The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/2397-7604.htm

Received 24 September 2017
Revised 14 December 2017
Accepted 14 December 2017

On predicting academic performance with process mining in learning analytics

Rahila Umer
Institute of Natural and Mathematical Sciences, College of Sciences, Massey University, Auckland, New Zealand
Teo Susnjak and Anuradha Mathrani
Massey University, Auckland, New Zealand, and
Suriadi Suriadi
Faculty of Science and Engineering, Queensland University of Technology, Brisbane, Australia

Abstract

Purpose – The purpose of this paper is to propose a process mining approach to help in making early predictions to improve students' learning experience in massive open online courses (MOOCs). It investigates the impact of various machine learning techniques in combination with process mining features to measure effectiveness of these techniques.
Design/methodology/approach – Student's data (e.g. assessment grades, demographic information) and weekly interaction data based on event logs (e.g. video lecture interaction, solution submission time, time spent weekly) have guided this design. This study evaluates four machine learning classification techniques used in the literature (logistic regression (LR), Naïve Bayes (NB), random forest (RF) and K-nearest neighbor) to monitor weekly progression of students' performance and to predict their overall performance outcome. Two data sets – one, with traditional features and second, with features obtained from process conformance testing – have been used.
Findings – The results show that techniques used in the study are able to make predictions on the performance of students. Overall accuracy (F1-score, area under curve) of machine learning techniques can be improved by integrating process mining features with standard features. Specifically, the use of LR and NB classifiers outperforms other techniques in a statistical significant way.
Practical implications – Although MOOCs provide a platform for learning in highly scalable and flexible manner, they are prone to early dropout and low completion rate. This study outlines a data-driven approach to improve students' learning experience and decrease the dropout rate.
Social implications – Early predictions based on individual's participation can help educators provide support to students who are struggling in the course.
Originality/value – This study outlines the innovative use of process mining techniques in education data mining to help educators gather data-driven insight on student performances in the enrolled courses.

Keywords Prediction, MOOCs, Machine learning, Learning analytics, Process mining, Education data mining

Paper type Research paper

1. Introduction

Massive open online courses (MOOCs) have become very popular among student communities, since it provides them an opportunity to register for courses offered by prestigious universities around the world. MOOCs provide a learning environment which attracts large number of learners having different goals and motivations. Courseera, edX

Journal of Research in Innovative Teaching & Learning
Vol. 10 No. 2, 2017
pp. 160-176
Emerald Publishing Limited
2397-7604
DOI 10.1108/JRIT-09-2017-0022

and Udacity are the three pioneers of MOOCs platform which are then closely followed by several around the world like Miriada and Spanish MOOC in Spain, Khan Academy in North America, Iversity in Germany, FutureLearn in England, Open2Study in Australia, Fun in France, Veduca in Brazil, Schoo in Japan and xuetangX in China.

MOOC environment has revolutionized education by centralizing global resources and restructuring the learning environment to bring it closer to students reach. Unlike traditional higher education learning environment, MOOCs provide an open access to courses to anyone with access to the internet. It provides free and open access to high-quality advanced courses comprising of video lectures, reading materials, quizzes, problem sets and forums for productive discussions to foster learning process and develop learning communities. The use of technology in teaching like forum, blogs (Ebner *et al.*, 2010), wiki or educational software has improved the learning process. The increased availability of recorded data from such environments has provided an opportunity to closely investigate student learning behaviors and work toward improving their learning process.

Although there are massive enrollments in courses offered via MOOCs, the completion rate and retaining of persistent students are rather low, often less than 20 percent (Kizilcec *et al.*, 2013). One of the criticisms in MOOCs is the low retention rate of the students which is heavily criticized. Therefore, predicting the likelihood of dropout is necessary, so that steps can be taken to retrain students by encouraging them in their learning activities.

Learning analytics (LA) has recently emerged as a new research lens that focuses on computational techniques to inform on students' practices. Every online interaction by students like click, page visited or video viewed while pursuing the course is recorded in the log history (Clow, 2013a). How can we get insights from the log history data so as to make pedagogical interventions to support student learning during the course? Campbell *et al.* (2007) identified five steps, namely, capture, report, predict, act and refine, as the central theme in LA. Once we have captured data that report students' interactions with the course, analysts works toward making predictions for pedagogical intervention, which is gradually refined.

Several works (Marquez-Vera *et al.*, 2013; Ye and Biswas, 2014; Bayer *et al.*, 2012; Manhães *et al.*, 2014; Martinho *et al.*, 2013; Simon *et al.*, 2006; Watson *et al.*, 2013) have suggested EDM techniques as the way forward to help in predictions of academic failure among students. In this study, the focus is on predicting students' performance through the traces they leave while pursuing a course. The aim is to apply data mining/machine learning algorithms to students' data, as students are progressing through a course, in order to predict which students are at risk of not satisfying course requirements, or are rather likely to fail. The identification of such students would then enable educators to carry out various forms of early intervention or provide additional and more tailored support as mitigation measures.

The study is guided by the following research questions:

RQ1. Which machine learning algorithms are effective at predicting students, who are at risk of failure on MOOCs data set?

RQ2. Is the integration of process mining features able to increase the effectiveness of the machine learning algorithms for the MOOCs problem?

The significance of our study is the integration of process mining to extend the features. Extended features are obtained as a result of process conformance testing. EDM techniques are then evaluated on two kinds of data sets, one with process mining features and other without. We used some of the widely used (Wu *et al.*, 2008) classifiers, namely, random forest (RF) (Breiman, 2001), logistic regression (LR), Naive Bayes (Cortes and Vapnik, 1995) and K-nearest neighbor (KNN) (Hechenbichler and Schliep, 2004), to answer the posed questions.

This section has laid the foundation of our research inquiry. The remainder of the paper is organized as follows. In Section 2, we present some of the related work in similar field. Section 3 discusses data sources and the limitations of data set used in the study. Section 4 describes the methods applied in our experiments. In Section 5, we present answers to the research questions and discuss experimental results. Finally, in Section 6, we make conclusion.

