Predicting Student's Performance in Education using Data Mining Techniques

Sara Fatima
Lecturer
Department Information Systems
King Abdulaziz University

Salma Mahgoub
Assistant Professor
Department of Information System King Abdulaziz University

ABSTRACT
In this data world, where users spawn their digital footprint and generate a huge amount of unstructured data continuously with each activity, data mining techniques help in discovering interesting patterns, establishing relationships and unravel the problems through analysis, in different aspects of life. Educational data mining is a multidisciplinary research area, in which data from various educational organizations, is explored and made operational, for various facets concerned with the students, like predicting academic performance, analyse the learning pattern, solving e-learning issues, predict employability, visualize the critical courses affecting performance, investigate the reasons for student's failure or drop out and thus make data-driven decisions to improve the institutions standards. This paper provides a brief overview of Data Mining tools and techniques, and its encroachment in the educational domain. It also proposes a simple framework using different variables which helps in predicting student's academic success using two different algorithms: Decision Trees and Bayesian Network. Finally, a comparative analysis of accuracy is done. The results show that Bayesian Network outperforms the Decision Tress and gives better accuracy.

General Terms
Data mining, classification algorithms, cluster analysis, predict

Keywords
EDM, Decision Trees, Higher Education (HE)

1. INTRODUCTION
Data mining is a collection of techniques based on advanced analytical methods for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so that they may be used in an enterprise decision making (G.K. Gupta, 2014). Data mining is also known as "Knowledge discovery in databases" (KDD). It is a continuously emerging and developing discipline, which has its deep roots taken from the statistics, artificial intelligence and database systems and thus, can be said as the intersection result of these three scientific sectors. In a broader sense, data mining is an attitude stating that business events should be based on learning that informed decisions are better than uninformed decisions, and that measuring outcome is advantageous to business (Berry & Linoff, 2004). Huge and vast amount of data which is collected from the real-life areas like business, finance, e-commerce, social networks, healthcare, banking, astronomy, biology, are explored sufficiently and mined to get meaningful information. Data mining is thus a promising and upcoming field which has attracted researchers and entrepreneurs down the years and opened the gateways of huge business opportunities, due to the eminent and massive data generated and it's requisite to change this data into valuable information. Data mining can also be viewed as a result of the natural evolution of information technology (Jiawei Han & Kamber, 2000). Fig 1 shows the evolution of database system technology.

Data mining has been applied to serve the various purposes like prediction and description, relationship marketing, customer profiling, customer segmentation, outlier's identification and fraud detection, website design and promotion, education, financial data analysis, telecommunication industry and many more sectors. Specific techniques are used for extracting novel and useful information from huge chunks of data like Association and Correlation rules, Supervised Classification, Cluster Analysis, Outlier Analysis and more.

2. DATA MINING PROCESS:
Data mining or KDD as a process consist of an iterative sequence of following steps (Han & Kamber, 2006):

- Data cleaning (to remove noise and inconsistent data)
- Data integration (where multiple data sources may be combined)
- Data selection (where data relevant to the analysis task are retrieved from the database)
- Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
- Data mining (an essential process where intelligent methods are applied to extract data patterns)
- Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
- Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

3. DATA MINING TOOLS:
Abundant tools are present for data mining tasks using artificial intelligence, machine learning to extract novel and useful information. Below are the list of few open source softwares that are widely used for data mining.

