International Journal of Educational Research Excellence (IJERE)
https://ejournal.ipinternasional.com/index.php/ijere

Volume 04, Issue 01, January-June 2025
e-ISSN:2830-7933
DOI: 10.55299/ijere.v4i1.1053

Multivariate Analysis of Determinants of Student Learning Achievement: Discriminant Analysis and Random Forest Approach

Nadrah Nadrah*), Indah Miftah Awaliah 2)

1) *Universitas Muhammadiyah Makassar, Makassar, Indonesia*
2) *Universitas Islam Negeri Alauddin Makassar, Makassar, Indonesia*

Correspondence: nadrah@unismuh.ac.id
Article history: received August 18, 2024; revised September 05, 2024; accepted May 01, 2025

Abstract
This study aims to identify factors that influence student learning achievement using the Random Forest method and linear discriminant analysis (LDA). The data used include variables such as gender, part-time work, days of absence, extracurricular activities, weekly hours of independent study, and grades from various subjects. The results of the analysis show that weekly hours of independent study are the most dominant factor influencing student academic achievement, followed by involvement in extracurricular activities. In addition, student attendance was also found to be an important factor, with a significant correlation between days of absence and part-time work and gender. These findings provide valuable insights for educators and policy makers to encourage independent learning practices, support student involvement in extracurricular activities, and reduce student absenteeism. Educational strategies that focus on these factors are expected to significantly improve student academic achievement.

Keywords: Learning Achievement, Independent Study Hours, Extracurricular Activities, Linear Discriminant Analysis, Random Forest

I. INTRODUCTION

Student academic achievement is one of the key indicators of a successful education system. Understanding the factors that influence student performance is crucial for educators, policymakers, and researchers in improving the quality of education. Various variables, including demographic, environmental, and academic factors, are thought to significantly impact student achievement [1], [2].

This study aims to analyze the determinants of student achievement using a multivariate approach, specifically Discriminant Analysis and Random Forest. Discriminant Analysis is an effective statistical method for classifying data into predefined categories based on certain characteristics [3]. Meanwhile, Random Forest is a robust and flexible machine learning algorithm that handles complex data and provides accurate results in prediction and classification [4].

Several previous studies have identified various factors that affect student performance, such as gender, part-time employment, the number of days absent, participation in extracurricular activities, and weekly self-study hours [5], [6]. Additionally, grades in subjects like mathematics, history, physics, chemistry, biology, English, and geography are considered important indicators of student achievement [7]. Higher stress levels are often correlated with lower academic performance [8]. Furthermore, the transition to virtual learning environments has introduced new challenges, such as reduced social interaction and increased screen time, which can affect students' mental health and academic performance [9].

In this study, we will use two complementary approaches to analyze the available data. Discriminant Analysis will help understand group differences among students based on their characteristics, while Random Forest will provide deeper insights into the relative contribution of each variable to student performance.

By combining these approaches, we hope to gain a more comprehensive understanding of the factors that influence student achievement. The findings from this study are expected to contribute to the development of more effective, evidence-based educational strategies and to assist educators in designing appropriate interventions to improve students' academic performance.

Volume 04, Issue 01, January-June 2025
e-ISSN:2830-7933
DOI: 10.55299/ijere.v4i1.1053

II. METHODS

A. Data Source

The data used in this research comes from a dataset available on the Kaggle platform, created by Mark Medhat [10]. This dataset contains various variables relevant to student academic performance, including gender, part-time employment, number of days absent, participation in extracurricular activities, weekly self-study hours, and grades in subjects such as mathematics, history, physics, chemistry, biology, English, and geography. This dataset was chosen for its comprehensive and structured information, allowing for in-depth analysis using Discriminant Analysis and Random Forest methods. By utilizing the data in Table 1, this research aims to identify key factors that influence student performance more accurately and holistically.

