THE PREDICTION OF STUDENT FAILURE USING CLASSIFICATION METHODS: A CASESTUDY

Mashael Al luhaybi, Allan Tucker and Leila Yousefi

Computer Science Department, Brunel University, London, UK

ABSTRACT

In the globalised education sector, predicting student performance has become a central issue for data mining and machine learning researchers where numerous aspects influence the predictive models. This paper attempts to apply classification algorithms to evaluate student's performance in the higher education sector and identify the key features affecting the prediction process based on a combination of three major attributes categories. These are: admission information, module-related data and 1st year final grades. For this purpose, J48 (C4.5) decision tree and Naïve Bayes classification algorithms are applied on computer science level 2studentdatasets at Brunel University London for the academic year 2015/16. The outcome of the predictive model identifies the low, medium and high risk of failure of students. This prediction will help instructors to assist high-risk students by making appropriate interventions.

KEYWORDS

Prediction, classification, decision tree, Naïve Bayes, student performance

1. INTRODUCTION

In recent years, there has been an increasing interest in applying data mining algorithms in various fields such as medicine, marketing, education, engineering so forth, due to its benefits in transforming huge amount of such data into useful knowledge. Data mining (DM), or in other words Knowledge Discovery in Databases (KDD), can be defining as a multi-disciplinary field in which several computing paradigms converge: decision-trees, artificial neural networks, rule induction, instance-based learning, Bayesian learning, logic programming, statistical algorithms, etc. The most well-known data mining techniques are Clustering, Classification, Association rule mining and Description and visualisation [1].

The growing availability of data in educational databases attracts many researchers to analyse and evaluate such data to enhance education and provide optimal solutions for associated issues. This emerging discipline is called Educational data mining (EDM) where we apply data mining (DM) techniques or develop new DM methods to explore educational data in order to understand student's learning process and their outcomes [2].

Within the education field DM seeks to analyse students learning by developing approaches that merge student's data and data mining algorithms to benefit the students and enhance their

David C. Wyld et al. (Eds) : IPPR, SOENG, DaMi, CSIT, AIS, CSE, CSIP, CCNET - 2018
pp. 79–90, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80506

learning process. However, student's performance plays a crucial role in students' academic achievement. The final grades obtained by the students throughout his/her academic study inspire their future. Therefore, it becomes essential to determine whether the students will pass or fail the module. If the predictive model can characterize the students with high risk of failure prior the examination then the academics can provide extra effort to improve students' performance and assist them to pass the module or obtain higher results.

In this connection, this study seeks to address the following:

- Factors affecting the prediction of the high risk of failure of students in higher education institutions and universities,

- Predictive data mining models using classification algorithms based on level 1 student final grades, modules related data and students admission datasets.

2. RELATED WORK

Researchers have been increasingly attempting to analyse students' datasets using data mining and machine learning algorithms in order to understand how students learn and to ultimately increase the performance of students and the quality of learning. However, a considerable amount of literature has been published on predicting the performance of the students based on different factors and attributes. These are summarized in Table 1.

Cortez and Silva [3] conducted a study to predict the performance of secondary school students based on demographic, social and past school grades. By means of Classification and Regression algorithms (Decision Trees, Random Forest, Neural Networks and Support Vector Machines), it was found that the past evaluation of the students were highly influenced with their performance. Also, there were other factors that correlated with the students' academic performance (such as: number of absences, parent's job and education, alcohol consumption).

Preliminary work on mining student datasets to predict their performance was undertaken by Al-Radaideh et al. [4]. They applied Classification algorithms ID3, C4.5, and Naïve Bayes on student's data that obtained via questionnaire for the academic members of C++ programming course at Yarmouk University, Jordan. The attributes included in this study were students demographic and tutors related data such as degree, gender and affiliated department. Weka mining tool was used in this investigation to develop the predictive models. The outcome expressed the correlation between the high school grades and students' academic performance.

Aher and L.M.R.J. [5] attempted to analyse the examination performance of final year students for undergraduate module using Weka mining tool. The algorithms Association Rule, Classification (ZeroR), Prediction and Clustering (DBSCAN) were applied on student's examination data to study the possibility of applying data mining on educational systems. The outcome of their result indicates the usefulness of data mining algorithms for higher education data especially to improve the students' performance.

