ISSN: 2583-9292 (ONLINE)
DOI: 10.21917/ijdsml.2025.0178

PREDICTING SCHOOL DROPOUTS WITH ENSEMBLE MODELS: A DATA- DRIVEN APPROACH TO EDUCATIONAL RETENTION

Rupali Ambalal Jadhav and Rupal Parekh
Master of Computer Application, Atmiya University, India

Abstract

Student attrition within schooling systems represents a persistent obstacle to both individual progress and broader societal improvement. This research presents a predictive model set to notify teachers of students most likely to drop out of school early, leveraging ensemble learning methods on a dense, multi-dimensional data set. The data portal consists of socio-demographic and educational aspects: residential area, first language, home occupation and level of education, number of family members, school distance, age, gender, level of education of mother, grade level, mode of transport to school, and number of siblings. Together, these measures chart the path to the dropout outcome. An ensemble algorithm suite of Random Forest, XGBoost, and Stacking Classifier leverages their capacity to capture complex, non-linear relationships, thus raising predictive accuracy. Model performance is evaluated by Accuracy, Recall, F1-Score, and ROC-AUC. Results indicate a consistent superiority of the ensemble techniques over standard algorithms producing actionable intelligence for teachers, school administrators, and policymakers. This question informs the build-out of future-oriented, evidence-based warning systems designed to reduce dropout rates and improve overall school performance.

Keywords:

Educational data mining, Dropout, Ensemble learning, Early Warning System, School Education

1. INTRODUCTION

Student dropout is a critical obstacle facing schools globally, with significant implications for the individual and for society at large. Dropouts are typically limited in their social and economic opportunities as a result of educational inadequacies, preventing them from reaching their long-term potential [1]. Simultaneously, society must endure a smaller semi-skilled workforce and elevated levels of welfare and unemployment benefit dependency [2].

One of the strongest strategies to address this concern is Early Warning Systems (EWS) for dropout students—data-based systems that identify students at risk and provide interventions in a timely manner. These systems are critical for students who do not have the benefit of access to psychological or academic assistance or awareness regarding the long-term consequences of dropping out of school [3] [4]. Governments across the globe have used such interventions to respond dropout. For instance,

Australia's Victoria Department of Education implemented the Student Mapping Tool (SMT) to identify disengagement indicators [5], whereas Wisconsin in the United States deployed the Dropout Early Warning System (DEWS) to predict and mitigate potential dropouts [6].

Concurrently, the discipline of educational data mining has become an influential driver of predictive analytics for education. By examining multi-dimensional, large-scale student data, it offers insights on patterns that lead up to dropout behavior. Ensemble learning algorithms like Stacking Classifiers, Random Forest, and XGBoost have been especially useful in this area, as they use many different models to identify intricate, non-linear relationships within educational data. Such models repeatedly out predict individual classifiers in prediction accuracy, providing schools with solid tools to improve student retention, maximize resource distribution, and guide policy-making. Therefore, incorporating ensemble learning into early warning systems can greatly improve attempts to deter student dropout and enhance overall education outcomes.

2. LITERATURE REVIEW

This section offers an integrated overview of recent research that examines student dropout prediction and retention from the perspective of predictive analytics and machine learning methods. A broad array of studies has assisted in illuminating how data-driven models can detect at-risk students and facilitate early intervention. The studies under review offer varied methods from conventional supervised learning paradigms to sophisticated AI designs, where each providing insights into student academic achievement, behavioral trends, and institutional aspects contributing to student attrition. Notwithstanding progress, some recurring gaps in research are evident throughout the literature. These are the absence of longitudinal evaluations for determining long-term effects, a lack of personalized intervention efforts, scarcity of cross-institutional validations, and operational implementation challenges such as data privacy issues and the absence of technical infrastructure. Each of the paragraphs below describes one study, summarizing its major findings and the particular research gaps it identified, thus establishing both the advances that have been made and those that still exist in developing successful dropout early warning systems.

