Review on Predictive Modelling Techniques for Identifying Students at Risk in University Environment

Nik Nurul Hafzan Mat Yaacob[1,*], *Safaai* Deris[1,3], *Asiah* Mat[2], *Mohd Saberi* Mohamad[1,3], *Siti Syuhaida* Safaai[1]

[1]Faculty of Bioengineering and Technology, Universiti Malaysia Kelantan, Jeli Campus, 17600 Jeli, Kelantan, Malaysia

[2]Faculty of Creative and Heritage Technology, Universiti Malaysia Kelantan,16300 Bachok, Kelantan, Malaysia

[3]Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100 Kota Bharu, Kelantan, Malaysia

Abstract. Predictive analytics including statistical techniques, predictive modelling, machine learning, and data mining that analyse current and historical facts to make predictions about future or otherwise unknown events. Higher education institutions nowadays are under increasing pressure to respond to national and global economic, political and social changes such as the growing need to increase the proportion of students in certain disciplines, embedding workplace graduate attributes and ensuring that the quality of learning programs are both nationally and globally relevant. However, in higher education institution, there are significant numbers of students that stop their studies before graduation, especially for undergraduate students. Problem related to stopping out student and late or not graduating student can be improved by applying analytics. Using analytics, administrators, instructors and student can predict what will happen in future. Administrator and instructors can decide suitable intervention programs for at-risk students and before students decide to leave their study. Many different machine learning techniques have been implemented for predictive modelling in the past including decision tree, k-nearest neighbour, random forest, neural network, support vector machine, naïve Bayesian and a few others. A few attempts have been made to use Bayesian network and dynamic Bayesian network as modelling techniques for predicting at-risk student but a few challenges need to be resolved. The motivation for using dynamic Bayesian network is that it is robust to incomplete data and it provides opportunities for handling changing and dynamic environment. The trends and directions of research on prediction and identifying at-risk student are developing prediction model that can provide as early as possible alert to administrators, predictive model that handle dynamic and changing environment and the model that provide real-time prediction.

1 Introduction

In the past several years people responsible for higher education administration have been subjected to immense pressure to improve quality of education in terms of good quality programs and good graduate for future employment both in public and industrial sectors [1]. At the same time, with current trends of world economy that force fierce competition among public and private high education institution (HEI) to cut cost and increase efficiency. In order to make business sustainable the HEI must compete to enrol and retain the maximum number of students every semester and throughout study year. To be more efficient and competitive the HEI must depend on full use of the data to the maximum level through analytics.

Among indicators of quality of education of HEI are attrition rate (no of drop out student per semester), student graduating on time, and employability. Statistics from some universities indicated that enrolment management of the universities has become more difficult as the attrition rate is not getting better. Records indicated that the attrition rate or drop out student of some local universities is between 200-400 students per semester. If the drop out student can be managed and improved the income of the university also can be improved and this makes university more competitive and sustainable. Many reasons for drop out have been identified in previous studies ranging from academic standing, financial reason, motivation, internal institutional problems [2], however the real factors vary from university to university and from time to time.

* Corresponding author: niknurulhafzan88@gmail.com

Data in HEI have been continuously collected through various management information systems and teaching and learning management system and every year HEI keep on adding new systems to automate more functions and processes resulting in growth rate of data increase exponentially. In the past data are collected, catalogued and archived and these data have been used to certain extend for monitoring and operational decision making. Recently, with emergence of big data analytics, data can be processed and analysed to discover patterns and trends and can be used for strategic decision making and to get insight of the future. This is made possible due to the volume, size and the growth rate of data in organization covering the historical data and the current data collected through various means to handle structured and unstructured data. In addition, the increasing computational powers of the computer both hardware and software is now accessible by administrators, instructors and students.

