Knowledge-Based Systems 161 (2018) 134–146

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Predicting academic performance by considering student heterogeneity

Sumyea Helal[*,1,a], Jiuyong Li[*,1,a], Lin Liu[1,a], Esmaeil Ebrahimie[a], Shane Dawson[b], Duncan J. Murray[c], Qi Long[d]

[a] School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, Australia
[b] Centre for Change and Complexity in Learning, University of South Australia, Australia
[c] Business Intelligence and Planning, University of South Australia, Australia
[d] Yunnan Open University, China

ARTICLE INFO

Keywords:
Student performance prediction
Student heterogeneity
Learning management system
Educational data mining
WEKA

ABSTRACT

The capacity to predict student academic outcomes is of value for any educational institution aiming to improve student performance and persistence. Based on the generated predictions, students identified as being at risk of academic retention or performance can be provided support in a more timely manner. This study creates different classification models for predicting student performance, using data collected from an Australian university. The data include student enrolment details as well as the activity data generated from the university learning management system (LMS). The enrolment data contain student information such as socio-demographic features, university admission basis (e.g. via entry exam or past experience) and attendance type (e.g. full-time vs. part-time). The LMS data record student engagement with their online learning activities. An important contribution of this study is the consideration of student heterogeneity in constructing the predictive models. This is based on the observation that students with different socio-demographic features or study modes may exhibit varying learning motivations. The experiments validated the hypothesis that the models trained with instances in student sub-populations outperform those constructed using all data instances. Furthermore, the experiments revealed that considering both enrolment and course activity features aids in identifying vulnerable students more precisely. The experiments determined that no individual method exhibits superior performance in all aspects. However, the rule-based and tree-based methods generate models with higher interpretability, making them more useful for designing effective student support.

1. Introduction

At present, educational institutions are facing a highly competitive environment. As such, there is a need to ensure that resources are utilised effectively and efficiently to improve the student learning experience and promote esteem factors, such as student retention and performance. A challenge herein is to conduct an in-depth analysis of student academic performance that can aid in developing a student support strategy, and improve teaching and learning practices. In this regard, institutions may be interested in understanding student academic performance predictors. However, this is a complex task to solve, and a huge number of factors, such as economic, social, demographic, cultural and academic background, may influence academic outcomes [1]. Discovering significant student academic performance factors requires an in-depth analysis of the data, which may be achieved through educational data mining (EDM), a knowledge-discovery process that can provide valuable insights from data originating from an educational setting [2].

One of the most popular data mining methods, namely classification [3], has been successfully applied to predict performance. Classification is a supervised process of organising objects with similar characteristics into classes. Classification approaches can be broadly categorised into white box models (e.g. decision tree and rule-based), and black box models (e.g. artificial neural networks) [4]. For example, Ibrahim and Rusli [5] developed a neural network-based approach to predict student grades. The authors determined that students' basic knowledge regarding the course, prior schooling and financial status were the most important factors influencing performance. Although

[*] Corresponding authors.

E-mail addresses: sumyeahelal@gmail.com (S. Helal), Jiuyong.Li@unisa.edu.au (J. Li), Lin.Liu@unisa.edu.au (L. Liu), esmaeil.ebrahimie@adelaide.edu.au (E. Ebrahimie), shane.dawson@unisa.edu.au (S. Dawson), duncan.j.murray@unisa.edu.au (D.J. Murray), 654620695@qq.com (Q. Long).

[1] Present address: Data Analytic Group, School of Information Technology and Mathematical Sciences, Mawson Lakes, SA-5095 Australia.

https://doi.org/10.1016/j.knosys.2018.07.042
Received 2 December 2017; Received in revised form 29 July 2018; Accepted 31 July 2018
Available online 31 July 2018

such black box models can achieve higher prediction accuracy, interpreting the findings is a challenging process, thereby retarding meaningful actions. Certain researchers [6,7] have proposed different techniques for improving interpretation in several black box methods. For example, Villagrá-Arnedo et al. [6] considered three features, namely multi-class, probabilistic and progressiveness, in the design of a classifier to improve its interpretation level. In contrast, white box models such as decision trees and rule-based approaches discover rules in an if-then structure, depicting the knowledge in a more comprehensible manner that can be used directly for further decision-making [2]. For example, a classification tree-based method has been demonstrated to identify the influencing factors separating academically successful and unsuccessful students. Kovačić [8] found that ethnicity, program (e.g. Bachelor of Business) and course block (e.g. first or second trimester) had a strong association with student success. A number of studies [9,10] have attempted to employ both the black box and white box techniques to predict student performance by considering student participation in different online course resources and activities.

Prior works have also sought to incorporate all data instances in order to construct accurate and comprehensive classification models. However, the developments of these forms of global models are unlikely to produce quality predictions that can be readily interpreted. For example, Gaševič et al. [11] demonstrated that university-wide predictive models often fail to account for the subtleties in course design that influence student motivation, learning strategies and performance. Hence, the predicted outcomes derived from a global model may not be useful for informing support and intervention strategies, as the identified factors deemed to influence student performance may vary considerably across student subgroups.

This study considers the use of student sub-models generated from specific student subgroups to identify students at risk of academic failure. We aim to analyse the existing differences in student characteristics, considering key demographic and academic features that may influence academic performance. Therefore, the study aims to construct prediction models in different sub-populations, taking into account student gender, age and attendance type. These attributes are selected owing to their discriminating ability of student involvement in different academic activities. We refer to a model constructed in a subgroup as a *sub-model*, whereas the model built with all data instances is known as the *base model*.

This study evaluates the performance of the proposed approach in terms of both predictive ability and interpretability of the developed models. Hence, we apply two black box and two white box classification methods for generating the sub-models. The black box methods are naïve-Bayes [12] and sequential minimal optimizer (SMO) [13], while the white box methods consist of J48 [3] and JRip [14]. The findings demonstrate that, in most cases, these sub-models outperform the base model.

In this study, we consider student demographic, academic and course activity features separately as well as jointly in order to aid the identification of “at-risk” students.

Our work offers the following contributions.

1. We propose the concept of exploiting the heterogeneity in student characteristics for identifying significant predictors of student academic performance. In this regard, this study generates student sub-populations based on key demographic and academic features for constructing student sub-models, and evaluates their usefulness in identifying vulnerable students.
2. We conduct experiments with four different classification methods in order to validate the effectiveness of the proposed approach. The results demonstrate that specific student sub-models attain superior results to the original model.
3. Our experimental work also compares the performance of the different classification methods in constructing prediction models by considering students enrolment and activity features both separately and jointly. The results indicate that models generated using both enrolment and activity features outperform models constructed from individual features.

The remainder of this paper is organised as follows. The related literature is presented in Section 2. Section 3 discusses the methodology followed in this study. In Section 4, firstly, the datasets and their pre-processing procedure are described, secondly, student sub-models are generated from their enrolment and course activity features, and finally, the results are discussed. The interpretability and usefulness of the generated sub-models are analysed in Section 5. Finally, Section 6 concludes the paper and suggests future works.

2. Related works

The ability to predict student performance and identify students at risk of failure is an expanding research area. Various data mining techniques have been successfully applied to predict student academic performance. Thiele et al. [15] found that students’ socio-demographic (e.g. ethnicity, gender and economic status) and academic (e.g. type of school and their performance in that school) features are associated with their academic performance. Other works such as those of Guarín et al. [16] and Strecht et al. [17] demonstrated the effectiveness of considering academic records along with socio-demographic information during specific candidature enrolment in terms of generating higher-performing models with higher prediction accuracy. Other research [18,19] identified specific courses that serve as significant indicators of student academic performance and claimed that courses are not equally informative for making accurate predictions. These works also detected the typical progression of student performance throughout their study year and related these with the indicator courses. Alternate data sources were incorporated by gathering student study (study duration and focus) and social behaviour (partying) data from smart phones and it was found that these are highly correlated to their GPA [20].

Student performance prediction for specific courses is also a thoroughly studied area in data mining. A number of researchers [11,21,22] have demonstrated the effectiveness of using data from an institution learning management system (LMS) (e.g. Moodle [23], BlackBoard [24], Desire2Learn [25]), which accumulates a vast volume of student information related to courses, study activities and outcomes. It was demonstrated by Macfadyen and Dawson [26] that LMS tracking data can be used to predict the student final grade. In this instance, the authors noted that the quantities of discussion posts and mail messages sent were significant indicators of the student final grade. Khan et al. [27] illustrated that student participation in different activities of web-based courses may also aid in enhancing performance. Other authors [28] only considered data from current courses to generate a prediction model by presenting the concept of the self-learner. This prediction type may be useful when considering new courses and no prior data exists. Other approaches have sought to combine both behavioural data (such as LMS activity logs) and prior grades with socio-emotional and psychological data. For example, Adejo and Connelly [29] proposed a framework to consider the use of psychosocial-personality (SPP) data from a self-report survey in order to predict performance.

Although a number of works have been proposed to predict student academic outcomes, these tend to consider the entire student population in order to generate the models. However, this study takes into account the different student sub-groups for predicting academic outcomes.

