DOI: https://doi.org/10.57233/ihcfs.052

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025 WEBSITE: fos.fugusau.edu.ng

1ST INTERNATIONAL HYBRID CONFERENCE OF FACULTY OF SCIENCE

(A PUBLICATION OF FACULTY OF SCIENCE, FEDERAL UNIVERSITY GUSAU, NIGERIA)

MACHINE LEARNING APPLICATIONS IN FORECASTING ACADEMIC PATHWAYS BASED ON STUDENT DEMOGRAPHIC PROFILES IN TERTIARY EDUCATION: A CASE STUDY OF NORTEASTERN NIGERIA

Ibrahim Hassan[1], Muhammad Al-amin Alkali[2], Ahmed Haruna Dokoro[3], Mustapha Umar[4] and Aliyu Ahmad[5]

[1,2,3,4] Department of Computer Science Technology, Gombe State Polytechnic Bajoga,

[5]Directorate of ICT, Federal University Gusau

Corresponding author: hassangsu@gmail.com

ABSTRACT

This study explores the role of student demographics in predicting course selection in tertiary institutions using machine learning models in addressing the issue of misaligned academic choices. It aims to guide the decision-making processes for students, educators, and policymakers by identifying optimal pathways for individual academic and career success by analyzing factors such as academic grades, gender, and family income. A dataset of 1,200 students from tertiary institution across the six states from northeastern Nigeria, were used to evaluate five machine learning models: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Neural Networks. Experiment was carried out in WEKA, using stratified 10-fold cross-validation, Random Forest and Neural Networks outperformed the others, with accuracies of 88.8% and 86.4%, respectively. The results show that academic performance, gender, and socioeconomic background significantly influence course selection, offering valuable insights for educators and policymakers to improve student guidance and reduce course mismatch.

Keywords: Artificial Intelligence, Machine Learning, WEKA, Demographics, Predictive models

1.0- INTRODUCTION

Student academic performance is a key indicator of a country's educational progress. Various factors, such as gender, age, the quality of teaching staff, and the students' learning processes, significantly influence academic outcomes. Predicting student success has become an area of growing interest in the field of education (Malik, et al., 2025). In essence, student performance reflects how well learners meet both short-term and long-term educational goals (Alsariera, et al., 2022).

Student dropout is widely recognized as one of the most challenging and impactful issues within the educational system (Kim & Kim, 2018). This issue leads to economic, social, academic, political, and financial harm affecting all key stakeholders in education, from students to government bodies and funding organizations. Part of the reasons that constitute this challenge include wrong selection of courses in higher schools (Niyogisubizo et al., 2022). To address this challenge, there is a pressing need for effective and efficient strategies to reduce dropout rates, this can be achieved by exploring demographic factors and educational choices, enabling a more personalized approach to student placement and guidance in higher education (Alboaneen, et al., 2022). Ultimately, the research serves as a tool for decision-making to students, educators, and policymakers by identifying the most suitable academic and career paths for individual success (Alqahtani, et al., 2023).

DOI: https://doi.org/10.57233/ihcfs.052

ISBN: 978-978-68-07775

The selection of courses in tertiary institutions is a crucial decision for students, as it often determines their future career trajectories (Kassaw, et al., 2024). In Nigeria, students must decide between attending a polytechnic, college of education, or university, each offering distinct educational pathways and career prospects. The selection of courses in tertiary institutions is a pivotal decision in a student's academic and professional journey. Demographics such as age, gender, socioeconomic background, and prior academic performance have been observed to influence these decisions. With the growing diversity of student populations, predicting course selection based on demographic factors has become increasingly complex. Machine learning (ML) models offer a powerful tool to analyze and predict trends by utilizing vast amounts of data, identifying patterns that are often imperceptible to humans. Machine learning models can effectively predict how student demographics impact course selection (Malik, 2025).