There are many definitions for analytics have been used by different authors, for example analytics is decision making based on data where information is used to support and justify decision at all levels of the company [3]. Gartner uses term analytics to describe statistical and mathematical data analysis that cluster, segment, scores and predict what scenario is most likely to happen [4]. The keywords that must present in definition are large data set, statistical or machine learning techniques and predictive modelling. Predictive analytics is a subset of analytics that focuses on predicting a set of technologies to uncover relationships and patterns within large volumes of data that can be used to predict behaviour or events [5].

2 Academic Analytics

Big data analytics in higher education institution is new, however it is needed to enable management of higher education institution to response to demand. The main challenges facing higher education institution today in response to pressure and demand for accountability and transparency are quality education programs and student success or student performance [1].

Two main factors directly related to student performance and success is student retention rate and graduate on time. Student retention and student performance can be improved through the use of tools and techniques such as analytics and predictive modelling. The use of big data analytics in academic world sometimes refer to as academic analytics.

Academic analytics combines selected institutional data, statistical analysis, and predictive modelling to create intelligence upon which students, instructors, or administrators can change academic behaviour. Researchers at Purdue University have begun to move Academic Analytics project beyond the research and pilot phases and today become of the major resources for academic analytics research [1],[6],[7].

Academic analytics sometime use interchangeably with learning analytics by some authors, but lately these two have been differentiated as follows: academic analytics is at institutional level and for the benefit of administrators and learning analytics is at course level and for the benefit of instructors and students or learners [3],[8],[9].

Academic analytics is a process for providing higher education institutions with the data necessary to support operational and financial decision making [5],[10], while learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs [11],[12].

The use of academic analytics is currently at infancy stage [3]. Academic analytics can be used for enrollment management and improvement of teaching and learning. Academic analytics also can help the university administration to predict stopping out students and also for early identification of at-risk students enrolled in learning programme [13].

One of important area academic analytic is identifying at-risk students through analytics and predictive modelling; combination of these two terms is referred to as predictive analytics. Predictive analytics can provide institution with better decision and actionable insights based on data. Predictive analytics aims at estimating likelihood of future events by looking into trends and identifying associations about related issues and identifying any risks or opportunities in the future [8]. Output from predictive model will be used for designing intervention program to assist at-risk students improve their performance. Author [3],[13],[14],[15], [16] said that early warning notification system can be developed for administrators, instructors and students for student performance monitoring.

3 Predictive Analytics for Identifying at-risk Students

Student can be categorized into two categories. That is success student and at-risk student. According to Smith et al. [17] identifying the students' risk not only has implications for the university in terms of retention, but also comes in a cost to students. Students who wish to progress in their degree will be affected by the failure of the programme. The consequences of failure are the additional costs, postponement in degree completion, the most likely change, perhaps decline in self-confidence, and they cannot proceed to advance into the second year.

National Centre for Education Statistics [18] estimated that among first-time, full-time students who started work toward a bachelor's degree at a four-year institution in 2008, only 60 per cent graduated within six years by 2014. At public institutions, the six-year graduation rate hovers around 58 per cent; at private, non-profit institutions it's 65 per cent, while at private, for-profit institutions, it's only 27 per cent.

Drop out student and student not graduating on time can be studied through identifying at-risk student using predictive analytics. Why identifying at-risk student is very important? This is because the administrator can detect which student is at-risk and the administrator can provide intervention programs for them. By this way we can improve retention among undergraduate students.

4 Predictive Modelling Techniques for identifying at-risk Student

Predictive model is created from data trained using machine learning techniques. Many authors have made comparison among several techniques and algorithms experimentally and findings have shown that some advantages and disadvantage of the techniques.

According to Daniel [8], predictive analytics can provide institutions with better decisions and actionable insights based on data. Predictive analytics aims at estimating likelihood of future events by looking into trends and identifying associations about related issues and identifying any risks or opportunities in the future. Predictive analytics could reveal hidden relationships in data that might not be apparent with descriptive models, such as demographics and completion rates. It can also be used to help look at students who are exhibiting risk behaviours early in the semester that might result to dropping out or failing a course. It can help teachers look at predicted course completion rate for a particular tools and content in the course are directly correlated to student success [19].

