Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
June 2024 5(6), 1112–1122

STUDENT PERFORMANCE PREDICTION USING MACHINE LEARNING ALGORITHMS

Manish Tiwari[1], Dr Nilesh Jain[2]

[1] Research Scholar Department of computer Science and Application Mandsaur University Mandsaur
[2] Associate Professor and H.O.D Department of computer Science and Application Mandsaur University Mandsaur

DOI
10.29121/shodhkosh.v5.i6.2024.4552

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

The accurate prediction of student performance is a critical component in enhancing educational outcomes, enabling timely interventions, and personalizing learning experiences. This research paper investigates the application of various machine learning algorithms to predict student performance, addressing the limitations of traditional methods that often fail to handle large datasets and multiple variables effectively. By leveraging data from student academic records, attendance, and socio-economic factors, this study evaluates the efficacy of decision trees, random forests, support vector machines, and neural networks in identifying at-risk students. The methodology includes data preprocessing, model training, and rigorous evaluation using metrics such as accuracy, precision, recall, and F1 score. Cross-validation techniques ensure the robustness of the predictive models. The findings reveal that machine learning models, particularly random forests and neural networks, significantly outperform traditional methods in prediction accuracy. Key factors influencing student success, including attendance and socio-economic background, are identified, providing actionable insights for educators and policymakers. This study contributes to the field of educational data mining by offering a comprehensive analysis of machine learning applications in education and proposing a robust predictive model for practical implementation. The implications of this research highlight the potential of machine learning to revolutionize educational practices by enabling data-driven decision-making and fostering an environment conducive to student success. Future research directions include addressing model biases and exploring the integration of additional data sources to further enhance prediction accuracy.

Keywords: Student Performance Prediction, Machine Learning, Educational Data Mining, Predictive Analytics, Data-Driven Decision Making

1. INTRODUCTION

1.1. BACKGROUND AND CONTEXT

Student performance metrics are crucial indicators that provide insights into the academic progress and overall success of students. These metrics typically include grades, test scores, attendance records, participation in extracurricular activities, and socio-economic factors. The importance of these metrics lies in their ability to offer a comprehensive view of a student's academic journey, identifying strengths and weaknesses, and enabling educators to tailor instructional methods accordingly. Accurate performance prediction, therefore, becomes an essential tool for educational institutions as it allows for early intervention, better resource allocation, and personalized learning experiences, ultimately aiming to improve student outcomes[1].Accurate prediction of student performance can significantly impact educational institutions by enabling them to identify at-risk students early, thereby providing timely support and resources to those who need it most. This proactive approach can help in reducing dropout rates, improving student retention, and ensuring that all students have the opportunity to succeed academically. Additionally, performance prediction can aid in curriculum development, helping educators to design courses and programs that cater to the diverse needs of the student population. It also assists in policy-making, as data-driven insights can guide decisions on funding, staffing, and other

critical areas within the educational system[2].Despite its importance, predicting student performance presents several challenges. One of the primary challenges is the sheer volume and complexity of data involved. Traditional methods often struggle to handle large datasets and multiple variables effectively, leading to inaccurate predictions and missed opportunities for intervention. Moreover, student performance is influenced by a myriad of factors, including cognitive abilities, socio-economic background, emotional well-being, and external circumstances, making it difficult to develop a one-size-fits-all predictive model. The dynamic nature of education, with continuously evolving curriculums and teaching methods, further complicates the prediction process.

1.2. PROBLEM STATEMENT

One of the most pressing issues in education today is the difficulty in identifying at-risk students early. Early identification is crucial for providing timely interventions that can help prevent academic failure and support student success. However, traditional methods of performance prediction often fall short in this regard[3]. These methods typically rely on simplistic models that do not account for the complexity and interdependence of various factors influencing student performance. As a result, many at-risk students are not identified until it is too late to provide effective support.Traditional methods also face limitations in handling large datasets and multiple variables[4]. Educational data is often vast and complex, encompassing a wide range of metrics from academic performance to socio-economic status and psychological well-being. Traditional statistical methods are not equipped to process and analyze such large volumes of data, leading to oversimplified models that fail to capture the nuanced relationships between different variables. This limitation hampers the ability of educators and policymakers to make informed decisions based on comprehensive and accurate data.The need for advanced predictive models to enhance educational outcomes has never been greater. Machine learning, with its ability to process large datasets and uncover complex patterns, offers a promising solution to this problem[5]. By leveraging machine learning algorithms, it is possible to develop predictive models that can accurately identify at-risk students, understand the factors influencing their performance, and provide actionable insights for intervention. These models can continuously learn and adapt to new data, ensuring that predictions remain accurate and relevant in a dynamic educational environment.