Machine Learning, a branch of Artificial Intelligence (AI) assists in creating personalized learning plans that take into account each student's unique learning style, interests, and goals, which enhances motivation and engagement, ultimately resulting in improved academic performance (Liu et al., 2023). Advancements in artificial intelligence (AI) and machine learning (ML) enable us to move beyond traditional computing to mimic and even surpass human intelligence. These technologies have already made a significant impact on education, equipping students with new skills and fostering a collaborative learning environment in higher education institutions (HEIs), with promising implications for the future (Kuleto, et al., 2021). Moreover, AI can stay current with the latest and most effective teaching strategies and methodologies due to ongoing advancements in educational research (Basu et al., 2019). This can lead to a situation where students are not truly understanding or mastering the material but are merely finding ways to exploit the system. Therefore, while those technologies have considerable potential to support teaching and learning, their use must be carefully managed and supplemented with strategies that encourage genuine understanding and critical thinking. It's also crucial to cultivate an educational environment that values academic integrity and discourages misuse of such technologies (Alqahtani, et al., 2023). These techniques are highlighted in the literature for their ability to predict outcomes using both quantitative and qualitative/categorical data. Neural Networks (NN), a machine learning method, develop computational models by iteratively adjusting based on training data to minimize errors, a process known as supervised learning. NN is widely applied for classification tasks due to its ability to predict categorical outcomes, such as whether a student completes a degree or not. Similarly, Support Vector Machines (SVM) are supervised learning models used for classification and prediction (Tomasevic et al., 2020), particularly useful for identifying data groupings, like determining if a student will remain at the university (Alsariera, et al., 2022). Decision Trees (DT) use observations to infer a target value through classification trees, with regression trees applied when the target is a continuous variable. Logistic Regression (LR) is employed to create statistical models for binary dependent variables, these algorithms are frequently used in retention

studies to analyze data and predict categorical dependent variables. (Cardona et al., 2023).

Outcome-based education is an educational framework centered on achieving specific learning outcomes. These outcomes serve as benchmarks to assess the degree to which students acquire the desired competencies, including knowledge, skills, and values, by the conclusion of a particular learning experience, student outcomes provide a more comprehensive measure of academic success than traditional assessment grades alone. (Namoun & Alshanqiti, 2021).

Therefore, the paper has used the techniques of machine learning to predict the placement of students of various tertiary institutions in the Northeastern Nigeria and the proposed course of study.

The relationship between demographic factors and academic choices has been widely studied in educational research. Age and gender are recognized as key influencers of academic preferences, with certain trends showing gendered inclinations toward specific fields (OECD, 2022).

The study of Basu et al. (2019), applies machine learning techniques to predict whether students offered admission to arts college will accept or reject the offer. Using four years of data on 11,001 students, the researchers treated the problem as a binary classification task and tested various algorithms, finding that logistic regression performed best with an AUC score of 79.6%. This research highlights how institutions can use machine learning to improve predictions of incoming student numbers, helping optimize resource allocation and manage tuition-based revenue. The model is broadly applicable and can be easily replicated by other academic institutions using available admissions data.

Namoun & Alshanqiti (2021) reviewed the prediction of student academic performance, shifting focus from traditional metrics like grades and GPAs to learning outcomes. They analyzed 62 studies (from 2010 to 2020) to identify intelligent techniques for predicting student success. The review covers methods for predicting outcomes, models used, and influencing factors. Regression and supervised machine learning models are commonly used, with key predictors like online learning activities, assessments, and academic emotions. The paper highlights challenges in using these techniques and advocates for a holistic approach to understanding student achievement, emphasizing competencies and skills alongside academic results.

Niyogisubizo et al. (2022) developed a stacking ensemble model using Random Forest, XGBoost, Gradient Boosting, and Feed-forward Neural Networks to predict university dropouts. The model, based on data from Constantine the Philosopher University (2016-2020), outperformed individual classifiers in terms of accuracy and AUC, which identified at-risk students early for timely interventions. The study highlights the potential of machine learning in improving educational outcomes, though challenges like class imbalance and model interpretability persist. The ensemble approach offers a solution by enhancing prediction performance and aiding in dropout prevention.

