I.J. Education and Management Engineering, **2015, 5, 12-19**
Published Online October 2015 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2015.05.02

Available online at http://www.mecs-press.net/ijeme

A Decision Tree Approach for Predicting Students Academic Performance

Kolo David Kolo[a], Solomon A. Adepoju[b], John Kolo Alhassan[b]

[a] *Department of Computer Science, Niger State College of Education, Minna*
[b] *Department of Computer Science, Federal University of Technology Minna, Nigeria*

Abstract

This research is on the use of a decision tree approach for predicting students' academic performance. Education is the platform on which a society improves the quality of its citizens. To improve on the quality of education, there is a need to be able to predict academic performance of the students. The IBM Statistical Package for Social Studies (SPSS) is used to apply the Chi-Square Automatic Interaction Detection (CHAID) in producing the decision tree structure. Factors such as the financial status of the students, motivation to learn, gender were discovered to affect the performance of the students. 66.8% of the students were predicted to have passed while 33.2% were predicted to fail. It is observed that much larger percentage of the students were likely to pass and there is also a higher likely of male students passing than female students.

Index Terms: Prediction, Data Mining, Performance, Decision Tree, Academic.

1. Introduction

Knowledge discovery in databases is the process of identifying valid, novel, potentially useful and ultimately understandable patterns in data**.** Data mining is a step in the knowledge discovery process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, finds patterns or models in data. In other words, the goal of knowledge discovery and data mining is to find interesting patterns and/or models that exist in databases but are hidden among the volumes of data [6]. Data mining is applied in various fields of human endeavors including education; this is referred to as educational data mining. Educational data mining is a discipline, which is concerned with developing methods for exploring the unique types and pattern of data that come from educational institutions as well as using those methods to better understand students, and the settings which they learn in.

There have been increasing research interests in using data mining in education. This new emerging field,

* Corresponding author. Tel.: +2348035956346
E-mail address: kolodave2@gmail.com, sa.adepoju@gmail.com

called Educational Data Mining is concerned with developing methods that discover knowledge from data originating from educational environments. Data mining methods or approaches includes Classifications, Clustering, Na ïve Bayesian, decision trees, neural networks and Fuzzy rules

All participants in the educational process (teachers, students, academic planners and administrators) could benefit by applying data mining on the data from the higher education system. Data mining represents the computational data process from different perspectives, with the goal of extracting implicit and interesting samples trends and information from the data. So, it can greatly help every participant in the educational process [1].

According to [17] the concept of educational data mining may be classified into the following classes: Prediction which includes the following sub classes: classification, regression, density estimation; Clustering which includes relationship mining, association, correlation mining, sequential pattern mining, causal data mining; distillation of data for human judgment and discovery with models. The following are areas of advantages of educational data mining in educational institutions [5]:

- Accurate analysis and visualization of data
- The ability to provide feedback for teachers/ instructors
- The ability to predict student performance
- The production of functional student behavioural models
- The ability to detect undesirable student behaviours
- It also helps to group students into classes based on various performance characteristics
- The ability to analyze the students' social networks
- Developing concept maps
- It also aid in the construction of curriculums
- It helps in the education planning and administration

Prediction of student performance is helpful in order to provide a student with the necessary assistance in the learning process [3]. Knowing the possible outcome of the learning process based on results of prediction can help an institution to change the outcome of the new set of students by adjusting the factors that contributed to the past performance. The ability to predict students' performance can also aid educational planners and administrators to adequately plan for the change in students population in any direction, i.e. increase of decrease. However, it is usually difficult to come up with a manual set of rules that are needed to predict students' performance. Hence there is a need to look for readily capable methods for dealing with the task. This research has looked specifically at personal, social and economic factors that may affect students' performance and used same to predict the performance of students in the coming session or semester as the case may be.

