Mining Parent Socio-Economic Factors to Predict Students' Academic Performance in Osun State College of Technology, Esa Oke

Ayinde A.Q[1], Odeniyi O.A[2] and Sarumi O.A[3]
Osun State College of Technology Esa Oke, Osun State, Nigeria[1,2,3]

Abstract

The instability in the academic performance of students in tertiary institutions has been an herculean task over the years between the government and stakeholders in the educational sector. From investigation conducted at the Department of Computer Science at Osun States College of Technology, Esa Oke , it revealed that there is a coherent relationship between parent level of literacy, parent salary and family size in predicting the students' academic performance. This research carried out a deterministic relationship that can be used to measure the students' academic performance. This work was based on Cross Industry for Standard Model for Data Mining (CRISP – DM) and data mining techniques were used to investigate the relationship between socio-economic on the performance of students using the data from the Department of Computer Science, Osun States College of Technology, Esa Oke as case study. An incremental model was designed and the analysis was carried out using lazy classifiers (IBk, KStar and LWL) in training and testing the students' data collected from the computer science department. The academic performance of students was measured by the students' final year cumulative grade point average (CGPA). The applied lazy algorithms performance was evaluated in terms of TP Rate, FP Rate and precision.

1. Introduction

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 namely: Data Selection, Pre-processing, Transformation, Data Mining, Interpretation/Evaluation and Knowledge Extraction.

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 mine the relationship between parent socio-economic factors and students' final year grade by the application of lazy classifiers to educational data in Department of Computer Science and Engineering, Osun State College of Technology, Esa Oke, Osun 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.

Ayinde, Adetunji, Odeniyi and Bello [14] applied Naïve Bayes and Decision Stumps algorithms in predicting the students' final year grade. 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 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.

Adeyemo and Kuyoro[15] investigating the effect of students socio-economic/family background on students academic performance in tertiary institutions using decision tree algorithms.The The attribute usage shows that parents marital status, sponsor, mother's education and father's occupation rank higher than other socio-economic factors. The percentage usage for marital status is 100%, sponsor 97%, mother's education 88% and father's occupation 84% indicating that parental background and education sponsor contribute immensely to the academic performance of students. Age on entry, secondary school location and total SSCE score also shows more than 30% contribution to the performance of students. The Boost decision tree shows that all the attributes have considerable effect on the performance of students with the parental information and education sponsor ranking highest. All the attributes contributed at least 30% to the outcome of the student grade class. This shows that parental background and education sponsor are important factors to be considered for the good performance of students in tertiary institutions.

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 College management needs are identified. In the ***Data Understanding Phase*** the students' educational results at HND II with their corresponding parent socio-economic factors such as attributes states in the figure 1.. During the ***Data Pre-processing Phase***, student data collected from the Department of Computer Science at Osun State College of Technology Esa Oke, Nigeria, were organized in a new data mart.

The research sample includes data about 350students, described by 15 attributes (student age of entry, gender, mother's educational qualification, marital status of parent, father's occupation, mother's occupation, father's salary, mother's salary, family size, student position in the family, type of secondary school attended, residence location, HND II CGPA and sponsor). 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 model that predicts the students' academic performance based on the parent socio-economic factor. The target variable was the "student final year grade", it was constructed as a categorical variable, based on the numeric values of the "student total college score'' attribute which has four distinct values - "Distinction" (4.00-5.00), "Upper Credit" (3.00-3.99), "Lower Credit" (2.50-2.99), "Pass"(1.99- 2.49).

During the ***Modelling Phase***, three different classification algorithms are selected and applied. Popular WEKA classifiers (with their default settings unless specified otherwise).The strength of the three WEKA lazy classifiers such as IBK, KStar and LWL are applied to the data collected from the department of Computer Science.

