Health Leadership and Quality of Life. 2024; 3:.424
doi: 10.56294/hl2024.424

ORIGINAL

Data Analysis and Prediction of Student Academic Performance

Análisis de datos y predicción del rendimiento académico de los estudiantes

Priya Darshini[1]

[1]Chanakya National Law University. Patna, India.

Citar como: Darshini P. Data Analysis and Prediction of Student Academic Performance. Health Leadership and Quality of Life. 2024; 3:.424. https://doi.org/10.56294/hl2024.424

Enviado: 20-03-2024 **Revisado:** 09-08-2024 **Aceptado:** 13-12-2024 **Publicado:** 14-12-2024

Editor: PhD. Prof. Neela Satheesh

Autor para la correspondencia: Priya Darshini

ABSTRACT

Introduction: predicting student performance across datasets with varying distributions remains a complex challenge in educational analytics. This study presents a novel approach to address this issue by utilizing transfer learning techniques to improve prediction accuracy.
Objective: the research leverages a comprehensive dataset from Kaggle, encompassing demographic details, social factors, and academic performance indicators, to uncover significant patterns and relationships that influence student outcomes. By analyzing these factors, the study provides valuable insights that enable students to assess their academic progress, refine their learning strategies, and enhance overall efficiency.
Method: the proposed methodology not only improves predictive accuracy but also bridges existing gaps in understanding student performance across diverse educational contexts.
Results: these findings can be applied to develop personalized support systems, empowering students with actionable recommendations tailored to their individual needs.
Conclusions: by addressing these challenges, the study contributes to a deeper understanding of student performance dynamics and highlights the potential of advanced predictive techniques to drive meaningful educational interventions.

Keywords: Education; Data Analysis; Performance; Demographic Data.

RESUMEN

Introducción: predecir el rendimiento de los estudiantes en conjuntos de datos con distribuciones variables sigue siendo un reto complejo en el análisis educativo. Este estudio presenta un enfoque novedoso para abordar esta cuestión mediante la utilización de técnicas de aprendizaje por transferencia para mejorar la precisión de las predicciones.
Objetivo: la investigación aprovecha un conjunto de datos exhaustivo de Kaggle, que abarca detalles demográficos, factores sociales e indicadores de rendimiento académico, para descubrir patrones y relaciones significativas que influyen en los resultados de los estudiantes. Al analizar estos factores, el estudio proporciona información valiosa que permite a los estudiantes evaluar su progreso académico, perfeccionar sus estrategias de aprendizaje y mejorar la eficiencia general.
Método: la metodología propuesta no solo mejora la precisión predictiva, sino que también salva las brechas existentes en la comprensión del rendimiento de los estudiantes en diversos contextos educativos.
Resultados: estos hallazgos pueden aplicarse para desarrollar sistemas de apoyo personalizados, que proporcionen a los estudiantes recomendaciones prácticas adaptadas a sus necesidades individuales.
Conclusiones: al abordar estos desafíos, el estudio contribuye a una comprensión más profunda de la dinámica del rendimiento de los estudiantes y destaca el potencial de las técnicas predictivas avanzadas para impulsar intervenciones educativas significativas.

Palabras clave: Educación; Análisis de Datos; Rendimiento; Datos Demográficos.

INTRODUCTION

Academic achievement plays a crucial role in educational systems, functioning as a vital benchmark for assessing students' learning capabilities and evaluating the effectiveness of school administration and teaching quality. It reflects not only individual progress but also the success of broader institutional policies and practices. Consequently, predicting academic performance has become a central focus in the field of educational management, evolving alongside advancements in data analysis and technology. The development of predictive models for academic achievement has revolutionized how educators and administrators approach teaching and decision-making.[1] By leveraging these models, educators gain insights into factors influencing student success, enabling them to identify areas where intervention is needed. This allows for personalized support, optimized teaching strategies, and targeted improvements to the educational process, ultimately fostering better learning outcomes for students.

Historically, research on predicting academic performance has relied heavily on statistical methods to analyze and interpret data. These approaches utilized information sourced from various channels, such as educational management systems, student records, or surveys. While these methods have provided valuable insights, their limitations often stem from a narrow focus on specific datasets, potentially overlooking complex interactions between diverse factors that influence student achievement.[2] Today, the integration of advanced analytics, machine learning, and other innovative methodologies has expanded the potential for more accurate and comprehensive predictions. By incorporating data from multiple sources—ranging from demographic information to behavioral and social patterns—educators and policymakers can gain a deeper understanding of the dynamics affecting academic success.[3] This holistic perspective equips institutions with the tools to create adaptive, student-centered learning environments.

