Prediction of Students' Dropout in MOOC Environment

Rahila Umer, Teo Susnjak, Anuradha Mathrani, and Suriadi Suriadi

***Abstract*—Massive Open Online Courses (MOOCs) is a new approach to online learning which provides a platform for learning in highly scalable and flexible manner. Many higher education institutes are developing and delivering a wide range of such courses. MOOCs are gaining popularity, however they are prone to early dropout and low completion rate. Students registering in MOOCs are different than traditional higher education students in terms of age, education background and motivation. These differences pose challenges in understanding their intent in registering for these courses. In order to improve students' retention in online learning environment, it is necessary to predict the likelihood of dropout. Timely and proper academic intervention could help struggling students during the course. In this paper, we used MOOCs dataset as a case study to predict student dropout based on the count of online activities. We used classification methods that have been utilized in the field of education domain and are suitable for imbalanced dataset. The machine learning algorithms used in our experiments are: Naive Bayes, Random Forest, Logistic Regression and K Nearest Neighbor. Our results show that techniques used in this study are able to make predictions of dropout, and Logistic Regression outperformed other classifiers with maximum accuracy.**

***Index Terms*—Learning analytics, MOOCs, machine learning, data mining, prediction.**

I. Introduction

In recent years, the rising popularity of Massive Online Open Courses (MOOCs) and online education environment has attracted a large number of participants. Such web-based education systems take advantage of the Internet capabilities to improve traditional education approaches by assisting learners in a flexible manner without the barrier of physical presence. Meanwhile, it also provides an opportunity for higher education institutes to expand their services to more students with the same time investment and less resources [1].

In order to provide an effective learning experience to students, it is necessary to monitor and observe the learning process and to provide timely support to the students. In online learning environments like MOOCs, students' participation level is higher than in a traditional classroom setting, therefore their interaction with online environment provides valuable information about their learning experience.

Despite the fact that they attract a large number of students, such online courses exhibit higher dropout rates than traditional education courses, often being more than 10-20% [2].

Student retention rate is considered as an indicator to measure the quality of an educational institute's service and similarly for online courses, the number of students completing courses is the key of acceptance and success of a course.

In order to reduce the number of dropouts, it is necessary to identify those students who are likely to dropout in a timely manner such that proper interventions to help such struggling students can be made

In a web-based education system, all activities of students are logged. These logs provide an opportunity to understand students' behavior by analyzing the digital traces they leave behind [3]. In a virtual learning environment, students' activities can be monitored by applying machine learning methods on log files obtained from learning management system databases for the purpose of identifying those students who are at-risk or are struggling through the course [4], [5].

Learning analytics and Educational Data Mining are two approaches that use data-driven techniques to address issues like dropout prediction and they involve five steps i.e. capture the data, report, predict, act and refine [6]. Higher number of dropout in the MOOCs environment increased emphasis on retention. Several research works have used machine learning techniques to address this issue [7]-[13].

In this study, the focus is on predicting students' dropout by using the event logs and to investigate the importance of such engagement activities and their correlation with the retention of the student. We used MOOCs dataset as a case study to predict student dropout based on the count of online activities. The machine learning algorithms used in the experiment are: Naive Bayes (NB) [14], Random Forest (RF) [15], Logistic Regression (LR) and K Nearest Neighbor (KNN) [16].

The remainder of the paper is organized as follows. In Section 2, we present the detail description of datasets used. Section 3 describes the method used in our study. In Section 4 we discuss the research question and discuss the results. Finally in Section 5 we make conclusions.

II. Data Description

In this study we used data from KDD Cup 2015 [17]. We chose five courses and used the event log of enrolled students. Events logs contain timestamps for following event types:

1. Problem –working on course assignment
2. Video-watching course video
3. Access-accessing other source than video and problem
4. Wiki- accessing course wiki
5. Discussion-accessing the course forum
6. Navigate-navigating other parts of course

Manuscript received August 23, 2017; revised November 14, 2017.

Rahila Umer, Teo Susnjak, and Anuradha Mathrani are with Institute of Natural and Mathematical Science, Massey University, Auckland, New Zealand (e-mail: r.umer@massey.ac.nz, T.Susnjak@massey.ac.nz, A.S.Mathrani@massey.ac.nz).

