Predicting Undergraduate Students' Demographics and Entry Status from Academic Performance Metrics: A Data Mining Approach

Anzuruni Maulidi Katunka[1]

Abstract

This study aims to investigate if students' demographics (gender and age) and student's entry status (diploma or form six) have an impact on students' academic performance in undergraduate studies. The study used the reverse approach of data mining techniques by predicting gender, age, and entry status from students' academic performance metrics using three machine learning algorithms (Decision Trees, Random Forests, and Naïve Bayes) whereby accuracy, precision, recall, and F1 score was used to evaluate the models. Furthermore, the study used datasets containing 348 academic records of first-year bachelor's degree students studying Records, Archives, and Information Management at Tanzania Public Service College, Dar es Salaam campus. On one hand, the results showed that there is no relationship between students' demographics (age and gender) and academic performance in undergraduate studies. On the other hand, the study revealed that there is a relationship between the entry status of students and their academic performance. Following these findings, the study recommends that educators, assessors, and moderators take action by aligning the curricula and teaching methods which reduce or eliminate the imbalance by favoring both diploma and form six students to have homogeneous performances.

Keywords: *Predictive modeling, Student demographics, Academic Performance, Gender Prediction, Educational Data mining, Performance metrics, Machine learning.*

1.0 Introduction

Education Data Mining (EDM) is the field that applies data mining techniques, statistical methods, and machine learning algorithms to discover hidden insights from data (Hussain et al., 2024). In the education system, students are characterized by different attributes and behaviors which may impact their academic performance. In the modern education system, exploring the impacting factors on students' academic performance is very important to evolving education strategies and applying the right and quick interventions (Orrego *et al.,* 2022).

In the education system, students are normally assessed and acquire different academic performance metrics such as assignments, tests, presentations, and semester examination scores. Through these scores, students acquire grades and GPAs. Based on examination

1 Assistant Lecturer, Tanzania Public Service College – Dar es Salaam Campus
anzuruni.katunka@yahoo.com

regulations a certain number of modules passed enables students to advance to the higher semester. Student demographics, such as sex, age, parents' job, parents' education, family size, and support have an academic impact on students' performance (Unal, 2020).

Various studies focused on the prediction of students' performance across various fields of education, particularly in science, technology, engineering, and mathematics, to build learning models that can forecast students' academic success or identify students who are at risk for quick and timely interventions. However other fields like Records, Archives, and

Information Management have been insufficiently studied. Also, the majority of these studies placed more emphasis on performance prediction using students' demographic data but still, there is a contradiction between the findings of existing studies on whether gender and age of students influence academic performance significantly as shown in the empirical review of this study, in section 2.2.

In recent years there has been an ongoing debate on whether students completing ordinary-level education are better off pursuing tertiary education at the certificate level than to diploma or to continue with advanced-level studies (Güre, 2023). These two different educational pathways eventually meet at the undergraduate level in higher education, making the entry status a crucial factor for academic performance prediction to reveal if there is homogeneous performance between the two groups.

This study aims to predict the combination of students' demographics data (gender and age) and entry status (diploma or form six) from performance academic metrics for undergraduate studies in the field of Records, Archives, and Information Management by using data mining techniques.

2.0 Literature Review

2.1 Theoretical Review

Different theories have been describing the impact of gender, age, and educational background (Entry status) on students' academic performance. The following is a theoretical review of the independent variables used in this study about academic performance.

2.2 Gender and Academic Performance

Social Role Theory: In society, there is the existence of different social gender roles that influence the behaviour of male and female individuals with the inclusion of their academic performance (Eagly & Wood, 2012). For example, females may be optimistic about doing extremely well in social sciences, while males may be motivated toward Science, Technology, Engineering, and Technology.

Gender Stereotyping Theory: Gender stereotypes on abilities such that males are better at mathematics as compared to females and females are better at communication skills may affect students' confidence and self-concepts. This may influence students' academic performance (Eccles *et al.,* 1990).