Alsariera et al. (2022) developed a model that predicts student academic performance for improving educational outcomes and future employment opportunities. Factors such as gender, age, and socio-economic

background influence performance, which also impacts the ranking of educational institutions. Machine learning (ML) techniques, including artificial neural networks (ANN), support vector machines (SVM), and decision trees (DT), are used to analyze large datasets and predict student success. However, challenges remain in handling complex data. By identifying patterns and predicting performance, ML enables early interventions to support students, though further research is needed to enhance predictive accuracy.

Alqahtani et al. (2023) explores the transformative impact of Artificial Intelligence (AI), specifically through Natural Language Processing (NLP) and Large Language Models (LLMs) like GPT-4 and BARD, on education and research. The paper highlights AI's capabilities in enhancing educational support, curriculum design, personalized learning, and key research tasks such as text generation, data analysis, and peer review. It also discusses AI's role in providing tailored feedback, grading, and career guidance. The authors address challenges like ethical concerns, algorithmic bias, and data privacy, emphasizing responsible AI use to balance technological advancements with human engagement for improved educational and research outcomes.

1.1- Aim and Objectives of the Study

The aim of this research is to develop and evaluate machine learning models that can predict a Secondary School student's future course of study, in Science, Humanities, or Management at tertiary institution, based on data from the six northeastern states. The objectives are to:

i- Gather and analyze student data, focusing on demographics, academic performance, socio-economic background, and personal interests.

ii- Determine the factors that influence a student's choice of academic field (Science, Humanities, or Management) in contrast to the educational pathway (College of Education, Polytechnic, or University).

iii- Assess the predictive accuracies of the developed models across different student demographics to ensure a fair and robust application. iv- Provide recommendations to educational stakeholders through the findings which can be used to support students in making academic choices.

This paper is structured as follows: It begins with an introduction and literature review on the application of machine learning in educational settings. The Materials and Methods section is divided into two parts: data preprocessing and exploration, followed by implementation steps, success metrics, and feature selection. The results of the predictions are presented using various metrics, and the paper concludes with a summary of findings, alternative techniques, and future research directions.

2- METHODOLOGY

The method employed in conducting this research, lies on machine learning (ML) life cycle in relation to the educational studies. This is illustrated through an expanded ML life cycle diagram (Figure 1), which we use as a conceptual framework to highlight the key insights from our study. Therefore, the materials and methods employed in the study are elaborated under the following headings:

2.1- Research Design

The study employs a quantitative research design using machine learning models to

DOI: https://doi.org/10.57233/ihcfs.052

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025 WEBSITE: fos.fugusau.edu.ng

predict course selection based on student demographics. Data was collected from students of tertiary institutions from the six northeastern states of Nigeria. The analysis focuses on determining whether students are more likely to pursue higher education at a polytechnic, college of education, or university, and which career pathways they are inclined to follow—such as Sciences, Humanities, or Management. The dataset was prepared and stored in CSV and saved in RFF file. The data was loaded into the WEKA explorer and the algorithms were chosen where the value of the cross validation was set. After running the experiment, the performance of each of the classifier was obtained from the confusion matrix

2.2- Sample Population

The targeted population consists of students of tertiary institutions in northeastern Nigeria. A sample of 1,200 students were selected through a stratified random sampling technique from the six states of the northeastern Nigeria. To ensure representation across various demographic categories, such as age, gender, and socioeconomic status. Demographic dataset of the targeted population was obtained through surveys, academic records, and interviews with the students. 200 respondents each from the states.

2.3- Data Attributes and Feature Selection

Datasets consist of irrelevant, redundant, and noisy data that can harm a model's performance and hinder the discovery of valuable patterns. Moreover, using all available features in a model is computationally costly and results in longer training times. With this, cleaning the data and removing of noise has become imperative. This process is considered the preprocessing stage.

