International Journal of Engineering Research and Technology.
ISSN 0974-3154 Volume 11, Number 2 (2018), pp. 319-331

http://www.irphouse.com

Analytics in e-Learning

Sandra Treesa Tom[1], Tulasi B[2]

[12]Department of Computer Science,
CHRIST (Deemed to be University), Bengaluru,
Karnataka, India

Abstract

Predictive analytics play an important role in the evolving dynamics of higher education. There has been a steady up rise in use of technology in the field of education. e-learning is seen as a futuristic approach of learning. Hence, the study of factors influencing success in e-learning courses is relevant to the current scenario. Use of predictive analytics in virtual learning environment would provide insight on learning patterns of students. The learning data available in the traditional teaching environment is different from the one, which is available from virtual learning. This paper tries to identify various attributes associated with e learning which can help in making the learning process effectual.

Keywords - Data analytics, e learning, Higher education, predictive analytics, Principle Component Analysis (PCA).

I. INTRODUCTION

Higher education is under huge influence to increase enrolment, reduce the student dropout, increase student retention, ensure students graduate, reduce costs, consistently enhance the learning experience and be responsible and clear among others [12] [10] [13]. These influences are pushing universities and colleges to develop and re-think the way things will be done and to drive towards new answers to solve these difficulties. Many universities and university market leaders are employing data evaluation and visualization to understand the problem and develop effective solutions [12] [5]. Attaining these insights and information requires not only a solitary statement from an individual system, but also the capability to gain access to share and explore institution-wide data that may be transformed into meaningful insights at every degree of the organization.

E learning is a new approach evolved from the technological era, which has provide an alternative approach in field of education [15]. In this fast growing world,

technologies can contribute more to the education system. That is where this research gets its importance. E learning works beyond the boundaries of old traditional classroom system [4] [9]. E learning has made a vast impact in the field of higher education.

The main aim of this study is to find various factors that can affect a student's learning. Identifying various attributes, which affect learning in a virtual environment, is more complex process than in a traditional classroom environment. Compared to the data, which is available in a college or a university, an E-learning modules learner's data has varied parameters.

For example, the demographic data available with the college or university of a region might not have encompass greater geographic location. However, in case of E learning the students from different regions would enrol for a particular course. The range of the demographic data is wide. To assess various factors like demographic data, educational qualification that can play major role in student learning is the main idea of this research.

I.I. Educational Data and Analytics

Data analytics is the buzzword in the present era. As the name suggests, data analytics is the process of analysing data. The purpose of analysing differs as per the requirements. Analytics can be applied to find solution for a particular problem, to predict the future, or to improve the future etc. The steps of data analytics include preparation, pre-processing, analysis and post processing [1]. It is a continuous process starting from collection of the data until the solution is obtained. The first step, preparation includes planning, data selection, data collection and feature generation. The data identification and collection is done in this stage. Data is collected as per the requirement and the data is pre-processed in the next step. Pre-processing includes cleaning, filtering, completion, correction, standardization and transformation of data. The erroneous data are removed here. The missing values can be filled by taking mean, replacing, fetching the nearest neighbour, filtering etc. The removal of missing values, invalid list, invalid value and outlier are performed in pre-processing. The data will be clean and noise free after pre-processing. Now the pre-processed data is carried to the analysis stage. Here the visualization, correlation, forecasting, regression, clustering or classification techniques are applied to find hidden patterns in the data. The feature and characteristics of the data is identified here. As per the characteristics of the data, the analysis method is chosen. For example, if the data is unlabelled then clustering techniques can be chosen for the analysis. If the collected data is labelled, then classification is a better option. As per the behaviour of the gathered data, various data analysis techniques can be applied. Once the analysis of data is completed, the result or the solutions can be arrived. Post processing is the step where the analysed result is interpreted, evaluated and documented. This is the stage where the results or the solutions will be reached to the public or respective persons.