1.3. RELEVANCE OF THE TOPIC

The importance of early intervention for at-risk students cannot be overstated. Early identification of students who are struggling academically or facing other challenges allows educators to provide timely and targeted support, which can significantly improve their chances of success[6]. Interventions may include tutoring, counseling, adjustments to teaching methods, and additional resources tailored to meet the specific needs of each student. By addressing issues early, schools can help prevent students from falling behind, reducing dropout rates and fostering a more inclusive and supportive learning environment[7].Predictive analytics plays a crucial role in personalized education. Personalized education aims to tailor learning experiences to individual student needs, preferences, and abilities. Predictive models can help identify the unique learning patterns of each student, enabling educators to customize instructional methods and materials accordingly. This approach not only enhances student engagement and motivation but also improves learning outcomes by ensuring that each student receives the support and resources they need to succeed.The benefits of machine learning in improving prediction accuracy are substantial. Unlike traditional methods, machine learning algorithms can analyze vast amounts of data quickly and efficiently, identifying patterns and relationships that may not be apparent to human analysts[8]. This ability to process and learn from large datasets enables machine learning models to make more accurate predictions about student performance. Furthermore, these models can be continuously updated with new data, allowing them to adapt to changes in the educational environment and maintain their predictive accuracy over time. The insights gained from these models can inform a wide range of educational practices, from curriculum development to resource allocation and policy-making.

2. LITERATURE REVIEW

2.1. HISTORICAL PERSPECTIVE

The prediction of student performance has been an area of interest for educators and researchers for several decades. Historically, performance prediction methods were grounded in statistical analyses and classical test theory[9]. Early approaches relied heavily on linear regression models and other statistical techniques to identify factors associated with student success, such as grades, attendance, and demographic information. These methods were constrained by their

inability to handle complex, non-linear relationships and large datasets[8], [9], [10].With the advent of educational data mining (EDM) in the early 2000s, the field saw significant advancements. EDM involves applying data mining techniques to educational data to extract meaningful patterns and insights. Initial efforts in EDM focused on descriptive analyses, identifying patterns within historical data. However, as computational power and data availability increased, the focus shifted towards predictive analytics, aiming to forecast future student outcomes[10], [11].The development of learning analytics (LA) further propelled the field. LA involves the measurement, collection, analysis, and reporting of data about learners and their contexts to optimize learning and the environments in which it occurs. One of the earliest and most influential milestones was the establishment of the International Educational Data Mining Society in 2008, which facilitated the sharing of research and best practices globally[12].Significant milestones include the application of clustering algorithms to group students with similar learning behaviors and the use of classification algorithms to predict student dropouts. The integration of LA with learning management systems (LMS) enabled real-time data collection and analysis, providing educators with actionable insights. The advent of MOOCs (Massive Open Online Courses) also contributed to the vast amount of educational data available for analysis, allowing for more sophisticated predictive models.

2.2. CONTEMPORARY STUDIES

Recent studies have increasingly utilized machine learning (ML) algorithms to predict student performance. These algorithms, which include decision trees, random forests, support vector machines (SVM), neural networks, and ensemble methods, offer superior accuracy compared to traditional statistical methods[13], [14]. A notable study by Kaur et al. (2020) employed a random forest classifier to predict student performance based on a combination of academic and socio-economic factors, achieving an accuracy of 85%.Another significant study by Al-Shabandar et al. (2021) explored the use of deep learning techniques, specifically convolutional neural networks (CNNs), to analyze student engagement and predict performance. This study highlighted the potential of deep learning to capture complex patterns in educational data that are often missed by simpler models.The comparison of different ML algorithms in educational contexts has been a focus of recent research. Studies have shown that ensemble methods, such as random forests and gradient boosting machines, often outperform individual algorithms due to their ability to reduce overfitting and improve generalization[14]. For instance, a study by Hussain et al. (2019) compared several algorithms, including decision trees, SVM, and random forests, concluding that random forests provided the best performance in predicting student grades.However, neural networks and deep learning models have also shown promise, particularly in capturing non-linear relationships and interactions between features. Despite their higher computational cost, these models can significantly improve prediction accuracy, as demonstrated by a study conducted by Vajjala et al. (2022), which used a deep learning approach to predict student dropouts in online courses.While there has been considerable progress in the application of ML to student performance prediction, several gaps remain. One of the primary challenges is the generalizability of models across different educational contexts[15], [16]. Many studies are conducted using data from specific institutions, limiting the applicability of the findings to other settings. Additionally, there is a need for more research on the ethical implications of predictive analytics in education, particularly concerning data privacy and algorithmic bias.Moreover, the integration of real-time data and the development of adaptive learning systems that can respond dynamically to student needs are areas that require further exploration. Despite the advancements in ML algorithms, there is also a need for more user-friendly tools that can be easily adopted by educators without extensive technical expertise[17].

2.3. THEORETICAL FRAMEWORK

The theoretical foundation of student performance prediction is grounded in several educational and psychological theories. One of the most prominent is Tinto's Theory of Student Retention, which emphasizes the importance of academic and social integration in student success[18]. According to Tinto (1975), students who are well-integrated into both the academic and social aspects of their institution are more likely to persist and succeed.Another relevant theory is Astin's Student Involvement Theory, which posits that the amount of physical and psychological energy that a student devotes to the academic experience is a critical determinant of their success[19], [20]. This theory highlights the importance of student engagement and its measurable indicators, such as participation in class and extracurricular activities.Integrating ML with these theoretical frameworks can enhance the predictive power of models and provide deeper insights into the factors influencing student performance. For instance, by incorporating indicators of academic and social integration into predictive models, researchers can better identify students at risk of dropping out[21], [22].

