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

Research Article

A Study on Predictive Modelling of Student Academic Performance using Machine Learning Method

Shoukath TK*[1], Midhunchakkravarthy[2]

[1]*Faculty of Computer Science and Multimedia, Lincoln University College, Malaysia*

**Corresponding Author Email: stkkodan@lincoln.edu.my*

[2]*Faculty of Computer Science and Multimedia, Lincoln University College, Malaysia*

Email: midhun@lincoln.edu.my

ARTICLE INFO

Received: 02 Oct 2024
Revised: 26 Nov 2024
Accepted: 11 Dec 2024

ABSTRACT

Predicting students’ academic performance is a crucial initiative in the field of education, as it allows educators and administrators to spot students who may require extra assistance, customize educational resources to meet individual requirements, and improve overall educational results. Conventional approaches to forecasting academic achievement, such as statistical analysis and expert evaluation, have certain drawbacks in terms of precision and scalability. The emergence of machine learning (ML) methods provides a possible alternative by utilizing extensive datasets and advanced algorithms to reveal patterns and generate more precise predictions. This investigation’s primary goal is to explore the predictive modelling of student academic performance by improving the accuracy of predictions through the utilization of machine learning techniques. The study was conducted utilizing the Python programming environment. The prediction of student academic performance was carried out utilising the Bidirectional long short-term memory (Bi-LSTM) based Weighted Cost Effective Random Forest algorithm. The study utilized the Deep Encoder CNN-Bi-LSTM for optimal feature extraction to foresee student academic performance. The extracted features were then classified using the Weighted Cost Effective Random Forest (WECRF) classifier, and the classification was evaluated in terms of accuracy, precision, specificity, sensitivity, and recall. The issues addressed include the class imbalance, computational complexity, cost, and huge dimensional issue, among others. The random forest method achieved Precision Score - 0.72, Recall Score - 0.68, F1 Score - 0.69, and Accuracy - 0.77 in this study. Moreover, the suggested technique facilitates the automated forecasting and enhancement of students' future academic performance. Keywords: Predictive Modelling; Students; Academic Performance; Machine Learning; Bi-LSTM; CNN-Bi-LSTM; Deep Learning.

Keywords: Student Academic Performance, Machine Learning

INTRODUCTION

Educational Data Mining (EDM) has greatly influenced the current advancements in the education industry. The extensive range of investigation has identified and implemented novel possibilities and also opportunities for technologically advanced learning systems tailored to the specific requirements of students. The EDM's cutting-edge approaches and application strategies are crucial in enhancing the learning environment. For instance, the EDM is indispensable in comprehending the student learning environment through the assessment of both the educational context and machine learning methods. The EDM field encompasses the investigation, study, and utilisation of Data Mining (DM) techniques, as stated in [1]. The field of DM utilizes a variety of interdisciplinary methods to achieve its objectives [2]. The system possesses a thorough approach to collecting significant and intellectual perceptions from unprocessed data; the DM cycle is depicted in the figure below. Analysing ML and statistical methods directly for

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

educational data allows for the identification of significant patterns that boost students' grasp and benefit academic institutions.