2.3 Age and Academic Performance

Theory of Cognitive Development: In the learning process, student progress through different stages of development in relation to their age. These stages of development influence their cognitive abilities and hence impact their academic performance (Barrouillet, 2015). For example, students of different ages may possess varying reasoning and problem-solving skills.

Readiness Theory: Students' development milestones may affect their readiness levels, hence affecting the effectiveness of engagement in curricular activities. Varying readiness levels due to age gaps may lead to academic performance variations (Pruitt, 2014).

2.4 Entry Status and Academic Performance

Cultural Capital Theory: Students from different educational backgrounds possess different levels of cultural capital such as communication skills, learning ability, and perceptions which may affect their academic performance (Bourdieu, 2011).

Constructivist Learning Theory: The quality of background education influences the Zone of Proximal Development (ZPD) which affects how students utilize previous knowledge to acquire new knowledge and skills (Nyikos & Hashimoto, 1997). In undergraduate studies, students who have different educational backgrounds converge. Some students acquire their entrance qualification through Form Six, while other acquires entry qualifications through diploma.

2.5 Empirical Review

In the field of education, different studies have been conducted to build predictive models to explore different attributes of students and academic performance. Durak and Bulut (2024) studied the relationship between students' performance by building a model to predict whether the performance outcome of students in the programming course is either high or low in the relationship among various features including programming level, thinking perspective in computation, programming empowerment, and students' success level. Random forest, K-nearest neighbors, decision trees, Naïve Bayes Classifiers, support vector machines, and logistic regression were used. The findings revealed that the computational thinking perspective, students' programming performance level, programming empowerment, and computational identity are impacted by student gender but are not impacted by student success level.

The study by Almutairi *et al.,* (2019) used an online data repository to perform students' academic performance predictions based on the behavioral, academic background, and demographic features using logistic regression, random forest, MLP, XGBoost, and ensemble learning. The study used behavioral features such as students' learning habits, educational background, and demographic features including age. The results of this study revealed that gender had an impact on students' performance where females outperform males.

The study by Mkwazu and Yan (2020) used datasets from the Sokoine University of Agriculture to build a predictive model by using a decision tree classifier to assist in predicting students' grades based on students' features such as Previous GPA, Programme registered, semester and academic year. The findings of their study showed that new students could be effectively classified based on academic features to help them during

the selection of the courses to avoid later drop-offs. However, the findings did not specify features that impacted the student's performance.

The research by Unal (2020) explored the prediction of students' performance in Mathematics and the Portuguese language by using different student attributes including sex and age to predict final grades. Three machine learning algorithms were used: Naïve Bayes, Decision tree, and Random forest with five levels grading system and binary level of grading system. The prediction model achieved high accuracy, especially with a binary level of grading in both subjects. The findings showed that both age and gender impacted the academic performance of students.

Age and gender influence learners' academic performance especially when the two features are concurrently considered (Kimeli *et al.,* 2019). However, between the two features, age seemed to have a more significant influence, resulting in relatively less performance for older trainees. The authors in this study empirically analyzed how gender and age can impact the examination results in adult education.

Kapinga and Amani (2016) examined the factors which influence academic performance in undergraduate studies at Mkwawa University College of Education in Tanzania, by using students' academic results secondary data. Features such as grade point average, entry grade points, communication skills grades, age, and gender. The study revealed a positive relationship between academic performance in final examinations and students' entry points. That is, students who entered undergraduate studies with high grades seemed to perform well as compared to those with low grades. However, gender had no impact on students' academic performance.

In a different approach from previous studies, this study used a reverse data mining approach by building a predictive model to predict students' age, gender, and entry status from students' academic performance metrics. The study intends to answer the question of whether students joining bachelor's degree programmes after completion of an advanced certificate of secondary education (form six) and students who join after the completion of an ordinary diploma perform differently or not in their undergraduate studies. It also intends to answer the question of whether students of different ages and genders have varying academic performance in undergraduate studies, specifically in the field of Records, Archives, and Information Management which has not been explored.