A comparative analysis has been conducted by Yadav and Pal [6] to predict the final exam performance for engineering students . They applied ID3, C4.5 and CART decision trees algorithms on student's datasets that include personal, social, psychological and environmental

factors for the prediction task. The obtained results reveal that C4.5 decision tree prediction model gives better result than ID3 and CART with accuracy of 67.77% for identifying the weaker students before the examination and that help them to improve their study for better exams results.

Another study was conducted by López et al. [7] to predict the final grades of the students based on their participation in the online forum using Weka mining tool. By means of Clusteing algorithms (EM, FarthestFirst, HierarchicalClusterer, sIB, SimpleKMeans, and XMeans) they found that students participation in the course forum is a predictive factor for predicting student final grade in a module.

As shown in Table 1, researchers have attempted to analyse students demographic, social and assessment data to predict the slow learning students in order to improve their performance and reduce failure rate prior the exam [8][11]. Also, there are several studies which compared Naive Bayes method with other classification methods to classify the students and identify their abilities, interests and weaknesses [9][10].

Table 1. Accuracy results based on Decision Tree and Naïve Bayes methods

Method	**Attributes**	**Accuracy**	**Authors**
Decision Tree	Past school grades (first and second periods), demographic and social data	76.70%	[3]
	High school dataset (Demographic, Personal Data and Admission data)	69.73%	[11]
	personal, social and psychological data	67.77%	[6]
	Personal and pre-university data	65.94%	[9]
	Demographic, personal and psychological data	61.53%	[10]
	Demographic, personal and tutors related data	38.05 %	[4]
Naïve Bayes	Demographic, CGPA and course assessments data	73%	[8]
	Demographic, social data and past grades (first and second periods)	65.13%	[11]
	Demographic, psychological and environmental data	63.59%	[10]
	Demographic and pre-university data	58.10%	[9]

Bayesian classification method was applied by Bekele and Menzel [12] to predict students' performance based on values of social and personal attributes. The empirical result revealed that Bayesian network classifier is a valuable method for predicting the students having satisfactory, or above/bellow satisfactory performance.

Another Bayesian classification method (in particular Naïve Bayes) was modelled by Bhardwaj and Pal [13] to predict the slow and the high learner's students. The study conducted on 300 student records for BCA module (Bachelor of Computer Applications) from five colleges at Awadh University, Faizabad, India. The attributes included in this investigation were demographic, academic and socio-economic that obtained from students questionnaire and the

database of the university. By means of Naïve classification approach, it was stated that student's performance in university level is dependent on Senior Secondary Examination grades, students living location, teaching mode and other potential factors such as (Mother's Qualification, Students Habit, Family annual income and family status).

However, the predictive data mining model presented in this paper is different from what excites in the literature as it does not involve social, psychological, environmental and personal factors to predict the academic performance of the students, it based on a combination of three data categories which are admission, module-related data and student's level 1 final grade.

3. DATA MINING PROCESS

In the educational sector, student overall grades of the Modules is an important factor to determine whether the student pass or fail the Module. The overall grade is calculated by adding the student assessment grades, course activities and final examination results. Therefore, we performed steps to predict students at high risk of failing the Module based on their final or overall grades and other aspects. These steps are as follow:

3.1. Data selection and Pre-processing

This study considers students and modules data obtained from the Admission and the Department of Computer Science databases at Brunel University London, UK. The integrated data considered in this investigation could be categorised into three categories, are as follows:

I. **Admission Data** the data relating to students information when they register at the university such as Student Enrolment Status, Student Route name, Fee Status, Student Mode of studying, Qualification on Entry, Location of Study, previous institution … etc (see Table 2)

II. **Level 1 Final Grades** the overall grades for all level 1 modules that were taken by Computer Science Students in the first year which are:
Information Systems and Organisations
Logic and Computation
Level 1 Group Project Reflection
Data and Information Assessment
Software Design
Software Implementation Event
Fundamental Programming Assessment

III. **Module-Related Data:** the data for the predicted module such as Module teaching mode, Tutor Code, Tutor Name, Student study mode, Assessment type and Absences

The attributes and the domain values for the selected attributes for the current study are defined in Table 2 for reference. A total of 129 student records (instances) for the year 2015/16 are involved in this investigation to develop the predictive model for the prediction of the students at high risk of failure in some of year 2 modules as the following:

- Algorithms and their Applications
- Usability Engineering
- Software Development and Management
- Year 2 Group Project

The predicted class attribute is **Overall Grade**, which is the final grade obtained by the student in the targeted module. It has five possible values A: Excellent, B:very Good, C: Good, D: Acceptable and F: Unacceptable or Fail, which have been merged later on to Low risk, Medium risk and High risk of failure to improve the classification results as explained in pre-processing section (see Table 3).

