JER
KUSOED

Journal of Education and Research
2019, Vol. 10, No. 2, pp. 27-51
https://doi.org/10.3126/jer.v10i2.32721

Article History: Received: 03 May 2020; Revised: 12 July 2020; Accepted: 25 August 2020

Original Research

Data Mining Applications Used in Education Sector

Sushil Shrestha* and Manish Pokharel
Digital Learning Research Lab, Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal

Abstract

The purpose of this work is to study the usage trends of Data Mining (DM) methods in education. It discusses different data mining techniques used for different types of educational data. The related papers were initially selected from the metadata containing words like *Online Learning (OL)* and *Educational Data Mining (EDM)*. The papers were then filtered on the basis of *DM algorithms*, *the purpose of study*, and *the types of data used*. The findings suggested that EDM is the most commonly used technique for the prediction of students' academic success, and the most used purpose is *classification*, followed by *clustering* and *association*. Further, this research also contains the study conducted on moodle data to find anomalies. K-means clustering was applied to find the optimal number of clusters on moodle data that consists of *log* and *quiz* dataset. The growth in the number of Internet users has increased learning through the online process. Hence, several activities are performed in OL systems, which generate a massive amount of data to be analysed to obtain useful information. Therefore, this type of research is very beneficial to academicians and instructors to identify the learner's behaviors and develop suitable models.

Keywords: *Online learning, LMS, data mining, educational data mining, learning analytics*

* Corresponding Author.
sushil@ku.edu.np

ISSN: 2091-0118 (Print) / 2091-2560 (Online)

Journal homepages: [1]http://www.kusoed.edu.np/journal/index.php/je
[2]https://www.nepjol.info/index.php/JER/index

Published by Kathmandu University School of Education, Lalitpur, Nepal.

Introduction

The use of information and communication technology (ICT) in education has been rapidly growing in recent years. The use of ICT in education is gradually turning the conventional classroom teaching environments into online learning (OL) environment. The OL system improves the learning experience of students and reduces the need for the direct involvement of the instructor. As the OL system has become accessible through the internet, students can enroll themselves in the courses from anywhere and can be involved in different learning activities. It provides a huge repository of data through the Learning Management System (LMS). To analyse these data, different DM algorithms need to be applied to obtain meaningful information and represent it in a way to facilitate the process of decision making to increase the effectiveness of the learning process. Different types of data are generated from different sources, such as *student attendance records*, *course information*, *curriculum*, and *classroom scheduled information*. Similarly, data of varied and diverse kinds are also produced from various other web-based applications deployed in an educational environment such as *educational games*, *virtual environments*, *discussion forums*, *notice board*, *interactive multimedia systems*, *online test/quizzes*, *user's activity logs*, and *various other learning contents and text*.

The paper summarises several works done in the field of educational data mining. It addresses mainly two research questions: *what are the different data mining algorithms applied in educational data?* and *what are the different types of data used in EDM?* Hence the aim of this research paper is to disseminate the information about different EDM algorithms and different types of educational data used for the analysis. The related research papers were selected based on the trends of DM methods in educational data. So, the data used in this research were from the education sector, which includes *students' personal records*, *previous academic records*, *log from the user's interaction with the system*, *midterm assessment records*, and *survey questionnaires*.

This research paper also includes an analysis done in the moodle data of an undergraduate course called Human-Computer Interaction (COMP 341), offered by the Department of Computer Science and Engineering at Kathmandu University, Nepal. The study was conducted to find outliers in the moodle data. To achieve this task, unsupervised learning technique, i.e., *k-means* clustering method, is used. The paper

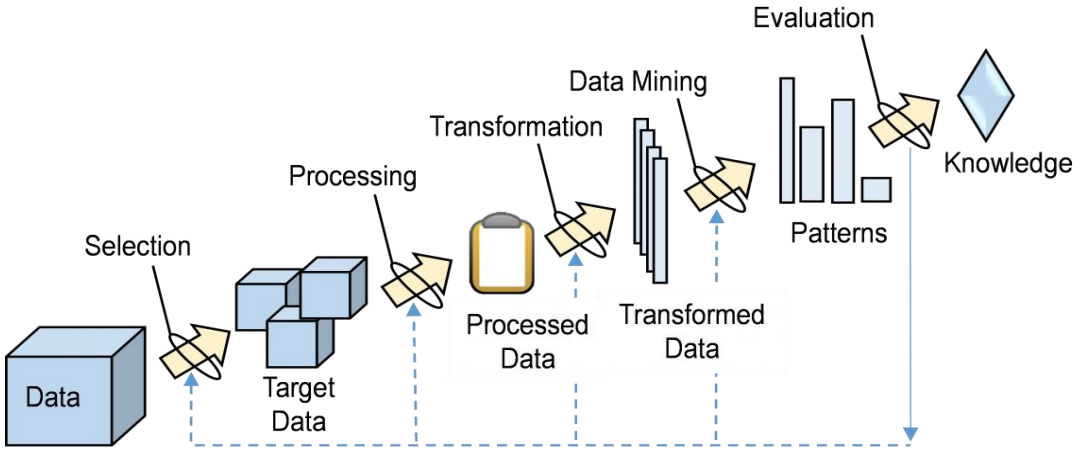
concludes with the research directions, which include the expansion of research that can be conducted in the future. The paper identifies *educationists, course instructors, online learning system administrators,* and *researchers* working in the field of EDM as the major stakeholders who will be benefited from this research paper.

Data Mining

Data Mining (DM) is the process which deals with the automatic extraction and analysis of data from large sets of data to explore previously unknown patterns (Maimon & Rokach, 2005). It is a core step of *Knowledge Discovery in Databases (KDD). KDD* is the process of extracting information from data in the context of large data sets. It is iterative and interactive and consists of six steps, as shown in Figure 1. It starts with understanding the domain and ends with discovering the knowledge from the patterns generated by using DM methods. The discovered knowledge from the KDD process is used for different purposes, such as *understanding student's behavior, assisting instructors, improving teaching methods,* and *evaluating and improving e-learning systems* (Romero et al., 2008). Hence DM involves methods to search for new and generalizable relationships and findings rather than attempting to test prior hypotheses (Collins et al., 2002).

