International Journal of Computer Applications Technology and Research
Volume 5–Issue 11, 698-704, 2016, ISSN:-2319–8656

Data Mining Model for Predicting Student Enrolment in STEM Courses in Higher Education Institutions

Stephen Kahara Wanjau
Directorate of ICT
Murang'a University of Technology
Murang'a, Kenya

George Okeyo
Department of Computing
Jomo Kenyatta University of Agriculture and Technology
Nairobi, Kenya

Richard Rimiru
Department of Computing
Jomo Kenyatta University of Agriculture and Technology
Nairobi, Kenya

Abstract: Educational data mining is the process of applying data mining tools and techniques to analyze data at educational institutions. In this paper, educational data mining was used to predict enrollment of students in Science, Technology, Engineering and Mathematics (STEM) courses in higher educational institutions. The study examined the extent to which individual, socio-demographic and school-level contextual factors help in pre-identifying successful and unsuccessful students in enrollment in STEM disciplines in Higher Education Institutions in Kenya. The Cross Industry Standard Process for Data Mining framework was applied to a dataset drawn from the first, second and third year undergraduate female students enrolled in STEM disciplines in one University in Kenya to model student enrollment. Feature selection was used to rank the predictor variables by their importance for further analysis. Various predictive algorithms were evaluated in predicting enrollment of students in STEM courses. Empirical results showed the following: (i) the most important factors separating successful from unsuccessful students are: High School final grade, teacher inspiration, career flexibility, pre-university awareness and mathematics grade. (ii) among classification algorithms for prediction, decision tree (CART) was the most successful classifier with an overall percentage of correct classification of 85.2%. This paper showcases the importance of Prediction and Classification based data mining algorithms in the field of education and also presents some promising future lines.

Keywords: Classification; data mining; enrollment; educational data mining; predictive modeling; STEM

1. INTRODUCTION

Over the past decade there has been a rapid growth in the number of higher education institutions in Kenya. One of the remarkable facts about these institutions is the rapid growth in data and this educational data is expanding quickly without any advantage to the educational management. To date, they have become data rich but information poor.

Every year, these institutions conduct admissions of new students whereby, the admission process results in the recording of large amounts of data in databases and student files in the registries. However, in most of the cases, this data is not put in a form of improving its use and results in wastage of what would otherwise be one of the most precious assets of these institutions [1]. By applying the various data mining techniques on this data, higher education institutions can get valuable information and predictions for the betterment of the admission process [2].

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 settings which they learn in [3], [4]. In the recent years, there has been increasing interest in the use of data mining to investigate scientific questions within educational research [4]. Scholars in educational data mining have used many data mining techniques such as Decision Trees, Support Vector Machines, Neural Networks, Naïve Bayes, K-Nearest neighbor, among others to discover many kinds of knowledge such as association rules, classifications and clustering [6],[11],[14], [18].

One area that is increasingly getting the attention of scholars in educational data mining is enrollment management, particularly in STEM disciplines. This is because, preparing an educated workforce to enter Science, Technology, Engineering and Mathematics (STEM) occupations is important for scientific innovations and technological advancements, as well as economic development and competitiveness [7]. In addition, the growth in the number of University students enrolling in STEM fields of study does not keep pace with the STEM labour market demand and on the other hand, there is indication that high school graduates' interest in and readiness for STEM fields of study have been in the decline [8], [9]. Given the growing need to attract more high school graduates into these specific postsecondary areas of study, research devoted to understanding the influences on students' academic choices in regard to postsecondary STEM majors is becoming essential [8].

Lichtenberger and George-Jackson [9] study identified several factors that impact high school students' interest in STEM fields which may be categorized into three main themes: students' interests and motivations, academic qualification, and educational contexts. These three themes largely correspond with two of the factors included in Perna and Thomas' model; namely internal and school-level contexts [10]. While other factors also relate to high school students' interests in and ability to enroll in STEM courses, these three themes highlighted appear to be very influential on students' participation and success in STEM fields as they plan to transition from high school to higher education institutions.

