International Journal of Informatics and Communication Technology (IJ-ICT)
Vol.10, No.1, April 2021, pp. 9~18
ISSN: 2252-8776, DOI: 10.11591/ijict.v10i1.pp9-18

Educational data mining in moodle data

Sushil Shrestha, Manish Pokharel
Department of Computer Science and Engineering, Kathmandu University, Nepal

Article Info

Article history:

Received Mar 3, 2020
Revised May 1, 2020
Accepted Dec 2, 2020

Keywords:

Classification
Educational data mining
Moodle
Prediction

ABSTRACT

The main purpose of this research paper is to analyze the moodle data and identify the most influencing features to develop the predictive model. The research applies a wrapper-based feature selection method called Boruta for the selection of best predicting features. Data were collected from eighty-one students who were enrolled in the course called Human Computer Interaction (COMP341), offered by the Department of Computer Science and Engineering at Kathmandu University, Nepal. Kathmandu University uses Moodle as an e-learning platform. The dataset contained eight features where Assignment.Click, Chat.Click, File.Click, Forum.Click, System.Click, Url.Click, and Wiki.Click was used as the independent features and Grade as the dependent feature. Five classification algorithms such as K Nearest Neighbour, Naïve Bayes, and Support Vector Machine (SVM), Random Forest, and CART decision tree were applied in the moodle data. The finding shows that SVM has the highest accuracy in comparison to other algorithms. It suggested that File.Click and System.Click was the most significant feature. This type of research helps in the early identification of students' performance. The growing popularity of the teaching-learning process through an online learning system has attracted researchers to work in the field of Educational Data Mining (EDM). Varieties of data are generated through several online activities that can be analyzed to understand the student's performance which helps in the overall teaching-learning process. Academicians especially course instructors who use e-learning platforms for the delivery of the course contents and the learners who use these platforms are highly benefited from this research.

Corresponding Author:

Sushil Shrestha
Department of Computer Science and Engineering
Kathmandu University, Nepal
Email: sushil@ku.edu.np

1. INTRODUCTION

Moodle (Modular Object-Oriented Dynamic Learning Environment) is an open-source Content Management System (CMS) software that allows instructors to provide and share course documents, assignments, quizzes, video materials, etc. for students' enabling quality online learning environment. It is a Virtual Learning Environment (VLE) or a virtual platform for users to communicate, learn and participate [1]. It is also called Learning Management System (LMS) that enables course instructors to keep track of students and assess their learning performance, evaluate the learning material such as assignments and quizzes, host text-based and video lectures. Moodle not only offers a platform to host video and text lectures, assignment, and quizzes but also offer collaborative learning with online discussion forum and keeps track of student activity via its logging system [2]. So, there is a growing use of the Moodle system for online learning. Students ineteract with the moodle system to assess the course contents and course materials. Due to this, students produce a huge amount of data through their interaction with the system for example;

students' enrolment, attendance records and examination results [3]. These varieties of data can be analyzed to extract meaningful information. So, it is important to implement the concept of Educational Data Mining (EDM) to extract useful knowledge from raw data. EDM is the process of converting raw data collected from the educational system into meaningful information [4]. Hence data mining in education also called EDM is one of the emerging research areas. It is useful in many different application areas as all the problems related to the educational environment can be managed or resolved using EDM such as *identifying weak/strong leaners, identifying learning needs, reducing dropout rates, enhancing academic achievement by improving learners' performance, improving teaching/learning processes* and so on. So, the analysis of user's online (OL) activities can be very useful to identify the hidden patterns and extract meaningful information to built the model useful for prediction. Rapid growth of educational data clears to the fact that distilling of huge amount of data requires appropriate data mining algorithm for the appropriate solution [5]. Data Mining is the extraction of knowledge from the large storage of data. Hence it is important to use EDM in order to utilize the reports that moodle keeps as well as to analyze and to build predictive models about the activity of the students that are using the system [4]. EDM uses computational approaches to analyze educational data in order to study questions [6]. It is concerned with developing methods that discover useful knowledge from data originating from educational environments. It utilizes DM methods to better understand student's performance in an educational system [7, 8]. There are several popular methods of EDM which can be applied in educational data such as *classification*, *clustering* and *regression*. Classification is a procedure in which individual items are placed into groups based on quantitative information regarding one or more characteristics inherent in the items and based on a training set of previously labeled items [6, 9]. Prediction of student performance, prediction of dropout and retention is the most popular application of classification algorithm in EDM. K-Nearest neighbors, Decision Tree, Naïve Bayes, Random Forest, etc. are the most used classification algorithm. Clustering is a technique of grouping the students according to their learning and interaction patterns [10]. Recommendation of resources, understanding, and preventing academic failure (exam failure) among university students are the common application of clustering in EDM. Hierarchical clustering and K-means are mostly used clustering algorithm. Regression is a data mining technique used to predict a range of numeric values (also called continuous values), given a particular dataset. In EDM, regression analysis has been used to predict a student's knowledge. Regression has also been applied for predicting whether the student will answer a question correctly enough, and also to create a model that illustrates the user's learning behavior [6]. So, in conclusion, EDM can be described as the process, which deals with the automatic extraction and analysis of data from large sets of data to explore previously unknown patterns [11].