Figure 1

Process of Knowledge Discovery in Database (Data Mining, 2019)



Educational Data Mining

Educational Data Mining (EDM) is the application of DM techniques to educational data (Romero et al., 2008). It is concerned with developing methods that discover useful knowledge from data originating from educational environments. It utilises DM methods to better understand student's performance through educational systems (De Morais et al., 2014; Rana & Garg, 2016). International Educational Data Mining Society described EDM as a new field of study that aims to prepare techniques for exploration of a specialised form of information received from the educational sector and the application of these techniques to improve understanding about students and the environment in which they imbibe knowledge (Siemens & Baker, 2012). It uses computational approaches to analyse educational data (Romero & Ventura, 2010). According to Romero and Ventura (2010), "EDM seeks to use these data repositories to better understand learners and learning, and to develop computational approaches that combine data and theory to transform practice to benefit learners" (p. 601). They classified the contribution provided by EDM activities into several categories such as analysis and visualization of data, Providing feedback for supporting instructors, Recommendations for students, Predicting student performance, Student modeling, Detecting undesirable student behaviors, Grouping students, Social network analysis, Constructing courseware, Developing concept maps and Planning and scheduling.

Educational Data Mining Techniques

There are several popular methods of EDM that can be applied in educational data, such as *classification*, *clustering*, and *regression*. Classification is a procedure of grouping of individual items based on quantitative information (Vasani & Gawali, 2014). Clustering is a technique of grouping the students according to their learning and interaction patterns (Romero & Ventura, 2013). This research paper implements the clustering technique in the moodle data of the course COMP 341. Regression is a DM technique used to predict a range of numeric values (also called continuous values), given a particular dataset. In EDM, regression analyses are 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 student's learning behavior (Romero & Ventura, 2010). Similarly, other methods like

text mining, *association rule mining*, *social network analysis*, *discovery with models*, *visualization*, and *statistics* are also used in EDM.

The process of EDM is conducted through the following steps (Rana & Garg, 2016):

1. *Data Cleaning*: Raw, noisy, and inconsistent data is cleaned by using various data cleaning methods such as smoothing (used to smooth the noisy data).
2. *Data Integration*: Data from varied sources are combined in a coherent data store.
3. *Data Selection*: Relevant data required for the analysis is selected from the database.
4. *Data Transformation*: Selected data is transformed into the forms appropriate for mining.
5. *Data Mining*: Data patterns are extracted using different intelligent methods of machine learning and artificial intelligence.

Further, Rana and Garg (2016) listed some of the applications of EDM as follows:

- *Data Analysis:* It included the analysis of educational data to assist the course administrators for student's academic performance related to decision making.
- *Students' Performance Prediction:* EDM can be used to predict student's performance using attributes such as grades, class performance, and position in the class.
- *Grouping Students:* Clustering algorithms like *Hierarchical clustering* and *K-means clustering* algorithms can be used to group the students according to their respective nature and academic performance.
- *Classification:* Classification algorithms such as *Decision Trees*, *Naïve Bayes*, and *Logistic Regression* can be used to predict student's performance.

Slater et al. (2017) focused on EDM tools and other tools frequently used to conduct EDM analysis. These authors had covered the primary tools used by some of the core research groups and/or organizations in the field. They had mentioned *prediction* as one of the objectives of the analysis. Fauvel and Yu (2016) mentioned that the predictive method could be further split into *a) sequential prediction and interpolation* and *b) supervised Learning and Descriptive*, which is further split into *a) clustering* and *b) exploratory analysis*. *Predictive methods* are used to obtain single or multiple variables

with predicted value from the predictor variable or group of variables. It is divided into three types: *classification, regression,* and *prediction* (Anoopkumar & Rahman, 2015). *Classification* is used to predict class labels in the form of perpetuating or discrete. The most commonly used classification approaches in EDM make use of *decision trees* and *logistic regression* (Anoopkumar & Rahman, 2015). Regression is utilised to derive a prediction from continuous variables. The most common regression approaches for EDM are *neural networks* and *linear regression* (Anoopkumar & Rahman, 2015). The next is the *prediction of density*, where predicted values are derived by using probability density function. Different kernel functions can be used in EDM to estimate the value of density, including Gaussian functions (Anoopkumar & Rahman, 2015).

The computational approach of EDM has led to research on learning analytics (LA). EDM and LA communities have emerged as alternatives to frequentist and Bayesian approaches for working with educational data (Romero & Ventura, 2007; Baker & Siemens, 2014). As the Society for Learning Analytics Research defines, *Learning Analytics* (LA) are the compilation, quantification, analysis, and notification of information related to students in relation to their individual characteristics so that the process of learning can be well understood and improved upon along with the surroundings in which it takes place (as cited in Siemens & Baker, 2012). The major goal of the LA goal is to improve student learning by giving a better environment **(**Simon, 2017).

Literature Review

El-Halees (2009) analysed students' learning behavior using EDM. The EDM methods of *association rules, classification, clustering*, and *outlier detection* were applied in this research. The study showed the usefulness of DM in higher education to improve student performance. To achieve this task, *association rules* were discovered from the data for excellent final grade students, *classification rules* were discovered using a decision tree, *clustering* of the student into a group was done using *Expectation-Maximization (EM),* and *outlier analysis* was conducted to detect outliers in data. Initial preprocessing of the data discovered that attendance, students' GPAs and lab grades were directly related to the final grades. Association rule suggested that students who failed in the final term also failed in midterms. A total of 37 outliers were found that could be used by the instructor to find out the students who need special attention. The

study collected all available users' usage data from moodle and applied DM techniques to discover hidden knowledge, where this knowledge can be used to improve the student's performance and identify a group of students who need special attention. In 2012, they conducted a study of graduate students' data of 15 years (1993-2007) (Abu Tair & El-Halees, 2012). This research was done to improve graduate students' performance and overcome the problem of low grades of graduate students.

Baradwaj and Pal (2011) applied the *decision tree* method using the *J48* algorithm for classification to extract knowledge to describe students' performance in the end-semester examination. The findings suggested that the Previous Semester Marks (PSM) had the highest gain ratio than in a class test, seminar, assignments marks, general proficiency, attendance, and lab work. The knowledge extracted by the decision tree was represented in the form of IF-THEN rules. This study used the data set of 50 students from the session 2007 to 2010.

In the same year, Kumar and Vijayalakshmi (2011) proposed two algorithms: *ID3* and *C4.5* (J48) for classification to predict the student's performance in the final exam based on the marks obtained in the first semester. For this purpose, the *C4.5* decision tree was implemented. The research outcome was the prediction of the number of students who were likely to pass or fail. This research also performed a comparative analysis of *C4.5* and *ID3* based on the accuracy comparing the result of the tree with the original marks obtained and the time taken to derive the tree. C4.5 was found to be more efficient than ID3.