2. Related work

Several studies have reported and provided promising results in prediction of students who are likely to fail in a given course. In most of these studies, the data used for prediction consist of non-academic information; all of which require extra effort to collect. Our study is the first of its kind which has used process mining to enhance existing features identified in the literature.

Khobragade (2015) proposed an approach where they have predicted the students' academic failure using decision tree, Naive Bayes and using classifiers that are based on induction rule and decision tree. Data used for classifications involved social, academic and background information of the student. These data have been collected through surveys. A total of 11 features were used for prediction after applying the feature selection algorithm. Classifiers have then been evaluated based on accuracy. Naive Bayes provided the best accuracy of above 87 percent. However, data were gathered through surveys which are time consuming and also involved methods that make overall prediction and do not consider early prediction.

Marquez-Vera *et al.* (2013) evaluated white-box classifiers. They used induction rules and decision tree for predicting academic failures for students in middle or secondary school. Detailed information of students' social background and academic information were used. Again the data collection process used here was very extensive and time consuming, since non-academic information was collected through surveys. The impact of different data pre-processing approaches was also analyzed on classification accuracy. The proposed methods show promising results for making prediction of overall academic performance of students.

Costa *et al.* (2017) presented a comparative study of EDM techniques to predict those students who are likely to fail in a programming course. The significance of this study is that these techniques used could predict the students' performance at early stages so that some intervention strategy could be made to help students. This study also analyzed the impact of data pre-processing methods and algorithms fine-tuning tasks on prediction results. The study showed that support vector machine outperformed other techniques in a statistical significant way, and data pre-processing and algorithm fine-tuning tasks can improve accuracy.

The work proposed by Ahmad *et al.* (2015) presents an approach where EDM techniques are used to predict academic performance of first-year students in a computer science course. EDM techniques used are decision tree, Naive Bayes and rule-based classification. The data used during the course of study include demographic data, previous academic records and other family-related information. Rule-based classifiers outperformed other methods and provided prediction accuracy of 71 percent.

Yukselturk *et al.* (2014) predicted students' dropout in an online course using EDM classifiers: Naive Bayes, decision tree, KNN and neural network. The data used for prediction consisted of demographic information, online technology self-efficacy scale, readiness for online learning questionnaire, locus of control scale and prior knowledge questionnaire. A total of ten features were used for predicting class label (dropout/not). The maximum prediction accuracy was obtained by KNN (87 percent). Data were collected through surveys and also did not provide prediction at early stages of course.

Another research, conducted by Boongoen (2017) in a Thai university, used a link-based cluster ensemble method as a data transformation framework for prediction. The research has compared several state-of-the art dimensionality reduction techniques.

Ye and Biswas (2014) used and extended standard features for MOOC analysis with higher granularity to make more accurate predictions for dropout and performance. Analysis was made using data collected from video lectures, weekly quizzes and peer assessments from the ten-week course. Standard features were extended using some detailed temporal features like when some assessment was started during the week, or when the first lecture was viewed.

The findings compared with existing studies showed that these features improved the prediction accuracy. The time when a student starts the peer assessment assignment was found to be a good predictor. Once the peer assessment score was available, the prediction performance improved. Analysis shows that the students who watched video and did not take quizzes were the ones who mostly dropped out. Overall results show that more precise temporal features and more quantitative information improved early prediction accuracies and false alarm rates as compared to using only assessment score features.

Bydzovska (2016) proposed an approach to predict the students' performance using course characteristics and previous grades. Two different approaches were used. In the first approach, classification and regression were used to predict performance using academic-related data and data about student's social behavior. The findings were significant with the small number of students. In the second approach, collaborative filtering techniques were used to predict the student's performance based on similarity of achievements. Classification algorithms, namely, support vector machine, decision tree, part, IBI, RF, Naive Bayes and rule-based classifier, were used, where support vector machine produced best predictions which were further improved by integrating social behavior data.

Cambruzzi *et al.* (2015) used learning analtyics to predict dropout rates in distance education. They developed a system that predicts dropout and supports to integrate pedagogical interventions and textual analysis to reverse identified dropout tendencies. The system was able to predict dropout with 87 percent precision, later the dropout was decreased by 11 percent by implementing specific pedagogical actions.

3. Data sources

In this study, we analyzed data obtained from Coursera for course "Principles of Economics" offered in Summer 2014. The data set consisted of assessments grades, solution submission time, video lecture interaction log, participant's demographic information, time spent weekly and final grades. The course was designed as an eight-week introduction to the study of economics. The total number of students was more than 3,000; however, we only included data of students who were registered at the time the course started (i.e. on June 24, 2014) and whose final score was not missing. We extracted data of total 167 students, out of which 40 students passed the course while rest had failed. Students with scores greater than 0.5 were considered passed. The final data set obtained was thus imbalanced in regard to the final grade distribution (Figure 1).

3.1 Data set 1 – standard features

The data set comprises of features like demographics, assessment grades, time spent on activities, video watch activities, etc. This has been characterized as "standard" features shown in (Table I).

3.2 Data set 2 – process mining features

A data set has been generated next using logs of weekly activities during the course. It includes features that reflect the differences in the behavior of students with respect to the

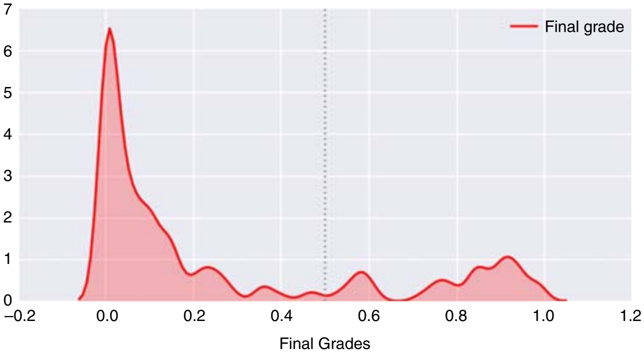


Figure 1.
Final grade distribution

Table I.
Features obtained from MOOCs data set: standard features

S.No.	Feature	Explanation
1	Age	Age in years
2	Education	Highest qualification
3	Gender	Female/Male/Null
4	Average score in weekly quiz	Average score in quizzes of particular week
5	Number of quizzes attempted	Average attempt for quiz in particular week
6	Quiz lag	Duration between first and last activity of quiz
7	Lecture lag	Duration between first and last activity of Lecture
8	Total lecture attended	Total lectures attended in particular week
9	Video activity count	Activity counts during video lecture (pause, play, stop, etc.)
10	Efforts in seconds	Total time spent in a particular week

behavior of top performing students in the course. These measures have been obtained as a result of process conformance testing (Van der Aalst *et al.*, 2012).