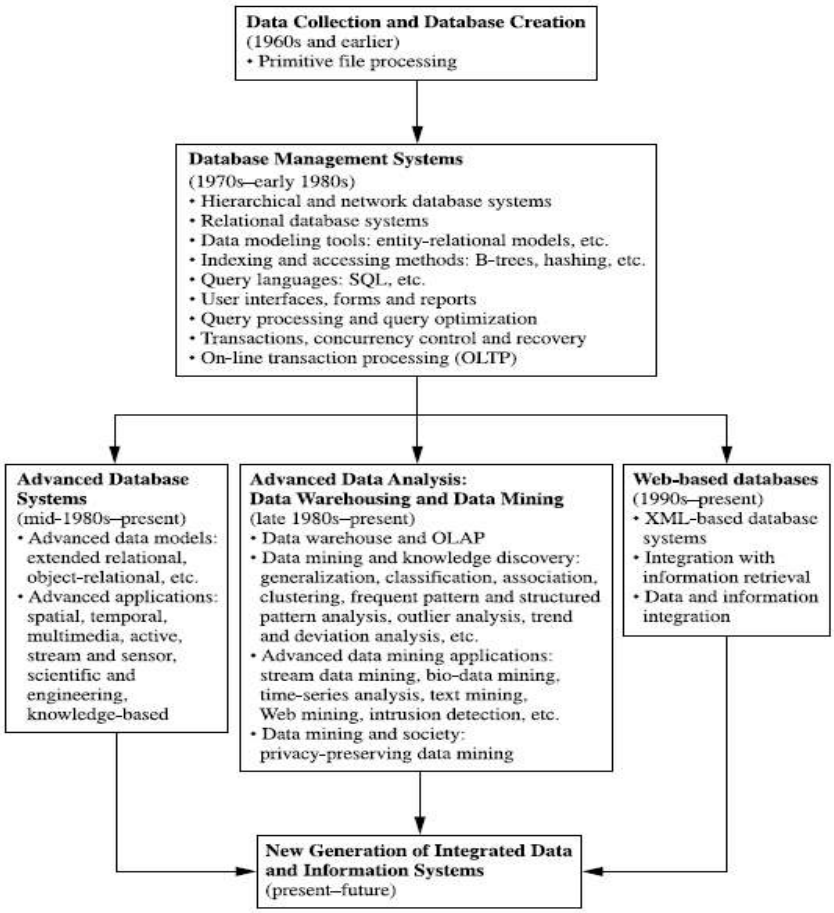

Fig 1: Evolution of Database System Technology

- RapidMiner: Formerly known as YALE, RapidMiner is written in Java Programming Language, which is offered as a service than a piece of software. It provides functionality like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation and deployment.
- WEKA: This software gives the user the liberty to customize as he/she wants, which makes it more favorable. Its Java-based version is more sophisticated and is used in different applications including visualization and algorithms for data analysis and predictive modeling.
- R-Programming: Its widely used among data miners for its more statistical and graphical features apart from various techniques which include linear and non-linear regression, classical statistical tests, time-series analysis, classification, clustering and others.
- Orange: It's the best tool for Python developers, as Orange is a Python based powerful tool with an additional eye for bio-informatics and text mining.
- KNIME: Written in Java and based on Eclipse, KNIME has attracted business intelligence and financial data analyst for its best graphical user interface and its easy extendable plugin feature.
- NLTK: It provides the pool for language processing tools like data mining, machine learning, data scraping, sentiment analysis and various other language processing tasks.

4. DATA MINING FUNCTIONALITIES:

- Prediction: This technique of data mining is used to foretell the future trends for the purpose of business intelligence. It thus helps in making better decisions. Linear Regression, Multivariate Linear Regression, Nonlinear Regression, Multivariate Nonlinear Regression are few techniques used.
- Classification: Classification is a data mining functionality that allocates items in the training data set to specific categories or class. It involves of predicting a certain outcome based on a given input. The classification algorithms then predict the outcome, by determining the relationship between the attributes. Different types of classification models are Decision Trees, Bayesian Classification, Neural Networks, Support Vector Machine (SVM), Classification based on Association.
- Clustering: Clustering divides the data into groups (clusters) based on their characteristics and similarities. Cluster analysis divides the data into clusters that are meaningful, useful or both. Clustering is applied in many arenas like pattern recognition, image analysis, information retrieval and more. Partitioning Methods, Hierarchical Agglomerative (divisive) methods, Density based methods, Grid-based methods, Model-based

methods are few examples.

Outlier Analysis: Outlier is a data point that is significantly different from the remaining data. An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. (Hawkins, 1980)

5. DATA MINING APPLICATIONS:

Data mining is practically applied in many fields like sales/marketing to better plan and organize the work so that it is more profitable. Banking/Finance sector use various tools of DM for credit analysis, fraudulent transactions, customer segmentation, cash management, and more. Health Care industry is able to find out high-risk patients, hospital rankings, diseases predication, at an earlier stage using diverse applications of DM. Students and Alumni challenges can be addressed effectively and further student's success/failure can be administered. Telecommunication industry uses DM methods to detect scams and enhance their efforts while working in highly changing and competitive environment. DM helps the retail industry by analyzing the customer's behavioral patterns, shopping trends, market volatility and thus facilitate in creating loyal customers for the company and generate huge profits.

