www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1163-1168

e-ISSN : 2583-1062
Impact Factor : 7.001

# IMPROVING STUDENT PERFORMANCE PREDICTION WITH SOCIO-ECONOMIC AND BEHAVIORAL DATA INTEGRATION - A SYSTEMATIC REVIEW

**Shaikh Mohd Afshaan[1], Dakhve Humaid[2], Hawaldar Ziya[3], Prof. Zaibunnisa L. H. Malik[4]**

[1,2,3]Department of Computer Engineering M.H. Saboo Siddik Polytechnic Mumbai, India.

mohammed.afshaan06@gmail.com

humaiddakhve@gmail.com

ziyahawaldar@gmail.com

[4]Guide, HOD Department of Computer Engineering M.H. Saboo Siddik Polytechnic Mumbai, India.

zebamalik@yahoo.com

DOI: https://www.doi.org/10.58257/IJPREMS36909

## ABSTRACT

This systematic review investigates the current land- scape of student performance prediction by integrating socio-economic and behavioral data with machine learning techniques. The review examines advancements, common datasets, method- ologies, and limitations in predictive models across various educa- tional settings. Ten selected papers provide insight into how socio- economic and behavioral data enhance the predictive accuracy of models, particularly using machine learning algorithms like Random Forest and Support Vector Machines (SVM). Future research directions include expanding datasets and incorporating real-time behavioral data for more robust predictions.

**Key words-** component, formatting, style, styling, insert

## 1. INTRODUCTION

In recent years, predicting student performance has become a focal point in educational research, driven by the increasing availability of large datasets from educational systems. Uni- versities and educational institutions are continuously seek-ing methods to improve student outcomes, reduce dropout rates, and provide timely interventions for at-risk students. Traditionally, models for predicting student performance were based largely on academic data, such as exam scores, GPA, or assignment grades. However, these models often overlook the complex socio-economic and behavioral factors that influence student success. Socio-economic data such as family income, parental education level, and financial aid status, combined with behavioral data like attendance patterns, online course engagement, and participation in class discussions, offer a more comprehensive view of a student's academic journey.

The integration of socio-economic and behavioral data into machine learning models has shown significant promise in enhancing predictive accuracy. This shift towards a more holistic approach is especially important for identifying at-risk students who may face challenges unrelated to their academic abilities. For instance, studies have demonstrated that students from lower socio-economic backgrounds are often more likely to face academic difficulties due to external pressures such as financial instability or lack of educational resources [1][5]. Similarly, students who display low levels of engagement in learning management systems (LMS) or have irregular attendance are more likely to perform poorly, regardless of their prior academic achievements [3][9].

In this systematic review, we focus on the impact of incorporating socio-economic and behavioral data into ma- chine learning models for student performance prediction. We explore the various machine learning algorithms used, such as Random Forest, Support Vector Machines (SVM), and Neural Networks, and examine their effectiveness in predict- ing student performance. Additionally, this review addresses the challenges and limitations of using non-academic data, including privacy concerns and data integration difficulties, which can affect the generalizability and ethical implications of these models.

Several recent studies have employed machine learning techniques to improve student performance prediction by integrating socio-economic and behavioral data. For example, the use of Random Forest and ensemble learning techniques has been shown to provide higher accuracy in predicting at-risk students compared to traditional models based on academic performance alone [2][6][9]. In many cases, the inclusion of socio-economic factors such as family background has proven to be a critical predictor of student success, highlighting the need for more inclusive predictive models that go

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1163-1168
e-ISSN : 2583-1062
Impact Factor : 7.001

beyond academic metrics [3][10].

Moreover, the incorporation of behavioral data, particularly

in online learning environments, offers early indicators of potential academic failure. By analyzing LMS logs, student interactions, and participation in discussion forums, predictive models can identify disengagement at an early stage and provide timely interventions [4][9]. This approach is especially relevant in the post-pandemic educational landscape, where online and blended learning environments have become more prevalent, requiring educators to adapt to new forms of student engagement monitoring [8][10].

The need for early interventions has never been more critical, with studies showing that students identified as at- risk early in their academic careers are more likely to improve their performance with targeted support [3][7]. As educational institutions increasingly embrace data-driven decision-making, integrating diverse data sources—both academic and non-academic—will be essential to developing more accurate, fair, and actionable predictive models. This review synthesizes the current research in this area, highlighting the potential for socio-economic and behavioral data integration to revolution- ize student performance prediction models.