Table 2. Attributes of the students

	Attribute	**Description**	**Domain Values**
Category 1: Admission Data	Enrolment Status	Students enrolment status	{EE}
	Programme Name	Student program name	{UG Computer Science}
	Route Name	The student chosen route	{Computer Science, Computer Science (Artificial Intelligence), Computer Science (Software Engineering), Computer Science (Digital Media And Games), Computer Science (Network Computing)}
	Route Code	The code of the student chosen route	Based on Rout Code at the University
	Through Clearing	Whether the student enrolled in the same course as the course she/he has applied for	{Y, N}
	Fee Status	Tuition fee status	{Home/EU, Overseas}
	Student MOA	Students study mode	{FT, FSK, FT120, PT80, PT20}
	Detailed Fee Status	Tuition fee status	{Home, European, Overseas}
	Fee	The amount of paid fees	Based on the amount of paid fees
	Gender	Student gender	{M, F}
	Country of Domicile	Student country	Based on Student country
	Age on Entry	The student's age when he/she enrolled at the university	Based on Student age
	Qualification on Entry	Students previous qualification	{Foundation degree, Foundation course at level J, Higher education (HE) access course, A/AS level, Level 3 quals, all are subject to UCAS Tariff, Other qualification at level 2, International Baccalaureate (IB) Diploma, Non-UK first degree}
	CRS Code indicates	Payment method for	{Y, N}

	LBIC	the course	
	Location of Study	Campus name	Based on Campus name
	Admissions - Core Grades Flag	Indicates admissions decision for registering the student in the course	{Achieved, Predicted}
	Previous Institution	Student previous school or institution	{UK State School, UK Independent School, Any Non-UK Institution, UK Higher Education Institution}
Category 2: Level 1 (1st Year) Final Grades	Information Systems and Organisations_Grade	Module Final Grade	{ A – Excellent, B - very Good, C - Good,D - Acceptable, F – Unacceptable}
	Logic and Computation_Grade	Module Final Grade	{ A – Excellent, B - very Good, C - Good,D - Acceptable, F – Unacceptable}
	Level 1 Group Project Reflection_Grade	Module Final Grade	{ A – Excellent, B - very Good, C - Good,D - Acceptable, F – Unacceptable}
	Data and Information Assessment_Grade	Module Final Grade	{ A – Excellent, B - very Good, C - Good,D - Acceptable, F – Unacceptable}
	Software Design_Grade	Module Final Grade	{ A – Excellent, B – very Good, C - Good, D - Acceptable, F – Unacceptable}
	Software Implementation Event_Grade	Module Final Grade	{ A – Excellent, B - very Good, C - Good,D - Acceptable, F – Unacceptable}
	Fundamental Programming Assessment_Grade	Module Final Grade	{ A – Excellent, B - very Good, C - Good,D - Acceptable, F – Unacceptable}
Category 3: Module-Related Data	Course MOA	Module teaching mode	{FT, FSK}
	Tutor 1 Code	The code of the tutor at the university	Based on tutor code
	Tutor 1	The name of the tutor of the Module	Based on tutor name
	Module	Module Code at Brunel University	Based on module code in the university
	MAB_SEQ	Assessment code	{1, 2}
	MAB_NAME	Assessment type	{Unseen Examination, Assessment, Post-Mortem Style Group Review, Assessment of ethical and professional behaviour, Open book in-class Programming Test, Group submission of a design document plusprototype, Individual viva voce, Programming Assignment, Coursework (Practical Assignment) }
	MOA	Student study mode	{full time, part time}

	Supervisor	Student supervisor name	Based on supervisor name
	Absences	The total number of absences during the semester	Based on Module attendance count
Class Attribute	Overall Grade	Student overall grade in the Module	{ A – Excellent, B - very Good, C - Good,D - Acceptable, F – Unacceptable}
Merged Class Attribute	Overall Grade	Student overall grade in the Module after merging	{ Low risk – A and B , Medium risk - C ,High risk - D and F }

We performed steps for the implementation of the classification and clustering algorithms to predict the academic performance of the students for some of year 2 computer science core modules which are: Algorithms and their Applications, Usability Engineering, Software Development and Management and Year 2 Group Project using Java API and Weka Mining tool.