Several studies in educational data mining have been conducted in an effort to address student enrollment [1], [8], [11], [12], [13], [14], [15], [18]. These studies have applied various data mining techniques on the data to get valuable information and predictions for the betterment of the admission process. However, there is a lack of a classification model that can be used to predict students' enrollment in Higher Education Institution, particularly in STEM courses.

Therefore, the main objective of this work was to build a classification model for predicting students' enrollment in STEM courses using data mining. Specific objectives were: to generate a data source of predictive variables, identify data mining techniques to study student enrollment in STEM, identify highly influencing predictive variables on the students' choice to enroll in STEM, and to evaluate the best classification algorithm.

The rest of the paper is organized as follows: Section 2 describes previous works on educational data mining in selecting student for enrollment. Section 3 describes the methodology used in this study. Section 4 presents the experiment conducted and results. Section 5 is a discussion of the findings and section 6 presents conclusion and recommendation for further works.

2. RELATED WORK

In the arena of educational data mining, there has been a recent surge in research paper and publishing. Several studies in the past have been conducted to investigate student enrollment using data mining techniques. Most notable of them are presented here.

In their work, Fong, Yain-Whar, Robert, and Aghai [11] used back-propogation algorithm and C4.5 algorithm for the student admission process. Their study proposed a hybrid model of neural network and decision tree classifier that predicts the likelihood of which University a student may enter, by analysing his academic merits, background and the University admission criteria from that of historical records.

The research conducted by Kovacic [12] presented a case study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success. The 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 accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively.

Moucary [13] carried out a study to find a reasonably accurate and reliable predictive tool that enables academicians (instructors and advisors) and administrators to decide about the enrollment of engineering students in Masters' studies or to succeed a Bachelor-of-Engineering program. The first objective of the study was aimed at discovering the relationship between the most affecting factors. Secondly, construct a predictive model that would endow both advisors and administrators with a powerful decision-making tool. The study used Matlab Neural Networks Pattern Recognition tool as well as Classification and Regression Trees (CART) with important cross validation and testing.

The research conducted by Padmapriya [14] applied data mining techniques to predict higher education admissibility of female students. The study focused on the development of data mining models for predicting student higher education admissibility by using two data mining algorithms for classification – a Decision tree algorithm and a Naive Bayesian Classifier. In this research, real data about 690 under-graduate students from Government arts college, Pudukkottai, India were used. The research is focused on the development of data mining models for predicting the students likely to go for higher studies, based on their personal, precollege and graduate-performance characteristics.

The study by Yadav and Pal [15] used data mining methodologies to select student for enrollment in a particular course. In this research, the classification task was used to evaluate previous student's performance. The decision tree method was used. Information like stream, marks in graduation, students performance etc. were collected from the student's management system, to predict the suitable student for enrollment in a particular course. The study findings showed that students past academic performance can be used to create the model using ID3 decision tree algorithm that can be used for prediction of student's enrollment in MCA course.

Gupta, Gupta, and Vijay [16] study explored the socio-demographic variables (age, gender, ethnicity, education, work status and disability) and study environment that may influence persistence or dropout of the students. They examined to what extent these factors, i.e. enrolment data help in pre identifying successful and unsuccessful students. Based on a data mining techniques such as feature selection, classification trees and logistic regression the most important factors for student success and a profile of the typical successful and unsuccessful students were identified. The empirical results showed that the most important factors separating successful from unsuccessful students are: ethnicity, course programme and course block. Among classification tree growing methods, Classification and Regression Tree (CART) was the most successful in growing the tree with an overall percentage of correct classification of 60.5%; both the risk estimated by the cross-validation and the gain diagram suggests that all trees, based only on enrolment data, are not quite good in separating successful from unsuccessful students, and the same conclusion was reached using the logistic regression. The case study was to build a data warehouse for a university student enrolment prediction data mining system. This data warehouse is able to generate summary reports as input data files for a data mining system to predict future student enrolment.

Priyanka and Ajit [1] study examined whether student's performance (past academic) can be used to construct a model using classification with a decision tree algorithm (ID3 and J48 decision tree algorithm). This study results helps students in selecting the course for admission according to his or her skills and academics.