6. DATA MININIG AND EDUCATION:

According to Journal of Educational Data Mining, EDM is defined as 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 in which they learn. It refers to techniques, tools and research designed to for automatically extracting meaning from large repositories of data generated by or related to people(student) learning activities in education settings (Nithya, Umamaheswari & Umadevi, 2016). Fig 2, shows the evolution of EDM from 1995 till 2022. Use of internet for education down the years has shaped a perspective and ever-expanding arena called e-learning or web-based education, thus generating huge chunks of data. In totality, they provide gigantic data sets, which can be used to extract knowledge and information. EDM can be considered as a helpful and facilitating tool not only to predict the student's success ratio but also to determine and dig up those factors which are the causes of student's failure. It can further assist the educational institutions to take on go for academically weak students and increase their concentration on them. Personal learning environments (PLEs) and personal recommendation systems (PRS) also directly relate to educational data mining. Personalized learning environments focus on providing the various tools, services, and artifacts so that the system can adapt to student's learning needs on the fly (Mödritscher, 2010). Data mining for an educational system is an iterative process for hypothesis development and testing. Fig.3 shows the application of data mining in an education system. Student's performance evaluation system can help in decision making for awarding scholarships or in other words targeting the right student. (Daud, Aljohani, Abbasi, Lytras, Abbas & Alowibdi, 2017)

Fig 2: Timeline of Significant Milestones in EDM (Courtesy: Baker & Inventado, 2014)

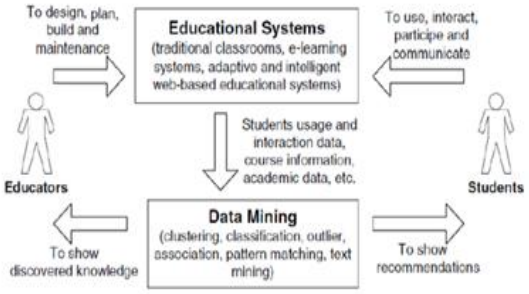

Fig 3: Cycle of Applying Data Mining in Education (Courtesy: Romero & Ventura, 2006)

6.1 Advantages of EDM in Higher Education (Kolo, Adepoju & Alhassan, 2015):

- Accurate analysis and predicting of data
- The ability to provide feedback for teachers/instructors
- The ability of predict student performance
- The production of student functional model
- The ability to detect undesirable student behavior
- It helps to group students into classes based on various performance characteristics
- The ability to analyze student's social network
- Developing concept maps
- It also aids in construction of curriculum
- It helps in educational planning and administration

6.2 Goals of EDM (Baker &Yacef, 2009):

- Predicting student's future learning behavior: Student's academic performance can be predicted by creating required models, using specific parameters like origin, knowledge level, motivations, likes, dislikes and more.
- Discovering and improving domain models: Using different methods and applications of EDM,

progress and enhancement of different models is possible. Ideal instructional sequences can be determined to engage the students and thus get to know more about their learning styles.

- Studying the effects of educational support: This goal is achieved through different learning systems.
- Advancing scientific knowledge about learning and learners: This aim is accomplished by structuring and incorporating various student's model, tools and software in EDM research.

