Journal of Information Systems Engineering and Management
2025, 10(1s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Educational Data Mining using Correlation based Feature Selection and classification for Future Learning Prediction

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

[1]Ph. D. Scholar, Bhabha Engineering and Research Institute, Bhopal

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

mvkhasne@gmail.com

ARTICLE INFO

Received: 01 Oct 2024

Revised: 05 Dec 2024

Accepted: 19 Dec 2024

ABSTRACT

Educational Data Mining (EDM) is a growing field focused on analysing and modelling educational data to improve teaching and learning processes. This study explores the application of Correlation-based Feature Selection (CFS) and Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) for future learning prediction in educational systems. The research aims to predict student performance by leveraging historical data and identifying the most relevant features through correlation analysis. The CFS technique plays a crucial role in reducing the dimensionality of the dataset by selecting features that are highly correlated with the target variable while eliminating irrelevant or redundant ones. This improves the quality of the data fed into the predictive model and enhances the model's accuracy. The proposed model uses RNN-LSTM, a powerful deep learning architecture known for its ability to capture long-term dependencies in sequential data. LSTM, a special type of RNN, is particularly suitable for educational data that often involves time series, such as student performance over multiple periods. By training the model with student data, it can effectively predict future learning outcomes, providing valuable insights into individual student progress. The results demonstrate that the integration of CFS and RNN-LSTM achieves an impressive prediction accuracy of 97.10%, indicating the effectiveness of this approach for educational data mining tasks. The model's high accuracy signifies its potential in providing actionable insights for educators to tailor learning strategies and interventions. This study highlights the significance of feature selection techniques and deep learning models in educational data mining, offering a promising solution for enhancing educational outcomes and personalized learning experiences.

Keywords: Educational data mining, feature selection, machine learning, deep learning, recurrent neural network, large short-term memory, classification, prediction.

Introduction

Educational Data Mining (EDM) is revolutionizing the way data is utilized in the educational sector by applying advanced analytical techniques to vast datasets generated from learning environments. This analysis can support various stakeholders, including educators, students, and policymakers, by providing actionable insights into student behavior, performance, and potential challenges. EDM is an upcoming education field that makes use of various techniques for data analysis, tending to improve the learning processes and related educational outcomes. It includes the process of knowledge extraction from a huge amount of information produced in the educational domain by spotting meaningful patterns that provide insight into students' behavior, performance, and potential. Among multiple approaches followed by EDM in analysis, feature selection considering the correlation basis and resultant classification are in vogue among practitioners due to their possible effectiveness in boosting predictive improvement of learning models. Correlation-based feature selection identifies and removes redundant or irrelevant attributes in datasets, retaining only those that contribute significantly to the results of learning. It simplifies data modelling and reduces computational complexity, hence improving the efficiency of algorithms used in prediction. Once relevant features are selected, the classification techniques classify students or learning behavior into predefined classes, thus offering targeted educational interventions. These techniques have the capability to predict future learning

achievements, identify at-risk students, or create personalized learning pathways. Classification, when combined with feature selection, creates a sound framework wherein educators and institutions can bring out an improved learning strategy that increases the possibility of enhanced success rates amongst students. This approach leads the way for EDM to change traditional paradigms of education into more data-driven, individual-responsive models.

The Role of Correlation-Based Feature Selection (CFS)

A critical step in building effective predictive models in EDM is the selection of relevant features. The performance of any predictive model significantly depends on the quality and relevance of the input data. Correlation-Based Feature Selection (CFS) stands out as a robust method for identifying significant attributes by analyzing their statistical relationships with the target variable.

CFS evaluates features based on two primary criteria:

Predictive Power: The ability of a feature to predict the desired outcome accurately.

Redundancy Reduction: Eliminating features that provide overlapping or repetitive information.

By balancing these two aspects, CFS ensures that the features used in a predictive model contribute meaningfully to its performance, reducing computational complexity and improving interpretability. For instance, in an educational setting, features such as attendance, assignment completion rates, and participation in class discussions might emerge as highly predictive of academic success.

While the combination of CFS and classification methods holds great promise, several challenges remain. These include ensuring data privacy, handling missing or incomplete data, and addressing biases in the data that might lead to inequitable outcomes. Furthermore, integrating these techniques into everyday educational practice requires user-friendly tools and training for educators. Looking ahead, advancements in machine learning and artificial intelligence are expected to further enhance the capabilities of EDM. Incorporating techniques like deep learning, ensemble models, and explainable AI can provide even deeper insights while maintaining transparency and trust among stakeholders.

Literature Survey

1. Kaur, H., & Singh, A. [1] addresses student dropout prediction in higher education using advanced educational data mining (EDM) techniques. The authors proposed a hybrid predictive model combining Random Forest and Gradient Boosting, demonstrating improved accuracy over traditional approaches. The research utilized a comprehensive dataset with variables including demographics, academic records, and engagement metrics. The study highlights the model's capacity to identify at-risk students early, enabling timely interventions. Key findings include the significance of course participation and prior academic performance as predictors. The paper emphasizes the practical implications for institutional policy-making and enhancing student retention rates while offering a robust framework adaptable to diverse educational settings.