Table 1, highlighted the most important features for predicting course selection in tertiary institutions, along with their descriptions and types:

Feature	Description	Type
Academic Grades (AG)	A student's grade: Excellent ‘1’, Good ‘2’, Average ‘3’, Poor ‘4’	Categorical (ordinal)
Family Income (FI)	High ‘1’, Middle ‘2’, Low ‘3’.	Categorical (ordinal)
Parental Education (PE)	Elementary ‘1’, BSc. ‘2’, MSc. ‘3’, PhD. ‘4’	Categorical (ordinal)
Gender	Male ‘1’, Female ‘2’	Categorical (binary/nominal)
Age	The student's age: 12 -18	Continuous (numerical)
Prior Coursework	Interested in: English ‘1’, Math ‘2’, Science ‘3’ General Studies ‘4’, All ‘5’	Categorical (nominal)

DOI: https://doi.org/10.57233/ihcfs.052

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025 WEBSITE: fos.fugusau.edu.ng

Geographic Location	Urban ‘1’, rural ‘2’.	Binary (Ordinal)
Parental/Guardian Influence	Highly Influential ‘1’, Influential ‘2’, Not Influential ‘3’	Categorical (nominal/ordinal)
Access to social media and Online Activity	Accessible ‘1’, non-accessible ‘2’	Binary (ordinal)
Advising and Counseling Services	Beneficiary ‘1’, non-beneficiary ‘2’	Binary (ordinal)

2.4- Developing the Models

The study focused on five ML models: Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR) Neural Network (NN) and Support Vector Machine (SVM). Designing the models has involved utilizing the widely used Java-based machine learning tool, Waikato Environment for Knowledge Analysis (WEKA). The data collected, were pre-processed, and cleaned for analysis. A stratified 10-fold cross-validation approach was used to evaluate the classifiers, the process splits the dataset into 10 equal parts. In each iteration, 90% of the data was used for training and 10% for testing, and after 10 iterations, the average performance metrics were calculated.

2.5- Models Performance Evaluation

A confusion matrix was employed in evaluating the performance of the models. It is a table used in machine learning and statistics to evaluate the performance of a classification algorithm.

Table 2. Confusion matrix.

	Predicted Accept	Predicted Reject
Accepted	True Positive (TP)	False Negative (FN)
Rejected	False Positive (FP)	False Positive (FP)

Accuracy is usually the first measure used to determine whether a machine learning algorithm is making enough correct predictions. Using accuracy alone as a metric to evaluate the performance of a classifier can often be misleading, therefore, other metrices which include: F1-score and Area Under the Curve (AUC) are used.

accuracy = *TP + TN / TP + FP + FN + true negative* … (1)

Precision (Positive Predictive Value): $P = \frac{TP}{TP+FP}$ … (2)

Recall (Sensitivity or True Positive Rate): $R = \frac{TP}{TP+FN}$ … (3)

DOI: https://doi.org/10.57233/ihcfs.052

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025 WEBSITE: fos.fugusau.edu.ng

2.6- Model Architecture

The development of the models is structured around the sequential stages of machine learning.

