Comparative Analysis of the Learning on KDD Cup 2015 Dataset

S. Nithya*
Research Scholar, Department of Computer Science, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, India.
E-mail: nithyaresearch19@gmail.com

Dr.S. Umarani
Professor, Department of Computer Science, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, India.
E-mail: ravania@gmail.com

Received August 14, 2021; Accepted November 30, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19050

Abstract

Massive open online courses (MOOCs) have delivered a high level of education around the world, but the significant dropout rate has impacted their educational efficiency. In MOOC, the researchers mainly focused on dropout prediction using various approaches. Our work emphasizes the real-time dataset of KDD CUP 2015 to extract certain main features, implemented by using various methods, and find out the different performance measures with different metrics. To measure early learner dropout, this work used a set of unique features extracted from real-time datasets. Individual user contributions and performance are frequently analyzed using the learning analytics framework. It also aids in the decision-making process for students in higher education. According to the data, the suggested model has a classification accuracy of 78 to 85%. The experimental results are also predicted, which one of the learning methods is mostly applicable to the dropout prediction in MOOC. In this evaluation, the performance measures are compared and the best one is reported.

Keywords

Classification, Learning Analytics, Performance Measure, Educational Datamining, Dropout prediction, Machine Learning.

Introduction

The rise of digital technology and advances in multimedia devices have resulted in a huge increase in the global number of active users, as well as the number of users' ways of learning have changed dramatically.

Massive open online courses (MOOCs) have exploded in popularity in recent years, attracting millions of participants. Since 2012, educators all over the world have favored and actively promoted MOOC as an extension of E-Learning. Coursera, Udacity, and edX should be the most well-known and relevant MOOC providers.

Educational data, which is a byproduct of learner-instructor interaction, has evolved into a multidisciplinary area of investigation involving researchers from a variety of fields.

Significant advances in technology-enhanced learning tools have led to a dramatic increase in virtual learning data, resulting in a slew of educational repositories with direct implications for higher education institutions. MOOC dropouts can be studied to help reduce dropout rates and increase MOOC value.

Learning analytics (LA) and Educational Data Mining (EDM) are two methods that use data-driven techniques to solve issues such as dropout prediction, and they consist of five steps: data collection, reporting, prediction, action, and refinement. The increased focus on retention in the MOOC community was due to the higher number of dropouts. Machine learning techniques have been used in a number of analyses to solve this issue.

For this paper, we're using data from the 2015 KDD Cup to model MOOC dropouts. The user activity log data, enrollment, time, source, and object make up the majority of the data in the input. This research used a variety of supervised machine learning methods to predict MOOC dropout.

The rest of the paper is organized as follows: In section 2, some previous MOOC dropout prediction research. Present the dataset definition and proposed prediction method in section 3. In section 4, experimental results on the KDD Cup 2015 dataset are presented. Then, come to a conclusion in section 5.

Literature Review

Online learning strategies have been the subject of modern educational paradigm. The significant rise of online courses on many platforms allows people to expand their expertise. The majority of learnings look at the dropout rate and retention rate of registered users. Analyze the intrinsic component, implement the various models, and assess the influence on performance measures.

Many recent researches in the literature have looked at the KDD Cup 2015 dataset to see how machine learning may be used to analyze and predict student performance.

L. Qui et al. (2018) present an integrated system for predicting MOOC dropouts with feature selection that encompasses feature development and extraction. To overcome the challenge of online dropout detection, the ensemble feature extraction (FSPred) method was used. FSPred can automatically extract and discover valid attributes from MOOC users' web usage data to improve forecast performance and reduce computational complexity. They suggest just one statistical analyses of student learning, and they utilized XuetangX's learning behavior log dataset. The experimental analyses lead to Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) models.

The student engaging in the courses will leave some click records of learning activities in MOOC. Each click record is a time-stamped log that records the student's various learning behavior information. Based on the students' learning behavior attributes, the proposed Student Dropout Prediction (SDP) model (C. Jin, 2020) can forecast a student's dropout status for a single MOOC course on a weekly basis. The Improved Quantum Particle Swarm Optimization (IQPSO) algorithm was used to determine three Support Vector Regression (SVR) parameters. It is an effective regression model that has been widely used in many prediction applications as an extension of SVM. The proposed SDP model's performance, accuracy (Acc), F1, and area under curve (AUC) are employed as estimation criteria. Test results on both observational and statistical data were used to develop the proposed SDP models.