TABLE I
TABLE OF KEY FACTORS THAT INFLUENCE STUDENT LEARNING ACHIEVEMENT MORE ACCURATELY AND HOLISTICALLY

gender	part_time_job	absence_days	extracurricular_activities	weekly_self_study_hours	math_score	history_score	physics_score	chemistry_score	biology_score	english_score	geography_score
male	FALSE	3	FALSE	27	73	81	93	97	63	80	87
female	FALSE	2	FALSE	47	90	86	96	100	90	88	90
female	FALSE	9	TRUE	13	81	97	95	96	65	77	94
female	FALSE	5	FALSE	3	71	74	88	80	89	63	86
male	FALSE	5	FALSE	10	84	77	65	65	80	74	76

In this study, researchers used several features to analyse the factors that influence student achievement. The first feature is gender, which is a qualitative variable with categories male and female. The second feature is part-time job, which indicates whether students have part-time jobs and consists of two categories: TRUE and FALSE.
Furthermore, the number of absence days is a quantitative variable that shows the number of days of student absence during a certain period. Participation in extracurricular activities is also a qualitative variable that shows whether students participate in extracurricular activities, with two categories: TRUE and FALSE.
Weekly self-study hours are a quantitative variable that measures the number of hours students spend on self-study each week. In addition, we also use grades in various subjects as quantitative variables. These include math scores, history scores, physics scores, chemistry scores, biology scores, English scores, and geography scores. These variables were chosen because they cover various aspects that are believed to have a significant influence on student learning achievement, both in terms of demographics, extracurricular activities, and academic performance in various subjects. By using the Discriminant Analysis and Random Forest methods, this study aims to identify and understand the contribution of each variable to student learning outcomes in more depth and comprehensively.

B. Discriminant Analysis

Discriminant analysis is a statistical method used to classify observations into predetermined categories based on a set of predictor variables. Following research papers such as [11], [12], [13], [14] and [15]. This technique is used to predict membership in categorical. The mechanism involves identifying linear combinations of predictor variables that optimally discriminate between different classes. This linear discriminant function can then be used to categorize new observations. The mathematical representation of the linear discriminant function for k classes is given by:

$$D_k(x) = \beta_{k0} + \beta_{k1}x_1 + \beta_{k2}x_2 + \cdots + \beta_{kp}x_p \quad (1)$$

Where $D_k(x)$ represents the discriminant score for class k, x1, x2, ..., xp are the predictor variables, and βk0, βk1, βk2, ..., βkp are the coefficients estimated from the data. These coefficients are determined to maximize the separation between classes while minimizing the variation within classes.

C. Random Forest

Random Forest is an effective machine learning method for classification and regression, as it combines predictions from multiple decision trees to improve accuracy and reduce the risk of overfitting. This method has been used in several previous studies such as [16], [17], [18], [19], and [20]. The process begins by dividing the data into two sets: training data

Volume 04, Issue 01, January-June 2025
e-ISSN:2830-7933
DOI: 10.55299/ijere.v4i1.1053

and test data. Training data is used to build the model, while test data is used to evaluate model performance. From the training data, multiple decision trees are formed, each using a random subset of the data and a random subset of the features. Each decision tree provides its own prediction based on the subset of data used to build it. Predictions from all decision trees are then combined to produce a final prediction. For classification, the final prediction is the class most frequently predicted by the trees (majority voting), while for regression, the final prediction is the average of all tree predictions.

$$h_i(x) = argmax_{c \in C} \sum_{j=1}^{N} I\left(h_{i,j}(x) = C\right) \quad (2)$$

Where $h_i(x)$ is the prediction of the i-th tree for input x, C is the class set, and I is the indicator function.

III. RESULTS AND DISCUSSION

The results of the discussion in this study are presented visually using box plots and biplots. Box plots are used to show the distribution and comparison of data between groups, while biplots are used to illustrate the relationships between variables and patterns found in the data.