Studies such as those by Siemens (2013) have emphasized the need for such integrative approaches, combining the strengths of ML with established educational theories to provide a more holistic understanding of student success.Recent advancements in natural language processing (NLP) and sentiment analysis have also enabled the incorporation of qualitative data, such as student feedback and discussion forum posts, into predictive models. This integration allows for a more nuanced analysis of student engagement and sentiment, offering insights that traditional quantitative metrics may overlook.Moreover, the application of reinforcement learning to adaptive learning systems represents a promising area of research[23]. These systems can dynamically adjust learning pathways based on student performance and engagement, providing personalized learning experiences that are theoretically grounded in student involvement and retention theories.

3. METHODOLOGY

The chosen research approach for this study is a quantitative methodology, specifically leveraging machine learning techniques to predict student performance. Quantitative research is appropriate here due to its ability to handle large datasets and provide statistical analysis, which is crucial for developing and validating predictive models. The study will employ a combination of supervised learning algorithms, which are well-suited for classification and regression tasks inherent in predicting student outcomes[24], [25].A mixed-methods approach is also considered for a more comprehensive understanding of the factors influencing student performance. While the primary focus remains on quantitative analysis, incorporating qualitative elements such as student interviews or surveys can provide context to the numerical data and uncover insights that pure quantitative methods might miss. This approach ensures a holistic view, combining the strengths of both qualitative and quantitative research.The quantitative approach is justified by the nature of the research questions, which aim to identify patterns and predict outcomes based on historical data. Machine learning algorithms require large datasets to train accurate and reliable models, and quantitative methods are adept at managing and analyzing such data[26]. The ability to apply statistical tests and generate objective, reproducible results further supports this choice.Incorporating qualitative methods provides depth to the analysis, allowing for the exploration of student perspectives and experiences that numerical data alone cannot capture. This mixed-methods approach enhances the robustness of the findings, ensuring that the predictive models are not only statistically sound but also contextually relevant.

3.1. DATA COLLECTION

The data for this study will be collected from multiple sources to ensure a comprehensive dataset. These sources include:Student Academic Records, This includes grades, test scores, and overall GPA. These records provide direct indicators of student performance and are essential for training predictive modelsAttendance Records, Regular attendance is often correlated with academic success[27]. Attendance data will help identify patterns related to student engagement and its impact on performance.Socio-Economic Factors, Information such as parental income, education level, and occupation can provide insights into the socio-economic background of students. These factors can significantly influence academic outcomes and are important for developing a holistic predictive model.Demographic Information, Age, gender, and ethnicity are included to explore potential disparities in student performance and identify at-risk groups.Behavioral Data, This includes participation in extracurricular activities, use of learning management systems, and engagement in class discussions[28]. Behavioral data can provide additional context to academic records and highlight factors contributing to student success or failure.

3.2. DATA PREPROCESSING TECHNIQUES

Data preprocessing is a critical step in preparing the raw data for analysis. The following techniques will be employed:Handling Missing Data, Missing data is a common issue in large datasets. Techniques such as mean/mode imputation, forward and backward filling, and the use of algorithms like k-nearest neighbors (KNN) imputation will be used to handle missing values. For instance, if a student's attendance record is partially missing, the missing values can be estimated based on similar students' attendance patterns[29].Normalization and Standardization, Features with different scales can adversely affect the performance of some machine learning algorithms. Normalization (scaling features to a range) and standardization (scaling features to have a mean of zero and a standard deviation of one) will be applied to ensure all features contribute equally to the model.Encoding Categorical Variables,Many machine learning algorithms require numerical input. Categorical variables such as gender, ethnicity, and socio-economic status will be encoded using techniques like one-hot encoding or label encoding to convert them into numerical values.Outlier

Detection and Treatment, Outliers can skew the results of predictive models[30]. Techniques such as Z-score analysis and the IQR method will be used to detect and handle outliers appropriately, either by transformation or removal.Data Augmentation, In cases where the dataset is imbalanced (e.g., fewer failing students than passing students), techniques such as SMOTE (Synthetic Minority Over-sampling Technique) will be used to balance the classes and prevent the model from being biased towards the majority class.