The criteria considered include students' exercise routines, message boards, multimedia, and website links. The data preprocessing technique is used to eliminate input variables with no value or the same value for all rows. For the first time, the author (M. Şahin,2020) an "Adaptive Neuro-Fuzzy Inference System (ANFIS) is used to forecast dropout rates in MOOCs". The suggested method makes use of both neural networks and fuzzy inference systems, resulting in extremely accurate predictions. Such input factors have no bearing on the accuracy of the prediction findings. By normalizing the data from all inputs with the min-max scaling function, the valuation of every attribute is transformed from zero to 1. Fivefold cross-validation is used to pick the benchmarked machine learning methods. MATLAB 2018 is used to carry out the 22 distinct techniques. Several machine learning approaches, such as Decision Trees (DT), LR, SVM, Ensemble Learning, and K-Nearest Neighbor (K-NN) methods, were used to create the various models.

The Convolutional Neural Network (CNN) model, combined with a feature vector for behavioral correlation of learning, was proposed by Y. Wen et al. (2020), to predict dropout students' learning. This model can be used to predict earlier failures once enough data has been collected. In dropout prediction, classification techniques like Naive Bayes

(NB), Linear Discriminant Analysis (LDA), LR, SVM, RF, and Gradient Boosted Decision Tree (GDBT) are frequently used. Standard metrics and tenfold cross validation are used to assess the outcomes.

Y. Zheng et al. (2020) create a two-dimensional CNN-based dropout prediction model that integrates Feature Weighting and Time Series (FWTS-CNN). Tutors had a better understanding of learners who relied on certain responses while using online courses. Using a sample of 39 courses obtained by Xuetang X, the model was compared to a basic classification algorithm with an accuracy of 87.1 percent.

Table 1 Comparative Analysis of the Assessment of the KDD CUP 2015 dataset

S. No	Paper Details	Parameters	Objectives	Techniques applied	Evaluation	Outperformed
1	Prediction of Students' Dropout in MOOC Environment. (R. Umer, 2017).	Age, education background and motivation	To forecast early dropout & completion rate.	NB RF, LR, and K-NN.	Accuracy (Acc), F1 measure	Logistic Regression. **Course E:** Accuracy --0.89.
2	Modeling MOOC Dropouts. (D. Peng & G. Aggarwal, 2015).	Course and Module Information, Event Log, Enrollment, Completion	MOOC dropouts using user activity data.	LR, SVM, GBDT, RF, and AdaBoost.	Weka tool: Acc, Recall, Precision F1-score, ROC, AUC.	GBDT: Accuracy -0.877
3	Understanding Dropouts in MOOCs. (W. Feng, J. Tang, & T. X. Liu, 2019).	Video, Forum, Assignment, web page clicking	To identify student dropouts.	CNN: Adam -ReLu Activation function.	AUC, F1 – Score.	CFIN: Context-aware Feature Interaction Network.
4	MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine. (J. Chen et al., 2019).	Enrollment, Object, Log, Date, Truth (x1-x23)	Learning Behavior: ELM algorithm for dropout prediction.	LR, SVM, Back Propagation neural network (BP), DT, LSTM, and ELM.	Accuracy, AUC, and F1-score.	Hybrid alg-DT-ELM
5	Predicting MOOC Dropout Based on Learner's Activity Features. (S. Ardchir et al., 2019).	Enrollment, Object, Log, Date, Truth (x1-x23)	Dropout prediction using activity log.	SVM, Scikit-learn python package	Accuracy	RBF Kernel: Accuracy - 88.9
6	Power of Attention in MOOC Dropout Prediction. (S. Yin, L. Lei, H. Wang, & W. Chen,2020).	4 Fields: a)Time b)Source c)Event d)Object	NLP Used with Viterbi Algorithm	Linear Regression, SVM, LR, CNN-LSTM model.	Accuracy	Sequence Problem.
7	Improving Prediction of MOOCs Student Dropout Using a Feature Engineering Approach. (S. Ardchir et al., 2020).	1)object.csv: 2)enrollment_train.csv 3) true_train.csv 4)log_train.csv	Analyzing behavior 1) All dataset features; 2) Only high weighted features	1) LR 2) AdaBoost 3) GBDT 4) RF	Accuracy, ROC curve.	GBDT yields a better score
8	Dropout Prediction in MOOCs Using Behavior Features and Multi-view Semi-supervised Learning. (W. Li et al., 2016).	Videos viewing behavior	Semi-Supervised Learning model	Multi training Algorithm, 1) LR 2) NB 3) SVM 4) DT	Accuracy, Precision, Recall, F-measure.	F-measure : 95%
9	Discovering Learning Behavior Patterns to Predict Dropout in MOOC. (B. Hong et al., 2017).	13 features	Dropout prediction using ML & cascading ML methods.	1) RF, 2) SVM, 3) Multinomial Logistic Regression (MLR).	Precision, Recall, F1-score, Accuracy, AUC.	C-RF : Accuracy – 88%
10	Student dropout prediction in massive open online courses by CNN. (L. Qiu et al., 2019).	Registration log, Time, Source, Event, Object.	Clickstream data of students learning behaviors.	1) LR, 2) NB, 3) DT, 4) SVM, 5) GBDT, 6) RF, 7) AdaBoost, 8) DP-CNN =>ReLu	Precision, Recall, F1 score, and AUC score.	Small dataset - LR and big dataset - DP-CNN

