Preprints.org

Article

Not peer-reviewed version

A Predictive Model for Calculating Student GPA Using Machine Learning Algorithms

Noor Ul-Aziz and Noor Ul Amin *

Posted Date: 11 November 2025

doi: 10.20944/preprints202511.0794.v1

Keywords: student performance; GPA prediction; machine learning; behavioral factor

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2025 doi:10.20944/preprints202511.0794.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

A Predictive Model for Calculating Student GPA Using Machine Learning Algorithms

Noor Ul-Aziz [1] and Noor Ul Amin [2,*]

[1] Department of Software Engineering University of Sialkot, Sialkot, Pakistan
[2] School of Computer Science, Taylor's University, Subang Jaya, Malaysia
* Correspondence: nooraminnawab@gmail.com

Abstract

This study focuses on analyzing students' educational performance and their behavior in relation to their grade point average (GPA), using a dataset that includes socio-behavioral and educational attributes. Machine learning techniques were applied to predict GPA and to develop a predictive model. Exploratory data analysis identified key correlations, and various algorithms were used for GPA prediction. The aim of this study is to assist policymakers in designing strategies to enhance educational outcomes and support student development. It highlights the importance for universities to identify students at risk of low GPA and to improve future predictions to help boost student performance.

Keywords: student performance; GPA prediction; machine learning; behavioral factor

1. Introduction

This dataset is a comprehensive educational dataset that contains a detailed analysis of 2,392 students' academic performance and other curricular activities. It includes 15 columns that cover various metrics such as demographic, behavioral, and educational (or academic) attributes [11].

Demographic features include Student ID, Age, Gender, and Ethnicity. The dataset also includes variables like Parental Education, Weekly Study Time, and Absences, which highlight study habits and parental background [12].

The core objective of this dataset is to evaluate and predict students' educational performance, particularly focusing on GPA (Grade Point Average) and grade class as outcome variables. It provides valuable insights into student performance and allows for multiple types of analysis, especially those involving demographic and behavioral factors [13–15].

The dataset also addresses key challenges in predicting student performance, such as determining their grade class. This problem is approached using various machine learning algorithms, including Random Forest, SVM, ANN, Naïve Bayes, AutoML, KNN, and Decision Tree [16–19]. By applying these algorithms using RapidMiner, their predictive accuracies were compared. The structure of this study follows a systematic research flow, as illustrated in Figure 1. The workflow begins with the Introduction, which sets the foundation by outlining the research problem and objectives [20,21]. This is followed by the Literature Review, which explores existing work and identifies gaps in prior research. The Proposed Methodology section details the techniques and models employed to address the research problem. The Results section presents the findings derived from experimentation and analysis. Finally, the paper concludes with the Conclusion, summarizing key insights and future research directions.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2025 doi:10.20944/preprints202511.0794.v1

Figure 1. Workflow diagram representing the structure of the research paper.

2. Literature Review

This paper focuses on analyzing student educational performance using machine learning techniques. Ghorbani and Ghousi [2] compared different resampling methods for handling class imbalance in predicting student performance using Support Vector Machines (SVM) and SMOTE, achieving the highest accuracy of 73% on an Iranian dataset. Their study enhances understanding in educational data mining and improves predictive modeling.

Mengash [1] applied Artificial Neural Networks (ANN) to analyze data from a Saudi university. Despite fixed admission criteria, only first-year students' performance was used to refine university decision-making using data mining techniques. Albreiki et al. [3] proposed a six-step framework data collection, statistical analysis, preparation, preprocessing, implementation, and evaluation for predicting academic success indicators such as GPA, persistence, and engagement without requiring advanced technical skills.

Yağc◎[4] analyzed Chinese university data using trained ANN models and recommended future refinements to improve prediction accuracy. The model identified critical factors such as exam results, gender, and institutional support, although it also highlighted challenges related to gender classification imbalances.

A study on 1,854 undergraduate students from a Turkish university [5] predicted final term grades using midterm grades, department, and faculty. It found midterm grades to be strong predictors of final outcomes, even without demographic data. Algorithms such as Random Forest, Neural Networks, and SVM achieved accuracies between 70–75%.

Namoun and Alshanqiti [6] surveyed machine learning approaches from 2009 to 2021, focusing on university databases and online learning platforms to predict student performance. Their findings emphasized the importance of incorporating both static and dynamic data to improve educational outcomes and reduce dropout rates.