In addition to the above four steps there are four more levels of data analytics 18]. They are descriptive analytics, diagnostic analytics, predictive analytics and prescriptive analytics. Descriptive analytics is the first level where the data is analysed to find what the data is about. Here detailed study of the data is completed to know what happened really. Once the researcher identifies the situation then he moves on to the next level, diagnostic analytics. In this level, why that particular situation happened is recognized. The causes of a particular problem are identified. The next level is predictive analytics, here based on the knowledge gathered from the previous levels the prediction is made. The detailed knowledge about the data, the current situation, and the causes for the situation will help in prediction. Prescriptive analytics is the final level where the goal to be achieved is identified first. Once the goal is set then the possible way to achieve the goal is identified.

Predictive analytics is introduced in every possible sectors to improve their quality and productivity. Analytics is been used in many domains like banking, trade markets, retail management, stock market, supply chain management. Education system [4] introduced predictive analytics to improve the quality of education.

In an educational institution where student data is available, the following steps can be taken to obtain deeper understanding about student learning. Data analysis of student's academic performance involvement in extra and co-curricular activities, demographic information about previous academic and no academic performances will provide pointers for both student and teachers to enhance the learning experience. Using analytics, the future performance of student's, institution's performance can be forecasted. It helps in implementing various new technologies in education sector. By predictive analytics intervention by student or teacher would be done at the right time to facilitate academic learning.

II. RELATED WORK

The growth of online learning gave birth to many online community colleges [15]. These colleges offer various online courses in their curriculum. Using this data, they build model to forecast and predict the student outcomes. Online learning is mostly related to the usage of online tools for learning. Online learning is the combination of e learning and blended learning. E learning is where the learning happens online. The teacher and the students interact via internet. Physical classroom doesn't exists in this type of learning. Assessments, assignment submission are done through online portal. Online learning is one of the major areas where the data analytics can be applied.

Learning analytics became a buzzword in the education system in present days. It can be defined as the measurement, gathering, examination and recording of data about learners and their situations, for purposes of understanding and improving learning and the environments in which it happens. The need of analytics increases day by day. One of the major area that focus in the education system is the student performance and success [1]. Analytics is the discovery on meaningful patterns in the collected data so that from those meaningful patterns new solutions are derived. Most of the researchers uses patterns to predict the student success or student's performance. By

learning the data in detail, the factors that can affect one's performance can be determined. Therefore, it helps the researchers to guide the student in correct path to increases his/her performance. While considering an educational institution, the student success plays a major role in the reputation of the institution. Data analytics has vast impact on increasing the student success and increasing the quality and productivity of an educational institution.

In addition to learning analytics, educational data mining also plays a major role in extracting patterns and predictions. Educational data mining is one of the area, where research is done in course management data to find the student behaviour pattern so that it can improve the institutional effectiveness. Combined data from student system and learning management system, the University of Phoenix implemented a model. The model helps the university to predict the student who might fail, so that the necessary action can be taken to improve the course success. Student retention is one of the main problem faced in the universities in other country. Many of the universities and colleges fail to retain their students [2]. The number of students dropping out increased day by day. If the institution can retain the students, it can improve the fiscal position of the universities or colleges.

Clustering and classification are the two techniques used in data mining and data analytics. Clustering is unsupervised while classification[3] is supervised learning. In case of clustering, any prior knowledge about the class is not required. It clusters or group the data according to the pattern within. Kmeans is one of the clustering algorithm that is widely used in data analytics in education. In paper [6] the authors modify the existing kmeans algorithm to make their model more accurate. The improved kmeans helped them in identifying the outliers and build an efficient model. The result of the clustering depends on the initial clusters. The improved kmeans was on based on grid density that help them to reduce the impact of outliers on the results.

Every educational institution struggles to improve its quality and productivity [7] for the better future of the institution and students. The data analytics plays important role in analysing the practices in the education sector and provides suggestions to improve the quality of education. This paper suggests sharing of data across institutions and provides necessary records and infrastructure to the institutions, faculty and students. Privacy and ethics are the some of the concerns that are faced by the institution while the data is been shared. The Australian education sector developed open and shared analytics curriculum for improving the education sector. This was done with the help of learning analytics and education data mining research communities.