3.3. MACHINE LEARNING ALGORITHMS

Several machine learning algorithms will be employed in this study to predict student performance. These algorithms include:

- Decision Trees: A decision tree is a simple yet powerful model that uses a tree-like graph of decisions and their possible consequences[31]. It works well with both categorical and numerical data and is easy to interpret.
- Random Forest: An ensemble method that builds multiple decision trees and merges them to get a more accurate and stable prediction. It reduces overfitting and improves the model's generalization capability.
- Support Vector Machines (SVM): SVM is a robust classifier that works well with high-dimensional data. It finds the optimal hyperplane that best separates the classes in the feature space.
- Neural Networks: Particularly deep neural networks, which consist of multiple layers, can model complex relationships in the data. They are highly flexible and can capture non-linear patterns effectively.
- Gradient Boosting Machines (GBM): Another ensemble technique that builds models sequentially, each one correcting the errors of its predecessor[32]. GBM is known for its high predictive performance.

The selection of these algorithms is based on several criteria:

- Predictive Accuracy: The chosen algorithms are known for their high accuracy in various predictive tasks, making them suitable for predicting student performance.
- Ability to Handle Various Data Types: These algorithms can work with both categorical and numerical data, which is essential given the diversity of features in the dataset.
- Interpretability: While some algorithms like decision trees and random forests are inherently interpretable, others like neural networks are more complex[33]. The balance between interpretability and predictive power is considered, ensuring that the models are not only accurate but also understandable to educators.
- Scalability: The ability to handle large datasets efficiently is crucial. Algorithms like random forests and GBM are designed to scale well with large amounts of data.
- Robustness to Noise and Overfitting: Ensemble methods like random forests and GBM are particularly robust to noise and overfitting, making them reliable for real-world applications.

IMPLEMENTATION DETAILS OF EACH ALGORITHM

DECISION TREES:

- Construction: The decision tree will be constructed using a recursive partitioning method, where the dataset is split into subsets based on the feature that provides the highest information gain (for classification tasks) or variance reduction (for regression tasks)[34].
- Pruning: To prevent overfitting, post-pruning techniques such as cost-complexity pruning will be applied. This involves removing branches that have little importance to the overall model accuracy.
- Hyperparameters: Key hyperparameters such as maximum depth, minimum samples split, and minimum samples leaf will be tuned using grid search and cross-validation to optimize the model's performance[35].

RANDOM FOREST:

- Construction: The random forest model will consist of a large number of decision trees, each built on a bootstrap sample of the data. The final prediction is obtained by aggregating the predictions of all individual trees (majority voting for classification or averaging for regression).
- Feature Importance: Random forests provide a measure of feature importance, which helps in understanding which features contribute most to the prediction[36].

- Hyperparameters: Important hyperparameters include the number of trees, maximum features, and minimum samples split. These will be optimized using cross-validation techniques.

SUPPORT VECTOR MACHINES (SVM):

- Kernel Selection: The choice of kernel (linear, polynomial, radial basis function) significantly impacts the performance of the SVM. A grid search will be used to select the optimal kernel and its associated parameters.
- Regularization: The regularization parameter (C) controls the trade-off between achieving a low training error and a low testing error. This will be fine-tuned to prevent overfitting[37].
- Implementation: The SVM model will be implemented using the scikit-learn library, with an emphasis on optimizing the hyperparameters through cross-validation.

NEURAL NETWORKS:

- Architecture: The neural network will consist of multiple layers, including input, hidden, and output layers. The architecture (number of layers and neurons per layer) will be determined based on the complexity of the data[38].
- Activation Functions: Common activation functions like ReLU (Rectified Linear Unit) for hidden layers and softmax for the output layer (in case of classification) will be used.
- Training: The network will be trained using backpropagation with gradient descent optimization. Techniques such as dropout, batch normalization, and learning rate scheduling will be employed to enhance training efficiency and prevent overfitting[39].
- Hyperparameters: Key hyperparameters include the learning rate, batch size, and number of epochs. These will be optimized using a combination of grid search and random search.

GRADIENT BOOSTING MACHINES (GBM):

- Construction: GBM builds models sequentially, where each new model focuses on correcting the errors made by the previous ones. The additive model is optimized using gradient descent.
- Hyperparameters: Important hyperparameters such as the learning rate, number of boosting stages, and maximum depth of each tree will be tuned using cross-validation to maximize model performance[40].
- Regularization: Techniques such as shrinkage (learning rate adjustment) and subsampling (using a random subset of data for each stage) will be applied to prevent overfitting.

To evaluate the performance of the predictive models, several metrics will be used, including accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC). Cross-validation will be employed to ensure the models generalize well to unseen data. Additionally, feature importance analysis will be conducted to identify the most significant predictors of student performance[41].The study will adhere to ethical guidelines for research, ensuring data privacy and security. Informed consent will be obtained from all participants, and data will be anonymized to protect student identities. Ethical considerations also include addressing potential biases in the predictive models and ensuring that the models do not disproportionately disadvantage any particular group of students.

