Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 27 No. 3 (2024)

Hybrid Machine Learning Approach using Real Time Data Processing for Educational Data Mining

[1]Mr. Mukul V Khasne, [2]Dr. Tripti Arjariya

[1]Ph. D. Scholar, Department of Computer Science and Engineering, Bhabha University, Bhopal

[2]Professor & Principal, Bhabha Engineering and Research Institute, Bhabha University, Bhopal

Article History:

Received: 05-03-2024

Revised: 30-04-2024

Accepted: *25-05-2024*

Abstract

Educational data produced by various systems, such as e-learning, student admission process, online student exam, and automated significance management solutions, may be successfully analyzed using educational data mining (EDM) methods to provide beneficial for the academic student's progress. A highly desired use of EDM is the prediction of standardized testing from previous academic data. In this intelligence, it is critical to establish an automated approach for predicting student success. The majority of existing research on academic prediction problem use traditional feature extraction and illustration approaches, in which retrieved characteristics are feed to machine learning classifier. Supervised learning has recently allowed academics to extract high-level characteristics from raw data automatically. The proposed enhanced feature representation approaches allow for better performance on real time heterogenous data. In this paper, we demonstrate the Long Short-Term Memory (LSTM) with Recurrent Neural Network (RNN) called as RNN-LSTM. By examining current research difficulties based on evolving feature classification and prediction, we applied the most sophisticated LSTM paired with an attentiveness process approach in this work. Academicians, institutions, and governmental organizations benefit from this study since it allows them to anticipate achievement early. In extensive experimental analysis proposed work compared with numerous state-of-the-art, the better recurrent neural network functionalities of LSTM matching with the hybrid mechanism give higher performance. Predictive performance of 90.50% has achieved using the proposed algorithm.

Keywords: Educational Data Mining, Supervised Learning , Feature Extraction, Coorelation mapping.

1. Introduction

In EDM approaches, many academics have looked at assessing student performance based on previous data. These studies focus on predicting student success in terms of scores, degrees, and pass/fail early on. However, attributed to the aforementioned fundamental constraints associated with prior efforts, predicting student academic achievement from noisy and huge datasets is a difficult task: utilizing machine learning approaches feature subset traditional modelling schemes preceded by a classifier, and (ii) inadequate selection of predictor variables reflecting student outcomes. The authors investigated the use of historical data to predict student performance using a machine learning-based

method. Different ML classifiers were utilized in the baseline investigation to estimate student quality in binary classifications, pass or fail. Predicting performance in terms of pass or fail, on the other hand, does not give a deeper understanding of the achievement of students. Another key flaw in their method is that it fails to evaluate the overall interdependence of predictor factors in student data. As a result, traditional machine learning classifiers are ineffective for estimating student achievement from academic data. Various types of assaults are launched against computer infrastructures. Because of its nature for sharing resources across users, the cloud environment suffers even more. In a cloud setting, a virtual machine shares hardware resources with some other virtual machines. In the cloud, a container even runs on the same operating system as other containers. Despite using isolation solutions, certain common resources, whether software or hardware, remain, potentially exposing the whole ecosystem to danger. In this part, we'll go through some of the threats that threaten a cloud environment and how they were employed in this study to produce aberrant data for testing purposes.

1.1 Background of EDM

Business intelligence is crucial and has a lot of promise for helping organizations concentrate on a most critical information in their database systems. In data processing and data analyses, software tools and techniques have been created to analyze data and make it more helpful to users who require it for decision-making. Techniques including classification, connection, regression, segmentation, forecasting, estimate, and visualization have been used in data analytics. Academic institutions may utilize categorization to analyze student performance or estimate the likelihood of several outcomes, including result, persistence, and course completion. Education systems may use data analytics to distribute resources and employees better, manage student performance more effectively, and increase the efficiency of their education administration systems.

1.2 Motivation

To address the limitations of the baseline research, we apply an enhanced feature extraction as well as selection approach followed by a deep neural network approach that has been effectively used in a variety of applications, including report recognition, intemperate association identification, and other areas. For student grade prediction, we suggest using feature extraction and the LSTM model.

i) It works as follows: co-relational functionalities extract an adequate high ranked characteristic creating a significant role in student score estimation in the attribute selection module, TF-IDF;

ii) LSTM has contemplated both past and prospective contextual information; and

iii) The hybrid feature selection method has utilized for essential features from the provided student data.