3. Methodology

This study considers heterogeneity in different student sub-populations and constructs classification models in these sub-populations for

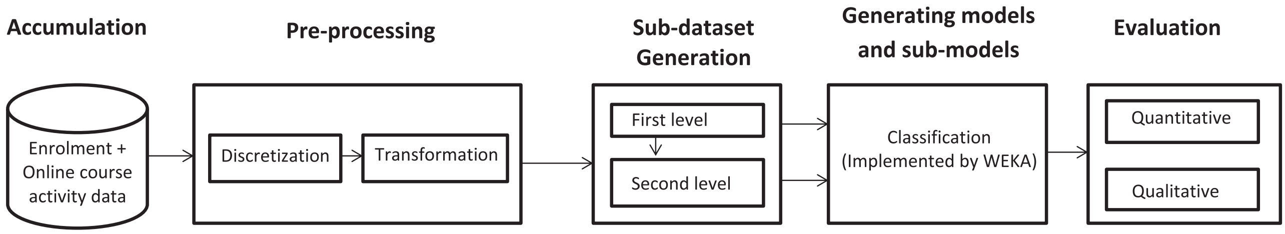


Fig. 1. Data mining approach for predicting student academic performance.

predicting academic outcomes. Firstly, the datasets are pre-processed; secondly, student subgroups are generated from the original datasets, considering certain significant demographic and student academic features; thirdly, different classification methods are applied to the sub-datasets to produce student sub-models; and finally, the sub-models are evaluated using different metrics in terms of their usability in decision-making. The approach applied in this study is illustrated in Fig. 1.

3.1. Collecting enrolment and LMS activity data

This study incorporates socio-demographic and academic data gathered during student enrolment, and activity data obtained from the university LMS - Moodle.

The enrolment dataset contains the socio-demographic (age, gender and economic status) and academic (attendance type and delivery mode) features of a student. In order to conduct experiments with this dataset, student performance is represented as the average mark in all the courses he/she has taken in a year. The data obtained from Moodle records the student participation in different activities (e.g. assignments, quizzes, forums and others) and resources (e.g. books and files). Each record of this dataset contains a student's frequency of involvement in different activities in a specific course. In the combined dataset, each student record possesses both the enrolment and participation features. As a LMS activity dataset record corresponds to a course taken by a student, there may be multiple records for different courses taken by the same student in the original dataset. Therefore, in the combined dataset, for the value of an activity feature for a student is the average counts of the student's participation in that particular activity in all courses he/she has taken during the year. The meanings of different enrolment and activity attributes can be found in Table 1.

3.2. Pre-processing of data

Data pre-processing is an important phase for preparing the data prior to applying data mining methods. In this study, pre-processing is conducted in two steps, namely discretisation and transformation. Most of the methods employed in this study operate only on categorical data; hence, all of the continuous attributes are discretised. Following discretisation, the datasets have been transformed into the appropriate format for ease of implementation. The pre-processing tasks conducted in this study are discussed below.

1. Discretisation: Discretisation is performed on the enrolment attributes AGE_NUM, AUST-SES and ATAR_rank[2] as well as on all activity attributes. All of the activity attributes are categorised into four quartiles, namely $Q1$, $Q2$, $Q3$ and $Q4$, where $Q1$ represents the lowest participation, and $Q4$, the highest.
2. Data transformation: Implementing the classification methods requires the data to be in the ARFF [30] format. The original enrolment and LMS activity dataset were in Excel format, and were transformed into the above-mentioned format, applicable to all executions.

3.3. Generating sub-datasets

The enrolment, activity and combined datasets are partitioned into several sub-datasets to form student subgroups. The dataset partitioning is performed in two steps, as follows.

1. Firstly, the enrolment, activity and combined datasets are partitioned according to student gender (male and female), age (normal and mature), attendance type (full-time and part-time) and attendance mode (internal and external). Hence, 8 sub-datasets are generated for each of the enrolment, activity and combined datasets, respectively.
2. Secondly, the female and male sub-datasets are further partitioned into another 6 sub-datasets according to student age, attendance type and attendance mode. The sub-datasets and their sizes can be seen in Table 2.

Initially, female, male, normal-aged, mature-aged, full-time, part-time, internal and external student sub-groups are generated for each dataset, which is termed as first-level sub-grouping throughout this paper. Thereafter, the male and female sub-groups are further sub-partitioned into normal-aged, mature-aged, full-time, part-time, internal and external sub-groups, which is referred to second-level sub-grouping.

3.4. Predicting student academic performance

We employ four classification methods for generating student sub-models. Among the classification approaches are two black box methods, namely naïve-Bayes and SMO. The remaining two approaches, J48 and JRip, are white box methods. For all executions, we use the implementations by the Waikato Environment for Knowledge Analysis (WEKA) [30]. WEKA, which was developed at the University of Waikato, New Zealand, is a software tool that provides collections of data mining and machine learning algorithms. The general features of the classification methods used in this study are discussed below.

Naïve-Bayes: This is a probabilistic classification method based on the Bayesian theorem. A naïve-Bayes classifier can be considered as the simplest Bayesian network classifier. This method is easy to implement and is particularly used with high-dimensional data. This classifier can be used with discrete or continuous attributes.

SMO: This method uses an optimisation algorithm for training a support vector machine (SVM) [13], and belongs to the type of functional classification that operates by revealing a function. The models generated by this method usually exhibit high classification accuracy. This method replaces missing values and can handle multi-class problems using pair-wise classification.

J48: This method generates a decision tree containing three

[2] *The primary criterion for entry into most undergraduate programs in any university of Australia, which represents a student's ranking relative to his/her peers upon completion of their secondary education.*

Table 1
Meaning of enrolment and course activity attributes used in experiments.

Type	Attribute	Meaning
Enrolment	GENDER	Students gender (Male/Female)
	SAS_ADM_CURRENT_SHORT_SAS	Admission basis for entry into the university (e.g. Mature age entry)
	AGE_NUM	Students age (If greater than 26 then Mature, otherwise Normal)
	ATTENDANCE_TYPE_DESC	How a student attends a course
	ATTENDANCE_MODE_SHORT_DESC	Whether a student attends in off-campus (External) or on-campus (Internal)
	AUST_SES	Australian Social Economic status (Determined by the living suburb)
	APPL_WAS_FIRST_PREF	Whether the program was preferred as first choice during enrolment
	IN_MULT_PROG_ANY_YR	Whether the student was admitted in multiple program in any year
	IN_MULT_PROG_THIS_YR	Whether the student is admitted in multiple program in current year
	PARENT_1_EDUCATION_CODE	Education status of a students male parent/guardian
	PARENT_2_EDUCATION_CODE	Education status of a students female parent/guardian
	HIGH_SCHOOL_STATE	The state in which the student attends his/her high school
	ATAR_rank	A students score of Australian Tertiary Admission Rank
LMS activity	BOOK_VIEW	Viewing book resources
	CHOICE_VIEW	Viewing choice activity
	COURSE_VISIT	Visiting course home page
	FORUM_ADD_DISCUSSION	Adding discussion in course forum
	FORUM_ADD_POST	Adding post in course forum
	FORUM_VIEW_DISCUSSION	Viewing discussion in course forum
	FORUM_VIEW_FORUM	Viewing forum activity
	LESSON_VIEW	Viewing lesson activity
	OUWIKI_EDIT	Editing course wiki
	OUWIKI_VIEW	Viewing course wiki
	QUIZ_ATTEMPT	Attempting quiz activity
	QUIZ_REVIEW	Reviewing quiz activity
	QUIZ_VIEW	Viewing quiz activity
	RESOURCE_VIEW	Viewing file resources

Table 2
Population sizes of different datasets.

Datasets	Sub-datasets	Cohort size		
		Enrolment	LMS activity	Combined
Full training set		2648	7052	2648
	Female	1986	5211	1986
	Male	662	1841	662
	Fulltime	2160	6123	2160
	Part-time	488	929	488
	Internal	2101	5955	2101
	External	547	1097	547
	Normal	1909	5339	1909
	Mature	739	1713	739
Female	Fulltime	1570	4428	1570
	Part-time	416	783	416
	Internal	1401	3987	1401
	External	585	1224	585
	Normal	1369	3820	1369
	Mature	617	1391	617
Male	Fulltime	590	1695	590
	Part-time	72	146	72
	Internal	513	1625	513
	External	149	216	149
	Normal	491	1519	491
	Mature	171	322	171

different node types: root, internal and leaf nodes. The root node is the top-most node in a tree. The root and internal nodes contain attribute test conditions, where each branch represents an outcome of the test and each leaf node represents a class level. Decision tree-based classifiers exhibit high accuracy and are easy to implement.

JRip: This is a rule-based classification method that generates comprehensible rules in an IF-THEN structure. An IF-THEN rule is expressed as *IF condition THEN conclusion*. Similar to the decision tree, this method is easy to implement, and the generated models are highly interpretable. This classifier type can easily handle missing values.

3.5. Evaluation of generated models

A number of criteria have been developed for measuring the predictive ability of a model. In this study, we use several metrics that are commonly encountered in the existing literature to assess the performance of different methods in terms of the generated models, including the following.

- Precision [31]: the fraction of true positive examples among all examples classified as positive by a classifier.
- Recall [31]: the fraction of true positive examples classified correctly by a classifier.
- F-measure [32]: the harmonic mean of the precision and recall of a classifier; that is, F-measure = 2 × precision × recall/(precision + recall).
- Kappa co-efficient [33]: compares the accuracy of a classifier with the accuracy a random classifier is expected to achieve. The Kappa value is less than or equal to 1, where 1 denotes perfect prediction by the classifier and 0 indicates no better than a random guess.
- AUC [34]: the area under the receiver operating characteristic (ROC) curve indicates the probability that a classifier will rank a randomly selected positive example more highly than a randomly selected negative example. An AUC value of 1 indicates a perfect classifier, while 0.5 implies that the classifier performs as random guesses.

4. Experiments

In this section, we follow the methodology described in Section 3 to construct and evaluate the prediction models generated for student subpopulations. We use student enrolment and LMS activity data separately as well as jointly, and employ four classification methods to generate student sub-models. First, the details of the datasets used in the experiments (Section 4.1) are provided. Then, in Section 4.2, the experiment results, including the performances of the different models constructed, measured using the metrics discussed in Section 3.5 are presented. Finally, in order to evaluate the model performances further, we use cross-validation to assess their prediction accuracy (Section 4.3).

Table 3
Summary of datasets.