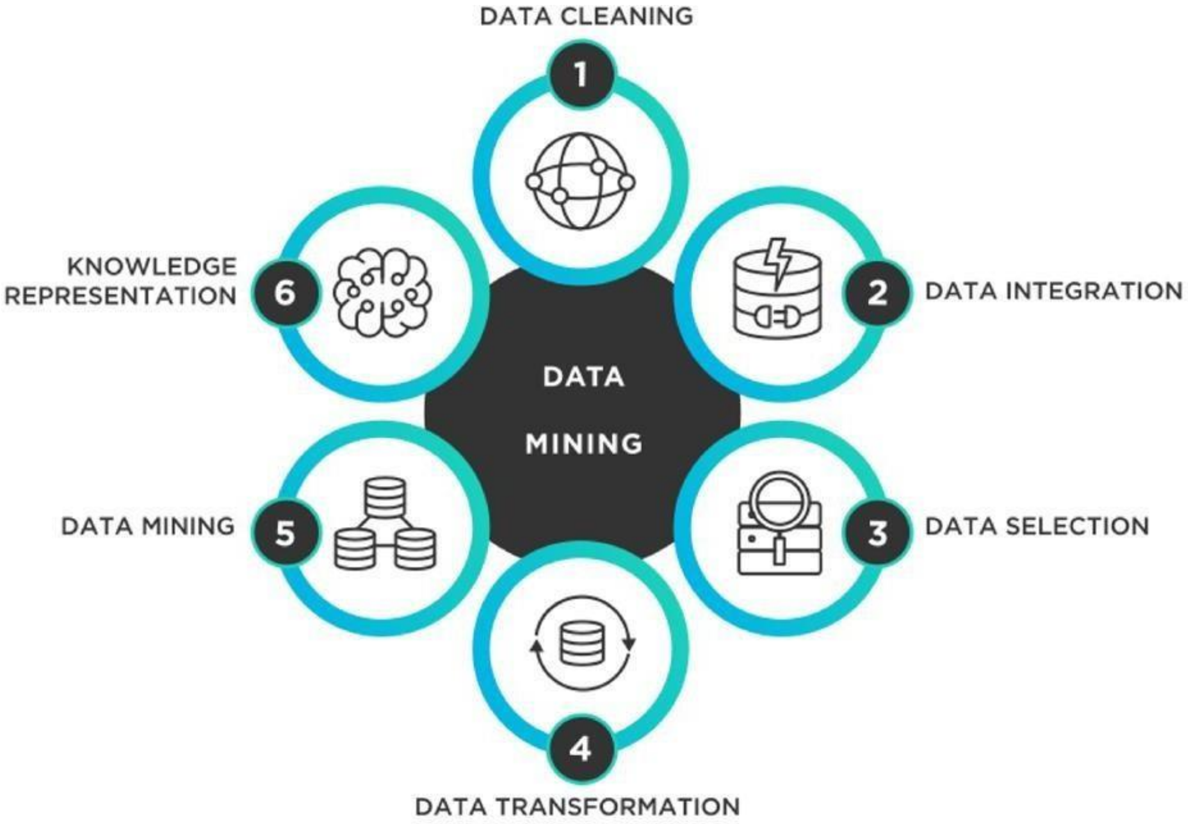

Figure 1: Typical cycle of Data Mining methodology [2]

Quality education is a crucial component of the Sustainable Development Goals (17-SDGs) as recognized by the UN [3]. It is essential to have in mind the importance of ensuring equal opportunities and equitable distribution when working towards sustainable development goals. The evaluation of students' dissertations in the pursuit of higher education is a matter of great importance that should be conducted on a global scale. The high rate of student attrition in academic institutions leads to a substantial and expensive loss of resources in educational settings [4]. This also has a negative impact on the evaluation alongside assessment procedures of academic organizations. The engineering programs have experienced greater reductions compared to other disciplines in science and art [5].

Therefore, this investigation looks to generate and improve predictive models of student academic performance by utilizing sophisticated machine learning techniques. The project aims to enhance the precision of performance prediction by utilizing a range of machine learning techniques and methodologies. This will facilitate the more efficient identification of students who are at danger of underperforming. This research is to investigate the influence of various attributes and data pre-processing approaches on the performance of the model. The goal is to offer educators and policymakers practical insights to enhance educational tactics and interventions. The project aims to enhance educational DM by showcasing the capabilities of machine learning in revolutionizing predictive modelling in academic settings, promoting a more data-focused approach to assisting student achievement.

Objectives of this study

1. To perform the student academic performance prediction from student performance dataset using Bi-LSTM based Weighted Cost-Effective Random Forest algorithm.
2. To perform feature extraction from pre-processed student performance data, Deep Encoder CNN- BiLSTM technique will be used to improve the accuracy and reduce the computational cost.

The subsequent section provides a detailed analysis of the previous literature that is pertinent to this specific research.

LITERATURE REVIEW

The table below contains the previous literature on predictive modelling of student academic performance, specifically focusing on enhancing prediction accuracy using machine learning techniques.

Table 1: Related Works

AUTHORS AND YEARS	METHODOLOGY	FINDINGS
Yıldız & Börekci (2020) [6]	To tackle a classification problem, this study estimated two classes (successful or unsuccessful based on exam results) from collected data. The study analysed the prediction accuracy of supervised classification algorithms and identified effective variables for class construction.	The results showed that the Neural Network algorithm (98.6%) scored highest. The accuracy rates of other algorithms include KNN (86.2%), Logistic Regression (78.4%), SVM (90.3%), Decision Tree (91.9%), Random Forest (90.0%), and Naive Bayes (81.7%).
Injadat, et.al.,(2020) [7]	The study analyses two university datasets to predict student performance at 20% and 50% course completion, using a multi-split approach with the Gini index and p-value to optimize a bagging ensemble of six machine learning algorithms.	The optimized bagging ensemble models show high accuracy in predicting student performance for both datasets, confirming the effectiveness of the approach.
Aouifi, et.al.,(2021) [8]	K-Nearest Neighbors (KNN) and Multilayer Perceptron Predicted student success using video clips and analysed viewing behaviour through educational data mining.	The KNN classifier yielded the highest results with an average accuracy of 65.07%, suggesting that learners' performance can be predicted based on video sequence viewing behaviour.
Cam, et.al., (2021) [9]	The study used Random Forest (RF), C5.0, CART, and Multilayer Perceptron (MLP) algorithms to predict first-year students' learning performance based on family background variables available before the semester starts.	the CART algorithm performed best, with key predictors being the mother's occupation, department, father's occupation, main source of living expenses, and admission status. These insights could help in early identification and intervention for students at risk.
A. Nabil et al., (2021) [10]	The study predicts student performance using first-year grades with models like deep neural networks (DNN), decision trees, and others. Resampling methods, including SMOTE and ADASYN, were compared to handle imbalanced data.	Addressing class imbalance through techniques like SMOTE improved model accuracy. The study found that balancing the class distribution improved the accuracy of SVM models to 86.9%.
Zeineddine et al., (2021) [11]	This paper proposes using Automated Machine Learning (AutoML) to predict student performance based on data available before the start of academic programs. The approach leverages	Using AutoML, the ensemble model achieved 83% accuracy in predicting failing students with pre-start data, surpassing the previous 70% limit after data balancing with SMOTE.

