STUDENT ACADEMIC PERFORMANCE PREDICTION USING SUPPORT VECTOR MACHINE

Article · May 2019

1 author:

Akinode John Lekan
The Federal Polytechnic, Ilaro
13 PUBLICATIONS **134** CITATIONS

✚IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY
STUDENT ACADEMIC PERFORMANCE PREDICTION USING SUPPORT VECTOR MACHINE

S.A. Oloruntoba[1] ,J.L.Akinode[2]
Computer Science Department, the Federal Polytechnic,ILARO,Ogun State, Nigeria
Computer Science Department, the Federal Polytechnic,ILARO,Ogun State, Nigeria

ABSTRACT
This paper investigates the relationship between students' preadmission academic profile and final academic performance. Data Sample of students in one of the Federal Polytechnic in south West part of Nigeria was used. The preadmission academic profile used for this study is the 'O' level grades(terminal high school results).The academic performance is defined using student's Grade Point Average(GPA). This research focused on using data mining technique to develop a model for predicting student performance based on 'O' level results and their first 3 semester at each semester. Data preprocessing was done to remove the results of rusticated and expelled student .Results obtained by comparing SVM with other ML techniques such as KNN,Decision trees, linear Regression shows that SVM outperforms other ML algorithms. The parameters of the SVM algorithm(kernel) was also tuned to improve its accuracy and result obtained shows that the RBF kernel with penalty(C=100) performs best.SVM and RBF gave the highest training accuracy of 94% and 97% predicting accuracy which outperforms other state of the art ML technique like KNN,decision trees etc

KEYWORDS: Student Perfomance,Prediction,Data Minning,Grade Point Average,SVM.

I.   INTRODUCTION
Educational data mining is an emerging research area that aims at developing models for exploring wealth of data from educational information system. The application of Educational Data Mining(EDM) traversed the educational sector and has become the new trend in data mining and knowledge discovery in Databases(KDD) field. it focuses on discovering important patterns and discovering useful knowledge from the academic based information system[3]. These include course management systems(moodle,blackboard),admissions systems, registration systems and other systems that are used for managing students at different levels of education from secondary school to Universities. The process involve in data mining and Knowledge discovery databases(KDD) is depicted in figure 1.

The main focus of any higher institution is to improve decision making at the managerial level and to impart education. Sound prediction of students' success in high institutions is one of the basis for improving the quality of education . Students performance is an important and integral part in higher institutions. This is because the quality of education in universities is based on its excellent record of academic achievements. Predicting students performance has become a daunting task due to the large volume of data in educational databases[6].

Educational data mining describes a process used to extract useful information and patterns from a huge educational database[7]. The extracted information and patterns can be effectively used in predicting students performance.Hence,this will assist educators at different strata of academic institutions to provide effective teaching approach and subsequently enhance student's academic performance. Furthermore, educators could also monitor their students achievements. Similalry,students could improve their learning activities, allowing the administration to improve the systems performance.

[5] reiterates that management of academic institutions should focus more on the profile of admitted students, getting aware of the different types and specific students' characteristics based on the received data. Similarly, they should consider if they have all the student's data needed to analyze the students at entry point for decision making.[5] submit that attributes and prediction methods are the two main factors in predicting students performances.

There exist several tasks that are used to build a predictive model, these include classification, regression and categorization. However, most popular task to predict students performance is classification. There are several data mining algorithms that have been applied to predict students performance. Among the algorithms used are Decision tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machine[6].

This research focused on using data mining technique to develop a model for predicting student performance based on their 'O' level results and their first 3 semester CGPA (Cumulative Grade Point Average) to predict their final CGPA.Secondary data was obtained from a Federal Polytechnic in the South West of Nigeria between 2015 and 2016 session . Information's like student's High School grade(WAEC or O level) and CGPA were collected from the student's file, to predict the performance at the end of the semester. This paper investigates the accuracy of SVM for predicting student performance.