Dataset	No. of Instances		Attributes
	Training	Test	
Enrolment	2648	1362	13
LMS Activity	7052	3916	14
Combined	2648	1362	27

4.1. Datasets

The datasets used in this paper were collected from 2011 to 2013 from a division (akin to a faculty comprising multiple disciplinary schools) in an Australian university regarding their first-year domestic undergraduate students. Three dataset types are employed in this work, namely enrolment data, activity data, and combined data containing both enrolment and activity features. For each course (the smallest unit of study in a program), a student's performance (passing or failing the course) is also included in the dataset as the experimental target variable.

Table 3 displays the training and testing datasets used in the experiments in Section 4.2, for training and testing the model performances, respectively. A training set includes the data (enrolment, activity, or combined) of years 2011 and 2012; a testing set contains the 2013 data. For the 10-fold cross-validation results presented in Section 4.3, we use all data from the three years (details can be found in Section 4.3).

4.2. Results

In this section, we construct student performance prediction models in each subgroup displayed in Table 2. Moreover, in order to investigate the effectiveness of constructing models in subgroups, we also build the base model in the entire population, and compare the sub-model and base model performances. These sub-models are constructed for each of the three datasets: (i) enrolment data, (ii) activity data obtained from the course Moodle and (iii) combined data containing both enrolment and activity features. The following subsections discuss the prediction ability of the models generated from each dataset, where the best results are boldfaced.

4.2.1. Predicting student performance using enrolment data

In this section, we discuss the performance of student sub-models in terms of identifying at-risk students using enrolment data. As indicated in Table 4, it is found that most sub-models achieve superior results to the base model in terms of the criteria discussed in Section 3.5. The sub-model representing external students performs best in identifying unsuccessful students. Among all methods, *SMO* achieves the best results in building this sub-model, with F-measure and Kappa values of 51%

Table 4
Mining enrolment data.

Method	Dataset						Female					Male				
		Precision	Recall	F-measure	Kappa	AUC	Precision	Recall	F-measure	Kappa	AUC	Precision	Recall	F-measure	Kappa	AUC
Naïve-Bayes	Original dataset	0.276	0.128	0.18	0.102	0.504	–	–	–	–	–	–	–	–	–	-
	Female	0.411	0.196	0.265	0.138	0.587	–	–	–	–	–	–	–	–	-	–
	Male	0.333	0.143	0.2	0.125	0.593	–	–	–	–	–	–	–	–	-	-
	Fulltime	0.495	0.194	0.278	0.167	0.617	0.409	0.153	0.222	0.12	0.605	0.45	0.385	0.435	0.256	0.641
	Part-time	0.444	0.353	0.393	0.186	0.681	0.429	0.286	0.343	0.109	0.648	0.318	0.538	0.4	-0.15	0.456
	Normal	0.471	0.25	0.327	0.134	0.649	0.457	0.246	0.319	0.19	0.653	0.39	0.355	0.392	0.21	0.568
	Mature	0.425	0.218	0.288	0.126	0.681	0.483	0.259	0.337	0.202	0.678	0.316	0.316	0.316	-0.016	0.547
	Internal	0.511	0.263	0.347	0.22	0.65	0.386	0.199	0.262	0.142	0.634	0.465	0.439	0.452	0.213	0.64
	External	**0.548**	**0.317**	**0.402**	**0.268**	**0.702**	0.458	0.208	0.286	0.116	0.655	**0.49**	**0.495**	**0.492**	**0.27**	**0.706**
J48	Original dataset	0.34	0.2	0.26	0.18	0.489	–	–	–	–	–	–	–	–	–	–
	Female	0.565	0.231	0.328	0.225	0.608	–	–	–	–	–	–	–	–	–	–
	Male	0.375	0.429	0.4	0.21	0.598	–	–	–	–	–	–	–	–	–	–
	Fulltime	0.542	0.326	0.407	0.25	0.704	0	0	0	0	0.5	0.367	0.515	0.44	0.31	0.668
	Part-time	0.444	0.353	0.393	0.18	0.581	0.424	0.333	0.313	0.119	0.541	0.313	0.385	0.345	-0.13	0.425
	Normal	0.306	0.172	0.22	0.013	0.523	0.571	0.075	0.133	0.067	0.497	0.492	0.642	0.557	0.326	0.68
	Mature	0.364	0.103	0.16	0.045	0.554	0	0	0	0	0.5	0.25	0.158	0.194	-0.108	0.402
	Internal	0.476	0.219	0.299	0.17	0.497	0	0	0	0	0.5	0.486	0.505	0.495	0.26	0.638
	External	**0.577**	**0.407**	**0.477**	**0.27**	**0.714**	**0.657**	**0.257**	**0.37**	**0.278**	**0.703**	**0.71**	**0.615**	**0.66**	**0.51**	**0.724**
SMO	Original dataset	0.322	0.066	0.1	0.075	0.526	–	–	–	–	–	–	–	–	–	–
	Female	0.385	0.122	0.18	0.17	0.576	–	–	–	–	–	–	–	–	–	–
	Male	0.385	0.156	0.222	0.146	0.569	–	–	–	–	–	–	–	–	–	–
	Fulltime	0.5	0.136	0.21	0.113	0.537	0.444	0.023	0.043	0.023	0.508	0.6	0.438	0.509	0.38	0.625
	Part-time	0.463	0.412	0.436	0.24	0.648	0.429	0.214	0.286	0.086	0.537	0.389	0.538	0.452	0.0137	0.507
	Normal	0.569	0.123	0.2	0.125	0.546	0.395	0.099	0.159	0.075	0.528	0.5	0.568	0.532	0.31	0.661
	Mature	0.42	0.13	0.2	0.18	0.566	0.333	0.019	0.035	0.009	0.503	0.267	0.211	0.235	-0.109	0.448
	Internal	0.385	0.156	0.22	0.14	0.584	0.333	0.068	0.113	0.047	0.516	0.486	0.486	0.486	0.26	0.626
	External	**0.508**	**0.513**	**0.51**	**0.28**	**0.751**	0.5	0.019	0.036	0.014	0.505	0.65	**0.538**	**0.59**	**0.48**	**0.676**
JRip	Original dataset	0.395	0.109	0.18	0.138	0.506	–	–	–	–	–	–	–	–	–	–
	Female	0.537	0.196	0.287	0.187	0.572	–	–	–	–	–	–	–	–	–	–
	Male	0.414	0.403	0.409	0.149	0.567	–	–	–	–	–	–	–	–	–	–
	Fulltime	0.419	0.14	0.21	0.1	0.54	0.684	0.147	0.242	0.33	0.564	0.343	0.492	0.467	0.43	0.662
	Part-time	0.33	0.32	0.32	0.06	0.532	0.5	0.214	0.3	0.129	0.555	0.382	0.351	0.37	0.42	0.642
	Normal	0.486	0.19	0.27	0.16	0.562	0.556	0.205	0.299	0.2	0.577	0.507	0.389	0.44	0.24	0.606
	Mature	0.545	0.231	0.324	0.193	0.579	0.563	0.167	0.257	0.16	0.563	0.222	0.105	0.143	-0.1	0.453
	Internal	0.405	0.234	0.297	0.08	0.539	0.583	0.174	0.268	0.19	0.57	0.43	0.43	0.43	0.17	0.507
	External	**0.517**	**0.341**	**0.41**	**0.27**	**0.719**	**0.455**	**0.094**	**0.156**	**0.056**	**0.523**	**0.421**	**0.8**	**0.552**	**0.47**	**0.683**

Table 5
Mining LMS activity data.