	Big Data Analytics and Machine Learning to enhance prediction accuracy, allowing for early intervention.	
Mubarak, A.A., et al., (2021) [12]	Developed a hyper-model of convolutional neural networks and long short-term memory (CONV-LSTM) to predict student dropout in MOOCs.	Addressed class imbalance, resulting in a more reliable prediction model with a high false-negative rate reduced by a cost-sensitive method.
Demeter, E., et al., (2022) [13]	Used Random Forest to predict graduation outcomes for first-time-in-college undergraduates using admissions, academic, and financial aid records. Identified predictors: credit hours earned, college and high school GPAs, estimated family contribution, and grades in required major courses.	Achieved 79% prediction accuracy. Identified at-risk students, especially those with high financial need and moderate degree progress, often overlooked by university protocols.
G. Feng et al., (2022) [14]	The study uses optimized K-means clustering, discriminant analysis, and convolutional neural networks (CNN) to predict student performance. A new statistic is introduced to accurately determine cluster numbers in K-means, with CNN handling data training and testing	The results show that the new statistic effectively determines the correct number of clusters in K-means, enhancing prediction accuracy. The combined approach improves the reliability of predicting student performance.
Wang X, et al., (2022) [15]	The study uses a machine learning model with pre-training and fine-tuning to predict learning performance in online education. It also implements a personalized feedback system, evaluated through a quasi-experiment with 62 participants.	The PT-GRU model improved learning performance and reduced cognitive load, achieving high accuracy and F1 scores. Personalized feedback, incorporating prediction results and suggestions, effectively enhanced learning outcomes and reduced cognitive load.
Hsing-Chung Chen., et al., (2022) [16]	Proposed a novel framework for predicting student performance using time-series weekly student activity data and VLE imbalanced data distribution. The framework combines CNN and LSTM models to extract spatiotemporal features from the activity data.	The combined CNN-LSTM model outperformed baseline models such as LSTM, SVM, and logistic regression (LR) in early prediction cases, emphasizing the importance of early identification of potential performance issues and providing targeted interventions. The analysis also highlights the value of visualizing predictions, student activity maps, and feature importance in understanding student behaviour and performance.
Feng, G. (2022) [17]	The paper uses clustering, discriminant analysis, and	The new statistic improves K-means clustering and

	convolutional neural networks to predict student performance. It introduces a new statistic for optimizing K-means clustering and evaluates the model's accuracy using cross-validation.	prediction reliability, with the model effectively forecasting student performance.
Yağc (2022) [18]	This study compared various machine learning methods—random forests, nearest neighbor, support vector machines, logistic regression, Naïve Bayes, and k-nearest neighbor—to predict final exam scores using a dataset of 1854 students from a Turkish state university.	Results indicated a 70-75% classification accuracy for the suggested model.
Wang, et al., (2023) [19]	Used probabilistic graphical models to represent and analyse the relationships between variables.	This approach effectively predicted dropout rates by modelling the dependencies among various factors influencing student retention.
S.TK, et al., (2023) [20]	The study uses Support Vector Machine (SVM) to predict student academic performance in an imbalanced dataset, leveraging Educational Data Mining (EDM) techniques to process large educational datasets.	SVM effectively predicted student performance, showing its value in identifying at-risk students and aiding institutions in improving educational outcomes.
Pallathadka et al., (2023) [21]	The study uses data mining techniques, including Naïve Bayes, ID3, C4.5, and SVM, to predict student performance based on prior course results. It employs the UCI student performance dataset and evaluates algorithms on accuracy and error rate.	The analysis identifies patterns in student performance data to enhance forecasting and reduce failure rates, demonstrating varying effectiveness across different machine learning algorithms.
Asselman, A., et al. (2023) [22]	This study proposed a novel Performance Factors Analysis (PFA) approach using machine learning models, including Random Forest, AdaBoost, and XGBoost, to boost predictive accuracy.	The scalable XGBoost model surpassed the other evaluated models, leading to substantial improvements in performance prediction compared to the original PFA algorithm.
Ilic, Milos, et al.(2024) [23]	The study uses machine learning (ML) and artificial neural networks (ANN) to predict student grades in an e-learning system. Key correlations among predictors were identified using the minimum redundancy–maximum relevance (mrMR) criterion to determine the most suitable AI method.	ANN, particularly the Levenberg–Marquardt algorithm with Bayesian regularization, outperformed ML methods in prediction accuracy. The findings underscore the importance of variable correlations in selecting the best model for student performance prediction.
Umamaheswari et al., (2024) [24]	This paper proposed an improved classifier that improves accuracy and	The proposed methodology outperformed existing models

	reduces overfitting and under fitting risks using advanced machine learning methods. This study identifies the main factors affecting student achievement. Student data-based classification is used to compare different classifiers.	in accuracy (84%), recall (95%), and F1 score (82%). This new approach can predict students' academic trajectory, enabling stakeholders to implement timely interventions.

Research Gap: Although there have been some advancements aimed at enhancing the prediction of student performance, numerous research fails to adequately consider the implementation of these novel techniques in practical educational environments. Moreover, there is a dearth of agreement on the most efficient measures for assessing student academic achievement and a restricted investigation into how various machine learning strategies might be enhanced and contrasted to identify the most dependable and precise prediction methods.