Universities are forced to adopt data analytics to improve their quality. The major concern of the universities are on the student success, reduce cost, dropout risk and academic performance [5]. Research is done to identify ways to decrease the education cost and to identify the students who are struggling to pass a particular course. By solving this two problems, the universities can reduce the number of students who are dropping out of the course due to financial problem and due to difficulty of the course. The universities uses predictive analytics for solving the problems. Predictive analytics uses variety of statistical techniques to analyze the

current and historical data to arrive at some conclusion or solution to a particular problem.

At-risk students [8] are those students who may fail in the exam. Those students can be identified by their performance and previous marks. To identify those students university of phoenix had designed a logical regression model. Not only in universities or in colleges, the online courses also faces at-risk student problem. The authors uses the time series data to identify the students. The authors used clustering approach to build the model. The model was built on dynamic data that was collected from the online courses, therefore the model is more accurate and feasibility is higher than other models. The researchers found that the behaviour of the students are different before and after holidays. The holiday effect also helped them in clustering the students into at-risk cluster.

Many researchers found after their studies that the environment for learning should be made convenient to the learners. Opting a course has impact on the learning. Therefore, education system introduced course management system to improve the efficiency of the faculty and other staff, which will show impact on the student performance and success [10]. Data mining is the appropriate tool for decision-making, finding solution, identifying the hidden patterns and relationships among huge data which help the organisation to take sufficient action for a solution.

Education sector is facing lots of problem to sustain in this new era. The few challenges [11] are course success, curriculum development, at-risk students, student learning outcomes, behaviour and process, student retention, academic performance, personalized learning, fiscal success, instructor performance, student performance, issues related to data tracking, gathering, evaluation, analysis, student success, lack of connection to learning sciences, optimizing learning environments, privacy and ethical issues, quality and productivity of educational institution. Blackboard analytics [13] is a group of researchers who have done various studies on this problem. According to their studies financial burden, lack of interest towards the student's success are the few reason for the student retention. They suggest that universities can provide more scholarships, concentrate more on at-risk students so that they can retain the students.

Web-based learning is one of the advanced types of learning which was introduced in education sector. Web-based learning system is where; there is no teacher to take classes. It is computer-aided teaching, which is independent of location and hardware [4]. Introducing the web-based learning has become a milestone in the growth of educational system. It has considerably gained importance. The students have been introduced to the new concept and then data of their performance are collected. The collected data is passed to the steps of data mining. Clustering, classification, pattern matching, text mining and outlier association are some of the data mining techniques that can be applied to the collected data. Using the discovered pattern of the data, the students get the recommendation or suggestion to improve their learning and the educators or the academics responsible will be able to view the discovered knowledge. This is how educational data mining is applied in education sector.

III. METHODOLOGY

The data for the research has been collected from the OULAD (Open University Learning Analytics Dataset) [17]. This dataset contains information about the students, courses and their interaction with the Virtual Learning Environment (VLE) for seven courses called modules. The course starts twice in a year i.e. in February and October. They are represented as 'B' and 'J' respectively. The dataset has three sections: student activities, module presentation and student demographics. All the seven files includes detailed information about more than twenty thousand students. Figure 1 illustrate the database schema of collected data.