Method	Dataset						Female					Male				
		Precision	Recall	F-measure	Kappa	AUC	Precision	Recall	F-measure	Kappa	AUC	Precision	Recall	F-measure	Kappa	AUC
Naïve-Bayes	Original dataset	0.294	0.388	0.33	0.23	0.603	–	–	–	–	–	–	–	–	–	–
	Female	0.447	0.658	0.532	0.374	0.797	–	–	–	–	–	–	–	–	–	–
	Male	0.597	0.735	0.659	0.48	0.834	–	–	–	–	–	–	–	–	–	–
	Fulltime	0.497	0.687	0.577	0.42	0.816	0.45	0.658	0.535	0.385	0.804	0.591	0.735	0.655	0.477	0.836
	Part-time	0.487	0.655	0.558	0.347	0.768	0.453	0.629	0.527	0.315	0.746	0.538	0.737	0.622	0.409	0.829
	Normal	0.508	0.688	0.584	0.427	0.818	0.456	0.648	0.535	0.382	0.797	0.511	0.694	0.579	0.425	0.752
	Mature	0.517	0.673	0.584	0.42	0.797	0.474	0.655	0.55	0.397	0.809	0.604	0.615	0.61	0.395	0.769
	Internal	0.483	0.677	0.564	0.41	0.807	0.416	0.641	0.504	0.355	0.789	0.602	0.731	0.66	0.487	0.834
	External	**0.657**	**0.762**	**0.7**	**0.53**	**0.8**	**0.649**	**0.759**	**0.7**	**0.52**	**0.879**	**0.711**	**0.76**	**0.71**	**0.54**	**0.799**
J48	Original dataset	0.363	0.285	0.32	0.31	0.605	–	–	–	–	–	–	–	–	–	–
	Female	0.633	0.466	0.537	0.436	0.8	–	–	–	–	–	–	–	–	–	–
	Male	0.722	0.628	0.672	0.51	0.816	–	–	–	–	–	–	–	–	–	–
	Fulltime	0.662	0.552	0.602	0.495	0.806	0.63	0.455	0.528	0.43	0.807	0.706	0.635	0.669	0.533	0.811
	Part-time	0.652	0.536	0.588	0.448	0.795	0.617	0.468	0.532	0.39	0.761	0.375	0.474	0.419	0.109	0.629
	Normal	0.645	0.608	0.626	0.51	0.802	0.604	0.471	0.529	0.423	0.797	0.614	0.528	0.515	0.414	0.681
	Mature	0.728	0.512	0.601	0.5	0.8	0.744	0.555	0.635	0.552	0.832	0.75	0.346	0.474	0.324	0.75
	Internal	0.631	0.587	0.608	0.5	0.793	0.594	0.449	0.511	0.415	0.793	0.707	0.627	0.664	0.528	0.814
	External	**0.736**	**0.623**	**0.675**	**0.53**	**0.861**	**0.791**	**0.607**	**0.687**	**0.565**	**0.886**	**0.721**	**0.671**	**0.695**	**0.568**	**0.817**
SMO	Original dataset	0.338	0.229	0.27	0.29	0.598	–	–	–	–	–	–	–	–	–	–
	Female	0.603	0.587	0.595	0.488	0.742	–	–	–	–	–	–	–	–	–	–
	Male	0.692	0.511	0.59	0.52	0.744	–	–	–	–	–	–	–	–	–	–
	Fulltime	0.629	0.641	0.635	0.52	0.723	0.591	0.601	0.596	0.49	0.747	0.385	0.695	0.539	0.489	0.778
	Part-time	0.608	0.738	0.667	0.52	0.736	0.568	0.677	0.618	0.46	0.745	0.63	0.655	0.629	0.459	0.728
	Normal	0.626	0.652	0.639	0.53	0.725	0.58	0.595	0.587	0.476	0.74	0.618	0.506	0.598	0.453	0.719
	Mature	0.718	0.519	0.602	0.5	0.726	0.744	0.555	0.635	0.552	0.751	0.676	0.481	0.562	0.381	0.678
	Internal	0.616	0.636	0.626	0.514	0.74	0.558	0.597	0.577	0.47	0.743	0.706	0.667	0.686	0.55	0.772
	External	**0.778**	**0.592**	**0.672**	**0.54**	**0.794**	**0.791**	0.607	**0.687**	**0.56**	**0.765**	**0.78**	**0.71**	**0.72**	**0.59**	**0.88**
JRip	Original dataset	0.426	0.217	0.29	0.31	0.581	–	–	–	–	–	–	–	–	–	–
	Female	0.627	0.465	0.534	0.43	0.695	–	–	–	–	–	–	–	–	–	–
	Male	0.711	0.651	0.64	0.48	0.668	–	–	–	–	–	–	–	–	–	–
	Fulltime	0.61	0.64	0.62	0.51	0.768	0.63	0.455	0.528	0.43	0.695	0.68	0.66	0.67	0.49	0.767
	Part-time	0.589	0.667	0.63	0.47	0.765	0.705	0.5	0.585	0.47	0.713	0.69	0.65	0.63	0.47	0.717
	Normal	0.643	0.604	0.623	0.51	0.757	0.58	0.48	0.52	0.4	0.695	0.43	0.58	0.49	0.38	0.694
	Mature	0.7	0.55	0.614	0.51	0.735	0.71	0.56	0.63	0.54	0.751	0.68	0.44	0.54	0.36	0.664
	Internal	0.61	0.65	0.63	0.51	0.708	0.595	0.429	0.5	0.403	0.682	0.55	0.48	0.49	0.34	0.763
	External	**0.76**	**0.67**	**0.71**	**0.58**	**0.87**	**0.73**	**0.63**	**0.67**	**0.56**	**0.77**	0.76	0.65	**0.71**	**0.58**	**0.779**

and 28%, respectively. This method also obtains a higher AUC of 75.1% for the external sub-model.

The performance of the second-level sub-models can be observed in the final two columns, *female* and *male*, in Table 4. The results demonstrate that not all, but some, specific sub-models outperform the corresponding base models (female/male). The *J*48 method achieves the highest precision for the female-external and male-external sub-models, with 65.7% and 71%, respectively. For the male-external sub-model, this method attains the highest score in terms of F-measure, Kappa and AUC values as well, with 66%, 51% and 72.4%, respectively.

4.2.2. Predicting student performance using LMS activity data

The experiments reveal that all of the sub-models attain superior results to the base model in all aspects, as illustrated in Table 5. Among these, the external sub-model performs better in detecting at-risk students, with F-measure, Kappa and AUC values above 66%, 53% and 78%, respectively. The JRip model secures the highest score in terms of F-measure, Kappa and AUC values, with 71%, 58% and 87%, respectively.

The performances of the second-level sub-models constructed from the female and male students can also be observed in Table 5. Only the female-mature and female-external sub-models exhibit superior prediction results to the female sub-model. The *J*48 and *SMO* models attain the highest precision of 79.1% for the female-external sub-model, which also results in a higher F-measure of 68.7%. However, the male-external sub-model attains superior performance to its base model (male). The male-external sub-model achieves the best result among all sub-models generated from the male and female students. The experiments also demonstrate that *SMO* performs best among all methods in generating this sub-model, with an F-measure of 72% and Kappa value of 59%, while the AUC value achieved by this sub-model is 88%.

4.2.3. Predicting student performance using combined data

The experiments performed on the combined dataset outperform those of the base model (see Table 6). The external sub-model secures the best prediction result compared to the other sub-models, as indicated by a precision above 80% for the different methods. The naïve-Bayes method achieves the highest precision and recall, at 86% and 67.8%, respectively, for the external sub-model, which consequently results in the highest F-measure value of 76% among all sub-models. Furthermore, this method obtains the highest performance in terms of Kappa and AUC values, at 61% and 89%, respectively.

As depicted in Table 6, specific female and male sub-models outperform the base model (female/male). The experiments demonstrate that the female-external, female-part-time and female-mature sub-models achieve superior results compared to the female model. The JRip model exhibits the highest performance in generating these sub-models in terms of both the F-measure and Kappa values, with F-measure values of 70%, 70% and 65%, and Kappa values of 59%, 58% and 55%, respectively, for the sub-models. This method also attains the

Table 6
Mining combined data.

Method	Dataset						Female					Male				
		Precision	Recall	F-measure	Kappa	AUC	Precision	Recall	F-measure	Kappa	AUC	Precision	Recall	F-measure	Kappa	AUC
Naïve-Bayes	Original dataset	0.682	0.374	0.48	0.43	0.644	–	–	–	–	–	–	–	–	–	–
	Female	0.802	0.431	0.561	0.477	0.832	–	–	–	–	–	–	–	–	–	–
	Male	0.7	0.529	0.603	0.454	0.825	–	–	–	–	–	–	–	–	–	–
	Fulltime	0.758	0.448	0.563	0.47	0.84	0.812	0.39	0.527	0.45	0.827	0.7	0.56	0.622	0.48	0.823
	Part-time	0.64	0.571	0.604	0.42	0.811	0.686	0.571	0.623	0.48	0.83	0.5	0.2	0.286	0.035	0.65
	Normal	0.793	0.485	0.602	0.51	0.856	0.814	0.409	0.545	0.464	0.839	0.532	0.505	0.523	0.325	0.813
	Mature	0.854	0.449	0.588	0.49	0.838	0.867	0.481	0.629	0.5	0.845	0.423	0.372	0.41	0.14	0.645
	Internal	0.671	0.515	0.583	0.47	0.833	0.614	0.435	0.509	0.408	0.801	0.704	0.553	0.62	0.47	0.824
	External	**0.86**	**0.678**	**0.76**	**0.61**	**0.89**	**0.825**	**0.604**	**0.67**	**0.53**	**0.786**	**0.802**	**0.624**	**0.683**	**0.548**	**0.853**
J48	Original dataset	0.646	0.254	0.37	0.31	0.699	–	–	–	–	–	–	–	–	–	–
	Female	0.831	0.262	0.399	0.324	0.801	–	–	–	–	–	–	–	–	–	–
	Male	0.754	0.387	0.511	0.38	0.771	–	–	–	–	–	–	–	–	–	–
	Fulltime	0.842	0.229	0.361	0.29	0.811	0.808	0.237	0.367	0.3	0.79	0.45	0.39	0.37	0.34	0.782
	Part-time	0.813	0.464	0.591	0.46	0.734	0.803	0.411	0.609	0.472	0.689	0.76	0.42	0.638	0.41	0.567
	Normal	0.78	0.265	0.396	0.3	0.818	0.74	0.216	0.335	0.26	0.728	0.457	0.383	0.43	0.25	0.724
	Mature	0.885	0.295	0.442	0.35	0.779	0.829	0.241	0.382	0.31	0.709	0.351	0.383	0.389	0.155	0.679
	Internal	0.855	0.174	0.289	0.227	0.795	0.593	0.099	0.17	0.118	0.737	0.788	0.398	0.529	0.4	0.793
	External	**0.806**	**0.453**	**0.58**	**0.45**	**0.825**	**0.913**	**0.436**	**0.68**	**0.53**	**0.798**	**0.857**	**0.587**	**0.7**	**0.55**	**0.811**
SMO	Original dataset	0.604	0.212	0.32	0.28	0.602	–	–	–	–	–	–	–	–	–	–
	Female	0.839	0.231	0.362	0.29	0.609	–	–	–	–	–	–	–	–	–	–
	Male	0.676	0.387	0.492	0.34	0.65	–	–	–	–	–	–	–	–	-	-
	Fulltime	0.75	0.301	0.43	0.34	0.635	0.75	0.169	0.276	0.216	0.577	0.452	0.44	0.43	0.36	0.681
	Part-time	0.676	0.446	0.538	0.365	0.668	0.8	0.526	0.637	0.54	0.709	0.462	0.35	0.36	0.3	0.617
	Normal	0.75	0.291	0.419	0.325	0.856	0.667	0.199	0.306	0.226	0.585	0.689	0.333	0.449	0.31	0.633
	Mature	0.886	0.397	0.549	0.456	0.838	0.955	0.489	0.653	0.54	0.691	0.334	0.297	0.278	0.13	0.501
	Internal	0.689	0.304	0.422	0.322	0.833	0.596	0.211	0.312	0.23	0.587	0.667	0.33	0.442	0.29	0.627
	External	**0.853**	**0.453**	**0.592**	**0.475**	**0.87**	**0.920**	**0.534**	**0.69**	**0.56**	**0.809**	**0.892**	**0.55**	**0.71**	**0.57**	**0.812**
JRip	Original dataset	0.54	0.22	0.29	0.28	0.601	–	–	–	–	–	–	–	–	–	–
	Female	0.86	0.25	0.39	0.32	0.621	–	–	–	–	–	–	–	–	–	–
	Male	0.63	0.3	0.4	0.3	0.628	–	–	–	–	–	–	–	–	–	–
	Fulltime	0.79	0.3	0.44	0.35	0.638	0.86	0.28	0.42	0.35	0.633	0.72	0.36	0.48	0.35	0.649
	Part-time	0.68	0.45	0.54	0.37	0.663	0.85	0.56	0.7	0.58	0.766	0.44	0.8	0.57	-0.03	0.483
	Normal	0.8	0.31	0.45	0.36	0.644	0.857	0.175	0.29	0.23	0.584	0.49	0.39	0.35	0.31	0.664
	Mature	0.79	0.42	0.55	0.37	0.646	0.89	0.44	0.65	0.55	0.713	0.67	0.17	0.28	0.14	0.56
	Internal	0.86	0.23	0.36	0.29	0.609	0.8	0.2	0.32	0.26	0.593	0.49	0.44	0.42	0.33	0.691
	External	**0.85**	**0.44**	**0.58**	**0.46**	**0.7**	**0.89**	**0.57**	**0.7**	**0.59**	**0.905**	**0.85**	**0.66**	**0.72**	**0.59**	**0.731**