Predictive model is created from data trained using machine learning algorithms. The main objectives of risk modelling using predictive analytics are to identify risk level and critical causal factors that directly related to the risks [20]. Several learning algorithms have been proposed in previous research including decision tree, nearest neighbor, regression, neural network, support vector machines, Bayesian, classification rules, and several variants of them. The suitability, performance and accuracy vary depending on nature, type and availability of data. Many authors made comparison among several algorithms experimentally and some findings and pros and cons have been analyzed and reported [2],[14],[21],[22],[23],[24],[25]. Several popular machine learning techniques will be introduced and discussed in the following section.

4.1 Decision Tree (DT)

Decision tree learns in a hierarchical way by repeatedly splitting data into separate branches that maximize the information gain of each split. DT is computationally cheap and easy to understand however DT are prone to over fitting and need to be pruned regularly. DT is considered among the popular techniques for identifying at-risk student (IARS) [2],[14],[21],[22]. The reason for popularity are easy to understand, easy to build, DT can handle both nominal and continuous input, build-in variable selection, DT are non-parametric making no assumption about distribution of input or target variables, and DT can handle missing data automatically.

4.2 Random Forest (RF)

Random Forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outfitting the class that is the mode of the classes or mean prediction (regression) on the individual tree. Dissanayake [23] concluded that Bayesian Neural Network and RF produced results of IARS suitable to be used by Principal Component Analysis with high accuracy.

4.3 k-Nearest Neighbour (KNN)

The KNN algorithm is very simple and very effective. The model representation for KNN is the entire training dataset. Predictions are made for new data point by searching through the entire training set for k most similar instances (the neighbor) and summarizing the output variables for those output variables. Marquez-Verra et al. [21] reported that KNN gave the highest accuracy for identifying at-risk students in a course in university.

4.4 Regression

Regression methods such as linear regression, logistics regression and also mixed regression model have been used extensively in the past for identifying at risk student [2],[14],[24],[25],[26]. Regression approach models the relationship between variables; dependent and independent variables. Regression is a supervised learning task for modelling and predicting continuous numeric variables. Several advantages for IARS have been reported due to its nature of providing relationship between variables linearly or non-linearly. This feature helps estimating predictors faster with good accuracy.

4.5 Neural Network (NN)

Neural network is biologically inspired by the structure of human brain. Widely used for data classification. The structure of NN algorithm has three layers; input layer, hidden layer and output layer. NN process past and current data to estimate future values, discovering any complex co-relation hidden in the data. Neural network tends to have high accuracy even if the data has a significant amount of noises. This is because the hidden layer can still discover relationships in the data despite noise. NN model have been implemented and compared with several other machine learning techniques for IARS [14],[22],[23],[24].

4.6 Naive Bayesian (NB)

Naive Bayesian is a simple but powerful algorithm for predictive modelling. NB model is based on conditional probabilities calculated directly from training data. NB is called naive because it assumes that each variable is independent but in reality it is not. Marbouti et al. [22] concluded that NB classifier and ensemble had the best results for IARS compare to 6 other classifiers.

4.7 Support vector machines (SVM)

SVM use a mechanism called kernel which calculate the distance between two observations. The SVM algorithm finds a decision boundary that maximizes the distance between the closest members of separate classes. SVM

can model non-linear decision boundaries and there are several different kernels to choose and we can even design our own kernels for specific purpose. SVM is a robust technique against overfitting and highly suitable for high-dimensional space. SVM also have been used for IARS because it is a reliable and a powerful classification tool [2],[21],[22].