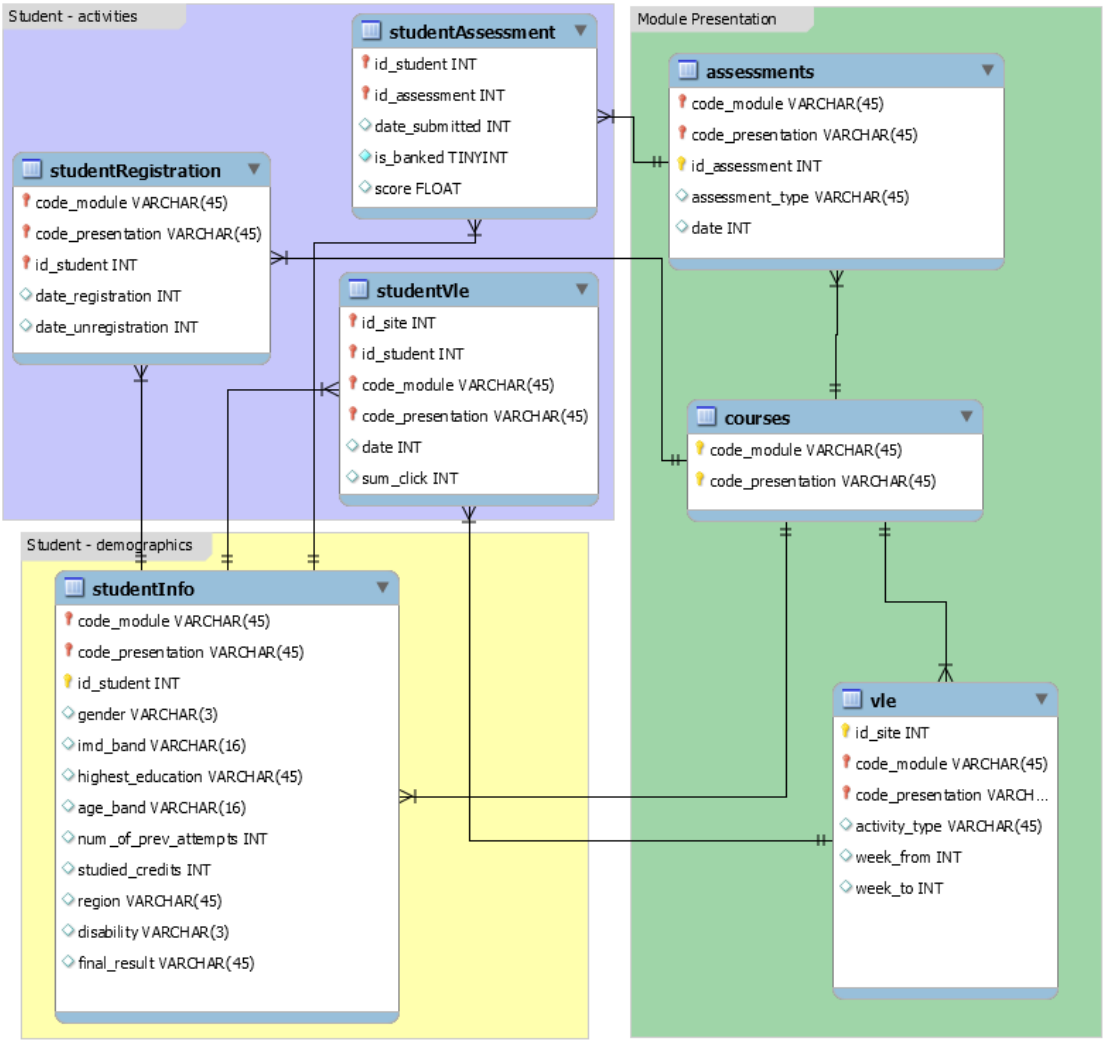

Fig. 1 Database Schema

The dataset contains student registration details, basic information about the student, assessment details, and details about the VLE that he/she is enrolled for. It also

contains the offered VLE and its course details and the assessment details for each course. There are three types of assessment in each course. TMA (Tutor Marked Assessment), CMA (Computer Marked Assessment) and Exam (Final Exam) marks. The number of assessments varies according to the course. Only the relevant data has been chosen from the entire dataset. The chosen data was course_module, code_presentation, id_student, region, higher_education, age_band, num_of_prev_attempts, studied_credits, and marks of TMA, CMA and Exam and the final_result.

Once the variables were identified the data was passed for pre-processing. The first step was to eliminate all null values. This can be done by taking the mean of the entire data and fill the null places or fill manually or delete that entire record. Once the null values are removed the next step was to fetch the erroneous data. The data, which was entered wrongly, was identified and removed. After completing pre-processing, 18121 records are retained in the dataset. The sample of pre-processed data is shown in figure 2. Kmeans clustering give results that are more efficient when it is applied to the numerical data rather than categorical data. For that, the categorical data values are converted to numerical values. Next, the pre-processed data is divided into training and testing dataset. The first 9060 records was labelled as training dataset and rest was used for the testing the model which was built by the training dataset.