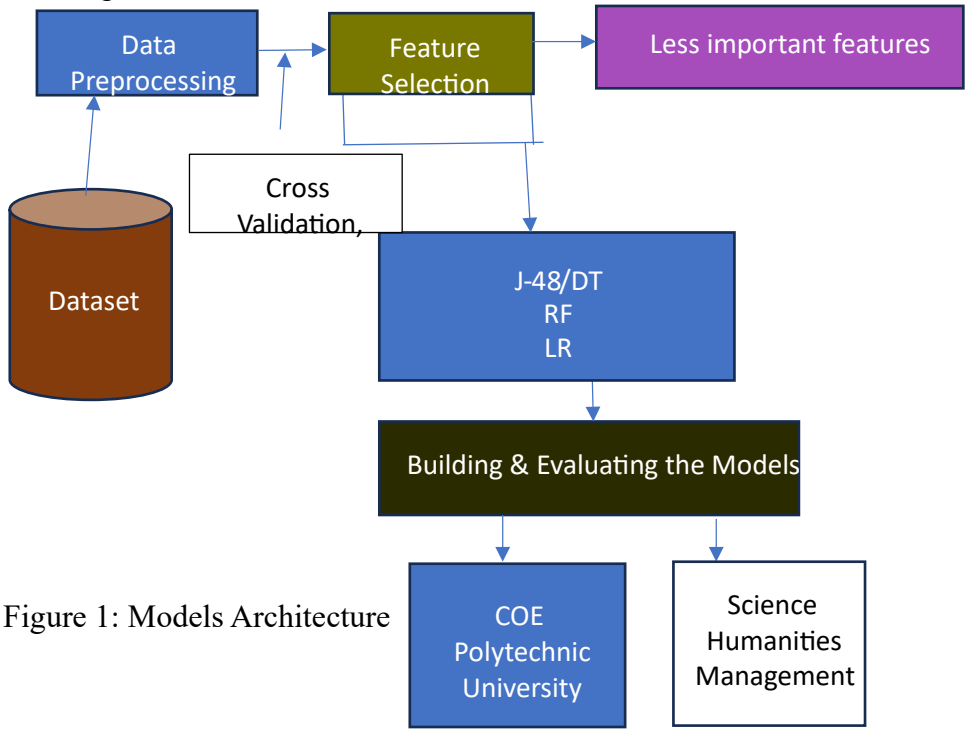

Figure 1: Models Architecture

3.0- RESULTS AND DISCUSSION

The results presented the performance of the five machine learning models in predicting course selection. The primary goal was to identify the most accurate model for predicting students' educational pathways based on their demographics.

Table 3: Performance of Classifiers

Classifier	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC(%)	Error (%)	Execution Time(ms)
DT	80.20	76.00	79.00	69.0	19.80	90.00
SVM	82.40	78.60	79.00	69.0	17.60	30.00
LR	80.00	70.60	80.40	72.0	20.00	40.00
NN	86.40	80.40	82.20	76.4	13.60	48.00