As a result, the suggested approach uses the enhanced feature selection and LSTM functions and the consideration layer to forecast scholars' ultimate results based on their previous hypothetical recital.

The residue of the paper is prearranged into the subsequent sections: Section 2 discusses many current EDM with machine learning approaches developed by earlier researchers. Section 3 describes an proposed system implementation and methodology design with algorithm. The section 4 demonstrates result and discussion with detail description of experimental setup including software and hardware requirements, configuration environment etc. In final section 5 we reviewed conclusion and future

work of proposed system with enhancement of future direction.

2 Review of Literature

The EDM is most essential for predict the students' performance to evaluate the future prediction. This section we discussed various existing systems developed by previous authors.

2.1 Existing Methodologies

For forecasting academic achievement, Inayat Khan et al. [1] proposed a deep learning-based technique. This study analyses data to forecast academic achievement (grades) that used a deep neural network model called the consideration Bidirectional Long Short-Term Memory (BiLSTM) infrastructure. This approach has utilised the most powerful BiLSTM combined with an attentiveness mechanism model to examine recent research difficulties centred on comprehensive feature classifiers. Academicians, institutions, and government agencies benefit from this study since it allows them to anticipate performance early. Compared to the current state-of-the-art, BiLSTM's better sequence learning capabilities paired with its attention mechanism produce greater performance. On actual streaming data, this approach obtained an accuracy of roughly 91.00 per cent.

Zaira-Jazmn Zárate-Santana et al. [2] Zaira-Jazmn Zárate-Santana et al. This research looked at how deep, and surface techniques dealing with academic stress and gender are connected. One thousand twelve university students responded to an online survey. Multivariate canonical correspondence analysis investigated the association between gender, sources of stress, and learning techniques. The work's unique feature is that it is the first to give a combined multivariate visual depiction of learning methodologies and dealing with academic stress, both of which have received little prior investigation. In terms of practical applications, the findings might be utilized to propose design methods and planning procedures in sustainability courses tailored to student profiles.

In an educational situation, Manus Ross et al. [3] establish the framework for constructing a system to effectively categories a student as interested or indifferent. Depending on the customer RGB-D sensor data, they employ machine learning like K-means as well as SVM to identify kids as active or uninterested. It's utilized to help teachers at all levels improve their teaching tactics and deploy individualized learning systems, which is a National Academy of Engineering Grand Challenge. This study uses machine learning algorithms in a classroom context. Instructors may utilize the data from these algorithms to get vital feedback on the success of their teaching tactics and pedagogies. Instructors may utilize this information to enhance their teaching methods, and students will benefit from increased learning and topic mastery. In the end, this will improve the students' capacity to work in their respective fields. In general, this activity may aid efforts in a variety of areas of education and teaching.

Dr. A. Senthil Kumar and K. Joshna [4] emphasize the need for data analytics in educational systems and provide various ways to meet these demands. Understanding the various components and their functions are required while implementing any system. Educational data analytics offers the capacity to uncover, evaluate, and anticipate useful information from educational data, allowing education management systems to plan, execute, and predict the future with more flexibility.

Eyman Alyahyan et al. [5] provide a clear set of rules for utilizing EDM to predict success. Although the research was confined to undergraduate students, the same ideas may readily be applied to graduate students. It was developed for someone who is inexperienced to data mining, computer vision, and AI. When it comes to predicting individuals' academic performance, which has been measured as student ability, the most common components reported in the study were prior educational performance, student characteristics, learning materials, and psychiatric qualities. Various machine learning algorithms have been used to predict student outcomes in terms of prediction strategies. Using the categorization strategy, you'll have a lot of success.

Ijaz Khan al. [6] assess the usefulness of machine learning algorithms in monitoring students' academic progress and alerting instructors to students who are at danger of receiving a poor grade in a course. Furthermore, the prediction model is changed into a visible shape, allowing the teacher to quickly plan the appropriate preventative actions. With the help of several machine learning techniques, it created a collection of prediction models. Decision trees outperform other models and are therefore translated into an easily understandable structure. The research's ultimate product is a series of supporting measures to closely monitor students' performance from the beginning of the course, as well as a set of preventative measures to provide extra attention to problematic students.