McLean [7] highlighted the growing use of predictive analytics to enhance continuous improvement in higher education. The study emphasizes the importance of data-driven approaches in refining institutional strategies. However, it points out a significant limitation: the absence of comprehensive models that align with and support broader organizational goals.

Dart [8] focused on developing predictive models specifically for student performance in engineering mathematics. While the study showed promise in targeted academic forecasting, it also exposed a critical challenge—many institutions struggle with implementing such models due to a lack of technical expertise and analytical capacity among staff.

Bacus and Cascaro [9] conducted a thorough analysis of how analytics impacts predictive learning in higher education. Their findings revealed positive correlations between data usage and educational outcomes. Nonetheless, a key limitation identified

was the lack of longitudinal studies, which are essential for assessing the long-term effectiveness of these predictive systems.

Bujang et al. [10] employed supervised machine learning techniques to predict student grades. Their model achieved high predictive accuracy, supporting the utility of machine learning in educational forecasting. However, the study highlighted the need for cross-institutional research to verify the generalizability and robustness of these models across varied academic settings.

Borna et al. [11] utilized AI-assisted clickstream data to evaluate and forecast student learning outcomes. Their work demonstrated that behavioral interaction data can provide valuable predictive signals. Yet, the study also noted a gap in integrating diverse data sources, such as academic, socio-economic, and psychological indicators, into a single cohesive model.

Berens et al. [12] applied machine learning algorithms to administrative student data to predict dropout risk. Their results underscored the potential of administrative datasets in modeling attrition. However, they called for further validation across institutions to ensure the findings are not context-specific and can be scaled universally.

Berens et al. [13] in another study, reinforced the application of machine learning to predict student dropout using historical student data. Although effective, they acknowledged that practical implementation in real-world academic environments remains a major barrier, especially due to limited infrastructure and technical challenges.

Nakale and Amugongo [14] presented a case study from the University of Namibia, exploring student attrition through predictive modeling. The research successfully identified at-risk students, yet it flagged a critical limitation: the absence of personalized interventions tailored to individual student needs.

Tirado et al. [15] proposed an AI-based architecture for operationalizing learning analytics in higher education. The model aimed to facilitate institutional decision-making. However, the practical deployment of such architectures is hindered by challenges including data privacy concerns and limited institutional resources.

Ismaili and Besimi [16] developed a data warehousing framework for predictive analytics targeting at-risk students. The system integrated institutional datasets to generate early alerts. Nevertheless, the study emphasized the limited focus on personalized support mechanisms, which are crucial for effective intervention.

Sghir et al. [17] provided a comprehensive review of predictive learning analytics over the past decade. Their work synthesized trends, methods, and outcomes across multiple studies. Despite its depth, the review noted a recurring gap: the scarcity of longitudinal studies that track impact over extended periods.

Tarmizi et al. [18] explored student attrition using big data analytics and data mining. Their results demonstrated the strength of large-scale data in identifying dropouts. However, they pointed to practical issues, such as insufficient funding, data infrastructure, and technical expertise, which obstruct real-world implementation.

Williams et al. [19] took a holistic approach by incorporating demographic, academic, and psychological variables into dropout prediction models. This enriched framework improved prediction quality, but the study also encountered implementation barriers due to limited staff expertise in managing complex analytical systems.

Hernández-de-Menéndez et al. [20] offered a state-of-the-art overview of learning analytics, evaluating cutting-edge techniques and models. Despite highlighting significant advancements, they noted the ongoing need for more integrated and comprehensive frameworks that unify varied data types and sources.

Ortiz et al. [21] investigated student retention using academic performance, admission data, and minority status as predictors. Their study validated the importance of demographic and equity-focused variables. Yet, it found limited research dedicated to designing personalized interventions based on these predictors.

Herodotou et al. [22] examined the use of personalized learning analytics (PLAs) for motivational interventions aimed at improving retention. The study showed positive short-term outcomes. However, it highlighted the lack of longitudinal follow-up to evaluate the lasting impact of such interventions on student success.