The paper is divided into five sections. The purpose and background of the conducted research work is presented in the Introduction. A quick review of the related research work is provided in Section 2, the methodology adopted for the research work is described in Section 3, the obtained results and the comparative analysis are given in Section 4. The paper concludes with a summary of the achievements and discussion of further work.
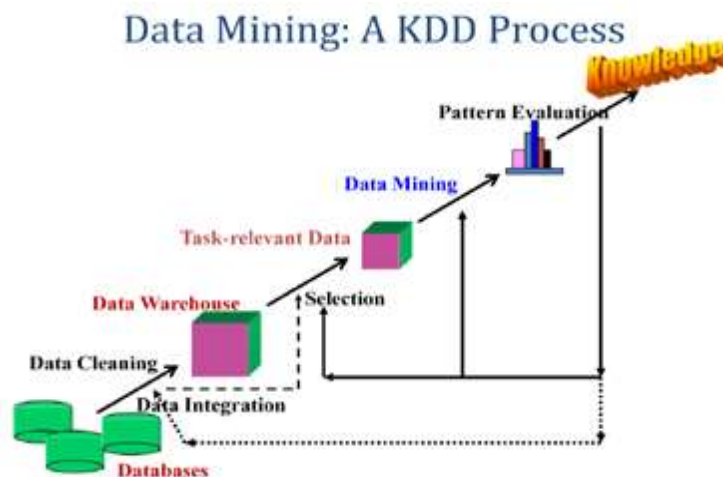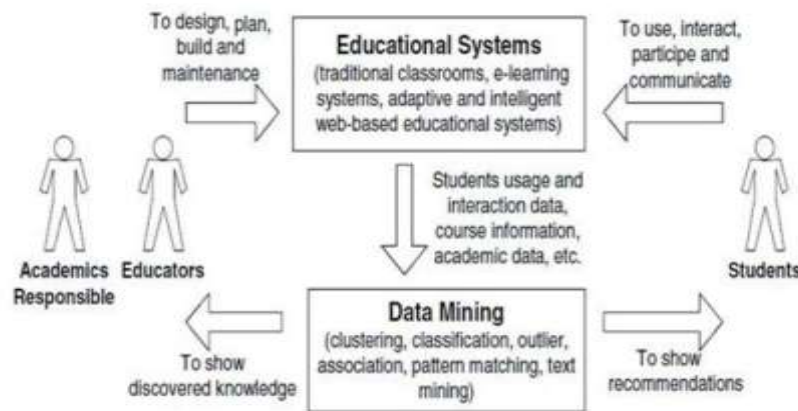

Figure 1:Data mining :A KDD process

Source: [22]

Figure 2. Educational Data Mining Cycle .Source:(Neha Choudary,2016)

II.   RELATED WORKS

In this section, we discussed the various approaches for predicting student performance and measurement using different data mining techniques, documented by various researchers.

[1] was arguably among the foremost authors who classified students by using genetic algorithms to predict their final grade.

[13] applied Support Vector Machine as their prediction technique in a small datasets.

Using the regression techniques,[2] predicted a student's marks (pass and fail classes) in their research work

 [14] in their paper, predicted a student's academic success (classified into low, medium, and high risk classes) using different data mining methods (decision trees and neural network).

[15] used a decision tree model to predict the final grade of group of student who took the C++ course in Yarmouk University in Jordan.

[25] compares different methods of data mining tools and techniques for classifying students on the basis of their Moodle usage data and the final score obtained in their corresponding courses. [24]   shows the enhancement of two machine learning model which are used to predict whether a student can answer correct question in an Intelligent Tutoring system

[16] using the decision tree ,predicted the result of the final exam to help professors determine students who needed help, in order to improve their performance and obtain pass grade.

[12] stated that Support Vector Machine has a good generalization ability and faster than other methods.

[4] applied three supervised data mining algorithms to assess the data of first year students to predict favorable outcome in a course and evaluating the performance based on certain factors like convienence,accuracy and approach of learning.[30] conducted an experimental survey to generate database for students for predicting the performance. The main focus is to identify the important predictive variables on higher secondary students, to determine the best algorithm and to predict the grade at higher education

The current research work by [28], compares two tools of data mining applied to data sets of small related to higher learning institutions and opined that the result will encourage higher education to inculcate data mining in their business processes. The author successfully predicted the success rate of students' enrolled.

RESEARCHERID

THOMSON REUTERS

ISSN: 2277-9655

[Oloruntoba S.A* *et al.,* 6(12): December, 2017]

Impact Factor: 4.116

IC™ Value: 3.00

CODEN: IJESS7

[18] demonstrated in their research the use of Support Vector Machine to predict risk of failing a course among student. In their work, Support Vector Machine method acquired the highest prediction accuracy in identifying students at risk of failing.