METHODOLOGY

The UCI (University of California Irvine ML repository) Repository's student performance dataset(https://archive.ics.uci.edu/ml/datasets/student+performancePREDICTION) was first pre-processed for scaling, normalization, and replacement of missing values. Additionally, the Deep Encoder CNN-Bi-LSTM model was used to extract the important features. This model reduced computational costs, addressed issues with class imbalance and sequential classification, enhanced accuracy, and enabled autonomous prediction. Additionally, the data was divided directly into testing and training sets. The Weighted Cost-Effective Random Forest (WCERF) method was used to classify the data and extract student performance information in terms of accuracy, specificity, precision, and recall. Using the resulting classification, a prediction of the student's performance was made in order to assist future academic growth.

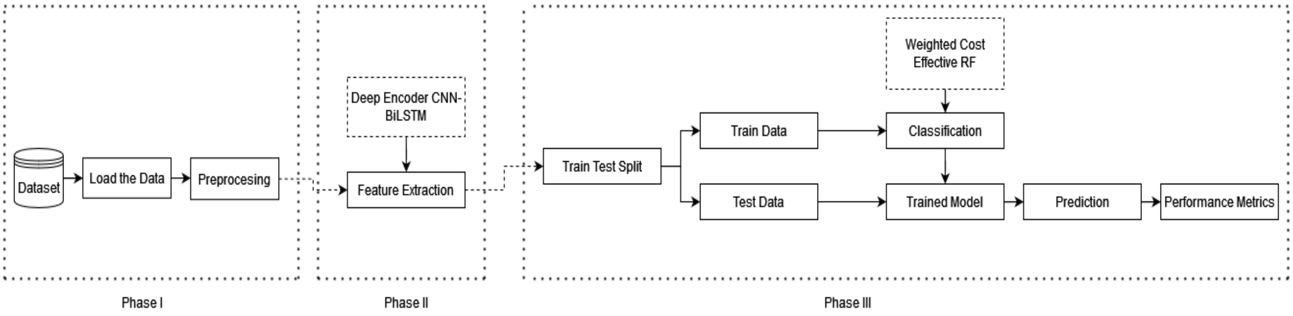

Figure 2: Workflow of this study

This research combines the Deep CNN with encoder technology and the Bi-LSTM Neural Network model to boost the performance of feature extraction. CNN is a Deep Learning (DL) technique consisting of convolutional layers. The convolutional layer's extract feature maps from the text dataset using different numbers of kernels. The pooling layers are subsequently followed by a reduction in these dimensions. Additionally, other types of pooling layers are introduced, including average pooling and max pooling. Subsampling layers are introduced between the convolutional and fully linked layers. The application of CNNs is suitable for datasets that involve many parameters and nodes and can be done automatically. The figure below illustrates the architecture of the Deep Encoder CNN feature extraction.

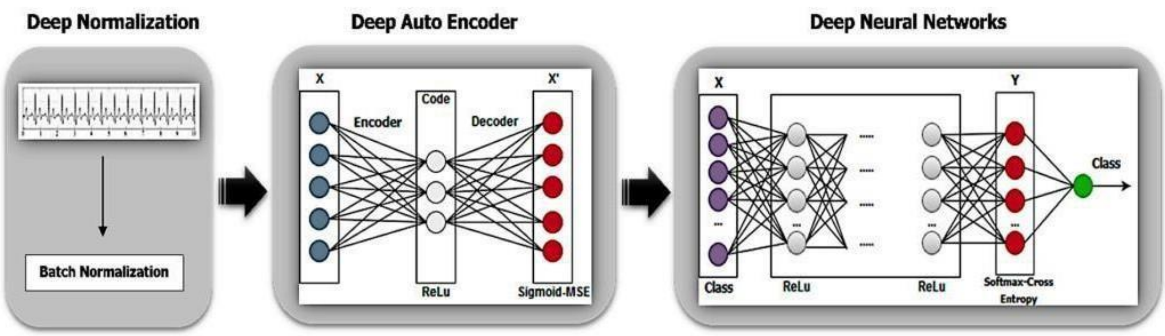

Figure 3: Deep Encoder CNN architecture

An auto-encoder is an algorithm that learns generic features through a training process that is based on greedy layer-wise training. Deep network topologies offer faster and exceptional prediction results, and when combined with Deep CNN, they provide performance increase. Dimensionality reduction involves extracting low-dimensional features from high-dimensional data using an unsupervised and non-linear approach. To further enhance performance, the Bi-LSTM method is integrated with the Deep Encoder based CNN technology. Bidirectional LSTMs are an extension of regular LSTMs that improve the performance of models on sequence classification issues.

Out of all the categorization techniques in machine learning, the random forest algorithm offers the highest level of accuracy. It is capable of handling large amounts of data with a high number of variables, potentially in the hundreds. It can automatically balance the dataset when there is a higher proportion of infrequent classes compared to other data classes. The WCERF model is illustrated in the figure below.

