Performance Evaluation of Naive Bayes and Decision Stump Algorithms in Mining Students' Educational Data

Ayinde A.Q[1], Dr Adetunji A.B[2], Bello M[3] and Odeniyi O.A[4]

**1. Computer Science Department, Osun State College of Technology,
Esa-Oke, 234035/South West, Nigeria**

**2 Computer Science and Engineering Department, LAUTECH,
Ogbomoso, 23402/South West, Nigeria**

**3. Computer Science and Engineering Department, LAUTECH
Ogbomoso, 23402/South West, Nigeria**

**4. Computer Science Department, Osun State College of Technology,
Esa-Oke, 234035/South West, Nigeria.**

Abstract

Educational data mining is an emerging trend, concerned with developing methods for exploring the huge data that come from the educational system. This data is used to derive the knowledge which is useful in decision making which is known as Knowledge Discovery in Databases (KDD). EDM methods are useful to measure the performance of students, assessment of students and study students' behavior etc. In recent years, Educational data mining has proven to be more successful at many of the educational statistics problems due to enormous computing power and data mining algorithms. The main objective of this research is to find out interesting patterns in the educational data that could contribute to predicting student performance .This paper describes how to apply the main data mining methods such as prediction and classification to educational data**.**

Keywords: *Educational Data Mining, Data Mining, Knowledge Discovery in Databases, Algorithms*

1. Introduction

The data mining has attracted a great deal of attention in the information technology industry, due to availability of large volume of data which is stored in various formats like files, texts, records, images, sounds, videos, scientific data and many new data formats. There is imminent need for turning such huge data into meaningful information and knowledge. The data collected from various applications require a proper data mining technique to extract the knowledge from large repositories for decision making. Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large volume of data [1].

Data mining and knowledge discovery in databases are treated as synonyms, but data mining is actually a step in the process of knowledge discovery. The sequences of steps indentified in extracting knowledge from data are shown in Figure 1.

The main functionality of data mining techniques is applying various methods and algorithms in order to discover and extract patterns of stored data. These interesting patterns are presented to the user and may be stored as new knowledge in knowledge base. Data mining and knowledge discovery applications have got a rich focus due to its significance in decision making. Data mining has been used in areas such as database systems, data warehousing, statistics, machine learning, data visualization, and information retrieval. Data mining techniques have been introduced to new areas including neural networks, patterns recognition, spatial data analysis, image databases and many application fields such as business, economics and bioinformatics. The main objective of this paper is to predict the students' grade by application of classifiers to educational data in Department of Computer Science and Engineering, Ladoke Akintola University of Technology, Ogbomoso, Oyo State, Nigeria. The first section is used to describe the history and current trends in the field of Educational Data Mining (EDM). The second section is the review of past works on educational data mining. The third section covers the research methodology. The fourth section covers the results and discussions. The fifth section covers the conclusion.

2. Reviews on Educational Data Mining

The educational data mining community [2] defines educational data mining as, “Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the setting which they learn in”. There are increasing research interests in using data mining techniques in educational filed. This new emerging field, EDM, concerns with developing methods that discover knowledge from data originating from educational environments.

Educational data mining techniques often differ from traditional data mining techniques, in explicitly exploiting the multiple levels of meaningful hierarchy in educational data.

EDM focuses on collection, archiving, and analysis of data related to students’ learning and assessment. The analysis performed in EDM research is often related to techniques drawn from variety of literatures [3], including psychometrics, machine learning, data mining, educational statistics, information visualization and computational modeling.

Reviews pertaining to not only the diverse factors like personal, socio-economic, psychological and other environmental variables that influence the performance of students but also the models that have been used for the performance prediction are available in the literature and a few specific studies are listed below for reference.

Walters and Soyibo [4] conducted a study to determine Jamaican high school students' (population n=305) level of performance on five integrated science process skills with performance linked to gender, grade level, school location, school type, student type, and socio-economic background (SEB). The results revealed that there was a positive significant relationship between academic performance of the student and the nature of the school.

Khan [5] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Hijazi and Naqvi [6] conducted as study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that 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" was framed. By means of simple linear regression analysis, it was found that the factors like mother’s education and student’s family income were highly correlated with the student academic performance.