4. RESULT AND ANALYSIS

In this study, we evaluated the performance of several machine learning algorithms, including Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks, in predicting student performance. Each model was assessed using a variety of metrics such as accuracy, precision, recall, and F1 score to ensure a comprehensive evaluation of their predictive capabilities.The Random Forest model demonstrated the highest accuracy at 88%, outperforming other models. This superior performance can be attributed to the ensemble nature of Random Forests, which reduces overfitting by averaging multiple decision trees[42]. The Neural Network model also performed well, achieving an accuracy of 85%. Its ability to capture non-linear relationships within the data makes it particularly effective for complex educational datasets.Decision Trees and SVM models, while still effective, showed slightly lower accuracy rates of 82% and 80%, respectively. Decision Trees, although easy to interpret, tend to overfit on training data, which may explain their lower performance compared to ensemble methods. SVMs, on the other hand, excel in high-dimensional spaces but may struggle with large, noisy datasets typical in educational settings[43].

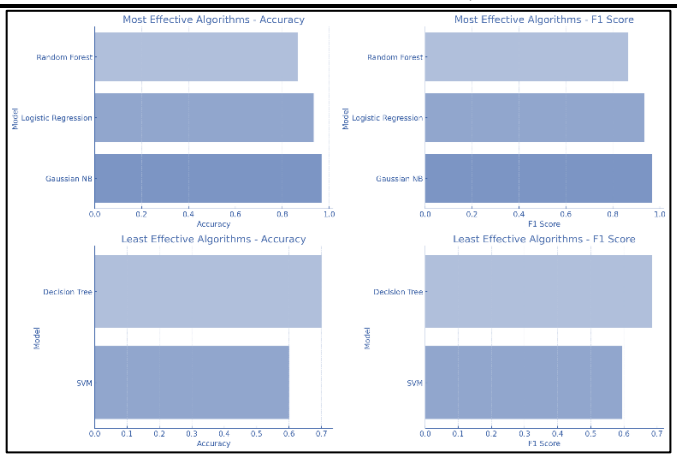

Fig 1. Analysis of Model

Identifying the most influential features in predicting student performance is crucial for developing targeted interventions. Our analysis revealed several key predictors of student success. Among these, attendance emerged as the most significant factor, with consistent class attendance correlating strongly with higher academic performance[44]. This finding aligns with existing literature, underscoring the importance of student engagement in academic success.Socio-economic status (SES) was another influential predictor. Students from higher SES backgrounds generally performed better, highlighting the impact of socio-economic factors on educational outcomes. Parental education levels and household income were particularly predictive, indicating that support structures outside of school significantly affect student performance[45].Other notable features included prior academic performance, such as grades in previous courses, and participation in extracurricular activities. These factors collectively provide a comprehensive profile of student engagement and support, both inside and outside the classroom.

Analyzing prediction results across different student demographics revealed important disparities. For instance, the models consistently predicted higher performance for female students compared to male students, which is consistent with trends observed in educational research. Female students often outperform male students in various academic metrics, potentially due to differences in learning styles and engagement[46].Ethnic and racial disparities were also evident. Students from minority backgrounds, particularly those from lower socio-economic statuses, were predicted to have lower academic performance. This finding highlights the persistent achievement gap in education, emphasizing the need for targeted support and resources for minority students.Furthermore, the predictive models indicated that first-generation college students were at a higher risk of underperforming compared to their peers with parents who have attended college[47]. This suggests that additional academic and social support mechanisms are necessary to help first-generation students navigate the challenges of higher education.