4.8 Bayesian network

BN models probabilistic relation among random variables. A BN assigns probability factor to various results based upon an analysis of a set of input data. A BN is taught using training data, once trained a BN can be queried to make prediction about new data that was not represented by the training set. BN has some properties that make determining the strengths of variables in influencing outcomes an ideal choice. According to Mesgarpour et al. [20], accuracy and efficiency of BN models depend on four main design choices; the frameworks of causal tree, the framework of the system state, inference approximation algorithm and finally the assignment and update method of prior probabilities. BN have applied have applied many different areas successfully [27].

4.9 Classification rules

The predictive model can be trained to extract and discover hidden relationships and patterns using as simple as logical IF-THEN rules. These rules can also be as complex as Repeated Incremental Pruning rule [21]. In many cases classification rules were used together with other machine learning techniques as a supplement or complement to improve performance and accuracy [22], [23] model.

4.10 Ensemble

Ensemble model combines two or more models to enable a more robust prediction, classification or variable selection. The most common form of ensemble is combining decision trees into a strong model.

In general, it is difficult to decide which techniques are better than the others because the final outcomes depend on many factors from the nature of the data to design, setting and configuration of solution itself. One technique may be better under one environment but may not be the best for different environments. This is may be the reason why many researchers preferred to use several techniques for solving prediction of at-risk students and then choose the best results. Table 1 shows comparison of major machine learning techniques for predicting at-risk students.

Table 1. Machine Learning Techniques for Predicting at-risk Student

Techniques	Strengths and Weaknesses	Authors
K-Nearest Neighbour	• No assumptions about data- useful, for example, for nonlinear data • Simple algorithm to explain and understand/interpret • Relatively high accuracy classifier • Versatile—useful for classification or regression • Computationally expensive • High memory requirement • Stores all training data • Prediction stage might be slow • Sensitive to irrelevant features and the scale of the data	Marquez-Verra et al. [21] Dissanayake [22] Marbouti et al. [23] Wolff et al. [24] Oliverra et al. [33] Abbott [34] Howard et al. [38] Kuzilek et al. [44] Shahiri et al. [49]
Decision Tree	• Does not require any domain knowledge. • Easy to comprehend. • Learning and classification steps are simple and fast. • Trees can be very non-robust. • Handle missing missing data automatically • Non-parametric	Lakkaraju et al. [2] Agnihotri and Ott [14] Marquez-Verra et al. [21] Marbouti et al. [23] Abbott [34] Marques et al. [37] Nandeshwar et al. [42] Singh and Singh [47] Shahiri et al. [49]
Random Forest	• One of the most accurate learning algorithms • Runs efficiently on large databases. • Can handle thousands of input variables without variable deletion. • Gives estimates of what variables are important in the classification.. • An effective method for estimating missing data and maintains accuracy • Good for unbalanced data sets. • Over fitting problems for some datasets with noisy classification/	Lakkaraju et al. [2] Dissanayake [22] Oliverra [33] Abbott [34] Howard et al. [38] Najdi and Er-Raha [46]