3.0 Methodology

This study utilized secondary data collected from the Academic and Registration Information System (ARIS) for first-year students studying the Bachelor's degree in Records, Archives, and Information Management. The dataset contained structured data with 11 features and 348 records (students). To prepare the data for analysis, three data pre-processing techniques were applied: data cleaning, data discretization, and data aggregation.

Data cleaning was used to standardize students' dates of birth to a uniform format, which was necessary for calculating their age, a key class variable in the study. Data discretization converted students' ages from numeric to categorical data by grouping them into four distinct age categories: 15–25, 26–36, 37–47, and 48–58. This transformation made the age data suitable for classification tasks. Data aggregation involved calculating the average

coursework scores across six subjects for each student, creating a new variable called the average course assessment score (AVCA). Similarly, the average semester examination score (AVSE) was computed for each student from their semester examination scores.

To predict the class variables, which were all categorical, decision trees, Naïve Bayes, and random forest classifiers were used and compared. The models were tested in three different modes: using the full training set, 10-fold cross-validation, and a train-test split. The evaluation of the models was based on four key metrics: accuracy, precision, recall, and F1-score. These metrics were chosen because the study focused on a classification task. The models were constructed and tested using the Waikato Environment for Knowledge Analysis (WEKA) software, while MS Excel and Python libraries such as Seaborn, Matplotlib, and Pandas were also used for data analysis and visualization.

4.0 Results

This section presents and discusses the evaluation results of the three models during both the training and prediction phases. The values of performance metrics are also presented. These metrics can be manually calculated using the formula shown in equations (2) – (5). Since WEKA was used, performance metrics were automatically calculated.

4.1 Results of Training Phase

The three predictive models were trained by using a decision tree (J48), Naïve Bayes, and Random forest in three different training modes (Full training data, 10-fold Cross-Validation, 66% - Split) for each class attribute. To train the models, 176 students' records from the first semester of their bachelor's degree in Records Archives and Information Management were used.

4.2 Results of Model Training using Decision tree (J48) algorithm

Table 1 below summarizes the outcomes of training the model with the decision tree (J48) algorithm by showing accuracy, precision, Recall, and F1 Score.

Table 1: Model Training results using J48

Training Mode and Class		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Full Training Data	Gender Class	66.5	64.8	66.5	60.6
	Entry Status Class	**75.0**	56.3	75.0	64.3
	Age Class	59.1	67.4	59.1	50.9
10-fold Cross-Validation	Gender Class	61.4	50.1	61.4	49.8
	Entry Status Class	**75.0**	56.3	75.0	64.3
	Age Class	43.75	36.7	37.3	46.5
Split 66% train, remainder test	Gender Class	58.3	55.4	58.3	50.0
	Entry Status Class	78.3	78.3	78.3	68.8
	Age Class	38.3	31.8	38.3	33.5

Source: Field Data (2024)

From Table 1 above, the result of training the model showed that when full data were used to train the model, the model achieved the highest accuracy when the entry status class variable was used. The model achieved an accuracy of 75% where gender and age class achieved 66.5% and 59.1% respectively. Also for 10-fold Cross-validation, the model achieved an accuracy of 75%, which was the highest as compared to the rest of the performance for age and gender classes. Again, when 66%-split was used the model achieved 78.3% for the entry status class which was the highest score. The model was trained poorly by achieving an accuracy of 38.3% when age was selected as the class variable.

The variations of model performance using four performance metrics (accuracies, precision, recall, and F1 Score) are depicted by the chart in Figure 1.

Figure 1: J48 Model Training Visualization Source: Field Data (2024)

4.3 Results of Model Training using Random Forest algorithm

The model was trained using a Random forest algorithm. The training results are shown in Table 2.