Literature review

Prediction accuracy in academic performance largely hinges on the careful selection of relevant indicators. Identifying appropriate input data is a foundational step in constructing reliable predictive models. Research highlights three primary categories of student-related features as consistent input parameters: historical academic performance, student engagement, and demographic data (Tomasevic et al., 2020).[4] Historical academic performance remains a robust predictor of future success. For instance, DeBerard et al. (2004)[5] demonstrated that high school GPA strongly correlates with college academic achievement. Similarly, Shaw et al. (2012) found that combined SAT scores account for approximately 28 % of the variance in first-year college GPA, underlining the reliability of standardized test scores as predictive indicators. Various studies also emphasize the role of prior academic records in forecasting future performance.[6] Student engagement has also been strongly linked to academic achievement. Hussain et al. (2019) observed a moderately strong positive correlation between student engagement and academic performance.[7] With the rise of modern teaching formats such as Massive Open Online Courses (MOOCs) and flipped classrooms, predictive models now utilize student behavior in learning management systems, including interactions with videos, assignment submissions, and participation in forums, to assess engagement and its impact on academic success. Advancements in educational technology, such as artificial intelligence tools (e.g., ChatGPT) and virtual reality, have revolutionized the learning landscape. These technologies, grounded in educational theories like constructivism and experiential learning, provide immersive and interactive experiences that enhance student engagement, motivation, and critical thinking skills. Consequently, they significantly contribute to improved academic outcomes by fostering deeper learning and active participation.[8]

Demographic factors have been widely studied as part of predictive models for academic performance, although their impact on accuracy varies. Some studies report that demographic variables account for approximately 60 % of prediction relevance, while others argue their role is comparatively limited. Beyond demographics, other factors, such as student collaboration, teacher-student communication, and psychological attributes like motivation and attitude, have also been explored.[9] Recent research underscores the critical role of psychological well-being and cognitive processes in education. Motivation and coping strategies, in particular, significantly shape students' learning approaches and influence their overall academic performance. Student achievement is, therefore, a multidimensional construct, encompassing cognitive, behavioral, skill-based, and emotional outcomes rooted in educational experiences. While there is consensus on key predictive indicators, the datasets chosen for student achievement analysis differ across studies, reflecting varying research goals and objectives. This lack of standardized guidelines for dataset selection poses a challenge to achieving consistent results.[10] The rapid advancement of artificial intelligence (AI) and machine learning (ML) technologies has

transformed predictive modeling in education, healthcare, and other fields. However, the complexity of neural networks, often referred to as their “black box” nature, has raised concerns about model transparency and interpretability. To address these issues, interpretable machine learning (IML) has emerged as a promising solution, enabling models to explain their decision-making processes in human-understandable terms.[11] IML techniques are broadly categorized into self-interpreting models and post-hoc interpretation methods. Self-interpreting models, such as linear regression, logistic regression, and decision trees, are inherently transparent due to their simpler structures.[12] In contrast, post-hoc interpretation methods, which can be model-independent or model-specific, analyze complex models but often demand additional computational and analytical resources. By incorporating IML, predictive models can offer valuable insights while maintaining trust and reliability in their applications.

METHOD

Dataset collection

The dataset for this study was sourced from Kaggle[13], a well-established platform known for hosting data science competitions and providing a collaborative environment for data scientists and machine learning practitioners.

Distribution of Exam Scores

This histogram shows the frequency distribution of exam scores among students, indicating how scores are spread across different ranges. A histogram is a type of bar chart that represents the frequency of data points within specified ranges (or bins). Each bar in the histogram corresponds to a range of scores, and the height of the bar indicates how many students achieved scores within that range. Frequency Distribution describes how often each score (or range of scores) occurs in the dataset. Spread of Scores provides insights into the overall performance of students. It can show whether most students scored low, high, or if the scores are evenly distributed.