Kristjansson, Sigfusdottir and Allegrante [10] made a study to estimate the relationship between health behaviors, body mass index (BMI), self-esteem and the academic achievement of adolescents. The authors analyzed survey data related to 6,346 adolescents in Iceland and it was found that the factors like lower BMI, physical activity, and good dietary habits were well associated with higher academic achievement.

Moriana et al. [11] studied the possible influence of extra-curricular activities like study-related (tutoring or private classes, computers) and/or sports-related (indoor and outdoor games) on the academic performance of the secondary school students in Spain. A total number of 222 students from 12 different schools were the samples and they were categorized into two groups as a function of student activities (both sports and academic) outside the school day. Analysis of variance (ANOVA) was used to verify the effect of extracurricular activities on the academic performance and it was observed that group involved in activities outside the school yielded better academic performance.

Bray [12], in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Srilanka. It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions.

Modeling of student performance at various levels is discussed in [10], [11], and [12]. Ma, Liu, Wong, Yu, and Lee [4] applied a data mining technique based on association rules to find weak tertiary school students (n= 264) of Singapore for remedial classes. Three scoring measures namely Scoring Based on Associations (SBA-score), C4.5-score and NB-score for evaluating the prediction in connection with the selection of the students for remedial classes were used with the input variables like sex, region and school performance over the past years. It was found that the predictive accuracy of SBA-score methodology was 20% higher than that of C4.5 score, NB-score methods and traditional method.

Kotsiantis, et al. [8] applied five classification algorithms namely Decision Trees, Perception-based Learning, Bayesian Nets, Instance-Based Learning and Rule-learning to predict the performance of computer science students from distance learning stream of Hellenic Open University, Greece. A total of 365 student records comprising several demographic variables like sex, age and marital status were used. In addition, the performance attribute namely mark in a given assignment was used as input to a binary (pass/fail) classifier. Filter based variable selection technique was used to select highly influencing variables and all the above five classification models were constructed. It was noticed that the Naïve-Bayes algorithm yielded high predictive accuracy (74%) for two-class (*pass/fail*) dataset.

Al-Radaideh, et al. [13] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. They used 12 predictive variables and a 4-class response variable for the model construction. Three different classification methods namely ID3, C4.5, and the Naïve Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models with the predictive accuracy of 38.33% for four-class response variable.

Cortez and Silva [9] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school students from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms such as Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Support Vector Machine (SVM) were applied on a data set of 788 students, who appeared in 2006 examination. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (*pass/fail*) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset.

From these specific studies, we observed that the student performance could depend on diversified factors such as demographic, academic, psychological, socio-economic and other environmental factors.

3. Research Methodology

The data mining research was based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) research approach. The open source software tool WEKA (Knowledge Flow Interface) was used for the research implementation. During the ***Business Understanding Phase*** the specific University management needs are identified. In the ***Data Understanding Phase*** the students' educational results at 200level and 500level with their corresponding biodata were collected for observation. During the ***Data Preprocessing Phase***, student data collected from the Department of Computer Science and Engineering at Ladoke Akintola University of Technology Ogbomoso, Oyo State, Nigeria, were organized in a new data mart.

The research sample includes data about 473students, described by 10 parameters (gender, age, mode of admission (MOA), religion, pre- degree score, student matriculation number, state of origin, 200 level CGPA, 500level GPA and student grade. The provided data is subjected to many transformations –removing parameters that are considered useless (e.g. fields with one value only),replacing fields containing free text with nominal variable (with a number of distinct values), transforming numeric to nominal variables, etc. The data is also being studied for missing values (very few and not important), and obvious mistakes (corrected).

The data mining task is to develop and validate a predictive model that predicts the students' university performance based on the student examination results. The target variable was the "student class", it was constructed as a categorical variable, based on the numeric values of the "student total university score'' attribute which has five distinct values - "First Class" (4.5-5.00), "Second Class Upper" (4.49-3.50), "Second Class Lower" (3.49-2.50), "Third Class" (2.49-1.50) and "Pass"(1.49- 1.00). The dataset contains 473 instances (10 classified as First Class, 188 classified as Second Class Upper, 182 classified as Second Class Lower, 64 classified as Third Class, and 30 classified as pass), each described with 10attributes (1 output and 9 input variables), nominal ,numeric and categorical.