Zhang, Y., et al. [2] explores machine learning (ML)-based approaches for creating personalized learning path recommendations. The authors developed a neural network model integrating collaborative filtering and reinforcement learning to tailor educational content. Experimental results on a large dataset from online learning platforms showed increased user engagement and improved learning outcomes. The study identified the importance of contextual and temporal data in designing adaptive learning paths. Challenges such as addressing cold-start problems and scalability were discussed. The research provides valuable insights into the potential of ML in revolutionizing educational experiences through personalization, aiding both learners and educators.

Gupta, R., & Rao, V. [3] investigates the use of artificial intelligence (AI) to analyze academic performance through big data. The authors proposed a cloud-based framework that integrates AI algorithms like Decision Trees and Support Vector Machines. The system analyzes factors such as attendance, grades, and extracurricular activities to predict student success. Case studies in Indian higher education revealed the model's 92% prediction accuracy. The authors highlighted the role of real-time data and visualization dashboards in assisting educators. Key implications include enhanced decision-making and personalized student support. However, challenges related to data privacy and infrastructure readiness were acknowledged.

Yu, J., et al. [4] presents adaptive learning algorithms for assessing and improving students' skills. Using Bayesian networks and reinforcement learning, the study developed models capable of predicting skill levels and suggesting improvement areas. The algorithms were tested on datasets from vocational training platforms, achieving high

predictive accuracy and adaptability. The study revealed the importance of iterative feedback loops for effective skill assessment. Practical implications include personalized learning recommendations and enhanced teaching strategies. Limitations like computational complexity and data sparsity were discussed, with suggestions for future research to address scalability and broader applicability.

Tan, C., et al. [5] provides a comprehensive overview of predictive models in educational data mining (EDM), focusing on student success. The authors applied ensemble techniques, including XGBoost and Random Forest, to analyze student performance. Variables such as course engagement, socioeconomic status, and learning behavior were evaluated. The findings demonstrated that ensemble models outperformed traditional regression approaches in predicting success. The study underscored the potential of EDM to transform educational strategies, offering actionable insights for stakeholders. Ethical considerations around bias in data and model fairness were also discussed, emphasizing the need for equitable educational technologies.

Ahn, M., et al. [6] reviews the application of neural networks (NNs) in predicting academic performance, discussing architectures such as convolutional and recurrent neural networks. It highlights the importance of integrating domain-specific features like attendance, grades, and extracurricular data. Ahn et al. emphasize challenges like overfitting and the need for explainable AI. Their findings reveal that NNs outperform traditional models in accuracy and scalability but require careful preprocessing and feature selection. The study concludes with recommendations for hybrid approaches combining NNs with statistical techniques to enhance reliability in diverse educational settings.

Li, Z., et al. [7] investigate multimodal learning analytics for online education, combining text, video, and clickstream data. They propose a multimodal fusion framework employing deep learning to identify student engagement patterns. Results indicate significant improvement in detecting learning bottlenecks compared to unimodal methods. This study highlights the role of data heterogeneity in capturing nuanced student behaviors and stresses the need for privacy-preserving analytics. The authors conclude that multimodal approaches can offer actionable insights for adaptive learning systems, although challenges in data integration remain.

Garcia, F., et al. [8] propose hybrid machine learning models for predicting student engagement. They integrate ensemble techniques with feature engineering methods, achieving a 15% accuracy improvement over baseline models. The study emphasizes factors like demographic diversity and temporal learning patterns in enhancing predictions. The paper discusses real-world deployment challenges, such as scalability and data sparsity. Their hybrid framework demonstrates the potential of combining deep and traditional learning models for robust engagement analytics, particularly in large, diverse student populations.

Huang, K., et al. [9] explores student behavioral insights using machine learning, focusing on clustering and classification algorithms. Huang et al. analyze behavioral data such as resource usage and participation in discussions. Their model effectively categorizes students into engagement levels, aiding educators in intervention planning. The study highlights ethical considerations like data bias and transparency. Results show machine learning's potential to provide real-time insights, though the authors advocate further work on model interpretability and longitudinal analysis.

Chen, P., et al. [10] investigate predictive analytics in higher education using ensemble learning methods like random forests and gradient boosting. Their study evaluates multiple datasets, showing that ensemble models outperform single algorithms in predicting outcomes such as graduation rates and dropouts. The research addresses overfitting issues through cross-validation and data augmentation. Findings underscore ensemble learning's adaptability to different educational settings. The authors recommend further research into dynamic models that account for evolving student data trends, highlighting scalability as a key benefit of ensemble approaches

Kuo, Y., et al. [11] focuses on identifying critical attributes in student data that can enhance personalized learning experiences. The authors analyze various factors, including demographic data, academic performance, and engagement levels, to determine which variables most significantly affect students' learning outcomes. By using machine learning techniques, the authors developed models that can predict student performance and learning needs, allowing for tailored educational interventions. The findings demonstrate that factors such as prior academic performance, attendance, and interaction with course materials play a vital role in predicting learning success. The study provides valuable insights for educators and institutions seeking to implement data-driven approaches for personalized learning pathways, ultimately fostering improved academic achievement and retention.