DOI: https://doi.org/10.57233/ihcfs.052

ISBN: 978-978-68-07775

BOOK OF PROCEEDINGS

IHCFS – 2025 WEBSITE: fos.fugusau.edu.ng

RF	88.80	80.80	82.20	74.8	11.20	30.00

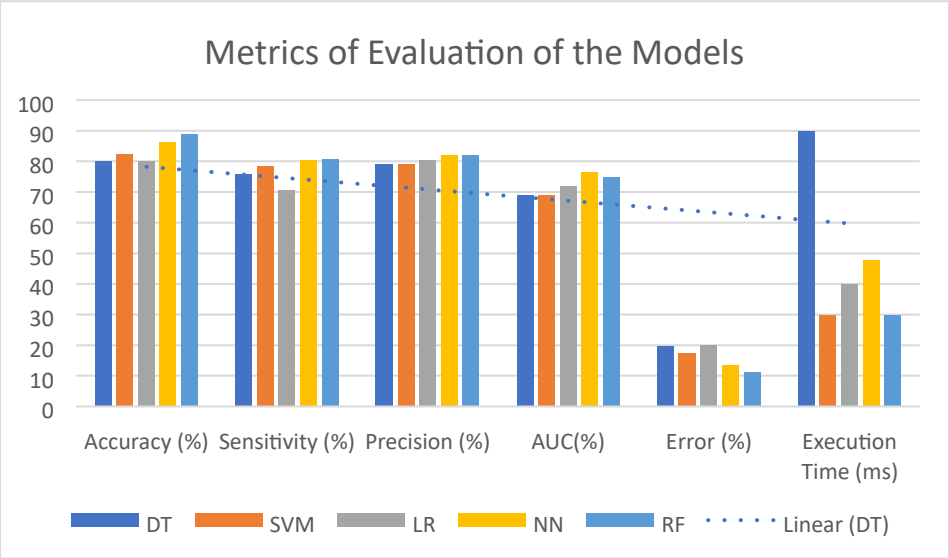

Figure 2: Evaluation of the Metrics of the Models

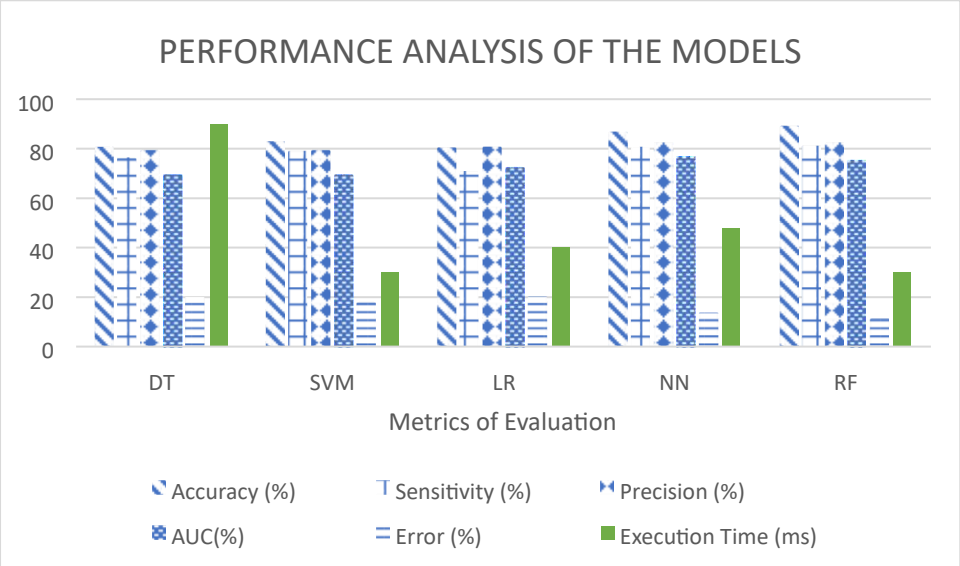

Figure 3: Performance Analysis of the Models

Preliminary results indicated that **Random Forest** and **Neural Networks** outperform other models in terms of accuracy, with Random Forest excelling in interpretability due to its ability to rank feature importance which align with (Alhassan et al., 2020). The most influential features across all models were **academic grades**, **gender**, and **family income**.

Academic grade was the strongest predictor of course selection, particularly for students aspiring to attend universities, which generally have higher entry requirements. **Family income** also played a critical role, with students from parents with higher income backgrounds more likely to choose prestigious institutions such as universities, while those from family with lower income backgrounds leaned toward polytechnics or

colleges of education. **Gender** differences were also observed, with males more likely to choose technical and vocational courses, while females favored courses in education and the social sciences, this is also similar to the study of (Cardona et al., 2023).

These findings align with existing literature, confirming that demographics significantly influence educational pathways like (Jokhan et al., 2022). The use of machine learning models offers a powerful tool for predicting these outcomes, providing educators and policymakers with valuable insights into student preferences and potential career trajectories.

4.0- CONCLUSION

This study successfully demonstrates the impact of student demographics on course selection in tertiary institutions using machine learning models. The findings suggest that key demographic factors, such as academic grades, gender, and family income, significantly influence students' educational choices and career trajectories which align with (Gong, et al., 2019). Random Forest and Neural Networks emerged as the most effective models for prediction, the results obtained is similar to (Alhassan et al., 2020). These insights can be leveraged by educators and policymakers to develop more personalized and equitable strategies for student placement, ultimately enhancing educational outcomes and reducing the rate of mismatched course selections thereby reducing the rate of school drop outs. Further research should explore the scalability of these models across different regions and institutions.

5.0- RECOMMENDATION

To further enhance the predictive accuracy and applicability of the study, it is recommended to expand the dataset by incorporating additional demographic and socio-economic variables, such as parental occupation and access to extracurricular activities. Additionally, exploring the integration of deep learning techniques alongside traditional machine learning algorithms could provide a more nuanced understanding of how student demographics influence course selection. Collaborating with educational institutions to implement predictive models in real-world student advising systems could also be explored, enabling a more tailored and personalized academic guidance process.