In the process conformance testing, given a normative model *M* and an event log *L*, difference between the process behavior and *L* can be explained. Conformance checking was performed using the model representing top student's weekly activities and log of other student's weekly activities. The log was replayed using the model to establish a precise relationship between event and model elements and to analyze the deviation of students from modeled behavior. Output of conformance testing is a fitness score that is assigned to each student (case). The fitness scores, obtained based on weekly logs of top performing students and other students, were used as features and integrated with standard features. The following steps have been performed to prepare data set with process mining features.

3.2.1 Step 1. Using the inductive miner method in ProM (Van der Aalst *et al.*, 2009), we generated a process model using activity logs of top performing students having grade more than 90 percent. The result of this step is a process model shown in Figure 2.

3.2.2 Step 2. By using "Replay a log on Petri net for performance/conformance analysis" method in Prom, the log model alignment was generated, as shown in Figure 3. Inputs to this method were process model of top performing students and log of activities of other students (Figure 4).

3.2.3 Step 3. The log model alignment generated and exported it in the comma separated value format. We extracted fitness scores for each student and integrated them with the standard features in data set 1. This helped characterize the fitness scores as process mining features in data set 2. Same process was repeated and data sets were created using weekly logs.

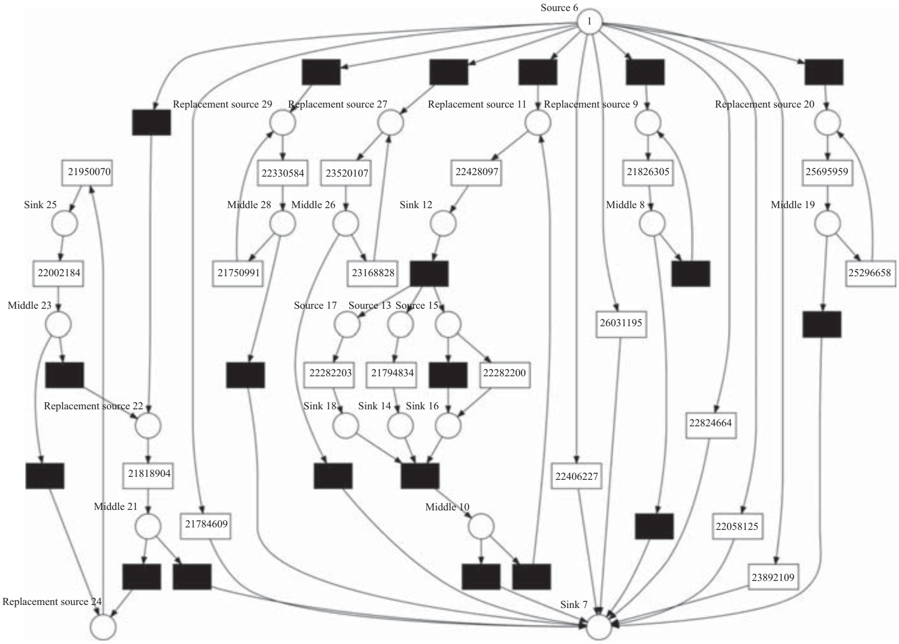
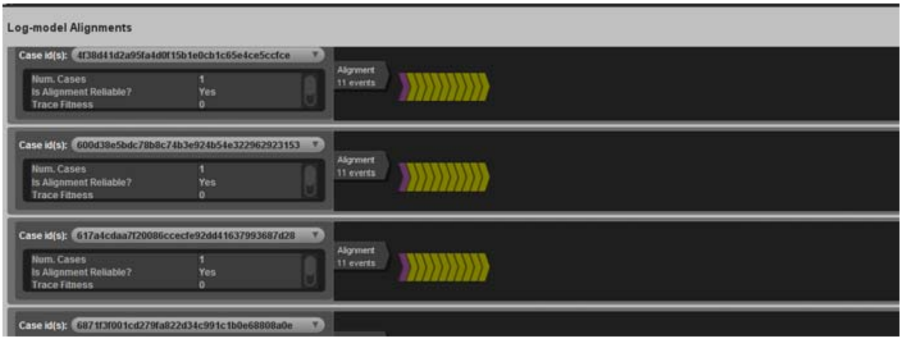


Figure 2. Process model of top performing students using inductive miner method based on log of week-1 to week-2



Log-model Alignments

Case id(s): 4f38d41d2a95fa4d0f15b1e0cb1c65e4ce5ccfce

Num. Cases	1
Is Alignment Reliable?	Yes
Trace Fitness	0

Alignment
11 events

Case id(s): 600d38e5bdc78b8c74b3e924b54e322962923153

Num. Cases	1
Is Alignment Reliable?	Yes
Trace Fitness	0

Alignment
11 events

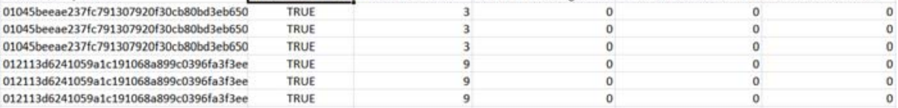
Case id(s): 617a4cdaa7f20086ccecfe92dd41637993687d28

Num. Cases	1
Is Alignment Reliable?	Yes
Trace Fitness	0

Alignment
11 events

Case id(s): 6871f3f001cd279fa822d34c991c1b0e68808a0e

Figure 3. Result of replaying log history on process model for conformance analysis



TRACE:concept:name	TRACEAL:IsReliable	TRACEAL:RawCost	TRACEAL:MoveLogFitness	TRACEAL:MoveModelFitnes	TRACEAL:TraceFitness
01045beeae237fc791307920f30cb80bd3eb650	TRUE	3	0	0	0
01045beeae237fc791307920f30cb80bd3eb650	TRUE	3	0	0	0
01045beeae237fc791307920f30cb80bd3eb650	TRUE	3	0	0	0
012113d6241059a1c191068a899c0396fa3f3ee	TRUE	9	0	0	0
012113d6241059a1c191068a899c0396fa3f3ee	TRUE	9	0	0	0
012113d6241059a1c191068a899c0396fa3f3ee	TRUE	9	0	0	0

Figure 4. Screen-shot of report generated after replaying log on process model

3.3 Limitations of data sets

MOOCs environments are different from traditional learning setups, which makes it challenging to analyze the data. Large amount of missing data, multiple number of attempts for assignment submission, multiple time registration and higher rate of dropout are some of the major challenges faced during the analysis of MOOCs data.