## 2. RESEARCH METHODS

### A. Research Questions

In this systematic review, the focus is on addressing the following research questions:

1) **What machine learning models are most effective** in predicting student performance when socio-economic and behavioral data are integrated? This question ex- amines the performance of different machine learn- ing models—ranging from traditional classifiers like Support Vector Machines (SVM) to ensemble learning techniques like Random Forest and newer approaches such as attention-based models. Understanding which algorithms are best suited to leverage non-academic data sources will highlight the most promising approaches for future work.
2) **How do different types of data sources** (e.g., Learning Management System (LMS) logs, socio-economic data, demographic information) influence the accuracy and ro- bustness of student performance prediction models? This question seeks to understand the relative contribution of various data types and the benefits of integrating them with academic performance data.
3) **What are the main limitations and challenges** in using socio-economic and behavioral data for student performance prediction? The inclusion of non-academic data can introduce challenges related to data quality, privacy concerns, model interpretability, and ethical con- siderations. This review seeks to identify these obstacles and explore potential solutions.

### B. Search Strategy

A comprehensive search strategy was designed to iden- tify the most relevant studies for this review. A variety of academic databases, including IEEE Xplore, ScienceDirect, and SpringerLink, were utilized to ensure a broad scope of research materials. The search terms used included: ”stu- dent performance prediction,” ”machine learning,” ”socio-economic data,” ”behavioral data,” ”educational data mining,” and ”learning analytics.” These terms were used both individ- ually and in combination to capture relevant literature. The timeframe for the review was limited to studies pub-lished between 2022 and 2024, focusing on recent advance- ments in the field. This allowed for an up-to-date under-standing of the current methodologies and technologies being implemented in educational data mining. Only peer-reviewed journal articles and conference papers were considered, ensur- ing a focus on high-quality empirical research.

To ensure relevance, an initial screening was performed based on the title and abstract of each paper. Studies that focused solely on academic data, without any integration of socio-economic or behavioral factors, were excluded. Arti- cles that presented theoretical frameworks without empirical validation were also excluded. The remaining papers were subjected to a full-text review to confirm their inclusion in this systematic review.

### C. Study Selection Criteria

The selection criteria for this systematic review were care- fully designed to ensure that the studies included were both relevant and of high quality. The criteria were as follows:

1) **Inclusion of machine learning models**: Only studies that applied machine learning models to predict student performance were included. These models could range from simple classifiers like Logistic Regression to more advanced models like Neural Networks and ensemble methods like Random Forest.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 04, Issue 11, November 2024, pp : 1163-1168
e-ISSN : 2583-1062
Impact Factor : 7.001

2) **Incorporation of socio-economic and behavioral data**: The key focus of this review was on the integration of non-academic data, such as socio-economic factors (e.g., family income, parental education) and behavioral data (e.g., student engagement, LMS activity). Studies that focused purely on academic data were excluded.
3) **Empirical studies**: Only studies that presented actual case studies or empirical research in educational settings were considered. This ensured that the findings were based on real-world data and not purely theoretical models.
4) **Publication date**: Only studies published between 2022 and 2024 were included to capture the most recent advancements in the field.
5) **Study relevance**: Studies were selected based on their relevance to student performance prediction, with a focus on those that incorporated diverse data sources to improve the prediction of at-risk students.

Ultimately, ten studies met these criteria and were included in the final analysis for this systematic review.

## 3. RESULTS

### A. Common Datasets Used

Across the ten studies, the use of **Learning Management System (LMS)** logs and **Student Information Systems (SIS)** data was ubiquitous. These datasets typically include stu- dent demographics, academic performance records (e.g., GPA, exam scores), and behavioral data (e.g., attendance, participa- tion in online discussions). In some cases, additional socio-economic data, such as family income, parental education level, and scholarship status, were integrated into the models to enhance predictive accuracy [2][5][9]. For example, the Open University Learning Analytics Dataset (OULAD) was frequently cited as a rich source of data, with studies leveraging its detailed LMS logs to analyze student interaction patterns and predict academic outcomes [1][9]. Other institutional datasets, such as those from Purdue University, also played a crucial role in providing longitudi- nal data for tracking student performance over time [6][7]. Additionally, some studies relied on smaller datasets from individual courses or universities, which provided insight into specific contexts but limited generalizability [8][10].