In 2013, a study was conducted to predict student's performance in university results on the basis of their performance in the Unit test, assignment, graduation, percentage, and attendance (Borkar & Rajeswari, 2013). Several methods, like *Association rule mining, Apriori algorithm,* and *Correlation coefficient,* were implemented. The finding suggested that: *To get good university performance, the student must be good at their assignment, attendance,* and *unit test*. To support the research evidence, analysis of generated association rules and correlation coefficients values were carried out. However, the evidence of the research was not as strong as the correlation coefficient values between different attributes were not identical to the associations obtained from the Apriori algorithm. Despite that, the evidence presented was well connected to the claims as the rules generated by association rule mining and correlation coefficient

values showed that the different attributes of students were dependent, which impact students' university performance.

Ratnapala et al. (2014) used EDM techniques to conduct a quantitative analysis of student's interaction with an e-learning system through instructor-led, non-graded, and graded courses. The finding suggested that the learning environment differentiation can change the student's online access behavior. The majority of the student population were not self-motivated to do self-learning. Lack of interest and motivation to carry online learning, and the main course study side by side was found. However, the datasets were not large enough, and the reason to use k-means clustering was also not well justified in this research.

Yukselturk et al. (2014) predicted dropout through data mining approaches in an online program. In this research, *3- Nearest Neighbor (NN)* and *Decision Tree (DT)* were found to be more sensitive. Though 3-NN and DT were said to be more sensitive, it does not clarify in what ways and also accuracy can be questioned since the dataset was not large enough.

Kadiyala and Potluri (2014) used the *k-means clustering* technique and *decision tree* technique for the analysis of students' academic performance. This study collected data of 200 students from their exam results and applied k means clustering method to group the students into three categories (i.e., low, medium, and high) based on students' performance in percentage. The result of clustering showed a low-performance group having a percentage less than 60, medium performance group having a percentage greater or equal to 60 and less than 85, and a high-performance group having a percentage greater than or equal to 85. The research also applied a decision tree to classify the patterns of students' performance in order to obtain specific knowledge to improve both the educational system and learners' performance.

In the same year, the research was conducted using an *EM algorithm* for clustering, which showed the groups of students with a similar characteristic of performance, and the *J48 classifier* was used for classification, which showed correctly classified instances with an accuracy of 96.6667 % (Prabha & Shanavas, 2014). This research explored the application areas of EDM in the OL system. The findings were: *adopting DM tools and techniques in academic institution helped in improving decision making, improve the services they provide, and increase the student grades and retention*. In

this research, although the MATHS TUTOR, an LMS environment for school students for 6th, 7th, and 8th grade, was designed and implemented in three schools, the dataset was collected only from 6th grade for analysis. Further, it only used the *EM* algorithm for clustering and *J48* for classification, which can be considered as the limitation of this research. A year later, Prabha and Shanavas (2015) conducted research to better understand how the students identify the settings in which they learn to improve education outcomes. Different methods such as *EDM Classification for Model Construction and Model usage*, *Prediction to develop Predictive model,* and *Clustering for Classification of clustering algorithm* techniques were used. The finding suggested that the classification of a students group according to their knowledge level with test marks will make easier for the teacher to concentrate the areas for weak students. To support the evidence, prediction model was developed with the use of classification algorithm (using "if-then rule"). Model construction using EDM and use of real datasets of 60 students from 6th grade logged into MATHS TUTOR were considered, which can be also taken as the limitation of this research.

In the same year, Kashyap and Chauhan (2015) conducted research focusing on the comparative analysis of various EDM techniques and Machine Learning (ML) algorithms. Different methods, such as *association*, *clustering,* and *classification,* were considered. The study showed that for classification, *Naïve Bayes classification* was the best algorithm in performance; for clustering, the *k-means clustering* algorithm was the best algorithm, and for the association, the *Apriori algorithm* was the best and more accurate as compared to other algorithms. As a continuation, Kashyap and Chauhan (2016) also conducted research on the comparative analysis of different ML techniques and compared the accuracy of different classification techniques. Decision Tree algorithms such as *C5.0* and *ID3* produced an accurate result for the classification of the structured educational dataset. To classify the unstructured educational dataset, *Support Vector Machine* (SVM), *Naïve Bayes Classification*, as well as *Neural Network* (NN) *Classification* produced accurate results in terms of several parameters such as speed and efficiency. Similarly, in bio-medical data analysis, decision tree algorithm *C5.0* provided a better result than the *C4.5* algorithm. Further, *Neural Network (NN) Classifier* produced a more accurate result than the *Decision Tree* and *Naïve Bayes Classifier* in terms of efficiency for the analysis of the Mammographic Mass dataset. In Bank Direct Marketing, *SVM Classifier* provided more accurate results in terms of

speed and efficiency as compared to other classifiers. Preethi and Goswami (2015) conducted research to study the students' performance using classification methods such as *Decision Tree (J48)* and *Bayesian (Naïve Bayes). The naïve Bayes* classifier had an accuracy of 74%, and the *J48* classifier had an accuracy of 73% in classifying instances. From the result of the *J48* prediction model, it suggested that the time difference between posts (in mins) greater than 3 had the highest number of predictions to obtain grade 'B'. The dataset of 100 students was collected from an online examination for this research.

In the same year, Kaur et al. (2015) conducted research on predicting and analyzing students' performance and identifying slow learners among students in academics. Different classification algorithms such as *Naïve Bayes, Multi-Layer Perception, SMO, J48,* and *REPTree* were considered for this research. Multi-Layer Perception algorithms were found to be the best classifier, with an accuracy of 75% than other classifiers. Aziz et al. (2015) applied the *Naïve Bayes classifier* to extract the hidden pattern of Students' Academic Performance (SAP) to identify the parameters that influence the students' academic success. The study showed that *Naïve Bayes* applying 3-fold cross-validation classified the instances with an accuracy of 57.4%. Among six different parameters, the *family income* had a high influence on SAP with 56.8% probability. Also, it showed that an average student category had a better classification with an accuracy of 68.5% than other categories such as poor and good.

Pratiyush and Manu (2016) applied one of the supervised learning algorithms called SVM for the classification task. The main goal of this research was to predict the students' placement results in a labeled class as Yes or No. The study collected data of 200 students with six independent attributes such as *Attendance, GPA, Reasoning Aptitude, Quantitative Aptitude, Communication Skills, Technical Skills,* and one dependent attribute (i.e., *Placement).* This study showed how the classification result of the students' placement gives a better perception of how a particular group of students should perform and what they should target on new educational trends to get placed in the future.

Liang et al. (2016) applied DM in the datasets of 39 courses (200,000 samples) from the Edx MOOC online learning platform for the prediction of dropout students enrolled in different courses. They developed four classification models *(SVM, Logistic*

Regression, Random Forest, and Gradient Boosting Decision Tree (GBDT)). Among these models, GBDT produced the highest accuracy of 88%.