[26] developed and tested Support Vector Machine Algorithm and Multiple Linear Regression .The adapted methodology applied for data set of student enrolled in engineering. The result examined showed that SVM produce higher accuracy to identify the students having low grades.

[23] studied how the performance of the students evolves during their year of studies. For clustering, progression of students is used .Thus, students in the same cluster have same progression.

[29] also developed a prediction model that depend on the participation of students through genetic programming by including learning analytics and educational data mining

[27] developed a methodology to determine the future career of University graduate students. the research aimed to determine the strategy to improve the performance and scheduling of exams by using different data mining

[9] employed data mining algorithm to predict the student course selection . The research outcome submits that a students' grade point average relative to the grades of the courses they are considering for enrolment was the most important factor in determining future course selections

III. DATA MINING METHODS

Data mining(DM) described a computational method of processing data which is mostly used in many areas that aim to obtain useful knowledge from the data (Klosgen and Zytkow, 2002). The DM techniques are used to build a model according to which the unknown data will try to identify the new information. There are several data mining methods that are used in obtain hidden knowledge from vast amount of data. These include; Decision theory, Neural Network, Bayesian Classifier-Nearest Neighbor and Support Vector Machine.

Decision Tree

Decision Tree is one of a popular technique for prediction. The technique have been used extensively by most of researchers because of its simplicity and comprehensibility to uncover small or large data structure and predict the value. Decision Tree classifiers are used in data mining to produce trees after studying the training set and will be used to create predictions. Decision tree classifiers are one of the admired and influential tools for classification. Normally, decision tree classifiers have a tree-like structure which starts from root attributes, and ends with leaf nodes. It also has several branches consisting of dissimilar attributes, the leaf node on each branch representing a class or a kind of class distribution. Decision tree algorithms explain the relationship with attributes, and the comparative significance of attributes. The benefit of decision trees are that they characterize rules which could simply be understood and interpreted by users, do not need complex data preparation, and do well for numerical and categorical variables. The core algorithm for constructing decision trees called ID3.

Neural Network

Neural network is arguably one of the popular technique used in educational data mining. The advantage of neural network is that it has the potential to detect all possible interactions between predictors variables[10]) .Neural network could also do a complete detection without having any doubt even in complex nonlinear relationship between dependent and independent variables[17] . Therefore, neural network technique is selected as one of the best prediction method.

Bayesian Classifier

This is a simple classification method that is based on the theory of probability(the Bayesian theorem) [17]. It is referred to as naive because it simplifies problems relying on two important assumptions: it assumes that the prognostic attributes are conditionally independent with familiar classification, and it supposes that there are no hidden attributes that could affect the process of prediction. This classifier represents the promising approach to the probabilistic discovery of knowledge, and it showcase a very efficient algorithm for data classification.

K-Nearest Neighbor

The k-Nearest Neighbor algorithms (k-NN) organize objects based on the neighboring training examples in

the feature space. K-NN is a kind of instance-based learning, or lazy learning, where the function is only approximated nearby and the entire calculation is delayed in anticipation of classification. The main problem of k-NN algorithm is that its accuracy can be strictly ruined by the existence of loud or inappropriate features. Likewise, its accuracy becomes unfortunate if the feature balance are not reliable with their importance

Support Vector Machine(Svm)
SVMs are described as a set of related supervised learning techniques used for classification and regression [20]. They are member of a family of generalized linear classification. An important property of SVM is , SVM simultaneously minimize the empirical classification error and maximize the geometric margin. Thus,SVM is also known as a Maximum Margin Classifiers. SVM is based on the Structural risk Minimization (SRM). SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes.

IV. DATA DESCRIPTION

The data for the model were collected from the students' Registration File obtained at the Department of Computer Science of a Federal Polytechnic in Nigeria for 2015-2016 academic year for graduated students. After eliminating incomplete data, the sample comprised 89 students who were at the time of researche present at the practice classes.

