

Received 27 July 2023, accepted 31 August 2023, date of publication 5 September 2023, date of current version 12 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3312150

RESEARCH ARTICLE

Deep FM-Based Predictive Model for Student Dropout in Online Classes

NUHA MOHAMMED ALRUWAIS 

Department of Computer Science and Engineering, College of Applied Studies and Community Services, King Saud University, Riyadh 11495, Saudi Arabia

e-mail: nalrowais@ksu.edu.sa

This work was supported by the Researchers Supporting Project, King Saud University, Riyadh, Saudi Arabia, under Grant RSPD2023R608.

ABSTRACT The student's high dropout rate is a severe issue in online learning courses. As a result, it is creating concerns for academics and administrators in the field of education. A practical method of preventing dropouts is predicting students' likelihood of dropping out. This study uses an explainable factorization machine and deep-learning approach to predict students' dropouts with two datasets, namely HarvardX Person-Course Academic Year 2013 De-Identified and MOOC datasets. With the solvable approach, the aim is to enable the interpretation of the predictive models to produce actionable insights for related online educational interventions. This approach creates a DeepFM-based prediction model for student dropout, which entails multiple processes, including data preparation, feature engineering, model construction, training, assessment, and deployment. Moreover, the DeepFM design combines a factorization machine with DNN models to forecast student dropouts. It examines performance metrics, including recall, F1 score, accuracy, precision, and AUC-ROC. After ten iterations and 64 batches, the DeepFM model accurately predicted student dropout from online courses with a 99% accuracy rate on validation data. It also outperformed other techniques because of its capacity to capture complicated non-linear connections between features, combine dense and sparse information, and consider the unique properties of online learning. This study illustrated using an explainable factorization machine learning and DNN approach called DeepFM to interpret the underlying reasons for predicting students' dropout from online classes. Moreover, this approach has the potential to be extended to additional Massively open online courses (MOOC) datasets to assist educators and institutions in identifying at-risk students and providing targeted interventions to enhance their learning results.


INDEX TERMS Student dropout, online class, DeepFM model, deep-learning, deep-neural networks, machine-learning.

I. INTRODUCTION

Online learning has grown significantly in popularity in recent years and now provides students all around the world with flexible learning choices [1]. The high student dropout rate is one of the difficulties online educational systems confront. Student dropout impacts individual students, educational institutions, and online course providers, who are severely concerned [2]. However, the lesser binding force of MOOCs compared to the conventional classroom has led to a waste of educational resources as many students have left

their courses due to internal or external circumstances [3]. Researchers have concentrated on the study of MOOC learners' dropout behavior prediction to lessen the occurrence of this phenomenon, with the hope of accurately identifying those students who are at risk of dropping out and taking proactive steps to help them continue learning to increase the course completion rate [4]. Therefore, in the realm of MOOC big data analytics [5] and educational data mining research [6], forecasting MOOC learners' dropout likelihood based on learning behavior has become a hot issue [7].

Academics have proposed several models based on empirical research to explain the factors contributing to online learners' loss and attempt to reduce the drop-out rate by

The associate editor coordinating the review of this manuscript and approving it for publication was Nkaepe Olaniyi .

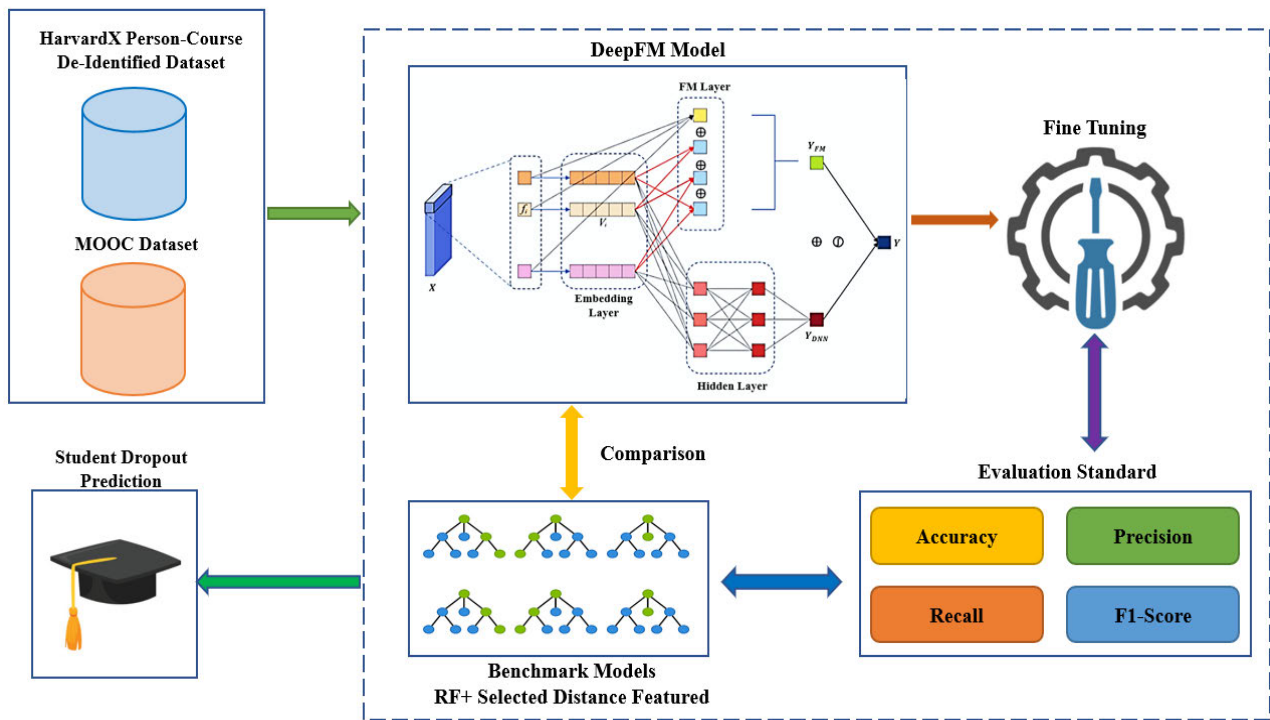


FIGURE 1. Framework of predictive model for student dropout in online class.

mitigating negative characteristics while improving positive features overall [8]. However, because there are significant individual variances among learners, macro-level development initiatives frequently fail owing to a lack of specificity. Understanding the numerous variables connected to dropping out is the foundation for lowering dropout rates [9]. The key to reducing dropout rates is to use these indicators to identify students who are likely to drop out and to put focused retention initiatives in place before students start showing drop-out signs. In this study, an ML approach is employed to create prediction models that are then trained with fresh data using data from the information systems of online education institutions [10], [11]. After training, the produced prediction model samples might be used to forecast future dropout behavior. Using this strategy, online education institutions may identify potential dropouts early and implement retention strategies before the dropout behavior manifests itself. It will help in lowering the dropout rate [12].

It is essential to comprehend the causes of student dropout in online courses to develop efficient intervention techniques and raise retention rates. Both conventional statistical models and machine learning strategies have been used to deal with this issue. Researchers have been looking into the issue of MOOC students quitting school. To create prediction models, several studies employed conventional classification techniques, including logical regression (LR) [13], [14], [15], KNN [14], and SVM [15], [16], [17]. Based on course data, [16] extracted 19 features from the click stream, homework test, and forum behavior perspectives. They then

built a sliding window model in conjunction with a machine learning algorithm to dynamically predict the dropout rate of students. To predict students' dropout behavior in the upcoming weeks, [15] extracted 19 student behavior characteristics from click stream data that could be represented as single real numbers. They then applied extensive logical regression and linear SVM methods and came to the conclusion that including forum data can significantly increase prediction accuracy [14]. To build a Gradient Boosting Decision Tree model to predict the likelihood of students quitting school in the next 10 days, [17] enrollment feature. Other researchers predicted dropout behavior using neural network techniques like CNN [6], [18] and LSTM [19], [20]. These methods, however, frequently need to adequately capture the intricate linkages and non-linear patterns seen in educational data.

Figure 1 depicts the structure of a prediction model for student dropout in online classrooms, which uses two distinct datasets: the HarvardX Person Course dataset and the MOOC dataset. Based on the information provided at enrolment, the dropout forecast was computed. In addition to making predictions, a machine learning model is interpreted using the DeepFM model's DNN and cutting-edge interpretable machine learning approaches. The model is then adjusted, and after passing the assessment criteria, it is finished. Then, the proposed model is contrasted with currently used techniques like RF and a selected distance feature. Finally, a prediction model for student dropout in online classrooms is developed.

This paper suggests a novel DeepFM-based prediction model for online course dropout. DeepFM stands for Deep Factorization Machine, which combines the advantages of deep neural networks and factorization machines. This hybrid model enables more precise predictions of student dropout by incorporating both linear and non-linear connections between features. The primary goal of this project is to develop a system for early intervention that can recognize students who are in danger of dropping out and offer prompt assistance to stop their disengagement. The proposed model attempts to improve the precision and dependability of dropout predictions in online education by utilizing deep learning and factorization techniques.

The DeepFM model, which has been developed exclusively for predicting student dropout in online courses, is the main contribution of this work. Factorization machines, which are excellent at capturing feature interactions, and deep neural networks, which can learn complicated representations from high-dimensional data, are combined to create DeepFM. This innovative model architecture offers a more thorough and precise method for dropout prediction. It shows a thorough feature engineering pipeline that includes feature selection, feature encoding, feature scaling, and data preparation. Thanks to this process, the input features are adequately cleaned and optimized for the DeepFM model. To enhance the model's prediction capability, feature engineering is done carefully for the dataset to extract the most pertinent facts.