The types of data collected can be categorized as follows:

- **Academic data**: GPA, exam scores, assignment grades.
- **Behavioral data**: Attendance records, online activity logs, submission times for assignments.
- **Socio-economic data**: Parental education, family income, geographical location, access to technology.
- **Engagement data**: Participation in discussion forums, clicks on LMS content, time spent on learning activities.

### B. Machine Learning Models

Several machine learning models were employed in the reviewed studies, with varying degrees of success. The most commonly used algorithms included:

- **Random Forest (RF)**: This model was highlighted in several studies for its ability to handle large datasets with high-dimensional features. Random Forest was par- ticularly effective in predicting at-risk students, with one study achieving an accuracy of **98.4%** when using both academic and socio-economic data [2][9].
- **Support Vector Machines (SVM)**: SVM was frequently used for its robustness in classification tasks. While it per-formed well in predicting academic performance, studies showed that its effectiveness improved when behavioral and socio-economic data were integrated [3][6].
- **Neural Networks (NN)**: Some studies applied Neural Networks to capture complex, non-linear relationships between different data sources. However, the performance of Neural Networks varied, with some models struggling to outperform simpler models like Random Forest in certain contexts [7][10].
- **Ensemble learning techniques**: A hybrid approach, combining multiple machine learning models, was also explored. These ensemble methods often outperformed individual models by leveraging the strengths of different algorithms, particularly when predicting performance in large-scale datasets [4][8].

Overall, the integration of socio-economic and behavioral data significantly boosted the accuracy of these models. The best- performing models were able to identify at-risk students early, allowing institutions to provide timely interventions [2][6][9].

### C. Impact of Socio-Economic and Behavioral Data

The inclusion of socio-economic and behavioral data had a profound impact on the accuracy of predictive models.

Behavioral data, such as student participation in LMS activities, provided early warning signs of disengagement, which were strong indicators of poor academic performance. For instance, students with low levels of participation in discussion forums or who submitted assignments late were more likely to struggle academically [3][6]. Similarly, **socio- economic factors** like family income and parental education level were found to be critical predictors of student success. Students from lower socio-economic backgrounds often faced additional challenges, such as limited access to learning re- sources, which affected their academic outcomes [5][9][10].

One study found that integrating socio-economic data im- proved the model's ability to predict at-risk students by **20-30%** compared to models that relied solely on academic data [9]. Behavioral data, particularly from online learning platforms, provided a more **real-time analysis** of student engagement, allowing educators to intervene before students fell too far behind [3][9].

### D. Challenges in Data Integration

Despite the benefits, there were several challenges asso- ciated with integrating socio-economic and behavioral data into predictive models. One of the most significant issues was **data privacy**. Collecting sensitive socio-economic data, such as family income or parental education, raised ethical concerns about how this information would be used and protected [4][9][10].

Additionally, many studies faced difficulties with **data di- versity**. Most of the datasets used in the reviewed studies came from **single institutions**, which limited the generalizability of the findings. This reliance on small, localized datasets meant that models trained on these data were less applicable to broader educational contexts [1][6].

Another challenge was the **lack of real-time data integra- tion**. While many studies successfully used historical data to predict student outcomes, the absence of real-time behavioral data limited the ability to provide timely interventions [1][9]. Real-time data could significantly enhance predictive models by allowing institutions to react to changes in student behavior as they happen, rather than retrospectively [7].

## 4. LIMITATIONS AND FUTURE WORK

While integrating socio-economic and behavioral data into student performance prediction models offers significant im- provements in predictive accuracy, several limitations remain. One of the key challenges is the **availability and quality of data**. Many studies rely on limited datasets from specific institutions, which restricts the generalizability of the models to other educational contexts. For instance, most datasets used in the reviewed studies were derived from Learning

Management Systems (LMS) or Student Information Systems (SIS) at single institutions, which can create biases and reduce the applicability of the models across diverse educational environments [2][6]. The **sample size** of some studies is also a concern, as small datasets may not capture the full spectrum of student behaviors and socio-economic factors.