The evaluation measures of learner performance using various machine learning approaches implemented in the KDD CUP 2015 dataset are shown in *Table1*. The focus of this article is to leverage a range of user participation characteristics to predict learner performance.

Materials and Methods

1. Dataset Depiction

The dataset, which includes 39 courses and 120542 enrolled users from the KDD CUP 2015(KDD CUP 2015 Dataset), demonstrates how to forecast dropouts in online courses. The proposed architecture for dropout prediction is shown in *Figure 1.*

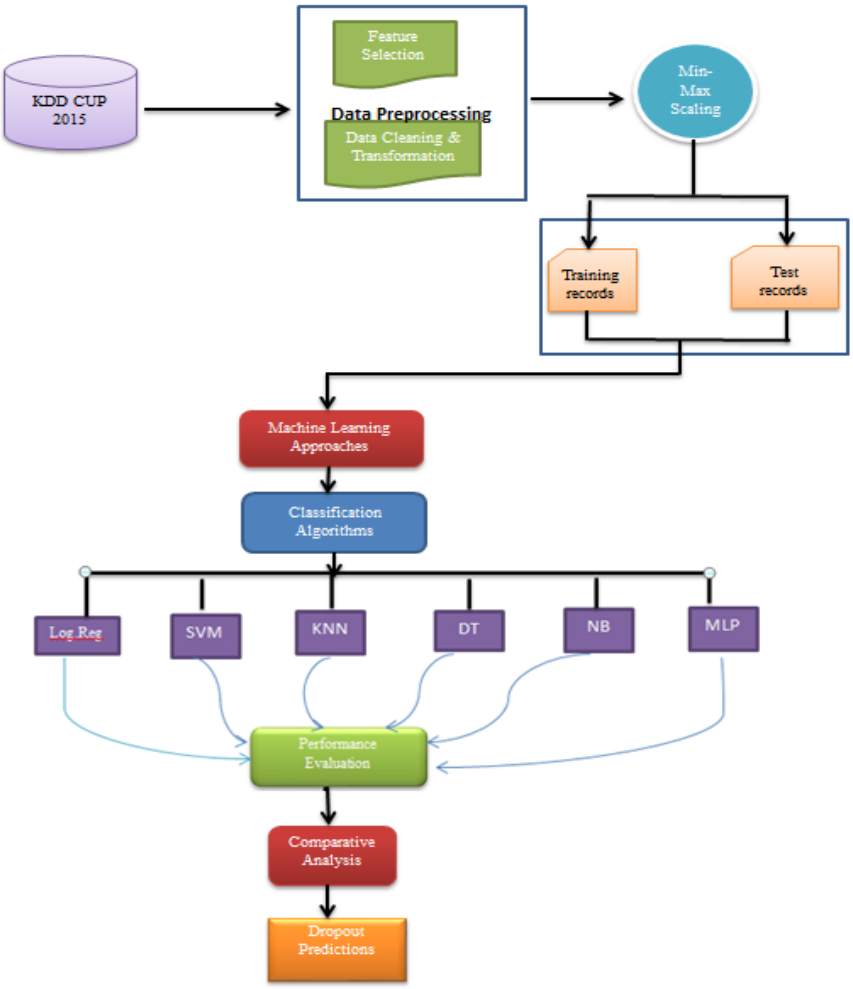

Figure 1. Proposed Architecture of the Dropout Prediction

The description of **Figure1** as a KDD CUP 2015 dataset has 54 features, according to the website. It then proceeds on to the pre-processing stage, where it selects features using