Some of the significant contributions of this study are:

1. The DeepFM model's performance is thoroughly assessed using the HarvardX Person-Course Academic Year 2013 De-Identified Dataset. Accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic (AUC-ROC) curve are only a few measures used in the evaluation.
2. The proposed model's advantage is shown in terms of prediction accuracy and efficacy in identifying students at risk of dropping out by comparing its performance to baseline models.
3. A fresh insight into the elements influencing student dropout in online programs is shown in addition to model performance.
4. Important variables are selected and examined, such as course content, level, length, workload, and student demographics, using feature importance analysis. These insights can assist educational institutions and course providers develop focused interventions to raise student retention rates by helping them comprehend the underlying mechanisms of student dropout.

The paper is organized as follows: Section II discusses some of the related works, and the description of the dataset and the preparation are introduced gradually in Section III. The primary properties of the dataset under investigation are also described in this section. The techniques for forecasting student dropout are covered in Section IV, which uses several deep-learning classifiers. However, the findings and the models' evaluation using several performance indicators

are covered in Section V. While Section VI contains the discussion, Section VII covers the conclusion, followed by references.

II. LITERATURE REVIEW

An in-depth analysis of dropout prediction in online learning environments has already been conducted. Student dropout has been predicted using a variety of machine learning approaches, including logistic regression, decision trees, and support vector machines, based on variables including course performance, engagement, and demographic data. However, these methods frequently fail to adequately capture educational data's intricate linkages and non-linear patterns.

Recent advancements in machine learning have made deep understanding the most advanced method for predicting dropouts [21]. Feature representations that are not linear are automatically utilized by Jiao's deep and fully linked feed-forward neural network [10]. Similar reasoning may be used by [11], who employed a recurrent neural network model with LSTM cells that stored attributes in adjacent states. While DNN requires iterative training and a sizable amount of training data, DL needs to be more precise than standard ML techniques. Additionally, because MOOC platforms are designed differently, current research uses a variety of learning patterns to predict dropouts [22]. In an online learning environment, a lack of a uniform description and comprehension of learning behaviors might lead to unintended inferences about behavior features with a higher potential for classification. Due to differences in learning behavior, the range of findings varies greatly. Dropout prediction relies heavily on feature selection [23]. However, only some of the commonly used features are dedicated to it. There is a need to utilize the scalability-related features as well. One of the of-ten-used scalable feature selection approaches is DeepFM, which combines well with deep and machine learning algorithms but requires the repetition of training [24], [25].

Factorization machines have become a potent tool for working with high-dimensional, sparse data and provide accurate feature interaction modeling. These models are appropriate for dropout prediction problems because they can capture both linear and non-linear correlations between features. The adaptability and efficacy of factorization machines have been effectively employed in recommender systems [26], click-through rate prediction, and personalized medicine.

Due to its capacity to automatically generate hierarchical representations from unstructured data, deep learning approaches, particularly deep neural networks, have demonstrated promising outcomes in a number of disciplines [27]. Deep learning models have been used to extract significant aspects from educational data, such as sequential learning patterns, temporal dynamics, and complex interactions, in the context of dropout prediction [28]. However, the majority of current deep learning models for dropout prediction ignore

the potential advantages of factorization methods in favor of recurrent neural networks (RNNs) or long short-term memory (LSTM) networks. In their study, an innovative two-phase ensemble-based strategy for forecasting students' grades in MOOCs was put out by [10]. Their method, which combined a Random Forest algorithm with specific distance characteristics, had a high accuracy of 97%. The study emphasized the value of taking distance-based aspects into account and gave a more in-depth understanding of the variables affecting student performance.

Based on input predictors (lectures, quizzes, labs, and videos) taken from Moodle records, [29] developed Random Forest models that predict student achievement with 96.3% accuracy. The laboratory and questionnaire results have the most substantial impact on the final grade, according to their analysis of the predictors' dependency on the target value.

Based on a weekly study, several researchers develop dropout prediction models [30], [31]. For categorization purposes, [30] employed logistic regression models. This article includes prediction performance and a tutoring action plan showing a 14% decrease in dropout rates. Building a prediction model utilizing data from the previous several weeks to determine if a student would drop out in the upcoming week is how [32] investigate Deep Learning approaches.

The goal of [33] was to develop a Random Forest model to predict student dropout from self-paced MOOC courses. Their program proved its capability to recognize pupils who are in danger of dropping out with an accuracy of 87%. The study emphasized the need of factoring in many aspects of student involvement, academic advancement, and demographic data when predicting dropouts.

The hybrid model DeepFM, which was put out in this work, has been effectively used in a number of fields, although its use in educational settings for dropout prediction is rather restricted. This study examines how well DeepFM predicts student dropout from online courses in an effort to close this gap. DeepFM offers the potential to increase the precision and interpretability of dropout prediction models in educational situations by merging factorization techniques and deep learning. The work by [10] concentrated on utilizing a Random Forest algorithm to analyze and forecast MOOC learners' dropout behavior. Their method identified students who were in danger of dropping out with a 91% accuracy rate. The study shed light on the root reasons for dropout behavior and emphasized the necessity of early intervention techniques to raise retention rates for students. In another work [34], different algorithms have different accuracies: For Neural Networks, the accuracy is around 0.94; for RF: The accuracy is 0.93; for SVM: The accuracy is around 0.93. Techniques applied are Machine and Deep Learning, SVM, RF, Decision Tree, Confusion Matrix, and Cross-Validation Approach. In [21], the initial dataset consists of gender, age, mother tongue, current

employment status, level of English language skills, number of study hours, and MOOC attendance. The module consists of two to five learning units; LightGBM is the highest-performing model, F1-score, and accuracy ranges from 96% to 93%, respectively. This study's [35] primary goal is to find the best modeling approach to discover dropout student predictors using 17,430 student data points from a private institution. The applied technique is Machine Learning, Ensemble Classifier Model, KNN, DT, NB, and Confusion Matrix. The results are that the decision tree achieved the highest accuracy of 98.90%. Then, NB attained an accuracy of 98.2%, followed by the KNN, which had 98.1% accuracy. In another work [9], first-year students were enrolled in five subjects: Health Sciences, Engineering, Law, Social Sciences, Arts and Humanities. Three-thousand and four hundred students were enrolled, out of which five hundred and forty were dropped out. The outcome is the dropout percentage is around 16% per the total enrollees and dropouts.

The authors of the study [36] utilized ML algorithms to anticipate student dropout in a MOOC for smart city professionals early on in order to solve the issue of poor completion rates. Based on data from the first week of the course, the findings reveal great accuracy, enabling efficient intervention and assistance.

The well-researched issue of MOOC dropout prediction was the main subject of another study [37]. The objective was to create models to categorize students according to whether they were more likely to complete a course. The majority of current research in this field, according to the author, uses student engagement data to build prediction models. Although these models are effective at predicting dropout rates, they cannot adequately explain why a student is likely to drop out. The author highlighted the interpretation of dropout predictions at the student and model levels as a crucial expansion topic. It entails not just comprehending the forecasts but also having knowledge of the underlying causes of those predictions.

The proposed research in this study aims to fill many knowledge gaps in the area of online learning environment dropout prediction. The insufficient capture of complex patterns in educational data by conventional machine learning techniques like logistic regression, decision trees, and support vector machines is one notable flaw. These methods have trouble capturing complex relationships and non-linear patterns, which could have an effect on the precision and potency of dropout prediction models. The report also emphasizes current developments in deep learning as a more sophisticated strategy for dropout prediction. However, it points out that several deep learning models for dropout prediction now in use tend to concentrate on RNNs or LSTM networks, ignoring the potential advantages of factorization techniques. Another difficulty is the variation between different MOOC sites. Each platform has a unique design, which causes variances in students' learning habits and behaviors. The work

TABLE 1. Past references with datasets, techniques/ methodology, and results.

Ref	Dataset	Technique/ Methodology	Results
[34]	<ul style="list-style-type: none"> - The dataset contains around two hundred and sixty student records for training and testing datasets with approximately ten parameters. - Datasets consist of exams, test grades, projects and assignments, results, graduate grades, and passing years. 	<ul style="list-style-type: none"> - Machine and Deep Learning, SVM, RF, Decision Tree, Confusion Matrix, and Cross-Validation Approach. 	<ul style="list-style-type: none"> - Different algorithms have different accuracies: For Neural Networks, the accuracy is around 0.94; for RF: The accuracy is 0.93; for SVM: The accuracy is around 0.93.
[21]	<ul style="list-style-type: none"> - The initial dataset consists of gender, age, mother tongue, current employment status, level of English language skills, number of study hours, and MOOC attendance. - The module consists of two to five learning units. 	<ul style="list-style-type: none"> - Machine Learning, AdaBoost Algorithm, LightGBM, GBM, Logistic Regression, DNN, Linear SVM, RF, RF, Ensemble Method. 	<ul style="list-style-type: none"> - LightGBM is the highest-performing model. - F1-score and accuracy range from 96% and 93%, respectively.
[35]	<ul style="list-style-type: none"> - This study's primary goal is to find the best modeling approach to discover dropout student predictors using 17,430 student data points from a private institution. - Three classifiers are used to calculate the students' dropout prediction. 	<ul style="list-style-type: none"> - Machine Learning, Ensemble Classifier Model, KNN, DT, NB, Confusion Matrix. 	<ul style="list-style-type: none"> - The decision tree achieved the highest accuracy of 98.90%. - Then, NB achieved an accuracy of 98.2%, followed by the KNN, which had 98.1% accuracy.
[9]	<ul style="list-style-type: none"> - In this experiment, first-year students were enrolled in five subjects: Health Sciences, Engineering, Law, Social Sciences, Arts and Humanities. - Three-thousand and four hundred students were enrolled, out of which five hundred and forty were dropped out. 	<ul style="list-style-type: none"> - Machine Learning Method, Feature Selection, ANN, SVM, KNN, Decision Tree, Logistics Regression. 	<ul style="list-style-type: none"> - The dropout percentage is around 16% per the total enrollees and dropouts.
[38]	<ul style="list-style-type: none"> - The experiment consists of around eight thousand and five hundred students. 	<ul style="list-style-type: none"> - Machine Learning, Neural Networks, - Gradient-boosted trees, CatBoost, XGBoost, Linear Discriminant Analysis, SHAP plot. 	<ul style="list-style-type: none"> - The first student has a dropout risk of around 0.3 (estimated probability). - The average of students taking the courses has a dropout probability of around 0.7.

emphasizes the need for a more uniform description and understanding of learning patterns across platforms since this diversity might affect how well dropout prediction algorithms function.