This research aimed to apply EDM tools and techniques to analyze students' OL data to develop a prediction model of student performance. One of the EDM methods i.e., *classification* is mainly used for students' performance analysis and prediction. The classification technique is a supervised learning algorithm that can be used to predict categorical class labels [12]. To build a predictive model, different classification techniques such as *K-Nearest Neighbour (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), CART decision tree,* and *Random Forest (RF)* are applied in this research. Further, it also includes a comparative analysis of these models to find the best classifier which helps in the early identification of strong and weak students in the course. Also, this research aims to find the parameters affecting the performance of students.

2. RELATED WORKS

In [13], the authors had applied three widely used decision tree learning algorithms such as ID3, C4.5, and CART for the classification task. The main goal of this research was to predict student performance in the final exam. For this study, data from 90 engineering students were collected from the Institute of Engineering and Technology at VBS Purvanchal University, Jaunpur (Uttar Pradesh). This study implemented the comparative analysis among ID3, C4.5, and CART in the Weka tool. The evaluation was done using 10-fold cross-validation based on accuracy and time of execution where it was found that C4.5 had the highest accuracy of 67.7778% than other classifiers such as ID3 and CART with the same accuracy of 62.2222%. However, the execution time to build the model was less for ID3 with 0.00 sec than for C4.5 with 0.03 sec and CART with 0.09 sec.

The random forest method was used in another research to examine the important variables [14]. The data were collected from the online discussion-based learning as the first blended learning class and lecture-based learning as the second blended learning class. This study compared the prediction model based on these collected data. The experimented result revealed that in discussion-based learning, active learner's participation in online forum affect student's achievement while in lecture-based learning, submitting tasks or downloading material as main online activities doesn't affect student's achievement but only log frequency does affect i.e. both cases indicated different important features for the prediction model. Similarly, in

another study, classification techniques such as Decision Tree, Naïve Bayes, and Rule-Based were applied for the data mining process [15]. The main goal of this study was to predict students' academic performance using the classification technique. Data from 497 students were collected that included their demographics, previous academic records, and family background information. From the experimented result, the study showed Rule-Based as the best prediction model with the highest accuracy of 71.3% than other techniques such as Naïve Bayes with 67.0% accuracy and Decision Tree with 68.8% accuracy [15].

In [16], four classification techniques such as Random Forest, Naïve Bayes, K-Nearest Neighbour, and Decision Tree were applied for the students' performance prediction. This study also conducted a comparative analysis of the four classifier models based on Accuracy, Precision, Recall, and F-Measure. For this study, data from 26 students were collected with a total of 11 features such as *CourseView*, *AssginView*, *Assign_submit_update*, *ResourceView*, *ForumView*, *PT I (Overall score in programming technique I)*, *PT II (Overall score in programming technique II)*, *Assignment (total score in all assignment)*, *LabTotal (total score in all lab work)*, *Midterm* and *Performance (Students' overall grade as low, medium and high)*. The experimented result revealed that Random Forest had a better performance with an accuracy of 76.9% than KNN (69.2%) and Decision Tree (61.5%) but it was outperformed by Naïve Bayes with an accuracy of 93.3%.

3. METHODS

This section discusses the research methods followed in this research.

3.1. Data collection

This study collected the data from students enrolled in the course COMP 341 called *Human Computer Interaction (HCI)* from the Moodle system of Kathmandu University, Nepal. The types of data used for the analysis were from students' OL activities such as *System Log* and *Quiz Grades*.

3.1.1. System log

The system log consists of data of each click made in the system by a user. This dataset consists of 14839 observations and 9 attributes of 128 active users in the system. Figure 1 shows the sample of the system log dataset.