4. TABLES

S/N	Attribute Name	Attribute Description	A.T
1	Age on entry	Student age on admission	continuous
2	Gender	Male or Female	Categorical
3	Mother's Educational Qualification	P,S,F.D,S.D, PHD,Professor	Categorical
4	Father's Eduactional Qualification	P,S,F.D,S.D, PHD,Professor	Categorical
5	Marital Status of Parent	M,D,S,W	Categorical
6	Father's Occupation	G.W, Private and Self employed	Categorical
7	Mother's Ocuupation	G.W, Private and Self employed	Categorical
8	Father's Present Salary	Low ,Average, High	Numerical
9	Mother's Present Salary	Low ,Average, High	Numerical
10	Family Size	Total No of children in the family	Numerical
11	Type secondary school attended	Private,state.federal	Categorical
12	Student position in the family	1st born,last born,only child and others	Categorical
13	Residence Location	Rural,Semi-Urban and Urban	Categorical
14	Students' HND II CGPA	Between 1.99 to 5.00	Categorical
15	Sponsor	Parent,Scholarship, self and others	Categorical

G.W=Government Work P=Primary, F.D= 1st Degree S.D= 2nd Degree M= Married S=Separated

D=Divorced W=Widowed.

In the model calibration, it was observed from the statistical analysis drawn from the pre-processed data that just three attribute has a tremendous effect on the students' academic performance which are parent salary, parent educational qualification and family size. The three attributes with the target attribute will be use to training the model to ensure that we have an optimal predictive accuracy of the applied lazy classifiers.

5. RESULTS AND DISCUSSION

Table 2: Using Parent Educational Qualification to Predict Students' Academic Performance

Grade	TP Rate	FP Rate	Precision
Distinction	IBK 0.567 KStar 0.783 LWL 0.856	IBK 0.864 KStar 0.984 LWL 0.657	IBK 0.564 KStar 0.453 LWL 0.621
Upper Credit	IBK 0.764 KStar 0.974 LWL 0.454	IBK 0.897 KStar 0.775 LWL 0.874	IBK 0.764 KStar 0.834 LWL 0.321
Lower Credit	IBK 0.657 KStar 0.213 LWL 0.756	IBK 0.756 KStar 0.453 LWL 0.934	IBK 0.657 KStar 0.564 LWL 0.432
Pass	IBK 0.875 KStar 0.9854 LWL 0.992	IBK 0.984 KStar 0.983 LWL 0.845	IBK 0.983 KStar 0.845 LWL 0.832

Table 3: Using Parent Salary to Predict Students' Academic Performance

Grade	TP Rate	FP Rate	Precision
Distinction	IBK 0.368 KStar 0.586 LWL 0.459	IBK 0.666 KStar 0.784 LWL 0.457	IBK 0.667 KStar 0.853 LWL 0.921
Upper Credit	IBK 0.764	IBK 0.697	IBK 0.969
	KStar 0.974 LWL 0.954	KStar 0.975 LWL 0.874	KStar 0.934 LWL 0.828
Lower Credit	IBK 0.958 KStar 0.818 LWL 0.859	IBK 0.756 KStar 0.453 LWL 0.934	IBK 0.877 KStar 0.864 LWL 0.772
Pass	IBK 0.978 KStar 0.959 LWL 0.698	IBK 0.984 KStar 0.983 LWL 0.948	IBK 0.683 KStar 0.945 LWL 0.983

Table 4: Table 2: Using Family Size to Predict Students' Academic Performance

Grade	TP Rate	FP Rate	Precision
Distinction	IBK 0.865 KStar 0.783 LWL 0.856	IBK 0.864 KStar 0.984 LWL 0.858	IBK 0.564 KStar 0.453 LWL 0.828
Upper Credit	IBK 0.864 KStar 0.978 LWL 0.854	IBK 0.897 KStar 0.975 LWL 0.974	IBK 0.767 KStar 0.939 LWL 0.928
Lower Credit	IBK 0.959 KStar 0.913 LWL 0.956	IBK 0.958 KStar 0.858 LWL 0.994	IBK 0.857 KStar 0.964 LWL 0.492
Pass	IBK 0.875 KStar 0.9854 LWL 0.972	IBK 0.984 KStar 0.983 LWL 0.945	IBK 0.783 KStar 0.948 LWL 0.932

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 lazy classifiers, including True Positive Rate and Precision (the average values for the 10-fold cross validation and split options) are presented in Table 2, 3 and 4. The results for the classifiers' performance on the four classes of grade are presented in the tables mentioned above.