Seidel and Kutieleh [23] implemented targeted retention strategies for first-year students using predictive analytics. Their approach was effective in reducing attrition rates, but concerns were raised regarding the ethical handling of student data and the operational complexity of integrating predictive tools into existing systems.

Shafiq et al. [24] conducted a wide-ranging literature review covering predictive analytics and educational data mining in student retention. The study mapped out methodological trends and implementation practices but concluded that more cross-institutional research is necessary to validate and extend current findings.

Most studies rely on small, publicly accessible datasets, and valuable data from ERP systems, AISHE records, or LMS platforms is underutilised because of privacy and accessibility concerns. As a result, access to actual institutional datasets is still restricted. [25]

3. DESCRIPTION OF THE DATASET

With a focus on socioeconomic, behavioural, academic, and demographic aspects, the dataset utilised in this study contains comprehensive information about specific students.

The Secondary_school_dropout dataset from Kaggle contains information on a variety of student characteristics, including gender, home language, household occupation, mother's education, household size, school distance, means of transportation, grade, and dropout status. This information can be used to identify factors that are associated with a higher risk of dropout and to develop interventions to prevent students from dropping out.

The dataset is well-suited for analysis using statistical methods. The categorical variables can be used for descriptive analysis, such as frequency counts and cross-tabulations. The numerical variables can be used for more in-depth analysis, such as linear regression and logistic regression. The results of these analyses can provide insights into the factors that influence

student dropout and can help to inform policy decisions about how to reduce dropout rates [26].

Table.1.

Feature Name	Description	Possible Values / Range	Type
location_name	Student's residential area	Rural, Urban	Categorical
home_language	Primary language spoken at home	English, Kiswahili, Other	Categorical
hh_occupation	Main occupation of the household	Unemployed, Self-employed, Other	Categorical
hh_edu	Highest educational qualification in the household	Primary, None, Other	Categorical
hh_size	Number of family members in the household	5, More than 5, Other	Categorical
school_distancekm	Distance from student's home to school	0.5–1 km, 2–3 km	Categorical
gender	Student's gender	Male, Female	Categorical
mothers_edu	Mother's education level	Primary, None	Categorical
grade	Current grade of the student	One, Two	Categorical

4. ML ALGORITHMS UTILIZED FOR PREDICTIVE CLASSIFICATION

This study utilizes advanced machine learning methods well-suited for dealing with complicated and skewed datasets to precisely predict at-risk students who will drop out. Random Forest and XGBoost ensemble methods are chosen based on their accuracy, capacity to minimize overfitting, and efficient handling of interactions between features. Random Forest constructs an ensemble of decision trees from random subsets of data and aggregates their predictions to provide enhanced generalization and feature importance for interpretability. XGBoost is a gradient boosting method that improves the performance of the model with tree pruning, regularization, and parallelization, making it very appropriate for structured educational data and stable for predicting student dropouts.

5. PERFORMANCE MATRIX

In this research, density plots are employed to visualize the distribution of important numerical attributes like student performance, attendance percentage, family income, and proximity to school for two groups of students: dropouts and persisting students. By separately plotting each group's distribution (through color-coded overlays), we can graphically evaluate the discrimination or overlap in values and assist in the identification of discriminatory features.