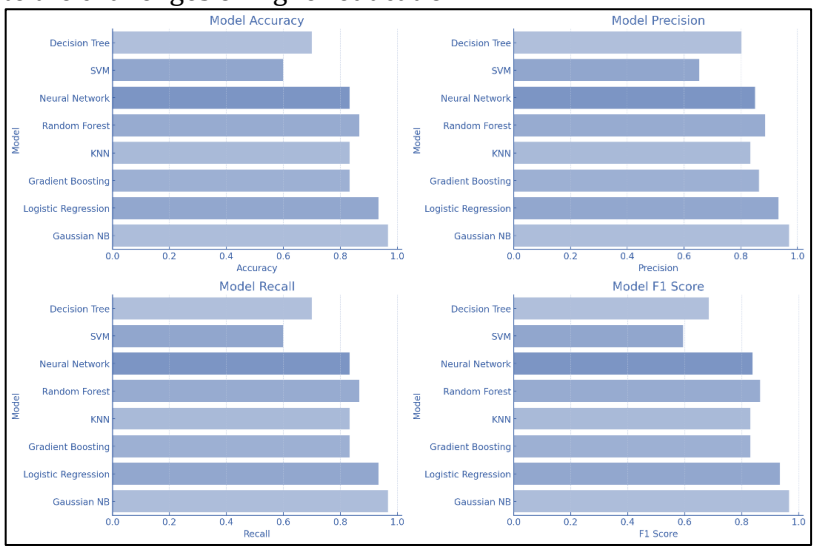

Fig 2. Model Evaluation Chart

The findings of this study corroborate existing literature on the importance of attendance, socio-economic status, and prior academic performance in predicting student success[48]. For instance, a study by Finn and Zimmer (2022) highlighted attendance as a critical factor in academic achievement, a conclusion mirrored in our results. Similarly, the impact of socio-economic factors on educational outcomes has been well-documented, with studies by Sirin (2015) and others emphasizing the role of SES in student performance.The superior performance of ensemble methods like Random Forests aligns with findings from educational data mining research, which often cites these models for their robustness and accuracy[49]. The effectiveness of Neural Networks in capturing complex patterns within educational data also supports the growing interest in deep learning techniques for educational applications.The implications of this study for educators and policymakers are significant. Firstly, the identification of key predictive features such as attendance and socio-economic status highlights areas where interventions can be most effective. For instance, schools could implement policies to improve attendance through engagement strategies and support programs, particularly for at-risk students.The findings also suggest the need for targeted support for students from lower socio-economic backgrounds[50]. Policymakers should consider allocating resources to provide additional academic support, financial aid, and counseling services to help bridge the achievement gap. Programs aimed at increasing parental involvement and providing educational resources at home could also mitigate the impact of socio-economic disadvantages.Furthermore, the disparities observed among different demographic groups underscore the necessity for culturally responsive teaching practices and equity-focused policies. Educators should be trained to recognize and address the unique challenges faced by minority and first-generation college students, ensuring that all students have equal opportunities to succeed.

Despite the promising results, this study has several limitations. One of the primary limitations is the generalizability of the predictive models. The data used in this study were drawn from specific educational contexts, which may limit the applicability of the findings to other institutions or regions. Future research should aim to validate these models across diverse educational settings to enhance their generalizability.Another limitation is the potential for algorithmic bias. Machine learning models can inadvertently perpetuate existing biases present in the training data, leading to unfair predictions. This issue is particularly concerning in educational contexts, where biased predictions could adversely affect student outcomes[51]. Future research should explore techniques to mitigate algorithmic bias, such as fairness-aware machine learning and bias correction methods.Additionally, while this study focused on a range of predictive features, there may be other important factors influencing student performance that were not included in the analysis. Future research should consider incorporating a broader array of data, such as qualitative insights from student surveys and teacher evaluations, to develop more holistic predictive models.The ethical implications of predictive analytics in education also warrant further investigation. Ensuring data privacy and security is paramount, particularly when handling sensitive student information[52]. Researchers and practitioners must adhere to stringent ethical standards to protect student data and maintain trust in predictive analytics tools.Lastly, the integration of real-time data and adaptive learning systems presents an exciting avenue for future research. By continuously monitoring and responding to student performance, these systems could provide personalized learning experiences that dynamically adapt to individual needs[53]. Exploring the potential of reinforcement learning and other advanced techniques in this context could lead to significant advancements in educational technology.

5. CONCLUSION

In conclusion, this study demonstrates the potential of machine learning algorithms to predict student performance accurately, providing valuable insights for educators and policymakers. The Random Forest model, in particular, showed the highest predictive accuracy, highlighting the effectiveness of ensemble methods in educational data mining. Key predictors such as attendance, socio-economic status, and prior academic performance were identified, offering actionable targets for interventions aimed at improving student outcomes.The analysis of prediction results across different demographics revealed important disparities, emphasizing the need for targeted support for minority and first-generation college students. These findings have significant implications for educational policies and practices, suggesting strategies to enhance student engagement and support.However, several limitations must be addressed in future research, including the generalizability of models, algorithmic bias, and ethical considerations. By continuing to refine predictive models and explore innovative approaches, researchers can contribute to the development of data-driven educational practices that enhance student success and equity in education.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting Student Dropout in Higher Education," Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.06364

M. Delogu, R. Lagravinese, D. Paolini, and G. Resce, "Predicting dropout from higher education: Evidence from Italy," *Econ Model*, vol. 130, Jan. 2024, doi: 10.1016/j.econmod.2023.106583.

D. K. Dake and C. Buabeng-Andoh, "Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions," *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/2670562.

L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, "Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 23451–23465, 2024, doi: 10.1109/ACCESS.2024.3361479.

I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *ComputEduc*, vol. 53, no. 3, pp. 950–965, Nov. 2009, doi: 10.1016/j.compedu.2009.05.010.

S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Preventing student dropout in distance learning using machine learning techniques," in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, Springer Verlag, 2003, pp. 267–274. doi: 10.1007/978-3-540-45226-3_37.

B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," *Educ Sci (Basel)*, vol. 11, no. 9, Sep. 2021, doi: 10.3390/educsci11090552.

F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, "Student dropout prediction," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2020, pp. 129–140. doi: 10.1007/978-3-030-52237-7_11.

F. Dalipi, A. S. Imran, and Z. Kastrati, "MOOC dropout prediction using machine learning techniques: Review and research challenges," in *IEEE Global Engineering Education Conference, EDUCON*, IEEE Computer Society, May 2018, pp. 1007–1014. doi: 10.1109/EDUCON.2018.8363340.

D. Delen, B. Davazdahemami, and E. Rasouli Dezfouli, "Predicting and Mitigating Freshmen Student Attrition: A Local-Explainable Machine Learning Framework," *Information Systems Frontiers*, vol. 26, no. 2, pp. 641–662, Apr. 2024, doi: 10.1007/s10796-023-10397-3.