Furthermore, current dropout prediction algorithms mainly rely on feature selection yet frequently ignore crucial scalability properties. According to the study, scalable feature selection methods, such as DeepFM, can be useful for enhancing predictive performance. Another significant research void is the interpretability of dropout forecasts. Although many models can accurately forecast dropout rates, they are difficult to interpret, leaving educators and administrators in the dark about the underlying causes of a student's propensity to drop out. To gather useful information for individualized educational interventions, the study underlines the significance of evaluating dropout predictions at both the student and model levels. Finally, it is acknowledged that the hybrid model DeepFM has only been used sparingly to predict dropouts in educational contexts. Although DeepFM has demonstrated success in a number of areas, its use for dropout prediction in online learning settings has not been thoroughly investigated.

By filling these knowledge gaps, the study hopes to enhance dropout prediction models and help educators and institutions spot at-risk students, resulting in more successful interventions and better learning outcomes for online learners. A detailed description of the proposed work is presented in the subsequent sections.

III. DATA COLLECTION

The HarvardX Person-Course Academic Year 2013 De-Identified Dataset (HMPC) [3] includes data on student enrolment and performance in HarvardX courses made available via the edX platform during 2013. The dataset contains de-identified data on specific students, such as age, gender, and place of residence, as well as data at the course level, such as course topics, start and end dates, and grades. According to Table 2, this dataset helps analyze how learners behave in virtual classrooms and create forecasting models for outcomes like course completion and dropout rates.

A. DATA DESCRIPTION

The HarvardX Person-Course Academic Year 2013 De-Identified Dataset (HMPC) is a dataset that includes details on student enrolment and performance in HarvardX courses made available via the edX platform during the academic year 2013. The dataset, accessible in CSV format, has 14 columns and 641138 rows.

- **course_id**: the unique identifier for each course.
- **userid_DI**: de-identified user identifier.
- **Registered**: the date when the student registered for the course.
- **Viewed**: the date the student first viewed the course on the system.
- **Explored**: the date the student first explored the course, such as by watching a video or attempting an assessment.

TABLE 2. Harvardx person-course academic year 2013 De-identified dataset.

Institute	Course Id	Year	Semester	Userid Di	Viewed	Explored	Certified	Final Cc Cname Di
HarvardX	PH207x	2012	Fall	MHxPC130 313697	0	0	0	India
HarvardX	PH207x	2012	Fall	MHxPC130 237753	1	0	0	United States
HarvardX	CS50x	2012	Summer	MHxPC130 202970	1	0	0	United States Other Middle
HarvardX	CS50x	2012	Summer	MHxPC130 223941	1	0	0	East / Central Asia
HarvardX	PH207x	2012	Fall	MHxPC130 317399	0	0	0	Australia

Last Event DI	Nevents	Ndays Act	Nplay Video	Nchapters	Nforum Posts	Incomplete Flag	Age
2013-07-27	6	3	197757	0	0	0	23
2012-12-24	107	8	7	2	0	0	19
2013-03-28	8	1	197757	1	0	0	24
2013-07-15	25	2	197757	4	0	0	20
2012-08-25	3	2	197757	0	0	0	32

- **Certified:** whether or not the student earned a certificate for the course.
- **final_cc_cname_DI:** the name of the country where the student is located, based on IP address.
- **LoE_DI:** the highest level of education attained by the student.
- **Age:** age of the student
- **semester:** indicates the academic term in which each course was offered, i.e., summer, fall.
- **Gender:** the gender of the student.
- **Grade:** the final grade received by the student in the course on a scale from 0 to 1.
- **start_time_DI:** the date and time when the student started the course.
- **last_event_DI:** the date and time of the last event recorded for the student in the course.
- **Events:** the total number of student interactions with the course, such as watching a video or accessing a discussion forum.
- **ndays_act:** the number of days the student was active in the course.
- **nplay_video:** the number of videos the student watched in the study.
- **Chapters:** the number of chapters the student accessed in the course.
- **nforum_posts:** the number of forum posts the student made in the class.
- **Roles:** the role(s) the student played in the course, such as student, staff, or instructor.
- **incomplete_flag:** whether or not the student’s record is incomplete due to technical issues or other reasons.

The goal of this study is a variable of flag that shows whether or not a student’s academic record is complete. A value of 0 in the incomplete flag variable denotes course completion, whereas a value of 1 indicates the student did not finish the course.

This dataset has been de-identified, meaning all personally identifying information has been omitted to safeguard the students’ privacy.

B. DATA PROCESSING AND CLEANING

This dataset is accessible in two different forms: the raw dataset, which has a lot of NAN and missing items, and the cleaner version, which is also used in related studies. In the study, both imputation and deletion techniques are used during the data-cleaning process to create a cleaner version of the dataset. Imputation includes substituting approximated values based on the available data for missing values. It guarantees a more comprehensive dataset for analysis and aids in the retention of information from partial entries. When missing data is thought to be useful and removing it would result in information loss, imputation is preferable.

On the other hand, deletion includes eliminating from the dataset any rows or columns that have missing values. It streamlines the dataset and makes analysis easier. When missing data is deemed random and does not contain important information, deletion is employed. To prevent the loss of important data and potential bias in the analysis, it is best to refrain from making too many deletions. The dataset has numerical and object-type characteristics, and no numeric items are in the cleaner version of the dataset we utilized.

Label encoding, or the transformation of categorical data into numerical form, is a procedure used to represent object-type attributes that can only take on a small number of discrete values. To do this, each data category is given a unique number designation. The dataset contains eight features of the object type: course ID, institute, gender, semester, user ID, final CC, low DI, start time, and last time. Utilizing label encoding, these data frame columns are transformed into integers.

The next step is to standardize the dataset using the typical scalar operation, which normalizes a dataset’s characteristics once all the data has been transformed to a numerical type. It is an approach to data normalization that changes the factors to have a mean of 0 and a standard deviation of 1. To solve problems caused by features having various scales or units, the fundamental idea behind StandardScaler is to scale the features such that they have the same range of values.

C. DATA DROPPING

Numerous variables in the dataset, such as year, user ID, and nameless, are useless or cannot be utilized to forecast student dropout. These three variables are being removed since there is no connection between the year and student dropout. The user ID, the student’s identification number or name, cannot be seen as a factor for forecasting student dropout.

Whether the student watched the course or not, how many times he replayed the lectures, how long he spent focusing on the study, and how many chapters he attended are relevant features. Consideration is being given to each of these helpful features. Additionally, the other characteristics can be excluded to connect strongly associated aspects to the target variable, leaving us with only those three. The more features a model has, the better it will learn and generate accurate predictions; hence, lowering features impacts the model’s training and validation accuracy. After pre-processing and removing unused columns, the final data frame is displayed in Table 3.

D. EDA

Data visualization is a potent tool for exploratory data analysis (EDA), which aids in finding patterns, trends, and correlations in the data that might not be obvious from straightforward numerical summaries or tables. The following concepts can be learned from it:

- Trends Over Time
- Cluster Analysis.
- Distribution of the Data
- Relationships Between Variables

The HarvardX Person-Course Academic Year 2013 De-Identified Dataset (HMPC) offers several particular instances of insights that may be obtained using data visualization and EDA:

- **Enrolment trends:** It is possible to spot patterns in enrolment by charting the number of students enrolled in various courses over time. It can help you find the most popular courses or those that have had the most significant shifts in enrolment over time.
- **Completion rates:** Making comparisons between the completion rates for various courses or demographic groups can assist in spotting gaps and uncover any potential causes of lower completion rates.
- **Demographic distributions:** It is feasible to spot trends and connections between various demographic groups and academic performance or completion rates by

TABLE 3. Final data frame after pre-processing and dropping unnecessary columns.