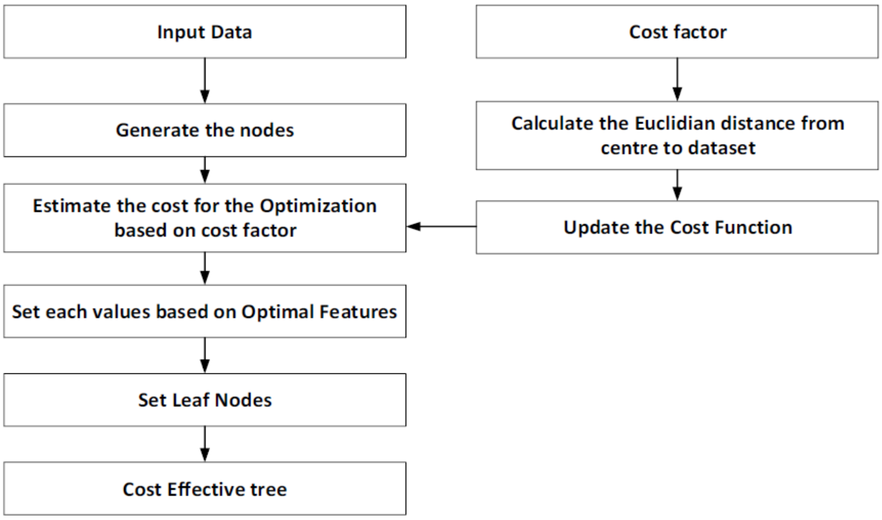
Input Data
Generate the nodes
Estimate the cost for the Optimization based on cost factor
Set each values based on Optimal Features
Set Leaf Nodes
Cost Effective tree
Cost factor
Calculate the Euclidian distance from centre to dataset
Update the Cost Function

Figure 4: WCERF workflow

The WCERF model utilizes the Euclidean distance to generate the cost function. This cost function is based on the actual sample distribution and aims to maximize the decreased values of feature costs by dividing the nodes. This algorithm also tackles the issue of unbalanced data. A weighted voting mechanism is utilized for the testing and prediction samples to determine the prediction outcome when voting for a random classifier. Optimizing the cost factor involves estimating the process cost and setting each value based on optimal feature extraction. The cost function has been further modified using the Euclidean distance. A cost-effective decision tree has been generated for classification. In addition, the WCERF classifier's performance is assessed alongside various ML techniques, counting Random Forest, and K-nearest neighbour, and Naïve Bayes methods.

RESULTS AND DISCUSSIONS

The machine learning classification procedure must exhibit high accuracy and efficiency in predicting the academic performance of students. It indicates the necessity for enhancing the achieved performance and the requirement for devising a unique strategy that optimizes prediction. The target samples of this study are depicted in the picture below.