Amrieh et al. (2016) applied different classifiers such as *Artificial Neural Network (ANN), Naïve Bayesian (NB),* and *Decision Tree (DT)*. This research was conducted to propose a new student's performance prediction model using a classification technique. The result showed that *learner's behaviour (features) and their educational achievement had a strong relationship* and one of the features (i.e., *visited resources)* was the most effective features. Using behavioral features, the accuracy of the prediction model achieved up to 22.1% improvement, while removing such features and using ensemble methods, the accuracy of the prediction model achieved up to 25.8% improvement. The accuracy of the prediction model was more than 80% through the testing and validation process. This study applied new data attributes/features called *students' behavioral features*, where these features related to students' interactivity in the OL system. The study also applied different ensemble methods such as *Bagging, Boosting,* and *Random Forest* to improve the accuracy of classifiers.

In the same year, Saa (2016) applied multiple classifier methods such as *C4.5*, *ID3, CART,* and *CHAID* to find a qualitative model to classify and predict the students' performance. The study showed the comparative analysis based on accuracy where *CART* had an accuracy of 40%, *CHAID,* and *C4.5* had an accuracy of 34.07% and 35.19%, *ID3* with the lowest accuracy of 33.33%, and *Naïve Bayes classifier* with an accuracy of 36.40%.

Nichat and Raut (2017) measured the student performance using two methods of classification techniques such as *Decision Tree Induction Algorithm* and *Decision Trees.* The finding suggests that the *C4.5* algorithm was more accurate and took less execution time than *ID3* with different data sizes and *early analysis of student's performance helped in time management*. The evidence provided in this research was enough for exploring the student's performance as *satisfactory* or *not satisfactory* and their weaknesses in a particular subject or field, which helped to predict the performance of the student activity.

El Moustamid et al. (2017) developed a system to analyse learners' profiles and indexing web videos. The system offered a rich database of courses with different levels to the learners. Different techniques like *classification, clustering,* and

association rules were applied. The research helped teachers to assess students to identify their gaps and find courses that match their levels without being lost in the large volume of videos available on the internet.

Almarabeh (2017) applied five classifiers: *Naïve Bayes*, *Bayesian Network*, *ID3*, *J48*, and *Neural Network* to predict and analyse students' performance in the university. The experimental result showed *Bayesian Network* as the best classifier with an accuracy of 92.0% than other classifiers such as *Naïve Bayes*, *J48*, *Neural Network*, and *ID3* with an accuracy of 91.11%, 91.11%, 90.2%, and 88.0%, respectively. For this study, students' data consisted of 225 instances and 10 attributes.

Al-Shehri et al. (2017) applied both *SVM* and *KNN* for the classification task to find the best prediction model based on their accuracy. So, the model developed was used to predict the students' grades. For this study, data of 375 students were collected. The experiment result showed that *SVM* achieved a slightly better result of 96% accuracy than *KNN* with 95% accuracy.

Kapur et al. (2017) used different classification algorithms such as *Decision tree (J48), Naïve Bayes, Random Forest, Naïve Bayes Multinominal, K-star,* and *IBk*. This research intended to study and compare all classification algorithms to find well-performing algorithms for students' final marks prediction. The experiment result showed that *Random Forest* had higher correctly classified instances with an accuracy of 76.666% than other methods. The dataset used in this study contained 480 entries of students with 16 attributes.

Costa et al. (2017) investigated the effectiveness of algorithms used for the early prediction of students who are likely to fail. Four prediction techniques, *SVM, Decision Tree via J48, Neural Network,* and *Naive Bayes,* were applied on the dataset of 424 undergraduate students. The study showed that the *SVM* technique was effective than the other three techniques with an efficiency of 92%.

Sarra et al. (2018) studied the usefulness of DM for determining students who are at higher risk of failure and more likely to drop out. For this purpose, they created a profile of students through *Bayesian Profile Regression (BPR)* on the basis of student's performance, motivation, and resilience with the data collected through an online questionnaire. The study suggested that *BPR* can be used for identifying students who

are at high risk of dropping out, and necessary steps could be taken by the instructor in hand.

Hussain et al. (2018) first applied features selection methods such as *correlation-based attribute evaluation, gain-ratio attribute evaluation, information-gain attribute evaluation, relief attribute evaluation,* and *symmetrical uncertainty attribute evaluation*. Then four different classification algorithms such as *J48, Random Forest, BayesNet,* and *PART* were implemented. The main aim of this research was to find highly influential attributes of students' academic performance and also compare four classification algorithms, such as *J48, Random Forest, BayesNet,* and *PART*. The experiment result showed that the *Random Forest Classification* method was the best-suited algorithm for the dataset with an accuracy of 99% (84.33% without selected features) than other classification methods such as *PART* (74.33%), *J48* (73%), and *BayesNet* (65.33%), with selected attributes. The datasets consist of 300 records with 24 attributes. Feature selection methods included 12 most influencing features.

Data Mining Application in Moodle Data

The data mining process involves successively processing raw data into more refined forms, enabling further processing of the data and the extraction of relevant information. It can be broken down into three core processes: *Data Preprocessing, Pattern Recognition, Interpreting Results* (Kamath, 2009). The datasets have many outliers or anomalies. Removal of these anomalies can help in the better prediction of student performance and behavior. Hence the study was conducted to detect the anomalies using the k-means clustering method.

The dataset used in this study was obtained from the computer students of Kathmandu University, Nepal, enrolled in the course COMP-341 (Human-Computer Interaction). There were two types of logs obtained from Moodle. The system log data from the Moodle had 14840 observations, including the column headers. There were nine columns for the active 128 users in the system. The columns present in the data were: *Time, User's full name, Affected user, Event context, Component, Event name, Description, Origin,* and *IP address*. The quiz grade dataset from Moodle had 105 observations, including the column headers. The dataset had 10 columns for each user altogether. The columns present in the data were: *First name, Surname, ID number, Institution Department, Email address, Assignment: Mini-Research Project Proposal*

"assignment" (Real), Quiz: quiz1 (Real), Quiz: COMP 341 MCQ (Real), and *Last downloaded from this course.* The data was prepared so that a minimal set of features were chosen for clustering the data. The total number of log count and the average grade of a student from Quiz 1 and 2 were taken as the most effective features. All kinds of a click in the moodle system were extracted from the log data of each student. After extracting the click count for different modules, the total number of click counts were calculated for each student (Grade count is out of 20).