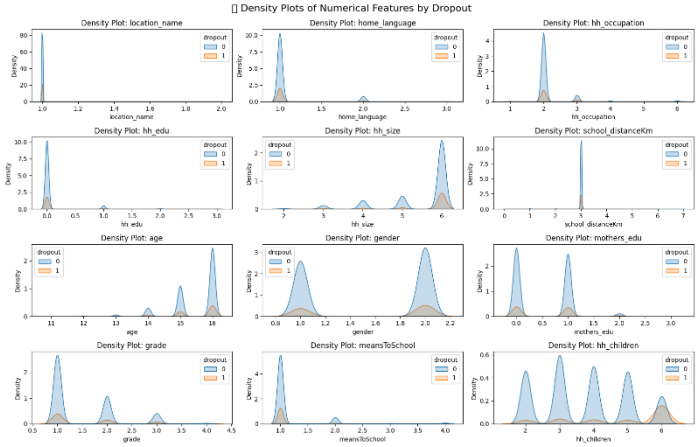

Fig.1. Density plot of Numerical Features by Dropouts

The analysis found that academic achievement (grade), educational level of parents (hh_edu, mothers_edu), family size (hh_size), household children (hh_children), and school distance (school_distanceKm) have distinct distributional separations between the two groups. These characteristics have higher dropout concentrations at lower education levels, bigger family sizes, and poorer academic grades, indicating their strong correlation with the risk of dropout. On the other hand, characteristics including gender, meansToSchool, and location_name showed lesser visualization difference and could have lesser contribution of predictive value per component. In summary, the density plots were helpful in understanding feature relevance and informed feature selection for predictive modeling.

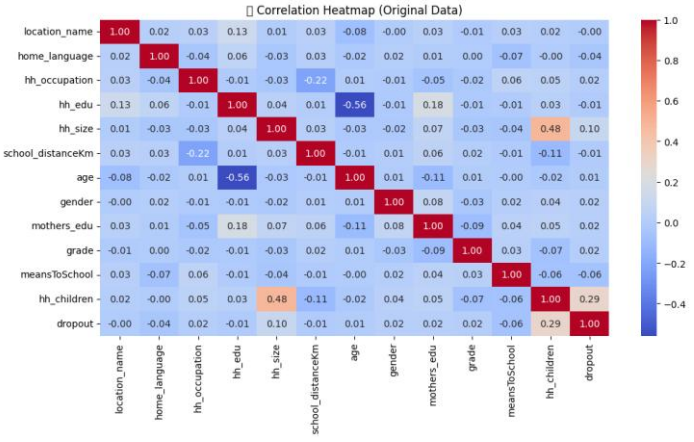

Fig.2. Correlation Heatmap

The heatmap of correlation indicates that all features except two have a very weak individual correlation with the dropout variable, meaning that no individual factor is any strong indicator of student dropout. All features are very weakly correlated with dropout except for the number of children in the household (hh_children) and household size (hh_size), which have the highest positive correlations with dropout (0.29) and (0.10),

respectively. This indicates that students from larger households might be more likely to drop out. Conversely, indicators like the mode of transportation to school (meansToSchool) and educational grade also have very weak negative correlations with school dropout, suggesting that improved access to school or better grades may diminish dropout chances slightly. Moreover, the moderate positive correlation between hh_size and hh_children (0.48) is found, whereas a negative correlation is found between household education level and age (–0.56), maybe reflecting delayed education among less-educated households' students.

🔍 Random Forest Model Evaluation Metrics:
Accuracy : 0.75
Precision: 0.89
Recall : 0.56
F1 Score : 0.69

Classification Report:

	precision	recall	f1-score	support
0	0.68	0.93	0.79	15629
1	0.89	0.56	0.69	15500
accuracy			0.75	31129
macro avg	0.78	0.74	0.74	31129
weighted avg	0.78	0.75	0.74	31129

Fig.3. Evaluation Matrix: Random forest

🔍 XGBoost Model Evaluation Metrics:
Accuracy : 0.74
Precision: 0.89
Recall : 0.54
F1 Score : 0.68

Classification Report:

	precision	recall	f1-score	support
0	0.67	0.94	0.78	15629
1	0.89	0.54	0.68	15500
accuracy			0.74	31129
macro avg	0.78	0.74	0.73	31129
weighted avg	0.78	0.74	0.73	31129

Fig.4. Evaluation Matrix: XGBoost

The Random Forest model demonstrates high accuracy for detecting dropouts but is poor in recall