The phenomena of synthetic intelligence's increase in university education teaching process is investigated by Stefan A. D. Popenici et al. [7]. It examines how modern technology affects how students understand and how educational establishments educate and adapt in the curriculum. Recent technical achievements and the rapid rate of technological in postsecondary learning are studied in order to foresee the future nature of postsecondary learning in a world where machine learning is interwoven in the fabric of these institutions. It outlines particular challenges for universities and colleges and classroom instruction when implementing these innovations for educating, research, student aid, and management.

Sundaresan Bhaskaran et al. [8] employ divide - and - conquer approach segmentation to construct an intelligent optimization technique that adapts to the requirements, interests, and degrees of skill of the students autonomously. The recommender studies and learns various methods and abilities of the students dynamically. To comprehend the different types of learners, divide as well as conquer strategic objective clustering is applied. The recommended loose collection linear information retrieval technique is utilised to extract the students' operational patterns. Based on the assessments of commonly encountered combinations, the system then generates intelligent recommendations.

Ajibola O. Oyedeji and colleagues [9] analyzing students' educational performance is crucial for higher education institutions and instructors to discover ways to enhance individual student performance. The project analyzed previous student results as well as their attributes such as age, regional distribution, family background, and study attitude, and used machine learning techniques to put this knowledge to the test. Three models were evaluated using the test and train data: Neural networks, regression analysis for supervised methods, and predicted values with machine learning. The best mean average error for supervised learning is linear regression.

The difficulty of measuring individual students' shifting knowledge states as they go through online courses is addressed by Robin Schmucker et al. [10]. This knowledge tracing issue, also known as student performance modelling, is an important step in developing adaptive online teaching systems.

It investigates how to use diverse kinds and vast volumes of log data from previous students to construct reliable machine learning models that predict future student knowledge states. This is the first research to leverage four big sets of student data recently made public by four different intelligent tutoring systems.

A conceptual framework presented by Maritza Mera-Gaona et al. [11] was proposed to elucidate the most important ideas in the feature selection process. The goal was to determine how to increase the classification algorithms' performance using the feature sets that train them. The qualitative approach used to develop the conceptual framework allows for identifying ideas and connections that explain the FS process and the consensus among various FS methodologies via a literature review. The authors were able to steer the creation of an app using the conceptual framework they created. Implementation framework with an ensemble of FS approaches for picking features. The chosen process is a collection of relevant characteristics that performs better than single algorithms in categorizing sets of features.

On the Open University (OU) dataset, Fatema Alnassar et al. [12] use three machine learning methods (Support Vector Classifier (SVC), k-Nearest Neighbour (k-NN), and Artificial Neural Network (ANN)) to solve the issue of student performance prediction. Three primary variables are examined in educational data: demographics, engagement, and performance. Compared to other approaches and current literature, the k-NN strategy emerged as the best for OU studies in the experimental study.

2.2 GAP Analysis

- IT produces low classification accuracy when data is real time complex such multi values attributes.
- Most of supervised classifiers works with single feature selection approaches due to this

3 Proposed system design

This work proposed student performance prediction using hybrid deep learning approach. Numerous feature extraction d selection techniques have been used for generate the background knowledge in training and testing respectively. RNN-LSTM is the classification algorithm have been used for predict the future performance from entire dataset using given techniques. We describe in details of this system in result section with numerous dataset