In this study, a total of 167 students' data have been used, of which the majority of students belonged to one class and also final data set was imbalanced. It has been found that in a typical MOOCs setting, students are active in the early weeks and become inactive or withdraw from the course in later weeks which result in a large number of missing values.

Figure 5 shows the average number of quizzes attempted by students during each week. Students attempted most of the quizzes in week-1 only. Figure 6 shows the average number of lectures watched during the course. It is evident that last three weeks were the most inactive weeks. The majority of the students either did not watch lectures or withdrew from the course. These indicate unequal patterns of participation. Figure 7 shows average time spent during the course. The graph shows week-7 to be the least active week. However, online engagement traces do not necessarily reflect all activities related to learning process. Due to this reason, measurement of success and participation in MOOC environment must be reconsidered (Clow, 2013b; Bergner *et al.*, 2015).

This study recognizes these limitations in the data sets; however, this subset (of 167 students) is representative of student participation and completion from a real-world MOOC environment. The data sets have been used to mainly demonstrate the effectiveness of data mining techniques using process mining features.

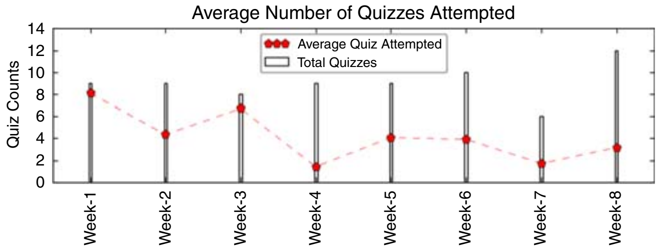


Figure 5.
Average number of quizzes attempted (failed or passed) by students during the course

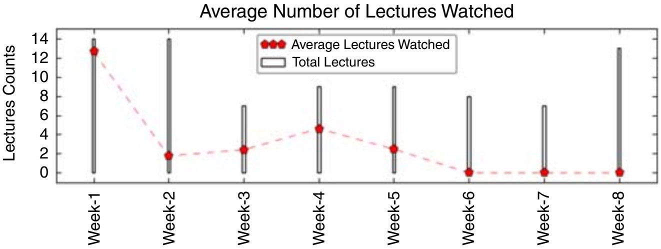


Figure 6.
Average number of lectures watched by students during the course

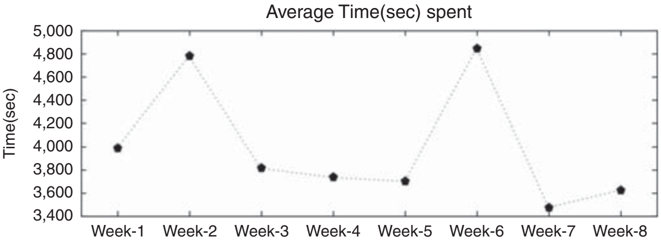


Figure 7.
Average number of time (seconds) spent during the course

4. Experimental design

The aim of this study is to compare the effectiveness of existing popular machine learning algorithms for early identification of students, who are likely to fail and to investigate the effect of process mining features in the performance of the techniques.

4.1 Classification methods

We used classification methods that have been utilized in the field of education domain and are suitable for imbalanced data set. Machine learning algorithms used in the experiment are as follows: Naive Bayes, RF, LR and KNN. In the following subsections, the classification methods used are briefly explained. Table II shows the parameters used for classifiers.

4.1.1 LR. LR is a parametric method, used in classifications wherein a sigmoid function is estimated based on the training data. This method is based upon the assumption that the probability of event occurring follows a logistic distribution. The distribution is defined as follows:

$$P\left(\text{outcome} = \text{Pass}/X\right) = \frac{1}{1+e^{-X^{t\beta}}}$$

where $X^{T\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$ and X is a set of measurements, $X = [x_1, x_2, \ldots, x_n]$.

Using this function, input space is partitioned into two regions. New instances are classified to the region they belong. The distribution is in the shape of an "*S*," which indicates that difference at the extreme ends will not effect much, as compared to the difference around center. LR is bounded by 0 and 1 to represent the probabilities. The upper and lower portions of "*S*" represent high probabilities and low probabilities of the same even, respectively.

This approach is widely used in the literature to predict the retention with high accuracy (Lin and Reid, 2009; Mertes and Hoover, 2014; Veenstra *et al.*, 2009; Dunn and Mulvenon, 2009).

4.1.2 Naive Bayes. Naive Bayes is a simple supervised method, which is a special form of discriminant analysis. It is based on the Bayes theorem (Benbassat, 1990), returns probability of prediction using the evidence derived from observed data. This method relies on two assumptions: all attributes are conditionally independent and contribute equally to the final outcome of the class and that there are no hidden attributes that can affect the process of prediction. The Naive Bayes classifier assigns to each instance the class value with the highest conditional probability. This method is extensively used since 1950s and is a very popular method especially in the domain of text mining. Naive Bayes performs surprisingly well in cases where even the attributes are not independent. This method is used in several studies in the domain of education data mining (Pittman, 2008; Zhang *et al.*, 2010; Khobragade, 2015; Nandeshwar *et al.*, 2011; Mashiloane and Mchunu, 2013; Sharma and Mavani, 2011; Costa *et al.*, 2017; Ahmad *et al.*, 2015).