4

	regression tasks.	
Naïve Bayes	• Easy to implement. • Requires a small amount of training data to estimate the parameters. • Good results obtained in most of the cases. • Can make probabilistic predictions. • Can't do regression.	Agnihotri and Ott [14] Marquez-Vera et al. [21] Marbouti et al. [23] Oliverra et al. [33] Abbott [34] Nandeswar et al. [42] Baker et al. [43] Singh and Singh [47] Shahiri et al. [49]
Bayesian Network **(BN)**	• Visually represent all the relationships between the variables. • Can handle incomplete data • Help to model noisy systems. • Can be used for any system model • The quality of the results of the network depends on the quality of the prior beliefs or model • Calculation can be NP-hard • Incomplete data sets can be handled well by BN • Causal relationships can be learned about via BN • BN promote the amalgamation of data and domain knowledge • BN avoid over fitting of data .	Mesgarspour et al. [20] Wolff et al. [24] Abbott [34] Subbiah et al. [36] Howard et al. [38] Kuzilek et al [44]
Logistic Regression	• More robust: the independent variables don't have to be normally distributed • Does not assume a linear relationship between the independent and dependent variable • May handle nonlinear effects • Does not require that the independents be interval and unbounded	Lakkaraju et al. [2] Agnihotri and Ott [14] Mesgarspour et al. [20] Abbott [34] Howard et al. [38] Norrish et al [39] Huang and Fang [40] Tran et al. [41] Baker et al. [43] Kuzilek et al. [44]
Linear Regression	• Assume linear relationship • Suitable for simple application • Space complexity is very low • It's a high latency algorithm. • Its very simple to understand • Good interpretability • Feature importance is generated at the time model building. • The algorithm assumes data is normally distributed in reality they are not. • Prone to outliers	Smith et al. [17] Marbouti et al. [23] Wolff et al [24] Oliverra et al. [33] Howard et al. [38] Norrish et al [39] Huang and Fang [40] Tran et al. [41] Baker et al. [43] Kuzilek et al. [44] Najdi and Er-Raha [46]
Support Vector Machine	• Works well with even unstructured and semi structured data like text, Images and trees. • The kernel trick is the real strength of SVM. • With an appropriate kernel function, we can solve any complex problems. • Choosing a "good" kernel function is not easy. • It scales relatively well to high dimensional data. • SVM models have generalization in practice; the risk of overfitting is less in SVM. • Training time for large datasets. • Difficult to understand and interpret the final model, variable weights and individual impact.	Lakkaraju et al. [2] Marquez-Verra et al. [21] Marbouti et al. [23] Abbott [34] Howard et al. [38] Shahiri et al. [49]
Neural Network	• Storing information on the entire network • Ability to work with incomplete knowledge • Having fault tolerance • Having a distributed memory • Parallel processing capability • Hardware dependence • Unexplained behavior of the network • Determination of proper network structure	Agnihotri and Ott [14] Dissanayake [22] Marbouti et al. [23] Oliverra et al. [33] Abbott [34] Howard et al. [38] Nandeswar et al. [42] Chen et al. [48] Shahiri et al. [49]

5

Classification Rules	• Use if-then rules searching technique • Facilitates identification of variables. • Helps to establish the relationship among various groups of variables.	Marquez-Verra et al. [21] Dissanayake [22] Marbouti et al. [23]
Ensemble	• Combination of more than one machine learning techniques • Improve accuracy of supervised learning tasks • Difficult to measure correlation between classifiers from different types of learners • Learning time and memory constraints • Learned concept difficult to understand	Agnihotri and Ott [14] Marbouti et al. [23] Abbott [34]