1	id_studen	region	highest_ec	age_band	studied_cr	Marks	final_result
2	11391	1	3	55	240	82	1
3	23629	1	1	25	60	30.8333	0
4	23798	11	2	25	60	94.4167	2
5	25107	1	1	25	120	79.25	1
6	26023	6	1	45	30	80.75	1
7	26247	8	3	45	60	8.8	0
8	26915	11	1	45	30	64.75	1
9	27935	1	1	25	60	25.0833	0
10	28046	13	3	45	70	45.5	0
11	28400	7	3	45	60	66.4	1
12	29335	12	1	25	60	92.4429	2
13	29411	2	2	25	60	69.0833	1
14	29769	2	2	25	60	41.0714	0
15	30091	10	2	25	60	81.6667	1

Fig 2. Pre-processed data

IV. IMPLEMENTATION

Course_module, code_presentation, id_student, region, higher_education, age_band, num_of_prev_attempts, studied_credits, and marks of TMA, CMA and Exam and the

final_result are the chosen attributes from the database for further research. Taking more attributes may affect the accuracy of the model hence dimensionality reduction has to be applied. Dimensionality reduction is process of eliminating the unnecessary variable for obtaining the principal variables. Feature extraction is the technique of transforming high-dimensional space data to a space of fewer dimensions. The data transformation can be linear and nonlinear.

PCA (Principal Component Analysis) one of the feature extraction techniques of dimensionality [19] reduction which was applied on the selected data. PCA is the linear technique of dimensionality reduction that was used to reduce the attributes. The steps for calculating the principal components includes various mathematical formulas.

Step 1: Select the pre-processed data on which PCA should be applied. For example, let us take a two dimensional data set. Let the values be X and Y. X and Y contains n values.

Step 2: Find the mean of the values $\overline{X}$and $\overline{Y}$ for the X and Y values respectively.

$$\bar{X} = \frac{\sum(X)}{n} \bar{Y} = \frac{\sum(Y)}{n}$$

Step 3: Subtract the mean. Itis necessary for performing PCA to ensure that the first principal component describes the direction of maximum variance. Find $X - \bar{X}$of all X values and $Y - \overline{Y}$ for all Y values

Step 4: Covariance matrix, to calculate covariance matrix the variance of X and Y and the covariance of X, Y should be calculated.

$$\text{Variance (X)} = \frac{\sum(X - \bar{X})^2}{n-1} \qquad \text{Variance (Y)} = \frac{\sum(Y - \bar{Y})^2}{n-1}$$

$$\text{Covariance (X, Y)} = \frac{\sum(X - \bar{X})\,(Y - \bar{Y})}{n-1}$$

$$\text{Covariance matrix} = \begin{bmatrix} Variance(X) & Covariance(X,Y) \\ Covariance(X,Y) & Variance(Y) \end{bmatrix}$$

Step 5: Eigen values and Eigen vectors

$$\lambda^2 + \big(variance(X) - variance(Y)\big)\lambda + (variance(X) * variance(Y) - covariance(X,Y))^2$$

To get the Eigen values the above equation should be solved and find the values of λ. Quadratic equation is implemented to sole the above equation.

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

The solution will give two values for λ i.e. λ_1 and λ_2. If

$$\text{Covariance (X,Y)} = \lambda_1 + \lambda_2$$

Then the calculated value λ_1 and λ_2 are correct. Let us take $\lambda = \lambda_1 + \lambda_2$. The Eigen vector is

$$\begin{bmatrix} \lambda - variance(X) & covariance(X,Y) \\ covariance(X,Y) & \lambda - variance(Y) \end{bmatrix}$$

Step 6: Choose components to form a feature vector: once Eigenvectors are identified from the covariance matrix, the next step is to arrange them by highestEigenvalue to the lowestEigenvalue.So that the components are organized in order of their importance. The number of Eigenvectors that you select will be according to the number of dimensions of the new data set. The goal of this step is build a matrix of vectors (feature vector). From the list of Eigen,vectors are selected to form a matrix with them in the columns:

$$FeatureVector = (eig_1, eig_2, ..., eig_n)$$

Step 7: Originate the new data set. Multiply the transpose of the FeatureVector with the original data set, transposed:

$$Final\ Data = RowFeatureVector \times RowDataAdjusted$$

Where RowFeatureVector is the matrix with the eigenvectors in the columns transposed.