highest AUC value for the female-external sub-model, at 90.5%. It is observed that the male-external sub-model exhibits superior performance in identifying unsuccessful students than the male model. The experiments demonstrate that the male-external sub-model attains the highest score in terms of both F-measure and Kappa, while JRip exhibits superior performance in generating this sub-model, with F-measure and Kappa values of 72% and 59%, respectively.

4.3. Evaluation of generated models using cross-validation and statistical test

In the previous section, we evaluated the prediction models by using the 2011 and 2012 samples to train the models, and the 2013 samples to test the performance. This setting is close to practical situations, where an institution uses historical data to construct models for future student performance prediction.

In this section, in order to evaluate the model performances further (purely from a model evaluation perspective), we use the stratified 10-fold cross-validation implemented in WEKA, with all three years of data. Specifically, for example, with the experiments on the enrolment data, during each run of the 10-fold cross-validation, the three years of enrolment samples are firstly divided into 10 equal-sized and disjoint subsets (folds), each containing roughly the same portion of samples for each class value as in the entire dataset. Then, in each iteration (of the 10 iterations), we sequentially use one fold of samples for testing and the remaining 9 folds for training. Complementary to the metrics used in the previous section, we use accuracy; that is, the ratio of correctly predicted samples over the total number of (testing) samples to measure the prediction performance. Five runs of the 10-fold cross-validation are carried out and the average accuracy and root mean squared error of the predictions are recorded.

Tables 7, 8 and 9 display the accuracy and error values obtained by the naïve-Bayes, J48, SMO and JRip methods for the enrolment, activity and combined datasets, respectively. We can observe that sub-models generated from the first-level subgrouping outperform the base model by achieving a higher accuracy. Certain specific student sub-models, such as full-time, normal-aged and internal, generated from the second-level subgrouping, achieve superior results to their corresponding base models (female and male).

In comparing the accuracy of a sub-model with the model (a second-level sub-model with a sub-model) given a method in Tables 7, 8 and 9, the significance of the accuracy difference is determined by a two-tailed *t*-test using the STAC tool [35]. The results are obtained from 20 runs of 10-fold cross-validations. The results demonstrate that all of the first-level sub-models generated from student sub-populations attain superior prediction results to their respective base models trained using all data instances. Furthermore, the experiments reveal that certain second-level sub-models outperform the female and male sub models. One reason for the second-level sub-models not performing as effectively as the first-level sub-models is owing to the small number of

Table 7
Cross-validation results for enrolment dataset. The accuracy of a sub-model built on a sub-dataset is compared to that of the model built on the entire dataset (and with a male or female sub-model for a second-level sub-model). The following notations are used to indicate the significance of the accuracy difference: “**” for p-value < 0.01, “*” for p-value < 0.05 and “ns” (not significant) for p-value > 0.05.

Method	Datasets			Female		Male	
		Accuracy (%)	Error	Accuracy (%)	Error	Accuracy (%)	Error
Naïve-Bayes	Original dataset	70.05	0.45	–	–	–	–
	Female	72.66 **	0.43	–	–	–	–
	Male	70.97 *	0.44	–	–	–	–
	Fulltime	73.67 **	0.43	**74.39** **	**0.42**	**72.48** **	**0.43**
	Part-time	70.63 *	0.44	62.86 (ns)	0.48	68.74 (ns)	0.47
	Normal	74.2 **	0.42	**74.03** **	**0.42**	**71.4** **	**0.43**
	Mature	71.75 *	0.44	70.24 (ns)	0.45	66.84 (ns)	0.48
	Internal	73.45 **	0.42	**74.08** **	**0.42**	**71.71** **	**0.43**
	External	71.19 *	0.44	67.52 (ns)	0.47	68.22 (ns)	0.49
J48	Original dataset	70.54	0.45	–	–	–	–
	Female	74.94 **	0.423	–	–	–	–
	Male	71.6 *	0.43	–	–	–	–
	Fulltime	76.97 **	0.41	**78** **	**0.4**	**72.56** **	**0.41**
	Part-time	72.44 *	0.42	65.36 (ns)	0.47	58.82 (ns)	0.54
	Normal	76.19 **	0.41	**76.98** **	**0.41**	**73.51** **	**0.4**
	Mature	71.3 *	0.43	69.52 (ns)	0.46	61.96 (ns)	0.5
	Internal	76.13 **	0.41	**77.15** **	**0.4**	**72.04** **	**0.42**
	External	71.3 *	0.43	69.55 (ns)	0.46	65.89 (ns)	0.49
SMO	Original dataset	72.04	0.53	–	–	–	–
	Female	73.96 **	0.51	–	–	–	–
	Male	71.89 *	0.52	–	–	–	–
	Fulltime	74.63 **	0.5	**76.13** **	**0.48**	**73.37** **	**0.51**
	Part-time	72.89 *	0.52	74.12 **	0.5	70.03 (ns)	0.54
	Normal	73.87 **	0.51	**76.51** **	**0.48**	**74.28** **	**0.5**
	Mature	73.06 **	0.51	71.07 (ns)	0.53	71.95 *	0.54
	Internal	73.32 **	0.52	**74.54** **	**0.5**	**75.01** **	**0.5**
	External	72.56 *	0.52	67.7 (ns)	0.56	72.93 *	0.53
JRip	Original dataset	73.79	0.44	–	–	–	–
	Female	74.33 **	0.43	–	–	–	–
	Male	73.24 **	0.42	–	–	–	–
	Fulltime	74.87 **	0.42	**76.8** **	**0.42**	**74.87** **	**0.41**
	Part-time	74.79 **	0.42	65 (ns)	0.47	73.65 *	0.43
	Normal	73.84 *	0.43	**76.37** **	**0.42**	**75.12** **	**0.4**
	Mature	73.98 *	0.43	71.9 (ns)	0.44	72.41 (ns)	0.44
	Internal	74.55 **	0.42	**76.42** **	**0.41**	**75.03** **	**0.4**
	External	74.02 **	0.43	69.43 (ns)	0.45	73.82 *	0.44

instances in the sub-sub-datasets.

5. Discussion

The knowledge derived from the prediction models may be helpful for educational institutions and course instructors to detect students who are at risk of academic failure early in their studies, so that proactive support strategies can be implemented in a timely manner. Hence, it is important to evaluate the interpretability of the discovered sub-models for identifying at-risk students in order to assess their usefulness for decision-making purposes.

This study demonstrates that the results derived from the sub-models produce a higher degree of accuracy than the base model. However, the male, part-time, external and mature sub-models exhibit superior performance in identifying low achievers in terms of both the F-measure and Kappa values. Similar findings are observed from the second-level subgrouping experiments. The female students predicted to be unsuccessful are either part-time, external or mature-aged. The results suggest that students who are male and either study part-time or attend a course on an external basis are predicted to possess a higher risk level for academic failure.

5.1. Comparing performance of different methods in generating student sub-models

In this study, we evaluate the generated models and sub-models in terms of their exactness (precision) and completeness (recall) in detecting unsuccessful students. However, no single method exhibits superior performance in terms of analysing the different datasets. For the first-level subgrouping experiments, SMO achieves the best results in mining the enrolment dataset, while JRip and naïve-Bayes perform best in mining the activity and combined data, respectively. For second-level subgrouping, *J*48 attains the best result for mining enrolment data, while SMO and JRip exhibit superior performance in mining the activity and combined datasets, respectively. Moreover, JRip performs best in predicting both the successful and unsuccessful students and correctly classifies above 83% of students for the models generated from the combined dataset, as illustrated in Fig. 2. This figure represents the percentage of correctly classified students (for both successful and unsuccessful students) of the external sub-model for the enrolment, activity and combined datasets.