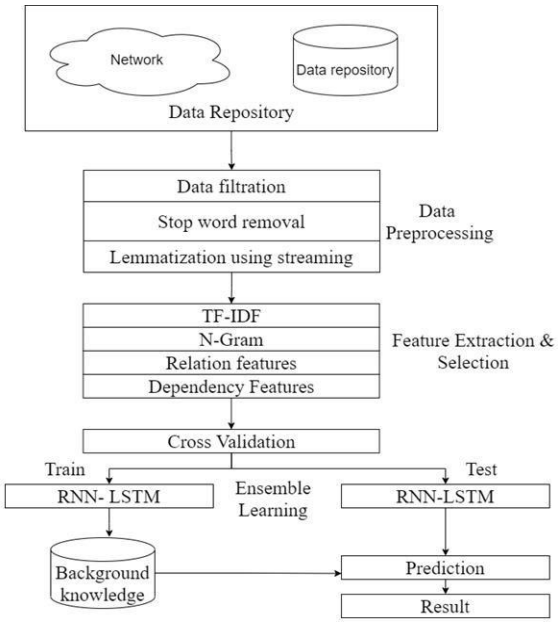

Figure 1: Proposed system architecture

The Figure 1 illustrates the execution of the proposed system with each layer. First, we collect data from various sources such as numerous web applications, real-time students' data and some synthetic data set from multiple social media sources. The data should be imbalanced, or it contains some miss classified instances. To eliminate such problems, it needs to pre-process and normalize before training. Numerous data filtration techniques have been used and data acquisition processes for removing null values. The stop word removal and lemmatization have been used for data balancing. This data is considered a normalized data set used for training and testing, respectively. The four different feature extraction techniques have been carried out to extract the features from the input data set, such as relational features, dependency feature TF-IDF and N-gram features etc. The extracted features feed to train Network and generator background knowledge concerning the algorithm. The proposed model utilizes RNN-LSTM as a classifier for model training and testing. A similar process has been followed for model testing and predicting the possibility according to the given data set.

4 Algorithm design

Input: Code line snippet that consist Term[1…..n]

Output: calculation of TF-IDF for each record

Step 1: Data_instance = {Token1, Token2, Token3…. Tokenn}

Step 2: for each (T into Data_Vector)

Calculate the Tf weight for T

Step 4: tf_score(t,d) = (term,document)

T_count=particular term

d_count = the term found in a entire documents

Step 5: idf_score = T_count □ sum(d_count)

Step 6: return (tf_score * idf_score)

Module Training

Input: Train_DB[ ] as training dataset, set of activation function AF[].

Output: Trained module in .PKL file for entire splited dataset

Step 1: Initialize the both algorithms Train_DB[], AF[], Iteration as epoch_size

Step 2: Extracted_Features_Set □ Extract_Features(Train_DB[])

Step 3: Selecetd_Features_Set [] □ optimization(Extracted_Features_Set)

Step 4: Train.pkl □ Build_Classifier(Selecetd_Features[])

Step 5: Return Train.pkl

Execution of Testing

Module Testing

Input: Test_DB [] as testing instance set or individual patient record, Training Background Knowledge Train.pkl, User defines threshold Th

Output: Output_Map <Predicted_class_label, Similarity_weight> optimized instance recommends by classifier.

Step 1: Read all testing records by using below equation

test_Feature (m)

$$= \sum_{m=0}^{n} (.\ \text{feature\_Set[A[i] ..... A[n]} \leftarrow \text{Test\_DB})$$

Step 2: Extract selected attribute features from entire test record $testFeature(m)$ $testFeature(m)$ by utilizing beneath function.

$$Extracted\_Feature\_Set\_x[t...n] = \sum_{x=1}^{n}(t) \leftarrow test\_Feature \sum_{x=1}^{n}(t) \leftarrow test\_Feature(m)$$

The feature vector is the set of extracted hybrid features from given input

Step 3: Extract all training instance from trained modules using below function

train_Feature (m)

$$= \sum_{m=1}^{n} (.\ \text{feature\_Set[A[i] ..... A[n]} \leftarrow \text{Train.pkl})$$

Step 4: feed the test instances or record set to testing classifier as $testFeature(m)$ $testFeature(m)$ using below equation.

$$Extracted\ Feature\_Set\_x[t.....n] = \sum_{x=1}^{n}(t) \leftarrow test\_Feature \sum_{x=1}^{n}(t) \leftarrow test\_Feature(m)$$

The Extracted_Feature_Set_x[t] comprehends feature vector for entire class labels.

Step 5: Now validate separately testing instance with all train features

$$Calc\_weight = CalcSim\ (Feature\_Set\_x \mid\mid \sum_{i=1}^{n} Feature\_Set\_y[y] \sum_{i=1}^{n} Feature\_Set\_y[y]\ )$$

Step 6: *Return calc_wSeight*

5 Results and Discussions

To validate the evaluation of the proposed bug forecast procedure, we decided on RNN classification algorithms utilized for fault prediction, including unlabelled datasets. The distribution of data, as seen in the Table 2, is used to evaluate various machine learning ability, and it contains measurements of accuracy, recall, and F-score [4].