6.3 Overview of popular EDM concepts & techniques:

- A framework was proposed to administer the student's academic performance using Learning Analytics techniques, which was effectively applied on the Secondary data collected from Department of Computer Science, University of Jos, Nigeria. Linear Regression was used to aid the Statistical Package for Social Sciences (SPSS) analysis tool. Hypothesis testing was then used to validate the model. The model provided with a 5% significance result. The limitations which were observed was the non- availability of primary data and central data repositories. Inconsistencies in determining which student's attributes contributes to the academic performances was also noted. (Oyerinde & Chia, 2017)
- Another research work showed the prediction of student's success factor, by collecting the scholarship holding student's data from different universities in Pakistan (on the data set of 690 students during the period of 2004 to 2011), which included the most important parameters like family income, family assets, family expenditure and student's personal information. Learning analytics and discriminative and generic classification are applied to the data set and results are obtained, which indicated that the above attributes are important predictors and has achieved F1-score of 86% ((Daud, Aljohani, Abbasi, Lytras, Abbas & Alowibdi, 2017). The aim of another study was to examine different formats of comment data (which the student gives after every lesson) to predict student's performance, using Latent Dirichlet Allocation (LDA) and Probability Latent Semantic Analysis (pLSA). A model was created using learning analytics (LA) and data mining approaches, which would use free-style comment data written by student after every class, and thus finally predict students' performance. The comment data was used to potentially eliminate barriers between students and their teachers. (Sorour, Goda & Mine, 2017).
- At the research conducted at King Saud University, KSA, the classification technique of data mining was used to predict students' performance. ID3 (Iterative Dichotomiser) decision tree algorithm was used to build the model, on records of 100 students at the university. The study also identified key courses which could be used as performance indicators. In ID3 algorithm a decision tree is constructed in top-down recursive approach. The training data set is iteratively split into smaller partitions. In each iteration ID3 decides which attribute best splits the data set. The splitting attribute selection is based on maximizing information gain. ID3 algorithm is simple and can be easily converted into understandable rules. (Altujjar, Altamimi, Turaiki, Razgan, 2016)
- In a study conducted on a group of students enrolled in different colleges in Ajman University of Science and Technology (AUST), UAE, the data set consisted to 270 records, which was collected anonymously through an online survey using Google forms. Multiple decision tree techniques and algorithms were used to and accuracy of the results were observed. CART (Classification and Regression Tree), CHAID Decision Tree, C4.5 and ID3 (Iterative Dichotomiser) were the different classification algorithms used for measuring the accuracy. CART showed the maximum accuracy in results (40%), followed by CHAID (34.07%) and C4.5 (35.1%). ID3 (33.33%) on other hand showed least precision in the results. The limitations which were highlighted during the research was the non-availability of central data repository for the student's data set. (Abu Sa, 2016)
- Objective of one of the research work was to show how prediction algorithm can be used to identify the most important attributes in the student's data set, at Universiti Sains Malyasia. Among the algorithms that were used Neural Network proved to have highest prediction accuracy (98%), followed by Decision Trees (91%). Support Vector Machine and K-Nearest Neighbour showed equal accuracy (83%) and Naïve Bayes demonstrated the least prediction accuracy (76%). Figure 4 illustrates the prediction accuracy grouped by algorithms since 2002-2015. (Shahiri, Husain & Rasid, 2015).
- An investigation was done to find out gap between Industry expectations and Institute produce of Technical Students. The paper identified parameters/skills on which the employability of engineering graduates' students depends and correlation of parameters with the statistical tool Chi Square Test in SPSS (Software Package of Social Sciences). A logistic regression mathematical model was developed which predicts the probability of Employability in campus placements, based on skills sets required in IT sector. Seven Parameters and 41 skills like aptitude, communication, technical expertise, personality and many such, were identified and test was conducted on the data set 2000 students initially, and then filtered to 362. The outcome shows that aptitude factor has the largest impact whereas personality factor has the least impact on the probability of employment. (Kalbande, Handa, 2015)
- Another case study based research was conducted at Universiti of Malaysia Pahang. It read data of 1000 students was collected to see the relationships between behavior of students and their academic performance. The variables used were Interest, Study Behavior, Engage Time, Believe and Family Support. The aim of the study was to apply the Kernel method as data mining technique to analyse the relationship between the selected predictors and student's academic performance. This is done by

using SSVM (Smooth Vector Machine) classification and Kernel K-Means Clustering. The study results showed a strong correlation between the mental condition of the student and their final academic performance, indicating that the psychometric predictors that were chosen, gave good contribution in predicting student's academic performance. (Sembiring, Sembiring, Hartama & Wani, 2011)

- The aim of another study was to examine different formats of comment data (which the student gives after every lesson) to predict student's performance, using Latent Dirichlet Allocation (LDA) and Probability Latent Semantic Analysis (pLSA). A model was created using learning analytics (LA) and data mining approaches, which would use free-style comment data written by student after every class, and thus finally predict students' performance. The comment data was used to potentially eliminate barriers between students and their teachers. Two limitations where highlighted in the study. Firstly, the study required a deeper understanding of those factors which would affect the student's academic performance like, students' attitude, skills and effort. Secondly, the work could not extract the common attributes from the students' comment data for more effective feedback. (Sorour, Goda & Mone, 2017)

7. PROPOSED FRAMEWORK

The proposed framework in Fig 5, consists of five major steps. The initial step is to gather the unstructured data. Subsequently, data cleaning process has to be followed, which involves omitting inaccurate, incomplete and unreasonable data, thus obtaining the structured data. On the resulting training data set, R tool is used to do data preprocessing, different attribute selection and classification algorithms are applied. Comparative analysis of efficient classification algorithm is thus done to predict the student's academic performance.