The data collected was for 2015 to 2016 session for students who had graduated from the OND(Ordinary National Diploma). All the predictor and response variables are given in the table 1 below:

As input to the model,12 variables are used, the names and coding are shown in Table 1

Table 1: Student related variables

S/N	Variable Name	Information	Values	Data type
1	Eng	English Result of the High school terminal Exam	A, B, C, D, E, F	Integer
2	Math	Mathematics Result of the High school terminal Exam	A, B, C, D, E, F	Integer
3	Phy	Physics Result of the High school terminal Exam	A, B, C, D, E, F	Integer
4	Bio	Biology Result of the High school terminal Exam	A, B, C, D, E, F	Integer
5	Agric	Agriculture Result of the High school terminal Exam	A, B, C, D, E, F	Integer
6	Eco	Economics Result of the High school terminal Exam	A, B, C, D, E, F	Integer
7	CHE	Chemistry Result of the High school terminal Exam	A, B, C, D, E, F	Integer
8	A	1st Semester result of a four semester program	Nominal (0-4)	Float
9	B	2nd Semester result of a four semester program	Nominal (0-4)	Float
10	C	3rd Semester result of a four semester	Nominal (0-4)	Float

11	D	program		
		4th Semester result of a four semester program	Nominal (0-4)	Float

V.   METHODOLOGY

Globally, most institution of higher learning adopt grading system to estimate and decide the academic performance of students.Similarly,we have adopted the same approach for the analysis and measurement of performance .

Proposed Approach

The initial step is to collect the data set required for the research work. The methodology is applied to a factual data containing information about the graduated student at the Department of Computer Science at the Polytechnic. The work flow of this paper is shown in figure 4.
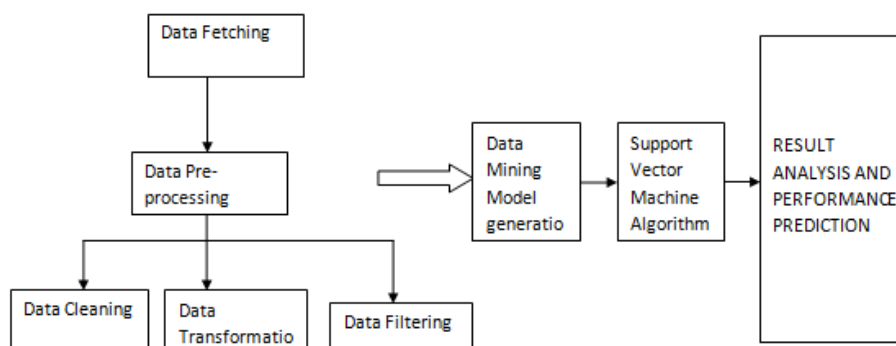

Figure 4:Work flow of Study

Once the data is obtained, it is transformed into required form for mining process, which is called pre-processing stage. it is an important step used in data mining process and it hinged on transforming the raw data into a proper format for resolving a particular problem..It has been discovered that the finer the pre-processing is done of the initial data, the more useful and suitable information is possible to discover.

Immediately, after the data is pre-processed ,we proceed to identify the incomplete ,incorrect and irrelevant data from our dataset and remove this erroneous and improperly formatted data..This phase is known as data cleaning phase. This process usually includes eliminating the typing errors or validating and correcting the valued of entities by cross checking it with accurate data set.

Once the data is complete and consistent in all respects, the next stage is to filter the data according to our requirement.

[23] highlights the importance of this step

1.   Data Fetching: Data fetching combines all the available data that can be used to resolve the data mining problem, into a set of instances

2.   Data Cleanning:At this stage, erroneous and irrelevant data are detected and discarded.

3.   Data Filtering: It helps to reduce the large amount of information available to us.

4   Data Transformation: This is the process of deriving new attributes from beforehand available attributes to assist in a better interpretation of information.

Algorithm Used

The proposed algorithm is the Support vector machine(SVM) which most suitable for small dataset.SVM is the newest technique for supervised learning. The SVM is used to carry out regression analysis on the ready data set.
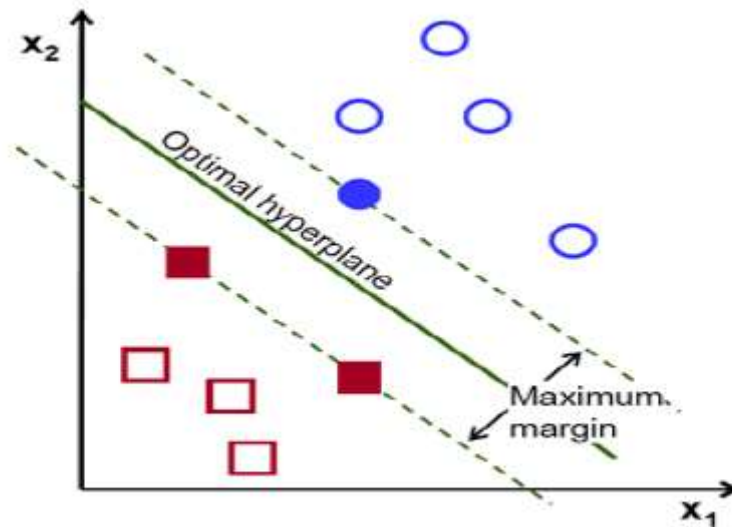
Figure5 Support Vector Machine [25]

