MOOC Dropout Prediction using FIAR-ANN Model based on Learner Behavioral Features

S. Nithya[1]*
Research Scholar, Department of Computer Science
SRM Institute of Science and Technology
Ramapuram Campus, Chennai, India

Dr. S.Umarani[2]
Professor, Department of Computer Science
SRM Institute of Science and Technology
Ramapuram Campus, Chennai, India

***Abstract*—Massive Open Online Courses (MOOCs) are a transformative technology in digital learning that incorporates new techniques through video sessions, exams, activities, and conversations. Everyone leads a successful life in their professional and personal skills learning courses during COVID-19. The research concentrated on employing video interaction analysis to characterize crucial MOOC tasks, including predicting dropouts and student achievement. Our work consists of merely generating and picking the best characteristics based on the learner behavior for evaluating the dropout measure. To locate the frequent objects for feature creation, an association rule-FP growth approach is applied. The neural network is implemented using frequent itemset-3, which is used for feature selection. The evaluation metrics are calculated by using the Multilayer Perceptron (MLP) method. The metric values were then compared to the proposed model and some base supervised machine learning models namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bayes (NB). The FIAR (Feature Importance Association Rule)-ANN(Artificial Neural Network) dropout prediction model was tested on the KDD Cup 2015 dataset and it had a high accuracy of over 92.42, which is approximately 18% better than the MLP-NN model. With the optimized parameters, we are solely focused on lowering dropout rates and increasing learner retention.**

Keywords*—Dropout prediction; data analytics; association rule mining; machine learning; artificial neural network

I. Introduction

Massive Open Online Courses are the result of the blending of learning and the online world. It broadens the educational system by providing knowledge via modern internet technologies [1]. Moreover, an innovative instructional environment would assist learners in saving their learning time.

Billions of individuals utilize MOOCs for a variety of purposes, including professional growth, career transition, syllabus formation, secondary education, skills training, [2] and more. Courses are created by institutions for MOOC providers such as Coursera, edx, Udacity, and Swayam. Coursera, on the other hand, has done a far better job of preserving the pandemic boost than its closest MOOC competitors. From 31% in 2020 to 39% in 2021, Coursera's proportion of non-university courses expanded.

Learner enrollment in online classes has increased during the last decade. Simultaneously, learner dropout rates should be raised. It's essential to consider the elements that contribute to inadequate retention rates [3]. The elements have to do with a learner's logs, access to materials, activities, and willingness to execute required actions in order to complete the program. Along with low dropout rates, course completion rates will rise.

This work has become focused on the analysis of learners' behaviors. In effective learning behavior, while the learner watches the video within the duration and submits the assessment on time. Moreover, analyzing the behavior of learners [4] with specific features such as accessing materials, discussion, course forum, and so on is significant.

Analytics has been popularized in recent years. The analytical approaches that extract relevant and meaningful enormous volumes of data and apply them to the educational system have revolutionized research [5] and it indicates as "Education Data Mining" (EDM), "Academic Analytics" (AA), and "Learning Analytics" (LA).

EDM has proven to be a good resource for revealing hidden information [6] and paradigms in data sources. Most educational institutions are being carried on e-learning platforms [7]. The huge amount of educational information recorded has laid the foundation for new research and analysis to better understand and increase learning performance.

Learning analytics is defined as the use of intellectual facts, learner-generated data, and models developed [8] to find knowledge and community interactions for prediction. To enhance learning progress and outcomes, learning analytics is associated with the gathering, processing, and interpretation of data sources. Learning Analytics predictive models employ any data mining, machine learning, or artificial intelligence technique, including classification, regression, prediction, and others, to evaluate the skills of learners, instructors, and universities.

Even though all attributes may be valuable in some circumstances, just a preferred number of attributes[9] are frequently used for identifying targets. In the KDD CUP 2015 dataset, an activity log is often used to collect several sorts of information regarding learners' behavior. In our research, behavioral variables were utilized to value the regularity of various learner behaviors, and we obtained the learners' parameters and used a variety of machine learning and

*Corresponding Author.

artificial neural network classifiers to produce a significantly high prediction rate.

Establishing feature extraction and selection methods[10], as well as applying an artificial neural network to develop the model in learning analytics. The following are some key aspects of the proposed work:

For effective learning analytics, we proposed the FIAR-ANN model in our research. When association rule mining is utilized for feature selection, the efficiency of the process can be increased.

The relevance of the features is measured using an FP-Growth technique, which then brings them back to the highest values in the features.

The results of our FIAR-ANN model's examination on the KDD CUP 2015 datasets indicate that it is effective. Furthermore, we compare different metrics to the baseline models. Experiments were carried out to verify the research approach.

The layout of successful work in regards discusses past research that focuses on learner behavior to alleviate the problem of learner dropout and details the many aspects of the KDD CUP 2015 dataset in the third part, which includes our processes for cleaning and transforming the raw data to make it suitable for the analysis. In the subsequent sections, we evaluate the gathered data in terms of forecasting learner dropout and compare it to some baseline approaches. Finally, summarize the key achievements and suggest some simple directions for future research.

II. Literature Review

Learners' performance prediction, behavior modeling, conversations, and retention have all been explored using data mining applications in online learning systems. Depending on the research setting, various studies have investigated multiple feature engineering strategies to extract additional sets of features.

The method of identifying a subset of relevant or essential features from raw data is known as feature selection, whereas feature extraction is the process of constructing a new variable from a collection of raw data.

On the KDD Cup 2015 dataset, Yafeng Zheng et al. [11] used the FWTS-CNN (Feature Weighting Time Series—Convolutional Neural Network) dropout prediction model, which used a decision tree to extract attributes out of a learner's record with a time series matrix [12] and then built a model from the weighted features using a convolutional neural network, and it had a better accuracy of over 87 percent.