Zhang et al. [8] explored how higher education institutions could use video learning and data mining techniques across student information systems and learning management systems, applying eight classification algorithms to predict performance.

A study focusing on Nigerian engineering students [7] calculated final CGPA based on performance in the first three years using data mining techniques. By applying k-Nearest Neighbors (kNN) and regression analysis, it achieved 89% prediction accuracy. This approach proved useful for identifying weak students and reducing failure and dropout rates.

Rehman et al. [9] conducted a study on 124 students, showing that those more active in model learning systems tended to achieve better final grades. The study also revealed that female students outperformed males. The results helped identify students' weaknesses and improve the teaching–learning process.

Allensworth and Clark [10] found that high school GPAs were a stronger and more consistent predictor of college graduation than ACT scores, which were shown to be influenced more by school effects. Their study suggested that educational systems should place less reliance on standardized tests and focus more on GPA-based assessments..

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2025 doi:10.20944/preprints202511.0794.v1

3. Proposed Methodology

This research proposes a machine learning-based methodology for predicting student GPA using various classification algorithms. The workflow was implemented in RapidMiner, with a dataset sourced from Kaggle. Figure 2 shows the detailed steps of machine learning workflow.

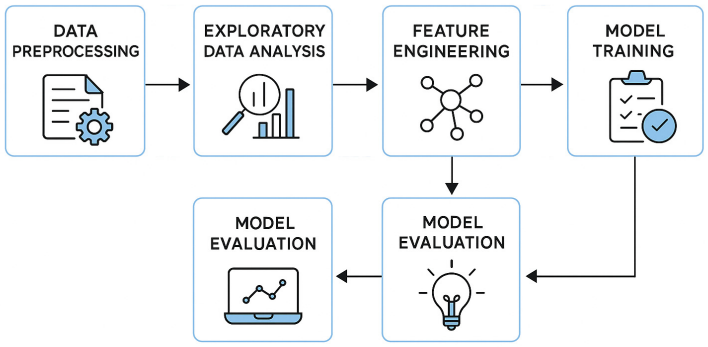

Figure 2. ML Workflow.

The methodology follows these major steps:

I. **Data Preprocessing**: Cleaning missing values, removing irrelevant columns, and normalizing features.
II. **Exploratory Data Analysis (EDA)**: Identifying relationships between key attributes.
III. **Feature Engineering**: Creating new attributes such as total activity count and transforming categorical variables into numerical ones.
IV. **Model Training**: Four classifiers were trained—KNN, Naïve Bayes, Artificial Neural Network (ANN), and AutoMLP. The dataset was split into training and testing subsets.
V. **Model Evaluation**: Accuracy was used as the primary metric to compare model performance.
VI. **Recommendation**: Insights were drawn from the analysis to assist educational institutions in identifying at-risk students and enhancing learning strategies [16].

Table 1 shows the different selected ML models with their parameter descriptions.

Table 1. ML Algorithms and Parameters.

Algorithm	Key Parameters Used
KNN	k=5, Distance: Euclidean
Naïve Bayes	Assumes independent features, Gaussian distribution
ANN	3 hidden layers, ReLU activation, 100 epochs
AutoMLP	Random/Grid Search, Auto-optimized architecture

3.1. Dataset Description

The dataset consists of 2,392 high school students and includes variables such as study habits, parental involvement, and extracurricular activities. The target variable is the Grade Class. After applying all algorithms, ANN achieved the highest accuracy of 77%, making it the best-performing model for predicting student academic outcomes. Table 2 shows dataset description.

Table 2. Dataset Description of Parameters.

Attribute	Description
StudentID	Unique identifier for each student
Age	Age of the student

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2025 doi:10.20944/preprints202511.0794.v1

Gender	Gender of the student (Male = 1, Female = 0)
Ethnicity	Ethnic background of the student
ParentalEducation	Educational level of the parents
GPA	Grade Point Average
GradeClass	Grade category (1, 2, 3, 4)
ParentalSupport	Level of support from parents
StudyTimeWeekly	Number of hours studied per week
Absences	Total number of absences
Tutoring	Whether the student receives tutoring (Yes/No)

Figure 3 shows a sample rapid miner workflow.