San Pedro, Ocumpaugh, Baker, and Heffernan [17] study predicted student outcomes from their interactions with the ASSISTments system, a free web-based mathematics tutoring system for middle-school mathematics. The study developed a prediction model to distinguish whether or not students who attend college will enroll in a STEM major. The study developed a logistic regression model predicting STEM major enrollment from combinations of features.

3. METHODOLOGY

This study utilized the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology suggested by Nisbet, Elder and Miner [19]. The methodology breaks down a data mining project in six phases which allow the building and implementation of a data mining model to be used in a real environment, helping to support business decisions. The process is described as cyclic as shown in Figure 1.

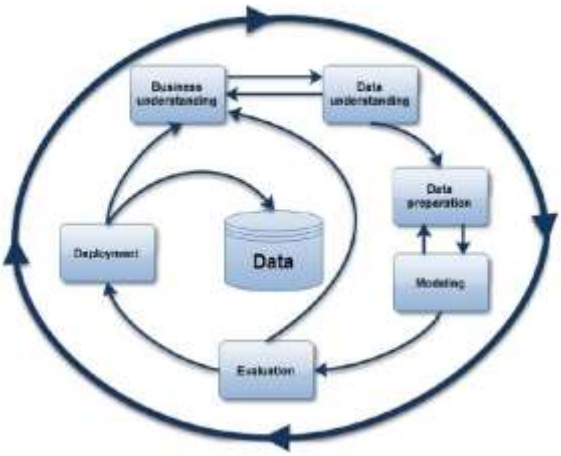

Figure 1: CRISP-DM Process Model for Data Mining (Adapted from Nisbet, Elder and Miner [33])

CRISP-DM has been adopted by some educational researchers as a method of choice for doing research in the newly emerged discipline of Educational Data Mining [4] [12]. A detailed explanation of the data mining process including model implementation and testing as applied in this study is presented in the next subsections.

3.1 Business Understanding Phase

This research was developed in the context of the Dedan Kimathi University of Technology, a Government sponsored University in Kenya. Studies at the University are organized into two academic semesters by year, each one spanning four months. Additionally, there is a four month semester, in which students break for long vacation, while others proceed for industrial attachment.

Every year, Universities in Kenya conduct admissions of new students. The admission process results in the recording of large amounts of data in databases while other enrollment data is stored in the student files in the academic registry. However, in most of the cases, this data is not put in a form of improving its use and results in wastage of what would otherwise be one of the most precious assets of these institutions.

3.2 Data Understanding Phase

The second phase is data understanding which begins with initial data collection. At this point, data collected from the respondents needs to be checked and understood. In order to be familiar with the data, the next step in data understanding is to identify data quality problem, get some insights about the data and detect interesting subsets to form hypotheses so as to uncover the hidden information within the data collected for the study [18].

The data set used in this study was collected through self-administered questionnaire survey at Dedan Kimathi University of Technology. The respondents used in this research were first, second and third year students enrolled in three undergraduate STEM courses in the School of Computer Science and Information Technology; two STEM courses in the School of Science, two STEM courses in the School of Engineering, and one STEM course in the School of Health Sciences at Dedan Kimathi University of Technology, Main Campus, during the second semester of 2016. The data set consisted of 232 instances. Taxonomy of predictive variables common to student enrollment in STEM courses is presented in Table 1.

Table 1: Taxonomy of Attributes of Student Enrollment in STEM Courses

Category	Predictor Variable / Attribute	Description
Student Interest and Motivations	Career Earning	Student's choice of STEM course is influenced by expected career earnings
	Career Flexibility	Student's choice of STEM course is influenced by expected career flexibility
	Self Efficacy	Student's choice of STEM course is influenced by one's belief in the ability to succeed in a STEM related Course
	Highest expected degree	Student's choice of STEM course is influenced by highest expected degree
Academic Qualifications and Preparations	High School Final Grade	Student's choice of STEM course is influenced by grade obtained in High School
	Subject Test Scores	Student's choice of STEM course is influenced by grade obtained in test scores in Science and Math subjects
	Career Awareness	Student's choice of STEM course is influenced by sensitization through forums such as career talks, career guidance
Educational Contexts	School Nature	Student's choice of STEM course is influenced by the nature of the school attended (Public or privately funded)
	Teacher Inspiration	Student's choice of STEM course is influenced by the former high school teacher
	Extra Curricular	Student's choice of STEM course is influenced by involvement on extra-curricular activities related to STEM e.g. Science Clubs
Socio-Demographic Descriptors	Family Income	Student's choice of STEM course is influenced by family's net income
	Family Size	Student's choice of STEM course is influenced by the family size
	Parent Career	Student's choice of STEM course is influenced by the career of the parent(s)
	Societal	Student's choice of STEM