SVM Algorithm:
    i.   Designate an optimal hperplane to maximize the margin
    ii.   Widen the above definition for non -linear separable problems
    iii.   Map the data to high dimensional space where it is simple to classify with linear decision and reformulate problem so that data is mapped completely to this space.

VI.   IMPLEMENTATION AND RESULT

Preprocessing of the data

Missing values of an attribute was cleaned by filling the value with average of fields on attribute. Meanwhile, the missing values that found at class involve that all field corresponding with that field of class was deleted. To simplify prediction process, GPA is used as the response variable.
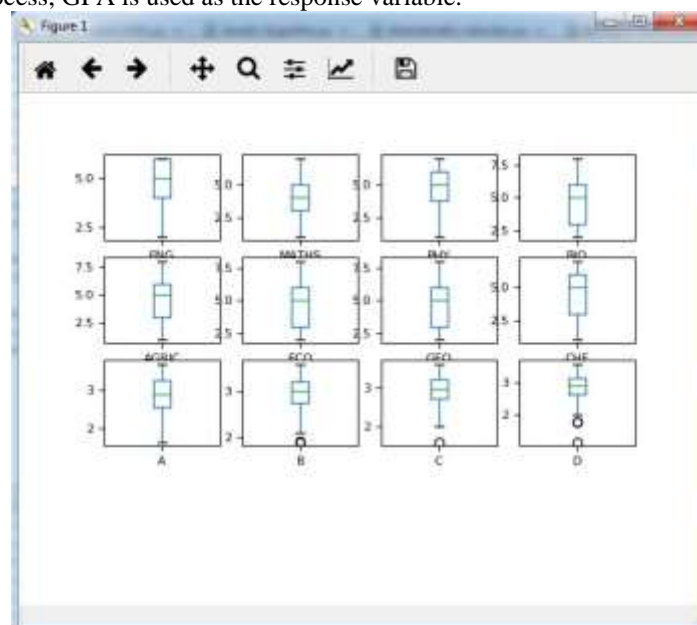

Figure 6: The Box plot of the sample data

Figure 7: Histogram of each attribute

The histogram shows that there is strong correlation between the attributes
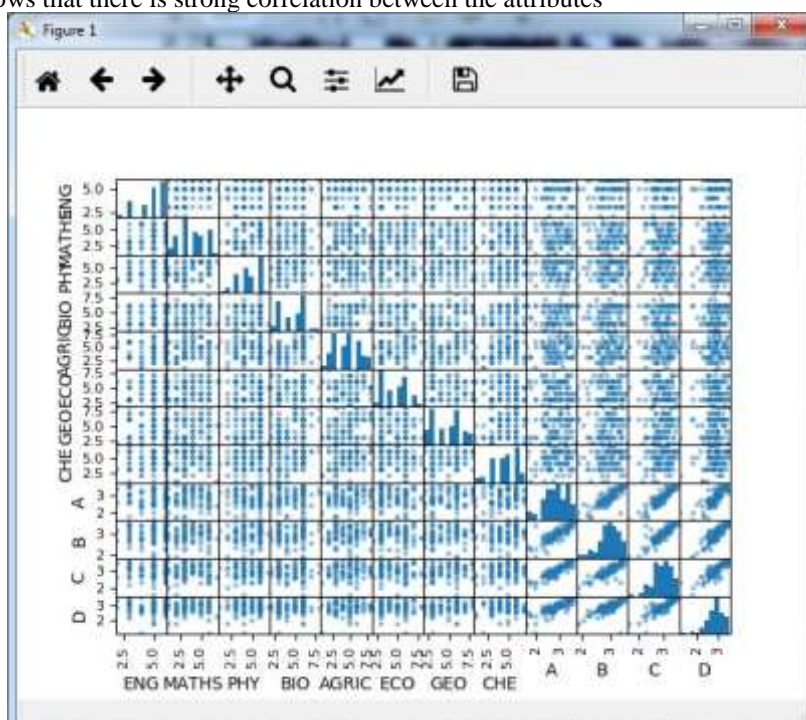

Figure 8: Scatter matrix

The scatter matrix also shows that there is strong correlation between the attribute

Experimental Result

The software chosen for the analysis was python Sklearn. Sklearn provides state of the art classification and regression algorithms such as linear Regression, K-nearest neighbor, Decision trees, Support Vector Regression. The dataset was divided into 70% training and 30% testing following[21].