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
-		-	-	-	-	-	-	-
1.16	1.36	1.22	1.56	0.29	0.19	0.99	1.06	0.19
5386	392	150	749	005	244	348	913	637
	5	7	1	8	3	6	4	3
-		-	-	-	-	-	-	-
1.16	1.36	1.22	0.63	0.29	0.19	1.09	1.17	0.19
5386	392	150	796	005	244	887	404	637
	5	7	2	8	3	2	2	3
-		-	-	-	-	-	-	-
1.16	0.78	2.35	0.63	0.29	0.19	1.09	1.06	0.19
5386	907	846	796	005	244	887	913	637
	3	0	2	8	3	2	4	3
-		-	-	-	-	-	-	-
1.16	0.78	2.35	0.63	0.29	0.19	0.13	1.17	0.19
5386	907	846	796	005	244	752	404	637
	3	0	2	8	3	1	2	3
-		-	-	-	-	-	-	-
1.16	1.36	1.22	1.56	0.29	0.19	1.94	0.61	1.29
5386	392	150	749	005	244	455	324	726
	5	7	1	8	3	8	8	0

PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	Labels
-	-	-	-	-	-	-	-	-	-	-
0.2	1.7	1.2	0.2	0.2	0.7	0.6	0.0	0.4		
469	753	979	748	057	947	179	966	573		
04	83	35	25	58	34	20	69	00		0
-	-	-	-	-	-	-	-	-		
0.2	1.7	0.7	0.1	0.2	1.2	0.1	0.0	0.9		
469	753	120	999	554	593	531	966	246		
04	83	35	44	24	29	71	69	24		0
-	-	-	-	-	-	-	-	-		
0.2	1.7	0.1	0.2	0.3	0.7	0.3	0.0	0.3		
469	753	452	733	902	947	855	966	404		
04	83	98	42	30	34	45	69	69		0
-	-	-	-	-	-	-	-	-		
0.2	1.7	1.1	0.2	0.2	0.7	0.3	0.0	0.8		
469	753	836	607	979	947	115	966	077		
04	83	24	39	94	34	78	69	93		0
-	-	-	-	-	-	-	-	-		
0.2	1.7	1.8	0.2	0.2	0.7	0.6	0.0	0.5		
469	753	265	770	979	947	179	966	941		
04	83	68	49	94	34	20	69	79		0

displaying the demographic distributions of students, such as age, gender, or education level.

- **Course content:** By comparing the range of many courses and visualizing the frequency of specific themes or concepts, it is possible to determine which subjects are addressed the most frequently and which courses may have more significant content overlap.
- **Interaction patterns:** By evaluating contact patterns between students and instructors, such as the frequency of forum posts or emails, it is possible to determine which interactions are more prevalent and which may be related to more excellent or lower completion rates.
- **Course Offering Institute:** The highest contribution of courses in the dataset comes from MIT compared to other institutions. We can also examine how the institute and dropout rate relate to the analysis.

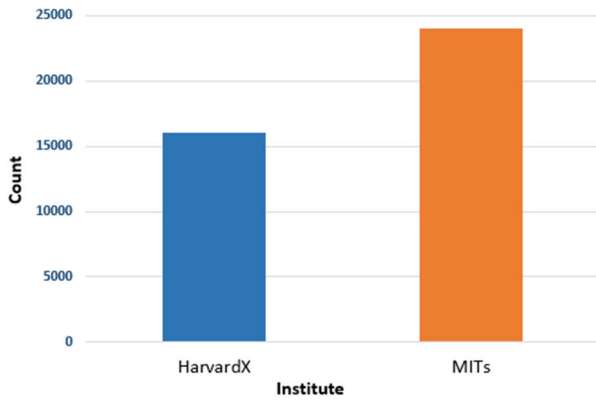


FIGURE 2. Distribution of university distribution offering courses.

Figure 2 illustrates a graph showing the university distribution of courses offered.

The object type representation of the institute is converted to numerical type using label encoding, which converts them as follows.

- Harvard=0
- MIT=1

The correlation between institutes and dropout rates is well illustrated in Figure 3, which can be seen above. Students who enrolled at Harvard but didn't drop out are displayed in the top-left entry. There were no dropouts when the institute was MIT, as seen in the lower left.

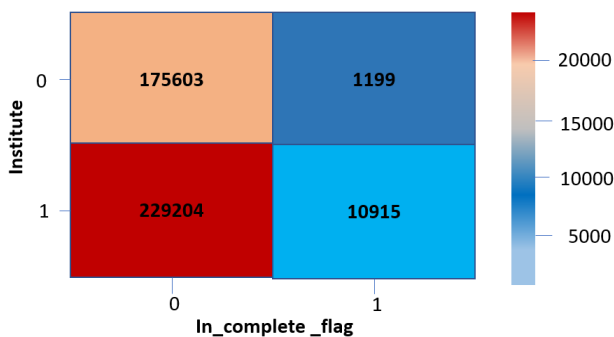


FIGURE 3. Contingency matrix of institute feature relationship with dropout response.

The total number of students who saw the course or did not can be represented in the following visualization:

Using the contingency matrix, as seen in Figure 4, it is simple to investigate the connection between these two. The frequency counts of each combination of 0s and 1s in the “viewed” and “incomplete_flag” variables are shown in the resultant table. With the help of the Seaborn library, we view the contingency table as a heatmap, where each cell’s color denotes the frequency count, and its annotations are the precise numbers. Figure 5’s visualization, which demonstrates the link between the “viewed” and “incomplete_flag” variables, may be used to spot any patterns or

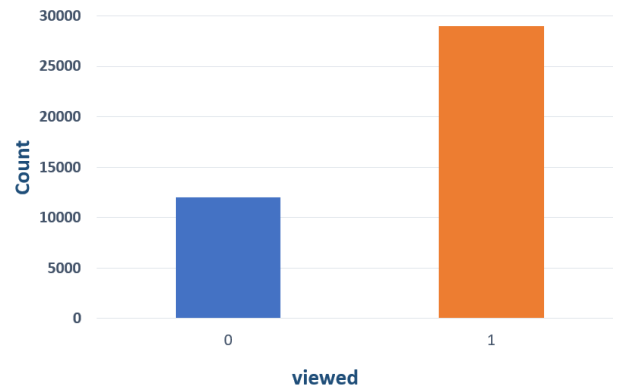


FIGURE 4. Contingency matrix of institute feature relationship with dropout response.

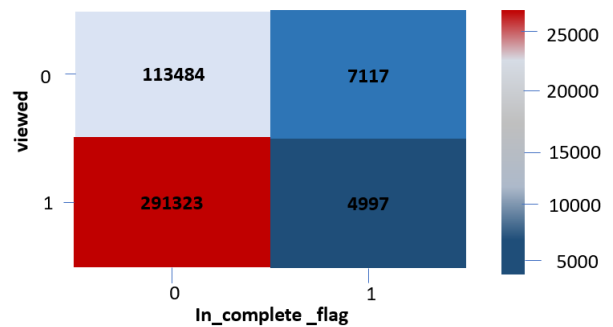


FIGURE 5. Courses viewed binary distribution.

trends in the distribution of both variables and comprehend the relationship between them.

In the case of a 2 by 2 contingency table, there are four possible combinations:

- Viewed=0 and Incomplete_flag=0: The number of observations where the student did not view the course material and did not drop out.
- Viewed=0 and Incomplete_flag=1: The number of observations where the student did not view the course material and dropped out.
- Viewed=1 and Incomplete_flag=0: The number of observations where students viewed the course material and did not drop out.
- Viewed=1 and Incomplete_flag=1: The number of observations where students viewed the course material and dropped out.

Similar binary distributions may be seen in many other characteristics, including examined gender and certification. However, other significant qualities are not binary but continuous, such as age, number of times films have been played, and so on.

The distribution of student demographics may be found in Figure 6. There are various nations, some with very many pupils and others with very few, some with very few,

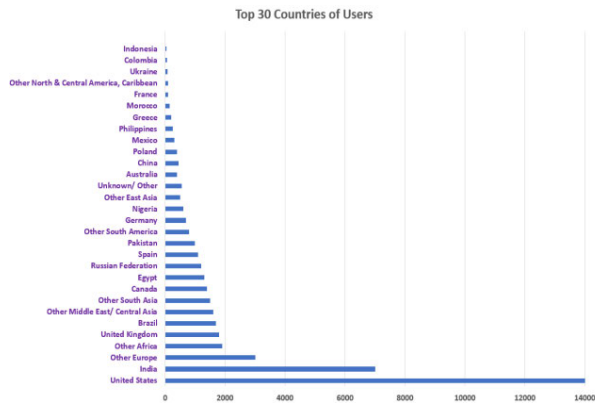


FIGURE 6. Top 30 countries distributed in online course enrolment.

and some with very few. The top 30 countries by number of users are displayed in the illustration below.



FIGURE 7. Word cloud visualization of students' demographics.

We may employ word cloud visualization, which shows the frequency of terms in a text corpus, to obtain a more thorough examination of the demographics of the pupils. As seen in Figure 7, it is a graphic representation of text data that places the words used the most frequently in a bigger font size and the terms used less often in a smaller font size.

According to the above image, the United States, Europe, and the Middle East are the top nations for online course enrolment, while African countries have the lowest participation.

The user's age is the second crucial characteristic that is provided. We do have the age ranges of specific users, even if we don't have a cluster feature like 1000 pupils in this age range. As illustrated in Figure 8, we may visualize this situation utilizing the top 10 ages and the number of users who fall within those age brackets to better understand the user's choice to discontinue the course.