Suriadi Suriadi is with Queensland University of Technology, Brisbane, Australia (e-mail: s.suriadi@qut.edu.au4).

doi: 10.18178/ijke.2017.3.2.085

7. Page-close-closing the web page

We counted the sum of total events performed in a day. So our dataset comprised of thirty variables for each day that represents the sum of total events performed by students. The last variable is either 0 or 1 which represents the status of the student i.e. not dropout or dropout respectively. There are a total of 40 courses in the dataset; however, we only considered those courses with a larger number of students – see Table I for a detailed description of the courses. Students enrolled in a course will be considered as a dropout if they leave no online activity for 10 days after the last day of that course.

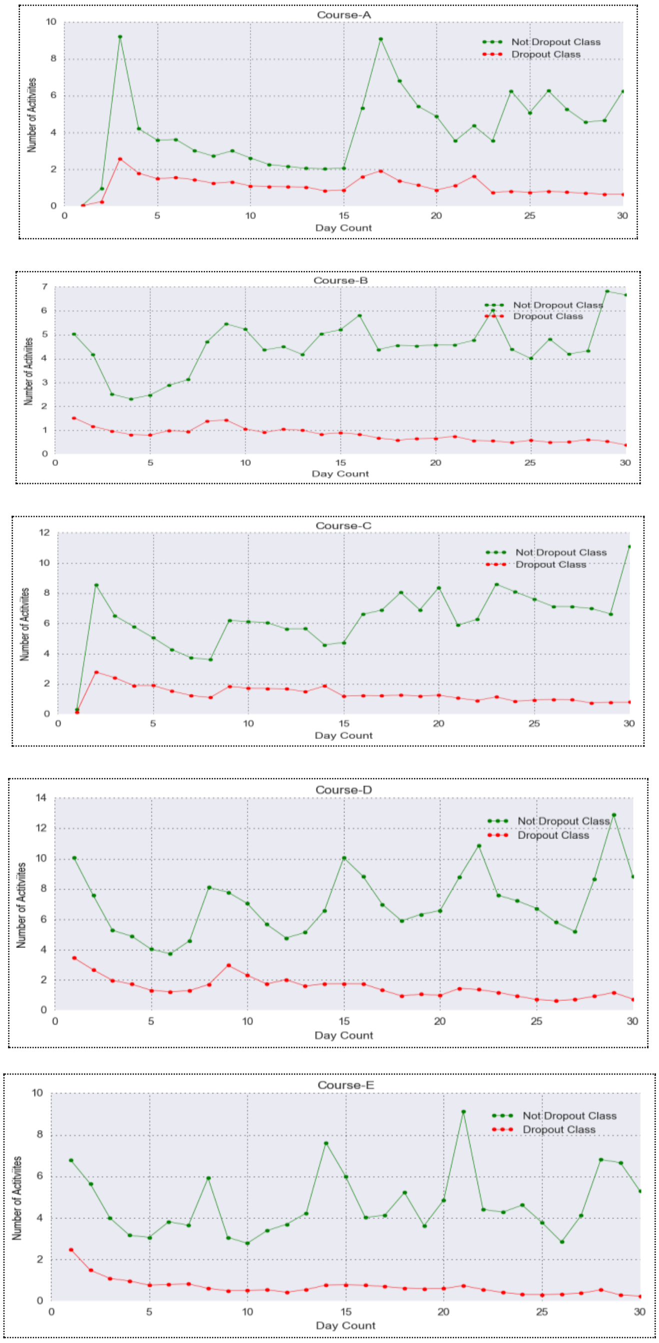


Fig. 1. Average number of activities performed by two groups of students during the 30-days course for courses A to E respectively.

Fig. 1 show the average number of activities performed by two groups of students (dropout and non-dropout) during the course. The particular duration of the course was 30 days. These figures show the difference of activity level between these two groups.

TABLE I: NUMBER OF STUDENTS ENROLLED IN THE SELECTED COURSES

Course	*Enrolled students*	*Drop-out*	*Not-Dropout*
A	120004	7186	3136
B	7775	6479	1296
C	10322	7186	3136
D	9382	6501	2881
E	3005	2597	408

III. EXPERIMENTAL DESIGN

The objective of the study is to identify the students who are most likely to drop out by using their log activities. Can the counts of log activities be used to predict the likelihood of students dropping out of the course? In our experiments, e used following machine learning algorithms: Na ïve Bayes (NB), Random Forest (RF), Logistic Regression (LR) and K Nearest Neighbor (KNN). These methods are widely using in Education Data Mining (EDM) and are considered well suited for such a domain. We performed this experiment for five datasets - one for each course. We considered five different courses with different level of engagement between two groups: shown in Figs. 1-5.