	Time	Event context	Component	Event name	Descriptio	Origin	IP address
2	1/05/18, 15:11	Quiz: quiz1	Quiz	Course module viewed	The user v	web	27.34.106.29
3	1/05/18, 15:09	Quiz: quiz1	Quiz	Quiz attempt reviewed	The user v	web	27.34.106.29
4	1/05/18, 15:09	Quiz: quiz1	Quiz	Quiz attempt submitted	The user v	web	27.34.106.29
5	1/05/18, 15:09	Course: COMP 341: Human Computer Interaction	System	User graded	The user v	web	27.34.106.29
6	1/05/18, 15:09	Course: COMP 341: Human Computer Interaction	System	User graded	The user v	web	27.34.106.29
7	1/05/18, 15:08	Quiz: quiz1	Quiz	Quiz attempt summary viewed	The user v	web	27.34.106.29
8	1/05/18, 15:07	Course: COMP 341: Human Computer Interaction	System	Course viewed	The user v	web	27.34.106.47
9	1/05/18, 15:07	Quiz: quiz1	Quiz	Course module viewed	The user v	web	27.34.106.47
10	1/05/18, 15:07	Quiz: quiz1	Quiz	Quiz attempt viewed	The user v	web	27.34.106.29
11	1/05/18, 15:07	Quiz: quiz1	Quiz	Quiz attempt reviewed	The user v	web	27.34.106.47
12	1/05/18, 15:07	Quiz: quiz1	Quiz	Quiz attempt submitted	The user v	web	27.34.106.47
13	1/05/18, 15:07	Course: COMP 341: Human Computer Interaction	System	User graded	The user v	web	27.34.106.47
14	1/05/18, 15:07	Course: COMP 341: Human Computer Interaction	System	User graded	The user v	web	27.34.106.47
15	1/05/18, 15:06	Quiz: quiz1	Quiz	Quiz attempt summary viewed	The user v	web	27.34.106.47
16	1/05/18, 15:06	Quiz: quiz1	Quiz	Quiz attempt viewed	The user v	web	27.34.106.29
17	1/05/18, 15:06	Quiz: quiz1	Quiz	Quiz attempt viewed	The user v	web	27.34.106.47
18	1/05/18, 15:06	Quiz: quiz1	Quiz	Quiz attempt viewed	The user v	web	27.34.106.47
19	1/05/18, 15:06	Quiz: quiz1	Quiz	Course module viewed	The user v	web	27.34.106.47
20	1/05/18, 15:06	Course: COMP 341: Human Computer Interaction	System	Course viewed	The user v	web	27.34.106.47
21	1/05/18, 15:05	Quiz: quiz1	Quiz	Quiz attempt viewed	The user v	web	27.34.106.29

Figure 1. System log dataset

3.1.2. Quiz grades

The grade dataset related to assessments taken during the term contains information about the scores of each student. There are 104 observations and 10 variables in the dataset. Figure 2 shows the sample of the quiz grades dataset.

3.2. Data preprocessing

In this step, only required features related to OL activities in the log system were collected from the Moodle system for the data mining process. The number of clicks by the students in the log components such as *Assignment.Click*, *Chat.Click*, *File.Click*, *Forum.Click*, *System.Click*, *Url.Click* and *Wiki.Click* was

considered as the independent feature and *Grade* as a dependent feature from the average grade scored by each student for the prediction task. A detailed description of the dataset is shown in Table 1.

	Email address	State	Started on	Completed	Time take	Grade/20.00
2	jivankandel@gmail.com	Finished	4/30/2018 12:46	4/30/2018 13:09	23 mins 5	16
3	lionelsuyog@gmail.com	Finished	4/30/2018 13:10	4/30/2018 13:18	8 mins 1 s	20
4	anuragadhikari4@gmail.com	Finished	4/30/2018 13:21	4/30/2018 13:35	13 mins 1	20
5	samyam_heli@yahoo.com	Finished	4/30/2018 13:28	4/30/2018 13:40	11 mins 2	20
6	beeshal55@gmail.com	Finished	4/30/2018 13:28	4/30/2018 13:39	10 mins 4	20
7	bodhita8@gmail.com	Finished	4/30/2018 13:31	4/30/2018 13:56	24 mins 2	19
8	neha215shrestha@gmail.com	Finished	4/30/2018 13:34	4/30/2018 14:05	30 mins 3	20
9	aintmea7@gmail.com	Finished	4/30/2018 13:52	4/30/2018 14:02	10 mins 2	17
10	rupesh.poudel07@gmail.com	Finished	4/30/2018 13:57	4/30/2018 14:12	14 mins 1	20
11	darkgamerbg15@gmail.com	Finished	4/30/2018 14:05	4/30/2018 14:14	8 mins 53	20
12	shoaibmdr@gmail.com	Finished	4/30/2018 14:09	4/30/2018 14:21	11 mins 5	20
13	jha.bibhushan1@gmail.com	Finished	4/30/2018 14:17	4/30/2018 14:26	8 mins 46	20
14	watshala.shrestha1997@gmail.com	Finished	4/30/2018 14:21	4/30/2018 14:37	16 mins 3	20
15	bipulthapa23@gmail.com	Finished	4/30/2018 14:27	4/30/2018 14:45	18 mins	20
16	rajshreerai931@gmail.com	Finished	4/30/2018 14:34	4/30/2018 14:46	11 mins 4	18
17	palludallu91@gmail.com	Finished	4/30/2018 14:35	4/30/2018 14:52	17 mins 4	20
18	aakashpaudyal@gmail.com	Finished	4/30/2018 14:44	4/30/2018 15:03	18 mins 2	20
19	anjeshojha67@gmail.com	Finished	4/30/2018 14:45	4/30/2018 14:52	6 mins 33	20
20	thesujal17@gmail.com	Finished	4/30/2018 14:49	4/30/2018 15:07	17 mins 3	20
21	binayachaudari@gmail.com	Finished	4/30/2018 14:58	4/30/2018 15:17	19 mins 5	19