A. Random Forest Analysis

The results of the Random Forest analysis show that the variable ***weekly_self_study_hours*** show the highest feature in the box plot of all subjects as shown in Fig. 1, Fig. 2, Fig. 3, Fig. 4, Fig. 5, Fig. 6 and Fig. 7. This means that of all the variables analyzed, ***weekly_self_study_hours*** are the most influential factor in determining student learning achievement. In simple terms, this shows that the amount of time students spends studying independently each week has a strong correlation with the grades they get in various subjects.

In the context of education, this finding is very important because it emphasizes the importance of self-study in improving student academic achievement. It can also provide valuable insights for teachers, parents, and education policy makers to encourage self-study practices among students. For example, schools can design programs or provide resources that support students to allocate more time to study outside of school hours. In addition, parents can be motivated to create a conducive environment at home that facilitates their children's self-study hours.

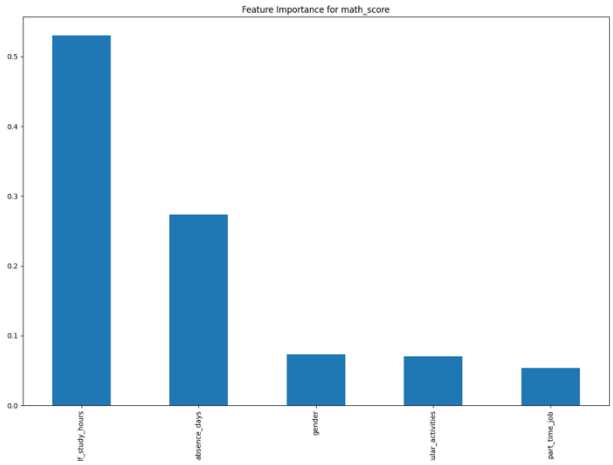

Fig. 1 Factors Affecting Mathematics Scores

Volume 04, Issue 01, January-June 2025
e-ISSN:2830-7933
DOI: 10.55299/ijere.v4i1.1053