Jing Chen and colleagues [13] developed DT-ELM (Decision Tree-Enhanced Learning Machine), a novel hybrid method that combines decision trees with extreme learning machines (ELM) that does not depend upon recurrent training. The first module, in particular, creates and extracts a number of attributes from learners' learning behavior records. The decision tree chooses features that can be classified well. It also gives the specified features more weight in order to improve their categorization abilities. MATLAB R2016b and Python 2 are used to carry out the tests. The success of DT-ELM is proved by experimental results on the benchmark KDD2015 dataset, which show that it outperforms various basic ML models on several metrics by a percentile.

Cong Jin et al.'s[14] work starts with a feature extraction approach based on the students' behavioral content. Based on the Support Vector Regression parameters, an improved quantum particle swarm optimization (IQPSO) technique [15]is used to estimate the SDP (student dropout prediction) model. MATLAB 7.0 was charged with supporting the analysis with a 2:1 ratio of training and test subgroups at 10-fold cross-validation. The suggested SDP model outperforms benchmark models such as logistic regression (LR), back propagation (BP), and others, according to experimental results using public data.

As compared to earlier feature selection approaches, Anwar UlHaq et al. [16] designed a unique approach to anticipate greater results utilizing similarity multi-filter feature selection (MFFS). The feature ranking module discovers relevant characteristics, while the clustering module minimizes redundant features. Empirical results from a range of real-time data sources support the hypothesis that merging a feature picked using diverse distributed approaches leads to more resilient extracted features and improves the accuracy rate.

Bo Wei and colleagues [17] suggested a new optimization method assigned to particle swarm optimization with learning memory (PSO-LM). The learning recall policy's goal would be to get significant insights into those who are fitter and develop faster, although the genetic operation is frequently employed to reconcile on a small and large scale. By using the Weka tool to test the model's efficacy, each attribute must be scaled between 0 and 1, and the quality of each component must be assessed using the k-nearest neighbor classifier with 10-fold cross-validation. When compared to wrapper-based feature selection algorithms based on global standard datasets, the analysis revealed that they were more efficient.

Anupam Khan et al. [18] state that the most familiar educational data mining technique for determining pedagogical components of learning and assessing student achievement is association rule mining.

The classification algorithm may be used rigorously to establish a forecast connection, and the interactivity allows learners to correlate behavioral characteristics of their actions to program accomplishments.

In the field of educational data mining, Shaveen Singh et al. [19] explores the use of feature selection approaches combined with association rule mining to identify essential course activities and locate more notable links within these parameters. The task at hand is to come up with the right combination of learning activities that use various methodologies to achieve the course's intended learning results. Subsection formation, subgroup analysis, terminating condition, and outcome checking are the four basic phases included in it. The Communication and Information Literacy dataset UU100 enrolled 2,172 students and included a variety of online actions. There had been no null data in the preprocessed dataset, which had 2172 occurrences with 19 features. To find valuable patterns and correlations among the features, an association rule mining technique is applied. The WEKA tool has been used to do data mining tasks by analyzing acquired data with various algorithms such as Naive Bayes, C4.5, and RBF Network, all of which have high prediction accuracy.

Abeer et al.[20]present a blended strategy for decreasing the high-dimensionality of DNA methylation data and extraction via the Kernel Density Estimation method, resulting in a considerably more accurate and quick-calculating method. The usefulness of the given hybridization technique is evaluated by the metrics of the proposed classifiers such as Naive Bayes, Random Forest, and SVM.

III. Materials and Method

This section delves into the framework of the proposed model for predicting online learning dropouts. Feature extraction and selection techniques are used alone or in combination [21] to improve performance, such as projected accuracy, visualization, and concision of learned content. The benefit of feature selection is that crucial information about a particular feature is preserved. However, only a small number of qualities are needed, and the distinctive features are diverse. Some of them attempt to forecast a learner's performance in binary classes such as dropout or continue the Programme and so on.

A. Problem Definition

The learner dropout rate is the largest myth in MOOC development. Our work identifies the best attributes for predicting dropouts based on a learner's activity and uses a multilayer perceptron to measure accuracy. The available dataset for the KDD CUP 2015 originates from "XuetangX," China's largest MOOC platform[22] and a popular tool for predicting MOOC attrition. As displayed in Table I, the collection contains four catalogs from 79,186 students enrolled in 39 courses.

TABLE I. Data Enactment

Catalog	Depiction
Date	Timespan of each course
Object	Module in a course
log_train	Behavioral record
true_train	Insight into the actual data of the training set's enrollments.

B. Dataset Revelation and Preprocess

The KDD CUP 2015 public dataset, which contains 72,142 records with information about seven learning behaviors, as shown in Table II, including observing visual aids, retrieving objects, learner interaction, laying the course, closing the pages, analytical thinking, and surfing the web, is the exploratory data file used in this article. The binary classifier is used in the outcome analysis, with "0" indicating that students will complete their studies and "1" indicating that they will drop out.

C. Framework

The paper provides a strategy for dropout prediction using a FIAR-ANN model that integrates feature importance and neural networks. As illustrated in Fig. 1, the general stages of this approach are database cleaning, character separation, parameter estimation, and comparison.

The analysis begins with preprocessing the public dataset, then using association rules for feature selection to generate a frequent itemset based on behavioral attributes. The ultimate estimated values are then obtained using multilayer perceptron in the ANN model, and it was assessed using a variety of evaluation criteria.

1) FIAR-ANN approach: Fig. 2 displays the technical aspects of the FIAR-ANN model design, which are divided into two sections in this article: identifying and grading the characteristics in the activity log; deploying the ANN model; and evaluating the performance metrics.

D. Identification and Grading of Features

Our research features were derived from learners' learning activity logs. In an online-learning scenario, association rules can be useful since they can find correlations among distinctive characteristics in a dataset. It is being used to link learners' actions to their results in order to figure out what is influencing their learning chances favorably or adversely.