Figure 2. Quiz grades dataset

Table 1. Attributes of the dataset

Features/Attributes	Descriptions	Data Types	Feature Categorization
Assignment.Click	Number of clicks on Assignment (i.e. related to HCI course assignment)	Continuous	0 – 10
Chat.Click	Number of clicks on Chat (i.e. HCI course chat)	Continuous	0 – 31
File.Click	Number of clicks on File (i.e. related to HCI course such as lecture slide, book, course content, example, guideline, etc.)	Continuous	0 – 59
Forum.Click	Number of clicks on Forum (i.e. related to announcements, survey, presentation etc. in HCI course)	Continuous	0 – 24
System.Click	Number of clicks on System (i.e. related to HCI course viewed)	Continuous	6 – 84
Url.Click	Number of clicks on Url (i.e. related to survey & videos in HCI course)	Continuous	0 – 13
Wiki.Click	Number of clicks on HCI wiki	Continuous	0 – 22
Grade	Average of Quiz1 and Quiz2 grades	Categorical	High (Grade > 16.5) Low (Grade <=16.5

3.3. Data visualization

The visualization of the data gives a quick review of the facts behind the analysis which consequently helps the researcher recognize the patterns of the students [17]. So, in this research, different charts are plotted for the visualization of data related to online learners. For example, *pie chart* to visualize users' most active participation in online learning, *correlation plot* to visualize the relationship between features i.e., between independent features (users' online activities in the system) and dependent feature (users' grade in the course), and *Boruta plot* to visualize the most influencing features of the students' performance in the course.

3.4. Feature selection

Feature selection is the process of selecting the relevant feature subset from the collection of whole features that contained several irrelevant features. So, in this step, irrelevant features were removed which consequently helped in building high accuracy prediction model. Likewise, there are other benefits such as reduce overfitting of data, reduce the complexity of computation, increase model classification accuracy, etc. [18]. The feature selection technique is very useful to analyze the relationship between the independent features and dependent features i.e., from the analysis, it helps to identify the most influencing independent features to the dependent feature [19]. The study in this research focuses to analyze the student OL features that have a greater impact on the students' performance. In this study, one of the wrapper-based feature selection method called *Boruta* was applied for the feature selection task. Boruta is a wrapper method built around the random forest classification algorithm [20]. The main advantage of using *Boruta* is that it decides

where a feature is important or not i.e. it select the statistically significant features. So, it helps to obtain all the important features from the dataset concerning the target feature.

3.5. Applying classification technique

Classification is a supervised learning technique used for the prediction of a predefined class or group. It is widely applied in students' performance prediction. There are many techniques such as Decision Tree, Naïve Bayes, SVM, Neural Network, K Nearest Neighbour, and Random Forest that can be applied to the students' dataset for the classification task [21]. In this study, K Nearest Neighbour, Naïve Bayes, SVM, Random Forest, and CART decision tree were applied for the classification task.

3.5.1. K-nearest neighbour

K Nearest Neighbor (KNN) is a classification technique used for classifying unknown objects based on the closest neighbor whose class is already known (i.e., it is an instance-based classifier that operates on unknown instances) [22]. This classifier retains the entire training set during learning and assigns a class to a new unknown instance represented by the major vote of its nearest neighbor label in the training set. KNN is the simplest algorithm that is easy to understand and implement for the classification task [23, 24].

3.5.2. Naïve bayes

Naïve Bayes (NB) is a simple probabilistic classifier that finds a probabilistic relationship between classes and their attributes [19]. NB algorithm is based on the Bayesian theorem that computes the probability of the target on a given predictor or attribute values. It is a better probabilistic classifier that can compute the most possible output based on the input and has proven to work satisfactorily in many application domains [21]. It is used when the dimensionality of input is high [12].

3.5.3. Support vector machine (SVM)

SVM is one of the supervised learning algorithms used for classification and regression. It is a new classification method used for both linear and nonlinear data [23]. It is one of the most popular classification techniques to predict accurate results for most of the classification and prediction problems [12].