Another significant limitation is the **lack of real-time data integration**. While many models provide accurate predictions based on historical data, they do not adapt to students' evolving behaviors or changing socio-economic conditions during the course of their studies. This limits the practical application of the models in dynamic learning environments, where real-time interventions could prevent students from falling behind [1][7]. For example, a student's behavior may change drastically during a semester due to personal or finan- cial difficulties, yet many models are unable to adjust their predictions accordingly. The inclusion of **real-time behavioral data** could offer more timely and effective interventions, but this requires more advanced data collection and processing capabilities that are currently lacking in many educational systems.

**Ethical and privacy concerns** are another major limitation in using socio-economic and behavioral data for predictive modeling. The collection and use of sensitive socio-economic data, such as family income, parental education, or mental health status, raise questions about data privacy and student consent [4][10]. Although predictive models offer valuable insights, they must be implemented in a way that ensures the **confidentiality of personal information** and adheres to legal regulations such as GDPR. Moreover, the potential for **bias** in predictions is a concern, particularly when models disproportionately affect students from lower socio-economic backgrounds. If not carefully managed, predictive models could reinforce existing inequalities in educational outcomes [9][10].

Finally, **model interpret ability** remains a challenge. Many machine learning algorithms, especially complex ones like Neural Networks, function as "black boxes," making it difficult for educators and policymakers to understand how specific predictions are made. This lack of transparency can reduce trust in the models and limit their adoption in educational

settings. Future research should focus on developing more **interpret-able models**, such as decision trees or explainable AI techniques, to ensure that educators can confidently act on the predictions [5][8].

**Future research** should prioritize:

- Expanding datasets to include a broader range of institu- tions, allowing for better generalization of results.
- Developing models that integrate real-time behavioral data for more responsive interventions.
- Addressing ethical concerns by ensuring transparency, student consent, and privacy protections in the use of socio-economic data.
- Improving model interpret-ability to enhance the trust and usability of predictive models for educators and decision-makers [1][4][9][10].

## 5. CONCLUSION

This review highlights the growing importance of in- tegrating socio-economic and behavioral data into stu- dent performance prediction models. While traditional models based solely on academic performance have their limitations, the inclusion of non-academic factors provides a more holistic view of a student's potential for success. By leveraging machine learning techniques such as Random Forest and Support Vector Machines, studies have demonstrated that predictive accuracy can be significantly improved, particularly when identifying at-risk students early in their academic journeys [2][6][9]. This allows educational institutions to intervene more effectively, providing targeted support and resources to those who need them most. However, significant challenges remain. The lack of data diversity, limited real-time integration, and eth- ical concerns surrounding the use of sensitive socio- economic data must be addressed before these models can be widely adopted. Additionally, the development of more interpret-able models is necessary to ensure that educators and decision-makers can confidently use these tools to improve student outcomes [5][9].

Future work should focus on expanding datasets, im- proving the real-time applicability of predictive models, and addressing privacy and fairness concerns in data collection and use. As educational institutions continue to adopt data-driven decision-making practices, the integra- tion of socio-economic and behavioral data will become increasingly critical for building more accurate, equitable, and actionable predictive models. By doing so, we can better support students from all backgrounds, ensuring they receive the resources and interventions necessary to succeed in their academic pursuits [1][5][10].

## 6. REFERENCES

[1] Priyambada, S. A., Usagawa, T., & Mahendrawathi, E. R. (2023). Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge.

[2] Santos, R. M., & Henriques, R. (2023). Accurate, timely, and portable: Course-agnostic early prediction of student performance from LMS logs.

[3] Chu, Y.-W., Hosseinalipour, S., Tenorio, E., Cruz, L., Douglas, K., Lan, A., & Brinton, C. (2022). Mitigating Biases in Student Per- formance Prediction via Attention-Based Personalized Federated Learning.

[4] Siafis, V., & Rangoussi, M. (2022). Educational Data Mining- based visualization and early prediction of student performance: a synergistic approach.

[5] Lu¨nich, M., & Keller, B. (2024). Explainable Artificial Intelli- gence for Academic Performance Prediction: An Experimental Study on Decision Trees and Fairness Perceptions.

[6] Jiao, P., Ouyang, F., Zhang, Q., & Alavi, A. H. (2022). Artifi- cial Intelligence-Enabled Prediction Model of Student Academic Performance in Online Engineering Education.

[7] Khairy, D., Alharbi, N., Amasha, M. A., Areed, M. F., Alkhalaf, S., & Abougalala, R. A. (2024). Prediction of Student Exam Performance Using Data Mining Classification Algorithms.