Table 2: Model Training results using Random forest

Test Mode and Class		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Evaluate Training Data	Gender Class	96.6	96.6	96.6	96.6
	Entry Status Class	**97.2**	97.1	97.2	97.1
	Age Class	96.0	96.1	96.0	96.0
10-fold Cross-Validation	Gender Class	50.0	59.2	49.0	49.4
	Entry Status Class	**59.1**	57.8	59.1	58.4
	Age Class	44.3	42.3	44.3	43.1
Split 66% train, remainder test	Gender Class	50.0	55.4	58.3	50.0
	Entry Status Class	**63.0**	61.1	63.3	62.2
	Age Class	38.3	38.7	38.3	38.5

Source: Field Data (2024)

When the model was trained using a Random forest, the prediction of entry status also showed high accuracy. The highest training accuracy was 97.2% in full training data. 10- fold Cross-validation and % - split accuracy were 59.1% and 63.0% respectively. The second class variable was Gender. Age was the last class which was least accurately trained. The variations of accuracies can be visualized well as shown by the chart in Figure 2.

4.4 Results of Model Training using Naïve Bayes algorithm

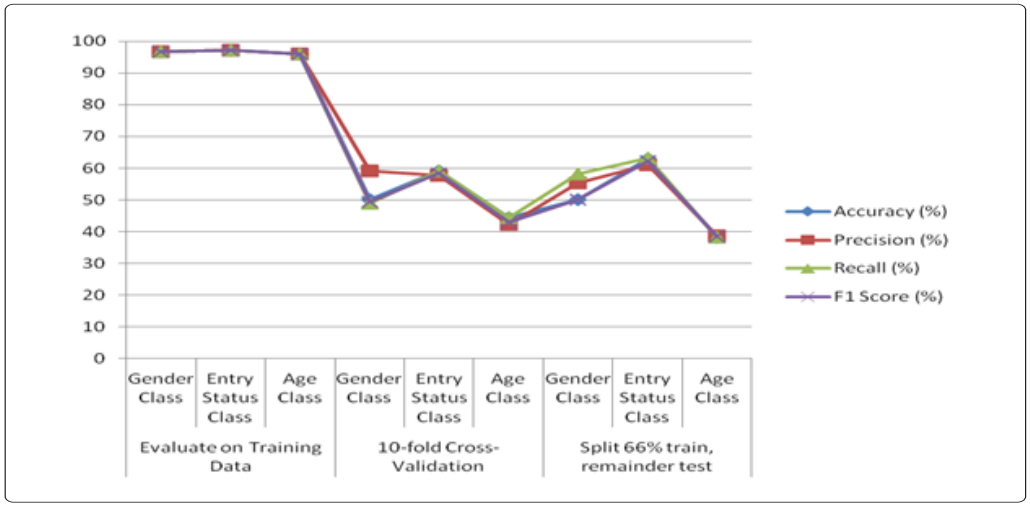

Figure 2: Random Forest Model training visualization

Source: Field Data (2024)

Table 3 below shows the outcomes of training the model using the Naïve Bayes algorithm. The result indicates that on training the model in all three modes, the model achieved higher accuracy when the class variable was the student's entry status. The accuracy of entry status in full training data, 10-fold Cross-validation, and % split were 75.0%, 74.4%, and 78.3% respectively. Also for gender and age classes, the models were trained with low accuracy whereas by age class showed the lowest accuracy.

In all three test modes, the model was poorly trained for age class variables. The accuracy of age in full training data, 10-fold Cross-validation, and % - %-split were 47.7 %, 47.2%, and 35.0% respectively.

Table 3: Model Training results using Naive Bayes

Test Mode and Class		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Evaluate Training Data	Gender Class	59.7	54.9	59.7	55.0
	Entry Status Class	**75.0**	56.3	75.0	64.3
	Age Class	47.7	47.3	47.7	46.4
10-fold Cross-Validation	Gender Class	59.1	53.9	59.1	54.1
	Entry Status Class	**74.4**	56.1	74.4	64.0
	Age Class	47.2	46.8	47.2	46.0
Split 66% train, remainder test	Gender Class	56.7	52.8	56.7	50.5
	Entry Status Class	**78.3**	61.4	78.3	68.8
	Age Class	35.0	27.1	35.0	30.5

Source: Field Data (2024)

The variations of accuracy, Precision, Recall, and F1 Score can be visualized in Figure 3.