3.5.4. CART decision tree

CART (Classification and Regression Trees) algorithm is a decision tree algorithm that can be used to build both classification and regression decision trees. It can handle both the categorical and numerical attributes. For example, CART is said to be a classification decision tree if it is used to predict a dataset into two classes. It is said to be a regression decision tree if it is used to predict a numerical variable. It has some advantages such as: it handles the missing values and uses the cost complexity pruning to remove the unreliable branches from the decision tree to increase the accuracy [13].

3.5.5. Random forest

Random Forest (RF) is a supervised machine learning algorithm used for classification, regression, and other tasks. Decision tree split each node using the best among the attributes while RF split each node using the best among the randomly chosen subset of predictors at the node. Hence RF performs well compared to other classification techniques such as SVM and neural network. Besides this, it is robust against overfitting [16].

3.6. Result evaluation

After the implementation of the classification technique on the students' dataset, the results of the classifier models were evaluated and reviewed to get the viewpoint of the result and find the best prediction model. This study evaluated models using four commonly used performance measure metrics such as *Accuracy*, *Recall*, *Precision*, and *F1 (F-measure)* where these matrices were calculated using the confusion matrix [17]. In classification problems, good accuracy in classification is the primary concern [18]. So, confusion matrix is the suitable method to determine the accuracy which is shown in Table 2. "A confusion matrix of size n x n associated with a classifier shows the predicted and actual classification, where n is the number of different classes" [18].

Table 2. Confusion matrix

		Predicted	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

Whereas,

a) *TP (True Positive):* Classifier correctly labeled cases as positive
b) *TN (True Negative):* Classifier correctly labeled cases as negative
c) *FP (False Positive):* Classifier incorrectly labeled cases as positive
d) *FN (False Positive):* Classifier incorrectly labeled cases as negative

Accuracy is the ratio of the number of all correct predictions to the total number of the dataset.

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

The *recall* is the ratio of the number of correct positive predictions to the total number of positives.

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

Precision is the ratio of the number of correct positive predictions to the total number of positive predictions.

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

F1 (F-measure) is a harmonic mean of precision and recall i.e. it is the weighted average of precision and recall which shows the relationship between them.

$$F - measure = \frac{2*P*R}{P+R}$$

Where *P* is Precision and *R* is Recall.

4. RESULTS AND DISCUSSION

The outcome of this research is divided into two parts. First is the *visualization of data* and second is the *implementation of classification techniques*. Figure 3 represents a sample of a preprocessed dataset of students collected from a moodle system of 81 students from the course called Human Computer Interaction (COMP 341), offered by the Department of Computer Science and Engineering of Kathmandu University, Nepal.

System.clicked	Course.viewed	Q_attempt.Date	Timetaken	Grade.20.00
55	2	1	19	20
41	1	1	30	18
2	2	0	17	19
34	3	1	10	17
47	5	1	16	20
38	7	1	7	20
28	1	1	14	20
6	2	0	11	20
28	1	1	13	20
54	2	1	26	13
10	4	0	10	20
59	6	1	20	20
6	3	0	26	20
4	2	0	18	15
29	2	1	9	20
38	2	1	19	19
43	1	1	18	20
40	9	1	11	20
30	1	1	9	20
32	1	1	25	19
2	1	0	25	20
3	2	0	21	20

Figure 3. A preprocessed dataset

4.1. Visualization of data

This section visualizes the moodle data related to students' log and quiz grades. Figure 4 shows the online activities such as *File.Click*, *System.Click*, *Forum.Click*, *Chat.Click*, *Assignment.Click*, *Wiki.Click* and *Url.Click* were plotted in a pie chart in percentage. It shows that *System.Click* (47.38%) and *File.Click* (30.82%) are in high number than other activities such as *Forum.Click* (12.34%), *Url.Click* (4.23%), *Assignment.Click* (2.94%), *Wiki.Click* (1.19%) and *Chat.Click* (1.1%). Table 3 shows the frequency (count) of users' online activities in moodle.

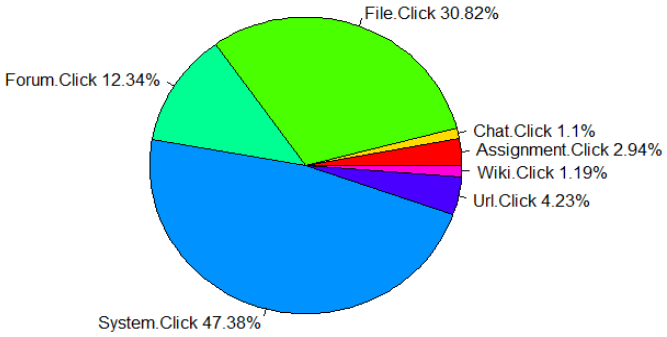

Figure 4. Pie chart of users' online activities

Table 3. Frequency of online activities

Activities	Count
Assignment.Click	228
Chat.Click	85
File.Click	2388
Forum.Click	956
System.Click	3671
Url.Click	328
Wiki.Click	92

Figure 5 shows the correlation plot between features. *File.Click*, *System.Click* and *Url.Click* has the highest significant relationship to the *Grade* than other features such as *Assignment Click*, *Chat.Click*, *Forum.Click* and *Wiki.Click*.