[8] Alalawi, K., Athauda, R., Chiong, R., & Renner, I. (2024). Eval- uating the student performance prediction and action framework through a learning analytics intervention study.

[9] Pek, R. Z., O¨ zyer, S. T., Elhage, T., O¨ zyer, T., & Alhajj, R. (2023). The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure.

[10] Pelima, L. R., Sukmana, Y., & Rosmansyah, Y. (2024). Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review.

[11] Zhen, Y., Luo, J., Chi, H. (2023). Prediction of Academic Performance of Students in Online Live Classroom Interactions An Analysis Using Natural Language Processing and Machine Learning. IEEE Xplore. March 23.

[12] Abdulrakeem, D., Shafiq, M., Mahmoud, M. (2023). Student Re- tention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review. IEEE Xplore. June 15.

[13] Alshannaq, A. (2023). Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification. IEEE Xplore. June 02.

[14] Reima, L. P., Muflih, N., Altamimi, L. (2024). Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review. IEEE Xplore. February 06.

[15] Nawanga, H., Chuan, Y. (2023). A Systematic Literature Review on Student Performance Prediction Research. IEEE Xplore. June 06.

[16] Sheehamer, A., Muflih, N., Reima, L. P. (2024). Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review. IEEE Xplore. February 02.

[17] Flanagan, B., Cheng, G. (2022). Early-warning Prediction of Student Performance and Engagement in Open Book Assessment by Reading Behavior Analysis. SpringerLink. August 19.

[18] Yui, Y. F., Zhen, Y., Luo, J. (2022). Long-Term Student Per- formance Prediction Using Learning Stability and Self-Adaptive Algorithm. SpringerLink. July 19.

[19] Khairy, D., Adebisi, O. (2024). Prediction of Student Exam Per- formance Using Data Mining Classification Algorithms. Springer- Link. March 06.

[20] Huang, Q., Liu, X. (2022). Improving Academic Performance Predictions with Dual Graph Neural Networks. SpringerLink. August 26.

[21] Duyang, F., Song, Y. (2023). Artificial Intelligence-Enabled Prediction Model of Student Academic Performance in Online Engineering Education. SpringerLink. August 21.

[22] Jiao, P., Wang, T., Zhang, Y. (2023). Integration of Artifi- cial Intelligence Performance Prediction and Learning Analytics to Improve Student Learning in Online Engineering Course. SpringerLink. January 21.

[23] Bilal, M., Hafeez, S. (2023). The Role of Demographic and Aca- demic Features in Student Performance Prediction. SpringerLink. July 16.

[24] Santos, R. M., Ferreira, A. (2023). Accurate, Timely, and Portable: Course-agnostic Early Prediction of Student Perfor- mance from LMS Logs. ScienceDirect. April 28.

[25] Purnayada, S. A., Brawijaya, I. A. (2023). Two-layer Ensemble Prediction of Students' Performance Using Learning Behavior and Domain Knowledge. ScienceDirect. December 16.

[26] Touka, K., Salah, H. (2023). Prediction of Student Perfor- mance in Abacus-Based Calculation Using Matrix Factorization. PDF.ScienceDirect. July 14.

[27] Wang, Y., Chu, L. (2023). Mitigating Biases in Student Per- formance Prediction via Attention-Based Personalized Federated Learning. ACM DL. October 17.

[28] Lurich, M., Mills, T. (2023). Explainable Artificial Intelligence for Academic Performance Prediction. ACM DL. June 03.

[29] Lameiri, A., Yu, Y. (2023). Student Academic Success Prediction Using Learning Management Multimedia Data with Convoluted Features and Ensemble Model. ACM DL. October 12.

[30] Pawa, M., Dastan, F. (2023). Enhancing Student Perfor- mance Prediction through Feature Selection: Insights from 'Office' Assessment Data During the COVID-19 Pandemic. PDF.ScienceDirect. October 26.

[31] Chuan, Y., Nawang, H. (2023). Student Performance Predictions for Advanced Engineering Mathematics Course With New Mul- tivariate Copula Models. IEEE Xplore. February 06.

[32] Sneeamer, A., Reima, L. (2024). Early Predicting of Students Performance in Higher Education. IEEE Xplore. February 24.