International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

"A Comparative Study of Behaviour Predictors for School Students in Indore Using Machine Learning Algorithms".

Yougal Kishore Sharma *1, Dr. Arpana Bharani *2
*1 (Research Scholars, Department of Computer Science, Dr. A. P. J. Abdul Kalam University, Indore)
*2 (Research Guide, Department of Computer Science, Dr. A.P.J. Abdul Kalam University, Indore)

ABSTRACT

Predicting student academic performance has become a key area in educational data mining, with machine learning techniques offering powerful tools for early intervention and decision-making. This study explores the application of classification models to forecast student success and behavioural outcomes, with the goal of improving academic support systems and reducing dropout rates. Two distinct datasets of student information were utilized, and three boosting-based machine learning algorithms - XGBoost, AdaBoost, and an Artificial Neural Network (DenseNet) - were implemented. Feature engineering techniques were applied to optimize input variables and enhance model effectiveness.

The results demonstrate that it is feasible to predict student behaviour and academic performance with significant accuracy using machine learning models. Among the evaluated methods, XGBoost and AdaBoost achieved the best predictive performance with an accuracy rate of approximately 88%. Conversely, the DenseNet-based neural network model produced the lowest accuracy, around 49%. These findings underscore the effectiveness of boosting methods for educational prediction tasks and highlight the role of machine learning as a practical approach to advancing educational research and institutional planning.

Keywords: student performance prediction, educational data mining, machine learning, boosting algorithms, XGBoost, AdaBoost, neural networks, feature engineering.

INTRODUCTION

Predicting student behaviour and academic performance using machine learning has emerged as a powerful approach for educational research and institutional planning. By analyzing data from various sources such as academic records, attendance logs, survey responses, and even digital footprints, machine learning algorithms can uncover hidden patterns, forecast outcomes, and generate insights that help institutions make informed decisions. These predictions allow educators to identify at-risk learners, address individual weaknesses, and provide targeted interventions that enhance both student success and institutional performance.

Education plays a crucial role in national development and constitutes a key driver of long-term socio-economic growth. However, persistent challenges such as student dropouts and failure in key subjects significantly impact literacy levels and overall academic achievement. Since educational institutions maintain extensive records of students, large databases containing academic and behavioural information offer potential opportunities for extracting actionable knowledge. Questions such as which students perform equally across subjects, what factors influence academic performance, which courses are most attractive to students, and whether performance can be reliably predicted can be addressed using data mining and machine learning. The application of these techniques provides valuable insights by identifying patterns and trends, thereby optimizing institutional success rates and educational planning.

Data mining, particularly classification methods, plays a central role in predicting student behaviour. Classification, a supervised learning process, categorizes datasets into predefined labels using algorithms such as Decision Trees, Support Vector Machines (SVM), Naïve Bayes, Random Forests, and logistic regression. More advanced approaches, including boosting algorithms and neural networks, further extend predictive capabilities. The process typically involves data collection from multiple sources, pre-processing and feature engineering to extract meaningful attributes, and the application of machine learning models to generate predictions.

In this study, emphasis is placed on comparing and evaluating machine learning algorithms to identify the most effective strategies for predicting student behaviour and success. This contributes not only to improving individual student outcomes but also to enhancing institutional planning, resource allocation, and dropout prevention.

RESEARCH GAP

While numerous studies have applied machine learning techniques to predict student academic performance, several limitations remain unaddressed:

- Limited focus on behavioural factors.
- Model comparison constraints.
- Underutilization of boosting methods.
- Feature engineering gaps.
- Dropout prediction as a secondary goal.

LITERATURE REVIEW

Many studies have been conducted in the field of educational data mining to determine how to predict how well students would perform using various data mining techniques such as clustering, classification, neural networks, and so on.In this section, we'll go over some of the approaches that have been used in various research studies. Associative analysis, clustering, statistical analysis, and link mining are only a few of the data mining techniques they investigated from various domains. Finally, they identified the top ten data mining algorithms that fit within the aforementioned criteria. Some of the top ten algorithms for identifying the best solutions are C4.5, k-Means, Support Vector Machine (SVM), Apriori, EM, PageRank, AdaBoost, K nearest neighbour method, Naive Bayes, and CART.Students' academic performance is predicted using real-time data sets from educational organisations using algorithms from the classification and clustering categories, which are discussed in detail in this work.