RowDataAdjusted is the mean-adjusted data transposed.

cor(X)

	region	highest_education	age_band	studied_credits	Marks	final_result
region	1	-0.001662413	-0.03681852	-0.001664687	-0.01796798	-0.015495594
highest_education	-0.001662413	1	0.11704877	0.022526542	0.14385187	0.158690161
age_band	-0.03681852	0.117048769	1	-0.092446123	0.08642368	0.086062145
studied_credits	-0.001664687	0.022526542	-0.09244612	1	-0.01629171	-0.008846808
Marks	-0.017967981	0.14385187	0.08642368	-0.016291706	1	0.885454969
final_result	-0.015495594	0.158690161	0.08606214	-0.008846808	0.88545497	1

pca1 <- princomp(X,scores = TRUE, cor = TRUE)

summary(pca1)

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	1.3993401	1.0471566	1.0021834	0.996722	0.9127427	0.33820814
Proportion of Variance	0.3263588	0.1827561	0.1673953	0.1655758	0.1388499	0.01906412
Cumulative Proportion	0.3263588	0.5091149	0.6765102	0.842086	0.9809359	1

Fig. 3. PCA results

The attributes will be ranked according to their impact to the end result. After applying PCA (figure 3) it was found that all the assessment marks have more impact on the end results. Region, higher_education, age_band and studied_credits also have impact on the end result in decreasing order. All these attributes were chosen for the clustering.

Clustering is unsupervised learning, that deals with the unlabelled data. k-means is a standard technique to clustering. This method is defined by its objective function that purposes to minimize the sum of all squared distances within a cluster, for all clusters. Number of cluster should be specified in kmeans. The algorithm clusters the entire data into k clusters. Each data is assigned to one of the k cluster based on their feature similarity. One feature of k-means that makes it dissimilar from other clustering algorithms is that the number of clusters is fixed when clustering happens. This can be reflected as a strength. The optimistic property of a fixed number of clusters is that in case of an anomaly data point, the k-means algorithm does not host new clusters; as an alternative, it categories the anomaly data point to its nearest cluster.

Kmeans clustering was applied on the pre-processed dataset. Kmeans algorithm choses k random data points as the centroids of k clusters. Using the distance measure specified distance matrix is developed and the data points are grouped to the nearest centroid. Once the first clustering or grouping is completed, the mean of each cluster is calculated. In the next step, the mean is made as the centroid. Then again, the distance matrix is generated and the data points are clustered to the nearest centroids. After the clustering the mean is calculated again. This whole process continues until the centroids does not change. Once the centroids are fixed, then the clustering is

completed and the results will be displayed. The grouping is done based on the behaviour of the data. The final_result contains three type of data: Fail, Pass and Distinction. So the number of cluster was chosen as three. The pre-processed data was divided into training and testing dataset. The model development and implementation was done on RStudio using R programming. R is an open source technology, and an important tool for statistics and machine learning. R has features, which is very relevant for the data science applications. R supports complex mathematical calculations, vector calculations, can run code without compliers and provides numerous support functions for data science. The kmeans model was trained using the training dataset and further tested with the testing dataset to determine the accuracy. Model was obtained with an accuracy of 89%.

V. RESULTS

Using PCA, the correlation between the various identified attributes was identified. From that it is clear that the attribute "marks" has the highest impact on the end result followed by higher_education, age, studied_credits and region. The research shows that the attribute like higher_education, region, and age_band have no direct impact on the final results of a student. The assessment marks have great impact on the results compared to other attributes. The predictive model using the identified attributes and k-means algorithm provided accuracy of 89%. Further investigations by using clustering techniques like spectral, dbscan can be done to enhance the model.

VI. CONCLUSION

E learning is futuristic learning. Research on this type of learning and its outcomes is essential to standardise the learning process. The research presented in this paper is the first step towards it. More approaches and techniques need to be explored to provide deeper insight about the data. The difference between the data, which is available from e-learning platform and from traditional universities, is great. The data spectrum is wide and versatile in e learning.