Table 2. Recent Research on Predictive Modelling Techniques of at-risk Student

Authors	Techniques compared	Findings/Comments
Marquez-Verra et al. [21]	Naïve-Bayes, SVM-SMO, K-NN, DT-classification rules, DT(C4.5). ICRM2-Interpretable classification rule mining	• Focus on early prediction of school dropouts • Ran 3 experiments: normal, with feature selection, and tested with imbalance data • Proposed early predictive analytic methodology • Proposed new technique ICRM2 outperformed all other techniques
Marbouti et al. [23]	Logistic regression, SVM, Decision tree, Neural network, Naïve Bayesian, K-NN	• Predict at-risk student of a university • Applied feature selection and created ensemble model • Naïve Bayes and ensemble models produce the best results
Olivera et al. [33]	Logistic regression, Artificial NN, Naïve Bayesian, Random forest	• Comparison of ML techniques in healthcare environment • The best models created were ANN and logistic regression, with AUC more than 74%. • All predictive models produced similar results.
Lakkaraju et al. [2]	Random forest, AdaBoost, Logistic regression, SVM, decision tree	• Predict at-risk student for school • Focus on early prediction of at-risk students based on changes in performance • Developed metric for evaluation of off-track student • Random forest was the best technique • Decision tree, AdaBoost and SVM performed poorly
Dissanayake [22]	K-NN, Classification tree, random forest, Neural Network, binomial GLM, Bayesian neural network	• Predict at-risk students of university • Used PCA for feature selection • Bayesian NN and random forest gave high prediction accuracy
Agnihotri and Ott [14]	Logistic regression, SVM, Neural network, Naïve Bayesian, decision tree, ensemble	• Build at-risk student model of high learning institution • Provide end-to-end solution • The models identified key factors of at-risk students • Compared all 5 techniques and LR and ensemble were among the highest recall
Wolff et al. [24]	k-NN, classification and regression tree, Bayesian network. Developed 4 models, 2 using K-NN, 1 model each for classification and regression, and Bayesian network.	• Predict at-risk students of distance learning programs nof Open University. • Focus on early detection and improve accuracy • Data from demography and VLE. • Predict at-risk student based on four model from 3 techniques.
Howard et al. [38]	Bayesian Additive Regressive Trees (BART), Random Forests, Principal Components Regression (PCR), Multivariate Adaptive Regression Splines (Splines), K-Nearest Neighbours, Neural Networks and, Support Vector	• Compared 8 machine learning techniques for early detection • Proposed a new technique, BART • BART gave better performance, accuracy, and eble to predict early

6

	Machine	

5 Bayesian Network Technique for predicting at-Risk Students

Section 4 discusses various machine learning techniques for predictive modelling of at-risk student based on classifier model. Most of the above techniques learn and discover relationships and pattern based on regression and classification tasks. Therefore, the prediction decision is made based on whether the student is classified as at-risk or not at-risk. BN can be used for preventive modelling, pattern recognition and regression. BN has several unique advantages compared to above machine learning algorithm; BN handle missing values very well and BN can be queried to make prediction. Training a BN is divided by two distinct stages; the first stage creates the structure of the BN and the second stage creates the probabilities between the BN's nodes. One of the most exciting prospects of using BN is the possibility of discovering causal structure in raw data. Training in BN is where training data is used to construct a BN. BN is constructed so that the probabilities produced by the final network closely match the training data [28].

Predictive modelling using BN have been studied for healthcare risk modelling Mesgarpour et al. [20], for prediction of future event of large scale complex event from IOT network [29], computer security analytics [30] and life sciences [50]. BN has also been studied for course selection of high school environment [31]. The most relevant and closest to academic analytics application is by Arcuria [26] in his thesis where dynamic BN predictive model have been created for predicting stopping out among students of community college. However, the fully dynamic BN was not successfully developed. Despite that the author has successfully developed the method and step model building process for developing dynamic BN for predictive model of at-risk student.

6 Trends and Directions

Number of student enrolment is an important factor that ensures the university operation is sustainable and making progress to the greater height. Management of student's enrolment can be improved by having early warning system (EWS) to detect and diagnose the at-risk student early in the process so that the faculty can provide intervention programs that help student to progress and improve their performance [38], [51]. In order to make EWS effective, it must me be developed and equipped with a model based on dynamic predictive techniques that enable the realization the goal of the university become reality. Chatti et al [45] proposed four-dimensional Learning Analytic Reference Model that consists of data and environment, stakeholders, objectives and methods as a framework and guideline form learning analytics implementation.

The main issues that hold back the success of EWS for IARS is how to develop models that provide prediction as early as possible that assist the administrators and instructor to introduced intervention programs to affected students as early as possible [13], [21], [23], [24], [31], [32], [38], [43]. The other direction is to create a dynamic model that is capable of handling changes and time varying environment. Having this model will enable EWS implement adaptive and real-time prediction which badly needed for early prediction. One of the possible ways toward this direction is by applying Dynamic Bayesian Network [20],[26].