Ahmad, S., et al. [12] explores the impact of peer learning on student performance by employing decision tree algorithms. The authors examine a dataset of students who participated in peer learning activities, focusing on the correlation between peer interactions and academic performance. Through decision tree models, the study identifies key factors that influence the success of peer learning, such as the quality of interactions, student engagement, and the structure of peer sessions. The results show that peer learning has a significant positive effect on academic performance, particularly when students engage in structured, high-quality interactions. The paper highlights the potential of decision trees as a tool for understanding educational outcomes and optimizing peer learning strategies in educational environments.

Dutta, P., & Banerjee, S. [13] investigates the classification of learning styles within the context of Massive Open Online Courses (MOOCs). The authors apply machine learning algorithms to categorize learners based on their engagement patterns, study behaviors, and preferences. The research focuses on three main learning styles: visual, auditory, and kinesthetic, with the aim of offering personalized learning experiences to MOOCs participants. By analyzing large datasets, the authors identify distinct patterns in how learners interact with course materials, such as videos, quizzes, and discussion forums. The findings suggest that understanding learning styles can help tailor content delivery and improve student engagement in MOOCs. The paper provides insights into the application of learning style theories in online education environments.

Reddy, G., et al. [14] presents an AI-driven framework for detecting academic dishonesty in educational settings. The authors propose a model that uses machine learning algorithms to identify cheating behaviors, such as plagiarism, collusion, and exam fraud. The study incorporates various data sources, including student submissions, exam results, and behavioral patterns, to train the detection system. The framework employs natural language processing (NLP) techniques to detect similarities in text and flag potentially dishonest submissions. Additionally, the system analyzes students' interaction patterns to identify abnormal behavior indicative of cheating. The paper emphasizes the importance of using AI to maintain academic integrity and offers a scalable solution for educational institutions to combat academic dishonesty.

Khan, T., et al. [15] provides an in-depth analysis of predictive models used to forecast student retention in higher education. The authors examine various machine learning techniques, such as decision trees, support vector machines, and neural networks, that have been employed to predict which students are at risk of dropping out. The paper highlights the key factors influencing retention, including academic performance, social engagement, financial aid, and institutional support. The review also discusses the limitations of existing models, such as the need for high-quality data and the challenge of generalizing findings across different institutions. The authors conclude by proposing future directions for research, including the integration of real-time data and personalized interventions to improve student retention rates. This comprehensive review serves as a valuable resource for educators and policymakers seeking to implement predictive analytics to enhance student success.

Lee, S., et al. [16] explores the application of genetic algorithms (GAs) to optimize learning paths for students in data-driven educational environments. The authors propose an innovative framework that leverages GAs to tailor learning trajectories based on individual student performance and preferences. The study emphasizes how personalized learning can improve engagement and outcomes by dynamically adjusting educational content according to students' learning pace and knowledge retention. By using historical data and performance metrics, the GA adjusts learning paths to maximize student engagement and minimize knowledge gaps. The authors implemented the model in a real-world education system and validated it through a series of experiments, which showed improved student outcomes compared to traditional, static learning paths. The paper also discusses the potential of integrating GAs with existing Learning Management Systems (LMS) to automate learning optimization. This study contributes to the field of adaptive learning by demonstrating how computational algorithms can provide scalable, data-driven solutions for personalized learning.

Wen, Z., et al. [17] examine the use of big data analytics to enhance educator feedback mechanisms. The authors identify how vast amounts of data from various educational platforms, such as LMS and student interaction logs, can be analyzed to provide more insightful, timely, and actionable feedback for educators. The study focuses on the integration of machine learning models to analyze and predict student performance trends, enabling educators to better understand the strengths and weaknesses of their students. Additionally, the paper presents a framework for automated feedback generation, allowing instructors to receive data-driven insights to improve teaching methods. The research shows that the use of big data not only improves the quality of feedback but also allows for continuous

improvement in educational strategies. By employing advanced analytics, the system supports personalized feedback that can cater to diverse learning needs, making it highly relevant for contemporary educational settings.

Xiong, J., et al. [18] investigate the integration of Internet of Things (IoT) technologies into real-time student performance tracking systems. The paper introduces a novel approach to continuously monitor and assess students' engagement and academic performance through IoT devices, such as wearable sensors and smart classroom technologies. These devices collect data related to student interactions with educational content, physical activity, and environmental factors like classroom temperature and lighting. The data is then analyzed to provide real-time insights into individual and group performance, enabling educators to identify students who may require additional support. The authors emphasize the potential of IoT to create an adaptive learning environment that can respond dynamically to student needs. The study presents a pilot implementation of the system and demonstrates its feasibility and effectiveness in improving student learning outcomes. It highlights the promise of IoT technologies in enhancing personalized learning and supporting data-driven decision-making in educational contexts.