Table 1

Click Counts for All Modules for Each Student

Assignment.Click	Chat.Click	File.Click	Forum.Click	System.Click	Url.Click	Wiki.Click	Grade
0	0	1	1	12	4	0	15.5
1	0	22	3	46	4	2	20
4	1	4	6	30	4	3	19
0	0	2	4	12	2	0	18.5

The data were then narrowed down to “Total.Click” and “Grade” for each student.

Table 2

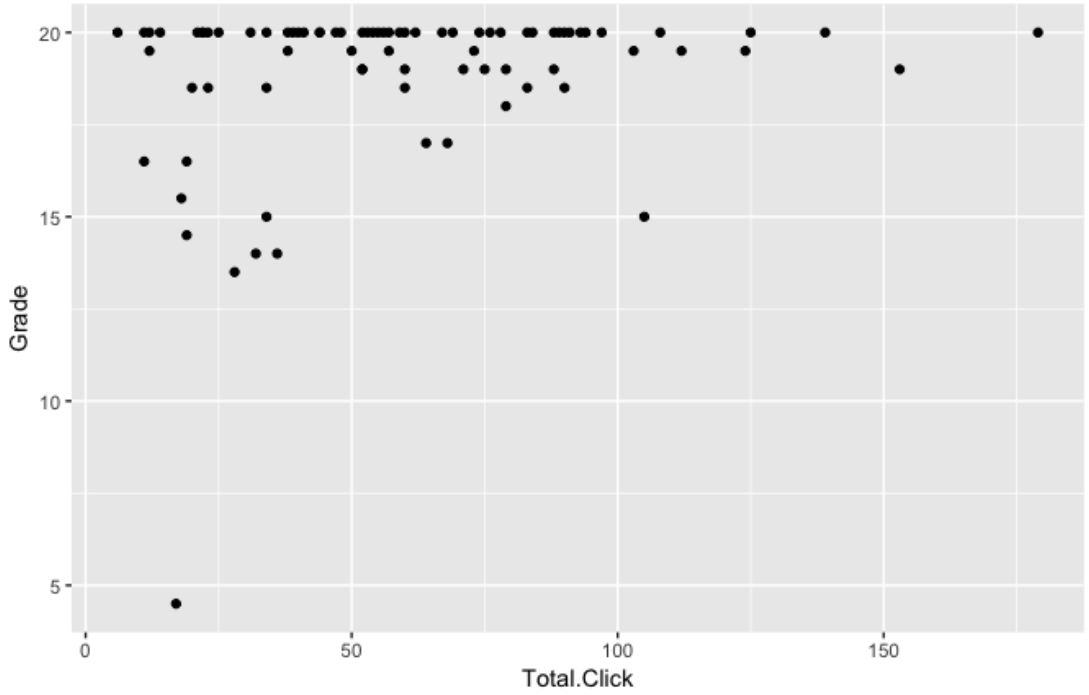
Total Click Counts and the Average Grade of Students

Total.Click	Grade
18	15.5
78	20
125	20
73	19.5

After pre-processing the data, a scatter plot of *Grade vs. Total.Click* was generated to study the relationship between these features. Figure 3 showed that most students, despite having low clicks performed above average, i.e., greater than 18.89. The students with the number of clicks above 100 always scored more than the average grade. Thus, the scatter plot shows a direct relationship between *Total.Click* and *Grade*. A higher number of interactions in the Moodle system resulted in a higher quiz score than normal.

Figure 2

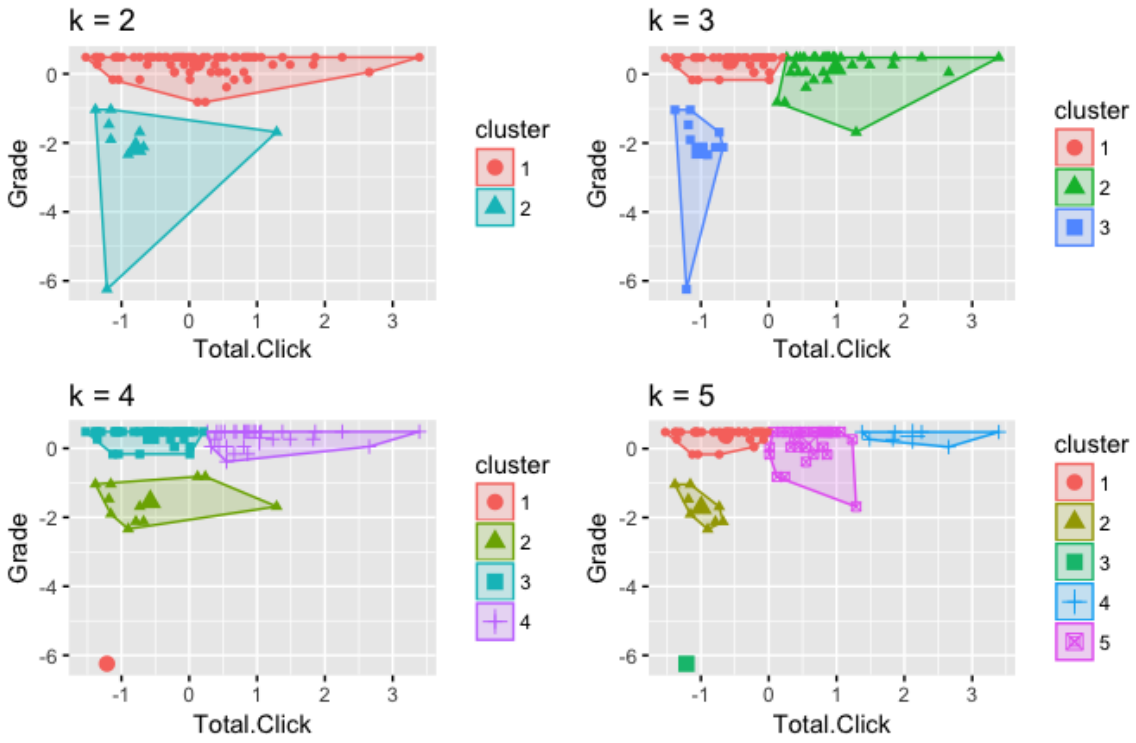
Grade vs. Total.Click Scatter Plot



Clustering is finding a group of objects such that the objects in one group will be like one another and different from the objects in another group. In EDM, clustering can be used to distinguish between student activities and their behaviors. In this study, students are clustered into groups according to their activity and exam scores. K-means clustering is one of the simplest and most commonly used clustering methods for splitting datasets into k groups. It is a popular clustering mechanism based on the distance between objects. The *'k'* in k-means is the number of clusters that k-means should generate. The dataset needs to be standardised (i.e., scaled) to make variables comparable. After standardization, the data is fed to the k-means algorithm. The cluster size can be increased by one each time to see if a cluster outside of the normal group of clusters was formed.

