Predicting Student Performance Based on Social and Academic Data using Data Mining

Janiah Suresh William, Dharshana Ramesh and Sriya Sanagala

Abstract: Predicting student performance using data mining and machine learning techniques is a critical area of research in educational data mining (EDM). This study utilizes academic, social, and behavioral data to develop predictive models that identify students at risk and propose efficient, early interventions. The machine learning algorithms employed were Decision Trees, Random Forests, SVMs, and Neural Networks for classifying student performance. The dataset consists of academic records, demographic attributes, and social engagement metrics, unifying the different aspects to represent the student's overall impression. Thus, it can reasonably be said that general methodology proposes all preprocessing operations such as data cleansing, PCA, and Chi-Square tests, feature selection, as well as model training using an 80:20 training-test split with hyperparameter tuning. Evaluation metrics such as accuracy, precision, recall, and F1 score were used to measure model success. Results demonstrate that ensemble models, in particular Random Forest and deep learning-based approaches, developed better classification in predicting student performance than naive-mechanism-based users. Besides, this study represents social and behavioral factors that influenced academic success, such as study habits, extra-curricular activities, and internet utilization. The results contribute to institutional academic development by enabling data-driven decision-making, personalized learning recommendations, and early intervention in helping students within the scope of performance. In this regard, future work will expand to datasets involving various institutions or even other relevant ones, improving deep learning models through their continued development, to provide real-time feedback through the creation of performance dashboards with continuous monitoring of student progress.

1. Introduction

The prediction of student performance represents a major field of focus in educational data mining, providing an opportunity to really enhance academic performance outputs and reduce dropout rates. Thus, educational institutions have been interested in the data mining techniques to provide insight into student performance, thereby enabling interventions and personalized learning approaches.

Many studies investigated the use of data mining techniques, especially classification methods such as Decision Trees, Naïve Bayes, and Support Vector Machines, because of their accuracy and simplicity, which classify students on the basis of their academic performance to do early identification of at-risk students. The much-acclaimed Random Forest and Artificial Neural Networks (ANN) methods had considerably increased precision after some feature selection processes, whereby unnecessary information was filtered out, thus strengthening the models' prediction success. The advancements in AI and Deep Learning like Convolutional Neural Networks and CNN-LSTM models also make the predictions more accurate in today's world.

Student characteristics such as their engagement in online learning become significant data in learning environments to analyse using CNN models because these models can capture nonlinear relationships. These may refer to any demandable parameter, including demographic traits, academic histories, behavioral patterns, and social interactions. Several studies have revealed that Naïve Bayes classified students' attention rates, assignment submission patterns, and participation level in onsite discussions as behavioral predictors for academic standing.

The hybrid model that combines clustering with regression has seen improved gains in GPA testing, where LassoCV outperformed correlation-based regression in most of the studies.Feature selection is key to better performance by decreasing complexity. Data mining tools such as WEKA, RapidMiner, and numerous Python libraries such as Scikit-Learn and TensorFlow are put into the framework for all these predictive models.

In this project, attempts will be made to analyze student performance using a few data mining algorithms and evaluate their indicative power on academic success. The contribution of the project in general will be the recognition of important features and predictive models for considering early intervention approaches, to which this project would contribute to students' success and dropout instances.

2. Literature Survey

Khan and Ghosh reviewed 140 studies of educational data mining (EDM) that focus on predicting student performance in classroom settings. They identified a variety of predictors and methodologies that are used with emphasis on the temporal aspect of predictions. Themeta-analysis found that practical and significant prediction accuracy is very well achievable during the course but predictions before course initiation would still require further efforts.

García et al explored the use of associative classification techniques to predict student academic performance.These technique employs the CBA algorithm, which is Classification Based on Associations, that integrates association rule mining with classification, and discovers linkage between student characteristics and academic performance. The results show that associative classification can discover patterns in educational data and provide educators with valuable insights to work on improving student performance.

Brijesh Kumar Baradwaj in his paper wrote about data mining techniques that are used for student performance analysis in higher education. The findings of the paper include predictions related to examination performance and the identification of at-risk students that would be based on findings from categorization techniques, chiefly the decision tree approach. The authors illustrated how Educational Data Mining (EDM) could identify levels needing better distinguishing teaching strategies and enhanced quality in education. The outlined model demonstrates how data mining has a tremendous potential in reforming such assessments in education to offer personalized advice to students.