The achieved results revealed that the IBk algorithm precision is highest with 97percent (Upper Credit) when using the parent salary for predicting the students' academic performance. KStar precision is highest for 96percent (Lower Credit) when using the family size for predicting the students' academic performance .Precision is highest with 98percent when using Parent Educational Qualification in predicting the students' academic performance. On the average the precision of each of the algorithms name IBk, KStar and LWL was 75percent, 83percent and 72percent respectively. The result of the algorithms precision revealed that the parent salary has the strongest effect in predicting the students' academic performance.

Further research efforts will be directed at achieving higher accuracy of the classifiers' by additional tuning of the target attribute, application of other incremental learning classifiers,the model can be train and test in batches to see if we can have a higher precision of the applied lazy classifiers.

.

6. CONCLUSION

This study identified parental conditions such as parents' education qualification, parents' salary and the family size as influencing socio-economic factors affecting the performance of students in tertiary institutions. This is due to the fact that students whose parents' have higher education qualification, parent salary (High income) and small family size performed best as revealed by the performance evaluation of the applied lazy classifiers and its was validated with the statistical analysis performed as a self verification mining in the data pre-processing stage. Majorly, it was observed that parents' with a low income and low educational qualification regardless of their residence location find it so difficult to finance the education of their children which led to their poor academic performance. It was vividly deduced that over 89percent of the students' came from an average home in which their parent are with low educational qualification and aveage monthly salary which when combined the monthly salary of both parent. Therefore, the parents socio-economic background determine to a very great extent the academic achievement and overall success of their children and parents are urged to liaise with government and stakeholders in educational sector ensure the provision of adequate resources that will enhance the academic performance of the students.

- *Ayinde A.Q is currently pursuing masters degree program in Computer Science at the Department of Computer Science and Engineering, LAUTECH, Nigeria. He is currently a lecturer in the Department of Computer Science, Osun State College of Technology, Esa Oke. His research area include Data mining, Data base and ICT, Soft Computing*

- *O.A Odeniyi obtained his B.Tech (Computer Science) from LAUTECH (1996). He held Post Graduate Diploma (PGD) in Education from National Teachers' Institute Kaduna (2006). He is presently a research student at the Department of Computer Science and Engineering, LAUTECH. He is a Lecturer I at Osun State College of Technology EsaOke. He is a member of Computer Professionals Registration Council of Nigeria (CPN) and Nigeria Computer Society (NCS).His research areas are soft computing, ICT, Data Mining.*
- *O.A Sarumi obtained his B.Sc (Computer Economics) from OAU Ile-Ife (2001).He is a research student in the Department of Computer Science and Engineering, Obafemi Awolowo University Ile-Ife (O.A.U) and he is a lecturer II in the Department of Computer Science at Osun State College of Technology EsaOke. He is a member of Computer Professionals Registration Council of Nigeria (CPN) and Nigeria Computer Society (NCS). His research areas are in soft computing, ICT and Database.*

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., BAK ER Ed. Elsevier, Oxford, UK, 2009.Forman, G. 2003. An extensive empiricalstudy 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 inScience Stream", Journal of Social Sciences, Vol. 1, No. 2, 2005, pp. 84-87.
[6] S. T. Hijazi, and R. S. M. 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 Psycholo gy,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 Decisio Trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University,
Jordan, 2006.
[14] Ayinde A.Q, Adetunji A.B, Odeniyi O.A and Bello M ''Performance Evaluation of Naïve Bayes and Decision Stumps algorithms in mining students' educational data. IJCSI International Journal of Comput er Science Issues, Vol. 10, Issue 4, No 1, July 2013 ISSN (Print): 1694- 0814 | ISSN (Online): 1694-0784
[15] Adeyemi A.B and Kuyoro S. O. 2010: Investigating the Effect of Students Socio-Economic/Family Background on Students Academic Performance in Tertiary Institutions Using Decision Tree Algorithms. Department of Computer Science, University of Ibadan.