Figure 3

Comparison of K-Means Clusters for k=2,3,4 and 5



The elbow method and average silhouette method were used to find out the optimal number of clusters. Determining optimal clusters requires an optimal value of '*k*'. The elbow method defines clusters, such as the total intra-cluster variation (total within-cluster sum of squares) is minimised. On the other hand, the average silhouette approach measures the quality of clustering, i.e., it determines how well each object lies within its cluster. A high average silhouette indicates good clustering (Air Force Institute of Technology, n.d., Average Silhouette Method section, para. 1).

Figure 4

Elbow Method for Determining the Optimal Number of Clusters

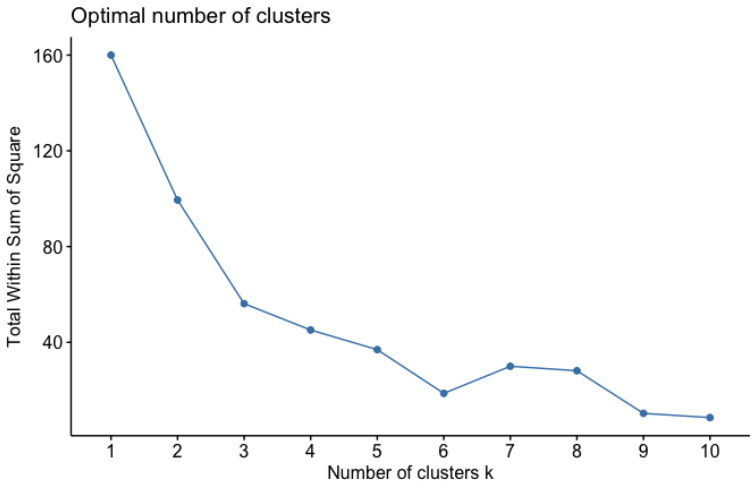


Figure 4 suggests that 3 is the optimal number of clusters as it appears to bend in the knee (or elbow).

Figure 5

Average Silhouette Method for Determining the Optimal Number of Clusters

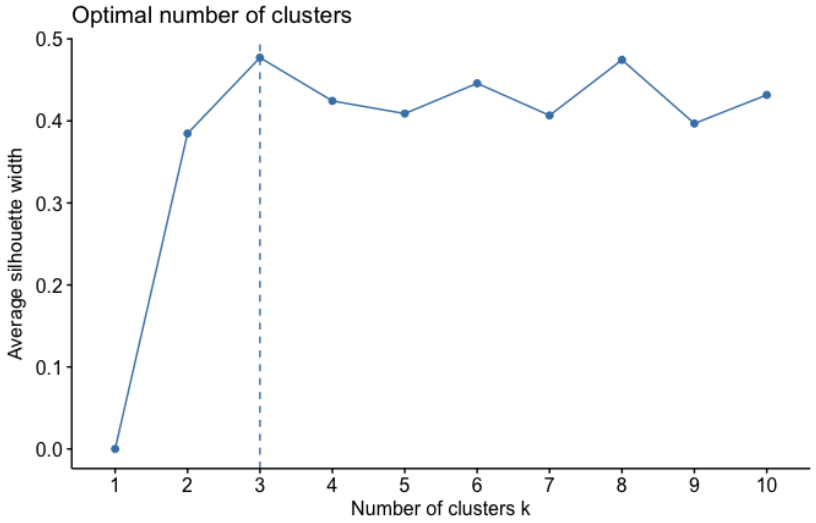
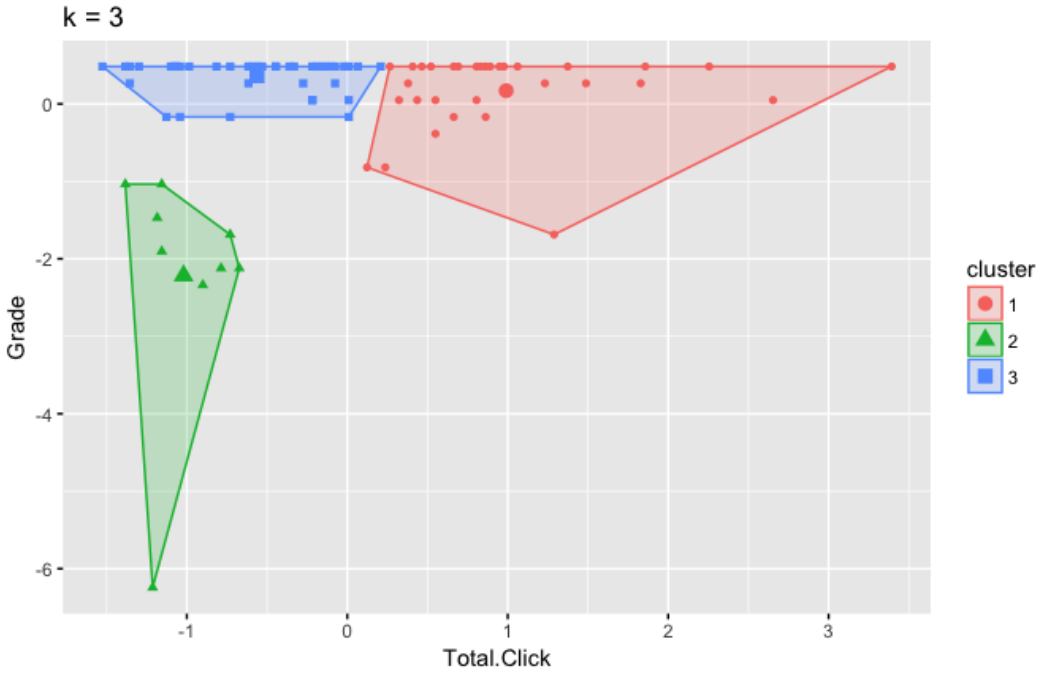


Figure 5 also suggests 3 as the optimal number of clusters.

As both approaches suggested 3 as the optimal number of clusters, the final analysis and clusters are extracted using *k=3*.

Figure 6

K-Means Clustering with k=3



The k-means algorithm split the group into 3 groups:

i. **Cluster 1:** Students with a high number of total clicks and high grades.
ii. **Cluster 2:** Students with a low number of total clicks and below-average grades.
iii. **Cluster 3:** Students with a moderate number of total clicks with high grades.

In the next step, the distance between the objects and cluster centers are calculated, and three largest distances from the result are identified as outliers. Table 3 shows that *cluster 1* has two outliers and *cluster 2* has one outlier.