7.1 Research Steps:

- Data Gathering: The data under study was collected from the Student Unit Department and was then transferred on to an excel sheet to make it more efficient. The student's data set consisted of the following attributes. Gender(Gen), Father Education(FEdu), Father Occupation(FOcu), Mother Education(MEdu), Mother Occupation(MOcu), Medium of Education in High School(MEhs), High school CGPA(HCGPA), Previous Semester Attendence(PSat), Previous Semester GCPA(PCGPA), Score in SE course (SSE). The data set consisted of 165 students record for the current year.

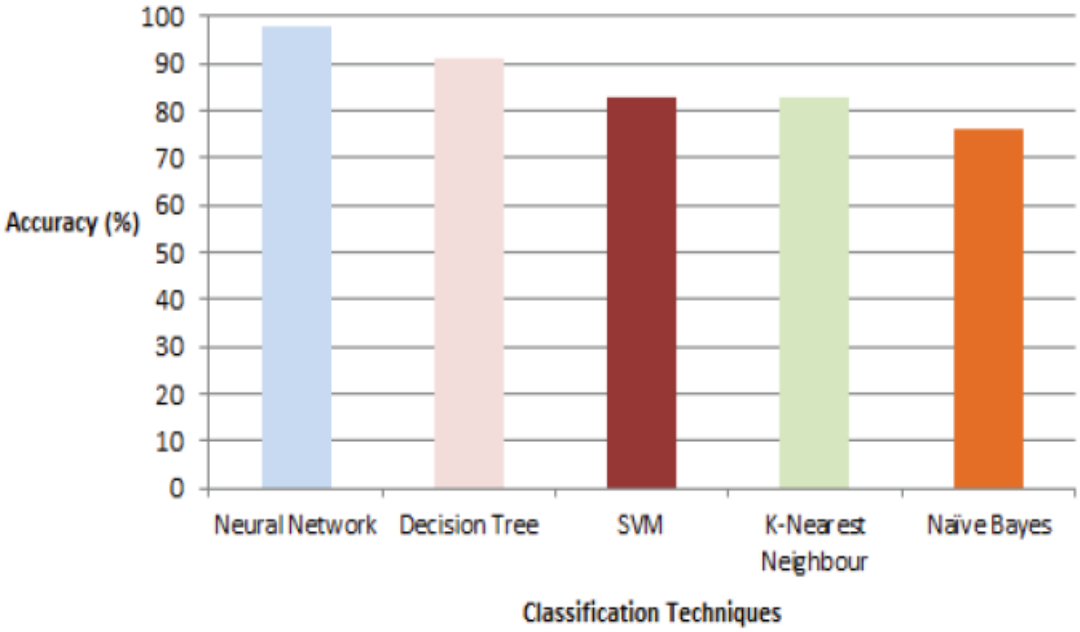

Fig 4: Prediction accuracy using different algorithms

Fig 5: Proposed framework

- Data Cleaning: This is an essential process in data science. Data cleaning removes the missing and unwanted values and makes the prediction process easier and efficient. After the data cleaning process, the resultant data set size was reduced to 108.
- Feature Selection: Feature Selection plays an important role in improving the model by removing those features from the data set which are uncorrelated and non-redundant. It further helps in perking up the model training and also reduces the complexity and makes the process of intercepting results simple. In this study, various features of R software was used to do feature selection. From the list of 10 attributes, 6 were included for analysis, which are Gen, MEhs, HCGPA, PSat, PCGPA, SSE.
- Analysis of Classification Algorithm: In this study, the algorithms which are used are: Decision Trees and Bayesian Network. Initially, all the attributes are used to check the accuracy rate. Later few selected attributes were used (based on the result of the feature selection step) to recheck the accuracy. The results are summarized in the table below

S.No	Attributes Selected	Algorithm used	Accuracy achieved
1	ALL	Bayesian Network	93.22%
		Decision Trees	91.34%
2	Gen, MEhs, HCGPA, PSat, PCGPA, SSE	Bayesian Network	96.6%
		Decision Trees	92.83%

7.2 Results Interpretation:

It can be observed from the above table that, when all the parameters and results are calculated Bayesian Network give 93.22% of accuracy whereas Decision Tress show an accuracy rate of 91.34%. In the next step when few parameters are selected, and the procedure is repeated, the accuracy is improved, and Bayesian Network outperforms the Decision Trees by 3.77% and hence it shows lots of improvement when specific features are used.