Since the number of students final grades classes is large with five possible values (A, B, C, D and F) and that will influence the performance of the predictive models, we merged students overall grades to reduce the number of classes for the targeted Modules using 'Merges many values' filter in Weka into three classes which are low risk, medium risk and high risk of failure classes. Low risk class is for students who have obtained A and B in the targeted module. Medium risk class is for students obtained C in the module. Whereas, the high risk class for students obtained D and F (see Table 3).

Table 3. Class Attribute regarding to student final grades

Class	**Grade Band**
Low risk	A, B
Medium risk	C
High risk	D, F

3.2. Clustering

Clustering is identifying groups of objects in which the objects of such groups are similar to one another in some aspects and different from the objects in the other groups [14]. Clustering is considered as the most applied unsupervised learning technique in data mining.

In educational data mining, clustering is applied to group the students according to their performance in the course into weak and strong students to help the weak students improve their studies [15] and [16]. Also, it used to identify the active and the non-active students based on their performance in course activities [5].

In our study we applied the simple K-Means clustering algorithm to each module using Java API in order to find interesting groups of student according to their final results (academic performance) in the predicted module. We obtained a number of three clusters which are cluster0, cluster1 and cluster2 providing adequate correlations of student groups with the class attribute Overall Grade (academic performance).

3.3. Classification

Classification, a form of supervised learning, is a very common data mining technique that is applied to map datasets into sets of classes [5]. To develop such models, the data undergo a process that consists of learning and classification. In the learning process, the training set is analysed using classification algorithms to generate logical rules based on the relation between the selected attributes. Consequently, the classification process identifies the accuracy of the model by applying obtained rules on the test sets to evaluate the classifier [13].

The machine learning algorithms applied for classification process in this study were naïve Bayes and C4.5 decision tree. Since the dataset was not large with only 129 student records, we encountered class implanting issue. To solve this, we applied the Synthetic Minority Oversampling Technique (SMOTE) to the minority class which was the Medium risk class in order to resample the dataset. When applying this technique, new minority class instances are created based on the percentage of SMOTE for the minority class.

We obtained the test results of the predictive models by10-fold cross validation evaluation method. The predictive models 'resulted from the classification process' illustrate ways to identify whether the student at high, medium or low risk of failure.

4. EXPERIMENTAL RESULTS

The student datasets used in this study WAS analysed using Java API and Weka Mining tool with two classification algorithms used to develop the predictive models, those were Naïve Bayes and C4.5 Decision tree. A comparison of accuracy of the selected classification algorithms is provided in Table 4 and Figure 1. In fact, Algorithms and their Applications Module obtained the highest accuracy result in both Naïve Bayes and C4.5 decision tree (see Table 4) comparing to other Modules. However, all the predictive models produced accurate results in terms of (69%-84%) compared to what found in the literature.

Table 4. Accuracy Comparison of predictive models

Module title	**Naïve Bayes Accuracy**	**J48 Decision Accuracy**
Algorithms and their Applications	88.48%	84.29%
Usability Engineering	70.31%	70.31%
Software Development and Management	69.11%	75.39%
Year 2 Group Project	87.33%	84.16%

The sensitivity analysis of the predictive models summarised in Table 5 illustrates the comparison of True Positive rate (TP) and the False Positive rate (FP) of the applied algorithms (Naïve Bayes and C4.5 Decision tree) on different modules. The highlighted probabilities in the following table indicate the highest TP rates and the lowest FP rates were found at high risk failure for each specific module. In particular, the probability of correctly detection of high risk failure in "Algorithms and their application" module is identified by the highest TP rate of 0.969 and 0.938 exploiting Naïve Bayes and C4.5 Decision tree, respectively.