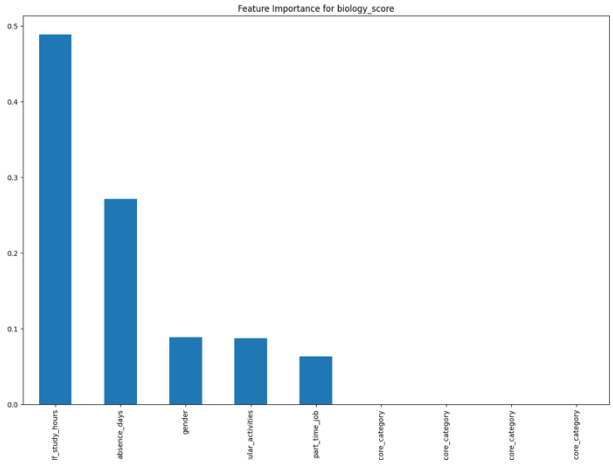

Fig. 2 Factors Affecting Biology Scores

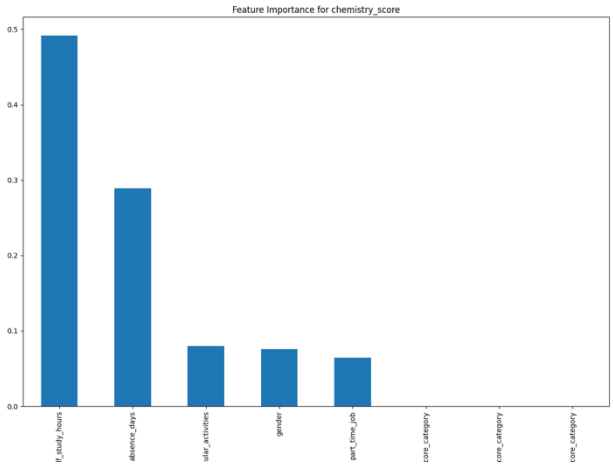

Fig. 3 Factors Affecting Chemistry Scores

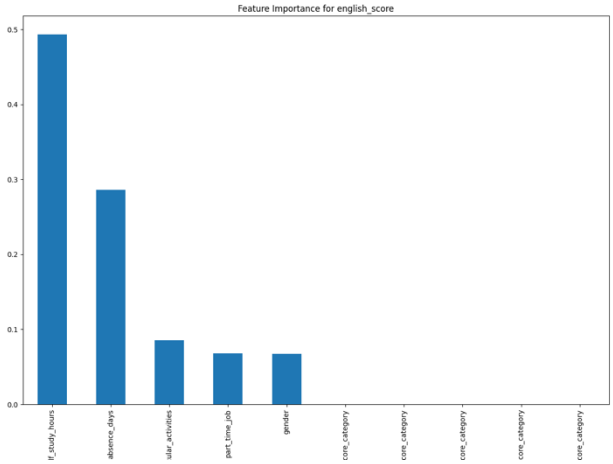

Fig. 4 Factors Affecting English Scores

International Journal of Educational Research Excellence (IJERE)
https://ejournal.ipinternasional.com/index.php/ijere

Volume 04, Issue 01, January-June 2025
e-ISSN:2830-7933
DOI: 10.55299/ijere.v4i1.1053

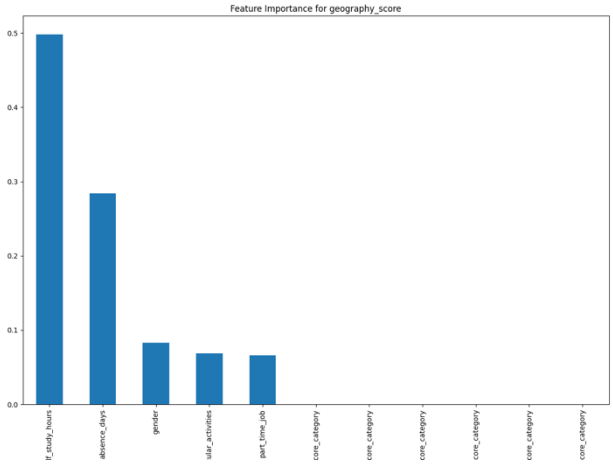

Fig. 5 Factors Affecting Geography Scores

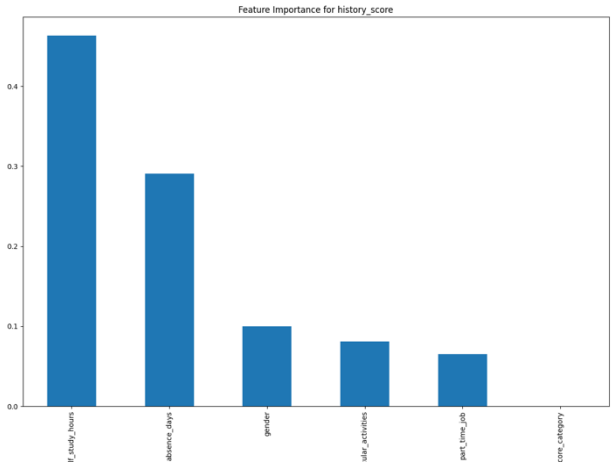

Fig. 6 Factors Affecting History Scores

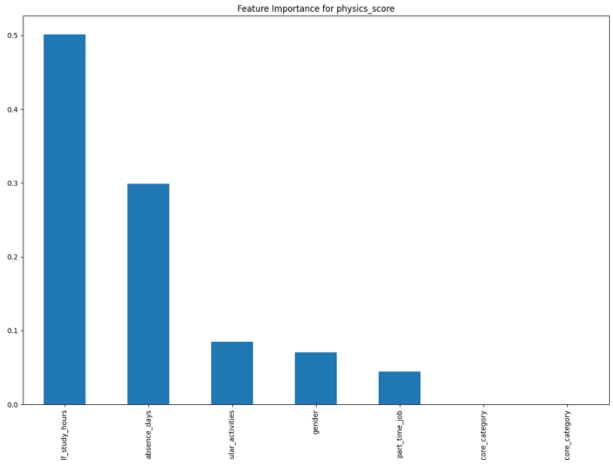

Fig. 7 Factors Affecting Physics Scores

Volume 04, Issue 01, January-June 2025
e-ISSN:2830-7933
DOI: 10.55299/ijere.v4i1.1053