Fekry, G. et al[1] This research is part of a study that uses pattern-recognition tools to try to understand out why people do what they do and how it relates to their own personal modalities. A group of distinct behaviours that students exhibit during group presentations at a higher level of school are chosen for this study. The first step in studying these behaviours is to examine how they appear and how the model's behaviour patterns match up with what we notice when we manually examine our own observations. They both make use of the same set of video files.

F. Orji et al [2] the techniques discovered a pattern of student engagement. They also discovered that student engagement and test scores are good predictors of academic performance. Assessment scores are determined by how well results do on quizzes and assignments.

N. Kondo, et al [3] the outcomes suggest that some learning behaviour that impacts learning can be detected by looking at log data that is kept online. The comparative relevance of explanatory variables obtained by this method can also be utilised in institutional research to determine which variable has the greatest impact on the learning outcome.

R. Cao et al [4] Machine learning can be utilised in a variety of industries, and the fact that it is likely to be employed makes it even more vital to teach machine learning. In the process of creating machine learning courses, CDIO teaching is used. This paper also uses student behaviour data as a real-world data source for practise connections. Students will learn how to use machine learning in this manner. They will accomplish this through more extensive theoretical study, real-world tests, and the entire application process.

M. M. Hassan et al[5] To analyse datasets and select significant qualities, we used machine learning algorithms. We also used Association Rule Mining methods to determine the most connected attributes in the dataset. The Apriori approach was used to implement Rule Mining, and we discovered 8 rules.

T. Purwoningsih, et al[6] In this research, we describe the features of e-learning students using exploratory data analytics (EDA) and machine learning. Based on their demographic profile data and learning behaviour, we then use this knowledge to forecast and advise how pupils will learn. Using analytics data to create e-Learning instructional designs that support online students succeed is a success that this study assists instructors with in the early stages.

J. Edmond Meku Fotso, et al[7] Our goal is to create a model that can predict how people will respond (interact) during the learning process, providing both students and teachers with a greater understanding of how people learn.

RESEARCH PROCESS

The research process in machine learning is a structured sequence of activities that enables researchers to explore problems systematically and develop robust predictive

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

models. In the context of predicting student behaviour, this process ensures that the models are not only accurate but also meaningful in an educational setting. The research process followed in this study is summarized in Figure 1 involves the following stages:

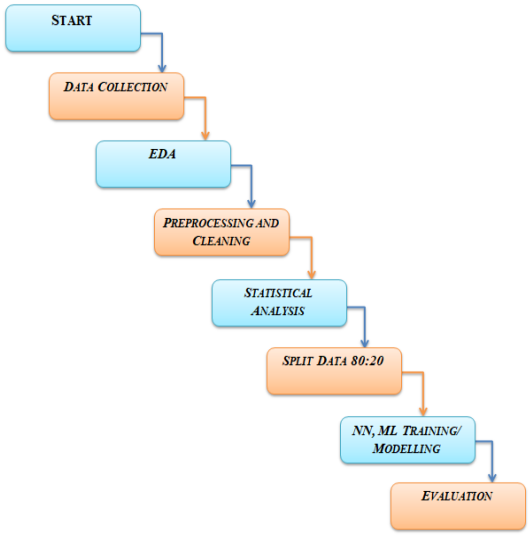

Figure 1: research process flow

CLASSIFICATION ML ALGORITHM

Machine learning is an emerging field that gives computers the ability to autonomously learn from examples given to them in the past. In order to construct mathematical models and provide predictions based on past data or information, machine learning makes use of a variety of algorithms. Image and audio recognition, email filtering, Facebook auto-tagging, recommender systems, and countless more applications are currently making use of it. This introductory machine learning course will walk you through the basics of the field and its many subfields, including supervised, unsupervised, and reinforcement learning. Models for classification and regression, clustering, hidden Markov chains, and sequential analysis will all be covered. There are people all around us in the actual world who can learn anything thanks to their incredible capacity for experience-based learning, and there are also computers and other machines that can carry out commands from us. Does this mean that machines can't learn from their own mistakes and previous data just like humans? Now they reach the part that Machine Learning plays.