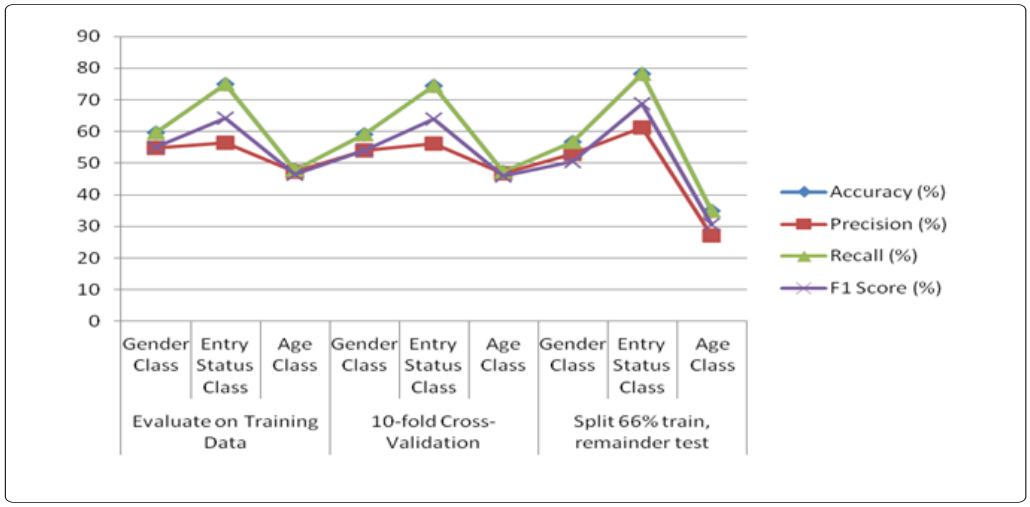

Figure 3: The variations of accuracy, Precision, Recall, and F1 Score

Source: Field Data (2024)

4.4 Results of Prediction Phase

After training the models, each of the three models was tested to determine their ability to predict new, unseen data—data that was not used during the training process. 172 new students' records of the second semester of their bachelor's degree in Records Archives and Information Management were used.

Table 4: Prediction summary using user-supplied test data

Algorithm used and Class		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
J48	Gender Class	48.0	53.5	48.8	49.7
	Entry Status Class	**75.0**	56.3	75.0	64.3
	Age Class	50.6	73.0	50.6	35.2
Random Forest	Gender Class	56.4	59.6	56.4	57.2
	Entry Status Class	**71.5**	66.0	71.5	67.5
	Age Class	48.8	40.8	48.8	41.4
Naïve Bayes	Gender Class	48.3	54.7	48.3	48.6
	Entry Status Class	**75.0**	56.3	75.0	64.3
	Age Class	51.2	53.1	51.2	37.5

Source: Field Data (2024)

Table 4 above shows the results of testing the models with all three algorithms. All three machine learning algorithms showed significant accuracy in the prediction of entry status. J48, Random forest, and Naïve Bayes achieved 75.0%, 71.5%, and 75.0% accuracy respectively. Therefore, it implies that entry status has a significant impact on students' performance.

All three machine learning algorithms showed in small amount of accuracy in the prediction of both gender and class. For gender class, J48, Random forest, and Naïve Bayes achieved 48.0%, 56.4%, and 48.3% accuracy respectively. For age class, J48, Random Forest, and Naïve Bayes achieved 50.6%, 48.8%, and 51.2% accuracy respectively. Therefore, it implies that gender and age have the least impact on students' performance metrics. All three algorithms were achieved with roughly an average of 50% accuracy.