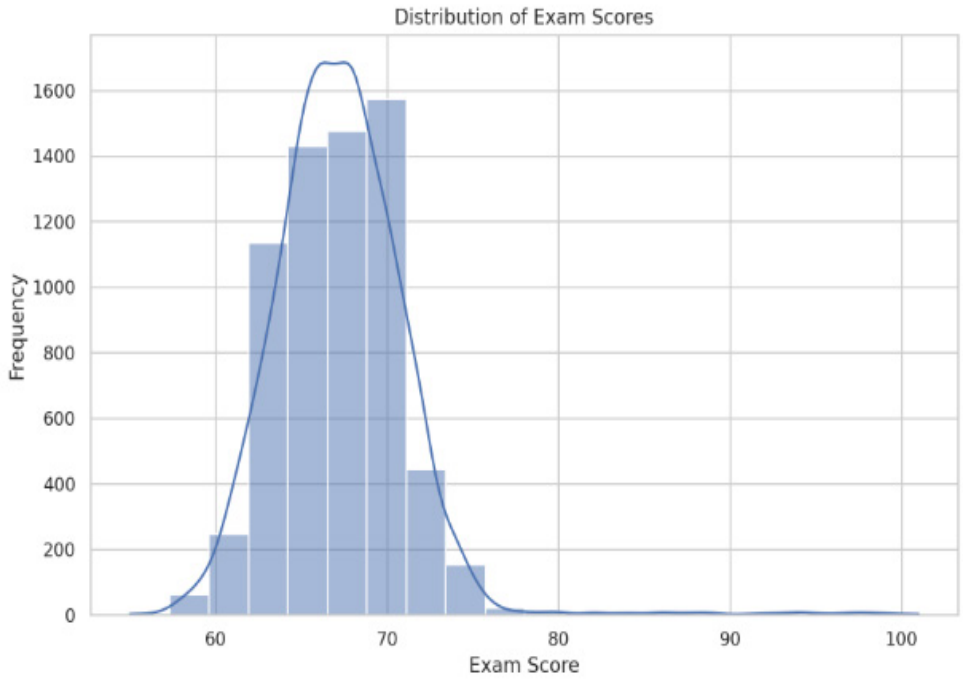

Figure 1. Distribution of exam scores[13]

Hours Studied vs. Exam Score

This scatter plot illustrates the relationship between the number of hours studied and the exam scores, with points colored by gender. It helps to visualize if more study hours correlate with higher exam scores.

Exam Score by Attendance Level

This box plot displays the exam scores categorized by attendance levels, providing insights into how attendance impacts performance” refers to the analysis of the relationship between students’ attendance and their exam scores using a box plot visualization. A box plot (or whisker plot) is a standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile (Q1), median (Q2), third

quartile (Q3), and maximum. It visually represents the central tendency and variability of the data, as well as potential outliers. The box plot serves as a powerful visual tool to analyze and interpret the relationship between attendance and exam performance. It allows educators and researchers to draw conclusions about the importance of attendance in academic success and to identify areas for potential improvement in student engagement and performance.