One branch of AI, known as "machine learning," focuses on teaching computers new skills by analysing existing data and drawing on their own experiences. Originally coined in 1959 by Arthur Samuel, the phrase machine learning has since become widely used. A simplified definition would be:

At a broad level, machine learning can be classified into three types:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

Machine learning uses two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data.

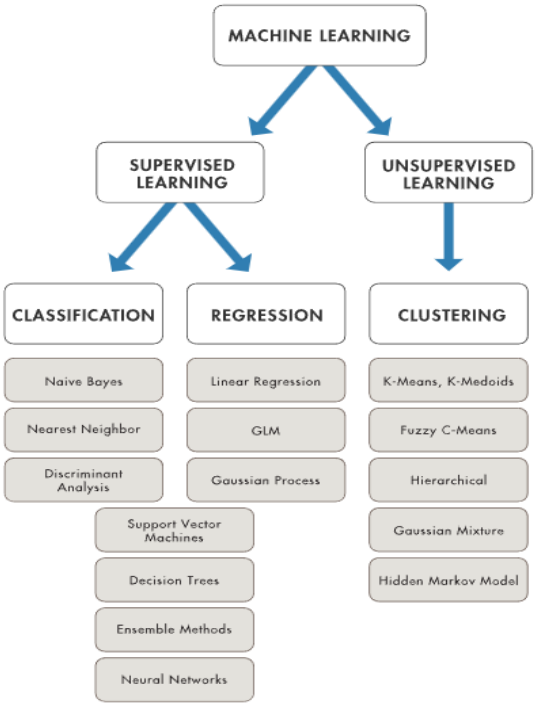

Figure 2: Machine learning techniques.

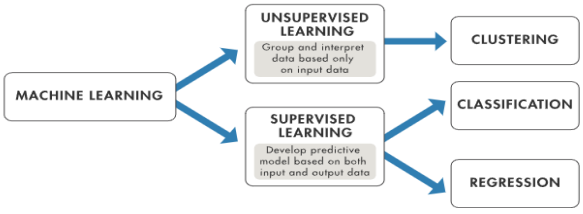

Figure 3: Machine learning techniques

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

RESULTS AND DISCUSSION

The proposed approach for behavior prediction leverages multiple advanced machine learning models, including Multi-Layer Perceptron's (MLP), Artificial Neural Networks (ANN) with DenseNet architecture, XGBoost, and AdaBoost, to create a robust and scalable predictive system. The methodology starts by gathering a wide range of data sources, such as sensor data, user interaction logs, social media activity, and demographic information. Preprocessing is performed to clean and standardize the data, ensuring that relevant features are extracted, which could represent behavioral patterns, preferences, or actions. MLP, being a feed-forward neural network with multiple hidden layers, is used to capture complex relationships in the data, learning non-linear patterns essential for behavior prediction.

ANN with DenseNet architecture is employed to optimize feature propagation through dense connections, improving the model's efficiency and accuracy, particularly when dealing with deep and complex datasets. XGBoost is incorporated for its superior performance in handling large datasets with high dimensionality, offering a robust solution for both classification and regression tasks through gradient boosting. AdaBoost, on the other hand, enhances the accuracy of the model by focusing on difficult-to-classify instances and improving prediction accuracy through sequential learning of weak learners. These models are trained using a combination of labeled datasets, with supervised learning methods applied for classification or regression, depending on the behavior being predicted.

Hyper parameter tuning and cross-validation are employed to optimize each model's performance and prevent over fitting. Furthermore, ensemble methods like bagging and boosting are integrated to increase the system's predictive power, providing more reliable and stable predictions over individual models. The final model is evaluated using precision, recall, accuracy, and F1 score, ensuring high performance. To enhance real-time prediction capabilities, a feedback loop is included in the system for continuous improvement, allowing the model to adapt to new data over time and enhance its predictive accuracy in dynamic environments. This multi-model approach aims to achieve high robustness and precision, making it applicable across various domains like personalized recommendations, emotion detection, and user behavior forecasting.

EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is a method for analyzing data sets in order that highlight their key properties. The use of statistical visualizations and several other information visualization techniques is commonplace in this context. Use a graph to represent the outcomes of the evaluation. By utilizing score selection, the final score column is selected and changed into classified features such as Good, Fair, and bad. A student is considered to be a good student if their final score falls between 15 and 20, a student is considered to be fair if their final score falls between 10 and 14, and a student is considered to be bad if their final score falls between 0 and 9. Moreover, a correlation matrix is utilized in this process. In order to gets a final grade depending on the degree of romantic compatibility between two individuals, with fair receiving a higher score. Additionally, it has been utilized to allocate study time according to age and the level of interest in pursuing higher education.

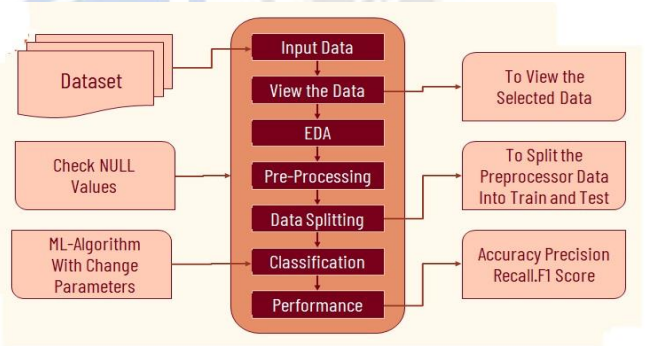

Figure 4: proposed flow diagram

CLASSIFICATION

Classification is the task of predicting a categorical label based on the given features. In the given example, the objective is to predict student achievement based on factors like study time, parental education, etc. Different algorithms, such as **Multilayer Perceptron (MLP), XGBoost, AdaBoost**, and **DenseNet**, can be used for classification, depending on the problem complexity. These classifiers learn the decision boundaries between different classes during training and make predictions for new, unseen data based on these learned patterns.

RESULTS ANALYSIS

The results of an analysis of cross-site scripting have been suggested in this section. The analysis was carried out with machine learning techniques, which preprocessed and encoded the data. The research was

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

carried out on a personal computer operating Windows 10 and equipped with 16.0 GB of RAM and a 2.6 GHz Intel Core i7-9750H processor. In each and every one of the studies, the Jupyter simulation environment was utilized, and Python was the primary programming language that was utilized. In addition to that, a number of Python libraries were utilized. In this section, we also cover the metrics that can be used to evaluate the performance of various machine learning models. EDA simulation, visualization Collect data from publicly available sources used for student behaviour prediction with three classes like [fair, good and poor]. The focus of the data provided by two Indore School in the fields of mathematics and Science subjects were used on the overall result, which is used to characterize student performance and generate labels for categorization and prediction with behaviour. It possesses a wide variety of features and columns. Choose the final grade based on score selection, and divide the students into three groups for behaviour likely: good, fair, and poor. A good student has a grade between 15 and 20, while a fair student has a grade between 10 and 14, and a poor student has a grade between 0 and 9. (A poor student). After that, proceed with EDA.

For the implementation of Student Behaviours Predictor we required the following libraries apart from basic libraries of python-

Sklearn

Tensorflow

Keras

Matplotlib

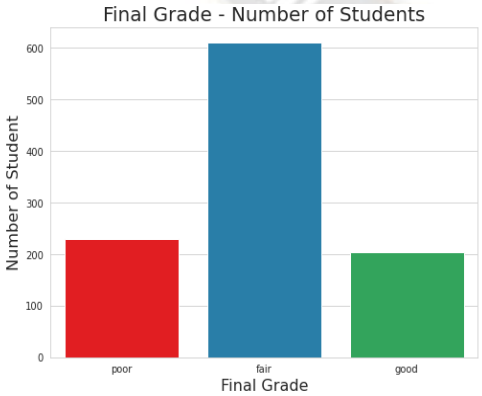

Figure 5: Final Grade Count Plot

This graph shows the grade. In order to categories features such as Good, Fair, and Poor, the final score columns is selected using score selection, with Fair receiving the highest value relative to Good and Poor.

EVALUATION METRICS

There are a lot of ways to measure the classifier's performance, which tally the number of accurate and inaccurate predictions that were generated based on values that is already known. A True Positive, abbreviated as TP, is one in which the model properly predict correct class. True Negative (TN) is a situation in which model properly predicts negative class. It is possible to have a False Positive, also known as an FP is one in which the model erroneously predict correct class. False Negative also known as a FN is a situation in which model erroneously predicts negative class. In the proposed work, following evaluation metrics are used for performance assessment.