B. Discriminant Analysis

The results of the biplot analysis using discriminant analysis as shown in Figure 8, Figure 9, Figure 10, Figure 11, Figure 12, Figure 13, and Figure 14. Show that the **weekly_self_study_hours** variable has the longest vector or arrow compared to other factors. This indicates that **weekly_self_study_hours** are the most significant variable in separating student achievement categories in the discriminant space. The length of the arrow indicates the strength of the variable's contribution to the variation in the data.

In addition, the position of the **weekly_self_study_hours** arrow which is close to **extracurricular_activities** indicate a positive correlation between the two variables. This means that students who have high weekly self-study hours tend to also be involved in extracurricular activities. The closeness of these two arrows indicates that these two factors together contribute significantly to student achievement.

The arrow indicating **absence_days** is the second longest factor after **weekly_self_study_hours**, indicating that the number of days absent also has a significant influence on student achievement. The correlation between **absence_days** and whether students have a part-time job and gender is seen from the adjacent arrow positions. This indicates that students who are often absent are also likely to have a **part_time_job** and there are differences in absence patterns based on **gender**.

In conclusion, this biplot shows that weekly self-study hours are the most dominant factor influencing student achievement, followed by involvement in extracurricular activities. Meanwhile, student attendance is also an important factor influenced by part-time jobs and gender.

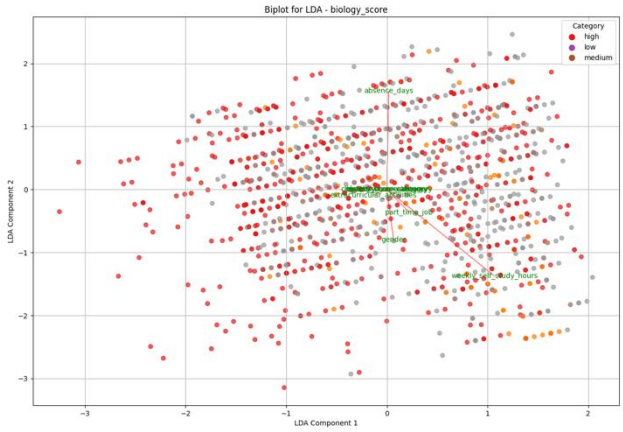

Fig. 8 Biology Lesson BiPlot Results

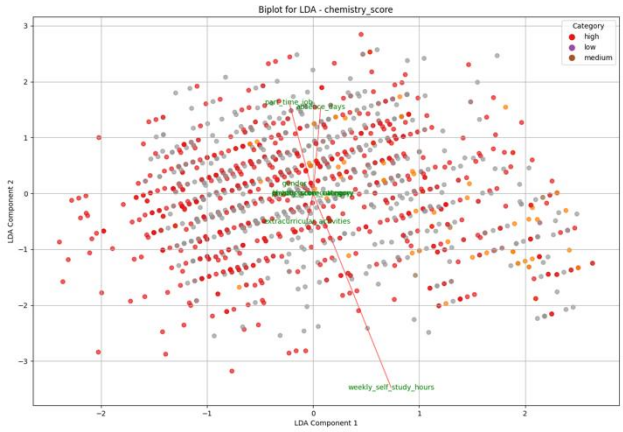

Fig. 9 Chemistry Lesson BiPlot Results

Volume 04, Issue 01, January-June 2025
e-ISSN:2830-7933
DOI: 10.55299/ijere.v4i1.1053