The objective of this study is to identify factors that will affect the academic performance of students and predict the academic performance of students as either pass or fail by using the decision tree and ordinal regression approaches of the SPSS software. Decision Tree technique has been found to be a very adequate technique to generate a comprehensive and precise analysis. Decision trees are used in data mining to study historical data and on the basis of the data analysis and its rules, one can predict the result [18]. SPSS combines the features of data mining activities which is a combination of Statistical analysis and database management. The SPSS package is easily available as a commercial off the shelf software and can easily be copied from one computer to another due to its robust portability. It also has a good graphics user interface thereby making it easy to learn and use by teachers/instructors who may want to use it for further analysis. The results are also easy to view and interpret for decision making.

2. Review of Related Work

In order to predict the performance of students the researcher took into consideration the work of other researchers that are in the same direction. Other researchers have looked at the work of predicting students' performance by applying many approaches and coming up with diverse results.

Three supervised data mining algorithms, i.e. Bayesian, Decision trees and Neural Networks which were applied by [1] on the preoperative assessment data to predict success in a course (to produce result as either passed or failed) and the performance of the learning methods were evaluated based on their predictive accuracy, ease of learning and user friendly characteristics. The researchers observed that that this methodology can be used to help students and teachers to improve student's performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of learning.

[12] described the process of knowledge discovery from databases using a practical example of a current actual problem. They developed two models based on decision tree which were successfully used to predict student success based on GPA criterion and time student needs to finish the undergraduate program (time-to-degree) criterion.

Bayesian classification method was also used by [4] in their work on student database to predict the students' grades on the basis of previous year performance. The researchers concluded that the study will help the students and the teachers to improve the grades of the student. The study also helps to identify those students which needed special attention to reduce failing ratio and taking appropriate action at right time.

[2] compared four different classifiers and combined the results into a multiple classifier. Their research divided the data into three (3) different classes weighing the features and using a genetic algorithm to minimize the error rate improves the prediction accuracy at least 10% in the all cases of 2, 3 and 9-Classes. In cases where the number of features is low, the feature weighting worked much better than feature selection. The successful optimization of student classification in all three cases demonstrates the merits of using the LON-CAPA data to *predict* the students' final grades based on their features, which are extracted from the homework data. However, the research in this case was based on an online course as opposed to the regular classroom class that the present study considers.

Furthermore, [3] observed that in the problem of prediction of performance, it is possible to automatically predict students' performance. Moreover by using extensible classification formalism such as Bayesian networks, which was employed in their research it becomes possible to easily and uniformly integrate such knowledge into the learning task. The researchers' experiments also show the need for methods aimed at predicting performance and exploring more learning algorithms.

[8] addressed the prediction of secondary school students' performance in two core subjects of mathematics and Portuguese by using their past score in the previous session and other demographic factors and employed four data mining methods of Decision trees, Random Forests, neural networks and Support Vector machines approach. The results show that the prediction was achievable provided the grades of the previous session were known. This confirms that the prediction of students' performance is premised on past performance and hence shows that a student's performance is closely related to the performance in previous course (most likely a prerequisite course).

[11] concluded that Decision trees are so popular because they produce classification rules that are easy to interpret than other classification methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for Student data to predict the student's performance in the end semester examination. The experimental results show that Classification and Regression Tree (CART) CART is the best algorithm for classification of data. From the study conducted by [7], by using a case study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success. Two algorithms CHAID and CART were applied on student enrolment data of information system students of open polytechnic of New Zealand to get two decision trees classifying successful and unsuccessful students. The result obtained showed that the accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively.

In their study [9] used classification task to predict the final grade of students. It was done by the use of ID3 decision tree method. Another study undertaken by [10] showed that Data Mining Techniques (DMT) capabilities provided effective improving tools for student performance. The study further showed how useful data mining can be in higher education particularly to predict the final performance of student. The researchers collected data from student by using questionnaire to find the relationships between behavioural attitude of student and their academic performance. Data mining techniques were then applied. They obtained the prediction rule model using decision tree as well as implementing the rules into Support Vector Machine (SVM) algorithm to predict the students' final grade. Also the students were clustered into groups using kernel k-means clustering. The study expressed the strong correlation between mental condition of student and their final academic performance.