Expectations	course is influenced by the societal expectations on the part of the students
Work Status	Student's choice of STEM course is influenced by the working status
Peer Influence	Student's choice of STEM course is influenced by peers
Financial Aid	Student's choice of STEM course is influenced by the ability to secure financial

3.3 Data Preparation Phase

3.3.1 Data Pre-processing

This stage involves preparing the data for analyses. Initially the datasets were collected in Ms Excel sheet and initial pre-processing was done manually by filling the missing values in the data set. Some irrelevant attributes were removed.

Feature selection was used as a method to select relevant attributes (or features) from the full set of attributes as a measure of dimensionality reduction. The objective of feature selection was to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information [20].

3.3.2 Data Transformation

This study used data mining software to investigate the most important variables that are associated with student enrollment and also to answer specific research questions. The open source software WEKA, offering a wide range of machine learning algorithms for Data Mining tasks, was used as a data mining tool for the research implementation. The WEKA workbench also contains a collection tools for data preprocessing, classification, regression, clustering, association rules, and visualization, together with graphical user interfaces for easy access to this functionality [21]. In this study, WEKA toolkit 3.6.13 was used.

The selected attributes were transformed into a form acceptable to WEKA data mining software. The data file was saved in Comma Separated Value (CSV) file format in Microsoft excel and later was converted to Attribute Relation File Format (ARFF) file inside WEKA for ease of analysis.

3.4 Modeling

During the modeling phase, modeling techniques were selected and applied to the dataset used in the study. This phase include selecting appropriate modeling technique, building the models and finally assessment of the model [12]. Subsequently, the model selection involves selecting appropriate techniques for the problem; refine the models whenever is necessary in order to meet the requirements [18].

The approach was to see how data mining techniques, more specifically classification techniques, could be used to determine whether the selected variables could predict enrollment in STEM. The study used WEKA to simulate the baseline performance accuracy of the classifiers in a more convenient manner to determine which algorithm is statistically better than the others. Classification algorithms for prediction were used namely; Decision Tree (CART), Naïve Bayes, k – Nearest Neighbor, Artificial Neural Networks(Multilayer Perceptron), Support Vector Machine (SMO) and Logistic Regression. These classifiers are available in the WEKA toolkit.

3.5 Evaluation and Deployment

In this phase, models were evaluated to assess the degree to which the model meets the business objectives and quality requirements [13]. This phase involves an iterative process of fitting different versions of models to training and testing data set, each time evaluating their predictive performance. The following metrics were used to determine the performance of the model: Time taken to build the model, Kappa statistics, mean absolute Error, Root Mean Squared Error, Relative Absolute Error, Prediction accuracy. The accuracy of the predictive model was calculated based on the percentage of total prediction that was correct.

Stratified 10-fold cross-validation evaluation model was used in the final analysis. In this method all the data was divided into 10 disjoint sets of equal size. Stratified 10-fold cross-validation (k = 10), also known as rotation estimation, is the most common [22] and universal [23] evaluation models, with lower sample distribution variance compared to the hold-out cross validation.

4. EXPERIMENT AND RESULTS

4.1 Attribute Selection

Two statistical methods were adopted to determine the importance of each predictor variable, namely, Chi-Square Attribute evaluation and Information Gain Attribute evaluation. To rank variables, Ranker Search method technique of WEKA was also applied. The output of the feature selection was a rank list of predictors according to their importance for further analysis.