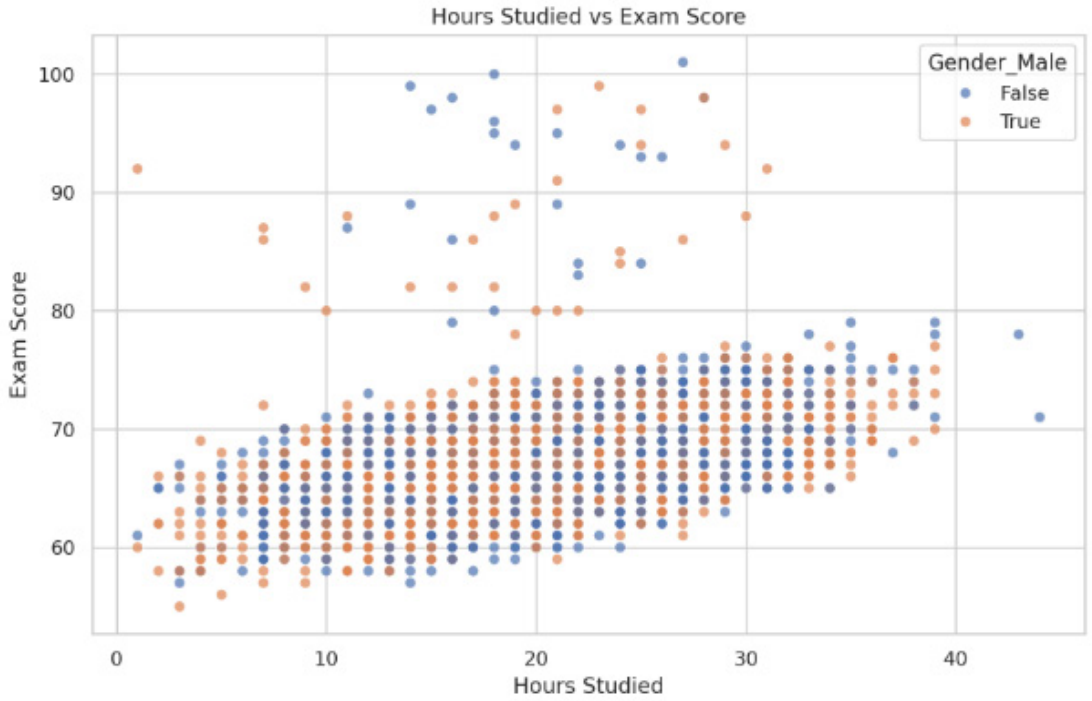

Figure 2. Hours studied and the exam scores

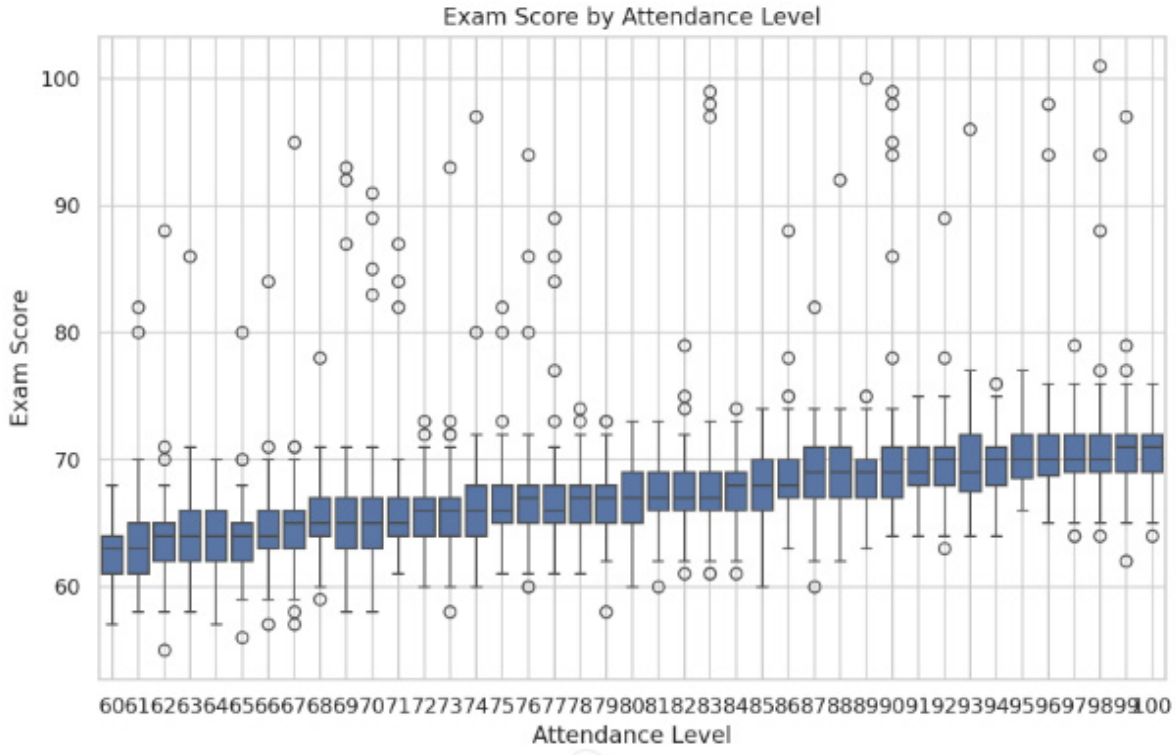

Figure 3. Exam Score by Attendance Level

RESULTS AND DISCUSSION

The study of predicting and analyzing student academic performance has become a crucial area of focus, leveraging data analytics and machine learning to transform educational outcomes. As education plays a pivotal role in societal progress, researchers have sought to understand the factors that influence academic success and develop predictive models to enable timely interventions. These advanced technologies empower educators and institutions to offer tailored support, enrich learning experiences, and allocate resources effectively. Central to this research is identifying the key drivers of academic performance. Factors such as consistent attendance, active engagement in discussions, and punctuality in submitting assignments are recognized as critical indicators of student success. Combining behavioral data with traditional academic records provides a holistic understanding of trends, enabling more precise analysis. This innovative approach promises a more inclusive and effective educational future.