Also, [16] used Iterative Dichotomiser 3 (ID3) decision tree algorithm to predict the university students' grade of a university in Nigeria. A prediction accuracy of 79,556 was obtained from the model. They further suggested the use of other decision based model to predict student's performance.

3. Methodology

The data collected was for a course CSC214 (Data Structures) which is a 2nd year course for NCE in computer science in Nigerian Colleges of Education. The course was chosen because of the familiarity of the researchers with the course and the grades considered were the semester results for the course CSC214 for three previous years. A questionnaire was also distributed to students to collect data about the other factors considered in the prediction such as the students' financial strength and motivation to study. The data collected from the result sheet was entered into SPSS Version 20 for analysis.

Table 1. Factors of the Prediction

Factor	Description	Attributes	SPSS value
Score	Student's grade	≥40 and <40	0 – 100
Status	Student's status	Pass, Fail	1,2
Gender	Students gender	Male, female	1,2
Finance	Financial strength	Low, Medium, High	1,2,3
Motivation	Attitude to learning	Low, Moderate, High	1,2,3

The factors that are considered in predicting the performance of students in this study are;

- Scores: this is the total scores obtained by the students in the previous session of the course under consideration. The scores' attributes are within the range of ≥40 and <40. The scores are however entered into SPSS as obtained from the results sheets (i.e. 0 – 100). For the purpose of this study,the classification was reduced from the six class classification of scores i.e. A, B, C, D, E, F to a two-class classification of Pass and Fail in order to reduce the margin of error for the research as it regards to the prediction
- Status: Status refers to the remark obtained for the scores of the students, i.e. 'Pass' or 'fail'. In which case a score less than 40 is entered as fail while a score greater or equal to 40 is entered as pass. The status is coded as 1 or 2; where 1 stands for pass and 2 stands for fail.
- Gender: Gender is used to classify the students as either male or female. This factor is quite important because it helps to know the effect of gender of the students on predicting. This is in agreement with [13] who established that gender plays a great role in the performance of female
- Students in a large public Turkish University. Though gender and age are predictors in the academic performance of students, gender is a better predictor [14].
- Finance: this refers to the financial status of students while in school, this factor is important because it affects the stability of the students as it relates to their comfort during the academic session. Finance is an avenue trough which the students are able to settle their bill. Students without adequate financial strength are affected adversely [15]. The attributes of the factor include, High, Medium mad Low are they are

coded in SPSS as 1, 2 and 3. Where 1 is for high, 2 is medium and 3 is Low.

- Motivation: this refers to a factor that pushes a student to work hard or serves as reinforcement for the student; such as studying in groups, rewards and prices etc. the attributes are high, moderate and low. These are coded as 1, 2 and 3; 1 for low, 2 for moderate and 3 for high.

In building the predictive decision tree using the SPSS program based on the criteria in Table 1, the SPSS predictive approach used is the CHAID approach. The dependent variable is Score which implies that the study is concerned with predicting the score of the student using the existing (previous) score. The factors that influence that scores are listed under the independent variables as status finance and gender while the variable chosen as the influence variable is the motivation.

Table 2. Model Summary

Specifications	Growing method	CHAID
	Dependent Variable	SCORES
	Independent Variables	STATUS, FINANCE, GENDER
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	STATUS, GENDER
	Number of Nodes	5
	Number of terminal Nodes	3
	Depth	2