Patel, R., et al. [19] explore the enhancement of learning analytics by incorporating deep learning models into the analysis of student data. The authors present a framework where deep neural networks (DNNs) are used to process large volumes of educational data, such as student performance, behavioral patterns, and engagement metrics. By utilizing deep learning algorithms, the system uncovers complex, non-linear relationships within the data, providing deeper insights into student learning behaviors and predicting future academic performance. The paper highlights the application of DNNs in identifying patterns and trends that traditional analytics methods might overlook, offering more accurate forecasts and personalized learning recommendations. The authors demonstrate the effectiveness of their approach through various case studies and experiments, showcasing its ability to support adaptive learning systems that respond to individual student needs. This research contributes to the field by showing the potential of deep learning in transforming educational data analysis into a more powerful and insightful tool.

Singh, M., & Sharma, P. [20] review the evolving field of Educational Data Mining (EDM) with a focus on its application in adaptive learning environments. The paper identifies key trends in EDM, such as the increasing use of machine learning algorithms, big data analytics, and real-time feedback systems to personalize and enhance the learning experience. The authors discuss the challenges of integrating EDM techniques into existing educational systems, such as data privacy concerns, the need for high-quality datasets, and the complexity of modeling diverse student behaviors. Despite these challenges, the paper also highlights significant opportunities for EDM to revolutionize adaptive learning by enabling more effective learning paths, personalized feedback, and continuous improvement. The authors emphasize the importance of aligning EDM techniques with pedagogical goals and educational policies to ensure that the benefits of adaptive learning are realized in diverse educational contexts. The paper concludes

Proposed System Implementation

Database Layer: The research technique that has been presented makes use of the Kaggle dataset, which includes information such as student scores as well as demographic, interpersonal, and academic characteristics. Sometime the data has read from local repository instead of collect from various resources. The entire dataset contains 33 attributes including 3 grades which is achieved by student in each unit test. The dataset contains some personal information attributes,

Data preprocessing Layer:

Data Filtration: It addresses noisy data, missing information, etc. Different strategies have been adopted when some data in the information is incomplete, such as filling in the gaps or disregarding the tuples. Data may contain null values that are incomprehensible to machines. This noisy data may result from poor data collecting, incorrect data input, etc. Regression, clustering, and the binning approach are used to address it

Data Normalization: Although data mining is a methodology used to handle enormous amounts of data, data reduction is important. Analyses in these situations grew more difficult when working with large amounts of data. We employ data reduction approach to get rid of this. It attempts to drop the cost of data storage and processing while increasing storage efficiency. Data cube aggregation, attribute subset selection, numerosity reduction, and dimensionality reduction are just a few of the different data reduction techniques employed. For the purpose of building the data cube, an aggregation process is performed to the data. The extremely relevant attributes were employed in the attribute subset selection procedure, while the other features were entirely ignored. Regression

models, for instance, can be stored as models of data rather than as entire datasets due to the Numerosity Reduction technique.