The results indicate that the experiments on the combined datasets achieve the best prediction in terms of both the F-measure and Kappa values. The first-level sub-grouping experiments reveal that the highest-performing sub-model achieves an F-measure of 51% for the enrolment dataset, while this proportion is increased up to 71% and 76% for the activity and combined datasets, respectively. It is also found that the sub-models that attain the best results perform 28%, 58% and 61% better than chance for the enrolment, activity and combined datasets, respectively.

Moreover, the findings demonstrate that the female-external or female-mature sub-models attain precisions of above 60% for the different classification methods when considering only the activity features. This proportion is increased to above 80% when student

Table 8
Cross-validation results for activity dataset. The accuracy of a sub-model built on a sub-dataset is compared to that of the model built on the entire dataset (and with a male or female sub-model for a second-level sub model). The following notations are used to indicate the significance of the accuracy difference: "**" for p-value $<$ 0.01, "*" for p-value $<$ 0.05 and "ns" (not significant) for p-value $>$ 0.05.

Method	Datasets			Female		Male	
		Accuracy (%)	Error	Accuracy (%)	Error	Accuracy (%)	Error
Naïve-Bayes	Original dataset	75.12	0.43	–	–	–	–
	Female	76.26 **	0.41	–	–	–	–
	Male	75.03 *	0.41	–	–	–	–
	Fulltime	79.12 **	0.38	**79.86** **	**0.39**	**78.65** **	**0.39**
	Part-time	76.06 **	0.41	76.45 *	0.42	75.45 *	0.42
	Normal	76.83 **	0.41	**78.08** **	**0.4**	**79.23** **	**0.4**
	Mature	75.32 **	0.42	76.78 *	0.42	76.05 *	0.41
	Internal	76.69 **	0.42	**78.44** **	**0.4**	**79.65** **	**0.4**
	External	75.65 *	0.42	75.27 (ns)	0.44	76.66 **	0.41
J48	Original dataset	75.09	0.37	–	–	–	–
	Female	81.4 **	0.34	–	–	–	–
	Male	78.35 **	0.35	–	–	–	–
	Fulltime	83.3 **	0.32	**83.23** **	**0.31**	**79.7** **	**0.34**
	Part-time	78 **	0.38	78.35 (ns)	0.4	77.6 (ns)	0.42
	Normal	82.2 **	0.33	**82.9** **	**0.32**	**80.48** **	**0.33**
	Mature	81.8 **	0.32	78.1 (ns)	0.41	73.62(ns)	0.43
	Internal	82 **	0.33	**82.84** **	**0.32**	**80.46** **	**0.32**
	External	79.9 **	0.35	80.2 (ns)	0.38	74.26(ns)	0.44
SMO	Original dataset	75.59	0.43	–	–	–	–
	Female	78.35 **	0.41	–	–	–	–
	Male	76.03 **	0.42	–	–	–	–
	Fulltime	79.48 **	0.4	**79.96** **	**0.4**	**78.49** **	**0.41**
	Part-time	77.28 **	0.42	77.64 (ns)	0.42	75.32 (ns)	0.44
	Normal	79.73 **	0.4	**79.14** **	**0.4**	**78.03** **	**0.41**
	Mature	78.1 **	0.42	76.32 (ns)	0.43	76.8 *	0.43
	Internal	80.02 **	0.39	**80.7** **	**0.39**	**79.25** **	**0.4**
	External	78.79 **	0.41	77.4 (ns)	0.42	76.23 *	0.43
JRip	Original dataset	76.66	0.42	–	–	–	–
	Female	78.5 **	0.39	–	–	–	–
	Male	77.1 **	0.41	–	–	–	–
	Fulltime	78.6 **	0.4	**79.58** **	**0.38**	**78.2** **	**0.4**
	Part-time	76.89 **	0.41	76.7 (ns)	0.42	76.5 (ns)	0.42
	Normal	78.3 **	0.4	**79.3** **	**0.38**	**78.64** **	**0.4**
	Mature	76.9 *	0.41	77.6 (ns)	0.41	76.84 (ns)	0.42
	Internal	77.66 **	0.4	**80.5** **	**0.37**	**79.13** **	**0.39**
	External	77.02 *	0.4	78.9 *	0.4	77.6 *	0.41

enrolment as well as activity features are considered. Furthermore, it is revealed from the LMS activity dataset experiment that the male-external sub-model achieves a precision of approximately 70%, while this amount increases to 80% when combined features are considered. The model generated from the combined dataset also attains the best result for correctly classifying both the successful and unsuccessful students, as illustrated in Fig. 2. This figure indicates that, for each method, the sub-models generated from the combined dataset achieve the best prediction result.

5.2. Interpretability and usability of discovered sub-models

The models developed by means of applying different classification methods aid us in predicting student academic outcomes. Identification of the influencing features affords the opportunity for course instructors to implement appropriate support measures. In this regard, the discovered knowledge should be easily understood and interpreted by course instructors and associated teaching staff. Four different classifier types are employed in this study to generate student performance models. The performance of the different methods is dissimilar across the different datasets. However, the black box techniques, namely naïve-Bayes and SMO, were unsuccessful in generating interpretable models for further use. In contrast, the white box techniques, namely JRip and J48, generate highly comprehensible models in rule and tree forms, respectively.

We describe certain findings provided by these white box techniques for the experiments on the combined dataset, for which they achieve the highest precision. JRip generates the following set of rules for the part-time sub-model.

As illustrated in Fig. 3, the rule set discovered by JRip takes an IF-THEN-ELSE form. The THEN is represented by the operator = >, while ELSE represents the default rule. The number in braces indicates the count of correctly and incorrectly classified instances by the rule, respectively. The first rule indicates that a student with lower participation in visiting the course home page is identified as low performer. The second rule states that a student who has a lower view in the course forum and is admitted based on professional qualification is predicted to be unsuccessful.

Consider another rule set discovered by the female-external sub-model, as illustrated in Fig. 4. The first rule indicates that students with lower participation in visiting the course home page and are not admitted into multiple programs in any year are predicted to be unsuccessful. The second rule states that students with a lower score in viewing discussions in the course forum and those who are admitted into multiple programs in the current year are most likely to be low performers. However, these findings are very obvious, as lower participation naturally leads to less involvement in different academic activities.

The *J48* algorithm generates a tree in which each branch represents an if-then rule. In a tree, the root and internal nodes are represented by *ellipse*, while the leaf nodes are represented by *rectangle*. This method discovers the following tree for the male-external sub-model, as illustrated in Fig. 5. The tree divides students into two major branches, according to the level of participation in commencing quiz attempts.

Table 9
Cross-validation results for combined dataset. The accuracy of a sub-model built on a sub-dataset is compared to that of the model built on the entire dataset (and with a male or female sub-model for a second-level sub model). The following notations are used to indicate the significance of the accuracy difference: "**" for p-value < 0.01, "*" for p-value < 0.05 and "ns" (not significant) for p-value > 0.05.

Method	Datasets			Female		Male	
		Accuracy (%)	Error	Accuracy (%)	Error	Accuracy (%)	Error
Naïve-Bayes	Original dataset	78.44	0.36	–	–	–	–
	Female	80.12 **	0.33	–	–	–	–
	Male	79.03 *	0.35	–	–	–	–
	Fulltime	81.18 **	0.32	**82.78** **	**0.31**	**81.23** **	**0.32**
	Part-time	80.68 **	0.33	78.18 (ns)	0.35	78.4 (ns)	0.35
	Normal	82.23 **	0.3	**81.97** **	**0.32**	**80.9** **	**0.33**
	Mature	80.67 **	0.32	79.05 (ns)	0.34	79.14 *	0.34
	Internal	82.79 **	0.3	**82.54** **	**0.31**	**81.69** **	**0.32**
	External	79.92 *	0.34	77.65(ns)	0.35	77.65 (ns)	0.36
J48	Original dataset	79.8	0.34	–	–	–	–
	Female	82.27 **	0.32	–	–	–	–
	Male	80.38 *	0.32	–	–	–	–
	Fulltime	82.87 **	0.3	**83.28** **	**0.3**	**81.9** **	**0.31**
	Part-time	80.26 *	0.32	75.8 (ns)	0.43	77.63 (ns)	0.41
	Normal	83.88 **	0.3	**82.9** **	**0.31**	**82.6** **	**0.3**
	Mature	80.13 *	0.32	78.76 (ns)	0.4	78.85 (ns)	0.4
	Internal	82.14 **	0.3	**83.14** **	**0.31**	**83.02****	**0.29**
	External	80.68 *	0.32	78 (ns)	0.4	79 (ns)	0.39
SMO	Original dataset	78.38	0.36	–	–	–	–
	Female	80.23 **	0.32	–	–	–	–
	Male	79.17 *	0.34	–	–	–	–
	Fulltime	81.23 **	0.31	**82.56** **	**0.29**	**80.78** **	**0.31**
	Part-time	78.79 *	0.35	79.3 (ns)	0.34	78.48 (ns)	0.34
	Normal	80.9 **	0.32	**81.97** **	**0.31**	**81.52** **	**0.3**
	Mature	79.3 *	0.34	76.46 (ns)	0.36	76.04 (ns)	0.36
	Internal	81.69 **	0.3	**82.04** **	**0.3**	**81.95** **	**0.3**
	External	79.67 *	0.33	77.32 (ns)	0.35	77.15 (ns)	0.35
JRip	Original dataset	74.33	0.43	–	–	–	–
	Female	81.81 **	0.37	–	–	–	–
	Male	79.38 **	0.4	–	–	–	–
	Fulltime	82.49 **	0.36	**83** **	**0.36**	**79.7** *	**0.39**
	Part-time	77.1 **	0.41	78.83 (ns)	0.4	74.29 (ns)	0.44
	Normal	82.2 **	0.37	**82.19** **	**0.36**	**80.11** *	**0.38**
	Mature	82.46 **	0.37	81.84 *	0.37	70.64 (ns)	0.45
	Internal	81.97 **	0.37	**82.6** **	**0.36**	**80.64** *	**0.39**
	External	79.58 **	0.4	80.28 (ns)	0.39	71.69 (ns)	0.45

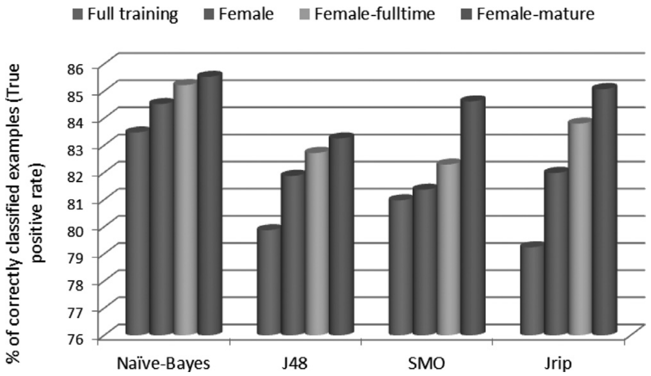


Fig. 2. Performance of different datasets for external sub-model in predicting student academic outcome.