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

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

$$F - Measure = \frac{2*Precision*Recall}{Recall+Precision} \frac{2*Precision*Recall}{Recall+Precision}$$

The performance evaluation has done of system with 10 synthetic datasets, using proposed classification techniques. In below Table 2 we demonstrate a number bug instances have detected in respective dataset with percentage of bug records in each dataset.

Experiment using RNN-LSTM (sigmoid)

In this experiment, we use the real-time student dataset to demonstrate RNN (Sigmoid) classification accuracy. Figure 2 illustrates the outcomes of similar trials using various cross-validation techniques. According to the findings, 15-fold cross-validation has the highest average classification accuracy of 95.10%.

The 5-fold cross validation also achieves 93.6% with RNN with sigmoid function. While Figure 2 describes with 10-fold data cross validation. Both functions achieve around similar accuracy during module testing.

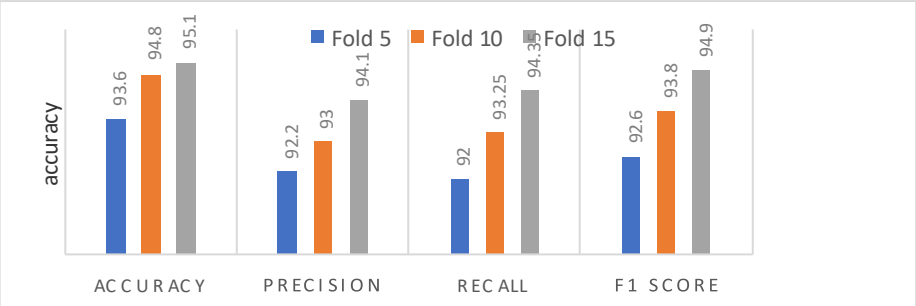

Figure 2: System validation with various cross validation using RNN-LSTM (sigmoid)

Experiment using Recurrent Neural Network (TanH)

Figure 3 displays RNN classification accuracy using the Cleveland dataset; similar tests were carried out with different cross validation and the results are shown in below figure 3. According to our findings, 15-fold cross validation delivers the greatest average classification accuracy of 93.55 percent and 94.90 percent for RNN utilizing Tanh.

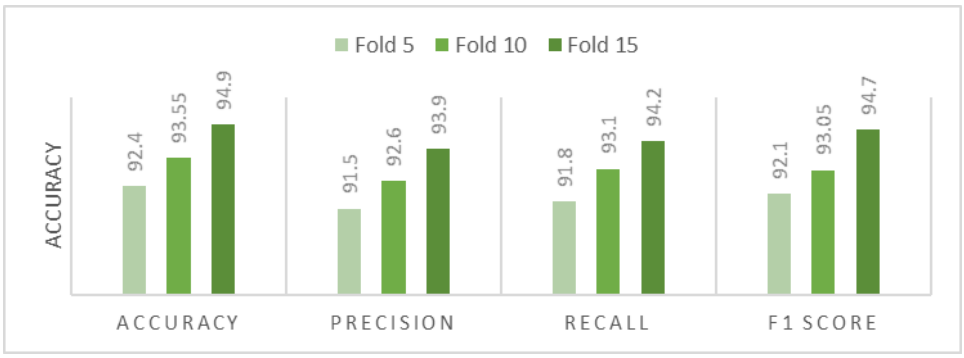

Figure 3: System validation with various cross validation using RNN-LSTM (Tanh)

Experiment using Recurrent Neural Network (ReLU)

In this experiment, we examined ReLU's classification accuracy using a real-time dataset; comparable tests were conducted using varied cross validation, and the results are shown in Figure 4. According to this study, 10-fold cross validation classification accuracy for RNNs is 95.30 percent and 97.10 percent, respectively.

The Figure 4 carried out 5-fold, 10-fold and 15-fold cross validation training of RNN (Tan h activation function).