various techniques, cleans the data, and transforms it into a certain format. After pre-processing the data, use min-max scaling to standardize the dataset values. Then training and testing sets are separated. Within this set, we implemented six classification algorithms like Logistic regression, SVM, KNN, DT, NB, and MLP using machine learning techniques. The performance evaluations are compared and the best one is displayed depending on the accuracy value.

In general, *Table.2* uses a true public dataset that takes into account the course's start and end dates, the number of enrolled users, information, course module, and source of an event, as well as the learners' dropout.

Table 2 The module and statistics of the dataset

Information	**Indication**
Enrolment	Registered learners (120,542 entries).
Object	Program and module relationships.
Log	Keeping track of one's behavior.
Date	Start and finish durations for the program.
Truth	Completion or dropout rate of learners.

This dataset has 120542 enrolled users, and the descriptive analysis yields 8,157,277 different learning behaviour items.

2. Feature Selection

Most analysis methods have been altered to allow researchers to define the quality of their abilities, which is a major problem in dropout prediction. As a result, the most important qualities of the source data are picked and used as inputs to create a unique and equivalent report.

A feature selector, like the Variance threshold, is used to exclude all low-variance features from this dataset. The Chi2 score is used in feature selection to choose positive features like frequencies or booleans that are heavily dependent on the outcomes. The 18 features are chosen as input and one class label is chosen as output as a result of this process.

3. Data Preprocessing

On this dataset, data cleaning and transformation activities must be performed. Missing values in an input variable are removed, as are the same values for all rows. To change the value ranges between 0 and 1, the data is normalized using the min-max scalar function.

Table 3 The characteristics of the dataset used for the analysis

Features	Description
enrollment_id	Unique Id of an enrolled user.
source	Link to the event (server or browser).
event	We established seven different event types
	Pbm - Practicing academic work
	Video - Listening training sessions
	Access - Accessing the other course materials.
	Wiki - Surf data via the web.
	Discussion - Sharing data on the message boards
	Navigate - Browsing through different areas of the program.
	Page_close – Exit the website
Output	Course completion (value 1) or dropout (value 0) of a user.

The dataset contains 72143 tuples after preprocessing, and the model should be formed using nineteen of the features listed in *Table 3*.

4. Methods

The most often used approaches for constructing models in the classification process are given here.

A. LR

In classification problems, the logistic method is frequently employed to find the prediction. The binomial logistic model is used to calculate the likelihood of a binary classification using one or more independent features (W. Lu et al., 2017). For input X, It is dependent on the logistic function also with the following weight (W) and bias (b) parameters:

$$f(x) = e^{w \bullet x} + b \,/\, 1 + e^{w \bullet x} + b \qquad (1)$$

In MOOCs, logistic regression with variables extracted manually is commonly used to forecast if users might stop learning.

B. SVM

The Support Vector Machine is the most commonly utilized in classification and regression analysis (SVM). It's a well-known supervised learning approach for analyzing data and detecting patterns. The distance between the acceptable and undesirable data nodes, as well as the hyperplanes, is measured by SVM. Kloft et al. employ SVM to forecast dropouts using attributes extracted from the weekly timeline of learner

clickstream data (M. Kloft, F. Stiehler, Z. Zheng, & N. Pinkwart, 2015). In this experiment, we will employ SVM with a linear kernel.

C. DT

For decision-making, the tree is constructed in a hierarchical structure. “The decision tree tests the associated feature qualities of the item to be classified, starting at the root node and continuing until it reaches the leaf node” (Y. Zheng et al., 2020), where output is then executed. The information gain and Gini index can all be used as the basis for optimal attribute partitioning. The Gini index is used as the foundation for division, with the number 2 denoting the tree's depth.

D. KNN

The KNN method uses a set of labeled records to compute the distance between them using various distance metrics, and then retrieves them using the value of k, which denotes the number of nearest neighbors. Academics utilize the CFA ratings for exam outcomes to forecast dropouts (C.G. Brinton & M. Chiang, 2015).