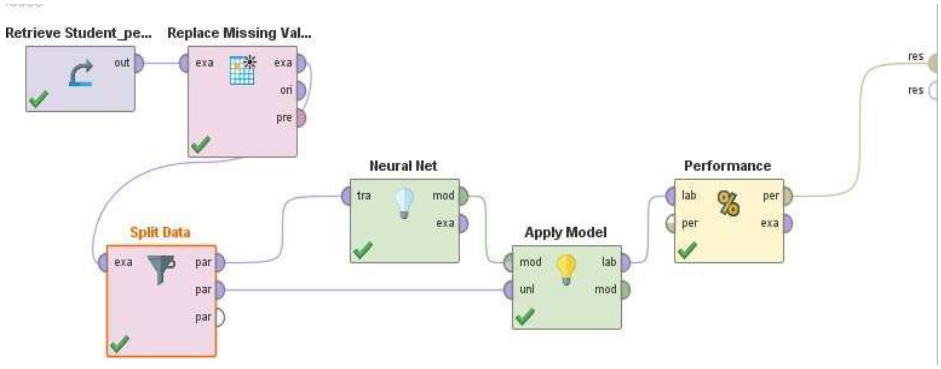

Figure 3. Rapid Miner Workflow.

4. Results

The primary objective of this study was to predict student GPA and understand the factors influencing academic performance using machine learning. The analysis revealed key influencers such as study time, parental support, and extracurricular involvement. These social and behavioral factors play a significant role in shaping educational outcomes.

We used predictive modeling, the system identifies students with lower expected performance, enabling targeted interventions to improve learning outcomes and reduce dropout rates. The study demonstrates that data-driven approaches can significantly enhance educational strategy and student success.

Table 3. Accuracy Comparison of ML Algorithms.

Algorithm	**Accuracy**
KNN	52%
AutoMLP	73%
ANN	**77%**
Naïve Bayes	75%

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2025 doi:10.20944/preprints202511.0794.v1

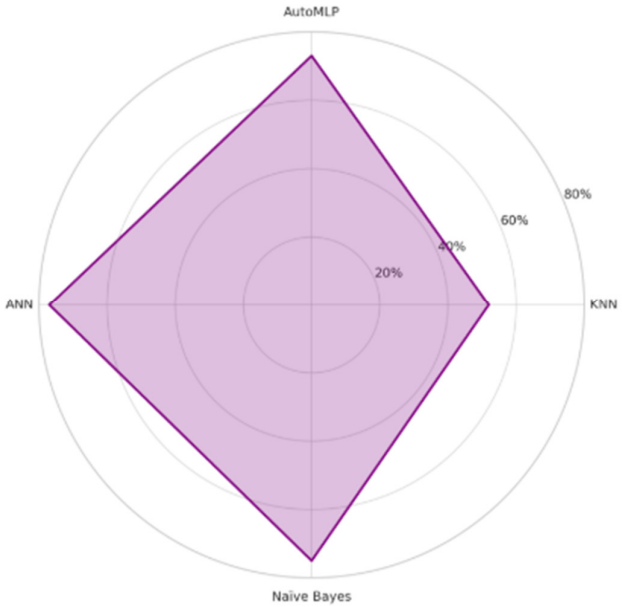

Figure 4. Polar Bar Chart Visualization of Results.

5. Conclusions

The main goal of this research paper is to understand student performance and identify the factors that affect GPA. By applying machine learning techniques, we not only predict students' GPA but also gain insights into their potential for educational and professional success. This study emphasizes the importance of data-driven approaches to enhance the educational system, particularly for supporting underperforming students. Among the tested models, the Artificial Neural Network (ANN) achieved the highest accuracy at 77%, making it the most effective in predicting student GPA. For future work, this model can be further refined to adapt to different educational systems and grading structures. Incorporating a more diverse dataset that includes social, cultural, and economic factors will enhance the model's robustness.