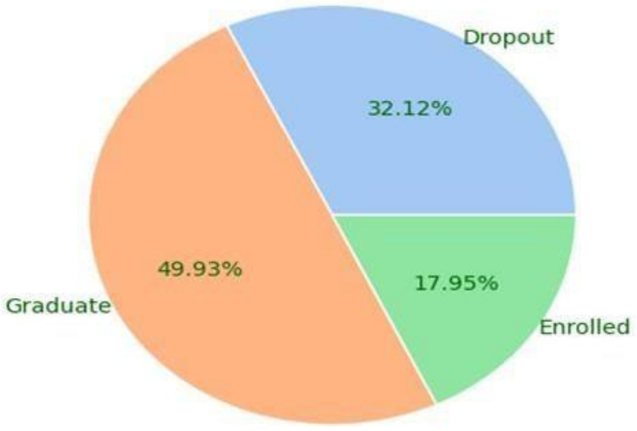

Figure 5: Target audience of this study

Here are the key features utilized in this research

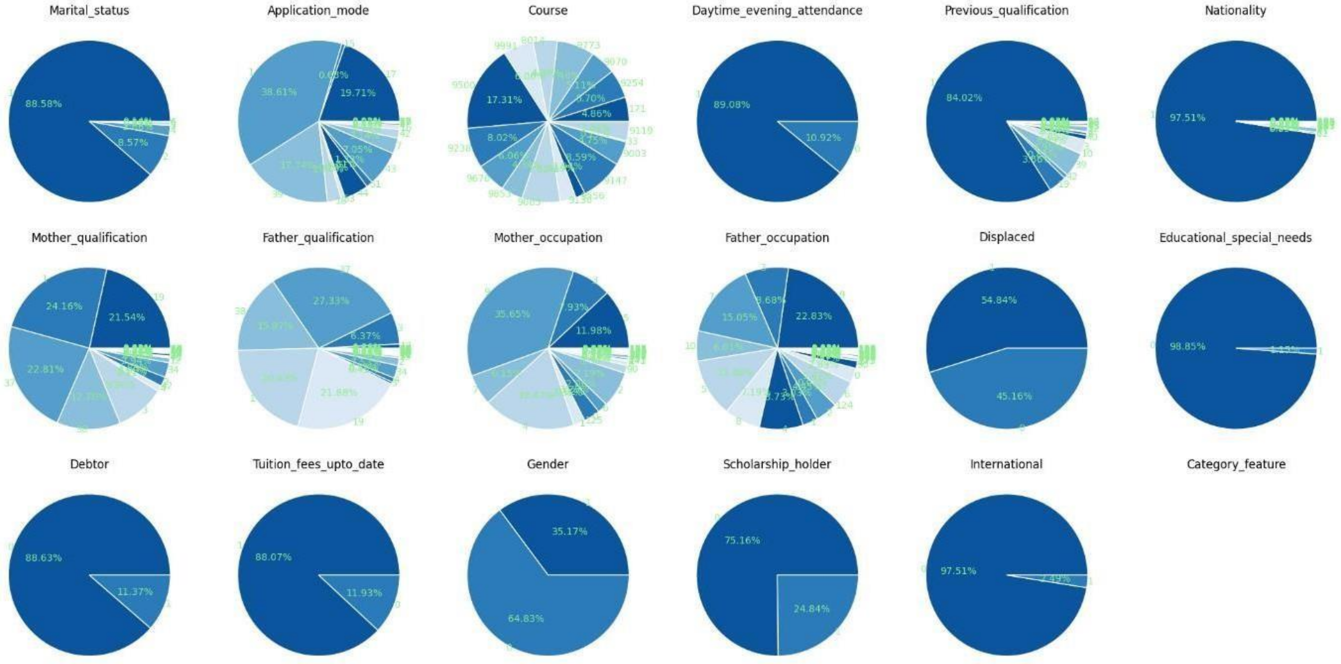

Figure 6: Key features

The correlation between the features and the expected value is illustrated in the figure below.

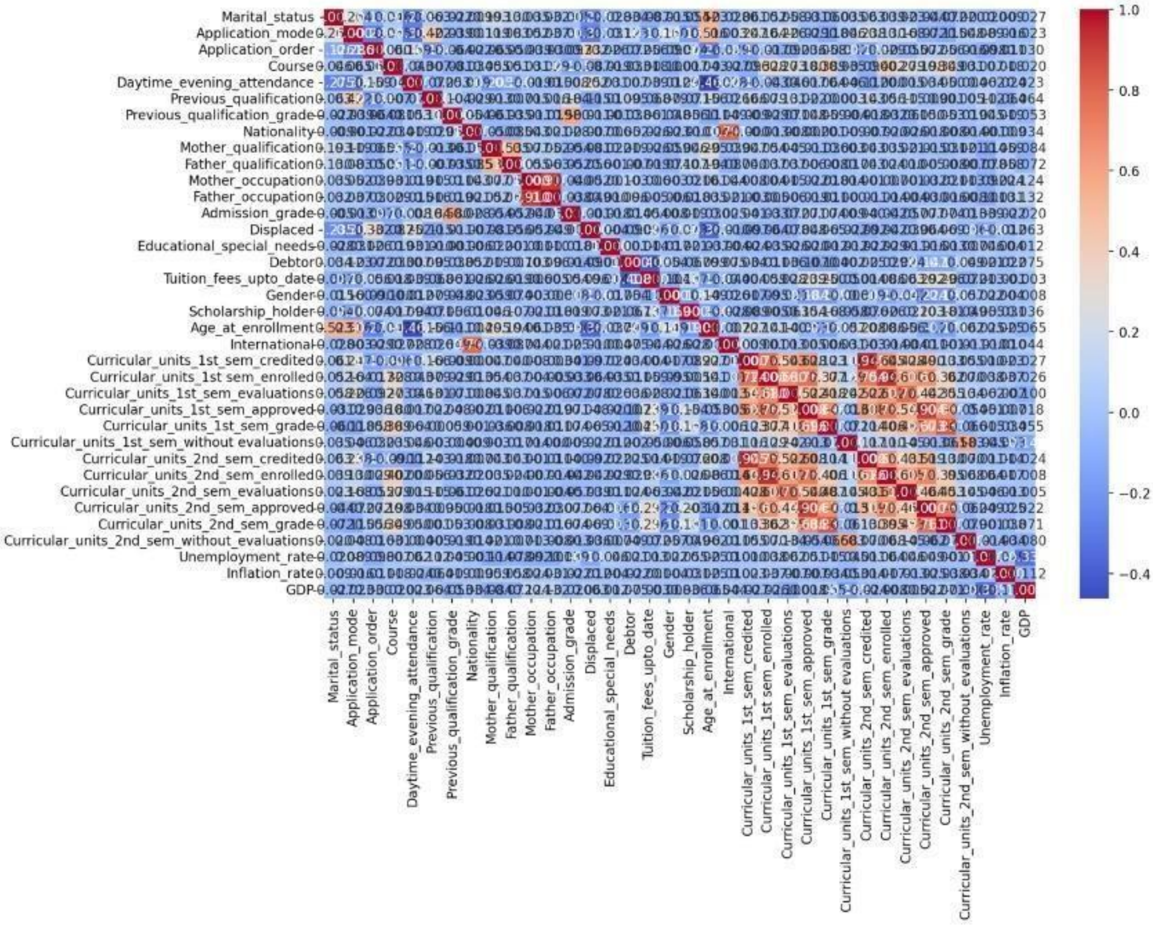

Figure 7: Relation between features and the predicted value

The graph below illustrates the relationship between several attributes.