8. CONCLUSION:

This paper provides a brief progress of Data Mining in Education and has put forward various tools and techniques used by various researchers in this arena. Educational Data Mining is an upcoming and promising field with broad prospectus. Student's academic performance prediction is very much crucial for every educational institution for various reasons like to determine student drop-out rate or failure ratio, which can be a supportive tool, facilitating the accreditation procedure at higher education level or at university. Thus, it can be concluded that EDM can be used for high quality research, seeing the immense growth of Educators and Students in Learning Analytics.

9. REFERENCES

[1] Abu Saa. A, (2016), "Educational Data Mining & Students' Performance Prediction", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5.

[2] Altujjar. Y,Altamimi. W, Turaiki.I & Razgan. M, (2016), "Predicting Critical Courses Affecting Students Performance: A Case Study", ScienceDirect, Procedia Computer Science 82, pg: 65 – 71.

[3] Asif. R, Merceron.A, Ali.S and Haider.N, (2017), "Analyzing undergraduate students' performance using educational data mining", Elsevier, Computers & Education 113 (2017) 177-194.

[4] Baker, R.S.& Yacef, K (2009), "The state of educational data mining in 2009: A review and future visions", JEDM-Journal of Educational Data Mining, Article 1, Vol 1, No 1.

[5] Berry M, Linoff G, (2004), "Data Mining Techniques: for marketing, sales, and customer relationship management", Wiley, Publishing, chapter 1, page 10.

[6] Baker.R & Inventado. P, (2014), "Learning Analytics", Chapter 4:"Educational Data Mining and Learning Analytics", pp 61-75.

[7] Daud. A, Aljohani. N, Abbasi.R, Lytras. M, Abbas. F & Alowibdi.J, (2017), "Predicting Student Performance using Advanced Learning Analytics", International World Wide Web Conference Committee (IW3C2), ACM 978-1-4503-4914-7/17/04.

[8] Gupta. G.K (2014), 3rd edition, "Introduction to Data mining with Case Studies, PHI Learning Private Limited, chapter 1, pag:3.

[9] Han. J, & Kamber. M, (2000), "Data mining: Concepts and Techniques", Morgan Kaufmann Publishing, chapter 1: Introduction page:6.

[10] Kalbande.N, Handa.C, (2015), "Developing a model to predict Employability of Engineering Students in Campus Placement for IT Sector", International Journal of Advance Research in Engineering, Science & Technology(IJAREST), ISSN(O):2393-9877, ISSN(P): 2394-2444

[11] Kolo. D, Adepoju. S & Alhassan. J (2015), "A Decision Tree Approach for Predicting Students Academic Performance", I.J. Education and Management Engineering, Published Online October 2015 in MECS (http://www.mecs-press.net) DOI: 10.5815/ijeme.2015.05.02.

[12] Khasanah. A & Harwati (2017), "A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques", IOP Conference Series: Materials Science and Engineering 215, doi:10.1088/1757-899X/215/1/012036.

[13] Nithya. P, Umamaheswari. B & Umadevi.A (2016), "A Survey on Educational Data Mining in Field of Education Department of CS", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 5, Issue 1.

[14] Sorour.S, Goda.K &Mine.T, (2017), "Comment Data Mining to Estimate Student Performance Considering Consecutive Lessons", Journal of Educational

Technology & Society, Vol. 20, No. 1, pp. 73-86.

[15] Shahiria. A,Husaina. W, Rashida.N,(2015), "A Review on Predicting Student's Performance using Data Mining Techniques", The Third Information Systems International Conference, Elsevier, (http://creativecommons.org/licenses/by-nc-nd/4.0/)

[16] Sembiring.S, Zarlis.M, Hartama.D,, Ramliana S & Wani.E, (2011), "Prediction of Student Academic Performance by an application of Data Mining", International Conference on Management and Artificial Intelligence, IPEDR vol.6.

[17] Romero.C, Ventura. S, (2007), "Educational Data Mining: A Survey from 1995 to 2005", Expert System with Applications, Vol.33, pp 135-146.

[18] Oyerinde O. D, Chia P.A, (2017), "Predicting Students' Academic Performances – A Learning Analytics Approach using Multiple Linear Regression", International Journal of Computer Applications (0975 – 8887) Volume 157 – No 4.