting patterns and detecting trends. Neural networks are best at identifying patterns or trend in data and well suited for prediction or forecasting needs. (Akinola et al, 2012). A typical neural network composed of three types of layers, namely, the input layer, hidden layer, and output layer.

Experimental

Figure 4.1: Visualization of Attributes

Source: WEKA 3.7.7

Result

S/N	Attributes	Information Gain	Gain Ratio		
		Value	Rank	Value	Rank
1.	Gender	0.0389	10	0.0453	11
2.	Age on entry	0.1689	*5*	0.0951	5
3.	Ethnicity	0.0609	8	0.0478	10
4.	Religion	0.0277	14	0.0673	9
5.	Family Size	0.1465	6	0.0745	7
6.	Sponsor	0.064	7	0.0681	8
7.	Father's education	0.0313	12	0.044	12
8.	Mother’s education	0.0359	11	0.0424	13
9.	Father’s occupation	0.4343	2	0.1088	2
10.	Mother’s occupation	0.374	3	0.1063	4
11.	Parent’s marital status	0.0588	9	0.0761	6
12.	Monthly Familyincome	0.0293	13	0.0397	14
13.	Extra Curricula activities	0.0423	4	0.1082	3

Source:

Corresponding Author: Name, Affiliation, Email, Tel

Performance of Decision Stump Model

Table 3.7 Summary result of decision stump absolute error analysis

Variable	**10- fold**	**Holdout**
Correct classification instance	38.8378%	38.9587%
Incorrect classified instance	61.1622%	61.0413%
Kappa statistics	0.0689	0.0635
Mean absolute error	0.1923	0.1923
Root mean squared error	0.3103	0.3102

Performance of Multilayer Perceptron (MLP) Model

Table 4.7 Summary result of Multilayer Perceptron absolute error analysis

Variable	**10-fold**	**Holdout**
Correctly classified instances	98.7595%	96/7843%
Incorrect classified instances	1.2405%	3.2157%
Kappa statistics	0.9785	0.9589
Mean absolute error	0.006	0.0091
Root mean squared error	0.07	0.0795

Performance of Native Bayes Model

Table 4.8 summary result of ZeroR absolute error analysis

Variable	**10-fold**	**Holdout**
correctly classified instances	36.4765%	96/7843%
incorrect classified instances	1.2405%	3.2157%
Kappa statistics	0.9785	0.9589
Mean absolute error	0.006	0.0091
Root mean squared error	0.07	0.0795

Conclusions

This research work indicate that social-economic attributes, mostly parental information and education sponsor are influencing factors of students' academic performance in tertiary institution the academic performance of students was measured by the students' first year cumulative grade point average (CGPA). Waikato Environment for Knowledge analysis

(WEKA) were used to generate the experimental result, a variation of decision tree discovered to be the best among the models, was used to predict new cases of students' performance.

Recommendations and suggestion for further studies

Future studies could use the same procedure and models with a larger dataset and more identified background factors of student at the first year academic level. This study focused on determining to what extent factors and previous academic achievement affect student’s academic performance in tertiary institution; the specific machine learning algorithm best model the student academic performance. The result generated from the machine learning can be used to design a system (i.e. graphical system). Any OOP language can be used to implement this for further research.

References

Adeyemo, A.B. and Kuye G., (2006).Mining Students Academic Performance using Decision Tree algorithms. Journal of information *Technology Impact 6(3):* 161-170.

Afemikhe, O.A (2007) Assessment and educational standard improvement, reflection from Nigeria.

Akinola O.S Akinkunmi, B.O and Alo, T.S (2012). Data Mining model for predicting computer programming profitency. African journal of computing & ICT C(n) 43-52

AI-Radaideh, Q.A, Al-Shawakfa, E.M., and Al-Najjar, M.I. (2006). Mining Students Academic Performance using Decision Tree in proceeding of the 2006 international Arab conference on information Technology (ACIT 2006)

Baker, R. S.J. D., andYacef, K. (2009). The state of educational data mining in 2009: *A review and future vision. Journal of educational Data Mining, 1, 3-17*

Bhardwaj and S.Pak *(2011).Data* mining, a prediction for performance improvement using *classification. Int. J. Computer Science and Information Security, Vol 9, no, 4*

Bousquest O. (2013) Machine learning Thoughts is there an optimal learning algorithm Available at: http//ml.typepad.commachinelearningthought/2005/08/is thereanopt.htmlRetrived15-10-2013

Coetez, P., and Silva, A. (2003).UaingdT mining to predict secondary school student performance. In proceeding of 5th Annual Futere Business Technilogy Conference, Porto, Portugal, 5-12

Han J, and Kamber M. (2008). Data mining: Concepts and techniques. Morgan Kaufmann Publisher, New Delhi.

Jadric, M., Garaca, Z. and okmusic, M. (2010) Student Dropout Analysis with Application of Data Mining methods.Managemental.Vol, 15, pp. 31-46

Kirkby, R., Frank, E and Reculemann, P. (2007) WEKA Explorer User Guide for version 3-5-5,26 January

Kuyoro S. O. Goga N. Awodele O. and Okolie S. O. (2013a),, '' Optimal Alogorithm for Predicting Stdent'sAcedemic performance", international Journal of computers and Technology., 4(1) No1: pp. 63-75

Romero C and Venture S. (2010): Educational Data Mining: A Review of the state of the Art, IEEE Trabsaction on systems, man and Cybernectic- Part C: Applications and Reviews 2010, No. 40/6, pp 01-13

McKenzie. K. and R. Schweitzer, (2001). Who succeeds at University? Factors predicting academic performance in first year Australian University Students.*Higher Educ. Res. Dev.*, 20: 21-33.

Walters Y.B, and Soyibo K. (2001): "An Analysis of High School Students' Performance on

Five Integrated Science Process Skills", *Research in Science & Technical Education,Vol. 19, No. 2, 2004, pp. 133-145.*

Corresponding Author: Name, Affiliation, Email, Tel