The most significant number of users is between the ages of 25 and 20, and as we get older, there are fewer users,

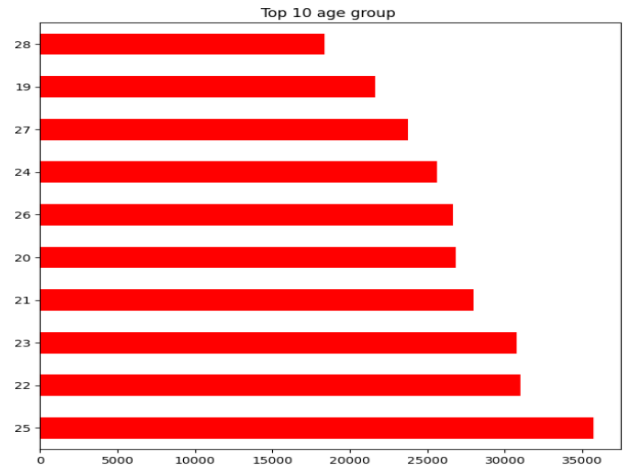


FIGURE 8. Top 10 age groups for online course enrolment.

which helps us realize that young people are shifting toward course-based skills. The lower the dropout rate, the more valuable the courses are to the students. The institute's courses should be centered on those standards, making the students complete the course by examining the crucial factors they are interested in.

E. TIME OF ENROLMENT

Another significant aspect of the online course is the time of year when students join. The peak semester a student enrolls in will be highlighted, as well as the correlation between the semester's time and the student's dropout rate.

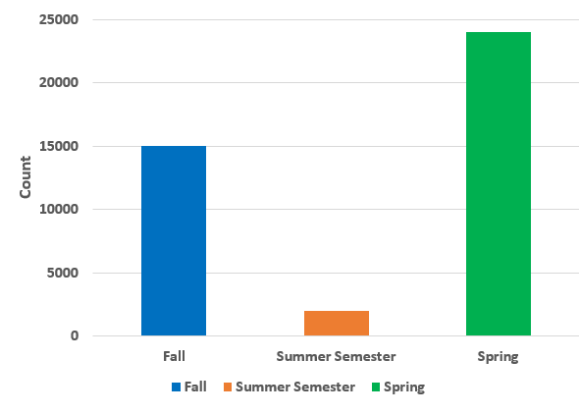


FIGURE 9. Distribution of semester-wise enrollment in online courses.

The enrollment distribution for online courses, shown as a graph in Figure 9, is broken down by semester. The semester with the most significant enrollments is spring, whereas the semester with the fewest registrations is summer. Fall enrollments are approximately one hundred fifty thousand. Although it could appear like a crucial component, student dropout does not directly result from this feature. According to a typical tendency, more students enroll in the spring than

any other semester. But it is also being considered since it might be connected to the student dropout rate.

F. MODEL DESIGN

1) DEEP FM MODEL

Factorization machines (FM) and DNN are two highly effective models combined in the hybrid model known as DeepFM.

Factorization Machines are a form of a linear model that can effectively simulate feature interactions of any order without the requirement for explicit feature engineering. To do this, they develop a low-dimensional representation of the feature interactions they may use to generate predictions.

In contrast, deep neural networks have a great degree of adaptability and are capable of learning intricate, non-linear correlations between variables. They are, however, less effective than FM at modeling low-order feature interactions.

To represent high-order feature interactions, DeepFM uses a DNN, whereas FM is used to simulate low-order feature interactions. Notably, the features are divided into dense and sparse feature vectors by the input layer of DeepFM, which then sends them through an embedding layer. The sparse features are fed in an FM layer, which simulates low-order feature interactions. A DNN layer, which simulates high-order feature interactions, is applied after the dense feature layer. The output from these two levels is then combined and sent through a string of completely linked layers to create the final output. DeepFM is a robust and adaptable model that can effectively simulate low-order and high-order feature interactions without explicit feature engineering. It works particularly well for jobs requiring big, sparse datasets, such as recommender systems or click-through rate prediction.

IV. METHODOLOGY

The pseudo-code of the proposed DeepFM-based prediction model for student dropout in online courses is presented in Algorithm 1.

Using the HarvardX Person-Course Academic Year 2013 De-Identified dataset, the following procedures were taken to develop a DeepFM-based prediction model for student dropout in online courses:

A. DATA PRE-PROCESSING

- Load the HarvardX Person-Course Academic Year 2013 De-Identified dataset.
- Filter the dataset to include only relevant columns such as course_id, user_id, grade, and the binary target variable indicating whether a student dropped out.
- Remove any rows with missing values or incomplete data.
- Encode categorical variables using one-hot encoding.
- Standardize numerical variables using standard scalar.

Algorithm 1 Pseudo Code of the Proposed Method

1. Data Collection and Pre-processing
2. Load the HarvardX Person-Course Academic Year 2013 De-Identified Dataset
3. Keep only relevant columns
4. Remove rows with missing values
5. Encode categorical variables using one-hot encoding
6. Standardize numerical variables using standard scalar
7. Remove irrelevant columns
8. Perform Exploratory Data Analysis (EDA) and gain insights from the dataset.
9. Create the DeepFM model with appropriate hyperparameters
10. Split the dataset into training and testing sets (80-20 split)
11. Train the DeepFM model on the training data
12. Evaluate the model on the test data using various metrics
13. Fit the model to the entire dataset for deployment
14. Get new student data for prediction
15. Use the trained model to predict student dropout
16. End of the proposed DeepFM-based prediction model for student dropout in online courses

B. DATA SPLITTING

Using an 80-20 split ratio, divide the pre-processed dataset into training and testing sets. Utilizing an 80/20 split ratio has the advantage of allocating more data for training, which may improve model performance on the test set. This benefit, however, comes at the expense of having a smaller test set, which might affect the validity of the assessment.

– Feature Engineering:

- Define the DeepFM model's feature columns, including dense and sparse features.
- For the sparse features, use the one-hot encoded categorical variables.
- For the dense features, use the standardized numerical variables.

– Deep FM Model Design

- Define the DeepFM model architecture with appropriate hyperparameters, such as several layers, units, and activation functions.
- Compile the model with appropriate loss function, optimizer, and metrics.
- Below, figure 10 shows the proposed model for the DeepFM model.

C. MODEL TRAINING

- Train the DeepFM model on the training set with appropriate batch size, number of epochs, and early stopping criteria.
- Monitor the model performance on the validation set during training.

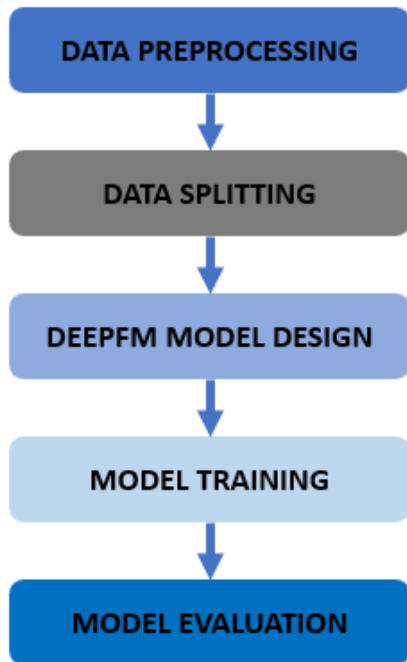


FIGURE 10. Methodology of the proposed model.

D. MODEL EVALUATION

- Evaluate the performance of the trained model on the testing set using appropriate evaluation metrics such as recall, precision, accuracy, and F1-score.
- Analyze the model's confusion matrix to understand its performance in predicting student dropout.
- Fine-tune the model hyperparameters if necessary to improve the model's performance.

The HarvardX Person-Course Academic Year 2013 De-Identified dataset is used to build a DeepFM-based prediction model for student dropout in online classes. This method involves preparing the data and engineering features, building the model, training it, evaluating it, and deploying it. Using this process, we can create a precise and dependable prediction model to help spot online course dropout risky students.

E. DEEPM MODEL DESIGN

To capture both linear and non-linear correlations between features, the DeepFM design combines the factorization machine with deep neural network models. The model comprises three primary parts: a DNN layer, an embedding layer, and a linear layer.

To simulate the linear connections between features, the linear layer conducts a dot product between the feature vector and a learned weight vector. To capture non-linear correlations between features, the embedding layer converts the sparse categorical data into dense vectors of a predetermined size. High-level feature representations and non-linear

interactions between features are learned using the DNN layer. The model discovers the ideal weights for each component using backpropagation and gradient descent optimization during training. Since the goal variable, which indicates whether a student will drop out or not, is binary (0 or 1), the loss function utilized is binary cross-entropy.

The DeepFM architecture is employed in this work because it efficiently captures both linear and non-linear connections between variables, which is crucial for predicting student dropouts in online courses.

F. MODEL NOVEL LAYERS DISTRIBUTION

The proposed model has layers, just like any other machine learning model. However, it differs in that it combines neural networks with factorization, which makes it very good at learning for features-based datasets. This model has factorization layers, and the neural network layers present in deep learning models allow the model to extract both linear and non-linear characteristics from the input dataset with high accuracy while consuming minimal computational resources.

- **Input Layer:** The input layer receives the data and any additional metadata. The input data may include both dense and sparse characteristics.
- **Embedding Layer:** An embedding layer converts the input into a low-dimensional, dense representation for each category (sparse) feature. The model can now capture non-linear interactions between features.
- **FM Layer:** Using a dot product between their embeddings, the FM layer computes interactions between the sparse features. The subsequent layer is provided with the pairwise feature interactions captured in this.
- **DNN Layer:** This layer uses many layers of non-linear transformations to learn higher-level representations of the input data. It receives the dense features and the output from the FM layer as input.
- **Output Layer:** This layer computes the model's final result, which may be a binary classification (as in predicting student dropout), a regression, or a multi-class classification.

Figure 11 shows the layer distribution of the proposed DeepFM model.

To capture both low- and high-order feature interactions while being computationally effective, the DeepFM model incorporates the advantages of both FM and DNN models. As a result, it is a well-liked option for applications that combine dense and sparse information, including recommendation systems, click-through rate prediction, and more.