E. Association Rules

The goal of association rule learning is to establish significant relationships between items in huge datasets [23] using a rule-based machine learning system. Let Ar = (ar_1, ar_2...ar_n) denote a group of n elements, and Tr = (tr_1, tr_2...tr_m) denote a repository of m transactions.

Each transaction tri contains a subset of Ar's available items. A rule is defined as X = Y, where X, Y I, i.e., X and Y (also known as itemsets) are subsets of the accessible items. The precedent and subsequent rules are typically referred to as X and Y, respectively.

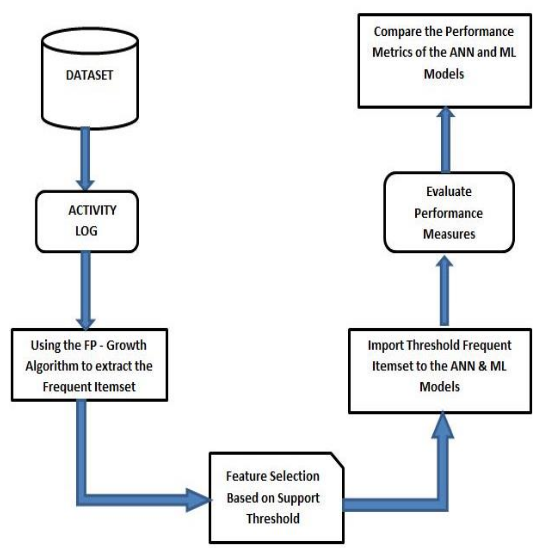

Fig. 1. The Layout of the Proposed Model.

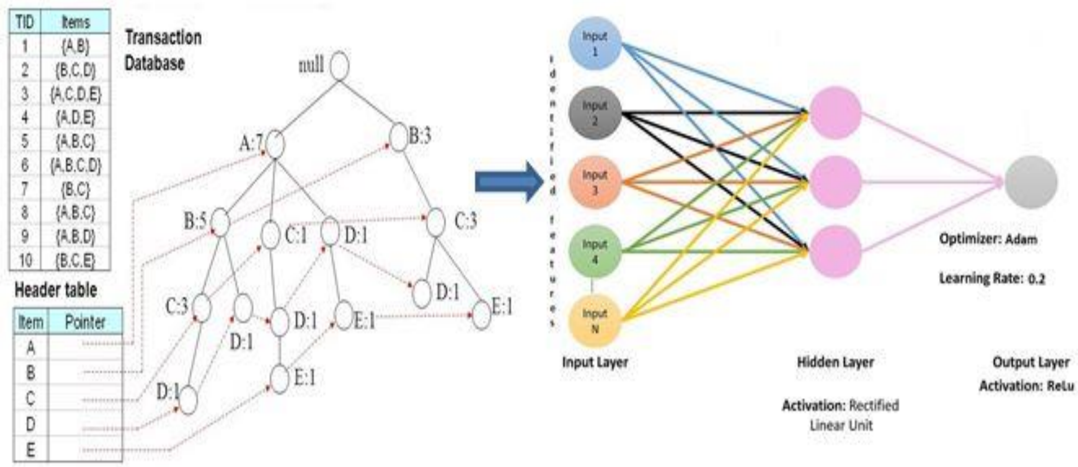

Fig. 2. Architectural Design of FIAR – ANN.

F. Frequent Pattern (FP)-Growth Algorithm

The frequent pattern growth algorithm is an optimized and reliable substitute for the Apriori algorithm, which additionally searches the transaction database for frequent items.

To produce the association rules, this method uses a divide-and-conquer strategy[23]. The approach instead focuses on a data model known as the FP-tree that saves metadata on items and interactions. The transaction database is examined once to generate the FP-tree[21], and the group of candidate itemset F is then calculated and organized in support values.

Numerous measurements were presented to identify the significance and instructiveness of an association rule.

The sections that follow give two measures.

Support:

The support for a set of transactions Tr by an itemset is determined as follows:

$$\text{Support}(X) = |tr \in Tr;\ X \subseteq tr| \,/\, |Tr| \qquad (1)$$

Confidence:

The proportion of the transaction's maximum set of frequent items and X values.

$$\text{Conf}(X \Longrightarrow Y) = \text{Support}\ (X \cup Y) \,/\, \text{Support}\ (X) \qquad (2)$$

G. ANN Model

An Artificial Neural Network is made up of numerous neuron nodes [24] that are divided into three levels: input, output, and hidden layers. In current technological research and development, ANN approaches are most commonly used to identify feature sets that enable the revelation of suitable predictions [25], considering both the maximum of commonly used formal metrics and the understandability of the model's behavior for knowledge extraction from data collection.

We trained on 80% of the feature data set and tested on 20%. The proposed approach is defined as a multilayer-perceptron with single hidden layers, as seen in Fig. 2. The Rectified Linear Unit function is used to activate the neurons in the hidden layer. The input layer contains far more neurons than there are sources in the data set. However, the output layer contains only one neuron with a ReLu activation function, which is suitable for classification problems because it distributes actual content in the range of 0 to 1.

$$\text{relu}(z) = \max(0, z) \qquad (3)$$

The computation costs have been reduced using an optimization strategy. With 10 epochs and a batch size of 10, the Adam optimizer developed to train artificial neural networks was employed and has reached the highest accuracy. Researchers really intended to broaden the outcomes of the research and establish that our method is applicable to future programs that emerge.

IV. Results and Discussion

Our work will look at the key elements of the FIAR-ANN model in this section, starting with creating frequent items using association rules and then inputting the selected parameters into the ANN model. The implementation of the FIAR-ANN model is broken down into two sections in this article: feature extraction and the enhanced ANN model. Fig. 2 depicts the process of implementation.

A. Experimental Framework

1) Software

- Windows 10 is the most popular version of Microsoft's operating system (Intel Core-i3 processor, 64-bit operating system).
- The Google Collaborator is a toolbox for the Scikit-learn Python package[26], and it was used to run the samples on a computer system with seven different sorts of events coming from two separate sources, as indicated in Table II.