7. Conclusion

In this paper, we reviewed the current issues in academic analytics focusing on modeling techniques for identifying and predicting at-risk students in the university environments. Almost all major machine learning techniques have been tested for predicting at-risk student including, decision tree, random forest, k-nearest neighbor, regression, neural network, support vector machine, Naïve Bayes, Bayesian Network, classification rules and a few others. Several new or improvement of existing techniques have been proposed and compared with other known techniques. In most of the cases, the proposed techniques always performed better than the previous techniques. Decision tree, random forest, Naïve Bayesian and regression are among the most popular modeling techniques for predicting stopping out student. Comparative studies also have shown that ensemble techniques out-performed others probably attributed to the results were generated from combination of several good techniques. Currently, most of the machine learning techniques used for predicting at-risk student was mainly based on classification, regression and association learning tasks. A few machine learning techniques such as Bayesian Network learnt based on probabilistic learning task. However Bayesian Network based learning may provide opportunities for dynamic and real time prediction of at-risk students.

8. References

[1] J. P. Campbell, and D. G. Oblinger,, EDUCAUSE Review, **42**, 4, 40–57 (2007).

[2] H. Lakkaraju, Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L., Kdd,1909–1918 (2015).

[3] S. A. Ferreira and A. Andrade, IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, **9**, 3, 98–105 (2014).

[4] Analytics - Definition by Gartner. (n.d.). Retrieved from https://www.gartner.com/it-glossary/analytics

[5] A. van Barneveld, K. E. Arnold, and J. P. Campbell, *Analytics in Higher Education* (2012).

[6] L. Iten, K. E. Arnold, and M. Pistilli, *Eleventh Annual TLT Conference*, Purdue University (2008).

[7] K. E. Arnold, EDUCAUSE Quarterly, **33**,1 (2010).

[8] B. Daniel, British J. of Edu. Tech.**, 46**, 5, 904–920 (2015).