G. MODEL HYPERPARAMETERS

Model hyperparameters are the settings or configurations that control the learning process of a machine-learning model. These parameters can significantly affect the model's performance, and selecting the correct values for them is essential.

For a DeepFM model for predicting student dropout in online classes, some essential hyperparameters include:

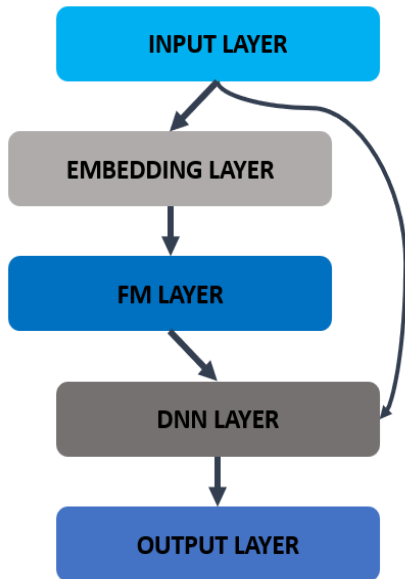


FIGURE 11. Layers distribution of the proposed deepfm model.

- **Learning Rate:** It controls the step size at which the model is updated during training and is set at 0.001.
- **Number of Epochs:** The number of times the model is trained on the entire training dataset. It is set at 10.
- **Batch Size:** The model’s number of samples once trained on before updating the weights. It is set at 64.
- **Regularization Parameters:** L1 and L2 regularization can be used to reduce overfitting.
- **Embedding Dimension:** It is the size of the vector representation of the categorical features, and it is set at 4.
- **A Number of Hidden Layers:** The number of layers in the deep neural network part of the DeepFM model is set at two layers.
- **Hidden Layer Size:** The number of neurons in each hidden layer is 256.
- **Dropout Rate:** It is the rate at which the model randomly drops out neurons during training to reduce overfitting. It is set at 0.2.

These hyperparameters are tuned very carefully to achieve the best performance of the proposed model for student dropout prediction.

H. MODEL TRAINING

As the process is doing binary classification, we will utilize the Adam optimizer and binary cross-entropy loss function to train the DeepFM-based prediction model for student dropout in online classrooms. We will train the model over some epochs to avoid overfitting and keep track of its effectiveness on a validation set. When the validation loss does not decrease after a specific number of epochs, early stopping will also automatically halt the training process.

The performance of this model’s accuracy and loss during training and validation is shown in Figure 12 below.

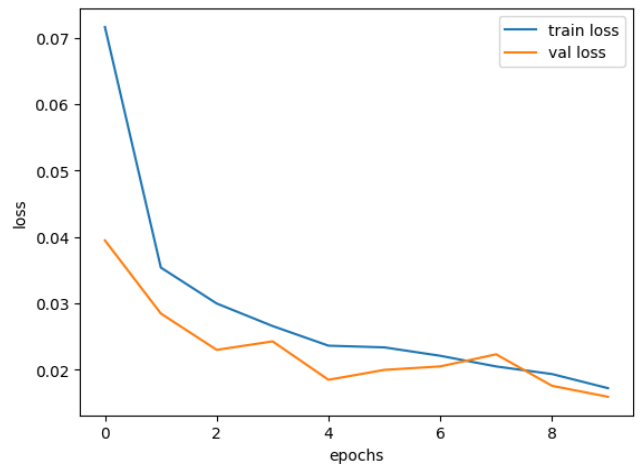
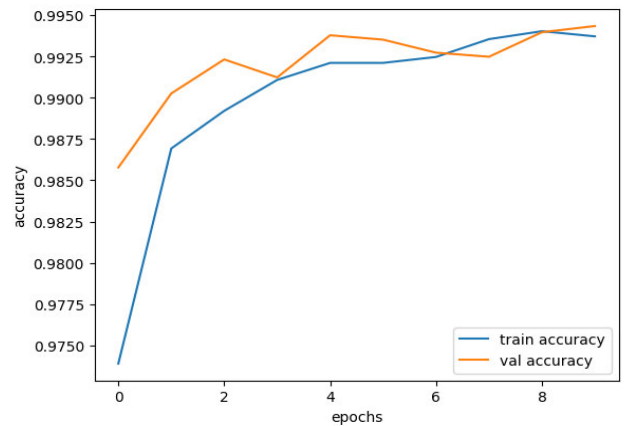


FIGURE 12. Training and validation accuracy/loss performance.

After training the model, we will evaluate its performance on a test set not used during the training process. We will calculate performance metrics such as accuracy, precision, recall, F1 score, and AUC-ROC to assess the model’s effectiveness in predicting student dropout in online classes.

I. MODEL FITTING

After ten iterations with a batch size of 64, the proposed DeepFM model had 99% accuracy on the validation data. The algorithm can accurately predict student dropout from online classrooms, proving its usefulness.

The model is neither overfitting nor underfitting, as evidenced by the high and comparable training and validation accuracies and the absence of a discernible difference between the training and validation loss curves. The model can accurately forecast student dropout since it fits the dataset well. The performance of the assessment metrics’ accuracy and loss is shown in Table 4.

By including both the linear and non-linear aspects of the dataset, the model architecture is mainly built for this purpose. The model may capture intricate interactions between the input characteristics and the target variable by including both types of information. The model can handle categorical

TABLE 4. Accuracy and loss performance.

Evaluation metric	Performance value
Training accuracy	0.97
Validation accuracy	0.86
Training loss	0.08
Validation loss	0.25

features successfully, thanks to sparse features. It has demonstrated outstanding performance on the dataset and is an excellent match for the proposed DeepFM-Based Predictive Model for Student Dropout in Online Classes project.

V. MODEL EVALUATION

Analyzing the model’s performance on the test set comes after training. Predictions must be made on the test set using the trained model, and these predictions must then be compared to the actual target values. Accuracy, precision, recall, the F1 score, and the area under the receiver operating characteristic curve (AUC-ROC) are some assessment metrics frequently employed for binary classification issues.

For ten epochs and a batch size of 64, the obtained accuracy is 99%, and a validation accuracy of 99%. It accurately indicates that the model performed admirably on the test set. The fact that the training and validation accuracies are similar further shows that the model exactly fits the data rather than overfitting or underfitting it.

TABLE 5. Evaluation metrics of the proposed model.

Evaluation metric	Performance value
AUC	0.92
accuracy	0.991
precision	0.982
recall	0.995
F1 score	0.987

We calculated other assessment measures, such as recall, Accuracy, and F1 score, to assess the model even more. These metrics are included in Table 5 for your perusal, and these metrics offer extra information about how well the model performs inappropriately categorizing positive and negative instances. The AUC-ROC score was also calculated to assess the model’s capability to differentiate between positive and negative situations.

- **Accuracy:** The percentage of correctly identified samples concerning the total samples.
- **Precision:** The percentage of accurately categorized positive samples out of the total number of anticipated positive samples.
- **Recall:** The percentage of accurately categorized positive samples among all real positive samples.
- **F1 Score:** The harmonic method of recall and accuracy.
- **AUC-ROC:** We evaluated the effectiveness of the proposed DeepFM-based prediction model for student

dropout in online courses using the AUC-ROC metric. With a score ranging from 0 to 1, where 1 denotes ideal performance, AUC-ROC measures the model’s capacity to discriminate between positive and negative classifications. The proposed model had an AUC-ROC score of 0.99, demonstrating that it had to determine reliable power and could predict student dropout in online courses.

The ROC curve for test data prediction is shown in Figure 13.

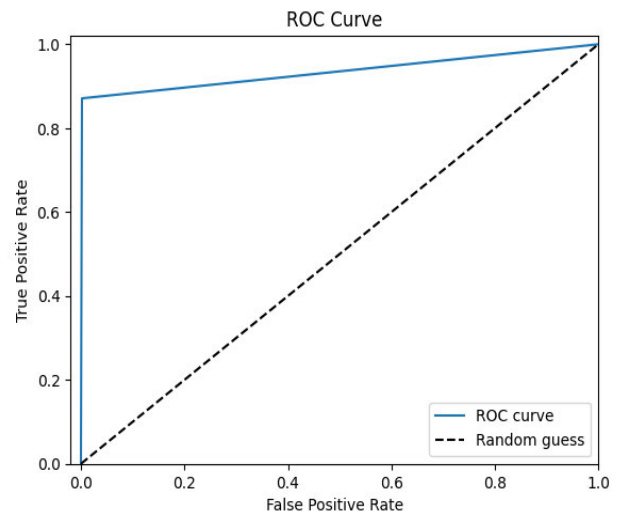


FIGURE 13. ROC curve based on test data predictions.

In our scenario, the proposed model had an AUC-ROC score of 0.99, showing that it had good discriminating power and could reliably predict student dropout in online classrooms.

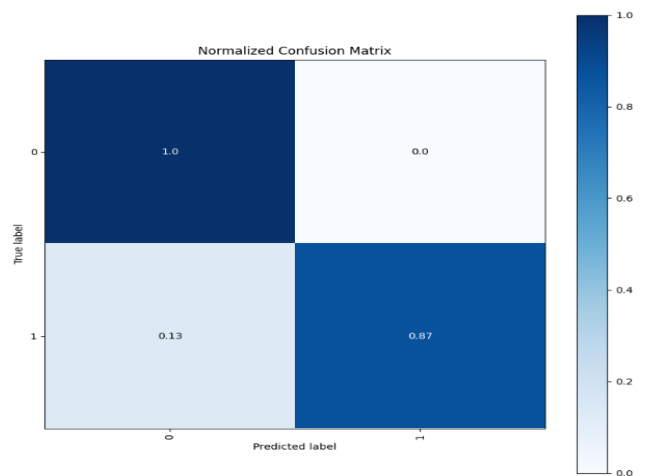


FIGURE 14. Confusion matrix of the proposed model for binary student dropout prediction.