References

1. H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.
2. R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.
3. B. Albreiki, N. Zaki, and H. Alashwal, "Systematic literature review of predicting student performance using machine learning techniques," *Education Sciences*, vol. 11, no. 9, 2021.
4. M. Yağc◉ "Educational data mining: Prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, 2022, doi: 10.1186/s40561-022-00192-z.
5. A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences*, vol. 11, no. 1, pp. 1–28, 2021, doi: 10.3390/app11010237.
6. E. M. Allensworth and K. Clark, "High school GPAs and ACT scores as predictors of college completion: Examining assumptions about consistency across high schools," *Educational Researcher*, vol. 49, no. 3, pp. 198–211, 2020, doi: 10.3102/0013189X20902110.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2025 doi:10.20944/preprints202511.0794.v1

7. Y. Zhang, A. Ghandour, and V. Shestak, "Using learning analytics to predict students' performance in Moodle LMS," *International Journal of Emerging Technologies in Learning*, vol. 15, no. 20, pp. 102–114, 2020, doi: 10.3991/ijet.v15i20.15915.
8. A. U. Rehman et al., "A machine learning--based framework for accurate and early diagnosis of liver diseases: A comprehensive study on feature selection, data imbalance, and algorithmic performance," *International Journal of Intelligent Systems*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/6111312.
9. T. M. Ali et al., "A sequential machine learning-cum-attention mechanism for effective segmentation of brain tumor," *Frontiers in Oncology*, vol. 12, Jun. 2022, doi: 10.3389/fonc.2022.873268.
10. H. Mir et al., "A novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques," *ESC Heart Failure*, Jul. 2024, doi: 10.1002/ehf2.14942.
11. Ahmed, Q. W., Garg, S., Rai, A., Ramachandran, M., Jhanjhi, N. Z., Masud, M., & Baz, M. (2022). Ai-based resource allocation techniques in wireless sensor internet of things networks in energy efficiency with data optimization. Electronics, 11(13), 2071.
12. Khan, N. A., Jhanjhi, N. Z., Brohi, S. N., Almazroi, A. A., & Almazroi, A. A. (2022). A secure communication protocol for unmanned aerial vehicles. CMC-Computers Materials & Continua, 70(1), 601-618.
13. Muzafar, S., & Jhanjhi, N. Z. (2020). Success stories of ICT implementation in Saudi Arabia. In Employing Recent Technologies for Improved Digital Governance (pp. 151-163). IGI Global Scientific Publishing.
14. Jabeen, T., Jabeen, I., Ashraf, H., Jhanjhi, N. Z., Yassine, A., & Hossain, M. S. (2023). An intelligent healthcare system using IoT in wireless sensor network. Sensors, 23(11), 5055.
15. Shah, I. A., Jhanjhi, N. Z., & Laraib, A. (2023). Cybersecurity and blockchain usage in contemporary business. In Handbook of Research on Cybersecurity Issues and Challenges for Business and FinTech Applications (pp. 49-64). IGI Global.
16. Hanif, M., Ashraf, H., Jalil, Z., Jhanjhi, N. Z., Humayun, M., Saeed, S., & Almuhaideb, A. M. (2022). AI-based wormhole attack detection techniques in wireless sensor networks. Electronics, 11(15), 2324.
17. Shah, I. A., Jhanjhi, N. Z., Amsaad, F., & Razaque, A. (2022). The role of cutting-edge technologies in industry 4.0. In Cyber Security Applications for Industry 4.0 (pp. 97-109). Chapman and Hall/CRC.
18. Humayun, M., Almufareh, M. F., & Jhanjhi, N. Z. (2022). Autonomous traffic system for emergency vehicles. Electronics, 11(4), 510.
19. Muzammal, S. M., Murugesan, R. K., Jhanjhi, N. Z., & Jung, L. T. (2020, October). SMTrust: Proposing trust-based secure routing protocol for RPL attacks for IoT applications. In 2020 International Conference on Computational Intelligence (ICCI) (pp. 305-310). IEEE.
20. Brohi, S. N., Jhanjhi, N. Z., Brohi, N. N., & Brohi, M. N. (2023). Key applications of state-of-the-art technologies to mitigate and eliminate COVID-19. Authorea Preprints.
21. Khalil, M. I., Humayun, M., Jhanjhi, N. Z., Talib, M. N., & Tabbakh, T. A. (2021). Multi-class segmentation of organ at risk from abdominal ct images: A deep learning approach. In Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021 (pp. 425-434). Singapore: Springer Nature Singapore.
22. Humayun, M., Jhanjhi, N. Z., Niazi, M., Amsaad, F., & Masood, I. (2022). Securing drug distribution systems from tampering using blockchain. Electronics, 11(8), 1195.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.