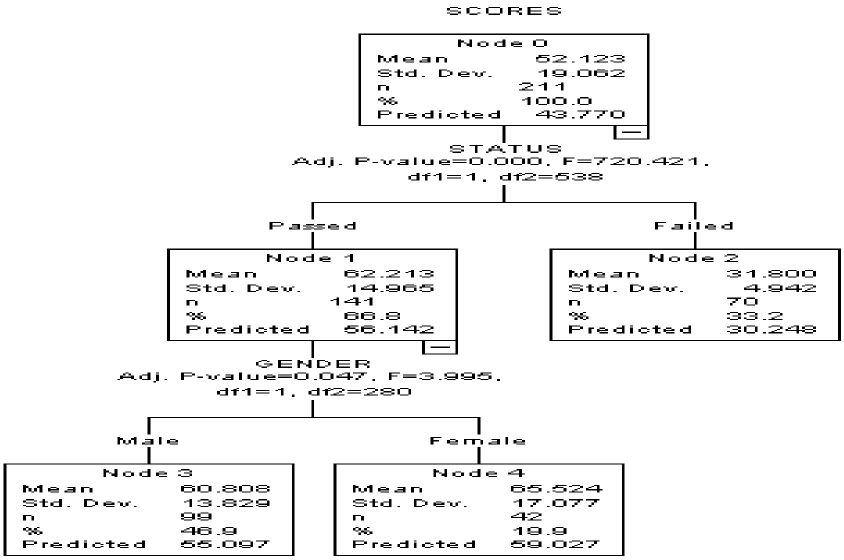

Fig. 1. Decision Tree Structure for Predicting Students' Academic Performance

4. Results and Discussion

Fig 1 shows the decision tree for the prediction of the students' performance in the study. The first test was based on the status of the students which was determined based on the number of passed and failed students. The result shows that 66.8% passed and 33.2% failed and based on the principle of decision trees it shows that the percentage of those who passed is higher and therefore the new consideration is the passed node (that is why the next decision tree node is under the passed node).

The next factor for consideration is the gender (Male, Female) and based on the principle of decision tree structure the gender division shows that 55.10% of the male students are predicted to pass while 59.03% of the female is predicted in the predictive analysis. Therefore this shows that the study is showing that more female students are predicted to do better in the course than the male students.

Table 3. Gain Summary for Nodes

Node	N	Percent	Mean
4	42	19.9%	65.52
3	99	46.9%	60.81
2	70	33.2%	31.80

Growing Method: CHAID, Dependent Variable: SCORES

Table 3 shows the summary of the gain for the nodes in the predictive decision tree structures. It shows that node tree has a higher gain than the other nodes at 46.9% which is an indication of the level of influence of the third note factor on the predictive analysis of the students' performance.

5. Conclusion and Recommendations

The research has shown that some factors such as finance level, motivation level, gender and grades obtained in previous (prerequisite) courses can affect the performance of students in academic institutions. This was also established in some of the reviewed literature in this paper. More so, a more automated system can be used to achieve the prediction of student's performance through data mining, by using software with such capabilities such as SPSS used in this work. The research would have produced a clearer and more accurate result if a higher version of SPSS was employed which could have shown a larger view of the tree model without affecting the MS Word window. More research involving additional factors other than the ones used in this research could still be carried out. This would have helped to have a clearer and more detailed predictive analysis.

The researcher wishes to recommend that the public and private educational institutions in Nigeria should pay more attention to efficient prediction analysis in education. This could be achieved by initiating and funding research into the prediction analysis in education, as it will aid educational administration and planning.

References

[1] Osmanbegovic E., Suljic M. "Data mining approach for predicting student performance" Economic Review-Journal of Economics and Business. Volume 10(1) (2012)

[2] Behrouz, M, Karshy, D, Korlemeyer G, Punch, W. "Predicting student performance: an application of data Mining methods with the educational web-based system" Lon-capa. 33rd ASEE/IEEE Frontiers in Education Conference. Boulder C.O. USA, (2003).

[3] Bekele, R., Menzel, W. “A bayesian approach to predict performance of a student (BAPPS): A Case with Ethiopian Students”. Journal of Information Science (2013).

[4] Bhardwaj, K., Pal, S “Data Mining: A prediction for performance improvement using classification”. International Journal of Computer Science and Information Security. Volume 9(4). (2011).

[5] Romero, C, Ventura, S. “Educational Data Mining: A Review of the State-of-the-Art. IEEE Transaction on Systems, Man, and Cybernetics, Part C” Applications and Reviews. Volume 40(6) (2012).

[6] Bae, E., Bailey, J: “COALA: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity”. Proceedings of the Sixth International Conference on Data Mining. Pp. 53 – 62. (2006).