CONCLUSIONS

The paper concludes that the integration of intelligent technologies in education significantly enhances the analysis and prediction of student performance. This advancement is essential for improving educational outcomes and providing timely support to student. The research emphasizes the limitations of traditional data processing methods in managing the rapid growth of educational data. It points out the necessity for more sophisticated approaches to avoid unreasonable evaluation results and to better monitor students' future performance The analysis highlights the importance of study habits and attendance in influencing student performance. While hours studied and attendance are significant factors, the variability in scores suggests that other elements, such as motivation, parental involvement, and access to resources, may also impact academic outcomes. Further investigation into these additional factors could provide a more holistic understanding of student performance. The conclusions suggest that further research is needed to refine these methodologies and explore additional intelligent technologies that can enhance educational data mining. This ongoing research is crucial for continuously improving the educational landscape and student outcomes

BILBIOGRAPHIC REFERENCES

1. Feng G, Fan M, Chen Y. Analysis and prediction of students' academic performance based on educational data mining. IEEE Access. 2022;10:19558–71.

2. Oyedeji AO, Salami AM, Folorunsho O, Abolade OR. Analysis and prediction of student academic performance using machine learning. JITCE (Journal Inf Technol Comput Eng. 2020;4(01):10–5.

3. Ahmad F, Ismail NH, Aziz AA. The prediction of students' academic performance using classification data mining techniques. Appl Math Sci. 2015;9(129):6415–26.

4. Tomasevic N, Gvozdenovic N, Vranes S. An overview and comparison of supervised data mining techniques for student exam performance prediction. Comput Educ. 2020;143:103676.

5. DeBerard MS, Spielmans GI, Julka DL. Predictors of academic achievement and retention among college freshmen: A longitudinal study. Coll Stud J. 2004;38(1):66–81.

6. El-Moussa OJ. The Challenge of Predicting College Success in Male and Female Saudi Arabian Students: An Examination of Whether High School GPA, Qudrat, and Tahsili Are Effective Predictors. The University of North Dakota; 2023.

7. Hussain AM ud DM, Gillani MASA. Association between the use of active learning strategies and classroom engagement among nursing students. J Heal Med Nurs. 2019;62(8):59–65.

8. Kumar D, Haque A, Mishra K, Islam F, Mishra BK, Ahmad S. Exploring the Transformative Role of Artificial Intelligence and Metaverse in Education: A Comprehensive Review. Metaverse Basic Appl Res. 2023;2:55.

9. Prakash A, Haque A, Islam F, Sonal D. Exploring the Potential of Metaverse for Higher Education: Opportunities, Challenges, and Implications. Metaverse Basic Appl Res [Internet]. 2023 Apr 26;2(SE-Reviews):40. Available from: https://mr.saludcyt.ar/index.php/mr/article/view/40

10. Alimul Haque * , Devanshu Kumar , Khushboo Mishra , Farheen Islam BKM. The Factor Affecting Student's Performance of E-Learning Environment Using Machine Learning Algorithm. In: MOL2NET'22, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 8th ed congress USEDAT-08: USA-Europe Data

Analysis Training Congress, Cambridge, UK-Bilbao, Basque Country-Miami, USA, 2022 [Internet]. sciforum,MDPI; 2022. Available from: https://sciforum.net/paper/view/12682

11. Alyoussef IY. E-Learning Acceptance: The Role of Task–Technology Fit as Sustainability in Higher Education. Sustainability. 2021;13(11):6450.

12. Zeba S, Haque MA, Alhazmi S, Haque S. Advanced Topics in Machine Learning. Mach Learn Methods Eng Appl Dev. 2022;197.

13. Student-performance-factors [Internet]. Available from: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

FUNDING

None.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHORSHIP CONTRIBUTION

Conceptualization: Priya Darshini.
Investigation: Priya Darshini.
Methodology: Priya Darshini.
Writing - original draft: Priya Darshini.
Writing - review and editing: Priya Darshini.