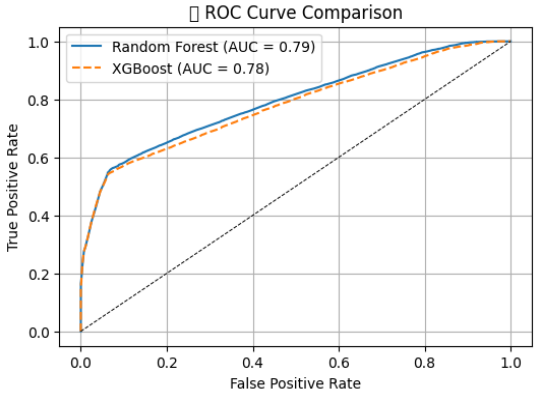

Fig.5. ROC Curve comparison

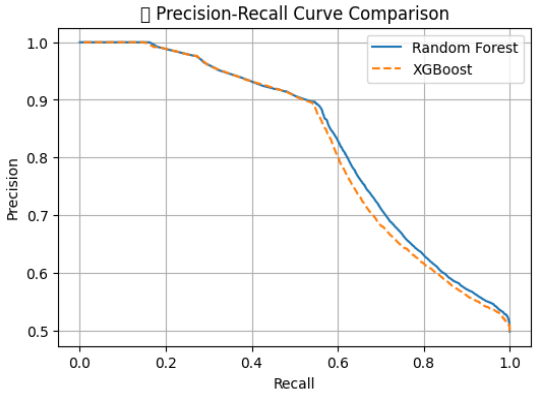

Fig.6. Precision Recall Curve comparison

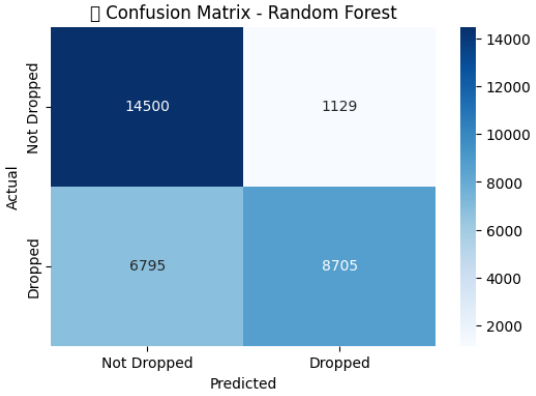

Fig.7 Confusion matrix – Random Forest

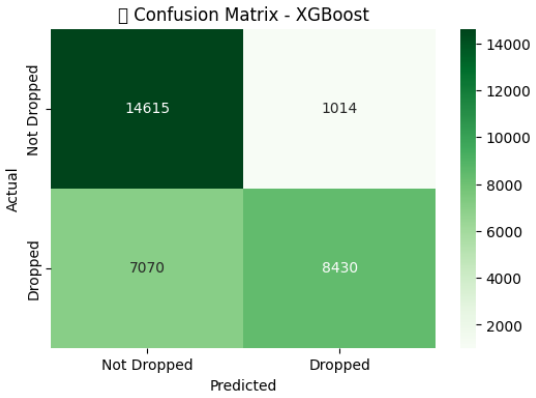

Fig.8 Confusion matrix – XGBoost

The accuracy of the Random Forest and XGBoost models were evaluated through ROC as well as Precision-Recall (PR) curve analysis to compare their classifying performance, especially in dealing with class imbalance. As the ROC Curve (Fig.5) illustrates, the Random Forest model had an Area Under the Curve (AUC) of 0.79, having a narrow margin over XGBoost's AUC of 0.78. This shows a slightly improved capability of the Random Forest classifier in separating dropout and non-dropout classes. Precision-Recall Curve (Fig.6) demonstrates that Random Forest always has higher precision at different recall levels than XGBoost, particularly in the range of mid to high recall. This implies a greater capacity of Random Forest in identifying true positives with fewer false positives. In

general, the visual analysis confirms that the Random Forest model exhibits higher predictive reliability in this educational dropout prediction contex.