E. NB

The Naive Bayes technique is a statistical classification approach based on the Bayes rule and the independent constraint of characteristic conditions. “Calculate the posterior probability using the class prior probability and the predictor prior probability” (J. Singh, S. Bagga, & R. Kaur, 2020). It is used to determine the sample's conditional probability for each class and to choose the class with the highest conditional probability as the sample's class. This study employed the Naive Bayes-Gaussian classifier to identify whether students withdrew from E-learning.

F. Ensemble method-RF

The method of ensemble machine learning, such a classic example, is the Random Forest. This algorithm can manage missing data in the classification process. To increase the quality of the classification optimization method (J. Singh, S. Bagga, & R. Kaur, 2020), it introduces a randomly chosen feature into the decision tree training process because the random forest generates a Bagging ensemble with a decision tree as the baseline. This classification model can also be used to evaluate categorical values.

G. MLP-NN

In this work, a feed-forward multilayer perceptron was used. The input neurons allocate all predictor variables. The yield of input neurons is employed in the hidden layer, as well

as the response variables that expound the output layer. Implementing the activation function as an identity function in the output layer (K. Coussement et al., 2020) directly results in a reduced dropout rate.

Experiment & Results

In this section, the planned work outputs were generated by feature selection approaches and a variety of classifiers. The scikit-learn python package was used in this work, with a cross-fold validation value of 5. The proposed system's primary performance measures are precision, recall, F1-score, and accuracy. Our measurements can be viewed in the following way:

i. Accuracy – Calculated by combining expected and actual dropout rates.
ii. Precision (P) – the ratio of accurately classified dropout occurrences
iii. Recall (R) – The proportion of accurately classified dropouts vs. the total number of dropouts.
iv. F1- Measure – A metric that combines P and R values to describe the suggested classifier's system results.
v. AUC – A scalar value that represents the classifier's system results.

Table 4 shows the test results for the various algorithms used in the experiment. In the comparison, the noted terms have the highest value.

Table 4 Compare the results with different classifiers

S. No	Metrics / Learning Method	Accuracy	Precision	Recall	F1-score
1	***LR***	0.78	0.78	1	0.88
2	***SVM(Lin)***	**0.85**	**0.86**	**0.97**	**0.91**
3	***DT***	0.78	0.78	1	0.88
4	***RF***	0.78	0.78	1	0.88
5	***KNN***	0.79	0.795	0.99	0.88
6	***NB***	0.78	0.78	1	0.9
7	***MLP-NN***	0.78	0.78	1	0.88

For a dataset with independent variables, the logistic regression method computes the outcome in binary form, i.e. 1/0, Yes/No. It seeks to fit the likelihood of an event into a logit function in order to predict it.

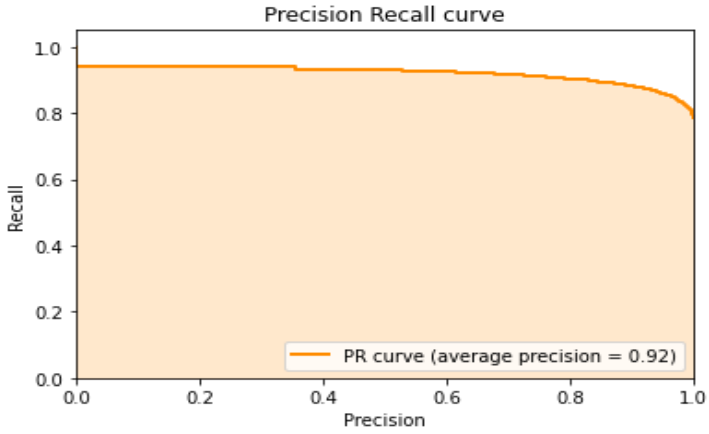

Figure 2 LR Precision-Recall Curve

LR has an average accuracy of 0.92%, as seen in *Figure 2*. When analyzing the variance between the precision and recall values (0 to 1) of this model, the precision values are present in the x-direction and the recall values are in the y-direction.The initial value of precision is 0.0, and the recall value starts at approximately 0.95. If the precision value is increased, the recall only slight difference in their ranges. At the point of 1.0 in the x-direction, the y value reached 0.8. Therefore, the average precision of this graph is calculated at 0.92.