Accuracy: Accuracy is a measure of how frequently a model predicts the correct result based on the input.

Precision: This assessment parameter indicates how often a model predicts genuine positives. $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$

Recall: This parameter gives information regarding how often a model predicts false negatives. $\text{Recall} = \frac{\text{TP+TN}}{\text{TP+FN}}$

F1 Score: Precision and recall are combined to calculate the F1 score. That is, a high F1 score suggests a low number of false positives and false negatives, implying that the model detects true elements accurately and is unaffected by false alarms. $\text{F1 score} = 2 * \frac{\text{Precision*Recall}}{\text{Presiion+Recall}}$

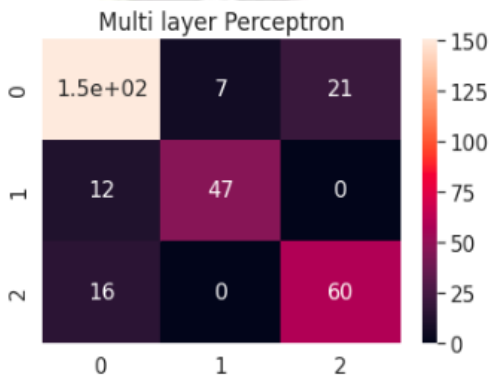

Figure 6: Confusion Matrix of MLP

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

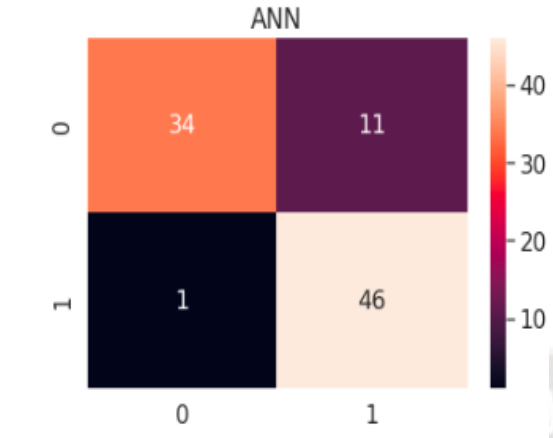

Figure 7: Confusion Matrix of ANN

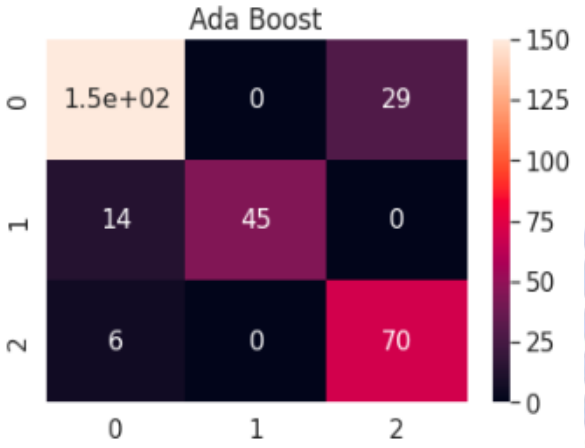

Figure 8: Confusion Matrix of Ada Boost

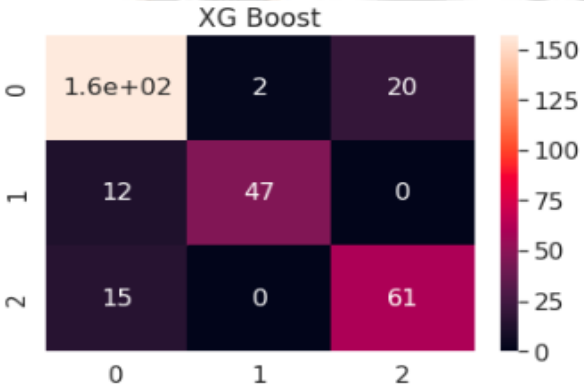

Figure 9: Confusion Matrix of XG-Boost

PERFORMANCE EVOLUTION

The evaluation of the model's performance can be seen in below table, which contains the models MLP, ANN DenseNet, Adaboost and Xgboost. The table's columns are labelled Precision, Recall, f1-Score and Accuracy. The ANN dense Net does has the lowest accuracy, which is 49, whereas the XgBoost and AdaBoost model does have the highest maximum accuracy, which is 88.