TABLE II. DATA SOURCE ATTRIBUTES ARE USED IN THE ANALYSIS

Attributes	Proclamation
enrollment_id	A registered participant's unique identifier.
source	Actions based on the server and browser.
event	We devised seven distinct activities as
	Pbm - Putting intellectual tasks to use
	Video - Participants are observant of the filmed content.
	Access - Getting access to a lot of the learning resources.
	Wiki - Browse the internet for information.
	Discussion - Using discussion boards to exchange intelligence
	Navigate - Exploring several aspects of the software.
	Page_close – Depart from the webpage.
Output	The success (value 0) or failure (value 1) of a participant in a course

2) Baseline models: Our work applied five classic machine learning models as baseline models, especially Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bayes (NB), to offer a point of comparison for the outcomes of the FIAR-ANN model.

3) Logistic regression: The extended linear regression model is the basis of the supervised machine learning classification strategy used in logistic regression. It uses the regression coefficients of one or more components to calculate the likelihood of occurrence.

A strategy for estimating class-based predictor factors (x) is logistic regression [27]. It enables us to determine the possibility (p) of components of a specific class. A binary result is classified as a dropout or non-dropout in this work and it is represented as a 0 or 1.

The classic logistic regression model for evaluating the result of an occurrence, given a variable (x), is $p = 1/ [1 + \exp(-y)]$.

In which $y = b_0+b_1*x$ (4)

The exponential function is exp ().

Given x, p is the probability of an event occurring.

The logistic function for the multiple parameters is as follows: $b_0 + b_1*x_1 + b_2*x_2 +... + b_n*x_n = \log [p/ (1-p)]$.

The predictive beta coefficients are b0 and b1.Increases in x will be proportional to increases in p if b1 is positive. On the other hand, a negative b1 indicates that increasing x could result in a considerable decrease in p.

4) Decision tree: A decision tree leads to increased demand for developing and depicting forecasting tactics. It's easy to learn and implement, and it's widely used for predictive analysis [28]. The basic goal is to divide a massive volume of information into smaller chunks. In predictive analytics, decision trees exhibit the prominent features in the datasets. The tree's root is at the top, with limbs descending. A node is a point on the limb wherever researchers divide the big bunch into shrinking units at every instance. A "leaf" is the term for the end node. In a decision tree, each limb represents a section, and each leaf node reflects the highlight attribute's result within a set range. The decision procedure begins at the root node, checks the related attribute of the item to be sorted, and chooses the outcome depending upon the level till it hits the leaf node.

5) Random forest: Bagging entails the use of many samples instead of a single sample. A trained model is a set of events that can be used to produce predictions. The decision trees' varied results make up the random forest algorithm. The final product will be chosen using a majority-voting procedure. Anomalies, as well as distortion, are less noticeable in Random Forest [29]. The Gini impurity is used for Random Forest class labels to minimize overfitting and bias errors, as well as prediction errors.

6) KNN: In the dropout prediction of this work, we utilize learner interaction within the dataset. In our KNN method, first choose a value for K. Using the Euclidean distance; calculate the distance between k neighbors [30]. Examine all of our neighbors to find which one is closest to our position. Our attribute is assigned to the class with the highest number. KNN looks for correlations between predictors and values within the dataset.

7) Naïve bayes: According to the Naive Bayes classifier [31], the availability of one variable in a class appears to be unconcerned with the appearance of other variables in the same class [32]. It's simple to set up and especially handy for large amounts of data.

The Bayes theorem allows us to derive the posterior probability P (c|x) from P(c), P(x), and P (x|c).

Have a note of the following equation.

P (c|x) denotes the posterior probability of a given class (c, target-dropout) given indicators (x, variables).

P(c) is the class prior probability.

P (x|c) denotes the probability of an indicator given a class.

P(x) is the indicators prior probability.

8) FIAR-ANN-Hyper parameters: The information is screened once for the rapid miner tool, and the set of frequent items, F, is then calculated and arranged in the size of the items with support values using FP-growth, as shown in Table III.

TABLE III. GENERATE A FREQUENT ITEMSET USING FP-GROWTH

Size	Support	Item 1	Item 2	Item 3
1	1	browser:access:UNKNOWN	-	-
1	1	enrollment_id	-	-
1	1	server:problem:problem	-	-
1	0.727	server:access:UNKNOWN	-	-
1	0.704	server:navigate:UNKNOWN	-	-
1	0.701	server:discussion:UNKNOWN	-	-
1	0.576	server:access:chapter	-	-
1	0.338	browser:video:video	-	-
1	0.326	browser:access:sequential	-	-
2	1	browser:access:UNKNOWN	enrollment_id	-
2	1	browser:access:UNKNOWN	server:problem:problem	-
2	0.727	browser:access:UNKNOWN	server:access:UNKNOWN	-
2	0.704	browser:access:UNKNOWN	server:navigate:UNKNOWN	-
2	0.701	browser:access:UNKNOWN	server:discussion:UNKNOWN	-
2	0.576	browser:access:UNKNOWN	server:access:chapter	-
2	0.429	browser:access:UNKNOWN	browser:problem:combinedopenended	-
2	0.338	browser:access:UNKNOWN	browser:video:video	-
2	0.326	browser:access:UNKNOWN	browser:access:sequential	-
2	1	enrollment_id	server:problem:problem	-
2	0.727	enrollment_id	server:access:UNKNOWN	-
2	0.704	enrollment_id	server:navigate:UNKNOWN	-
2	0.701	enrollment_id	server:discussion:UNKNOWN	-
2	0.576	enrollment_id	server:access:chapter	-
2	0.429	enrollment_id	browser:problem:combinedopenended	-
2	0.338	enrollment_id	browser:video:video	-
2	0.326	enrollment_id	browser:access:sequential	-
2	0.727	server:problem:problem	server:access:UNKNOWN	-
2	0.704	server:problem:problem	server:navigate:UNKNOWN	-
2	0.7	server:problem:problem	server:discussion:UNKNOWN	-
2	0.576	server:problem:problem	server:access:chapter	-
2	0.429	server:problem:problem	browser:problem:combinedopenended	-
2	0.338	server:problem:problem	browser:video:video	-
2	0.326	server:problem:problem	browser:access:sequential	-
2	0.671	server:access:UNKNOWN	server:navigate:UNKNOWN	-
2	0.672	server:access:UNKNOWN	server:discussion:UNKNOWN	-
2	0.55	server:access:UNKNOWN	server:access:chapter	-
2	0.429	server:access:UNKNOWN	browser:problem:combinedopenended	-
2	0.32	server:access:UNKNOWN	browser:video:video	-
2	0.315	server:access:UNKNOWN	browser:access:sequential	-
2	0.65	server:navigate:UNKNOWN	server:discussion:UNKNOWN	-
2	0.576	server:navigate:UNKNOWN	server:access:chapter	-
2	0.428	server:navigate:UNKNOWN	browser:problem:combinedopenended	-