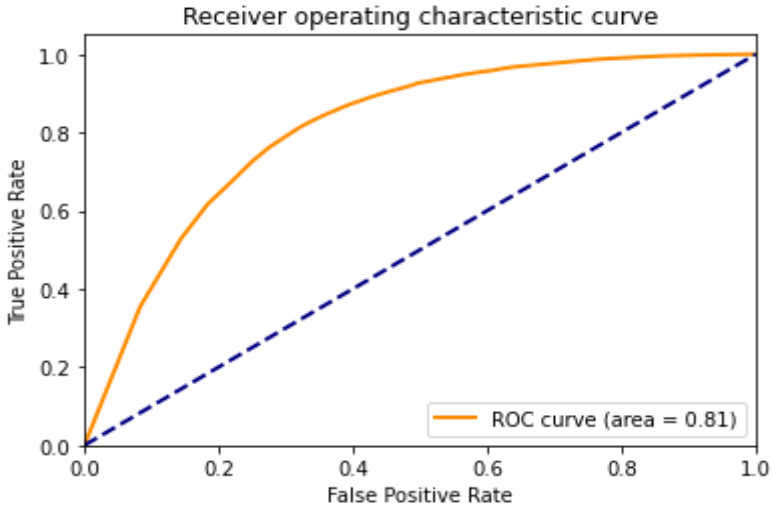

Figure 3 ROC of logistic regression

The proportion of true-positive and false-positive rates is represented by the ROC curve (0.0 to 1.0). A false-positive rate and a true-positive rate are proposed by the direction of x and y. The curve originally increased the false positive rate while also increasing the true positive rate, but after reaching a point of 0.81 in Logistic Regression, the true positive rate remained constant, as seen in *Figure 3.*

Although the precision of the KNN classifier is substantially higher than that of the other classifiers, the best precision is achieved by SVM. The value of category loss should be reduced using the Adam optimizer and the learning rate is 0.1.

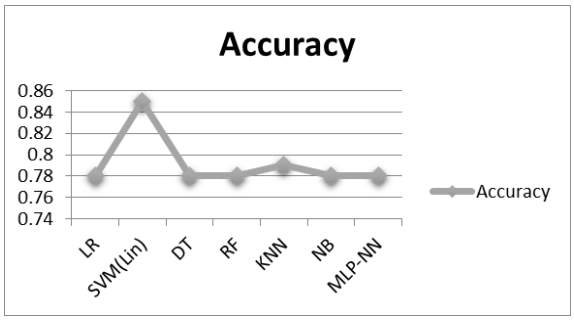

Figure 4 The outcomes of using various classifiers

In this analysis, we compared the various models with their accuracy range of 78-86%. We generated a graph between the learning method and the accuracy values based on their findings. The learning methods are LR, SVM, DT, RF, KNN, NB, and MLP-NN, and the accuracy values are 78, 85, 78, 78, 79, 78, and 78. SVM-Linear Kernel, which has an accuracy of 85 percent and is shown in *Figure 4,* is the best algorithm.

Conclusion & Future Direction

In this work was intended to predict the students' dropout rate in the real-time dataset of KDD CUP 2015. It was implemented by using various machine learning methods and finding out the different performance measures. A sample of 39 programs gathered from XuetangX University is used to evaluate the performance measures in the proposed model. Using a variance threshold, a chi-square approach, and other classification algorithms, we extract certain key features. We use numerous machine learning algorithms such as LR, SVM, DT, RF, K-NN, NB, and MLP in an experimental procedure. The performance measures were compared and analyzed in order to determine whether the SVM with the accuracy rate (85%) was the best. In the future, it will look at ways to increase the dropout forecast effect in MOOCs. This is achieved by collecting data from a variety of sources, including assignment submission and performance data, message boards and interaction data, and then widening the analysis and usage of Artificial Neural Networks in MOOCs.

References

Qiu, L., Liu, Y., & Liu, Y. (2018). An integrated framework with feature selection for dropout prediction in massive open online courses. *IEEE Access, 6,* 71474-71484. http://doi.org/10.1109/ACCESS.2018.2881275.