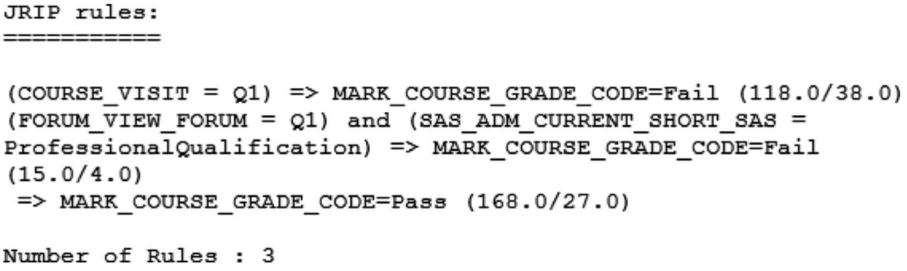
Students with higher contributions to the course online activities are predicted to pass. Students with lower participation are categorised according to whether or not they are admitted into multiple programs in any year of their candidature. Students who are admitted into multiple programs are predicted to pass. Students who are not admitted into multiple programs are further sub-divided into four sub-categories, according to their participation frequency in viewing book resources. Students with lower participation are predicted to fail, while the remainder are predicted to pass.

Fig. 6 depicts a decision tree generated from the external sub-model that classifies students into four major types according to their frequency of visiting the course home page, which indicates that students with medium or high visits pass the course. Among the students with the lowest visits of the course home page, the model discriminates them by considering whether or not they are admitted into multiple programs in any year. The first sub-category, in which students are not admitted into multiple programs, predicts students to fail. Finally, the second sub-category partitions students according to their attendance type, where part-time students are predicted to fail, and the remainder, to pass.

Another tree (Fig. 7) representing the female-part-time sub-model indicates that students with medium to higher participation in viewing discussions of the course forum pass the course. This tree further partitions students according to their ATAR with the lowest score in this activity. It demonstrates that students with a low to good ATAR fail in the course. The tree sub-divides students with a high ATAR according to their participation in viewing file resources. Students with higher participation pass, while the remainder fail in the course.

5.3. Major findings and their usefulness

Different white box methods have revealed a number of student academic outcome factors; among these, several are common across all methods. For example, it has been found that students with lower participation in either viewing the course home page or viewing a forum/discussion are most likely to be low performers. The experiments also demonstrate that students who are admitted into multiple

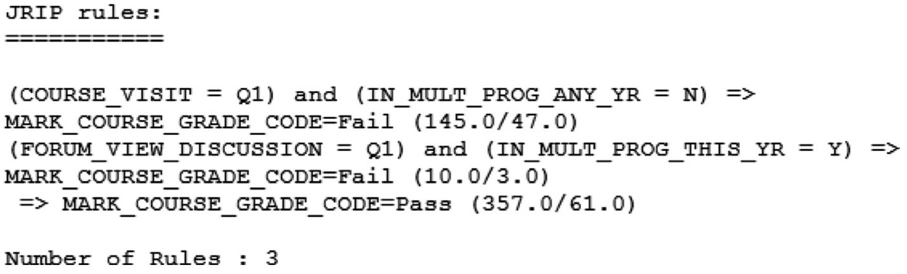


```
JRIP rules:
===========

(COURSE_VISIT = Q1) => MARK_COURSE_GRADE_CODE=Fail (118.0/38.0)
(FORUM_VIEW_FORUM = Q1) and (SAS_ADM_CURRENT_SHORT_SAS =
ProfessionalQualification) => MARK_COURSE_GRADE_CODE=Fail
(15.0/4.0)
 => MARK_COURSE_GRADE_CODE=Pass (168.0/27.0)

Number of Rules : 3
```

Fig. 3. Rules discovered by JRip method for part-time sub-model.



```
JRIP rules:
===========

(COURSE_VISIT = Q1) and (IN_MULT_PROG_ANY_YR = N) =>
MARK_COURSE_GRADE_CODE=Fail (145.0/47.0)
(FORUM_VIEW_DISCUSSION = Q1) and (IN_MULT_PROG_THIS_YR = Y) =>
MARK_COURSE_GRADE_CODE=Fail (10.0/3.0)
 => MARK_COURSE_GRADE_CODE=Pass (357.0/61.0)

Number of Rules : 3
```

Fig. 4. Rules discovered by JRip method for female-external sub-model.

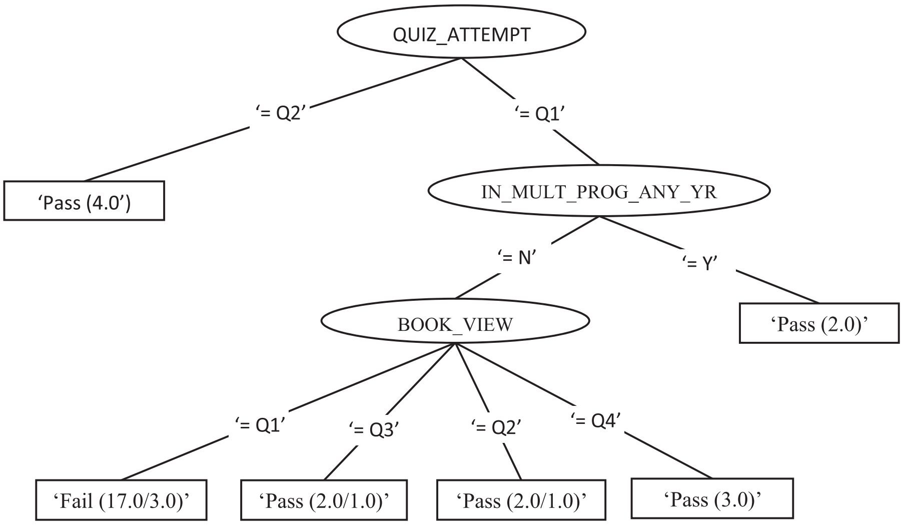


Fig. 5. Decision tree of male-external sub-model generated by J48.

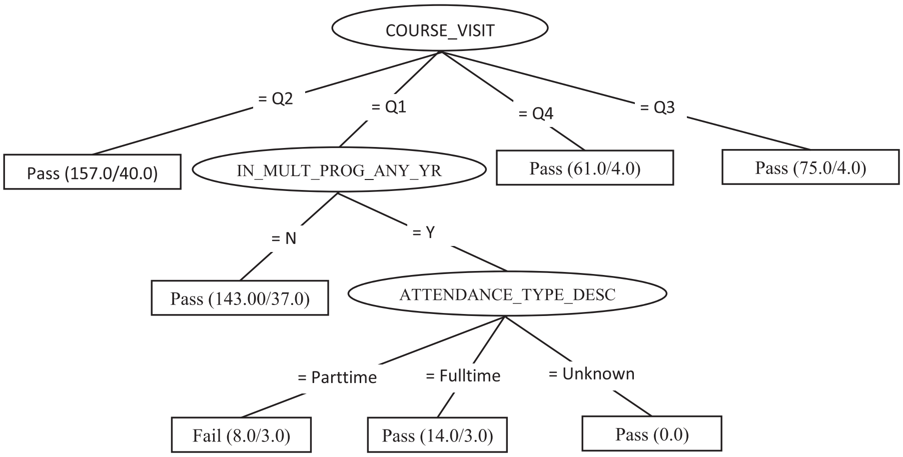


Fig. 6. Decision tree of external sub-model generated by J48.

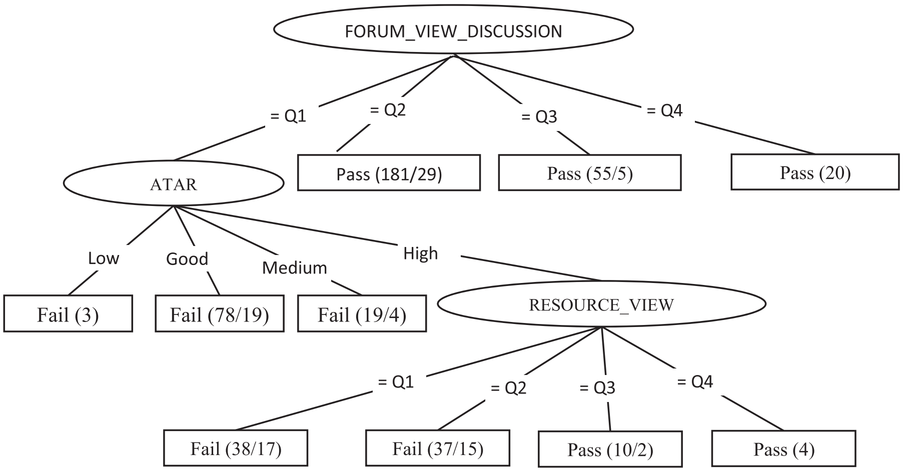


Fig. 7. Decision tree of female-part-time sub-model generated by J48.

programs in the current year or any year of their candidature are predicted to be low achievers. Certain additional factors are determined by individual methods. It is identified that students with lower participation in quiz activities or lower frequencies of viewing book or file resources are mostly unsuccessful. Furthermore, it is discovered that students with a poor academic background (ATAR, admission basis), belong to a lower social status (e.g. economic status or parent education) or study part-time are often time-poor and, as such, do not reach their academic potential. However, it is found in certain occasions that, when a student possesses a lower score in an activity, he/she can still demonstrate high performance in that course if he/she has a strong academic background and exhibits higher participation in other activities.