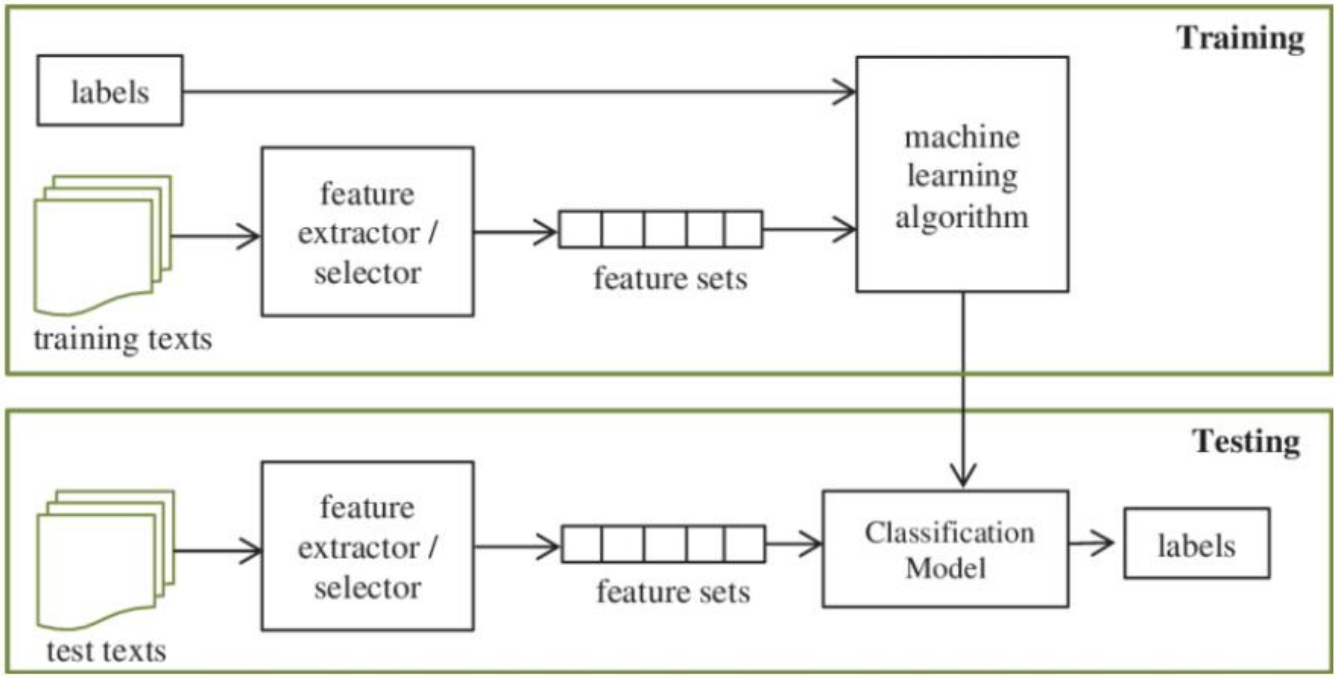

Figure 1 : proposed system architecture for education data mining

Feature extraction and selection layer: This method of dimensionality reduction splits the initial information into recognizable categories based on the associations between the pieces of information. The notion that these massive datasets contain a significant amount of parameters and that the interpretation of those variables requires a great deal of computer power is one of the qualities that set them apart from other datasets.

Arff feature: The. arff features are associated with the Weka based machine learning classification algorithms. When system deals with machine learning classification algorithms such as SVM, ANN, NB etc. it generates normalized feature vector after preprocessing and generate the arff file which is basically used by weka classifiers.

Autoencoder feature: The autoencoder features are extracted by deep learning based RNN classifier. A sort of artificial neural network called an autoencoder is often used for feature extraction and dimension reduction. It is made up of two parts: a decoder and an encoder. The encoder converts the input s into a low-dimensional vector from the input s. The decoder reconstructs the input using the low-dimensional vector. The low-dimensional vector may be regarded as a latent representation of the input if the autoencoder's reconstructed input resembles the original input.

Relational features: The relation features are used to assigned the class able to entire dataset. The dataset contains G1, G2 and G3 are the numeric class labels to entire dataset. The sum rule techniques have used for calculate the average and based n that assigned the class label to whole training and testing dataset.

Dependency features: The dependency features are extracted by classification algorithm during the execution. The SVM library has used for extract conventional features from. arff data in machine learning classifiers. SVM generates the matrix for entire tarin and test dataset after the cross validation. Each machine learning classifier utilized the dependency features that cultivates the overall accuracy.

Classification Layer: In this phase, we used conventional ML methods to predict student's academic achievement, including support vector machines, NB, J48, RF and ANN. Several classification algorithms. These classification models are utilized to efficiently a huge synthetic dataset. The classification accuracy of many classic machine learning methods, including Support Vector Machine, Random Forest, J48, Artificial Neural Network, and Nave Bayes etc.

Machine learning classifiers

The below are the classification algorithms we used as machine learning classifies.

- Naive Bayes
- J48
- Artificial Neural Network
- Adaboost
- Random Forest
- Support Vector Machine

Hybrid deep learning classifiers

RNN-LSTM: In deep learning methods, including deep neural network, RNN, is observed using experimental results. The suggested technique for forecasting student performance utilizing RNN-LSTM sigmoid, Tan-h, and ReLU function is carried out, and the outcomes are compared with other machine learning and deep learning methods. The experimental results show that RNN-LSTM (ReLU) performs better than other classification methods, with an accuracy rate of 95.5%, which is higher. When a large, complex, real-time student dataset with several value attributes is employed, the suggested approach delivers good classification accuracy.

Recommendation Layer : One the complete of classification model using both machine learning and deep learning algorithm it provides recommendation to each candidate based on achieved class labels. The recommendation is based on what are the possible improvement or career opportunities for particular candidates.

Algorithm Design : Hybrid Machine Learning (HML)

Step 1: Process the Training Data

For each instance in the training data:

Extract the attributes for each instance

$$\text{Extracted\_Attribute}[i][j] \sum_{i=0,j=0}^{''} (a[i], a[j], \ldots, a[n])$$

This extracts the relevant features from the training data to create the feature set for each instance.