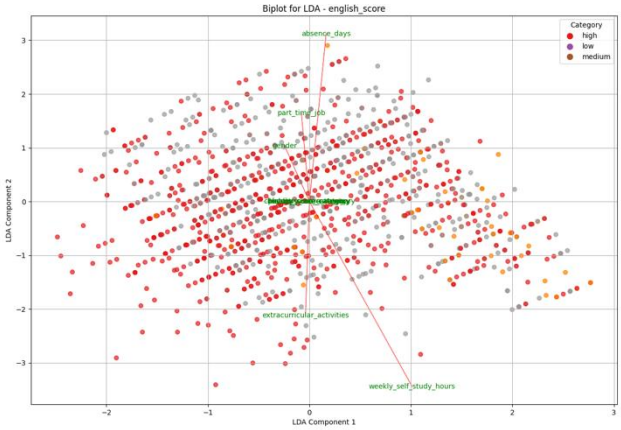

Fig. 10 English Lesson BiPlot Results

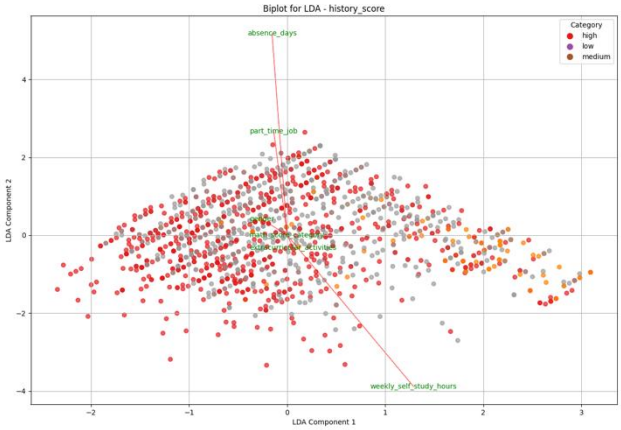

Fig. 11 History Lesson BiPlot Results

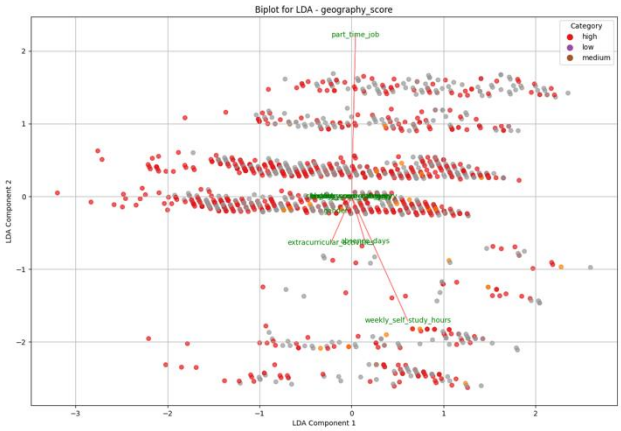

Fig. 12 Geography Lesson BiPlot Results

Volume 04, Issue 01, January-June 2025
e-ISSN:2830-7933
DOI: 10.55299/ijere.v4i1.1053