VII. REFERENCES

[1] Huebner, R. A Survey of Educational Data-Mining Research. Research in Higher Education Journal, 19,2013.

[2] Alhadad, S., Arnold, K., Baron, J., Bayer, I., Brooks, C., Little, R., Rocchio, R., Shehata, S. and Whitmer, J. (2015). The Predictive Learning Analytics Revolution: Leveraging Learning Data for Student Success. EDUCAUSE Center for Applied Research.

[3] Anirban, S. (2014). Big Data and analytics in higher education: Opportunities and challenges. International Journal of Research in Computer and Communication Technology, 3(11).

[4] Arnold K. E and Pistilli M. D, 2012, 'Course signals at Purdue: Using learning analytics to increase student success', In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge.

[5] Barber R and Sharkey M, 2012, 'Course correction: using analytics to predict course success', In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge

[6] Romero and S. Ventura, 2006, 'Educational data mining: A survey from 1995 to 2005', Expert System with Applications.

[7] Calvet Liñán, L. and Juan Pérez, Á. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. RUSC. Universities and Knowledge Society Journal, 12(3), p.98.

[8] Fan, Z. and Sun, Y. (2016). Clustering of College Students Based on Improved K-Means Algorithm. 2016 International Computer Symposium (ICS).

[9] Hung, J., Wang, M., Wang, S., Abdelrasoul, M., Li, Y. and He, W. (2017). Identifying At-Risk Students for Early Interventions—A Time-Series Clustering Approach. IEEE Transactions on Emerging Topics in Computing, 5(1), pp.45-55.

[10] Kumar, B. and Pal, S. (2011). Mining Educational Data to Analyze Students Performance. International Journal of Advanced Computer Science and Applications, 2(6).

[11] Nunn, S., Avella, J., Kanai, T. and Kebritchi, M. (2016). Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review. Online Learning, 20(2).

[12] Siemens, G., Dawson, S. and Lynch, G. (2013). Improving the Quality and Productivity of the Higher Education Sector. Society for Learning Analytics Research.

[13] Sousa JD, T., Adams, P., Culverhouse, L., McCoin,, M., Whitmer, J. and Cook, A. (2015). Analytics for Higher Education: Solving Six Critical Issues with Blackboard Analytics. Blackboard Analytics.

[14] Verbert K, Manouselis N, Drachsler H and Duval E, 2012, 'Dataset-driven research to support learning and knowledge analytics', Educational Technology & Society

[15] Vernom S., Lange, A. and Huston, D. (2012). Predictive Modeling to Forecast Student Outcomes and Drive Effective Interventions in Online Community College Courses. Online Learning, 16(3).

[16] Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017). Open Learning Analytics | OU Analyse | Knowledge Media Institute | The Open University. [online] Analyse.kmi.open.ac.uk. Available at: https://analyse.kmi.open.ac.uk/open_dataset [Accessed 1 June. 2017].

[17] Declues, J. (2017). Four Types of Big Data Analytics and Examples of Their Use. [online] Ingrammicroadvisor.com. Available at: http://www.ingrammicroadvisor.com/data-center/four-types-of-big-data-analytics-and-examples-of-their-use [Accessed 3 Sep. 2017].

[18] I Smith, L. (2002). A tutorial on Principal Components Analysis.

[19] Cornerstoneondemand.com. (2017). The Four Stages of Workforce Analytics: What Level Are You? [Infographic]. [online] Available at: https://www.cornerstoneondemand.com/rework/four-stages-workforce-analytics-what-level-are-you-infographic [Accessed 22 Aug. 2017].

[20] Optimizationgroup.com. (2017). The Different Levels of Analytics – Optimization Group. [online] Available at: http://www.optimizationgroup.com/the-different-levels-of-analytics/ [Accessed 20 Jul. 2017].

[21] En.wikipedia.org. (2017). Data analysis. [online] Available at: https://en.wikipedia.org/wiki/Data_analysis [Accessed 18 Mar. 2017].