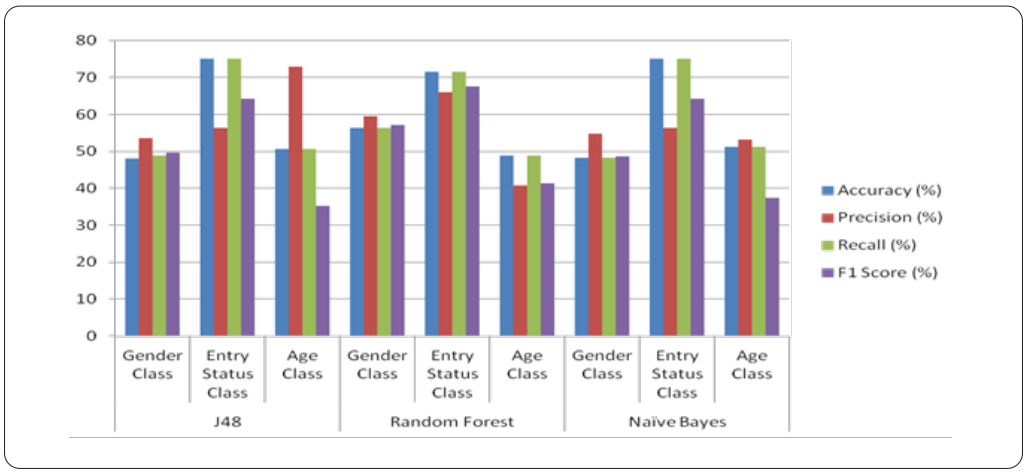

Figure 4: Comparison of entry status prediction by different models/algorithms

Source: Field Data (2024)

Figure 4 above displays the performance of each algorithm for each class variable. The graphs show higher performance trends with entry status. For all three models, all four performance metrics are high in the prediction of entry status.

4.5 Discussions

The experimental results using all three machine learning algorithms (Naïve Bayes, Decision tree, and Random Forests) were performed with insignificant accuracy for classifying students' gender, in both the training phase and in prediction phase. The results imply that student gender does not influence academic performance. This finding is compatible with the finding of Durak and Bulut (2024), which revealed that performance in programming courses is not influenced by gender. The findings also are compatible with the findings of Kapinga and Amani (2016) which revealed the inexistence of gender impact on the academic performance of students. Therefore, both male and female students should follow proper learning strategies in order to improve their academic performance and ignore the gender stereotypes that may negatively affect their academic performance.

The experimental results also showed lower performance in using all three machine learning algorithms in both the training phase and the prediction phase for classifying age class variables. This result also signifies the inexistence impact of age on students' academic performance. This finding is similar to the finding of Almutairi *et al.,* (2019) which uncovered that age had no impact on the academic performance of students. On the other side, the finding in this study is antagonistic to the finding of Unal (2020) which explored the prediction of student's performance in Mathematics and revealed the influence of age on performance in mathematics and the Portuguese language. Undergraduate students of different ages have to adhere to studying principles and methodology to have good performance regardless of their age.

The results of this finding showed that students' entry status (Diploma of form six) influenced students' academic performance. Naïve Bayes Classifier, Decision tree, and random forests achieved significant accuracy in training the model as well as in testing the

model. Undergraduate students who joined with diplomas and those who joined from form six seemed to have heterogeneous performance. There is limited literature that directly examines the performance of form six and diploma in undergraduate studies. The findings of this study slightly match the findings of Kapinga and Amani (2016) which revealed that there is a positive relationship between the educational background of students such as grades scored in lower levels and final examination scores.

5.0 Conclusion and Recommendations

This study is focused on the prediction of gender, age, and entry status of students (diploma or for six) by using four students' academic metrics: Average course work score (AVCA), Average Semester Exam Score (AVSE), number of passed modules (NPASS) and number of failed modules (NFAIL), at Tanzania Public Service College, Dar es Salaam campus for undergraduate students who study a bachelor degree of Records, Archives and Information Management. Decision trees, random forest, and Naïve Bayes machine learning algorithms were used in developing and validating the models. Based on the findings and discussion presented, the study concludes that the entry status of students significantly impacts students' academic performance in undergraduate studies. However, students' academic performance is less influenced by gender and age of students.