2	0.302	server:navigate:UNKNOWN	browser:video:video	-
2	0.326	server:navigate:UNKNOWN	browser:access:sequential	-
2	0.57	server:discussion:UNKNOWN	server:access:chapter	-
2	0.414	server:discussion:UNKNOWN	browser:problem:combinedopenended	-
2	0.314	server:discussion:UNKNOWN	browser:video:video	-
2	0.313	server:discussion:UNKNOWN	browser:access:sequential	-
2	0.374	server:access:chapter	browser:problem:combinedopenended	-
3	1	browser:access:UNKNOWN	enrollment_id	server:problem:problem
3	0.727	browser:access:UNKNOWN	enrollment_id	server:access:UNKNOWN
3	0.704	browser:access:UNKNOWN	enrollment_id	server:navigate:UNKNOWN
3	0.701	browser:access:UNKNOWN	enrollment_id	server:discussion:UNKNOWN
3	0.576	browser:access:UNKNOWN	enrollment_id	server:access:chapter
3	0.429	browser:access:UNKNOWN	enrollment_id	browser:problem:combinedopenended
3	0.338	browser:access:UNKNOWN	enrollment_id	browser:video:video
3	0.326	browser:access:UNKNOWN	enrollment_id	browser:access:sequential
3	0.727	browser:access:UNKNOWN	server:problem:problem	server:access:UNKNOWN
3	0.704	browser:access:UNKNOWN	server:problem:problem	server:navigate:UNKNOWN
3	0.7	browser:access:UNKNOWN	server:problem:problem	server:discussion:UNKNOWN
3	0.576	browser:access:UNKNOWN	server:problem:problem	server:access:chapter
3	0.429	browser:access:UNKNOWN	server:problem:problem	browser:problem:combinedopenended
3	0.338	browser:access:UNKNOWN	server:problem:problem	browser:video:video
3	0.326	browser:access:UNKNOWN	server:problem:problem	browser:access:sequential
3	0.671	browser:access:UNKNOWN	server:access:UNKNOWN	server:navigate:UNKNOWN
3	0.672	browser:access:UNKNOWN	server:access:UNKNOWN	server:discussion:UNKNOWN
3	0.55	browser:access:UNKNOWN	server:access:UNKNOWN	server:access:chapter
3	0.429	browser:access:UNKNOWN	server:access:UNKNOWN	browser:problem:combinedopenended
3	0.32	browser:access:UNKNOWN	server:access:UNKNOWN	browser:video:video
3	0.315	browser:access:UNKNOWN	server:access:UNKNOWN	browser:access:sequential
3	0.65	browser:access:UNKNOWN	server:navigate:UNKNOWN	server:discussion:UNKNOWN
3	0.576	browser:access:UNKNOWN	server:navigate:UNKNOWN	server:access:chapter
3	0.428	browser:access:UNKNOWN	server:navigate:UNKNOWN	browser:problem:combinedopenended
3	0.57	browser:access:UNKNOWN	server:discussion:UNKNOWN	server:access:chapter
3	0.414	browser:access:UNKNOWN	server:discussion:UNKNOWN	browser:problem:combinedopenended
3	0.314	browser:access:UNKNOWN	server:discussion:UNKNOWN	browser:video:video
3	0.313	browser:access:UNKNOWN	server:discussion:UNKNOWN	browser:access:sequential
3	0.374	browser:access:UNKNOWN	server:access:chapter	browser:problem:combinedopenended
3	0.727	enrollment_id	server:problem:problem	server:access:UNKNOWN
3	0.704	enrollment_id	server:problem:problem	server:navigate:UNKNOWN
3	0.7	enrollment_id	server:problem:problem	server:discussion:UNKNOWN
3	0.576	enrollment_id	server:problem:problem	server:access:chapter
3	0.429	enrollment_id	server:problem:problem	browser:problem:combinedopenended
3	0.338	enrollment_id	server:problem:problem	browser:video:video
3	0.326	enrollment_id	server:problem:problem	browser:access:sequential
3	0.671	enrollment_id	server:access:UNKNOWN	server:navigate:UNKNOWN