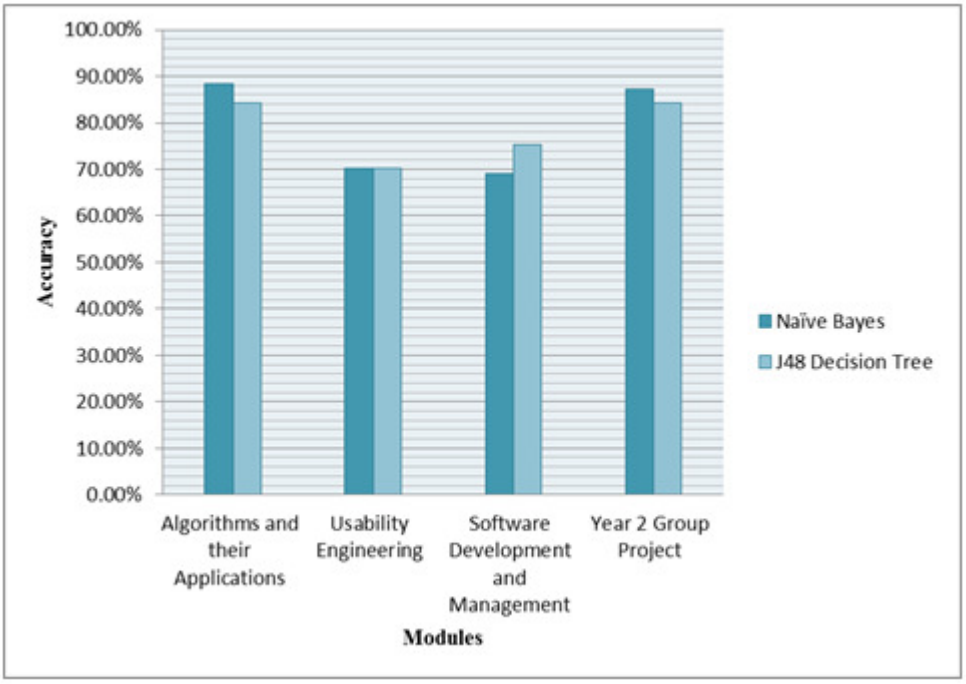

Figure 1. Accuracy Comparison of the predictive models

Table 5. TP Rate and FP Rate Comparison of the predicted modules

Module Title	Class	Naïve Bayes		C4.5 Decision Tree	
		TP Rate	FP Rate	TP Rate	FP Rate
Algorithms and their Applications	low risk	0.821	0.081	0.776	0.121
	medium risk	0.867	0.076	0.817	0.076
	high risk	**0.969**	**0.016**	**0.938**	**0.039**
Usability Engineering	low risk	0.671	0.205	0.714	0.213
	medium risk	0.192	0.066	0.192	0.066
	high risk	**0.865**	**0.219**	**0.833**	**0.208**
Software Development and Management	low risk	0.806	0.327	0.921	0.500
	medium risk	0.379	0.160	0.517	0.117
	high risk	**0.391**	**0.095**	**0.043**	**0.012**
Year 2 Group Project	low risk	0.732	0.072	0.610	0.056
	medium risk	0.882	0.039	0.882	0.059
	high risk	**0.920**	**0.083**	**0.902**	**0.147**

Figure 2 presents the best preform C4.5 decision tree model that predicts the students at high risk of failure. Student Overall Grade is the predicted feature in this classification model, and only a number of features were considered (8 of 33). Interestingly, remarkable result to emerge from the predictive model is that, student qualification has a high impact on the prediction of the high risk of failure students. Furthermore, some of level1 Modules final grades are highly influencing the prediction result. These Modules are Information Systems and Organisations Module, Logic and Computation Module and Software Implementation Event Module which are the core Modules of year 1 of computer science program.