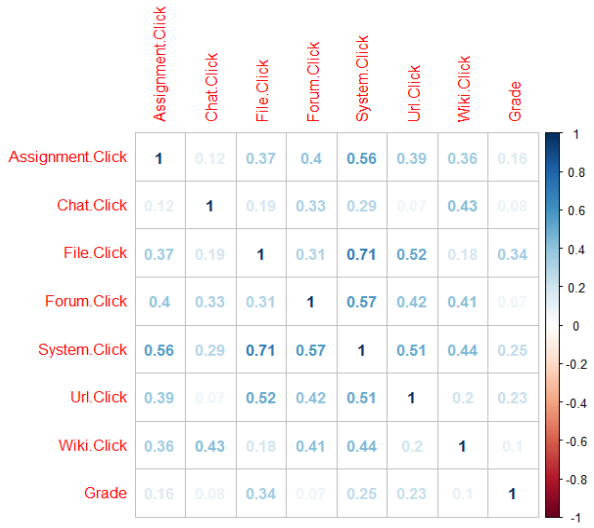

Figure 5. Correlation plot between features

Figure 6 shows the Boruta result plot for the variable importance. Blue box plots represent the minimum, average, and maximum Z score of shadow attributes. These are not actual attributes but are used by the Boruta algorithm to decide whether a variable is important or not based on the default P-value of 0.01 confidence level for the significance test. Red box plots represent Z score for rejected variables importance and green box plots represent Z score for confirmed variables importance that is good predictors to include in a feature classification model. The yellow box plot represents a tentative variable that Boruta is unable to decide whether a variable is important or not [21]. Table 4 shows the features with a mean Imp and a decision of confirmed and rejected feature importance by fixing the tentative feature (i.e. *Url.Click*). Figure 6 and Table 4 shows that *File.Click* and *System.Click* are the features confirmed to be significant while other

features such as *Assignment.Click, Chat.Click, Forum.Click, Url.Click* and *Wiki.Click* are rejected to be significant (i.e., insignificant).

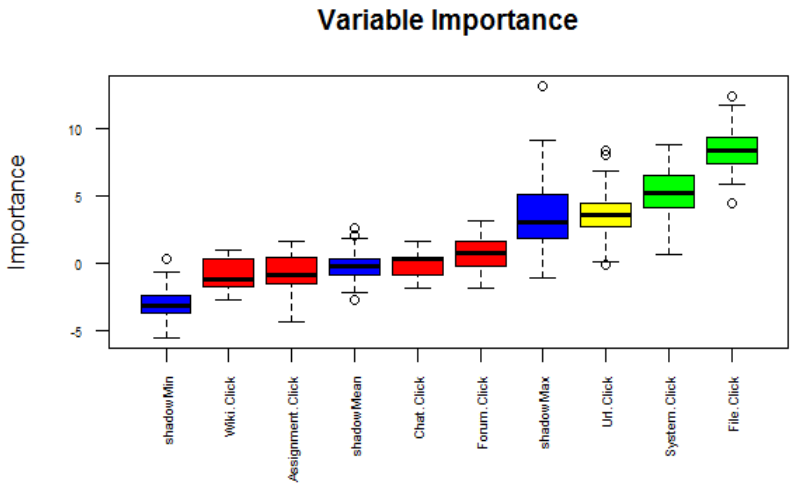

Figure 6. Boruta result plot for variable importance

Table 4. Variable with meanImp score and decision

Features	meanImp	decision
Assignment.Click	-0.76998092	Rejected
Chat.Click	0.06111646	Rejected
File.Click	8.45510448	Confirmed
Forum.Click	0.80458942	Rejected
System.Click	5.27863130	Confirmed
Url.Click	3.64773971	Rejected
Wiki.Click	-0.82471889	Rejected