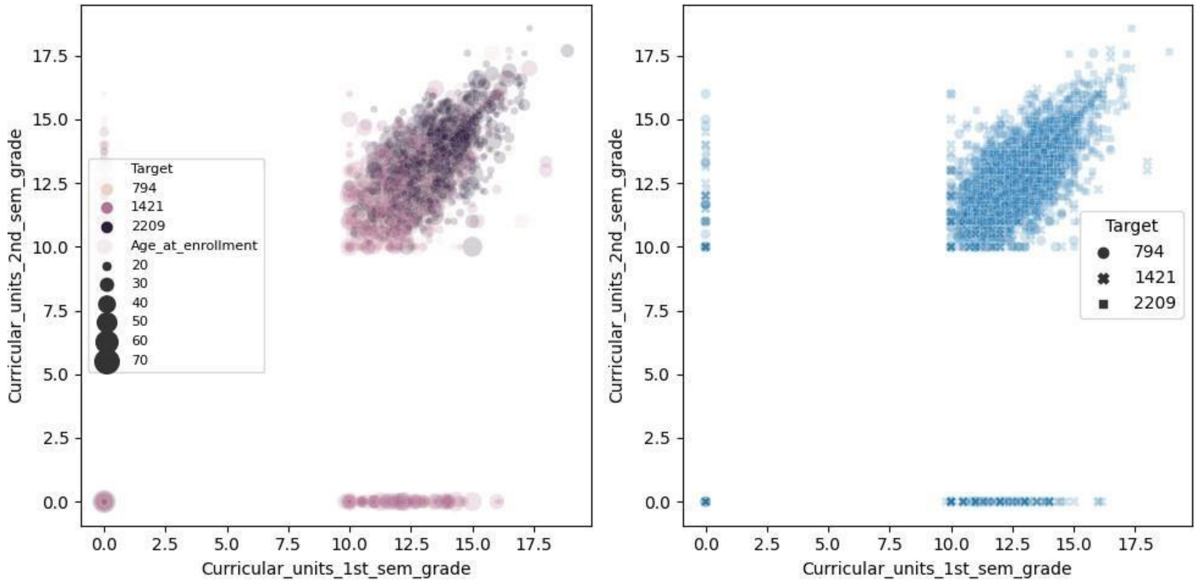

Figure: 8: Relationship between different features

The Bi-LSTM based Weighted Cost-Effective Random Forest Algorithm is a model that combines the advantages of Bi-LSTM networks with a Random Forest classifier. This hybrid methodology seeks to utilize the sequential learning skills of Bi-LSTM for extracting features and the ensemble learning qualities of Random Forest for accurate classification, especially when working with imbalanced datasets. The Bi-LSTM model is able to capture sequential dependencies and patterns, making it particularly valuable for jobs that include sequential data. The Weighted Cost-

Effective Random Forest algorithm tackles the issue of uneven class distribution by providing greater significance to minority classes. The study's findings indicate that the Random Forest Classifier achieved Precision Score - 0.72, Recall Score - 0.68, and F1 Score - 0.69.

The Deep Encoder CNN-BiLSTM technology effectively integrates CNNs and BiLSTM networks to proficiently handle and extract distinctive characteristics from sequential input. This hybrid technique is especially effective for challenges involving sequence-to-sequence learning and sequence categorization. The Deep Encoder CNN-BiLSTM technique provides a strong foundation for acquiring knowledge from sequential data by leveraging the CNNs' capacity for capturing spatial patterns and BiLSTMs in modelling temporal relationships. The RandomForestClassifier with the class weight parameter set to 'balanced' has Precision Score - 0.47, Recall Score - 0.41, and F1 Score - 0.36. The data analysis plot is depicted in the figure below.