4.1.3 RF. RF uses a standard machine learning technique called a "decision tree." Decision trees build a classification model by a recursive binary partition of a labeled data

Classifier	Training setting	Implementation source
KNN	$K = 3$	Scikit-learn (Pedregosa *et al.*, 2011)
Random forest	Estimator = 10	Scikit-learn (Pedregosa *et al.*, 2011)
Naïve Bayes	Gaussian default setting	Scikit-learn (Pedregosa *et al.*, 2011)
Logistic regression	default settings	Scikit-learn (Pedregosa *et al.*, 2011)

Table II. Machine learning parameters

set into increasingly homogeneous nodes. Homogeneity is measured by the Gini index, which is defined as:

$$G = \sum k.P(k).(1-p(k))$$

where $P(k)$ is the proportion of observations in the kth class.

At each step, an optimization is carried out to select, in each node, the feature and the numeric threshold or group of values if the variable is categorical that would produce the lowest G value if used to divide the node. This process continues until it is not possible to reduce the Gini index in any node. The final output is a classification tree with completely homogeneous lower nodes. However, this is not always the case, and the predominant class is used to label the node, the other cases being classification errors. On the basis of these errors, the tree is pruned to allow a higher generalization capacity. Small modification in a data set affects the results of classification in a case of single tree. However, this limitation could be overcome using ensemble learning techniques to obtain a better performance. Bootstrapped sample of the available instances is used to generate unpruned trees in a large number (500-2,000). In order to add randomness and to decrease correlation, each node division is carried out with a randomized subset of the predictors. In ensemble techniques, correlation is not a desirable property because the different results make sense to the voting system. New instance is classified to the class it belongs, based on the aggregate number of votes given by multiple trees. This method is widely used in the prediction of student's performance (Bydžovská, 2016; Mashiloane and Mchunu, 2013; Marquez-Vera *et al.*, 2013).

4.1.4 KNN. KNN (Hechenbichler and Schliep, 2004) is a classification method that estimates the class for every new instance using the k-closest instances, by calculating a distance metric, from the training set. Class probabilities for the new instance are estimated as the proportion of training set neighbors in each class. Ties are broken randomly or by including the k + 1 closest neighbor in the calculation. K is the number of neighbors, an important parameter to be considered when using this method. A small value leads to an increase in the probability of over-fitting, while too large a value causes a high-bias classification. This simple algorithm has been successful in a large number of classification problems (Gray *et al.*, 2014; Minaei-Bidgoli *et al.*, 2003; Mayilvaganan and Kalpanadevi, 2014; Yukselturk *et al.*, 2014).

4.2 Evaluation measures

In order to compare the performance of each classifier, F1-score and area under curve (AUC) were used. Due to the imbalanced nature of data set, overall accuracy might be misleading.

4.3 F1-score

F1-score is widely used in binary classification problems. F1-score is the harmonic mean between Precision and Recall:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

where TP is the number of positive instances correctly classified as positive; FP is the number of negative instances incorrectly classified as positive; and FN is the number of positive instances incorrectly classified as negative:

$$F1-\text{score}=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$$

4.4 AUC

A receiver operating characteristic curve is a way to compare diagnostic tests. It is a plot of the true positive rate against the false positive rate. The AUC is a number between 0 and 1:

$$\text{False positive rate}=\frac{\text{FP}}{(\text{FP}+\text{TN})}$$

$$\text{True positive rate}=\frac{\text{TP}}{(\text{TP}+\text{FN})}$$

4.5 Training procedure

To estimate the generalization capability of the model to future data set, ten-fold cross-validation technique was used. This technique splits the original data set into ten subsets of equal size, preserving the original ration of minority and majority class instances. One subset is left for the validation and rest are used for training the model. This process is repeated ten times using different subsets for training and validation each time. In the end, average results across each iteration are computed. Performance of classification methods described in Section 4.1 was evaluated next. These methods are used for the prediction of student's final outcome of the course as Pass or Fail on two data sets that are discussed in Section 3. Prediction was based on the learners' demographics and dynamic data of the previous week. Each data set was divided on weekly basis. After each week, prediction was made based on the available data of current and previous weeks. In definition, week-3 data set means that it consists of the all available data till week-3 which includes week-2 and week-1 data as well. We assume that prediction accuracy improves as more data become available in upcoming weeks. For instance, prediction after week-4 means that we used all available data till week-4, which might include scores of assessments and quizzes, which are part of final score and ultimately improve accuracy. Prediction accuracy at early stages is important so that timely interventions can be made to help students.

5. Experimental results

This section first presents the experimental results which are discussed next. Table III displays the result of data set 1 which includes features like demographic and grading scores.

Data set	LR	RF	NB	K.NN
Week-1	0.77	0.712	0.788	0.724
Week-2	0.879	0.866	0.89	0.848
Week-3	0.794	0.803	0.618	0.678
Week-4	0.8	0.799	0.715	0.722
Week-5	0.836	0.858	0.764	0.75
Week-6	0.836	0.84	0.743	0.747
Week-7	0.85	0.842	0.503	0.75
Week-8	0.841	0.866	0.536	0.801
Rank(mean)	*1.75*	1.875	3.125	3.25

Table III. Comparative results of the effectiveness of machine learning algorithms on the data set using standard features and mean ranks of classifiers from highest (1) to lowest (*N*)

Table IV displays the result of data set 2 which enriches the standard features with process mining features. Next, we answer the research questions in the light of our results of analysis:

- *RQ1.* Which machine learning algorithms are effective at predicting students, who are at risk of failure on MOOCs data set?

In order to answer this question, prediction was performed using four machine learning techniques on two data sets. Table III shows the results of effectiveness of machine learning algorithms using data set 1 (using standard features only) to predict students who are likely to fail. The results show that maximum F1-score obtained is 0.78 by Naive Bayes classifier after week-1. For week-2, F1-score improved to 0.89 by Naive Bayes classifier. After week-2, F1-score of all classifiers drops. In MOOC environment, it is normal that students are active in first week and also the assessments are easy to score high compared to the later weeks. After week-4, we observe continuous growth in F1-score for almost all classifiers. Maximum score achieved after week-8 is 0.86 by RF. Different classifiers performed differently for each week data, but overall RF and LR performed better than rest of the techniques. The performance of the models looks promising, till the mid of the course (after week-4) F1-score reaches to 80 percent by LR.