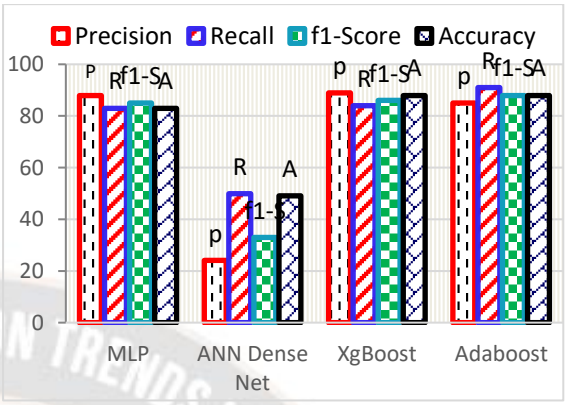

Figure 10. Performance of Models

CONCLUSION

This study successfully demonstrated the application of machine learning and deep neural network models to predict student academic performance and behavioural patterns in higher education. The proposed deep neural network algorithm involved key steps such as network initialization, data preprocessing, and the construction of hidden layers through meaningful feature extraction and weight assignment. Optimization techniques, including the Adam and RMSProp optimizers, were employed to enhance model accuracy and computational efficiency.

The experimental results highlight the practical value of these predictive models in assisting educational institutions with forecasting student outcomes, thereby facilitating early interventions and informed policy planning. Among the evaluated models—MLP, ANN DenseNet, XGBoost, and AdaBoost—XGBoost demonstrated superior overall performance across critical evaluation metrics such as accuracy, precision, recall, and F1-score, establishing its robustness and reliability. AdaBoost also showed strong, balanced performance. In contrast, MLP and ANN DenseNet exhibited less consistent results; notably, ANN DenseNet achieved high recall but lower precision and accuracy, indicating its potential usefulness for tasks prioritizing recall over balanced classification performance. These findings underscore the strengths of boosting-based techniques for educational data mining while identifying areas where deep learning architectures may require further enhancement

FUTURE SCOPE

Building on these findings, the research can be extended in several promising directions:

Hyper parameter Tuning : Employing advanced optimization strategies such as Randomized Search, Grid Search, or Bayesian Optimization could further refine model parameters, particularly improving ANN DenseNet and MLP performance.

Model Hybridization: Developing ensemble approaches that combine multiple models may yield improved accuracy and robustness, especially in addressing class imbalance or optimizing across multiple metrics.

Enhanced Feature Engineering: Further exploration of sophisticated feature extraction, selection, and creation of composite features could improve prediction quality, especially for deep learning models.

Deep Learning Architectures : Investigating architectures such as Convolutional Neural Networks (CNNs) for enhanced feature extraction or Recurrent Neural Networks (RNNs) for sequence data may enhance model capabilities, particularly for underperforming models.

Data Augmentation: Application of synthetic data generation or augmentation techniques can help mitigate challenges posed by limited or imbalanced datasets, reducing over fitting and improving generalization.

Real-World Deployment: Validating these models on live educational systems or other domains like healthcare diagnostics or predictive maintenance could assess practical effectiveness and guide further refinements for industry applications.

REFERENCES

[1]. Fekry, G. Dafoulas and M. Ismail, "Automatic detection for students behaviors in a group presentation," 2019 14th International Conference on Computer Engineering and Systems (ICCES), 2019, pp. 11-15, doi: 10.1109/ICCES48960.2019.9068128.

[2]. F. Orji and J. Vassileva, "Using Machine Learning to Explore the Relation between Student Engagement and Student Performance," 2020 24th International Conference Information Visualisation (IV), 2020, pp. 480-485, doi: 10.1109/IV51561.2020.00083.

[3]. N. Kondo, M. Okubo and T. Hatanaka, "Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data," 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 2017, pp. 198-201, doi: 10.1109/IIAI-AAI.2017.51.

[4]. R. Cao and L. Sun, "Design and Practice of Machine Learning Course Based on CDIO and Student Behavior Data," 2020 15th International Conference on Computer Science & Education (ICCSE), 2020, pp. 553-556, doi: 10.1109/ICCSE49874.2020.9201853.