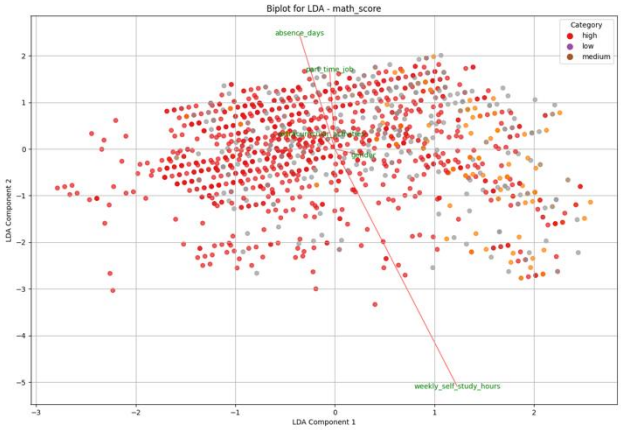

Fig. 13 BiPlot Results of Mathematics Lesson

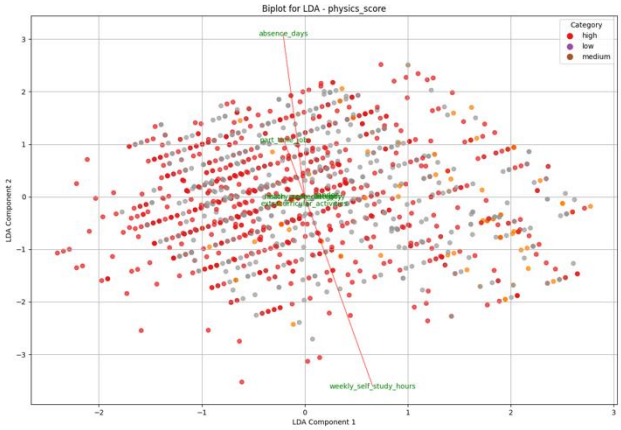

Fig. 14 Physics Lesson BiPlot Results

IV. CONCLUSIONS

Based on the results of the analysis that has been done, this study has succeeded in identifying significant factors that influence student learning achievement using the Random Forest method and linear discriminant analysis (LDA). The variable **weekly_self_study_hours** are proven to be the most dominant factor in determining student academic achievement. This finding indicates that students who spend more time studying independently tend to have better achievements in various subjects. In addition, the LDA biplot analysis shows that this variable is also positively correlated with **extracurricular_activities**, which means that students who are involved in extracurricular activities tend to also allocate more time for independent study.

The second significant factor is **absence_days**, which shows that the level of student attendance also has a major effect on their learning achievement. The correlation between **absence_days** with the variables **part_time_job** and **gender** indicate that students with part-time jobs and certain absence patterns based on gender have different academic achievements. These findings provide important insights for educators and policymakers to encourage independent learning practices, support student engagement in extracurricular activities, and reduce absenteeism through strategies tailored to student circumstances.

Volume 04, Issue 01, January-June 2025
e-ISSN:2830-7933
DOI: 10.55299/ijere.v4i1.1053

Overall, this study highlights the importance of independent learning hours, extracurricular engagement, and attendance in improving student achievement. Educational strategies that focus on improving these factors can have a significant positive impact on student academic achievement.