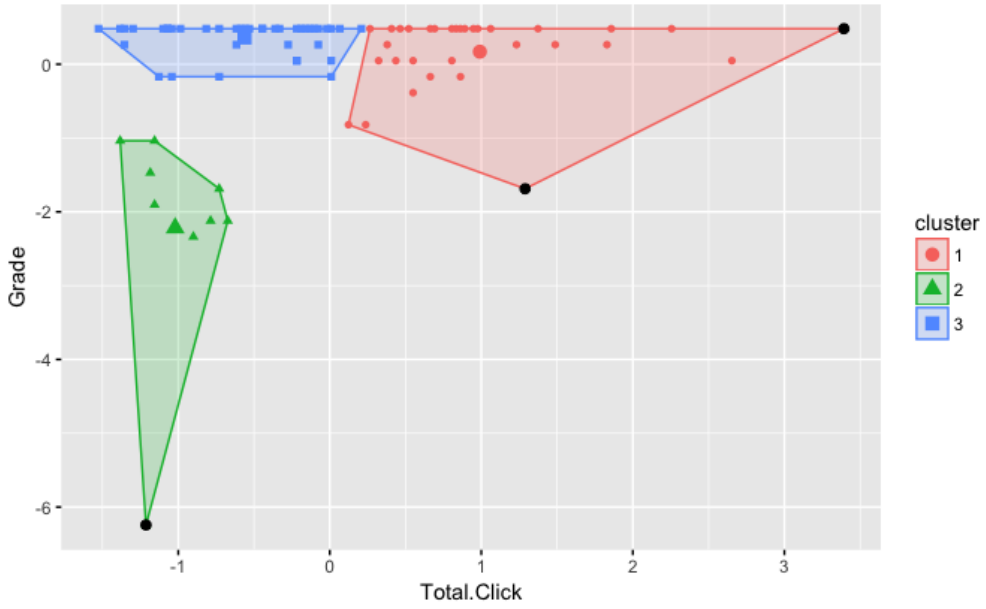
Table 3

Outlier Data Points From K-Means Clustering

Total.Click	Grade	Cluster	Distance From Cluster Center
105	15.0	Cluster 1	30
179	20.0	Cluster 1	46
17	4.5	Cluster 2	53

Figure 7

K-Means Cluster Plot With Outliers



Discussion and Conclusion

Clustering of quiz and log data of 128 students enrolled in the course COMP 341resulted in some interesting results. The main reason to use clustering was to group data and to identify outliers. The clustering method *k-means* was used to find the outliers. Out of 128 data points, 3 data points were flagged as outliers by this clustering technique. The result shows that 3 students did not conform to the normal distribution

of the data. In terms of clustering, these data points were far off from the clusters that were formed and had a large intra-cluster distance, among other data points within the cluster.

Educational Data Mining is one of the emerging and promising area in the field of Information Technology (IT). Due to the growth of OL users, EDM is growing its popularity as it is a suitable approach to identify the learner's behaviors and build the predictive model. This paper includes several pieces of research done in the field of EDM. The paper summarises different DM algorithms used in various types of educational data. It also consists of the study done on the moodle data of the course COMP 341. The study implements the k-means clustering method to find the anomalies in the dataset.

The emerging research on EDM has led to the researchers on the integration of educational theories and developing an environment for the notification of information, also called Learning Analytics. Liu et al. (2017), in the paper on a data-driven personalization system, have mentioned that there was a little focus on pedagogical and pastoral contents of learning. Formulation or improvement of DM algorithms can also be made where LA can be very useful. Learning Analytics focuses on how these algorithms can be deployed and integrated into learning designs. It can provide visible improvements for students (Liu et al., 2017). Simon (2017) mentioned that LA could be data-driven, and the learners' log data can be studied to optimise learning. The author had further stated the major goals of LA, such as *it helps students find more personalised ways to learn, help teachers to improve students learning,* and *help the student in their learning by giving a better environment*. This paper also described different learning theories, such as *behaviourisms, cognitivism, constructivism,* and *social constructivism,* and highlighted that the integration of learning theories could *provide a conceptual framework, allows to interpret what is observed*, and *also provides a solution to the problem which occur during learning*. Wong et al. (2019) reasoned that a good understanding of *how learning occurs, how learning can be supported*, and *how student characteristics influence learning* is needed if the goal of LA is to understand and optimise learning. Based on a recent review of papers published in the Review of Educational Research (RER) journal over the last century, Murphy and Knight (2016) found that learning sciences have been guided by three predominant theoretical lenses: behavioral, cognitive, and contextual (Wong et al.,

2019). Wong et al. (2019) further explained that one learning theory might be more suitable for understanding learning in one environment than another. The author also highlighted that researchers were missing theories to explain social factors that influence learning in groups.

ORCiD

Sushil Shrestha https://orcid.org/0000-0003-2225-2578
Manish Pokharel https://orcid.org/0000-0003-1357-0531

References

Abu Tair, M. M., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: A case study. *International Journal of Information and Communication Technology Research, 2*(2), 140-146.

Air Force Institute of Technology. (n.d.). *K-means cluster analysis*. https://afit-r.github.io/kmeans_clustering

Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. *International Journal of Modern Education and Computer Science*, *9*(8), 9. https://doi.org/10.5815/ijmecs.2017.08.02

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., & Olatunji, S. O. (2017). Student performance prediction using support vector machine and k-nearest neighbor. *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 1-4. https://doi.org/10.1109/CCECE.2017.7946847

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, *9*(8), 119-136. http://doi.org/10.14257/ijdta.2016.9.8.13

Anoopkumar, M., & Rahman, A. M. (2015). A comprehensive survey on educational data mining and use of data mining techniques for improving teaching and predicting student performance. *Advances in Innovative Engineering and Technologies*, 55-84.