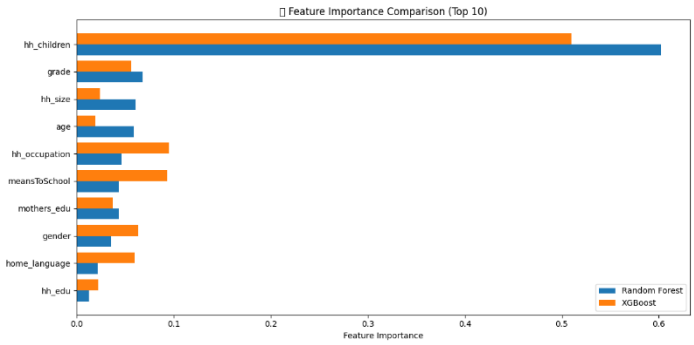

Fig.9. Feature Importance mechanism

The feature importance plot points out demographic and socio-economic attributes, especially the household number of children, as the most prominent markers for student dropout risk. The ensemble models both give prominence to variables related to household and family, pointing out that domestic pressures and financial issues need to be given special attention in dropout prevention. Random Forest gives greater importance to fewer features, while XGBoost distributes importance over more attributes, demonstrating varying learning mechanisms.

6. CONCLUSION

This study offers a holistic methodology for student dropout prediction using ensemble learning models, specifically Random Forest and XGBoost, coupled with socio-economic and academic factors in-depth analysis. Through exploratory data visualization methods like density plots and correlation heatmaps, the research discovers that household factors in particular, namely the number of children (hh_children), household size (hh_size), and parents' education levels, have significant impacts on dropout behaviors. These features exhibit clear distributional differences between the dropout and persisting student groups, which attest to their high discriminative power. Although both models achieved good performance, Random Forest had marginally higher precision and AUC. Feature importance analysis reinforced the prominence of family-based features. The results highlight the value of data-driven approaches in identifying at-risk students and guiding targeted dropout prevention efforts.

REFERENCES

[1] R.W. Rumberger, "High School Dropouts: A Review of Issues and Evidence", *Review of Educational Research*, Vol. 57, pp. 101-124, 1987.

[2] J.S. Catterall, "On the Social Costs of Dropping Out of School", *The High School Journal*, Vol. 71, pp. 19-30, 1987.

[3] R. Balfanz and D.J. Maciver, "Preventing Student Disengagement and Keeping Students on the Graduation Path in Urban Middle-Grades Schools: Early Identification and Effective Interventions", *Educational Psychologist*, Vol. 42, pp. 223-235, 2007.

[4] M. Dynarski and P. Gleason, "How Can We Help? What We Have Learned from Recent Federal Dropout Prevention Evaluations", *Journal of Education for Students Placed at Risk*, Vol. 7, pp. 43-69, 2002.

[5] S. Lamb and S. Rice, "Effective Strategies to Increase School Completion Report: Report to the Victorian Department of Education and Early Childhood Development; Communications Division", Department of Education and Early Childhood Development, 2008.

[6] J.E. Knowles, "Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin", *Journal of Education for Students Placed at Risk*, Vol. 7, pp. 18-67, 2015.

[7] Mary McLean, "Continuous Improvement in Higher Education: A Change Model Using Predictive Analytics to Achieve Organizational Goals", Available at https://fisherpub.sjf.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1307&context=education_etd, Accessed in 2017.

[8] Sarah Dart, "Developing Predictive Models of Student Success in Undergraduate Engineering Mathematics Courses", Available at https://aaee.net.au/wp-content/uploads/2020/07/AAEE2019_Annual_Conference_paper_38.pdf, Accessed in 2021.

[9] John A. Bacus and R.J. Cascaro, Rhodessa, "Impact of Predictive Learning Analytics in Higher Education: A Systematic Literature Review", *Proceedings of International Conference on Educational and Information Technology*, pp.1-6, 2024.

[10] S.D.A. Bujang, A. Selamat and O. Krejcar, "A Predictive Analytics Model for Students Grade Prediction by Supervised Machine Learning", *IOP Conference Series Materials Science and Engineering*, Vol. 1051, No. 1, pp. 1-6, 2021.