REFERENCES

[1] R. S. Vieira dan M. Arends-Kuenning, "Affirmative action in Brazilian universities: Effects on the enrollment of targeted groups," Economics of Education Review, vol. 73, hlm. 101931, Des 2019, doi: 10.1016/j.econedurev.2019.101931.
[2] H. Jamshidifarsani, S. Garbaya, T. Lim, P. Blazevic, dan J. M. Ritchie, "Technology-based reading intervention programs for elementary grades: An analytical review," Computers & Education, vol. 128, hlm. 427–451, Jan 2019, doi: 10.1016/j.compedu.2018.10.003.
[3] W. R. Klecka, Discriminant analysis. Sage, 1980.
[4] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, hlm. 5–32, 2001, doi: 10.1023/A:1010933404324.
[5] S. J. Min dan D. Y. Wohn, "All the news that you don't like: Cross-cutting exposure and political participation in the age of social media," Computers in Human Behavior, vol. 83, hlm. 24–31, Jun 2018, doi: 10.1016/j.chb.2018.01.015.
[6] A. L. Duckworth, C. Peterson, M. D. Matthews, dan D. R. Kelly, "Grit: Perseverance and passion for long-term goals.," Journal of Personality and Social Psychology, vol. 92, no. 6, hlm. 1087–1101, 2007, doi: 10.1037/0022-3514.92.6.1087.
[7] K. R. Wentzel dan D. B. Miele, Ed., Handbook of Motivation at School, 0 ed. Routledge, 2009. doi: 10.4324/9780203879498.
[8] M. Jwaifell, "A Proposed Model for Electronic Portfolio to Increase both Validating Skills and Employability," Procedia - Social and Behavioral Sciences, vol. 103, hlm. 356–364, Nov 2013, doi: 10.1016/j.sbspro.2013.10.345.
[9] E. J. Sintema, "Effect of COVID-19 on the Performance of Grade 12 Students: Implications for STEM Education," EURASIA J MATH SCI T, vol. 16, no. 7, Apr 2020, doi: 10.29333/ejmste/7893.
[10] "Student scores." Diakses: 21 Juli 2024. [Daring]. Tersedia pada: https://www.kaggle.com/datasets/markmedhat/student-scores?resource=download
[11] H. Du dan H. Li, "Multi-view Canonical Representation Discriminant Analysis," dalam 2023 International Conference on Communications, Computing and Artificial Intelligence (CCCAI), Shanghai, China: IEEE, Jun 2023, hlm. 114–118. doi: 10.1109/CCCAI59026.2023.00029.
[12] X. Li dan H. Wang, "On Mean-Optimal Robust Linear Discriminant Analysis," dalam 2022 IEEE International Conference on Data Mining (ICDM), Orlando, FL, USA: IEEE, Nov 2022, hlm. 1047–1052. doi: 10.1109/ICDM54844.2022.00129.
[13] J. Kim, Y. Lee, dan Z. Liang, "The Geometry of Nonlinear Embeddings in Kernel Discriminant Analysis," IEEE Trans. Pattern Anal. Mach. Intell., hlm. 1–14, 2022, doi: 10.1109/TPAMI.2022.3192726.
[14] N. Nagananda dan A. Savakis, "GILDA++: Grassmann Incremental Linear Discriminant Analysis," dalam 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA: IEEE, Jun 2021, hlm. 4448–4456. doi: 10.1109/CVPRW53098.2021.00502.
[15] A. F. Lapanowski dan I. Gaynanova, "Compressing large sample data for discriminant analysis," dalam 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA: IEEE, Des 2021, hlm. 1068–1076. doi: 10.1109/BigData52589.2021.9671676.
[16] S. Abdullah dan G. Prasetyo, "EASY ENSEMMBLE WITH RANDOM FOREST TO HANDLE IMBALANCED DATA IN CLASSIFICATION," JFMA, vol. 3, no. 1, hlm. 39–46, Jun 2020, doi: 10.14710/jfma.v3i1.7415.
[17] M. Y. Aldean, P. Paradise, dan N. A. Setya Nugraha, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac)," INISTA, vol. 4, no. 2, hlm. 64–72, Jun 2022, doi: 10.20895/inista.v4i2.575.
[18] N. Amini, T. H. Saragih, M. R. Faisal, A. Farmadi, dan F. Abadi, "IMPLEMENTASI ALGORITMA GENETIKA UNTUK SELEKSI FITUR PADA KLASIFIKASI GENRE MUSIK MENGGUNAKAN METODE RANDOM FOREST," JIP, vol. 9, no. 1, hlm. 75–82, Nov 2022, doi: 10.33795/jip.v9i1.1028.
[19] Z. A. Dwiyanti dan C. Prianto, "Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest," JTI, vol. 17, no. 2, hlm. 127–137, Okt 2023, doi: 10.36787/jti.v17i2.1136.
[20] T. C. Herdiyani dan A. U. Zailani, "SENTIMENT ANALYSIS TERKAIT PEMINDAHAN IBU KOTA INDONESIA MENGGUNAKAN METODE RANDOM FOREST BERDASARKAN TWEET WARGA NEGARA INDONESIA," JTSI, vol. 3, no. 2, hlm. 154–165, Sep 2022, doi: 10.35957/jtsi.v3i2.2920.