Table 2: Attribute Ranking Using Chi-Square and Information Gain

S/No	Attribute	Chi-Squared		Information Gain	
		Value	Rank	Value	Rank
1	Career Earning	0.42282	10	0.001502	10
2	Career Flexibility	2.24527	3	0.007203	3
3	Self Efficacy	0.00428	14	0.000013	14
4	Final Grade Math	4.24799	1	0.015424	1
5	Grade	1.49757	5	0.005989	5
6	Physics Grade	1.32498	6	0.004730	6
7	Chemistry Grade	1.06611	8	0.003308	8
8	Pre university awareness	1.76538	4	0.006340	4
9	Teacher Inspiration	2.75538	2	0.009265	2
10	Extracurricular	0.227	12	0.000706	12
11	Family Income	0.25797	11	0.000801	11
12	Parent	1.17257	7	0.003944	7

13	Career Societal Expectation	0.04511	13	0.000140	13
14	Financial Aid	0.97523	9	0.003228 1	9

Among the fourteen attributes used in this study, it was discovered that final grade, teacher inspiration, career flexibility, pre-university awareness and mathematics grade are the best five attributes. These attributes outperformed other attributes in their contribution to the outcome of students' choice to enroll in STEM courses in higher education institution. A record 232 students was used as dataset. Table 3 shows the best attributes used and a short statistical summary for each of them. These attributes were then used for further analysis.

Table 3: Selected Attributes

S/ No	Attribute	Data Type	Possible Values
1	Career Flexibility	Nominal	{Yes, No}
2	Final Grade	Categorical	{A,A-,B+,B,B-,C+}
3	Math Grade	Categorical	{A,A-,B+,B,B-,C+}
4	Pre - University awareness	Nominal	{Yes, No}
5	Teacher Inspiration	Nominal	{Yes, No}
	CLASS (Response Variable)		{Successful, Unsuccessful}

4.2 Training Experiments and Results

A two-way random partition was done to generate a training set and a test and validation set. Training dataset was used to build the models. The number of instances we used for training was 151. Six Classification algorithms including; Decision Tree (CART), Naïve Bayes, k – Nearest Neighbor, Artificial Neural Networks (Multilayer Perceptron), Support Vector Machine (SMO) and Logistic Regression were used. Table 4 shows the results obtained from the models after training.

Table 4: Training and simulation error table

	Kappa Statistics	***Mean Absolute Error (MAE)***	***Root Mean Squared Error (RMSE)***	***Relative Absolute Error (RAE)***	***Root Relative Square Error (RRSE)***
Decision Tree (CART)	0	0.396	0.4453	99.3041 %	99.9944%
Naïve Bayes	-0.0588	0.3909	0.4712	98.0041%	105.8043%
Artificial Neural Network (MLP)	-0.0757	0.0385	0.5059	96.4273%	113.559%
k-Nearest Neighbor	-0.0912	0.3942	0.5165	98.839%	115.9807%
Support Vector Machine	-0.0242	0.284	0.5329	71.1983%	119.6631%
Logistic Regression	-0.0731	0.3835	0.482	96.1609%	108.237%

The result demonstrated that all the models scale down reasonably well as the instance number decreases. This implies a certain amount of available data is required to boot-start the classifiers.

The values of MAE and RAE for the Logistic Regression model tend the least when compared to the values of the other models. The values of RMSE and RRSE for the Decision Tree (CART) model tend the least when compared to the values of the other models. This result reveals that both the Logistic Regression algorithm and Decision Tree (CART) are more suitable for the prediction of students' enrollment in STEM since the lesser the error value the better the prediction.

4.3 Testing and Validation Results

The models obtained from the training data were rerun using the test and validation data sets to evaluate the performance of the resultant models. A total of 81 instances were used. Table 5 shows the results of this experiment.

Table 5: Comparison of Evaluation Measures (Confusion Matrix)

S/ No	Classifier	Correctly Classified Instances	Correctly Classified Instances
1	Decision Tree (CART)	78	3
2	Naïve Bayes	73	8
3	Artificial Neural Network (MLP)	72	9
4	k-Nearest Neighbor	76	5
5	Support Vector Machine	75	6
6	Logistic Regression	76	5

The results in table 5 shows that the number of correct positive predictions of the decision tree (CART) is higher compared to other models indicating that the decision tree (CART) model predicts students' enrollment in STEM cases better.