Four regression algorithms were used to analyze the dataset. These include ; Linear Regression (LR), Lasso Regression (LASSO) and ElasticNet (EN), Classification and Regression Trees (CART), Support Vector Regression (SVR) and k-Nearest Neighbors (KNN).

All the algorithms use their default tuning parameters. Comparing the algorithms and displaying the Mean Squared Error (MSE) and standard deviation for each algorithm was shown below

S/N	Algorithm	MSE	Standard Deviation
1	Linear Regression	-0.042915	0.021999
5	Decision tree	-0.083867	0.051384
6	SVR	-0.170131	0.059279
2	LASSO	-0.157838	0.059085
3	Elastic Net	-0.157838	0.059085
4	KNN	-0.119799	0.038263

Table2: Comparative analysis of algorithms using MSE and Standard deviation

From the table above, SVR has the lowest MSE which shows the wrong predictions all the algorithms are performing (0 is perfect).

The parameters of the Support Vector Regression (SVR) are now tuned to improve the classification accuracy such as the penalty parameters (C) and kernel function. The percentage accuracy of the different kernels with C = 10, 100, 1000

C= 10

S/N	Kernel	Training	Testing
1	Linear	77	79
2	Poly	94	65
3	RBF	94	97

Table 3: Accuracy of different SVM Kernel when C=10

C= 100

S/N	Kernel	Training	Testing
1	Linear	78	75
2	Poly	94	51
3	RBF	94	98

Table 4: Accuracy of different SVM Kernel when C=100

C= 10000

S/N	Kernel	Training	Testing
1	Linear	45	49
2	Poly	94	51
3	RBF	94	96

Table 5: Accuracy of different SVM Kernel when C=1000

VI.   CONCLUSION AND FUTURE WORK

In this paper, Support Vector Machine data mining algorithm was applied to predict student success at the end of their study based on their 'O' level result and CGPA obtained at each semester Predicting student performance can be useful to the managements in many contexts. For identifying excellent students for scholarship programs, admissions, and to help level advisers to quickly identify students who are unlikely to graduate. From the results it is proven that Support vector regression algorithm remains the state of the art algorithm for predicting student performance. SVM gives 98% prediction for 89 instances which is relatively higher than other classifier and the MSE error rate is very low. For future work, an efficient method for tuning the performance of the algorithm such as particle swarm optimization could even provide better results