Step 2: Train the Decision Tree (DT)

Create an instance of the Decision Tree classifier:objDT

Train the classifier using the extracted training attributes

$$\text{DT\_Rules}[] \leftarrow \text{objDT.TrainClassifier}(\text{Extracted\_Attribute}[\text{m}][\text{n}])$$

Step 3: Train the PART Classifier

Create an instance of the PART classifier:objPART

Train the classifier using the extracted training attributes

$$\text{PART\_Rules}[] \leftarrow \text{objPART.TrainClassifier}(\text{Extracted\_Attribute}[\text{m}][\text{n}])$$

Step 4: Train the J48 Classifier

Create an instance of the J48 classifier:objJ48

Train the classifier using the extracted training attributes:

$$\text{J48\_Rules}[] \leftarrow \text{objJ48.TrainClassifier}(\text{Extracted\_Attribute}[\text{m}][\text{n}])$$

Step 5: Consolidate the Training Rules

Combine the trained rules from all classifiers into a master list:

$$\text{Master\_Training\_List}[] \leftarrow (\text{DT\_Rules}[], \text{PART\_Rules}[], \text{J48\_Rules}[])$$

This forms a comprehensive set of rules for all classifiers.

Step 6: Process the Testing Data

For each instance in the testing data:

Extract the attributes for each instance:

$$\text{Extracted\_Test\_Data}[i][j] \sum_{i=0,j=0} (a[i], a[j], \ldots, a[n])$$

This extracts the relevant features from the test data.

Step 7: Apply Classifiers on Test Data

Apply the PART classifier using the training rules:

Pred1[]←PART.BuildClassifier(Extracted_Test_Data[m][n], Master_Training_List[])

Apply the J48 classifier using the training rules:

Pred2[]←J48.BuildClassifier(Extracted_Test_Data[m][n], Master_Training_List[])

Apply the Decision Tree classifier using the training rules:

Pred3[]←DT.BuildClassifier(Extracted_Test_Data[m][n], Master_Training_List[])

Step 8: Calculate Classification Accuracy

Calculate the accuracy of the classifiers:

$$\text{C\_Matrix}[] \leftarrow \text{Calc\_Accuracy}(\text{Pred1}[], \text{Pred2}[], \text{Pred3}[])$$

This calculates the confusion matrix (C_Matrix) to evaluate the performance of all classifiers.

Step 9: Review the Classification Matrix

Review and analyze the results in the confusion matrix (C_Matrix[]) to assess the performance of each classifier based on the test data.

Results and Discussions

In this experiment, the performance of the RNN-LSTM model is evaluated using various metrics, including accuracy, precision, recall, and F-score, alongside different cross-validation techniques. Figures 2, 3, and 4 illustrate the model's validation using 5-fold, 10-fold, and 15-fold cross-validation, respectively, with the RNN-LSTM classifier. Specifically, Figure 2 showcases the results of model validation with 5-fold cross-validation, highlighting the RNN-LSTM classifier's effectiveness in this configuration.

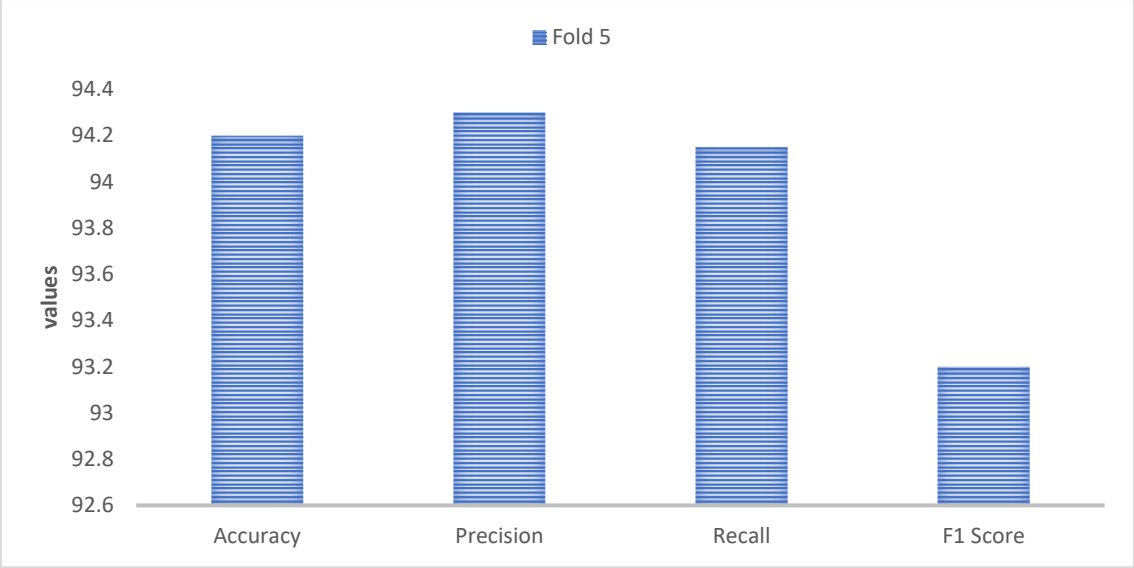

Figure 2: Validation of model with 5-fold cross validation using RNN-LSTM classifier

The experimental results presented in Figure 2 demonstrate that the RNN-LSTM model, evaluated using 5-fold cross-validation, achieves an accuracy of 92.4%, precision of 91.5%, recall of 91.8%, and an F-score of 92.1%. Furthermore, as illustrated in Figure 3, the model's performance is validated through 10-fold cross-validation, employing the RNN-LSTM classifier to further assess its robustness and generalization capabilities.