By learning the significant socio-demographic and academic factors, an educational institution can detect at-risk students at an early stage (prior to beginning their course or program), and take necessary steps to support students exhibiting these features, such as monitoring their progress by conducting a routine assessment of their studies throughout the term. Moreover, the institution can provide additional academic support; for example, forming smaller groups of such students to allow them to take several extra classes along with a small weekly seminar on a specific topic. When detecting the influencing LMS activity features, a course teacher should direct his/her attention to the group of students with a very high chance of failing, and also encourage them to participate in such activities because of their strong association with student academic performance in a course. However, in this case, it is not possible to conduct early identification of vulnerable students. The course instructor must wait until a specific period (e.g. the fourth or sixth week from starting the course) to detect student participation in specific activities. By learning the influencing factors, an institution can identify vulnerable students possessing specific socio-demographic or academic features, and advise the course instructors to monitor their course progress.

This study exhibits certain limitations; for example, although it employs different classification methods with student enrolment and LMS activity data, the experiments are confined to the data of domestic students for a specific university division. Moreover, when mining the LMS activity data, we only consider individual attributes for all of the modules (e.g. quiz, forum, etc.); we do not consider the combined features corresponding to a particular module (e.g. students' overall participation in a forum or resource module).

6. Conclusion

In this paper, we have proposed the concept of exploiting heterogeneity for obtaining improved prediction models. The experimental results demonstrated the effectiveness of using student sub-populations in predicting student academic performance. It was shown that the generated sub-models outperform the base model. The experiment indicated that the sub-model generated from the external student subgroup achieves the best performance.

Furthermore, it has been demonstrated that it is useful to investigate second-level subgroups. For example, the experiments indicated that the female-mature, female-external and female-part-time sub-models attain superior prediction results to the female model (a first-level sub-model). Moreover, the male-external sub-model outperforms the male model. These results indicate that, although not all of the second-level sub-models achieve superior predictions, several can still provide insights into student performance, and thus assist in the design of more targeted student support.

The experiments revealed that no single method exhibits superior performance in all aspects for predicting student performance. However, it is crucial to use the discovered knowledge for predicting vulnerable students with higher predictability. In this regard, the white box techniques, namely J48 and JRip, contribute significantly by generating comprehensible output in the form of a tree and rule, respectively. Furthermore, by considering combined features, a superior prediction result is obtained in identifying unsuccessful students compared to considering the features separately.

Further research should be conducted to address the limitations of this paper. Firstly, in order to generalise the results, an investigation should be carried out to identify the risk indicators of international students, as they may possess some different features from domestic students, such as diverse ethnic origins, funding opportunities, native languages and other factors. Secondly, it would be very useful to consider the combined features for a particular module and the categorisation features (e.g. social and information) in terms of student participation in LMS activities. Furthermore, we plan to generate student profiles using unlabelled data to discover interesting student clusters and their characteristics.

References

[1] Y. Zhang, S. Oussena, T. Clark, H. Kim, Use data mining to improve student retention in higher education - a case study, Proceedings of the Twelveth International Conference on Enterprise Information Systems ICEIS, (2010), pp. 190–197.

[2] C. Romero, M.I. López, J.M. Luna, S. Ventura, Predicting students' final performance from participation in on-line discussion forums, Comput. Edu. 68 (0) (2013) 458–472.

[3] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[4] W. Klösgen, J.M. Zytkow (Eds.), Handbook of Data Mining and Knowledge Discovery, Oxford University Press, Inc., New York, NY, USA, 2002.

[5] Z. Ibrahim, D. Rusli, Predicting students' academic performance: comparing

artificial neural network, decision tree and linear regression, Proceedings of Annual SAS Malaysia Forum, Kuala Lumpur, Malaysia, (2007), pp. 1–6.

[6] C.J. Villagrà-Arnedo, F.J. Gallego-Durn, F. Llorens-Largo, P. Compa-Rosique, R. Satorre-Cuerda, R. Molina-Carmona, Improving the expressiveness of black-box models for predicting student performance, Comput. Human Behav. 72 (2017) 621–631.

[7] L. Rosenbaum, G. Hinselmann, A. Jahn, A. Zell, Interpreting linear support vector machine models with heat map molecule coloring, J. Cheminf. 3 (1) (2011) 11.

[8] Z.J. Kovačić, Predicting student success by mining enrolment data, Res. Higher Edu. J. 15 (0) (2012) 1–20.

[9] C. Romero, P.G. Espejo, A. Zafra, J.R. Romero, S. Ventura, Web usage mining for predicting final marks of students that use Moodle courses, Comput. Appl. Eng. Edu. 21 (1) (2013) 135–146.

[10] M. Blagojević, Z. Micić, A web-based intelligent report e-learning system using data mining techniques, Comput. Electr. Eng. 39 (2) (2013) 465–474.

[11] D. Gašević, S. Dawson, T. Rogers, D. Gašević, Learning analytics should not promote one size fits all: the effects of instructional conditions in predicting academic success, Int. Higher Edu. 28 (2016) 68–84.

[12] G.H. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, (1995), pp. 338–345.

[13] J.C. Platt, Fast Training of Support Vector Machines Using Sequential Minimal Optimization, in: B. Schölkopf, Christopher J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods, MIT Press, Cambridge, MA, USA, 1999, pp. 185–208.

[14] W.W. Cohen, Fast effective rule induction, Proceedings of the Twelfth International Conference on Machine Learning, (1995), pp. 115–123.

[15] T. Thiele, A. Singleton, D. Pope, D. Stanistreet, Predicting students' academic performance based on school and socio-demographic characteristics, Stud. Higher Edu. 41 (8) (2016) 1424–1446.

[16] C.E.L. Guarín, E.L. Guzmán, F.A. González, A model to predict low academic performance at a specific enrollment using data mining, IEEE Revista Iberoamericana de Tecnologias del Aprendizaje 10 (3) (2015) 119–125.

[17] P. Strecht, L. Cruz, C. Soares, J. Moreira, R. Abreu, A comparative study of classification and regression algorithms for modelling students' academic performance, Proceedings of the Eighth International Conference on Educational Data Mining, (2015), pp. 392–395.

[18] R. Asif, A. Merceron, S.A. Ali, N.G. Haider, Analyzing undergraduate students' performance using educational data mining, Comput. Edu. 113 (2017) 177–194.

[19] J. Xu, K.H. Moon, M. van der Schaar, A machine learning approach for tracking and predicting student performance in degree programs, IEEE J. Sel. Top Signal Process. 11 (5) (2017) 742–753.

[20] R. Wang, G. Harari, P. Hao, X. Zhou, A.T. Campbell, Smartgpa: how smartphones can assess and predict academic performance of college students, Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15, (2015), pp. 295–306.

[21] G. Cobo, G. García, E. Santamaría, J.A. Morán, J. Melenchón, C. Monzo, Modeling students' activity in online discussion forums: a strategy based on time series and agglomerative hierarchical clustering, Proceedings of the International Conference on Educational Data Mining, (2011), pp. 253–258.

[22] M. Jovanovic, M. Vukicevic, M. Milovanovic, M. Minovic, Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study, Int. J. Comput. Intell. Syst. 5 (3) (2012) 597–610.

[23] Moodle, https://moodle.org/. Online accessed: 10-Feb-2018.

[24] BlackBoard, http://www.blackboard.com//. Online accessed: 10-Feb-2018.

[25] Desire2Learn, http://www.brightspace.com/. Online accessed: 10-Feb-2018.

[26] L.P. Macfadyen, S. Dawson, Mining LMS data to develop an early warning system for educators: a proof of concept, Comput. Edu. 54 (2) (2010) 588–599.

[27] T.M. Khan, F. Clear, S.S. Sajadi, The relationship between educational performance and online access routines: analysis of students' access to an online discussion forum, Proceedings of the Second International Conference on Learning Analytics and Knowledge, (2012), pp. 226–229.

[28] M. Hlosta, Z. Zdrahal, J. Zendulka, Ouroboros: early identification of at-risk students without models based on legacy data, Proceedings of the Seventh International Learning Analytics & Knowledge Conference, LAK '17, (2017), pp. 6–15.

[29] O. Adejo, T. Connolly, An integrated system framework for predicting students' academic performance in higher educational institutions, Int. J. Comput. Sci. Inf. Technol. 9 (2017) 149–157.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newslett. 11 (1) (2009) 10–18.

[31] J.W. Perry, A. Kent, M.M. Berry, Machine literature searching x. machine language; factors underlying its design and development, J. Assoc. Inf. Sci. Technol. 6 (4) (1955) 242–254.

[32] D.C. Blair, Information retrieval, J. Am. Soc. Inf. Sci. 30 (6) (1979) 374–375.

[33] J. Cohen, A coefficient of agreement for nominal scales, Edu. Psychol. Meas 20 (1) (1960) 37–46.

[34] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.

[35] I. Rodríguez-Fdez, A. Canosa, M. Mucientes, A. Bugarín, STAC: a web platform for the comparison of algorithms using statistical tests, Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), (2015), p. 1.