A. Evaluation Measures

To evaluate the performance of each machine learning techniques on test set, three performance criteria were used.

1) Accuracy

The overall accuracy is used to measure how good the model is for correctly predicting two groups of students (non-dropout or dropout) and is calculated in following way.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

2) F1-Score

F1-score is the harmonic mean of precision and recall and is considered a better performance measure for classification, when the dataset is imbalanced.

$$\text{F1} - \text{Score} = \frac{2\text{x Precision x Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \text{TP}/\ (\text{TP} + \text{FP})$$

$$\text{Recall} = \ \text{TP}/\ (\text{TP}+\text{FN})$$

3) ROC area under curve

A Receiver Operating Characteristic (ROC) Curve is a way to compare diagnostic tests. It is a plot of the true positive rate against the false positive rate. The Area under Curve (AUC) is the number between 0 and 1.

- False Positive Rate: FP/ (FP+TN)

- True Positive (TP) is the number of positive instances correctly classified as positive.
- False Positive (FP) is the number of negative instances incorrectly classified as positive.
- False Negative (FN) is the number of positive instances incorrectly classified as negative.
- True Negative (TN) is the number of negative examples that are correctly classified as negative.

B. Training Procedure

To estimate the generalization capability of the model for future dataset, 10-fold cross validation technique was used. Performance of the classification methods are then evaluated using overall accuracy, F1-score and using ROC curve. These classification methods are used for the prediction of students' final status in one of two classes: Dropout or Non-dropout. Prediction was based on the count of activities that students perform daily during the course.

IV. Experimental Results

In this section we will answer the following research question in the light of analysis.

Research Question: Which machine algorithm predicts the likelihood of students dropping out with high accuracy?

TABLE II: Performance of Different Classification Algorithms for MOOC Dataset

Course	*Logistic Regression*		*K Nearest Neighbor*		*Naive Bayes*		*Random Forest*	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
A	0.839	0.829	0.803	0.780	0.808	0.807	0.806	0.753
B	0.88	0.875	0.858	0.833	0.858	0.861	0.858	0.813
C	0.840	0.832	0.798	0.784	0.818	0.814	0.797	0.776
D	0.833	0.824	0.785	0.768	0.816	0.812	0.793	0.771
E	0.893	0.876	0.874	0.848	0.867	0.866	0.881	0.839

Table II shows the performance comparison for different classifiers used (i.e., NB, RF, LR and KNN). In the dataset, since counts of positive and negative class are not same, this makes it an imbalanced dataset. So a baseline classifier would give more weight to the majority class and gives a biased result. We compared the results and based on the F1 score and area under ROC curve, we chose the best classifier.

Results show that Logistic Regression outperforms other algorithms both on basis of overall accuracy and F1-score.

For all datasets, overall accuracy is 1 to 2% more than F1-score, however due to imbalanced dataset we will consider F1-score as the final metric for comparison.

Maximum accuracy obtained is 0.89 and F1-score is 0.876 for course E, which is a small dataset compared to the other datasets, it contains record of almost 3000 and difference of activity level between two groups is quite huge which makes classification easy to separate two groups with high accuracy.

Other classifiers' performance is similar, usually 3% less accurate than Logistic Regression. For all courses observed similar results and maximum F1 score is obtained by Logistic Regression. Overall maximum F1-score obtained is 0.875 for course B and course E. One similarity between courses B and E is that, in both courses, activity level for drop out students is very low and remains almost constant after first week. However, non-drop out students are active and their activity level is not constant throughout the duration of the course.