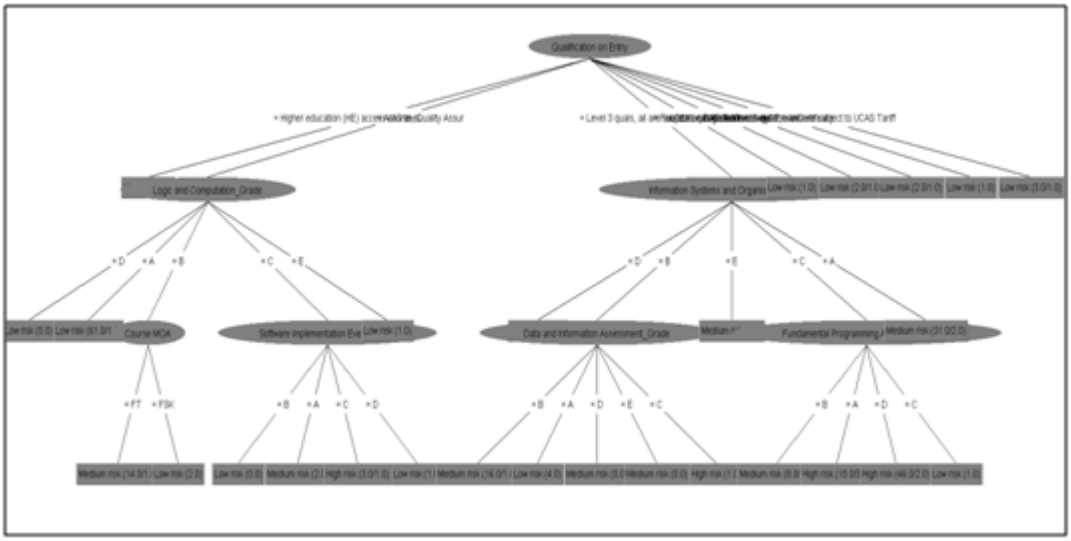

Figure 2. Algorithms and their Applications C4.5 Decision Tree Output

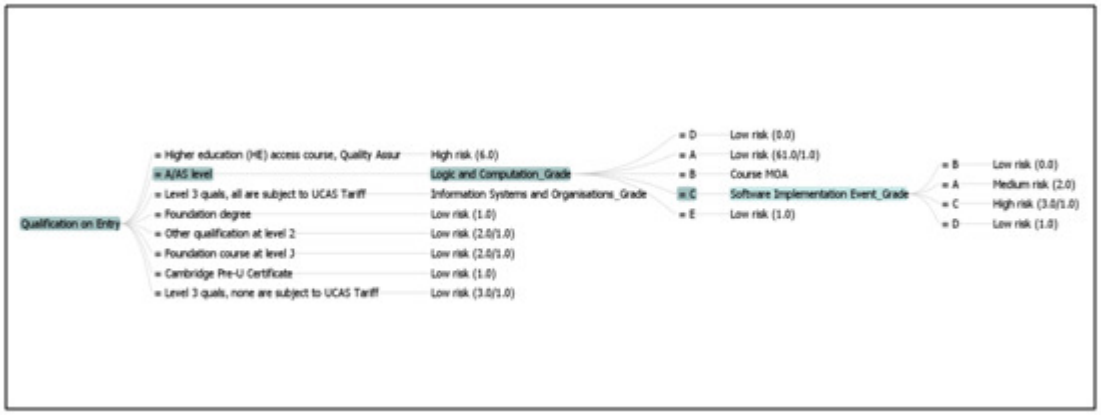

Figure 3. Algorithms and their Applications C4.5 Prefuse Tree Output

From the Prefuse tree in Figure 3 'which is Weka visualization tool that uses Prefuse toolkit to best explore the generated tree' we can extract some interesting rules that ended to high risk students. These rules indicate the influence of student's qualification on their academic performance in Algorithms and their Applications Module, for example:

1. **if** Qualification on entry **=** Higher Education (HE) access course **then** high risk**;**

2. **if** Qualification on entry **=** A/AS Level ∧Logic and Computation_Grade = C ∧ Software Implementation Event _Grade = C **then** high risk**;**

5. CONCLUSION AND FUTURE WORK

This study is an attempt to apply C4.5 and Naïve Bayes classification methods to analyse level 2 students' academic performance based on their admission, course related data and level 1 final grades. The main goal of the current investigation was to develop a predictive data mining model

for students' academic performance in university level so to identify the high risk of failure students. The second aim was to identify the key features affecting the predictive model.

By applying C4.5 and Naïve Bayes algorithms we revealed that Naïve Bayes performs better than C4.5 decision tree algorithm in predicting the students at high risk of failing the Module with an accuracy result of 88.48% for Naïve Bayes and 84.29% for C4.5 algorithm. Another major finding was that student qualifications on entry have high impact on students' academic performance. Moreover, some of level1 Modules final grades are influencing the results of the students in level2 Modules.

These findings provide the following insights for future investigation in Education Data Mining. The prediction of students' performance could be influenced by other factors or features. We are attempting to investigate other student's features that may influence the prediction process and provide better accuracy results. Moreover, different classification algorithms could be applied to obtain better predictive models using the same dataset.

ACKNOWLEDGEMENT

We would like to thank Dr Simon Kent, the director of education in the Department of Computer Science at Brunel University London for his professional cooperation throughout this study. Also, special thanks to Ms. Sara Brown, the student programmes manager at the department for her assistant in providing students datasets.