Aziz, A. A., Ismail, N. H., Ahmad, F., & Hassan, H. (2015). A framework for students'academic performance analysis using naïve bayes classifier. *Jurnal Teknologi, 75*(3), 13-19. https://doi.org/10.11113/jt.v75.5037

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61-75). Springer. https://doi.org/10.15215/aupress/9781771991490.01

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications, 2*(6), 63-69. https://arxiv.org/ftp/arxiv/papers/1201/1201.3417.pdf

Borkar, S., & Rajeswari, K. (2013). Predicting students' academic performance using education data mining. *International Journal of Computer Science and Mobile Computing, 2*(7), 273-279.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & 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*, *73*, 247-256. https://doi.org/10.1016/j.chb.2017.01.047

Data Mining. (2019). Retrieved 20 January 2019 from https://behavior.lbl.gov/?q=node/11

De Morais, A. M., Araujo, J. M., & Costa, E. B. (2014, October). Monitoring student performance using data clustering and predictive modelling. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, 1-8. https://doi.org/10.1109/FIE.2014.7044401

El-Halees, A. M. (2009). *Mining student's data to analyze e-Learning behavior: A case study*. https://bit.ly/2SEPsZ7

El Moustamid, A., En-Naimi, E., & El Bouhdidi, J. (2017, March). Integration of data mining techniques in e-learning systems: Clustering Profil of Lerners and Recommender Course System. *BDCA'17: Proceedings of the 2nd international Conference on Big Data, Cloud and Applications*, Art. 97. https://doi.org/10.1145/3090354.3090453

Fauvel, S., & Yu, H. (2016). *A survey on artificial intelligence and data mining for MOOCs.* https://arxiv.org/ftp/arxiv/papers/1601/1601.06862.pdf

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science, 9*(2), 447-459. DOI: 10.11591/ijeecs

Kadiyala, S., & Potluri, C. S. (2014). Analyzing the student's academic performance by using clustering methods in data mining. *International Journal of Scientific & Engineering Research*, *5*(6), 198-202. https://bit.ly/38mpwKO

Kamath, C. (2009). *Scientific data mining: A practical perspective.* Society for Industrial and Applied Mathematics. https://bit.ly/38xPY4j

Kapur, B., Ahluwalia, N., & Sathyaraj, R. (2017). Comparative study on marks prediction using data mining and classification algorithms. *International Journal of Advanced Research in Computer Science, 8*(3), 632-636. https://doi.org/10.26483/ijarcs.v8i3.3066

Kashyap, G., & Chauhan, E. (2015). Review on educational data mining techniques. *International Journal of Advanced Technology in Engineering and Science, 3*(11), 308–316.

Kashyap, G., & Chauhan, E. (2016). Parametric Comparisons of Classification Techniques in Data Mining Applications. *International Journal of Engineering Development and Research*, *4*(2), 1117-1123.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science, 57*, 500-508. https://doi.org/10.1016/j.procs.2015.07.372

Kumar, S. A., & Vijayalakshmi, M. N. (2011). Implication of classification techniques in predicting student's recital. *International Journal of Data Mining & Knowledge Management Process*, 1(5), 41-51. doi: https://doi.org/10.5121/ijdkp.2011.1504

Liang, J., Yang, J., Wu, Y., Li, C., & Zheng, L. (2016, April). Big data application in education: Dropout prediction in edx MOOCs. *2016 IEEE Second International Conference on Multimedia Big Data*, 440-443. https://doi.org/10.1109/BigMM.2016.70

Liu, D. Y. T., Bartimote-Aufflick, K., Pardo, A., & Bridgeman, A. J. (2017). Data-driven personalization of student learning support in higher education. In *Learning analytics: Fundaments, applications, and trends* (pp. 143-169). Springer. https://doi.org/10.1007/978-3-319-52977-6_5

Maimon, O., & Rokach, L. (2005). Decomposition methodology for knowledge discovery and data mining. In *Data mining and knowledge discovery handbook* (pp. 981-1003). Springer. https://doi.org/10.1007/0-387-25465-X_46

Murphy, P. K., & Knight, S. L. (2016). Exploring a century of advancements in the science of learning. *Review of Research in Education, 40*(1), 402-456. https://doi.org/10.3102/0091732X16677020

Nichat, A. A., & Raut, D. A. B. (2017). Predicting and analysis of student performance using decision tree technique. *International Journal of Innovative Research in Computer and Communication Engineering, 5*(4), 7319-7327.

Prabha, S. L., & Shanavas, A. M. (2014). Educational data mining applications. *Operations Research and Applications: An International Journal (ORAJ), 1*(1), 23-29.

Prabha, S. L., & Shanavas, A. M. (2015). Application of educational data mining techniques in e-Learning - A case study. *International Journal of Computer Science and Information Technologies, 6*(5), 4440-4443.

Pratiyush, G., & Manu, S. (2016). Classifying educational data using support vector machines: A supervised data mining technique. *Indian Journal of Science and Technology, 9*(34), 1-5.

Preethi, N., & Goswami, D. (2015). A review on role of data mining techniques in enhancing educational data to analyze student's performance. *International Journal of Computer Science and Information Technology Research, 3*(1), 123-129.

Rana, S., & Garg, R. (2016). Evaluation of student's performance of an institute using clustering algorithms. *International Journal of Applied Engineering Research, 11*(5), 3605-3609.

Ratnapala, I. P., Ragel, R. G., & Deegalla, S. (2014, December). Students' behavioral analysis in an online learning environment using data mining. *7th International Conference on Information and Automation for Sustainability,* 1-7. https://doi.org/10.1109/ICIAFS.2014.7069609

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems With Applications, 33*(1), 135-146. https://doi.org/10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601-618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12-27. https://doi.org/10.1002/widm.1075

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384. https://doi.org/10.1016/j.compedu.2007.05.016

Saa, A. A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications, 7*(5), 212-220.

Sarra, A., Fontanella, L., & Di Zio, S. (2019). Identifying students at risk of academic failure within the educational data mining framework. *Social Indicators Research, 146*(1-2), 41-60. https://doi.org/10.1007/s11205-018-1901-8

Siemens, G., & Baker, R. S. D. (2012). Learning analytics and educational data mining: Towards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 252-254. https://doi.org/10.1145/2330601.2330661

Simon, J. (2017). A priori knowledge in learning analytics. In *Learning analytics: Fundaments, applications, and trends* (pp. 199-227). Springer. doi: https://doi.org/10.1007/978-3-319-52977-6_7

Vasani, V. P., & Gawali, R. D. (2014). Classification and performance evaluation using data mining algorithms. *International Journal of Innovative Research in Science, Engineering and Technology*, *3*(3), 10453-10458. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1085.2475&rep=rep1&type=pdf

Wong, J., Baars, M., de Koning, B. B., van der Zee, T., Davis, D., Khalil, M., & Paas, F. (2019). Educational theories and learning analytics: From data to knowledge. In *Utilizing learning analytics to support study success* (pp. 3-25). Springer. https://doi.org/10.1007/978-3-319-64792-0_1

Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and E-Learning*, *17*(1), 118-133. https://doi.org/10.2478/eurodl-2014-0008

To cite this article:
Shreshta, S., & Pokharel, M. (2020). Data mining applications used in education sector. *Journal of Education and Research, 10*(2), 27-51. https://doi.org/10.3126/jer.v10i2.32721