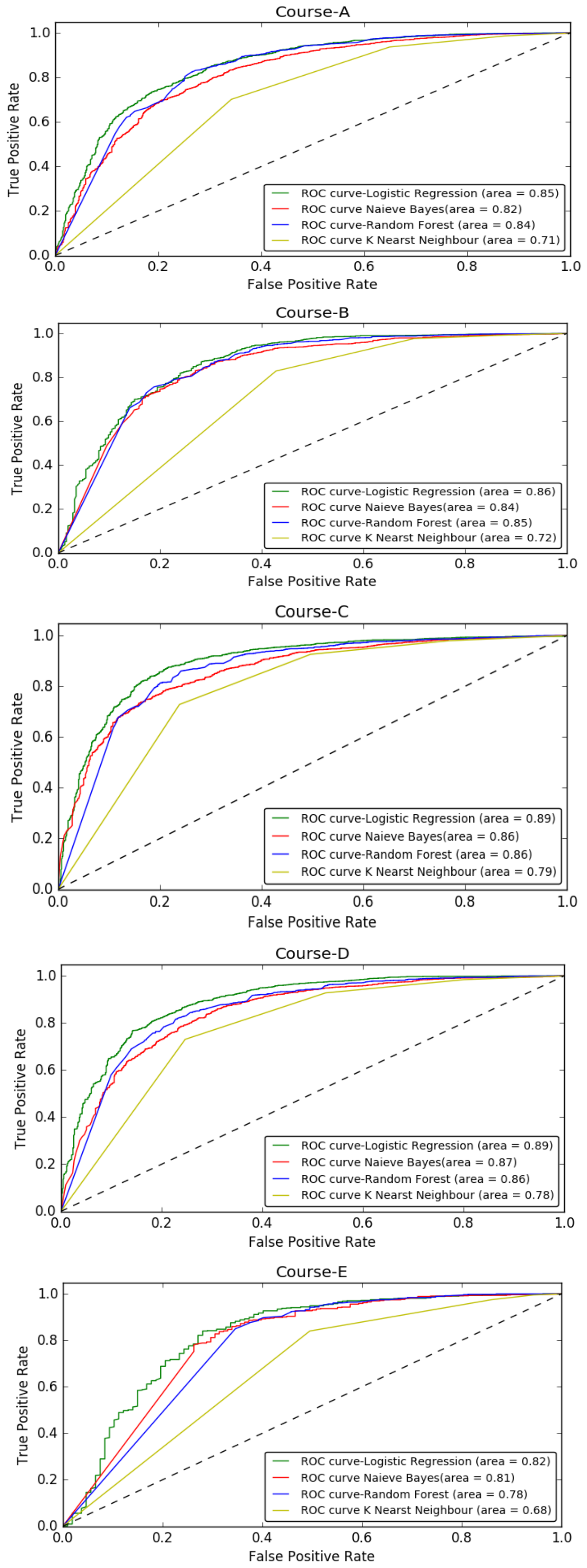


Fig. 2. Performance comparison of machine learning algorithm using ROC.

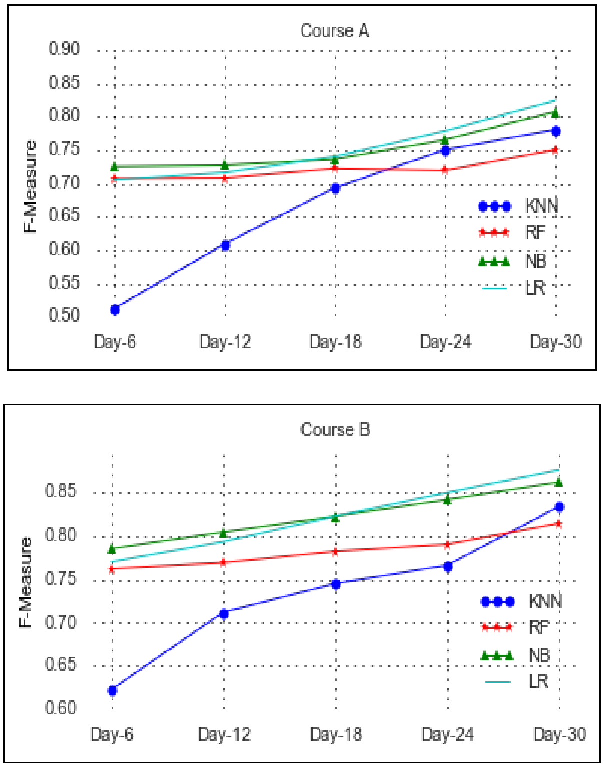
Further comparative analysis between the classifiers considered ROC. ROC curves for each of the classifiers are shown in Fig. 2. The x axis is the false positive rate or in our case it is the percent of students that continued the course and we identified as likely to drop out. The Y axis is the true positive rate or the percent of all drop outs we identified that we correctly identified as likely to drop out.