[9] A. K. Waljee, P. D. R. Higgins, and A. G. Singal, Clinical and Translational Gastroenterology, **5**, 1 (2014).
[10] P. J. Goldstein and R. Katz, *Educause Center for Analysis And Research (Ecar)* (2005).
[11] J. Grau-Valldosera and J. Minguillon, LAK 2011-1st Int. Conf. on Learning Analytics and Knowledge, Banff, AB, Canada (2011).
[12] R. Ferguson, Int. J. of Tech. Enhanced Learning, **4**, 5 (2012).
[13] M. L. Fonti, Doctoral dissertation. Nova Southeastern University. Retrieved from NSUWorks, College of Engineering and Computing (2015).
[14] L. Agnihotri and A. Ott, *Proceedings of the 7th Int. Conf. on Edu. Data Mining.*, London (2014).
[15] L. P. Macfadyen, S. Dawson, A. Pardo, and D. Gasevic, Research and Practice in Assessment, **9** (2014).
[16] L. Najdi and B. Er-Raha, Int. J. of Comp. App., **156**, 6, 25–30 (2016).
[17] M. Smith, L. Therry, and J. Whale, Higher Education Studies, **2**,4 (2012).
[18] G. Kena *et al., "The Condition of Education 2014",* (2014).
[19] L. D. Singell and G. R. Waddell, Research in Higher Education, **51**, 6, 546–572(2010).
[20] M. Mesgarpour, T. Chaussalet, and S. Chahed, *4th* Dagstuhl, Germany, **37**, 89–100 (2014).
[21] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, Expert Systems, **33**,1,107–124 (2016).
[22] H. U. Dissanayake, Master Thesis, St. Cloud State University (2016).
[23] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, Comp. & Edu, **103**, 1–15, (2016).
[24] A. Wolff, Z. Zdrahal, D. Herrmannova, J. Kuzilek, and M. Hlosta, *LAK14 - 4th Int. Conf. on Learning Analytics and Knowledge,* Indianapolis, Indiana, USA (2014).
[25] P. Arcuria, PhD Thesis, Arizona State University, (2014).
[26] S. Muthukumaran, PhD Thesis, Simon Fraser University (2010).
[27] J. Heaton, *Forecasting & Futurism*, 7, 6-10 (2013).
[28] X. Zhu, F. Kui, and Y. Wang, Int. J. of Distributed Sensor Networks, **9**, 12 (2013).
[29] M. H. Mohd Yusof and M. R. Mokhtar, Int. J. of Advanced Sc., Eng. and Infor. Tech.,**6**, 6 (2016).
[30] S. Asadianfam, M. Shamsi, and S. Asadianfam, *Int. J. of Computer-Aided Tech. (IJCAx)* **2**, 3 (2015*)*.
[31] R. Finnie, T. Fricker, E. Bozkurt, W. Poirier, and D. Pavlic, *Toronto:The Higher Education Quality Council of Ontario* (2017).
[32] S. Sivakumar, S. Venkataraman, and R. Selvaraj, Indian J. of Sc. and Tech., **9**, 4 (2016).
[33] A. R. Olivera *et al.*, Sao Paulo Medical Journal, **135**, 3, 234–246 (2017).
[34] D. Abbott, *Applied Predictive Analytics,* Wiley, Indianapolis Indiana, USA (2014).
[35] *EliteDataScience* (2017). [Online]. Available : https://elitedatascience.com/machine-learning-algorithms.
[36] M. Subbiah, M. R. Srinivasan, and S. Shanthi, *Int. J. of Sc. and Tech. Edu. Research*, **1**, 2, 32–38, (2011).
[37] J. Marques, D. Hobbs, S. Northwest, S. Graf, and S. Northwest, *Nuevas Ideas en Informática Educativa TISE,* 120–124 (2014).
[38] E. Howard, M. Meehan, and A. Parnell, *The Internet and Higher Education* **37**, 66 -75 (2018).
[39] M. Norrish, P. Kumar, and T. Heming, *J. of Contemporary Medical Education***, 2,** 4, 199, (2014).
[40] S. Huang and N. Fang, American Society for Engineering Education, p. 17 (2010).
[41] T.O. Tran, H.T. Dang, V.T. Dinh, T.M.N. Truong, T.P.T. Vuong, and X.H. Phan, *Cybernetics and Information Technologies*, **17**, no. 2, pp. 164–182, (2017).
[42] A. Nandeshwar, T. Menzies, and A. Nelson, *Expert Systems with Applications*, **38**, 12, 14984–14996, (2011).
[43] R. S. Baker, D. Lindrum, M. J. Lindrum, and D. Perkowski, *Journal of Educational Data Mining, Article 1*, **1**, No 1 (2009).
[44] J. Kuzilek, M. Hlosta, D. Herrmannova, Z. Zdrahal, J. Vaclavek, and A. Wolff, "*LAK15-5th Int. Conf. on Learning Analytics and Knowledge*, Poughkeepsie, New York, USA (2015).
[45] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs, *Int. J. of Tech. Enhanced Learning*, **4**, 5/6, p. 318 (2012).
[46] L. Najdi and B. Er-Raha, *Int. J. of Comp. Apps.* , **156**, 6, 25–30, (2016).
[47] M. Singh and D. J. Singh, Int. J. of Comp. Sc. and Net. **2**, 4 (2013).
[48] J.F. Chen, H.N. Hsieh, and Q. Do, *Algorithms*, **7**, 4, 538–553 (2014).
[49] A. M. Shahiri, W. Husain, and N. A. Rashid, *Procedia Computer Science*, **72**, 414–422 (2015).
[50] L. E. Chai, M. S. Mohamad, S. Deris, C. K. Chong, Y. W. Choon, and S. Omatu, *International Journal of Bio-Science and Bio-Technology*, **6**, 1, 41–52 (2014).
[51] M. Vahdat, A. Ghio, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg, *Computational Intelligence* (2015).

8