3	0.672	enrollment_id	server:access:UNKNOWN	server:discussion:UNKNOWN
3	0.55	enrollment_id	server:access:UNKNOWN	server:access:chapter
3	0.429	enrollment_id	server:access:UNKNOWN	browser:problem:combinedopenended
3	0.32	enrollment_id	server:access:UNKNOWN	browser:video:video
3	0.315	enrollment_id	server:access:UNKNOWN	browser:access:sequential
3	0.65	enrollment_id	server:navigate:UNKNOWN	server:discussion:UNKNOWN
3	0.576	enrollment_id	server:navigate:UNKNOWN	server:access:chapter
3	0.428	enrollment_id	server:navigate:UNKNOWN	browser:problem:combinedopenended
3	0.302	enrollment_id	server:navigate:UNKNOWN	browser:video:video
3	0.326	enrollment_id	server:navigate:UNKNOWN	browser:access:sequential
3	0.57	enrollment_id	server:discussion:UNKNOWN	server:access:chapter
3	0.414	enrollment_id	server:discussion:UNKNOWN	browser:problem:combinedopenended
3	0.314	enrollment_id	server:discussion:UNKNOWN	browser:video:video
3	0.313	enrollment_id	server:discussion:UNKNOWN	browser:access:sequential
3	0.374	enrollment_id	server:access:chapter	browser:problem:combinedopenended
3	0.671	server:problem:problem	server:access:UNKNOWN	server:navigate:UNKNOWN
3	0.672	server:problem:problem	server:access:UNKNOWN	server:discussion:UNKNOWN
3	0.55	server:problem:problem	server:access:UNKNOWN	server:access:chapter
3	0.429	server:problem:problem	server:access:UNKNOWN	browser:problem:combinedopenended
3	0.32	server:problem:problem	server:access:UNKNOWN	browser:video:video
3	0.315	server:problem:problem	server:access:UNKNOWN	browser:access:sequential
3	0.65	server:problem:problem	server:navigate:UNKNOWN	server:discussion:UNKNOWN
3	0.576	server:problem:problem	server:navigate:UNKNOWN	server:access:chapter
3	0.428	server:problem:problem	server:navigate:UNKNOWN	browser:problem:combinedopenended
3	0.302	server:problem:problem	server:navigate:UNKNOWN	browser:video:video
3	0.326	server:problem:problem	server:navigate:UNKNOWN	browser:access:sequential
3	0.569	server:problem:problem	server:discussion:UNKNOWN	server:access:chapter
3	0.414	server:problem:problem	server:discussion:UNKNOWN	browser:problem:combinedopenended
3	0.314	server:problem:problem	server:discussion:UNKNOWN	browser:video:video
3	0.313	server:problem:problem	server:discussion:UNKNOWN	browser:access:sequential
3	0.374	server:problem:problem	server:access:chapter	browser:problem:combinedopenended
3	0.622	server:access:UNKNOWN	server:navigate:UNKNOWN	server:discussion:UNKNOWN
3	0.55	server:access:UNKNOWN	server:navigate:UNKNOWN	server:access:chapter
3	0.428	server:access:UNKNOWN	server:navigate:UNKNOWN	browser:problem:combinedopenended
3	0.315	server:access:UNKNOWN	server:navigate:UNKNOWN	browser:access:sequential
3	0.545	server:access:UNKNOWN	server:discussion:UNKNOWN	server:access:chapter
3	0.414	server:access:UNKNOWN	server:discussion:UNKNOWN	browser:problem:combinedopenended
3	0.306	server:access:UNKNOWN	server:discussion:UNKNOWN	browser:video:video
3	0.302	server:access:UNKNOWN	server:discussion:UNKNOWN	browser:access:sequential
3	0.374	server:access:UNKNOWN	server:access:chapter	browser:problem:combinedopenended
3	0.57	server:navigate:UNKNOWN	server:discussion:UNKNOWN	server:access:chapter
3	0.412	server:navigate:UNKNOWN	server:discussion:UNKNOWN	browser:problem:combinedopenended
3	0.313	server:navigate:UNKNOWN	server:discussion:UNKNOWN	browser:access:sequential
3	0.374	server:navigate:UNKNOWN	server:access:chapter	browser:problem:combinedopenended

After preprocessing, the dataset contains 72,142 tuples, and the model must be built by selecting features with a minimum support value of greater than 0.5 and a maximum itemset size of 3, as shown in Table IV.

In Table V demonstrates the ReLu activation functions, which are present in the hidden and output layers, including the Adam optimization method and a learning rate of 0.2 with a 10 epoch rate, which is applied to ANN approaches.

B. Evaluation Metrics

The performance measures are evaluated metrics[33]like accuracy, precision, recall, and F1 – score, Training & Test Score. In this analysis, the class label is used as a binary classification method.

C. Comparison and Analysis with the baseline Model

ML approaches are used in the majority of dropout forecasts. The values in Table VI demonstrate the prediction accuracy of the overall learner behavioral features [26], and in Table VII, the selected features by our technique are significantly better than those of the benchmark method, indicating that our feature extraction method is effective.

TABLE IV. SELECTED LEARNER ACTIVITY PARAMETERS

Features	Description
enrollment_id	The learners' unique enrollment number.
server:problem: problem	The number of issues that the server is encountering determines the behavior.
server:access: UNKNOWN	The behavior is defined by the number of requests received from the server.
server:navigate: UNKNOWN	The size of server-based navigations to other regions of the course is used to calculate the behavior.
server:discussion: UNKNOWN	The behavior is determined by the number of users who access the course forum from the server.
browser:access: UNKNOWN	The learner's behavior is defined by the amount of browser accesses obtained.
Output	Label of the dataset

TABLE V. APPLIED CONSTRAINTS

Constraints	Implications
Learning rate	0.2
Epochs	10
Activation function	ReLu
Optimizer	Adam

TABLE VI. USING BASELINE MODELS, CONTRAST THE OVERALL LEARNER BEHAVIORAL RESULTS

S.No	Metrics/ Learning Method	Accuracy	Precision	Recall	F1-score
1	LR	0.78	0.78	1	0.88
2	DT	0.78	0.78	1	0.88
3	RF	0.78	0.78	1	0.88
4	KNN	0.79	0.79	0.99	0.88
5	NB	0.78	0.78	1	0.9

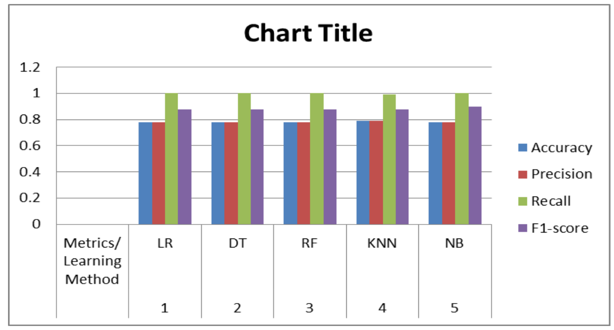

Fig. 3. The Results of Employing Several Classifiers to Predict Overall Learner Behavior.