[5]. M. M. Hassan, Z. J. Peya, S. Zaman, J. H. Angon, A. I. Keya and A. U. Dulla, "A Machine Learning Approach to Identify the Correlation and Association among the Students' Drug Addict Behavior," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225355.

[6]. T. Purwoningsih, H. B. Santoso and Z. A. Hasibuan, "Data Analytics of Students' Profiles and Activities in a Full Online Learning Context," 2020 Fifth International Conference on Informatics and Computing (ICIC), 2020, pp. 1-8, doi: 10.1109/ICIC50835.2020.9288540.

[7]. J. Edmond Meku Fotso, B. Batchakui, R. Nkambou and G. Okereke, "Algorithms for the Development of Deep Learning Models for Classification and Prediction of Behaviour in MOOCS.," 2020 IEEE Learning With MOOCS (LWMOOCS), 2020, pp. 180-184, doi: 10.1109/LWMOOCS50143.2020.9234363.

[8]. C. -Y. Huang-Fu, C. -H. Liao and J. -Y. Wu, "Comparing the performance of machine learning and deep learning algorithms classifying messages in Facebook learning group," 2021 International Conference on Advanced Learning Technologies (ICALT), 2021, pp. 347-349, doi: 10.1109/ICALT52272.2021.00111.

[9]. M. Adnan et al., "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models," in IEEE Access, vol. 9, pp. 7519-7539, 2021, doi: 10.1109/ACCESS.2021.3049446.

[10]. P. Kumar, V. Kumar and R. Sobti, "Predicting Joining Behavior of Freshmen Students using Machine Learning – A Case Study," 2020 International Conference on Computational Performance Evaluation (ComPE), 2020, pp. 141-145, doi: 10.1109/ComPE49325.2020.9200167.

[11]. D. Kelly and B. Tangney, "'First Aid for You': getting to know your learning style using machine learning," Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05), 2005, pp. 1-3, doi: 10.1109/ICALT.2005.1.

[12]. F. D. Pereira, E. H. T. Oliveira, D. Fernandes and A. Cristea, "Early Performance Prediction for CS1 Course Students using a Combination of Machine Learning and an Evolutionary Algorithm," 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), 2019, pp. 183-184, doi: 10.1109/ICALT.2019.00066.

[13]. M. N and J. S, "Intellectual Behaviour of Student Based on Education Data Determined by Opinion Mining," 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), 2021, pp. 559-564, doi: 10.1109/INDIACom51348.2021.00099.

[14]. H. Wan, Q. Yu, J. Ding and K. Liu, "Students' behavior analysis under the Sakai LMS," 2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE), 2017, pp. 250-255, doi: 10.1109/TALE.2017.8252342.

[15]. S. Lai, B. Sun, F. Wu and R. Xiao, "Automatic Personality Identification Using Students' Online Learning Behavior," in IEEE Transactions on Learning Technologies, vol. 13, no. 1, pp. 26-37, 1 Jan.-March 2020, doi: 10.1109/TLT.2019.2924223.

[16]. K. J. de O. Santos, A. G. Menezes, A. B. de Carvalho and C. A. E. Montesco, "Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout," 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), 2019, pp. 207-208, doi: 10.1109/ICALT.2019.00068.

[17]. T. Purwoningsih, H. B. Santoso and Z. A. Hasibuan, "Online Learners' Behaviors Detection Using Exploratory Data Analysis and Machine Learning Approach," 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019, pp. 1-8, doi: 10.1109/ICIC47613.2019.8985918.

[18]. A. Balakrishnan, "Learning Student Models through an Ontology of Learning Strategies," International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), 2007, pp. 133-135, doi: 10.1109/ICCIMA.2007.171.

[19]. K. Thakur, K. Lal and V. Kumar, "Ensemble method to predict impact of student intelligent quotient and academic achievement on placement," 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), 2021, pp. 249-253, doi: 10.1109/ICIEM51511.2021.9445323.

[20]. C. Lwande, L. Muchemi and R. Oboko, "Behaviour Prediction in a Learning Management System," 2019 IST-Africa Week Conference (IST-Africa), 2019, pp. 1-10, doi: 10.23919/ISTAFRICA.2019.87648