Performance across classifiers is comparable. Minimum AUC (area under curve) is 0.71 by KNN and maximum AUC (0.85) is gained 0.85 by Logistic Regression in course A. For all courses we get similar kind of results. Maximum AUC achieved is 0.89 for course C and D by logistic regression. Course C and D are among the largest datasets with almost 10000 students enrolled and activity level for not dropout students is more than the other course.

The above results show that event log data can be a strong signal for predicting students' dropout. However, these results can be regarded as baseline which can be further improved by integrating more features. Nevertheless, it is useful in the cases when we need to make early predictions during the first or second week of the courses where we did not have assessments but just the event logs of students when they interact with learning management system.

Identification of probable dropout students is only helpful when accurate prediction is made as early as possible, ideally before the mid of the course so that timely interventions can be made. Given this context, we performed predictions after every six days. A new dataset was divided into five sets: Day-6, Day-12, Day-18, Day-24, and Day-30. Here Day-12 means the event logs till 12th day of the course.

Fig. 3 shows the performance of machine learning techniques for making predictions of dropout during the course. Y axis show the F1-socre and X-axis show the dataset used for prediction.



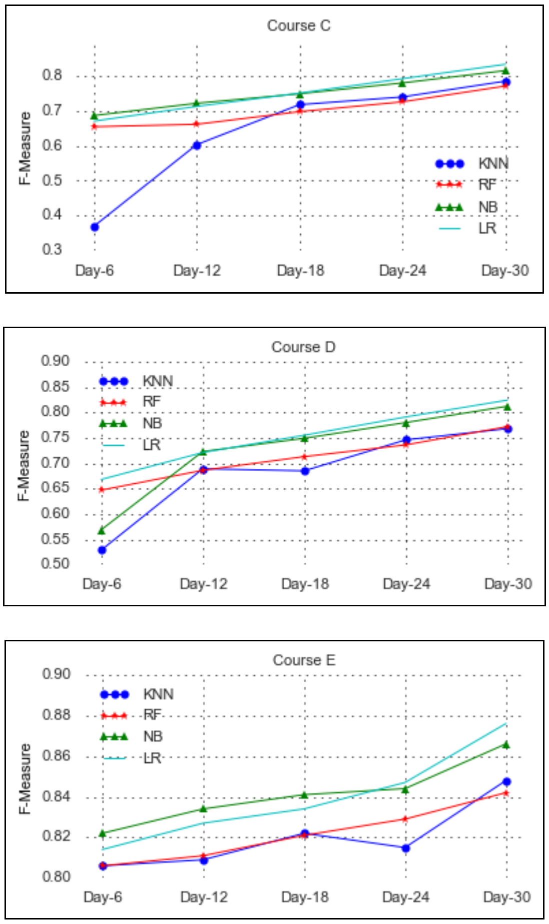


Fig. 3. Comparative results of the machine learning algorithms for prediction of dropout.

Overall results show that prediction accuracy improves with time as more engagement data becomes available. The best scores that we achieved are for course E, minimum F1-score achieved is 0.81 after 6 days which increased till 0.87 by the end of the course. Logistic regression and Naive Bayes performed better than Random forest and K Nearest Neighbor. As time increased, the prediction accuracy increased faster with KNN in comparison to other classifiers.

V. Discussion and Conclusions

In this work, we used event logs of five MOOCs courses and used to predict students that are most likely to have dropped out. Machine learning algorithms used for the classification are Random Forest, Logistic Regression, K Nearest Neighbor and Naive Bayes. Our results show that techniques used in this study are able to make predictions of dropout. However, it can be further improved by integrating more features that are directly linked to the learning process like assessments, quizzes grades etc. Nevertheless, it is useful in the cases when we need to make early predictions during the first or second week of the courses where assessments do not yet exist, but instead, the event logs of students' interaction with learning management system is available. We wanted to investigate the fact that students who are more engaged in the course are less likely to dropout.

We used event logs of five different courses, each of duration of one month and different engagement level of two groups of students (dropout and not drop-out). Results show that prediction accuracy is better in courses where there is a significant difference between the engagement levels of two groups. Logistic Regression outperformed other classifiers on this problem domain.

REFERENCES

[1] A. Cohen and R. Nachmias, "A quantitative cost effectiveness model for web-supported academic instruction," *The Internet and Higher Education,* vol. 9, no. 2, pp. 81-90, 2006.

[2] V. Carter, "Do media influence learning? Revisiting the debate in the context of distance education," *Open Learning*, vol. 11, no. 1, pp. 31-40, 1996.