Fig. 3 shows that the results of several basic supervised machine learning algorithms are used as input data for the whole learner's behavioral actions. In comparison to other models, the KNN model achieves the best accuracy rate of 79%.

To begin, we compared the models to a preset baseline model using the FIAR-ANN model. Table VII shows the findings for the various machine learning algorithms used in the research. The experimental findings with the best values are indicated as strong.

Compare the accuracy level with the selected features by using FIAR-ANN method, entire learner behavior and whole attributes in the dataset as 92%, 78% and 72%. Therefore the vital features produces the best result related to the others.

In terms of the four metric values, the FIAR-ANN model outperforms the other five models. Fig. 4 depicts the evaluation metrics for each model.

TABLE VII. RESULTS IN HYPER PARAMETERS COMPARED TO THE BASELINE MODEL

S.No	Metrics/ Learning Method	Accuracy	Precision	Recall	F1-score
1	LR	0.85	0.86	0.96	0.95
2	DT	0.83	0.84	0.97	0.93
3	RF	0.84	0.86	0.95	0.94
4	KNN	0.86	0.88	0.95	0.95
5	NB	0.84	0.84	0.97	0.8
6	**FIAR-ANN**	**0.92**	**0.93**	**0.99**	**0.91**

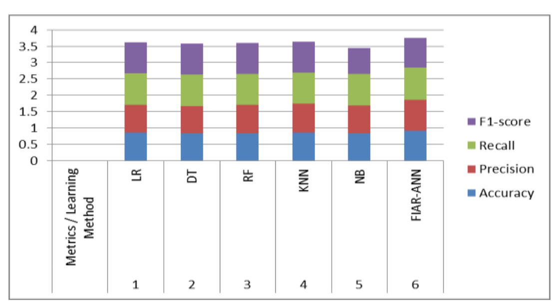

Fig. 4. The Extracted Feature Results are Compared to the Baseline Model's Results.

The findings show that certain machine learning models and the selecting input described in this work are better suited to predicting the dropout problem in MOOCs than others.

According to the findings.

1) In the large-data MOOC, the FIAR-ANN model developed in this research focuses on solving the dropout prediction problem and improving the baseline technique.

2) Using learner behavioral data from the KDD CUP 2015 dataset, the FIAR-ANN model, which has 92 percent accuracy, can be used to predict dropout rates for new programs.

V. Conclusion

Researchers devised a number of methods to predict learner dropout in online programs. From primary behavioral data, we identify and retrieve a number of interpretive behavior aspects. The frequent candidate itemset is generated using an association rule mining-FP growth method. The itemset contains the most often observed learner behavior. Then select the parameters that are present in the three most common items. An artificial neural network approach is applied for the evaluation of the selected parameters. Our proposed method for assessing the efficacy of the KDD CUP 2015 dataset parameters and their values was applied to several machine learning approaches. Then compare the ANN and ML techniques' performance measures. The FIAR-ANN model incorporates the consequences of behavioral characteristics of dropouts, promptly enhancing the dropout prediction accuracy. In this work, we are limited to evaluate learner behavioral activities on the computer related courses , but enhance the work in future with other essential characteristics and also in other discipline courses. Put more emphasis on the attributes that are relevant to different types of platforms in the future.

References

[1] S. Nithya and S. Umarani, “A Novel Analysis of MOOC’s Prediction,” vol. 17, no. 10.

[2] E. W. Winarni, E. P. Purwandari, and S. Hafiza, “Automatic Essay Assessment for Blended Learning in Elementary School,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 12, no. 1, pp. 85–91, 2022, doi: 10.18517/ijaseit.12.1.11835.

[3] K. Seery, A. Barreda, S. Hein, and J. Hiller, “Retention strategies for online students: A systematic literature review,” J. Glob. Educ. Res., vol. 5, no. 1, pp. 72–84, Jun. 2021, doi: 10.5038/2577-509X.5.1.1105.

[4] Z. Xie, “Bridging MOOC Education and Information Sciences: Empirical Studies,” IEEE Access, vol. 7, pp. 74206–74216, 2019, doi: 10.1109/ACCESS.2019.2921009.

[5] S. Ranjeeth, T. P. Latchoumi, and P. V. Paul, “A Survey on Predictive Models of Learning Analytics,” Procedia Comput. Sci., vol. 167, no. 2018, pp. 37–46, 2020, doi: 10.1016/j.procs.2020.03.180.

[6] M. M. Rahman, Y. Watanobe, R. U. Kiran, T. C. Thang, and I. Paik, “Impact of Practical Skills on Academic Performance: A Data-Driven Analysis,” IEEE Access, vol. 9, pp. 139975–139993, 2021, doi: 10.1109/ACCESS.2021.3119145.

[7] R. Eusoff, A. M. Zin, and S. M. Salleh, “A Flipped Classroom Framework for Teaching and Learning of Programming,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 12, no. 2, pp. 539–549, 2022, doi: 10.18517/ijaseit.12.2.14909.

[8] F. Hlioui, N. Aloui, and F. Gargouri, “A withdrawal prediction model of at-risk learners based on behavioural indicators,” Int. J. Web-Based Learn. Teach. Technol., vol. 16, no. 2, pp. 32–53, 2021, doi: 10.4018/IJWLTT.2021030103.