Table IV shows the results of classifiers using process mining features. Using process mining features, F1-score increases for almost all weeks. After week-5 and week-6, F1-score drops but still maximum score is 0.87 by Naive Bayes. The results show that overall all classifiers performed well in prediction; however, the Naive Bayes method outperforms all methods by scoring maximum accuracy of 0.89 after week-8.

In order to measure the significance of above findings, we used the Friedman's test (Demšar, 2006) methodology for comparison of multiple classifiers over multiple data sets. The Friedman's test is a non-parametric test used to compare observations repeated on same subjects. Chi-square with k-1 degree of freedom is the test statistic for the Friedman's test, where k is the number of repeated measures. When the p-value is small ($p < 0.05$), null hypothesis is rejected. The goal of this test is to see that there is any significance difference among the performance of machine learning techniques in our experiment. Null hypothesis of our study is "There is no difference among the performance of multiple classifiers."

After applying the Friedman's test, p-values obtained are 0.02 and 0.001 for data set 1 and data set 2, respectively. As p-values are less than 0.05, null hypothesis is rejected. We conclude that there is significance difference between the performance of classifiers. The calculation of mean ranks of classifiers (from highest to lowest) shows that the LR and Naive Bayes scored highest ranks for data set 1 and data set 2, respectively, and thus outperformed other classifiers on these data sets.

Table IV. Comparative results of the effectiveness of machine learning algorithms on the data set using process mining features and mean ranks of classifiers from highest (1) to lowest (*N*)

Data set	LR	RF	NB	K.NN
Week-1	0.831	0.817	0.829	0.816
Week-2	0.831	0.833	0.861	0.796
Week-3	0.842	0.852	0.872	0.808
Week-4	0.87	0.845	0.871	0.825
Week-5	0.878	0.892	0.878	0.844
Week-6	0.854	0.889	0.88	0.835
Week-7	0.865	0.868	0.879	0.828
Week-8	0.879	0.886	0.89	0.848
Rank(mean)	2.56	2.0	*1.43*	4

Figures 8-11 show the performance of classifiers when compared with second metric, i.e. AUC. The results are almost similar like in case of F1-score. All classifiers performed better with process mining features than standard features alone:

RQ2. Is the integration of process mining features able to increase the effectiveness of the machine learning algorithm for the MOOCs problem domain?

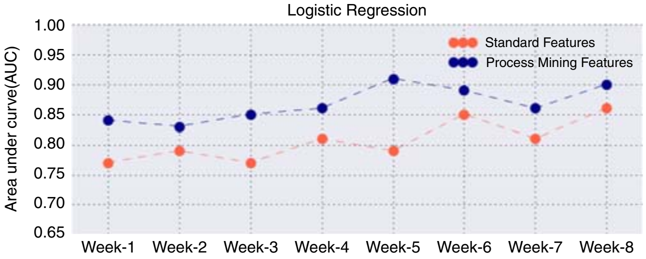


Figure 8. Comparative results of logistic regression method on the data set consisting of process mining features and without process mining features

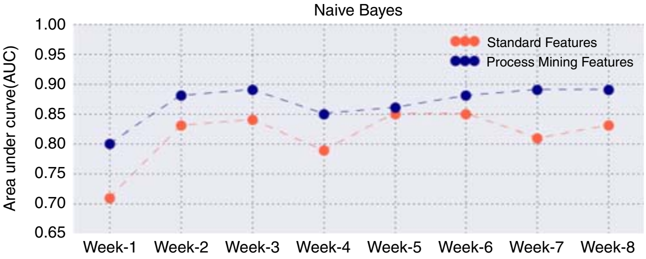


Figure 9. Comparative results of Naive Bayes method on the data set consisting of process mining features and without process mining features

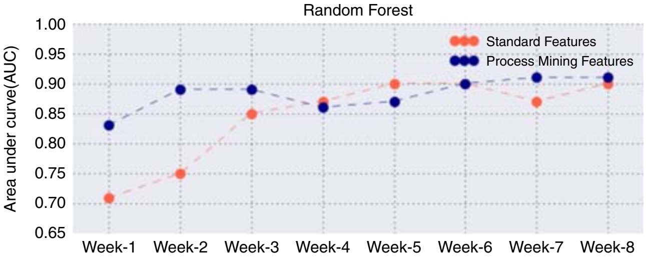


Figure 10. Comparative results of random forest method on the data set consisting of process mining features and without process mining features

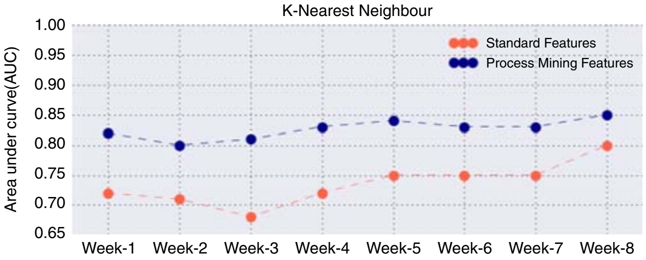


Figure 11. Comparative results of *K*-nearest neighbor method on the data set consisting of process mining features and without process mining features

In order to answer this question, we performed same prediction experiment on data set 2, which consists of features used in data set 1 and additional features obtained in the result of conformance testing. Table IV displays the F1-score obtained after evaluating machine learning algorithms on data set 2. F1-score improves after each week as expected. Figures 8-11 show the comparative results of the effectiveness of machine learning algorithms when applied on data set 1 with standard features and data set 2 which contains process mining features as additional features to the standard features. The results showed that for all weeks, F1-score improved using process mining features except for week-2 for some methods. In order to measure the significance of these results, paired *t*-test was applied on the results. Following *p*-values were obtained as a result of *t*-test: *p*-value (LR) = 0.052; *p*-value (RF) = 0.02; *p*-value (KNN) = 0.007; and *p*-value (Naive Bayes) = 0.01. According to Gorunescu (2011), in order to present a significance difference, *p*-value should be normally less than 0.05. Based on this discussion, we conclude that all classifiers present a statistically significant improvement in F1-score when process mining features were integrated with standard features, except LR.