[3] D. Clow, "An overview of learning analytics," *Teaching in Higher Education,* vol. 18, no. 6, pp. 683-695, 2013.

[4] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Computers & Education,* vol. 53, no. 3, pp. 950-965, 2009.

[5] V. A. Romero-Zaldivar, A. Pardo, D. Burgos, and C. D. Kloos, "Monitoring student progress using virtual appliances: A case study," *Computers & Education*, vol. 58, no. 4, pp. 1058-1067, 2012.

[6] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic analytics," *Educ. Rev.*, vol. 42, pp. 40-57, 2007

[7] C. Márquez Vera, C. R. Morales, and S. V. Soto, "Predicting school failure and dropout by using data mining techniques," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 8, no. 1, pp. 7-14, 2013.

[8] C. Ye and G. Biswas, "Early prediction of student dropout and performance in MOOCSs using higher granularity temporal information," *Journal of Learning Analytics*, vol. 1, pp. 169-172. 2014.

[9] J. Bayer, H. Bydzovská, J. Géryk, T. Obsivac, and L. Popelinsky, "Predicting drop-out from social behaviour of students," *International Educational Data Mining Society*, 2012.

[10] L. M. B. Manhães, S. M. S. da Cruz, and G. Zimbrão, "WAVE: An architecture for predicting dropout in undergraduate courses using EDM." in *Proc. of the 29th Annual ACM Symposium on Applied Computing,* ACM, 2014.

[11] V. R. Martinho, C. Nunes, and C. R. Minussi, "Prediction of school dropout risk group using neural network," in *Proc. 2013 Federated Conference on Computer Science and Information Systems (FedCSIS)* IEEE, 2013.

[12] S. Fincher, A. Robins, B. Baker, I. Box, Q. Cutts, M. de Raadt, and M. Petre, "Predictors of success in a first programming course," in *Proc. of the 8th Australasian Conference on Computing Education-Volume 52, Australian Computer Society, Inc.*. 2006, pp. 189-196.

[13] C. Watson, F. W. Li, and J. L. Godwin, "Predicting performance in an introductory programming course by logging and analyzing student programming behavior," in *Proc. of 2013 IEEE 13th International Conference on Advanced Learning Technologies (ICALT),* 2013.

[14] E. A. Freeman, G. G. Moisen, J. W. Coulston, and B. T. Wilson, "Random forests and stochastic gradient boosting for predicting tree canopy cover: Comparing tuning processes and model performance," *Canadian Journal of Forest Research,* vol. 46, no. 3, pp. 323-339, 2015.

[15] K. Hechenbichler and K. Schliep, *Weighted k-Nearest-Neighbor Techniques and Ordinal Classification*, 2004.

[16] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proc. of the Eleventh Conf. on Uncertainty in Artificial Intelligence*.

[17] KDD Cup 2015 data set description. [Online]. Available: https://goo.gl/xYZZq6.

Rahila Umer is a graduate from University of Georgia, America. She is currently a Ph.D candidate at Massey University, Auckland, New Zealand. Her research interest is in in the area of data mining, process mining and machine learning.

Teo Susnjak is a lecturer in information technology in the Institute of Natural and Mathematical Sciences at Massey University, Auckland, New Zealand. His research interests include data science, machine learning, data mining, pattern recognition, artificial intelligence, expert systems, decision support systems, software engineering.

Anuradha Mathrani is a senior lecturer in information technology in the Institute of Natural and Mathematical Sciences at Massey University, Auckland, New Zealand. Anuradha holds an engineering degree in electronics and telecommunications, a master's degree in management sciences, and a Ph.D in information technology. Her research interests include software assessment and governance methods, quality and reliability measurements, distributed software architectures, application lifecycle management and technology enhanced teaching/learning practices.

Suriadi Suriadi is a senior research fellow at Queensland University of Technology. Before this, from 2014-2016, he was a lecturer at the College of Sciences of Massey University, New Zealand. He obtained his Ph.D degree in the discipline of information security in late 2010 from the Queensland University of Technology (QUT). Since 2007, he has been involved in a number of research projects in the area of information security. From 2011 to late 2014, he was a research fellow within the business process management discipline at Queensland University of Technology, Brisbane, Australia. He enjoys working in collaborative, cross-domain research projects that allow the application of research outcomes to address real-world problems. His main research interests are in the area of process mining, data analytics, and information security.