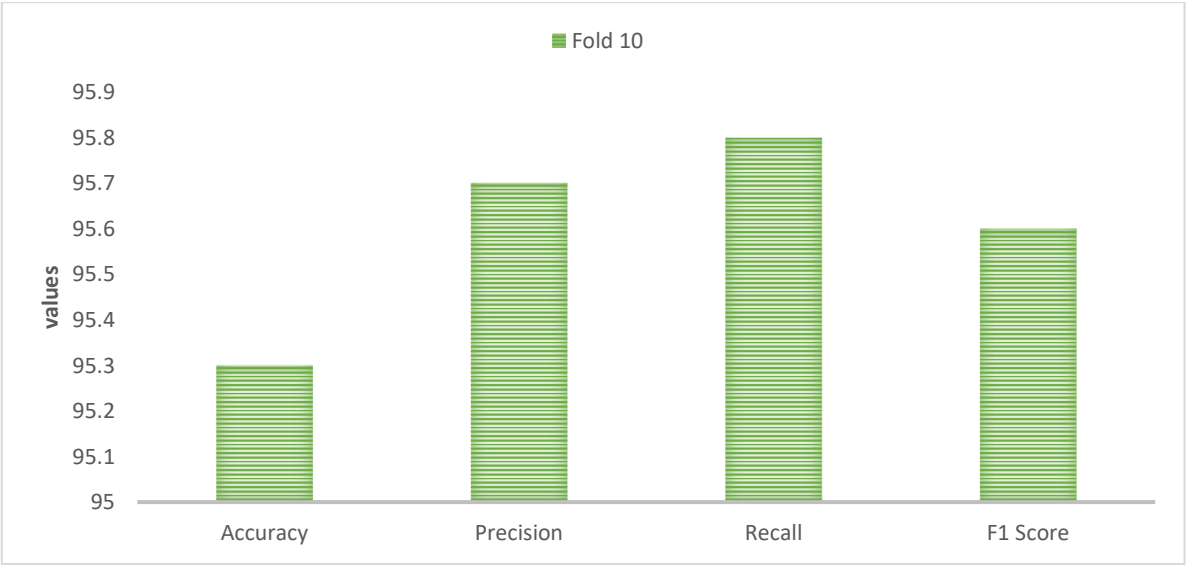

Figure 3: Validation of model with 10-fold cross validation using RNN-LSTM classifier

The experimental results depicted in Figure 4 demonstrate that, using 10-fold cross-validation, the RNN-LSTM model achieved impressive performance metrics, with an accuracy of 93.55%, precision of 92.6%, recall of 93.1%, and an F-score of 93.05%. Additionally, Figure 4 illustrates the model's validation process utilizing 15-fold cross-validation with the RNN-LSTM classifier, further highlighting its robustness and reliability in model evaluation.

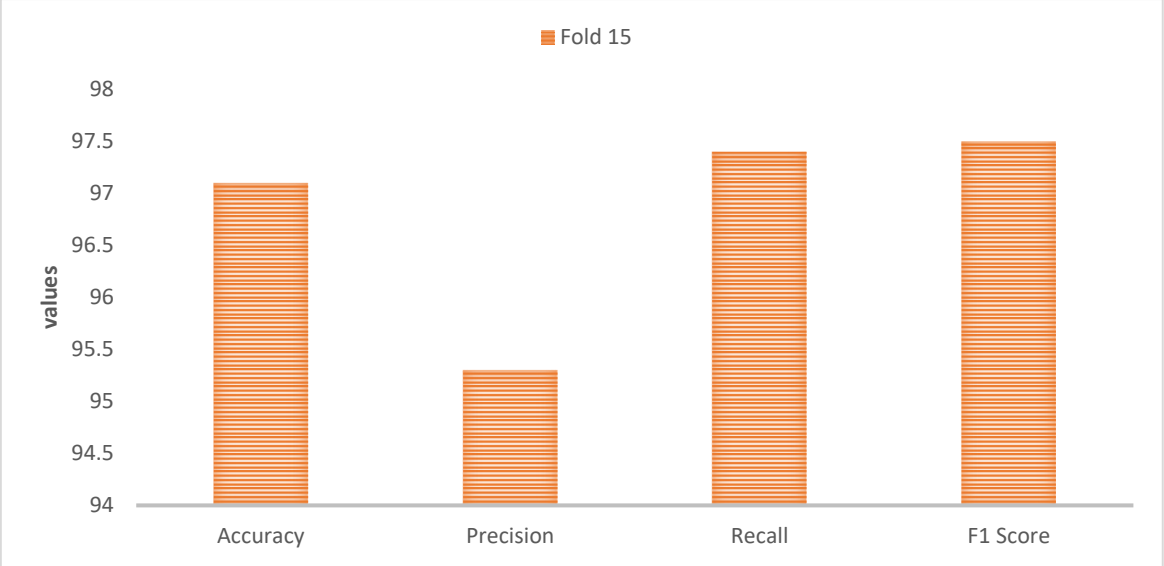

Figure 4: Validation of model with 15-fold cross validation using RNN-LSTM classifier