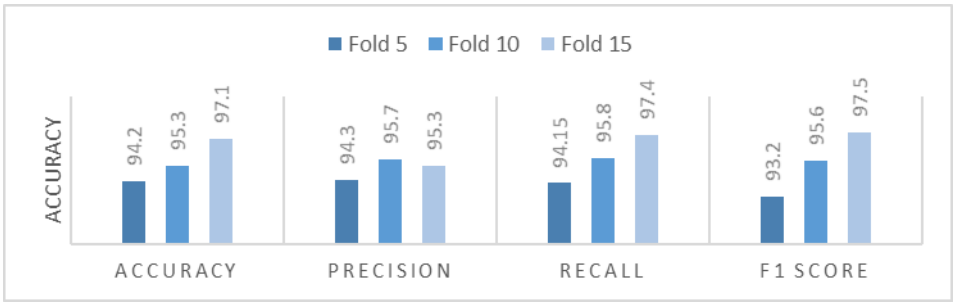

Figure 4: System validation with various cross validation using RNN-LSTM (ReLU)

A proposed deep learning classification method utilizing a machine learning algorithm is shown in Figure 4. The outcome of several cross-validation is shown in this diagram. For student performance prediction, we employed a minimum of three hidden layers. We infer that RNN with sigmoid delivers superior detection accuracy than the other three activation functions based on this experiment.

6 Conclusion

In this paper we are able to use the conceptual model they created to drive an integrated model capable of picking characteristics using a variety of feature selection approaches. The chosen procedure is a

collection of relevant aspects that performs better than single techniques in categorizing types of features. This work proposed the prediction of student performance using hybrid deep learning algorithms for educational data mining. In the traditional EDM, most systems face accuracy issues due to reading and redundant feature selection and overfitting problems during the classification. The multivariate input data has been taken for execution, and extract numerous features such as TF-IDF, correlation co-occurrence, dependency-based features etc. for achieving better classification and prediction accuracy. The proposed RNN-LSTM provides better accuracy as 95.50% by various cross-validation, which is better than traditional educational data mining students prediction systems. For the future enhancement to explore the system with ensemble deep learning algorithms on the real-time as well as the synthetic dataset.

References

[1] Yousafzai, Bashir Khan, et al. "Student-performulator: student academic performance using hybrid deep neural network." Sustainability 13.17 (2021): 9775.

[2] Zarate-Santana, Zaira-Jazmín, et al. "Learning Approaches and Coping with Academic Stress for Sustainability Teaching: Connections through Canonical Correspondence Analysis." Sustainability 13.2 (2021): 852.

[3] Ross, Manus, et al. "Using support vector machines to classify student attentiveness for the development of personalized learning systems." 2013 12th international conference on machine learning and applications. Vol. 1. IEEE, 2013.

[4] Kumar, A. Senthil, and K. Joshna. "Student's Performance Analysis with EDA and Machine Learning Models." (2021).

[5] Alyahyan, Eyman, and Dilek Düştegör. "Predicting academic success in higher education: literature review and best practices." International Journal of Educational Technology in Higher Education 17.1 (2020): 1-21.

[6] Khan, Ijaz, et al. "An artificial intelligence approach to monitor student performance and devise preventive measures." Smart Learning Environments 8.1 (2021): 1-18.

[7] Popenici, Stefan AD, and Sharon Kerr. "Exploring the impact of artificial intelligence on teaching and learning in higher education." Research and Practice in Technology Enhanced Learning 12.1 (2017): 1-13.

[8] Bhaskaran, Sundaresan, Raja Marappan, and Balachandran Santhi. "Design and analysis of a cluster-based intelligent hybrid recommendation system for e-learning applications." Mathematics 9.2 (2021): 197.

[9] Oyedeji, Ajibola Oluwafemi, et al. "Analysis and prediction of student academic performance using machine learning." JITCE (Journal of Information Technology and Computer Engineering) 4.01 (2020): 10-15.

[10] Schmucker, Robin, et al. "Assessing the Knowledge State of Online Students--New Data, New Approaches, Improved Accuracy." arXiv preprint arXiv:2109.01753 (2021).

[11] Mera-Gaona, Maritza, et al. "Framework for the Ensemble of Feature Selection Methods." Applied Sciences 11.17 (2021): 8122.

[12] Alnassar, Fatema, et al. "How Well a Student Performed? A Machine Learning Approach to Classify Students' Performance on Virtual Learning Environment." 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM). IEEE, 2021