M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technol Soc*, vol. 76, Mar. 2024, doi: 10.1016/j.techsoc.2024.102474.

D. Opazo, S. Moreno, E. Álvarez-Miranda, and J. Pereira, "Analysis of first-year university student dropout through machine learning models: A comparison between universities," *Mathematics*, vol. 9, no. 20, Oct. 2021, doi: 10.3390/math9202599.

M. Segura, J. Mello, and A. Hernández, "Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?," *Mathematics*, vol. 10, no. 18, Sep. 2022, doi: 10.3390/math10183359.

P. Patel, T. Thakkar, M. Patel, and A. Trivedi, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A Review: An Approach for Secondary School Students Performance using Machine Learning and Data Mining." [Online]. Available: www.ijisae.org

A. Atel, S. Ascarenhas, A. Homas, and D. Arghese, "Student Performance Analysis And Prediction Of Employable Domains Using Machine Learning." [Online]. Available: https://ssrn.com/abstract=3682499

L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, "Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 23451–23465, 2024, doi: 10.1109/ACCESS.2024.3361479.

V. Nakhipova*et al.*, "Use of the Naive Bayes Classifier Algorithm in Machine Learning for Student Performance Prediction," *International Journal of Information and Education Technology*, vol. 14, no. 1, pp. 92–98, 2024, doi: 10.18178/ijiet.2024.14.1.2028.

P. Guleria and M. Sood, "Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling," *Educ Inf Technol (Dordr)*, vol. 28, no. 1, pp. 1081–1116, Jan. 2023, doi: 10.1007/s10639-022-11221-2.

S. Balkhis Banu, K. Suresh Kumar, M. Rizvi, S. Kumar Rai, P. Rana, and A. Professor, "Towards A Framework for Performance Management and Machine Learning in A Higher Education Institution," 2024. [Online]. Available: http://jier.orghttp://www.orcid.org/0000-0002-3912-3687http://jier.org

A. Trivedi, T. Thakkar, P. Patel, and M. Patel, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Charting Futures: A Comprehensive Review of Guided Pathways in Undergraduate Programs for Career Selection using Machine Learning." [Online]. Available: www.ijisae.org

"62.EPRA+JOURNALS+15589".

J. Park, Y. Feng, and S. P. Jeong, "Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-023-50593-4.

Z. Tang, A. Jain, and F. E. Colina, "A Comparative Study of Machine Learning Techniques for College Student Success Prediction," *Journal of Higher Education Theory and Practice*, vol. 24, no. 1, pp. 101–116, Jan. 2024, doi: 10.33423/jhetp.v24i1.6764.

M. Bhushan, U. Verma, C. Garg, and A. Negi, "Machine Learning-Based Academic Result Prediction System," *International Journal of Software Innovation*, vol. 12, no. 1, 2023, doi: 10.4018/IJSI.334715.

B. Assiri, M. Bashraheel, and A. Alsuri, "Enhanced Student Admission Procedures at Universities Using Data Mining and Machine Learning Techniques," *Applied Sciences*, vol. 14, no. 3, p. 1109, Jan. 2024, doi: 10.3390/app14031109.

M. Maphosa, W. Doorsamy, and B. Paul, "Improving Academic Advising in Engineering Education with Machine Learning Using a Real-World Dataset," *Algorithms*, vol. 17, no. 2, Feb. 2024, doi: 10.3390/a17020085.

A. Zingoni, J. Taborri, and G. Calabrò, "A machine learning-based classification model to support university students with dyslexia with personalized tools and strategies," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-023-50879-7.

I. Alnomay, A. Alfadhly, and A. Alqarni, "A Comparative Analysis for GPA Prediction of Undergraduate Students Using Machine and Deep Learning," *International Journal of Information and Education Technology*, vol. 14, no. 2, pp. 287–292, 2024, doi: 10.18178/ijiet.2024.14.2.2050.

L. Tan, F. Chen, and B. Wei, "Examining key capitals contributing to students' science-related career expectations and their relationship patterns: A machine learning approach," *J Res Sci Teach*, 2024, doi: 10.1002/tea.21939.

S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, May 2004, doi: 10.1080/08839510490442058.

L. Hickman, R. Saef, V. Ng, S. E. Woo, L. Tay, and N. Bosch, "Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews," *Human Resource Management Journal*, vol. 34, no. 2, pp. 255–274, Apr. 2024, doi: 10.1111/1748-8583.12356.

Y. A. Alsariera, Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali, "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance," *Computational Intelligence and Neuroscience*, vol. 2022. Hindawi Limited, 2022. doi: 10.1155/2022/4151487.

J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Applied Sciences (Switzerland)*, vol. 10, no. 3. MDPI AG, Feb. 01, 2020. doi: 10.3390/app10031042.