The study also evaluated the performance of the classifiers using the following metrics: Classification Accuracy, Miscalculation Rate, ROC Area (AUC), Precision and Speed. Table 6 shows the results obtained.

Table 6: Performance Measures of the Classifiers using 10-fold Cross Validation

	Accuracy	*Miscalculation Rate*	*ROC Area (AUC)*	*Precision*	*Speed*
Decision Tree (CART)	85.18	14.81	0.884	0.835	0.05 Sec
Naïve Bayes	75.30	24.69	0.571	0.701	0.02 Sec
Artificial Neural Network (MLP)	70.37	29.62	0.576	0.664	0.13 Sec
k-Nearest Neighbor	77.77	22.22	0.6	0.733	0 Sec
Support Vector Machine	76.54	23.45	0.512	0.682	0.02 Sec
Logistic Regression	75.30	24.69	0.544	0.701	0.02 Sec

5. DISCUSSION OF THE RESULTS

The results of the feature selection revealed that out of the original number of attributes, five of them, namely, final high school score, teacher inspiration, career flexibility, pre-university awareness and mathematics grade are the best five attributes in predicting enrollment in STEM courses in higher education institutions.

From the evaluation results shown in Table 6, it is interesting to note that, for the case study considered, the Decision Tree classifier (CART) achieved classification accuracy of 85.2%, with a precision rate of 83.5% and a ROC Area (AUC) of 88.4%. The response time of CART classifier was 0.05 Seconds. The k-Nearest Neighbor classifier achieved a classification accuracy of 77.8%, with a precision rate of 73.3%, and a ROC Area (AUC) of 60%. The response time of this classifier was 0 Seconds. The Support Vector Machine classifier achieved a classification accuracy of 76.5%, with a precision rate of 68.2%, and a ROC Area (AUC) of 51.2%.

Both Naïve Bayes and Logistic Regression classifiers achieved classification accuracy of 75.3%, with a precision rate of 70.1%. However, the Naive Bayes was found to have a better ROC Area (57.1%) than Logistic Regression classifier (54.4%). In terms of classification speed, Support Vector Machine, Naïve Bayes and Logistic Regression classifiers had a response time of 0.02 Seconds. The Artificial Neural Network (Multilayer Perceptron) classifier achieved a classification accuracy of 70.4 %, with a precision rate of 66.4%, and a ROC Area (AUC) of 57.6%. The response time of this classifier was 0.13 Seconds.

The number of correctly classified instances is often called accuracy or sample accuracy of a model. Hence in terms of the accuracy results, the decision tree (CART) prediction model performed better than the rest. These results support previous scholars' argument that predictive models deserve specific evaluation methods for its performance evaluation. However, from the table, it is seen that k-Nearest Neighbor takes the shortest time in building the model compared to others and the Artificial Neural Network (Multilayer Perceptron) takes a longer time.

6. CONCLUSION AND FURTHER WORKS

This study sought to build a data mining model to predict enrollment of students in STEM. The results have indicated that data mining techniques can be applied in predicting student enrollment in STEM courses in higher education institutions. Among all data mining classifiers considered, the Decision Tree classifier (CART) achieved the best classification accuracy of 85.2% and therefore proves to be potentially effective and efficient classifier algorithm. The study has also discovered that final high school score, teacher inspiration, career flexibility, pre-university awareness and mathematics grade are the best five attributes in predicting enrollment in STEM courses in higher education institutions.

Further investigation with regards to these attributes should be conducted covering more institutions. This work will be further improved by designing a predictive/recommender system based on the findings of this work. Additionally, there are a lot of different algorithms that have been developed in the different data mining techniques that need to be considered in the future.

Finally, next steps also include reviewing and understanding the implications of the predictive methodology on STEM enrollment policy, practices and technologies in a University span with a wide range of potential changes to business practices, policy concerns and practical implementation issues.

7. REFERENCES

[1] Priyanka, S., & Ajit, K. (2013). Prediction using Classification Technique for the Students' Enrollment Process in Higher Educational Institutions. *International Journal of Computer Applications, 84*(14), 37-41.

[2] Delavari, N., & Beikzadeh, M. (2008). Data Mining Application in Higher Learning Institutions. *Informatics in Education, 7*(1), 31-54.