5.1 Feature importance

In order to identify the relevant predictor variables, we used variable importance measure produced by the RF classifiers. We trained an RF model with 10,000 trees on data set 2 and rank the ten features by their respective importance measures. We chose data set 2, as it comprises of both standard features and process mining features. We wanted to investigate which features are more informative to the target variable. Table V shows the top ten important features for each week. The results show that features that measure time spent weekly and video watch activities were the most important features for all weeks. Second most important features are related to process mining.

6. Discussion and conclusion

Educational data mining provides an insight from educational data. However, most of the EDM studies used traditional data mining techniques. This work describes a possibility to integrate process mining approaches in order to achieve high prediction accuracy. The use of features obtained from process mining approach for the purpose of prediction of students' performance is novel.

We took Coursera MOOC as a case study with the focus on predicting student's performance through the traces they leave while pursuing a course. Data mining/machine learning algorithms were applied to weekly generated student data, as students are progressing through a course, in order to predict which students are at risk of not satisfying course requirements, or are rather likely to fail.

This study conducted a comparative analysis of four techniques (LR, RF, Naive Bayes and KNN). These techniques were evaluated on using two data sets, one with standard features used in the literature and second with features obtained from process conformance testing. The impact of process mining features has been analyzed on the effectiveness of mentioned techniques. The results show that techniques used in the study are able to predict the performance of students at early stage. By integrating process mining features with traditional features, effectiveness of the some techniques has improved. LR and Naive Bayes classifiers outperform other techniques in a statistical significant way for data set 1 and data set 2, respectively. We also measured the importance of features using RF classifier. The results show that process mining features were among top ten important features for all data sets; however, features related to time spent weekly and video watch activities were most important features among all.

The significance of our study is the use of process mining to enrich the features, and the results show that overall performance is statistically significantly improved using

Table V. Feature importance by random forest classifier for data set 2

Rank	Week-1	Week-2	Week-3	Week-4	Week-5	Week-6	Week-7	Week-8
1	LecLagw1	Vid-Act-w2	Timespentw3	VidActw4	VidActW4	VidActw4	VidActw4	VidActw4
2	VidActw1	VidActw1	VidActw2	Timespentw3	VidActw5	VidActw5	VidActw5	VidActw5
3	Timespentw1	LecLagw1	VidActw1	Timespentw4	Timespentw3	Timespentw3	Timespentw3	Timespentw8
4	Tracefitness	Timespentw1	Quizattempw3	VidActw2	Timespentw4	Timespentw6	Timespentw6	Timespentw3
5	Movelogfit	Timespentw2	LecLagw1	VidAc-w1	Queue-state	Timespentw4	Timespentw7	Timespentw6
6	Movemodelfit	Queuestate	Timespentw2	Tracefitness	VidActw2	VidActw2	Timespentw4	Timespentw7
7	Queuestate	Movemodelfit	Queuestate	Movemodelfit	Movelogfit	Tracefitness	Queuestate	Timespentw4
8	Rawfitcost	Movelogfit	Tracefitness	Movelogfit	Tracfitness	Queuestate	Tracefitness	Queuestate
9	Quizattmw1	Tracefitness	Movelogfit	Queuestate	Movemodelfit	Movemodelfit	Movemodelfit	Movemodelfit
10	Tracelength	Quizlagw2	Movemodelfit	Timespentw2	VidActw1	Movelogfit	Movelogfit	Tracefitness

process mining features. The limitation of this study is the missing values and the small size of the data. This study recognizes these limitations in the data sets; however, this subset (of 167 students) is representative of student participation and completion from a real-world MOOC environment. The data sets have been used to mainly demonstrate the effectiveness of data mining techniques using process mining features.

References

Ahmad, F., Ismail, N.H. and Aziz, A.A. (2015), "The prediction of students' academic performance using classification data mining techniques", *Applied Mathematical Sciences*, Vol. 9 No. 129, pp. 6415-6426.

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T. and Popelínský, L. (2012), "Predicting drop-out from social behaviour of students", *Proceedings of the 5th International Conference on Educational Data Mining*, pp. 103-109.

Benbassat, G. (1990), "An essay towards solving a problem in the doctrine of chances. La Reconnaissance Automatique De La Parole", *Rev. Laryngol. Otol. Rhinol.*, Vol. 111 No. 4, pp. 389-392.

Bergner, Y., Kerr, D. and Pritchard, D.E. (2015), "Methodological challenges in the analysis of MOOC data for exploring the relationship between discussion forum views and learning outcomes", *Proceedings of the 8th International Conference on Education Data Mining*, pp. 234-241.

Boongoen, N.I.-O.T. (2017), "Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings", *International Journal of Machine Learning and Cybernetics*, Vol. 8 No. 2, pp. 497-510.

Breiman, L. (2001), "Random forests", *Machine Learning*, Vol. 45 No. 1, pp. 5-32.

Bydžovská, H. (2016), "A comparative analysis of techniques for predicting academic performance", *Proceedings of the 9th International Conference on Educational Data Mining*, pp. 306-311.

Cambruzzi, W., Rigo, S.J. and Barbosa, J.L.V. (2015), "Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach", *Journal of Universal Computer Science*, Vol. 21 No. 1, pp. 23-47.

Campbell, J.P., DeBlois, P.B. and Oblinger, D.G. (2007), "Academic analytics", *Education Review*, Vol. 42, October, pp. 40-57.

Clow, D. (2013a), "An overview of learning analytics", *Teaching in Higher Education*, Vol. 18 No. 6, pp. 683-695.

Clow, D. (2013b), "MOOCs and the funnel of participation", *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, p. 185.

Cortes, C. and Vapnik, V. (1995), "Support-vector networks", *Machine Learning*, Vol. 20 No. 3, pp. 273-297.

Costa, E.B., Fonseca, B., Santana, M.A., de Araújo, F.F. and Rego, J. (2017), "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses", *Computers in Human Behavior*, Vol. 73, pp. 247-256.

Demšar, J. (2006), "Statistical comparisons of classifiers over multiple data sets", *Journal of Machine Learning Research*, Vol. 7, January, pp. 1-30.