4.2. Implementation of classification techniques

Moodle data was collected from the course called Human Computer Interaction (COMP 341). The dataset is divided into two parts i.e. 60% to the training set and 40% to the testing set. For the classification task, training data are learned by the classification algorithms to construct a model. Testing data are then used to train a model to estimate the accuracy of the classifier model. 5-fold Cross-Validation (CV) method is applied. Models are evaluated based on the commonly used performance measure metrics such as *Accuracy, Recall, Precision,* and *F1*. Figure 7 shows the performance of five classifier models without feature selection method where it is found that *SVM* has the highest *accuracy* of 93.94% than other classifiers such as *KNN* (87.88%), *RF* (84.85%), *CART* (84.85%), and *NB* (69.70%) on the testing dataset with all features such as *Assignment.Click, Chat.Click, File.Click, Forum.Click, System.Click, Url.Click* and *Wiki.Click* as independent features and *Grade* as a dependent. Also, comparatively, other metrics of *SVM* such as *Recall* and *F1* shows the highest result of 100% and 96.88%, *Precision* and *accuracy* of 93.94% which suggest that 31 of 33 students are correctly classified to the right class labeled as *High* and *Low* while the remaining 2 students are incorrectly classified. Further *SVM* with 100% *Recall* suggests that 100% of cases are correctly classified to the total number of unclassified and correctly classified cases. Likewise, *SVM* with 93.94% *Precision* suggests that 93.94% of cases are correctly classified to the total number of misclassified and correctly classified cases.

Figure 8 shows the performance of five models based on Boruta wrapper-based feature selection method. From Figure 6 and Table 4, most influencing features are found to be *File.Click* and *System.Click,* which is selected as the independent features for the model and *Grade* as a dependent. Although *SVM* and *CART* show the same accuracy result of 93.94% and 84.85% in without feature selection case, other classifiers such as *KNN, NB,* and *RF* shows different accuracy results. *KNN* has an *accuracy* of 90.91% (87.88% in the case of without feature selection), *NB* has an *accuracy* of 81.82% (69.70% in the case of without feature selection) and *RF* has an *accuracy* of 78.79% (84.85% in the case of without feature

selection). This suggests that the selected most influencing features increase the accuracy of the prediction model. However, *RF accuracy* is reduced which suggests that it can handle high dimensionality of data for the prediction.

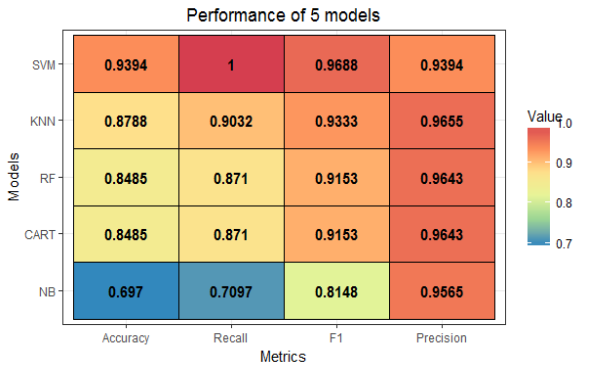

Figure 7. Performance of 5 classifier models without feature selection method

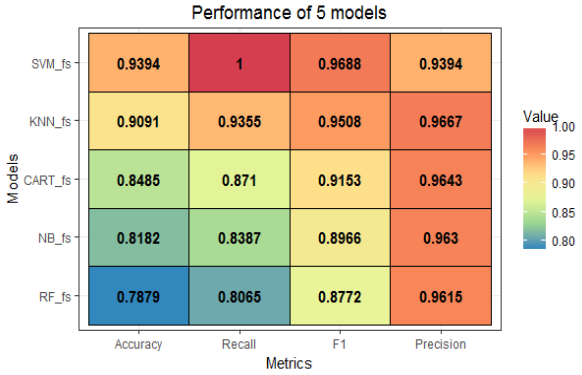

Figure 8. Performance of 5 classifier models with feature selection method

The result shows that *SVM* has the highest *accuracy* of 93.94% than other classifiers' accuracies such as *KNN* (87.88%), *RF* (84.85%), *CART* (84.85%), and *NB* (69.70%) on the testing dataset before feature selection methods. After the feature selection method, two features *File.Click* and *System.Click* is found to be the most influencing features for the prediction model. So, for the further experiment, *File.Click* and *System.Click* is considered as independent features and *Grade* as a dependent. The result with selected features shows that *SVM* and *CART* have the same accuracy as before feature selection. However, other classifiers such as *KNN, NB,* and *RF* show different accuracy results.

5. CONCLUSION

The finding of this research suggests that the selected most influencing features increase the accuracy of students' performance prediction model i.e. *there is a strong relationship between users' online activities and their performance*. The study helps in early identification of students' performance and course-related problems that they can plan early with a proper decision such as improvement of teaching/learning processes, proper counseling to the weak learners. As a result, it will improve the students' performance and reduce the number of students' dropout in the course which consequently helps in the academic achievement of the institution. Future work can be implementing the concept of learning analytics (LA), which focuses on informing and empowering instructors and learners. Both EDM and LA can be applied to educational data to get meaningful information for the improvement of the teaching-learning process. To achieve this task, a user model can be developed that fit with the instructional design strategies. Different types of learner's data related to their background, motivation, learning styles, and cognition can be linked with their log data. The focus can be on explanatory models that highlight the relationship between these data.