VII. REFERENCES

1. B. Minaei-Bidgoli ; D.A. Kashy ; G. Kortemeyer ; W.F. Punch,2003,Predicting student performance: an application of data mining methods with an educational Web-based system Frontiers in Education, 2003. FIE 2003 33rd Annual
2. S.B. Kotsiantis ; P.E. Pintelas, 2005,Predicting students marks in Hellenic Open University Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on
3. Amjad Abu Saa,2016,Educational Data Mining & Students' Performance Prediction Information Technology Department Ajman University (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016 212 | P a g e www.ijacsa.thesai.org
4. Edin Osmanbegović &, Mirza Suljić ,2012,Data Mining Approach For Predicting Student Performance Economic Review – Journal Of Economics And Business, Vol. X, Issue 1, May 2012
5. Dorina Kabakchieva ,2013, Predicting Student Performance By Using Data Mining Methods For Classification Cybernetics And Information Technologies
6. Amirah Mohamed Shahiri , Wahidah Husain , Nur'aini Abdul Rashid ,2015, A Review on Predicting Student's Performance using Data Mining Techniques, Procedia Computer Science 72 (2015) 414 – 422 Available online at www.sciencedirect.com
7. Angeline D. M. D., 2013,Association rule generation for student performance analysis using apriori algorithm, The SIJ Transactions on Computer Science Engineering & its Applications (CSEA) 1 (1) (2013) p12–16.
8. *Ankita Katare and Shubha Dubey2017,* A Study of various Techniques for Predicting student Performance under Educational Data Mining *International Journal of Electrical, Electronics* ISSN No. (Online): 2277-2626 *and Computer Engineering* 6(1): 24-28(2017)
9. Ognjanovic ,D. Gavesic and,S. Dawson , 2016,Using insttutional data to predict student course selections in higher education ,The internet and Higher Education,vol 29 ,pp 49-62
10. G. Gray, C. McGuinness, P. Owende, An application of classification models to predict learner progression in tertiary education, in: Advance Computing Conference (IACC), 2014 IEEE International, IEEE, 2014, pp. 549–554.
11. I. Hidayah, A. E. Permanasari, N. Ratwastuti, Student classification for academic performace
12. S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, E. Wani, 2011,Prediction of student academic performance by an application of data mining techniques, in: International Conference on Management and Artificial Intelligence IPEDR, Vol. 6, 2011, pp. 110–114
13. W. Hamal ainen, ¨ M. Vinni, 2006, Comparison of machine learning methods for intelligent tutoring systems, in: Intelligent Tutoring Systems, Springer, 2006, pp. 525–534.
14. Superby, J. Vandamme, J., Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006). Jhongli, Taiwan, pp37-44.
15. Al-Radaideh, Q., Al-Shawakfa, E. & AlNajjar, M. (2006), Mining Student Data Using Decision Trees, International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Available on: http://titania.addu.edu.ph/researches/DE CISION%20SUPPORT/Mining%20Student %20Data%20Using%20Decision%20Trees .pdf [pristup 10.januar 2012.]
16. Kumar S. A. & Vijayalakshmi M. N. (2011), Efficiency of Decision Trees in Predicting Student's Academic Performance, First International Conference on Computer Science, Engineering and Applications, CS and IT 02, Dubai, pp. 335-343.
17. P. M. Arsad, N. Buniyamin, J.-l. A. Manan,2013, A neural network students' performance prediction model (nnsppm), in: Smart Instrumentation, Measurement and Applications (ICSIMA), 2013 IEEE International Conference on, IEEE, 2013, pp. 1–5.
18. G. Gray, C. McGuinness, P. Owende, 2014,An application of classification models to predict learner progression in tertiary education, in: Advance Computing Conference (IACC), 2014 IEEE International, IEEE, 2014, pp. 549–554
19. Witten, I.H. & Frank E. (2000), Data Mining – Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann, San Francisco.
20. V. Vapnik.1995, The Nature of Statistical Learning Theory. NY: Springer-Verlag.

21. Ritesh Agrawal,2012 ,Dividing Data Into Training And Testing In R https://ragrawal.wordpress.com/2012/01/14/dividing-data-into-training-and-testing-dataset-in-r/ retrieved on 19th Dec.2017

22. Jiawei Han and Micheline Kamber,2006 ,Data Mining: Concepts and Techniques, 2nd ed. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6

23. R .Asif. A., and M.K Pathan, 2014, Predicting student academic performance at degree level: A case study,International Journal of Intelligient Systems and Applications vol 7 no1, pp. 49-61.

24. Mavrikis M,,2008,Data-driven modelling of student's interactions in an ILE In EDM,pp.87-96

25. Romero, C. Ventura ,P.G Espejo C,Hervas,Data,2008, Minning Algorithms to Classify Students,In EDM, pp.8-17

26. John M. Mativo& Shaobo Huang,2014, Prediction of students' academic performance: Adapt a methodology of predictive modeling for a small sample size", , vol. 00, no. , pp. 1-3, doi:10.1109/FIE.2014.7044287

27. Renza Campagni,Donatella Merlini, Renzo Sprugnoli,Maria CeciliaVerri ,2015,Data mining models for student careers,Expert Systems with Applications, Volume 42, Issue 13, 1 August, Pages 5508-5521

28. Srecko Natek a&, Moti Zwilling,2014 Student data mining solution–knowledge management system related to higher education institutions, Expert Systems with Applications 41 (2014) 6400–6407

29. Xing Wanli , Guo Rui, Petakovic Eva , Goggins Sean,2015 Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory, Computers in Human Behavior 47 (2015) 168–181

30. V.Ramesh, P.Parkavi & K.Ramar,2013, Predicting Student Performance: A Statistical and Data Mining ApproachInternational Journal of Computer Applications (0975 – 8887) Volume 63– No.8

CITE AN ARTICLE