6.0- ACKNOWLEDGEMENT

The authors sincerely acknowledge the financial support provided by the Tertiary Education Trust Fund (TETFund), which significantly facilitated the successful conduct of this research. The support was particularly instrumental in data collection across the six states in Northeastern Nigeria, thereby enhancing the quality and reliability of the dataset used for this study.

REFERENCES

Alboaneen, D., Almelihi, M., Alsubaie, R., Alghamdi, R., Alshehri, L., & Alharthi, R. (2022). Development of a Web-Based Prediction System for Students'Academic Performance. *MDPI*, 1-19.

Alhassan, A., Zafar, B., & Mueen, A. (2020). Predict Students' Academic Performance based on their Assessment Grades and Online Activity Data. *(IJACSA) International Journal of Advanced*

Computer Science and Applications, 11(4), 185- 195.

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., Saleh, K. b., . . . Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research . *Research in Social and Administrative Pharmacy*, 1236-1242.

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtan, A. A., & Ali, N. (2022). Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance. *Hindawi Computational Intelligence and Neuroscience*, 1-11.

Azam Malik, Deeba Abrar, Noor Alam Khan, & Abdulmalik M. Alhussami. (2025). Sustainable development goals: students' perception on quality education, gender equality, and sustainability. *Discover Sustainability, 6*, 1-18. doi:doi.org/10.1007/s43621-02501288-5

Basu, K., Basu, T., Buckmire, R., & Lal, N. (2019). Predictive Models of Student College Commitment Decisions Using Machine Learning. *MDPI*, 1-18.

Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J. (2023). Data Mining and Machine Learning Retention Models in Higher Education. *Journal of College Student Retention: Research, Theory & Practice, 25*(1), 51-75.

Gong, B., Nugent, J. P., Guest, W., Parker, W., Chang, P. J., Khosa, F., & Nicolaou, S. (2019).

Influence of Artificial Intelligence on Canadian Medical Students' Preference for Radiology Specialty: A National Survey Study. *Academic Radiology, 26*(4), 566-572.

Jokhan, A., Chand, A. A., Singh, V., & Mamun, K. A. (2022). Increased Digital Resource Consumption in Higher Educational Institutions and the Artificial Intelligence Role in Informing Decisions Related to Student Performance. *sustainability*, 1-17.

Kassaw, C., Demareva, V., Negash, M., & Sime, Y. (2024). Psychosocial determinants of academic achievement in Ethiopian higher education students, 2024. Systematic review and meta-analysis. *Heliyon, 10*, 1-13.

Kuleto, V., Ili'c, M., Dumangiu, M., Rankovi´c, M., Martins, O. M., Păun, D., & Mihoreanu, L. (2021). Exploring Opportunities and Challenges of Artificial Intelligence and Machine Learning in Higher Education Institutions. *sustainability*, 1-16.

Liu, L. T., Wang, S., Britton, T., & Abebe, R. (2023). Reimagining the machine learning life cycle to improve educational outcomes of students. *SOCIAL SCIENCES COMPUTER SCIENCES, 120*(9), 1-12.

Mpiti, V. S., Ncanywa, T., & Asaleye, A. J. (2025). Do Educators' Demographic Characteristics Drive Learner Academic Performance? Examining the Role of Gender, Qualifications, and Experience. *Education Scinces, 15*, 1-23. doi:doi.org/10.3390/educsci15040487

Mpiti, V. S., Ncanywa, T., & Asaleye, A. J. (2025). Do Educators' Demographic Characteristics Drive Learner Academic Performance? Examining the Role of Gender, Qualifications, and Experience. *Education Sciences, 15*, 1-23.

Namoun, A., & Alshanqiti, A. (2021). Predicting Student Performance Using Data Mining and Learning Analytics

DOI: https://doi.org/10.57233/ihcfs.052

ISBN: 978-978-68-07775

Techniques: A Systematic Literature Review. *MDPI: Applied Sciences*, 1-28.

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence* , 1-12.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computer and Education*, 1-18.