During the ***Modeling Phase***, two different classification algorithms are selected and applied. Popular WEKA classifiers (with their default settings unless specified otherwise) are used, including a common decision tree algorithm (Decision Stump) and Bayesian classifiers (Naïve Bayes).

4. Tables, Figures and Equations

4.1 Table and Figure

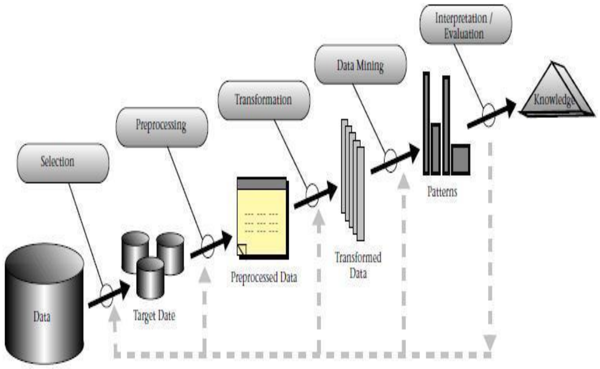

Fig. 1 The steps for extracting knowledge from data.

Table1: Performance evaluation of the classifiers

	1^{ST} CLASS		***2^{ND} CLASS UPPER***		***2^{ND} CLASS LOWER***		***THIRD CLASS***		***PASS***	
	NB	***DS***	***NB***	***DS***	***NB***	***DS***	***NB***	***DS***	***NB***	***DS***
TPrate	***1***	***0***	***0.6***	***0.9***	***0.8***	***0.8***	***0.8***	***0.9***	***0.6***	***0***
FPrate	***1***	***0***	***0.0***	***0.5***	***0.1***	***0.7***	***0.1***	***0.1***	***0.0***	***0***
PRN	***1***	***0***	***0.7***	***0.4***	***0.7***	***0.6***	***0.6***	***0.6***	***0.6***	***0***

N.B: Naïve Bayes D.S: Decision Stump TP Rate: True Positive Rate FP Rate: False Positive Rate PRN: Precision

4.2 Equations

$$Bayes\ theorem:\quad P(h|D) = \frac{P(D|h)P(h)}{P(D)} \qquad (1)$$

$$Naive\ Bayes\ Classifer: VNB = arg_{max}\ P(V_j)\prod P\left(a_i\ \ V_j\right) \qquad (2)$$

$$CART: \emptyset\left(\frac{s}{t}\right) = 2P_i\ P_R \sum_{j=1}^{\#\ classes} P\left(\frac{j}{t_L}\right) - P\left(\frac{j}{t_R}\right) \qquad (3)$$

4. Results and Discussions

The WEKA Knowledge flow application was used at this stage. Each classifier was applied for two testing options - cross validation (using 10 folds) and percentage split (2/3 of the dataset used for training and 1/3 – for testing). The results for the overall accuracy of the applied classifiers, including True Positive Rate and Precision (the average values for the 10-fold cross validation and split options) are presented in Table I. The results for the classifiers' performance on the five classes are presented on Table.1

The achieved results revealed that the Bayesian classifier (Naïve Bayes) performs best because it was able to predict for all the grades while the decision tree classifier (Decision Stump) cannot predict for First Class Grade and Pass Grade. On the average, the Naïve Bayes precision was above 77percent while Decision Stump precision was 57percent on the average.

Decision Stump inability to predict for First Class Grade and Pass Grade were due to insufficient dataset (training data and test data) and improper calibration of the cross validation fold maker.

Further research efforts will be directed at achieving higher accuracy of the classifiers' prediction by additional transformations of the student educational record, reconstruction of the target variable, tuning of the classification algorithms' parameters, working with larger dataset for training and testing etc.

5. Conclusion

Frequently used Decision tree classifiers and Bayesian classifiers are studied and the experiments are conducted to choose two classifiers for retention data to predict the student's educational performance. On working on student's performance, many attributes have been tested, and some of them are found effective on the prediction. The 200level CGPA was the strongest attribute and the 500level GPA had little effect on the student's final grade. Machine learning algorithms such as the Naïve Bayes and Decision Stump can learn effective predictive models from the student educational data accumulated from the preceding years, because the precision value for student performance was 77%. The practical results show that we can produce an accurate prediction list for the student performance, purposely by applying the predictive models to the records of the new set of 200level students. This study will also work to identify those students which need extraordinary attention to perform well in their discipline.