Jin, C. (2020). MOOC student dropout prediction model based on learning behavior features and parameter optimization. *Interactive Learning Environments,* 1-19. http://doi.org/10.1080/10494820.2020.1802300

Şahin, M. (2021). A comparative analysis of dropout prediction in massive open online courses. *Arabian Journal for Science and Engineering, 46*(2), 1845-1861. http://doi.org/10.1007/s13369-020-05127-9.

Wen, Y., Tian, Y., Wen, B., Zhou, Q., Cai, G., & Liu, S. (2019). Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs. *Tsinghua Science and Technology, 25*(3), 336-347. http://doi.org/10.26599/TST.2019.9010013

Zheng, Y., Gao, Z., Wang, Y., & Fu, Q. (2020). MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series. *IEEE Access, 8,* 225324-225335. http://doi.org/10.1109/ACCESS.2020.3045157

Umer, R., Susnjak, T., Mathrani, A., & Suriadi, S. (2017). Prediction of students' dropout in MOOC environment. *International Journal of Knowledge Engineering, 3*(2), 43-47. http://doi.org/10.18178/ijke.2017.3.2.085

Peng, D., & Aggarwal, G. (2015). Modeling mooc dropouts. *Entropy, 10*(114), 1-5.

Feng, W., Tang, J., & Liu, T.X. (2019). Understanding dropouts in MOOCs. *In Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 517-524. http://doi.org/10.1609/aaai.v33i01.3301517

Chen, J., Feng, J., Sun, X., Wu, N., Yang, Z., & Chen, S. (2019). MOOC dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine. *Mathematical Problems in Engineering.* http://doi.org/10.1155/2019/8404653

Ardchir, S., Talhaoui, M.A., Jihal, H., & Azzouazi, M. (2018). Predicting MOOC Dropout Based on Learner's Activity. *International Journal of Engineering & Technology, 7*(4.32), 124-126.

Yin, S., Lei, L., Wang, H., & Chen, W. (2020). Power of Attention in MOOC Dropout Prediction. *IEEE Access, 8,* 202993-203002. http://doi.org/10.1109/ACCESS.2020.3035687

Ardchir, S., Ouassit, Y., Ounacer, S., Jihal, H., Goumari, M.Y.E., & Azouazi, M. (2019). Improving Prediction of MOOCs Student Dropout Using a Feature Engineering Approach. In International Conference on Advanced Intelligent Systems for Sustainable Development, 146-156. http://doi.org/10.1007/978-3-030-36653-7_15.

Li, W., Gao, M., Li, H., Xiong, Q., Wen, J., & Wu, Z. (2016). Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning. *In international joint conference on neural networks (IJCNN), 3*130-3137. http://doi.org/10.1109/IJCNN.2016.7727598.

Hong, B., Wei, Z., & Yang, Y. (2017). Discovering learning behavior patterns to predict dropout in MOOC. *In 12th International Conference on Computer Science and Education (ICCSE),* 700-704. http://doi.org/10.1109/ICCSE.2017.8085583

Qiu, L., Liu, Y., Hu, Q., & Liu, Y. (2019). Student dropout prediction in massive open online courses by convolutional neural networks. *Soft Computing, 23*(20), 10287-10301. http://doi.org/10.1007/s00500-018-3581-3

"KDD CUP 2015 Dataset." https://data-mining.philippe-fournier-viger.com/the-kddcup-2015-dataset-download-link/

Lu, W., Wang, T., Jiao, M., Zhang, X., Wang, S., Du, X., & Chen, H. (2017). Predicting student examinee rate in massive open online courses. *In International Conference on Database Systems for Advanced Applications*, 340-351. http://doi.org/10.1007/978-3-319-55705-2_27

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. *In Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs*, 60-65. http://doi.org/10.3115/v1/w14-4111.

Brinton, C.G., & Chiang, M. (2015). MOOC performance prediction via clickstream data and social learning networks. *In IEEE conference on computer communications (INFOCOM)*, 2299-2307. http://doi.org/10.1109/INFOCOM.2015.7218617

Singh, J., Bagga, S., & Kaur, R. (2020). Software-based prediction of liver disease with feature selection and classification techniques. *Procedia Computer Science, 167*, 1970-1980. http://doi.org/10.1016/j.procs.2020.03.226

Coussement, K., Phan, M., De Caigny, A., Benoit, D.F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*, 135. http://doi.org/10.1016/j.dss.2020.113325