[3] Karim, M., & Rahman, R. M. (2013, April). Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. *Journal of Software Engineering and Applications, 6*, 196-206.

[4] Kabakchieva, D., Stefanova, K., & Kisimov, V. (2011). Analyzing University Data for Determining Student Profiles and Prediciting Performance. *Conference Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), 6- 8 July 2011* (pp. 347-348). The Netherlands: Eindhoven.

[5] IEDMS. (2009). International Educational Data Mining Society.

[6] Sahay, A., & Mehta, K. (2010). Assisting higher education in assessing, predicting, and managing issues related to student success: A web-based software using Data Mining and Quality Function Deployment. Academic and Business Research Institute Conference, (pp. 1-12). Las Vegas.

[7] Sarala, V., & Krishnaiah, J. (2015). Empirical Study Of Data Mining Techniques In Education System. nternational Journal of Advances in Computer Science and Technology (IJACST), 15-21. Retrieved from http://www.warse.org/pdfs/2014/iccsie2015sp03.pdf

[8] Wang, X. (2013, September). Modeling Entrance Into STEM Fields of Study Among Students Beginning at Beginning at Community Colleges and Four-Year Institutions. Research in Higher Education, 54 (6), 664-669.

[9] Lichtenberger, E., & George-Jackson, C. (2013). Predicting High School Students' Interest in Majoring in a STEM Field: Insight into High School Students' Postsecondary Plans. *Journal of Career and Technical Education, 28*(1), 19-38.

[10] Perna, L., & Thomas, S. (2006). *A Framework for Reducing the College Success Gap and Promoting Success for All.* National Post Secondary Education Coperative. Retrieved from http://web.ewu.edu/groups/academicaffairs/IR/NPEC_3_Perna_Thomas_Report.pdf

[11] Fong, S., Yain-Whar, S., Robert, P., & Aghai, B. (2009). Applying a Hybrid Model of Neural Network and Decision Tree Classifier for Predicting University Admission. *IEEE*.

[12] Kovačić, Z. (2010). Early Prediction of Student Success: Mining Students Enrolment Data. *Proceedings of Informing Science & IT Education Conference (InSITE)* (pp. 647-665). Wellington, New Zealand: InformingScience.org.

[13] Moucary, C. E. (2011). Data Mining for Engineering Schools: Predicting Students' Performance and Enrollment in Masters Programs. *International Journal of Advanced Computer Science and Applications, 2*(10), 1-9. Retrieved from http://www.ijacsa.thesai.org

[14] Padmapriya, A. (2012, November). Prediction of Higher Education Admissibility using Classification Algorithms. *International Journal of Advanced Research in Computer Science and Software Engineering, 2*(11), 330-336. Retrieved from http://www.ijarcsse.com/

[15] Yadav, S., & Pal, S. (2012, March). Data Mining Application in Enrollment Management: A Case Study. International Journal of Computer Applications, 41(5), 1-6.

[16] Gupta, S., Gupta, S., & Vijay, R. (2013, March). Prediction Of Student Success That Are Going To Enroll In The Higher Technical Education. *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), 3*(1), 95-108.

[17] San Pedro, M., Ocumpaugh, J., Baker, R., & Heffernan, N. (2014). Predicting STEM and Non-STEM College Major Enrollment from Middle School Interaction with Mathematics Educational Software.

[18] Siraj, F., & Abdoulha, M. (2009). Uncovering hidden information within university's student enrolmentdata using data mining. MASAUM Journal of Computing, 1(2), 337-342.

[19] Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications.* Amsterdam: Elsevier.

[20] Goswami, S., & Chakrabarti, A. (2014). Feature Selection: A Practitioner View. *International Journal of Information Technology and Computer Science, 11*, 66-77. doi:10.5815/ijitcs.2014.11.10

[21] Sudhir, B., & Kodge, B. (2013). Census Data Mining and Data Analysis using WEKA. International Conference in "Emerging Trends in Science, Technology and Management, (pp. 35-40). Singapore.

[22] Payam Refaeilzadeh, L. (2009). Cross-Validation. *Encyclopedia of database systems*.

[23] Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40-79.