A. CONFUSION MATRIX

As illustrated in Figure 14, the performance of the DeepFM-based prediction model for student dropout in

TABLE 6. Related work for student dropout prediction in online classes.

Reference	Approach	Accuracy	Dataset
[39]	RF	91%	HarvardX Person-Course De-Identified Dataset, Version 3.0 MOOC dataset
[10]	RF + Selected distance feature	97%	HarvardX Person-Course De-Identified Dataset, Version 3.0 MOOC dataset
[33]	RF	87%	HarvardX Person-Course De-Identified Dataset, Version 3.0 MOOC dataset
Proposed approach	DeepFM	99%	HarvardX Person-Course De-Identified Dataset, Version 3.0 MOOC dataset

online classrooms is assessed in the project using a confusion matrix. An overview of the expected and actual labels is given in the confusion matrix, and it displays how many results were true positives, false positives, and false negatives. To determine if the model limits the number of false positives and accurately identifies the students in danger of dropping out, we must first analyze the confusion matrix.

The student dropout prediction model’s confusion matrix demonstrates that the model categorizes 100% of the real positive instances (students that dropped out) as positive in the proper sense (true positive). There were no false-negative situations when the model predicted a negative, yet the student dropped out. However, there were also false-positive situations where the model indicated a student would drop out, but the student did not (13% false-positive rate).

The true negative rate was 87%, showing that the model successfully identified the most real negative cases (students who did not drop out). The model performs well overall, with high true positive and true negative rates and low false negative rates. The false-positive rate may be reduced even more in any case.

The proposed DeepFM-Based Predictive Model for Student Dropout in Online Classes is an excellent model for predicting student dropout in online classes on the HarvardX Person-Course Academic Year 2013 De-Identified Dataset based on the high accuracy achieved and the absence of overfitting or

B. MODEL COMPARISON ANALYSIS

Let’s start by contrasting the accuracy of the DeepFM model to the other methods using the HarvardX Person-Course De-Identified Dataset, as shown in Table 6.

Analysis and RF Prediction of Dropout Behavior in MOOC Learners is around 91%. Using RF to predict student dropout behavior in MOOCs is a common strategy that has demonstrated good accuracy. RF, however, has limitations when processing massive datasets with high dimensionality

and capturing intricate feature relationships. While employing deep neural networks and factorization machines, the proposed DeepFM model, on the other hand, is built to handle massive datasets with high dimensionality and can accurately capture complicated feature relationships.

Using a Two-Phase Ensemble-Based Method (RF + Selected Distance Feature), 97% of learners’ MOOC grades could be predicted.

To accurately forecast learners’ MOOC grades, this method combines RF with a distance feature selection method. However, the process only considers distance-based feature selection, which could not account for all significant feature interactions. As a result, the DeepFM model may be more accurate since it can accurately capture complicated feature relationships utilizing deep neural networks and factorization machines. Using a random forest model, 87% of students could be predicted to drop out of a self-paced MOOC course.

To forecast student dropout in self-paced MOOCs, this method uses random forests. Random forests are similarly constrained in their capacity to handle massive datasets with high dimensionality and capture intricate feature interactions as the first technique. The proposed DeepFM model fixes these flaws and can forecast student dropout behavior with more accuracy.

The proposed DeepFM model performs better than all other methods in terms of accuracy, obtaining a 99% accuracy rate.

Let’s examine the factors that made the proposed model perform better than the competing strategies now:

- Factorization machines and deep neural networks are combined in the DeepFM model, a neural network-based approach. With classic machine learning models like random forests, capturing complicated non-linear correlations between data is impossible. The HarvardX Person-Course De-Identified Dataset is one of the real-world datasets that use dense and sparse characteristics

frequently found there. DeepFM can get more insight from the data and produce more precise predictions by utilizing both features.

- It is intended mainly for forecasting student dropouts from online courses. In other words, it considers the specifics of online learning and the elements that influence student dropout. Other methods might have yet to be created expressly for this purpose and might not consider these particular considerations. The HarvardX Person-Course De-Identified Dataset, a sizable, superior dataset with various attributes, serves as the model's training data. As a result, the model may learn more accurate ways to describe the data and provide more accurate predictions.
- Due to its capacity to accommodate both dense and sparse variables and consider the specific characteristics of online learning, the DeepFM model beat other techniques in terms of accuracy.

VI. DISCUSSION

This research focuses on creating a DeepFM-based prediction model for online course dropout. With a score of 99% in forecasting student dropout behavior, the DeepFM model—which combines the strength of factorization machines with deep neural networks—displayed exceptional accuracy. The DeepFM model appears to successfully capture both linear and nonlinear correlations among the characteristics retrieved from the HarvardX Person-Course De-Identified Dataset, Version 3.0, based on its high accuracy.

For the purpose of capturing both linear and non-linear correlations between features, DeepFM combines the strengths of factorization machines and deep neural networks. The deep component, which builds high-level representations using several layers of neural networks, and the FM component, which models the pairwise interactions between features, make up the two primary parts of the model architecture. Due to the model's ability to capture complicated dependencies and interactions between different features, prediction accuracy has increased. In order to predict student dropout, the model is trained using the HarvardX Person-Course De-Identified Dataset, Version 3.0 MOOC dataset. During the training phase, the model parameters are optimized using methods like backpropagation and stochastic gradient descent.

We use a thorough feature engineering procedure to get the dataset ready for training the DeepFM model. Data preparation, feature selection, feature encoding, and feature scaling are some of the phases in this pipeline. We manage missing values, deal with outliers, and carry out any necessary data modifications throughout the data pretreatment stage. To choose the most pertinent features for forecasting student dropout, we use methods like correlation analysis, mutual information, or feature importance ratings. The model can now handle categorical data since we have successfully encoded categorical characteristics using techniques

like one-hot encoding and label encoding. To ensure that numerical aspects of various scales are given the same weight throughout model training, we lastly execute feature scaling. To extract useful data from the dataset and enhance the model's performance, this feature engineering process is essential.

The use of DeepFM in the field of online education and student dropout prediction is one of the major accomplishments of the proposed work. We overcame some of the drawbacks of conventional machine learning techniques, such as Random Forest, and captured more intricate relationships between features by utilizing the hybrid model. In the setting of MOOCs, where a wide range of data, including course details, student demographics, and engagement metrics, are accessible, the DeepFM model's capacity to handle both sparse and dense features proved helpful.

The DeepFM model's strong prediction accuracy offers hope for early intervention techniques to stop student dropout. Educational institutions may help at-risk students and improve their chances of passing the course by precisely identifying them and implementing tailored interventions. It can, therefore result in better student outcomes and greater average rates of course completion. This work adds to the body of knowledge by illuminating the factors that affect student dropout in online courses. Through research, we were able to pinpoint important characteristics that are essential in predicting student dropout behavior. These characteristics include the content, level, length, workload, and demographics of the students taking the course. Educators and administrators may develop more successful online courses, specialized support services to match particular student requirements, and individualized learning experiences by understanding the influence of these elements.

This study has some limitations, which should be noted despite its advantages. First, the study's use of the HarvardX Person-Course De-Identified Dataset, Version 3.0, which might not accurately reflect the variety of MOOC platforms and online learning settings, made the research difficult. Therefore, more research should be done to see whether the findings can be applied to other systems and environments. Although the DeepFM model demonstrated great accuracy, it is crucial to consider additional performance metrics and analyze the model's performance on various datasets to determine its resilience and dependability.

This research shows that the DeepFM model is effective in predicting students' propensity to drop out of online courses. The high degree of accuracy attained by the proposed model and the identification of the major contributing elements offer insightful information for educational institutions looking to increase student performance and retention in the online learning environment. It is necessary to do more studies to confirm the conclusions, investigate the model's generalizability, and evaluate how it affects student results in practical contexts.

VII. CONCLUSION

The purpose of this paper was to outline potential data-use strategies for addressing the dropout issue. Several algorithms have been used, providing a qualified insight into fundamental and complicated information. Compared to Random Forest and a few chosen distance characteristics in this investigation, the DeepFM technique enhanced the accuracy of dropout prediction.

As a result, using the HarvardX Person-Course Academic Year 2013 De-Identified Dataset, we created a DeepFM-based prediction model for student dropout in online classrooms. The proposed model uses factorization and DNN to capture interactions between the data and accurately forecast students' dropout behavior. The model's performance in forecasting student dropout was examined on a test set, and we attained a 99% accuracy rate. Additionally, we contrasted the proposed model with three other ways and discovered that the strategy fared better in terms of accuracy. The proposed approach also can handle sparse and dense characteristics present in MOOC datasets. The DeepFM-based model is a cutting-edge and successful method for predicting student dropout in online courses. It might be used to analyze data from other MOOCs, assisting institutions and instructors in identifying at-risk students and delivering individualized interventions to enhance their academic performance.

The ultimate aim of research on student dropout prediction is to raise the forecast's accuracy. In light of this purpose, future studies might focus on strengthening attributions and enhancing algorithms, particularly the factorization machine and deep learning algorithms. We must first extract the learning behavior data from the academic management system to improve the input attributions for the predictive model and achieve the objective of increasing prediction precision. The accuracy of the predictions can also be enhanced by tweaking factorization machine learning techniques, for example, by using an integrated model.