[7] Kovacic, Z. “Early prediction of student success: Mining student enrollment data” Proceedings of Informing Science & IT Education Conference. (2010).

[8] Cortez P, Silva A. Using data mining to predict Secondary school student performance. Journal of information science Volume 2(6). (2013).

[9] Ahmed, A. B. E, Ibrahim S. E.. "Data Mining: A prediction for Student's Performance Using Classification Method." *World* Journal of Computer Application and Technology Volume 2(2) (2014).

[10] Sembiring S, Zarlis, M, Hartama, D. Ramliana S, Elvi W. “Prediction of student academic performance by an application of data mining techniques.” International Conference on Management and Artificial Intelligence IPEDR Volume.6, (2011).

[11] Surjeet K, Yadav, Bharadwaj, B. Pal B.” Data Mining Applications: A comparative Study for Predicting Student’s performance.” International journal of innovative technology & creative engineering. Volume 1(12). (2012).

[12] Mladen D., Mirjana P. B., Vanja Š., “Improving University Operations with Data Mining: Predicting Student Performance”, International Journal of Social, Behavioral, Educational, Economic and Management Engineering Volume 8(4), 2014.

[13] Meltem, D. “Gender difference in academic performance in a large public university in Turkey”. Economic Research center working papers in economics. 4(17). Pp. 22-23, (2004).

[14] Abubakar, R. B. and Oguguo, O. D. “Age and gender as predictors of academic achievements of College mathematics and science students.” Proceedings of the International Conference of teaching, learning and change. International Association of Teaching and learning. (2011).

[15] Nnamani, C. N, Dikko, H. G and Kinta, L. M. “Impact of students’ financial strength on their academic performance: Kaduna Polytechnic experience”. African Research Review 8(1), (2014).

[16] Ogunde A.O., Ajibade D.A. “A data Mining System for Predicting University Students F=Graduation Grade Using ID3 Decision Tree approach”, Journal of Computer Science and Information Technology, Volume 2(1) (2014).

[17] Ryan S.J.D. Baker, Kalina Yacef. “The State of Educational Data Mining in 2009: A Review and Future Visions”, Journal of Educational Mining, Volume 1(2009).

[18] Undavia, J. N., Dolia, P. M.; Shah, N. P. “Prediction of Graduate Students for Master Degree based on Their Past Performance using Decision Tree in Weka Environment”. International Journal of Computer Applications; Volume 74 (21), (2013).

Authors’ Profiles

Kolo David Kolo holds a Bachelor of Technology (B.Tech) Degree in Mathematics/Computer Science from Federal University of Technology (FUT) Minna, Nigeria, in 2000, Postgraduate Diploma in Education (PGDE) from the Ahmadu Bello University (ABU) Zaria, Nigeria and Master of Technology (M.Tech) degree in Computer Science from the Federal university of Technology, Minna, Nigeria respectively in 2015. He works as a lecturer in the Department of Computer Science, College of Education, Minna, Nigeria. His area of interest includes database management and mobile learning.

Solomon. A. Adepoju holds B.Tech (Maths/Computer Science) and M.Sc (Computer Science) from the Federal University of Technology, Minna and University of Ibadan respectively. His research interests include Human Computer Interaction, Web Mining and ICT4D is a member of Computer Professional Association (CPN) as well other international association.

Dr. J. K. Alhassan obtained B.Tech (Mathematics/ Computer Science) at the Federal University of Technology, Minna. M.Sc (Computer Science) at University of Ibadan, Nigeria in 2006, and PhD (Computer Science), at Federal University of Technology, Minna, Nigeria. He carried out part of his PhD research at United Institute of Informatics Problems, National Academy of Sciences of Belarus (UIIP NASB) Minsk, Republic of Belarus. He is currently the Acting Head, at the Department of Cyber Security Science, Federal University of Technology, Minna, Niger State, Nigeria. His research interest includes Artificial Intelligence, Data Mining, Internet Technology, Database Management System, Software Architecture, Machine Learning, Human Computer Interaction and Computer Security. He is a member of Computer Professionals (Registration Council) of Nigeria (CPN).