Dunn, K.E. and Mulvenon, S.W. (2009), "A critical review of research on formative assessments: the limited scientific evidence of the impact of formative assessments in education", *Practical Assessment, Research and Evaluation*, Vol. 14 No. 7, pp. 1-11.

Ebner, M., Lienhardt, C., Rohs, M. and Meyer, I. (2010), "Microblogs in higher education – a chance to facilitate informal and process-oriented learning?", *Computers & Education*, Vol. 55 No. 1, pp. 92-100.

Gorunescu, F. (2011), *Data Mining: Concepts and Techniques*, Vol. 12, Springer Science & Business Media.

Gray, G., McGuinness, C. and Owende, P. (2014), "An application of classification models to predict learner progression in tertiary education", *IEEE International Advance Computing Conference*, pp. 549-554.

Hechenbichler, K. and Schliep, K. (2004), “Weighted k-nearest-neighbor techniques and ordinal classification”, Discussion Paper No. 399, Collaborative Research Center 386.

Khobragade, L.P. (2015), “Students’ academic failure prediction using data mining”, Vol. 3 No. 5, pp. 2321-7782.

Kizilcec, R.F., Piech, C. and Schneider, E. (2013), “Deconstructing disengagement: analyzing learner subpopulations in massive open online courses”, *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge (LAK ’13)*, pp. 170-179.

Lin, J.J.J. and Reid, K.J. (2009), “Student retention modelling: an evaluation of different methods and their impact on prediction results”, *Research in Engineering Education Sysmposium*, pp. 1-6.

Manhães, L.M.B., da Cruz, S.M.S. and Zimbrão, G. (2014), “WAVE: an architecture for predicting dropout in undergraduate courses using EDM”, *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pp. 243-247.

Marquez-Vera, C., Morales, C.R. and Soto, S.V. (2013), “Predicting school failure and dropout by using data mining techniques”, *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, Vol. 8 No. 1, pp. 7-14.

Martinho, V.R.C., Nunes, C. and Minussi, C.R. (2013), “Prediction of school dropout risk group using neural network”, *Federated Conference on Computer Science and Information Systems*, pp. 111-114.

Mashiloane, L. and Mchunu, M. (2013), “Mining for marks: a comparison of classification algorithms when predicting academic performance to identify ‘students at risk’ ”, in Rajendra, P. and Kathirvalavakumar, T. (Eds), Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 8284, Springer International Publishing, pp. 541-552.

Mayilvaganan, M. and Kalpanadevi, D. (2014), “Comparison of classification techniques for predicting the cognitive skill of students in education environment”, *IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1-4.

Mertes, S.J. and Hoover, R.E. (2014), “Predictors of first-year retention in a community college”, *Community College Journal of Research and Practice*, Vol. 38 No. 7, pp. 651-660.

Minaei-Bidgoli, B., Kashy, D., Kortemeyer, G. and Punch, W. (2003), “Predicting student performance: an application of data mining methods with an educational web-based system”, *33rd Annual Frontiers in Education Conference, Vol. 1*, pp. T2A_13-T2A_18.

Nandeshwar, A., Menzies, T. and Nelson, A. (2011), “Learning patterns of university student retention”, *Expert Systems with Applications*, Vol. 38 No. 12, pp. 14984-14996.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. (2011), “Scikit-learn: machine learning in python”, *Journal of Machine Learning Research*, Vol. 12, October, pp. 2825-2830.

Pittman, K. (2008), “Comparison of data mining techniques used to predict student retention”, PhD thesis, Nova Southeastern University.

Sharma, M. and Mavani, M. (2011), “Accuracy comparison of predictive algorithms of data mining: application in education sector’ ”, *Advances in Computing, Communication and Control*, Springer, Berlin and Heidelberg, pp. 189-194.

Simon, F.S., Robins, A., Baker, B., Box, I., Cutts, Q., Raadt, M.D., Haden, P., Hamer, J., Hamilton, M., Lister, R., Petre, M., Sutton, K., Tolhurst, D. and Tutty, J. (2006), “Predictors of success in a first programming course”, *Proceedings of the 8th Australian Conference on Computer Education, Vol. 52*, pp. 189-196.

Van der Aalst, W., Adriansyah, A. and Van Dongen, B. (2012), “Replaying history on process models for conformance checking and performance analysis”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 2 No. 2, pp. 182-192.

Van Der Aalst, W.M.P., Van Dongen, B.F., Ğunther, C., Rozinat, A., Verbeek, H.M. W. and Weijters, A.J.M.M. (2009), “Prom: the process mining toolkit”, CEUR Workshop Proceedings, Vol. 489, Ulm, September 8.

Veenstra, C.P., Dey, E.L. and Herrin, G.D. (2009), "A model for freshman engineering retention", *Advances in Engineering Education*, Vol. 1 No. 3, pp. 1-23.

Watson, C., Li, F.W.B. and Godwin, J.L. (2013), "Predicting performance in an introductory programming course by logging and analyzing student programming behavior", *Proceedings – 2013 IEEE 13th International Conference on Advanced Learning Technologies*, pp. 319-323.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J. and Steinberg, D. (2008), "Top 10 algorithms in data mining", *Knowledge and Information Systems*, Vol. 14 No. 1, pp. 1-37.

Ye, C. and Biswas, G. (2014), "Early prediction of student dropout and performance in MOOCs using higher granularity temporal information", *Journal of Learning Analytics*, Vol. 1 No. 3, pp. 169-172.

Yukselturk, E., Ozekes, S., Türel, Y.K., Education, C., Ozekes, S., Türel, Y.K. and Education, C. (2014), "Predicting dropout student: an application of data mining methods in an online education program", *European Journal of Open, Distance and E-Learning*, Vol. 17 No. 1, pp. 118-133.

Zhang, Y., Oussena, S., Clark, T. and Hyensook, K. (2010), "Using data mining to improve student retention in HE: a case study", *Proceedings of the 12th International Conference on Enterprise Information Systems, Vol. 1*, pp. 190-197.

Corresponding author
Rahila Umer can be contacted at: r.umer@massey.ac.nz