[11] Mahdi Reza Borna, Hanan Saadat and Elham Akbari, "Analyzing Click Data with AI: Implications for Student Performance Prediction and Learning Assessment", *Frontiers in Education*, Vol. 36, No. 2, pp. 1-15, 2024.

[12] Johannes Berens, Kerstin Schneider and Julian Burghoff, "Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods", *Journal of Educational Data Mining*, Vol. 11, No. 3, pp. 1-41, 2019.

[13] Samuel Nghidengwa and L.M. Amugongo, "Predicting Student Attrition: A Case Study of the University of Namibia Bachelor of Accounting (Chartered Accountancy) Programme", *Proceedings of International Conference on Emerging Trends in Networks and Computer Communications*, pp. 1-12, 2023.

[14] Alba Morales, Paul Mulholland and Miriam Fernndez, "Towards an Operational Responsible AI Framework for Learning Analytics in Higher Education", Available at: https://doi.org/10.48550/arXiv.2410.05827, Accessed in 2024.

[15] Burim Ismaili and Adrian Besimi, "A Data Warehousing Framework for Predictive Analytics in Higher Education: A Focus on Student at Risk Identification", *SEEU Review*, Vol. 19, No. 2, pp. 43-57, 2024.

[16] N. Sghir, A. Adadi and M. Lahmer, "Recent Advances in Predictive Learning Analytics: A Decade Systematic Review (2012-2022)", Education and Information Technologies, Vol. 28, pp. 8299-8333, 2022.

[17] Syaidatus Syahira Ahmad, Nurzeatul Hamimah Abdul and S.A. Rahman, "A Review on Student Attrition in Higher Education Using Big Data Analytics and Data Mining Techniques", *International Journal of Modern Education and Computer Science*, Vol. 45, No. 2, pp. 1-13, 2019.

[18] C. Williams, S. Freitas and S. Dziurawiec, "Predicting Student Attrition with "Big Data": Considering Demographic, Study-Based and Psychological Factors using Two Large Datasets", *Australian Psychologist*, Vol. 53, No. 1, pp. 68-69, 2018.

[19] M. Hernndez-De-Menndez and C. Morales-Menendez, "Learning Analytics: State of the Art", *International Journal on Interactive Design and Manufacturing*, Vol. 16, No. 3, pp. 1209-1230, 2023.

[20] A. Ortiz, J.V. Brown, M. Bains and S.L. Goffar, "Association Between Ethnic-Racial Minority Status, Admissions Data, and Academic Performance in Student Retention from a Physical Therapy Program in a Minority-Serving Institution", *Journal of Allied Health*, Vol. 51, No. 4, pp. 250-255, 2022.

[21] C. Herodotou, G. Naydenova and B. Rienties, "How Can Predictive Learning Analytics and Motivational Interventions Increase Student Retention and Enhance Administrative Support in Distance Education?", *Journal of Learning Analytics*, Vol. 72, No. 2, pp. 1-23, 2020.

[22] Ewa Seidel and Salah Kutieleh, "Using Predictive Analytics to Target and Improve First Year Student Attrition", Australian *Journal of Education*, Vol. 61, No. 2, pp. 1-21, 2017.

[23] Dalia Abdulkareem, Marjani Mohsen, Riyaz Ahamed Ariyaluran and D. Asirvatham, "Student Retention using Educational Data Mining and Predictive Analytics: A Systematic Literature Review", *IEEE Access*, Vol. 10, No. 2, pp. 1-29, 2022.

[24] Adnan Anwar Qureshi, Tufail Ahmed Qureshi, Saad Nasir Khan and Z.L.H. Malik, "A Review on Machine Learning Approaches for Predicting Student Dropout Rates", *Proceedings of International Conference on Educational and Technology*, pp.1-6, 2024.

[25] Kaggle Dataset, "Secondary School Student Dropout", Available at https://www.kaggle.com/datasets/edgargulay/secondary-school-student-dropout, Accessed in 2025.