[9] J. Shobana and M. Murali, “Adaptive particle swarm optimization algorithm based long short-term memory networks for sentiment analysis,” J. Intell. Fuzzy Syst., vol. 40, no. 6, pp. 10703–10719, 2021, doi: 10.3233/JIFS-201644.

[10] G. Gavisiddappa, S. Mahadevappa, and C. M. Patil, “Multimodal biometric authentication system using modified relief feature selection and multi support vector machine,” Int. J. Intell. Eng. Syst., vol. 13, no. 1, pp. 1–12, 2020, doi: 10.22266/ijies2020.0229.01.

[11] Y. Zheng, Z. Gao, Y. Wang, and Q. Fu, “MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series,” IEEE Access, vol. 8, pp. 225324–225335, 2020, doi: 10.1109/ACCESS.2020.3045157.

[12] K. Sharma, K. Mangaroska, N. van Berkel, M. Giannakos, and V. Kostakos, “Information flow and cognition affect each other: Evidence from digital learning,” Int. J. Hum. Comput. Stud., vol. 146, no. September 2020, p. 102549, 2021, doi: 10.1016/j.ijhcs.2020.102549.

[13] J. Chen, J. Feng, X. Sun, N. Wu, Z. Yang, and S. Chen, “MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine,” Math. Probl. Eng., vol. 2019, 2019, doi: 10.1155/2019/8404653.

[14] C. Jin, “MOOC student dropout prediction model based on learning behavior features and parameter optimization,” Interact. Learn. Environ., vol. 0, no. 0, pp. 1–19, 2020, doi: 10.1080/10494820.2020.1802300.

[15] J. Shobana and M. Murali, “An efficient sentiment analysis methodology based on long short-term memory networks,” Complex Intell. Syst., vol. 7, no. 5, pp. 2485–2501, 2021, doi: 10.1007/s40747-021-00436-4.

[16] A. U. Haq, D. Zhang, H. Peng, and S. U. Rahman, “Combining Multiple Feature-Ranking Techniques and Clustering of Variables for Feature Selection,” IEEE Access, vol. 7, pp. 151482–151492, 2019, doi: 10.1109/ACCESS.2019.2947701.

[17] B. Wei, W. Zhang, X. Xia, Y. Zhang, F. Yu, and Z. Zhu, “Efficient Feature Selection Algorithm Based on Particle Swarm Optimization with Learning Memory,” IEEE Access, vol. 7, pp. 166066–166078, 2019, doi: 10.1109/ACCESS.2019.2953298.

[18] A. Khan and S. K. Ghosh, Student performance analysis and prediction in classroom learning: A review of educational data mining studies, vol. 26, no. 1. Education and Information Technologies, 2021.

[19] S. Singh and S. P. Lal, “Using feature selection and association rule mining to evaluate digital courseware,” Int. Conf. ICT Knowl. Eng., 2013, doi: 10.1109/ICTKE.2013.6756286.

[20] A. A. Raweh, M. Nassef, and A. Badr, “A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation,” IEEE Access, vol. 6, pp. 15212–15223, 2018, doi: 10.1109/ACCESS.2018.2812734.

[21] S. Umarani, T. R. Chaithanya, and M. Divya, “Deployment of P-Cycle in Optical Networks : A Data Mining Approach Deployment of P-Cycle in Optical Networks : A Data Mining Approach,” no. September 2016, 2019.

[22] “KDD CUP 2015 Dataset.” https://data-mining.philippe-fournier-viger.com/the-kddcup-2015-dataset-download-link/.

[23] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, “Relationship between student engagement and performance in e-learning environment using association rules,” EDUNINE 2018 - 2nd IEEE World Eng. Educ. Conf. Role Prof. Assoc. Contemp. Eng. Careers, Proc., pp. 1–6, 2018, doi: 10.1109/EDUNINE.2018.8451005.

[24] A. Zuiderwijk, Y. Chen, and F. Salem, “Implications of the use of artificial intelligence in public governance : A systematic literature review and a research agenda,” Gov. Inf. Q., no. March, p. 101577, 2021, doi: 10.1016/j.giq.2021.101577.

[25] C. H. Yu, J. Wu, and A. C. Liu, “Predicting learning outcomes with MOOC clickstreams,” Educ. Sci., vol. 9, no. 2, 2019, doi: 10.3390/educsci9020104.

[26] S. Nithya. and Dr.S.Umarani, “Comparative Analysis of the Learning on KDD Cup 2015 Dataset,” Webology, vol. 19, no. 1, pp. 705–717, 2022, doi: 10.14704/web/v19i1/web19050.

[27] D. Peng and G. Aggarwal, “Modeling MOOC Dropouts,” Entropy, vol. 10, no. 114, p. 49944, 2015.

[28] M. Şahin, “A Comparative Analysis of Dropout Prediction in Massive Open Online Courses,” Arab. J. Sci. Eng., 2020, doi: 10.1007/s13369-020-05127-9.

[29] S. Dass, K. Gary, and J. Cunningham, “Predicting student dropout in self-paced mooc course using random forest model,” Inf., vol. 12, no. 11, 2021, doi: 10.3390/info12110476.

[30] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, “Analyzing and predicting students’ performance by means of machine learning: A review,” Appl. Sci., vol. 10, no. 3, 2020, doi: 10.3390/app10031042.

[31] D. Romahadi, A. A. Luthfie, W. Suprihatiningsih, and H. Xiong, “Designing Expert System for Centrifugal using Vibration Signal and Bayesian Networks,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 12, no. 1, pp. 23–31, 2022, doi: 10.18517/ijaseit.12.1.12448.

[32] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Munoz-Merino, and C. D. Kloos, “Prediction in MOOCs: A Review and Future Research Directions,” IEEE Trans. Learn. Technol., vol. 12, no. 3, pp. 384–401, 2019, doi: 10.1109/TLT.2018.2856808.

[33] 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, 2020, doi: 10.26599/TST.2019.9010013.