REFERENCES

[1] Hand, D. J., Mannila, H., and Smyth, P. (2001). Principles of Data Mining.MIT Press.

[2] Baker, R. (2010) Data Mining for Education. In McGaw, B., Peterson, P. and Baker, E. (Eds.) International Encyclopaedia of Education (3rd edition), vol. 7, pp. 112-118.Elsevier, Oxford, UK.

[3] Cortez, P., Silva, A., (2008) Using data mining to predict secondary school student performance. Presented at the 5th Annual Future Business Technology Conference, EUROSIS, pp. 5–12.

[4] Al-Radaideh, Q., Al-Shawakfa, E., Al-Najjar, M., (2006) Mining Student Data Using Decision Trees (PDF Download Available). Presented at the International Arab Conference on Information Technology (ACIT'2006), Jordan.

[5] Aher, S., L.M.R.J., L., (2011) Data Mining in Educational System using WEKA. Presented at the International Conference on Emerging Technology Trends (ICETT), International Journal of Computer Applications® (IJCA), pp. 20–25.

[6] Yadav, S., Pal, S., (2012) Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. World of Computer Science and Information Technology Journal (WCSIT) 2, 51–56.

[7] López, M.I., Luna, J.., Romero, C., Ventura, S., (2012) Classification via clustering for predicting final marks based on student participation in forums. Presented at the The 5th International Conference on Educational Data Mining, ERIC, pp. 148–151.

[8] Mayilvaganan, M. and Kalpanadevi, D., (2014) Comparison of Classification Techniques for predicting the performance of Students Academic Environment in: 2014 IEEE Conference on Communication and Network Technology (ICCNT), pp. 113-118

[9] Kabakchieva, D., (2013) Predicting Student Performance by Using Data Mining Methods for Classification. Cybern. Inf. Technol. 13, 61–72. doi:10.2478/cait-2013-0006

[10] Kaur, G., Singh, W., (2016) Prediction Of Student Performance Using Weka Tool. Vidya 17, 8–16.

[11] Kaur, P., Singh, M., Josan, G., (2015) Classification and prediction based data mining algorithms to predict slow learners in education sector. Presented at the 3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015), Elsevier B.V, pp. 500–508.

[12] Bekele, R., Menzel, W., (2005) A BAYESIAN APPROACH TO PREDICT PERFORMANCE OF A STUDENT (BAPPS): A Case with Ethiopian Students. Presented at the IASTED International Conference on Artificial Intelligence and Applications, part of the 23rd Multi-Conference on Applied Informatics, Innsbruck, Austria.

[13] Bhardwaj, B.K., Pal, S., (2012) Data Mining: A prediction for performance improvement using classification. ArXiv12013418 Cs.

[14] El-Halees, A., (2009) Mining Students Data to Analyze Learning Behavior: A Case Study (PDF Download Available). Dep. Comput. Sci. Islam. Univ. Gaza PO Box 108.

[15] Hogo, M.A., (2010) Evaluation of e-learning systems based on fuzzy clustering models and statistical tools. Expert Syst. Appl. 37, 6891–6903. doi:10.1016/j.eswa.2010.03.032

[16] Perera, D., Kay, J., Koprinska, I., Yacef, K., Zaiane, O., (2009) Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. IEEE transactions on knowledge and data engineering 21, 759–772. doi:10.1109/TKDE.2008.138

AUTHORS

Mashael Al-luhaybi is a Lecturer in eLearning and Distance Education at Umm Al-Qura University, Saudi Arabia. She is currently a PhD candidate in Machine learning in particular Educational Data Mining at Brunel University London, UK. She obtained her MSc from the University of Brighton, UK in 2011.She is interested in predicting student academic performance and detecting their learning behaviour.

Allan Tucker is a Senior Lecturer at Brunel University London, United Kingdom. He is the Head of Intelligent Data Analytics (IDA) Research Group at Brunel University. His research interests lie in modelling of brain function, human and animal behaviour. He obtained his PhD from Birkbeck, University of London.

Leila Yousefi is a PhD candidate in Machine Learning at Brunel University London, UK. She obtained her MSc from Azad University of Qazvin, Iran. She is a member of the Intelligent Data Analytics (IDA) Research Group at Brunel University London. Her research interest is in Artificial Intelligence in Medicine and Data Mining.