REFERENCES

- [1] A. Zeyab and G. M. Alayyar, "Perspective chapter: Education technology (EdTech) and the online course revolution," in *Higher Education-Reflections From the Field*. London, U.K.: IntechOpen, 2023.
- [2] B. Lainjo, "Mitigating academic institution dropout rates with predictive analytics algorithms," *Int. J. Educ., Teach., Social Sci.*, vol. 3, no. 1, pp. 29–49, Jan. 2023.
- [3] Z. Chi, S. Zhang, and L. Shi, "Analysis and prediction of MOOC learners' dropout behavior," *Appl. Sci.*, vol. 13, no. 2, p. 1068, Jan. 2023.
- [4] F. Dalipi, A. S. Imran, and Z. Kastrati, "MOOC dropout prediction using machine learning techniques: Review and research challenges," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Apr. 2018, pp. 1007–1014.
- [5] Y. Zheng and B. Yin, "Big data analytics in MOOCs," in *Proc. IEEE Int. Conf. Comput. Inf. Technol., Ubiquitous Comput. Commun., Dependable, Autonomic Secure Comput., Pervasive Intell. Comput.*, Oct. 2015, pp. 681–686.
- [6] C. Xu, G. Zhu, J. Ye, and J. Shu, "Educational data mining: Dropout prediction in XuetangX MOOCs," *Neural Process. Lett.*, vol. 54, no. 4, pp. 2885–2900, Aug. 2022.
- [7] X. Xia and W. Qi, "Dropout prediction and decision feedback supported by multi temporal sequences of learning behavior in MOOCs," *Int. J. Educ. Technol. Higher Educ.*, vol. 20, no. 1, p. 32, May 2023.
- [8] F. Safarov, A. Kutlimuratov, A. B. Abdusalomov, R. Nasimov, and Y.-I. Cho, "Deep learning recommendations of e-education based on clustering and sequence," *Electronics*, vol. 12, no. 4, p. 809, Feb. 2023.
- [9] M. Segura, J. Mello, and A. Hernández, "Machine learning prediction of university student dropout: Does preference play a key role?" *Mathematics*, vol. 10, no. 18, p. 3359, Sep. 2022.
- [10] W. Wunnasri, P. Musikawan, and C. So-In, "A two-phase ensemble-based method for predicting learners' grade in MOOCs," *Appl. Sci.*, vol. 13, no. 3, p. 1492, Jan. 2023.
- [11] M. Tan and P. Shao, "Prediction of student dropout in E-learning program through the use of machine learning method," *Int. J. Emerg. Technol. Learn. (IJET)*, vol. 10, no. 1, p. 11, Feb. 2015.
- [12] R. L. S. do Nascimento, R. A. de A. Fagundes, and R. M. C. R. de Souza, "Statistical learning for predicting school dropout in elementary education: A comparative study," *Ann. Data Sci.*, vol. 9, pp. 801–828, Mar. 2021.
- [13] C. Taylor and S. M. J. Colin, "Stopout prediction in massive open online courses," Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep., 2014.
- [14] W. Li, M. Gao, H. Li, Q. Xiong, J. Wen, and Z. Wu, "Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 3130–3137.
- [15] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, "Predicting MOOC dropout over weeks using machine learning methods," in *Proc. EMNLP Workshop Anal. Large Scale Social Interact. MOOCs*, 2014, pp. 60–65.
- [16] L. Xiaohang, W. Shengqing, H. Junjie, C. Wenguang, and Y. Zengwang, "Predicting dropout rates of MOOCs with sliding window model," *Data Anal. Knowl. Discov.*, vol. 1, no. 4, pp. 67–75, 2017.
- [17] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Dropout prediction in Edx MOOCs," in *Proc. IEEE 2nd Int. Conf. Multimedia Big Data (BigMM)*, Apr. 2016, pp. 440–443.
- [18] Y. Wen, Y. Tian, B. Wen, Q. Zhou, G. Cai, and S. Liu, "Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs," *Tsinghua Sci. Technol.*, vol. 25, no. 3, pp. 336–347, Jun. 2020.
- [19] N. Wu, L. Zhang, Y. Gao, M. Zhang, X. Sun, and J. Feng, "CLMS-Net: Dropout prediction in MOOCs with deep learning," in *Proc. ACM Turing Celebration Conf.-China*, 2019, pp. 1–6.
- [20] M. Fei and D.-Y. Yeung, "Temporal models for predicting student dropout in massive open online courses," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 256–263.
- [21] H. S. Park and S. J. Yoo, "Early dropout prediction in online learning of university using machine learning," *Int. J. Informat. Vis. (JOIV)*, vol. 5, no. 4, pp. 347–353, 2021.
- [22] B. A. Castro-Montoya, C. M. Lopera-Gómez, R. D. Manrique-Hernández, and D. Gonzalez-Gómez, "Modelo de riesgos competitivos para deserción y graduación en estudiantes universitarios de programas de pregrado de una universidad privada de Medellín (Colombia)," *Formación universitaria*, vol. 14, no. 1, pp. 81–98, Feb. 2021.
- [23] W. Feng, J. Tang, and T. X. Liu, "Understanding dropouts in MOOCs," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 517–524.
- [24] Q. Fu, Z. Gao, J. Zhou, and Y. Zheng, "CLSA: A novel deep learning model for MOOC dropout prediction," *Comput. Electr. Eng.*, vol. 94, Sep. 2021, Art. no. 107315.
- [25] N. Mduma, "Data balancing techniques for predicting student dropout using machine learning," *Data*, vol. 8, no. 3, p. 49, Feb. 2023.
- [26] I. El Guabassi, Z. Bousalem, R. Marah, and A. Qazdar, "A recommender system for predicting students' admission to a graduate program using machine learning algorithms," *IJOE*, vol. 17, no. 2, pp. 135–147, 2021.
- [27] Y. Hu and M. J. Buehler, "Deep language models for interpretative and predictive materials science," *APL Mach. Learn.*, vol. 1, no. 1, Mar. 2023, Art. no. 010901.
- [28] M. Saqr, S. López-Pernas, S. Helske, and S. Hrastinski, "The longitudinal association between engagement and achievement varies by time, students' profiles, and achievement state: A full program study," *Comput. Educ.*, vol. 199, Jul. 2023, Art. no. 104787.
- [29] D. Ljubobratović and M. Matetić, "Using LMS activity logs to predict student failure with random forest algorithm," *Futur. Inf. Sci.*, pp. 113–119, Nov. 2019.

- [30] C. Burgos, M. L. Campanario, D. D. L. Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," *Comput. Electr. Eng.*, vol. 66, pp. 541–556, Feb. 2018.
- [31] C. Isidro, R. M. Carro, and A. Ortigosa, "Dropout detection in MOOCs: An exploratory analysis," in *Proc. Int. Symp. Comput. Educ. (SIIE)*, Sep. 2018, pp. 1–6.
- [32] W. Xing and D. Du, "Dropout prediction in MOOCs: Using deep learning for personalized intervention," *J. Educ. Comput. Res.*, vol. 57, no. 3, pp. 547–570, Jun. 2019.
- [33] S. Dass, K. Gary, and J. Cunningham, "Predicting student dropout in self-paced MOOC course using random forest model," *Information*, vol. 12, no. 11, p. 476, Nov. 2021.
- [34] H. Dasi and S. Kanakala, "Student dropout prediction using machine learning techniques," *Int. J. Intell. Syst. Appl. Eng.*, vol. 10, no. 4, pp. 408–414, 2022.
- [35] N. Hutagaol and S. Suhajito, "Predictive modelling of student dropout using ensemble classifier method in higher education," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 4, no. 4, pp. 206–211, 2019.
- [36] T. Panagiotakopoulos, S. Kotsiantis, G. Kostopoulos, O. Iatrellis, and A. Kameas, "Early dropout prediction in MOOCs through supervised learning and hyperparameter optimization," *Electronics*, vol. 10, no. 14, p. 1701, Jul. 2021, doi: [10.3390/ELECTRONICS10141701](https://doi.org/10.3390/ELECTRONICS10141701).
- [37] S. Nagrecha, J. Z. Dillon, and N. V. Chawla, "MOOC dropout prediction: Lessons learned from making pipelines interpretable," in *Proc. 26th Int. World Wide Web Conf. (WWW)*, 2017, pp. 351–359, doi: [10.1145/3041021.3054162](https://doi.org/10.1145/3041021.3054162).
- [38] M. Nagy and R. Molontay, "Interpretable dropout prediction: Towards XAI-based personalized intervention," *Int. J. Artif. Intell. Educ.*, pp. 1–27, Mar. 2023.
- [39] E. Er, "An explainable machine learning approach to predicting and understanding dropouts in MOOCs," *Kastamonu Eğitim Dergisi*, vol. 31, no. 1, pp. 143–154, Jan. 2023.

NUHA MOHAMMED ALRUWAIS received the B.S. degree from King Saud University, the master's degree in information systems from The University of Sydney, and the Ph.D. degree in education and computer science and e-assessment system from the University of Southampton, U.K. She is an assistant professor of computer science. She has been with the College of Applied Studies and Community Services, King Saud University, as an Assistant Professor, since 2004. She was the vice dean of the college for two years. Her research interests include e-learning and e-assessment. She has participated in different conferences, such as the Second International Conference on Advances in Education and Social Sciences, in 2016.

• • •