The findings in this study suggest that educational institutions, assessors, and moderators to take action on how to incorporate both diploma and form six students. Since diploma and form six students attend the same classes in undergraduate programs, measures should be taken to have homogeneous performances. For instance, aligning the curriculums and teaching methods that favour both groups may help to reduce or eliminate the imbalance.

References

Almutairi, S., Shaiba, H., & Bezbradica, M. (2019). Predicting students' academic performance and main behavioural features using data mining techniques. In *Advances in Data Science, Cyber Security and IT Applications: First International Conference on Computing, ICC 2019, Riyadh, Saudi Arabia, December 10–12, 2019, Proceedings, Part I 1*, 245-259. Springer International Publishing.

Barrouillet, P. (2015). Theories of cognitive development: From Piaget to today. *Developmental Review*, *38*, 1-12.

Bourdieu, P. (2011). The forms of capital. (1986). *Cultural theory: An anthology*, *1*(81-93), 949.

Dean, S., & Illowsky, B. (2013). *Principles of business statistics*. Connexions, Rice University.

Drousiotis, E., Shi, L., & Maskell, S. (2021). Early predictor for student success based on behavioural and demographical indicators. In *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17*, 161-172. Springer International Publishing.

Durak, A., & Bulut, V. (2024). Classification and prediction-based machine learning algorithms to predict students' low and high programming performance. *Computer Applications in Engineering Education*, *32*(1), e22679.

Eagly, A. H., & Wood, W. (2012). Social role theory. *Handbook of theories of social psychology*, *2*, 458-476.

Eccles, J. S., Jacobs, J. E., & Harold, R. D. (1990). Gender role stereotypes, expectancy effects, and parents' socialisation of gender differences. *Journal of Social Issues*, *46*(2), 183-201.

Güre, Ö. B. (2023). Investigating the Performance of Feature Selection Methods in Classifying Student Success. International Journal of Education Technology and Scientific Researches, 8(24), 2695-2728.

Hussain, M. M., Akbar, S., Hassan, S. A., Aziz, M. W., & Urooj, F. (2024). Prediction of Student's Academic Performance through Data Mining Approach. *Journal of Informatics and Web Engineering*, *3*(1), 241-251.

Kapinga, O., & Amani, J. (2016). Determinants of students' academic performance in higher learning institutions in Tanzania. *Journal of Education and Human Development*, *5*(4), 78-86.

Kimeli, C. M., Charles, O., & Douglas, N. M. (2019). Empirical analysis of age and gender as predictors of performance in examination among adult learners. *European Journal of Education Studies*.

Mallak, S., Kanan, M., Al-Ramahi, N., Qedan, A., Khalilia, H., Khassati, A. & Al-Sartawi, A. (2023). Using Markov chains and data mining techniques to predict students' academic performance. *Inf. Sci. Lett*, *12*(9), 2073-2083.

Mkwazu, H. R., & Yan, C. (2020). Grade prediction method for university course selection based on decision tree. In *Proceedings of the 2020 International Conference on Aviation Safety and Information Technology*, 593-599.

Nyikos, M., & Hashimoto, R. (1997). Constructivist theory applied to collaborative learning in teacher education: In search of ZPD. *The Modern Language Journal*, *81*(4), 506-517.

Orrego Granados, D., Ugalde, J., Salas, R., Torres, R., & López-Gonzales, J. L. (2022). Visual-predictive data analysis approach for the academic performance of students from a Peruvian University. *Applied Sciences*, *12*(21), 11251.

Pruitt, D. G. (2014). The evolution of readiness theory. In *Handbook of international negotiation: Interpersonal, intercultural, and diplomatic perspectives*, 123-138. Cham: Springer International Publishing.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160.

Su, Y. S., Lin, Y. D., & Liu, T. Q. (2022). Applying machine learning technologies to explore students' learning features and performance prediction. *Frontiers in Neuroscience*, *16*, 1018005.

Ünal, F. (2020). Data mining for student performance prediction in education. *Data Mining-Methods, Applications and Systems*, *28*, 423-432.

Ye, N. (2013). *Data mining: theories, algorithms, and examples*. CRC press.