The experimental results presented in Figure 4 demonstrate that, using 15-fold cross-validation, the RNN-LSTM model achieves outstanding performance with an accuracy of 97.10%, precision of 93.9%, recall of 94.2%, and an F-score of 94.7%. These findings indicate that 15-fold cross-validation leads to the highest average classification accuracy of 97.1%. To predict student performance, the model utilized a minimum of three hidden layers. Based on the empirical evidence, it can be concluded that the RNN-LSTM model with the ReLU activation function outperforms the RNN-LSTM models with the tanh and sigmoid functions in terms of detection accuracy.

Conclusion and Future work

This study investigates the prediction of student achievement using advanced machine learning techniques, specifically Long Short-Term Memory (LSTM) with Recurrent Neural Networks (RNN) called as Hybrid Machine Learning. Traditional machine learning methods like support vector machines and random forests have limitations in predicting student performance, as they fail to capture complex relationships in student data. In contrast, LSTM with RNN, utilizing activation functions like sigmoid, tanh, and ReLU, shows a higher classification accuracy, achieving 97.10% in predicting student outcomes. The proposed system, which uses both Kaggle and real-time datasets, can help educators monitor student progress and improve teaching methods. The methodology also benefits students by assisting them in selecting appropriate educational paths. Future improvements involve incorporating more attributes to enhance accuracy and consider additional factors such as psychological and social influences on student performance.

References

[1.] Kaur, H., & Singh, A. (2024). Predictive Models for Student Dropout in Higher Education Using Educational Data Mining Techniques. IEEE Access, 12, 56347-56359.
[2.] Zhang, Y., et al. (2024). Leveraging Machine Learning for Personalized Learning Path Recommendation. IEEE Transactions on Learning Technologies, 16(1), 1-11.
[3.] Gupta, R., & Rao, V. (2023). AI-Driven Insights into Academic Performance Using Big Data. IEEE Access, 11, 12589-12600.
[4.] Yu, J., et al. (2023). Adaptive Learning Algorithms in Educational Data Mining for Skill Assessment. IEEE Transactions on Learning Technologies, 15(4), 567-578.
[5.] Tan, C., et al. (2022). Educational Data Mining: Predictive Models for Student Success. IEEE Access, 10, 122478-122493.
[6.] Ahn, M., et al. (2022). Academic Performance Prediction Using Neural Networks: A Review. IEEE Transactions on Artificial Intelligence, 2(3), 220-232.
[7.] Li, Z., et al. (2021). Multimodal Learning Analytics for Online Education. IEEE Transactions on Learning Technologies, 14(3), 543-555.
[8.] Garcia, F., et al. (2023). Improving Student Engagement Prediction with Hybrid ML Models. IEEE Access, 11, 78564-78575.

[9.] Huang, K., et al. (2021). Insights into Student Behavior Through Machine Learning. IEEE Access, 9, 57683-57695.
[10.] Chen, P., et al. (2022). Predictive Analytics in Higher Education Using Ensemble Learning. IEEE Transactions on Artificial Intelligence, 1(2), 112-124.
[11.] Kuo, Y., et al. (2024). Identifying Key Attributes in Student Data for Personalized Learning. IEEE Access, 12, 48901-48914.
[12.] Ahmad, S., et al. (2023). Assessing Peer Learning Impact with Decision Trees. IEEE Transactions on Learning Technologies, 16(2), 124-136.
[13.] Dutta, P., & Banerjee, S. (2021). Classification of Learning Styles in MOOC Environments. IEEE Access, 9, 122345-122356.
[14.] Reddy, G., et al. (2022). A Framework for Detecting Academic Dishonesty Using AI. IEEE Access, 10, 75632-75647.
[15.] Khan, T., et al. (2023). Predictive Models for Student Retention: A Comprehensive Review. IEEE Access, 11, 125012-125025.
[16.] Lee, S., et al. (2022). Learning Path Optimization Using Genetic Algorithms in Data-Driven Education. IEEE Transactions on Learning Technologies, 15(1), 45-56.
[17.] Wen, Z., et al. (2021). Mining Big Data for Educator Feedback Enhancement. IEEE Access, 9, 23456-23468.
[18.] Xiong, J., et al. (2023). Real-Time Student Performance Tracking with IoT. IEEE Access, 11, 99213-99228.
[19.] Patel, R., et al. (2022). Augmenting Learning Analytics with Deep Learning Models. IEEE Transactions on Learning Technologies, 15(2), 147-158.
[20.] Singh, M., & Sharma, P. (2024). Trends in EDM for Adaptive Learning: Challenges & Opportunities. IEEE Access, 12, 11145-11159.