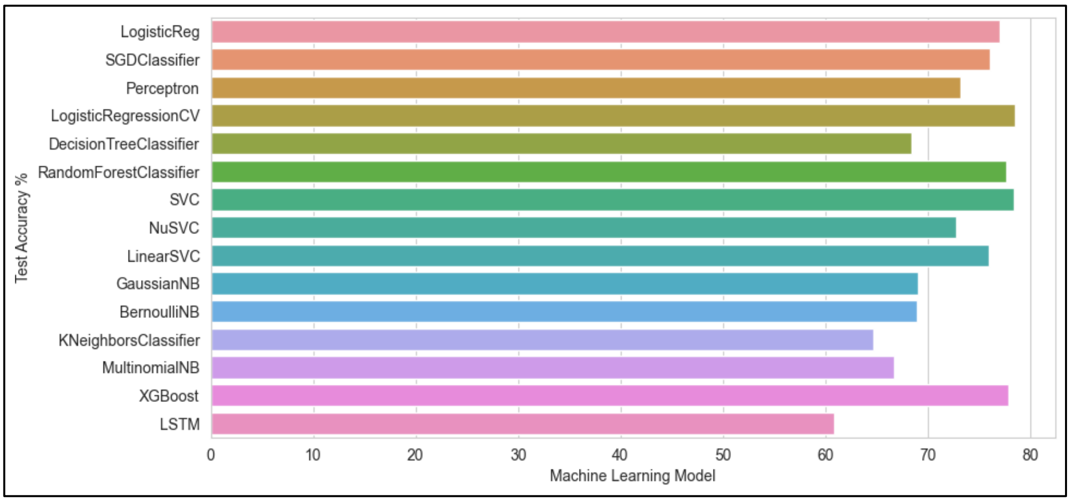

Figure 9: Prediction Accuracy of ML Techniques

CONCLUSION

Finally, employing machine learning approaches to predict student academic achievement greatly improves the accuracy of predictions in comparison to conventional statistical methods. Through the utilization of sophisticated algorithms, it becomes feasible to discern intricate patterns and connections within educational data. These models can integrate a diverse array of variables, such as socio-economic considerations, prior academic records, and behavioural indications, to offer a more thorough and nuanced comprehension of student performance. The Weighted Cost Effective Random Forest technique addresses the problem of imbalanced class distribution by assigning higher importance to minority classes. The study's results demonstrate that the Random Forest Classifier attained Precision Score - 0.72, Recall Score - 0.68, and F1 Score - 0.69. For feature extraction, the CNN-BiLSTM model was employed. The Random Forest Classifier, with the class weight parameter set to 'balanced', attained Precision Score - 0.47, Recall Score - 0.41, and F1 Score - 0.36. The enhanced precision of these forecasts empowers educators and administrators to detect kids at higher failure risk at an earlier stage and customize interventions with greater effectiveness, ultimately leading to improved educational results and more optimized allocation of resources.

REFERENCES

[1] Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, *2019*.

[2] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552.

[3] Hussain, S., & Khan, M. Q. (2023). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of data science*, *10*(3), 637-655.

[4] Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q. (2019). Predicting students' performance using machine learning techniques. *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences*, *27*(1), 194-205.

[5] Mulà, I., Tilbury, D., Ryan, A., Mader, M., Dlouhá, J., Mader, C., ... & Alba, D. (2017). Catalysing change in higher education for sustainable development: A review of professional development initiatives for university educators. *International journal of sustainability in higher education*, *18*(5), 798-820.

[6] Yıldız, M., & Börekci, C. (2020). Predicting academic achievement with machine learning algorithms. *Journal of educational technology and online learning*, *3*(3), 372-392.

[7] Injadat, M., Moubayed, A., Nassif, A.B. and Shami, A., 2020. Multi-split optimized bagging ensemble model selection for multi-class educational data mining. Applied Intelligence, 50(12), pp.4506-4528

[8] El Aouifi, H., El Hajji, M., Es-Saady, Y. and Douzi, H., 2021. Predicting learner's performance through video sequences viewing behavior analysis using educational data-mining. Education and Information Technologies, 26(5), pp.5799-5814

[9] Huynh-Cam, T.T., Chen, L.S. and Le, H., 2021. Using decision trees and random Forest algorithms to predict and determine factors contributing to first-Year University students' learning performance. Algorithms, 14(11), p.318.

[10] A. Nabil, M. Seyam, and A. Abou-Elfetouh, "Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks," IEEE Access, vol. 9, pp. 140731–140746, 2021, doi: 10.1109/ACCESS.2021.3119596.

[11] Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, *89*, 106903.

[12] A. A. Mubarak, H. Cao, and I. M. Hezam, "Deep analytic model for student dropout prediction in massive open online courses," Computers & Electrical Engineering, vol. 93, p. 107271, 2021.

[13] Demeter, E., Dorodchi, M., Al-Hossami, E. et al. Predicting first-time-in-college students' degree completion outcomes. High Educ 84, 589–609 (2022). https://doi.org/10.1007/s10734-021-00790-9

[14] G. Feng, M. Fan and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," in IEEE Access, vol. 10, pp. 19558-19571, 2022, doi: 10.1109/ACCESS.2022.3151652.

[15] Wang X, Zhang L, He T. Learning performance prediction-based personalized feedback in online learning via machine learning. Sustainability. 2022 Jun 23;14(13):7654.

[16] H. C. Chen et al., "Week-Wise Student Performance Early Prediction in Virtual Learning Environment Using a Deep Explainable Artificial Intelligence," Applied Sciences 2022, Vol. 12, Page 1885, vol. 12, no. 4, p. 1885, Feb. 2022, doi: 10.3390/APP12041885.

[17] G. Feng, M. Fan, and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," IEEE Access, vol. 10, pp. 19558–19571, 2022.

[18] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1),

[19] Wang, H., Li, X., and Wu, Y., 2023. Application of Bayesian networks for predicting student dropout rates. Computers & Education, 193, p.104718

[20] S. TK and Midhunchakkravarthy, "Academic Performance Prediction of At-Risk Students using Machine Learning Techniques," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 1222-1227, doi: 10.1109/ICACITE57410.2023.10183199.

[21] Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials today: proceedings*, *80*, 3782-3785.

[22] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," Interactive Learning Environments, vol. 31, no. 6, pp. 3360–3379, Aug. 2023, doi: 10.1080/10494820.2021.1928235.

[23] Ilic, M., Kekovic, G., Mikic, V., Mangaroska, K., Kopanja, L. and Vesin, B., 2024. Predicting Student Performance in a Programming Tutoring System Using AI and Filtering Techniques. IEEE Transactions on Learning Technologies.

[24] Umamaheswari, P., Vanitha, M., Devi, P. V., Theporal, J. G., & Basha, B. R. (2024). Student success prediction using a novel machine learning approach based on modified SVM. *Multidisciplinary Science Journal*, *6*.