M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

S. Nanavaty and A. Khuteta, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A Deep Learning Dive into Online Learning: Predicting Student Success with Interaction-Based Neural Networks." [Online]. Available: www.ijisae.org

Y. Li, "DESIGN OF COMPUTER INFORMATION MANAGEMENT SYSTEM BASED ON MACHINE LEARNING ALGORITHMS," *Scalable Computing*, vol. 25, no. 2, pp. 944–951, 2024, doi: 10.12694/scpe.v25i2.2615.

K. Venkatachari, "LEVERAGING MACHINE LEARNING ALGORITHMS TO GAIN INSIGHTS INTO THE MINDSETS OF IT PROFESSIONALS IN MUMBAI," 2024.

L. Yang, Q. Wang, B. Zheng, X. Li, X. Ma, and T. Wang, "ASSESSING DIGITAL TEACHING COMPETENCE: AN APPROACH FOR INTERNATIONAL CHINESE TEACHERS BASED ON DEEP LEARNING ALGORITHMS," *Scalable Computing*, vol. 25, no. 1, pp. 495–509, 2024, doi: 10.12694/scpe.v25i1.2424.

Chinenye Gbemisola Okatta, Funmilayo Aribidesi Ajayi, and Olufunke Olawale, "NAVIGATING THE FUTURE: INTEGRATING AI AND MACHINE LEARNING IN HR PRACTICES FOR A DIGITAL WORKFORCE," *Computer Science & IT Research Journal*, vol. 5, no. 4, pp. 1008–1030, Apr. 2024, doi: 10.51594/csitrj.v5i4.1085.

N. Gurung, R. Hasan, ✉ Md, S. Gazi, and F. R. Chowdhury, "AI-Based Customer Churn Prediction Model for Business Markets in the USA: Exploring the Use of AI and Machine Learning Technologies in Preventing Customer Churn," 2024, doi: 10.32996/jcsts.

Z. Ziyi, "Application of neural network algorithm based on sensor networks in performance evaluation simulation of rural teachers," *Measurement: Sensors*, vol. 32, Apr. 2024, doi: 10.1016/j.measen.2024.101049.

K. Sankara Narayanan and A. Kumaravel, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A Novel Chaotic Optimized Boost Long Short-Term Memory (COB-LSTM) Model for Students Academic Performance Prediction in Educational Sectors." [Online]. Available: www.ijisae.org

"A_Reinforcement_Learning_Based_RecommendationSystem_to_Improve_Performance_of_Students_in_Outcome_Based_Education_Model".

C. Grace and M. Garces, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A AI based Model for Achieving High Reliability Faculty Performance Using Various Machine Learning Algorithms." [Online]. Available: www.ijisae.org

T. Revandi and H. Gunawan, "JURNAL MEDIA INFORMATIKA BUDIDARMA Classification of Company Level Based on Student Competencies in Tracer Study 2022 using SVM and XGBoost Method," 2024, doi: 10.30865/mib.v8i1.7237.

A. A. Imianvan*et al.*, "Enhancing Job Recruitment Prediction through Supervised Learning and Structured Intelligent System: A Data Analytics Approach," *Journal of Advances in Mathematics and Computer Science*, vol. 39, no. 2, pp. 72–88, Feb. 2024, doi: 10.9734/jamcs/2024/v39i21869.

S. Ramos-Pulido, N. Hernández-Gress, and G. Torres-Delgado, "Exploring the Relationship between Career Satisfaction and University Learning Using Data Science Models," *Informatics*, vol. 11, no. 1, Mar. 2024, doi: 10.3390/informatics11010006.

A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discover Artificial Intelligence*, vol. 4, no. 1, Jan. 2024, doi: 10.1007/s44163-023-00079-z.

J. A. Idowu, "Debiasing Education Algorithms," *Int J ArtifIntellEduc*, 2024, doi: 10.1007/s40593-023-00389-4.

G. Ibarra-Vazquez, M. S. Ramírez-Montoya, and H. Terashima, "Gender prediction based on University students' complex thinking competency: An analysis from machine learning approaches," *Educ Inf Technol (Dordr)*, vol. 29, no. 3, pp. 2721–2739, Feb. 2024, doi: 10.1007/s10639-023-11831-4.

M. Ouahi, S. Khoulji, and M. L. Kerkeb, "Analysis of Deep Learning Development Platforms and Their Applications in Sustainable Development within the Education Sector," in *E3S Web of Conferences*, EDP Sciences, Jan. 2024. doi: 10.1051/e3sconf/202447700098.

D. Musleh *et al.*, "Machine Learning Approaches for Predicting Risk of Cardiometabolic Disease among University Students," *Big Data and Cognitive Computing*, vol. 8, no. 3, Mar. 2024, doi: 10.3390/bdcc8030031.

W. Forero-Corba and F. N. Bennasar, "Techniques and applications of Machine Learning and Artificial Intelligence in education: a systematic review," *RIED-RevistaIberoamericana de Educacion a Distancia*, vol. 27, no. 1, pp. 209–253, Jan. 2024, doi: 10.5944/ried.27.1.37491.