REFERENCES

[1] R. S Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions", *JEDM Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3-16, 2009.

[2] S. B. Aher and L. M. R. J. Lobo, "Course recommender system in e-learning". *International Journal of Computer Science and Communication*, vol. 3, no. 1, pp. 159-164, 2012.

[3] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining". *IEEE Access*, vol. 5, pp. 15991-16005, 2017.

[4] A. Algarni, "Data mining in education", *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 6, pp. 456-461, 2016.

[5] D. Krpan and S. Stankov, "Educational data mining for grouping students in e-learning system", In *Proceedings of the Iti 2012 34th International Conference on Information Technology Interfaces*, pp. 207-212, 2012.

[6] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601-618, 2010.

[7] A.M. De Morais, J. M. Araujo and E. B. Costa, "Monitoring student performance using data clustering and predictive modelling", In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, pp. 1-8, IEEE, 2014.
[8] S. Rana and R. Garg, "Evaluation of student's performance of an institute using clustering algorithms", *International Journal of Applied Engineering Research*, vol. 11, no. 5, pp. 3605-3609, 2016.
[9] P.G. Espejo, S. Ventura and F. Herrera, "A survey on the application of genetic programming to classification", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 2, pp. 121-144, 2010.
[10] C Romero and S. Ventura, Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12-27, 2013.
[11] O. Maimon and L. Rokach, "Decomposition methodology for knowledge discovery and data mining", In *Data mining and knowledge discovery handbook*, pp. 981-1003, Springer, Boston, MA, 2005.
[12] S.S. Nikam, "A comparative study of classification techniques in data mining algorithms", *Oriental Journal of Computer Science & Technology*, vol. 8, no. 1, pp. 13-19, 2015.
[13] S. K. Yadav and S. Pal, "Data mining: A prediction for performance improvement of engineering students using classification" *World of Computer Science and Information Technology Journal WCSIT*, vol. 2, no. 2, pp. 51-56, *arXiv preprint arXiv: 1203.3832*, 2012.
[14] J. H. Kim, S. Seodaemum-gu, Y. Park, J. Song and I. H. Jo, "Predicting students' learning performance by using online behavior patterns in blended learning environments: Comparison of two cases on linear and non-linear model", *International Conference on Educational Data Mining (EDM 2014)*, 2014.
[15] F. Ahmad, N. Ismail and A. A. Aziz, "The prediction of students' academic performance using classification data mining techniques", *Applied Mathematical Sciences*, vol. 9, no. 129, pp. 6415-6426, 2015.
[16] Y. Abubakar, and N.B. Ahmad, "Prediction of Students' Performance in E-Learning Environment Using Random Forest", *International Journal of Innovative Computing*, vol. 7, no. 2, pp. 1-5, 2017.
[17] E. A. Amrieh, T. Hamtini and I. Aljarah, "Mining educational data to predict Student's academic performance using ensemble methods", *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119-136, 2016.
[18] S. Visa, B. Ramsay, A.L. Ralescu and E. Van Der Knaap, "Confusion Matrix-based Feature Selection", *MAICS*, *710*, 120-127, 2011.
[19] A. Acharya and D. Sinha, "Application of feature selection methods in educational data mining", *International Journal of Computer Applications*, vol. 103, no. 2, pp. 34-38, 2014.
[20] A. Mueen, B. Zafar and U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques", *International Journal of Modern Education and Computer Science*, vol. 11, no. 11, pp36-42, 2016.
[21] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package", *J Stat Softw*, vol. 36, no. 11, pp. 1-13, 2010.
[22] M. Al-Saleem, N. Al-Kathiry, S. Al-Osimi and G. Badr, "Mining educational data to predict students' academic performance", *In International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, Cham, pp. 403-414, 2015.
[23] H. Bhavsar and M. H. Panchal, "A review on support vector machine for data classification", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET*), vol. 1, no. 10, pp-185, 2012.
[24] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier," International Journal of Information Engineering and Electronic Business, vol. 8, no. 4, pp. 54-62, *arXiv preprint arXiv:1610.09982*, 2016.

BIOGRAPHIES OF AUTHORS

Mr. Sushil Shrestha is an Assistant Professor in the Department of Computer Science and Engineering at Kathmandu University, Nepal. He is a Lead Researcher of Digital Learning Research Laboratory at Kathmandu University, Nepal. He is also a PhD scholar and doing PhD in the field of Educational Data Mining and Learning Analytics.

Dr. Manish Pokharel is a Professor in the Department of Computer Science and Engineering at Kathmandu University, Nepal.