References

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial intelligence, pp. 37-54, 1997.

[2] Baker R.S.J.D., "Data Mining For Education. In International

Encyclopedia of Education (3rd edition)", B. MCGAW, PETERSON, P., BAKER Ed. Elsevier, Oxford, UK, 2009.Forman, G. 2003. An extensive Empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3, pp. 1289-1305, 2003.

[3] www.educationaldatamining.org.

[4] Y. B. Walters, and K. Soyibo, "An Analysis of High School Students' Performance on Five Integrated Science Process Skills", Research in Science & Technical Education, Vol. 19, No. 2, 2001, pp.133 – 145.

[5] Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream", Journal of Social Sciences, Vol. 1, No. 2, 2005, pp. 84-87.

[6] S. T. Hijazi, and R. S. M. Naqvi, "Factors Affecting Student's Performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

[7] Y. Ma, B. Liu, C.K. Wong, P.S. Yu, and S.M. Lee, "Targeting the Right Students Using Data Mining", Proceedings of KDD, International Conference on Knowledge discovery and Data Mining, Boston, USA, 2000, pp. 457-464.

[8] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques", Applied Artificial Intelligence, Vol. 18, No. 5, 2004, pp. 411-426.

[9] P. Cortez, and A. Silva, "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp.5-12.

[10] A. L. Kristjansson, I. G. Sigfusdottir, and J. P. Allegrante, "Health Behavior and Academic Achievement Among Adolescents: The Relative Contribution of Dietary Habits, Physical Activity, Body Mass Index, and Self-Esteem", Health Education & Behavior, (In Press).

[11] J. A. Moriana, F. Alos, R. Alcala, M. J. Pino, J. Herruzo, and R. Ruiz, "Extra Curricular Activities and Academic Performance in Secondary Students", Electronic Journal of Research in Educational Psychology,Vol. 4, No. 1, 2006, pp. 35-46.

[12] M. Bray, The Shadow Education System: Private Tutoring And Its Implications For Planners, (2nd ed.), UNESCO, PARIS, France, 2007.

[13] Q. A. AI-Radaideh, E. M. AI-Shawakfa, and M. I. AI-Najjar, "Mining Student Data using Decision Trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

A.Q Ayinde obtained his B.Tech (Computer Science) from LAUTECH (2008).He is a research student in the Department of Computer Science and Engineering, Ladoke Akintola University of Technology (LAUTECH) and a Lecturer in the Department of Computer Science at Osun State College of Technology, Esa-Oke. His research areas include Data Mining, ICT, Soft Computing and Database. He is a member of the following professional bodies Redhat Linux (RHCE Certified) Microsoft (MCITP Certified) and Information and Technology in Infrastructure Library (ITIL Certified).

A.B Adetunji obtained her B.Sc (computer Science) from University of Ibadan (1988), M.Sc (Computer Science) from Obafemi Awolowo University Ile- Ife (2002) and obtained her PhD in Computer Science from LAUTECH (2010).She is a Senior Lecturer, presently working in Computer Science and Engineering Department, LAUTECH. She is a member of Computer Professionals Registration Council of Nigeria (CPN). Her research areas include Database, Artificial Intelligent and Data Mining.

M. Bello obtained his B.Tech (Computer Science) from LAUTECH (2008).He is a research student in the Department of Computer Science and Engineering, Ladoke Akintola University of Technology (LAUTECH) Presently working as a Programmer in Mathematical Department at Kogi State University, Ayingba. His research areas include Data Mining, ICT, Artificial Intelligent and Database. He is a member of the following professional bodies Redhat Linux (RHCE Certified) and Microsoft (MCITP Certified)

O.A Odeniyi obtained his B.Tech (Computer Science) from LAUTECH (1996).He obtained his PGD in Education from National Teachers' Institute Kaduna (2006).He is a research student at the Department of Computer Science and Engineering in LAUTECH. He is a Lecturer 1 at the Department of Computer Science, Osun State College of Technology, Esa Oke. He is a member of the Computer Professionals Council of Nigeria (CPN) and Nigeria Computer Society (NCS). His research areas include Soft Computing, ICT, Artificial Intelligence and Database.