Fan Yang in a research article, "SCB-Dataset: A Dataset for Detecting Student Classroom Behavior," introduces the SCB-Dataset5, specifically for aiding deep learning techniques in analysis of classroom behavior. The study attempts to provide a baseline object-detection algorithm based on the YOLOv7 series, aiming to combat scale variations and occlusions challenges in classroom settings. The SCB-Dataset5 will be of great value toward actually advancing automated systems that routinely monitor and evaluate the efficacy of teaching in an AI-based approach in educational settings.

Bunengi Henry Dagogo, in "The Student Behavior and Its Relationship on Academic Achievement: A Study of Nigeria High Schools," examines the influences of student behavior on academic performance in Nigerian high schools. A sample of 170 students, teachers, and counselors was drawn from the Obia/Akpor local government area in River State, and statistical analyses were performed to investigate the link. Positive correlational findings between study habits and academic performance were reported. It is recommended that students be assisted in developing good study habits to improve their academic performance.

Ahmed Kord's study explores how EDM and ML can be used to predict students' academic performance and recommend suitable courses. The responsiveness of 11 ML algorithms and 3 DL models concluded that SVC was the best-performing. The study also developed a recommendation system to allow students to select suitable courses in order to improve academic performance. The potential of predictive models and recommendation systems to enhance student experiences and facilitate success is highlighted.

M. Abdallah NamounBlogs-researched on the paper "Predicting Academic Performance Using Machine Learning Techniques: A Survey", which reviews the application of various machine learning techniques in predicting student academic performance. Discusses difficulties in predicting student outcomes along with data-driven approaches to enhancing educational processes. A wide range of machine learning algorithms have been surveyed including decision trees and neural networks, these being effective in predicting student's grades and identifying at-risk students. The paper also pleads for the importance of selecting relevant features from educational data to improve the accuracy of predictions.

Sarah A. Alwarthan researched different methods under machine learning. Some of the main features considered while modeling prediction are students' academic performance, their demographic information, and their previous qualification. Random Forest and ensemble models have proven to have the highest prediction accuracy; nevertheless, this review identified important gaps in the literature. The review ends with a discussion of some challenges faced by researchers that include issues with data quality as well as the need for further investigating influences among diverse student populations.

YOMNA M. I. HASSAN and ABEER ELKORANY conducted research on student modeling evaluate a student's knowledge, behavior, and performance by using various models to predict possible outcomes. The Community of Inquiry (CoI) framework enables one to select relevant attributes, with an emphasis on social, cognitive, and teaching presence. Time-series clustering has been used to group students according to their behaviors, which leads to an improved GPA prediction accuracy. . Thus, it is necessary to have optimized and generalizable models that better predict the academic performance of students with the least complexity in the data.

Saman Amjad and others conducted a study based on the application of data mining techniques to analyze the influence of social media on high school students' academic performances. Based on the use of machine learning classifiers including decision trees, random forests, and support vector machines, the study concludes that students who use social media all week tend to perform worse than those who use it only on weekends. The study shows that technology and social media usage patterns can significantly affect student outcomes, with random forests showing the highest performance in the prediction of these patterns.

Haiwei Chen and other researchers attempted to model urban air quality prediction using real-time sensor data with modern deep learning techniques. The researchers modeled air quality prediction through the LSTM networks by improving prediction accuracy with various environmental parameters. Through its capability of capturing complex temporal dependencies, the model has performed better than traditional methods, thus becoming proficient at forecasting air quality trends. Such approaches would further contribute to the next generation of smarter environmental monitoring systems.

Javier López-Zambrano et al conducted a study aiming to correlate EI with academic performance, and to show the relevant role of EI in the various settings for education purposes. Basically, the research presents EI as complementary to any learning and socialization that contribute to academic performance. An array of EI models is elaborated related to the child population and their performance, which emphasizes the necessity for all educational systems to embed the EI principles in their structuring in order to facilitate overall academic excellence.

EDM and LA were employed to predict the performance of students and narrow it down to specific areas for improvement. In his review of 82 studies on early prediction studies, López-Zambrano et al. (2021) showed the efficiency of EDM approaches. In another review, Batool et al. (2023) evaluated 260 studies in which ANN and RF formed the most popular algorithms considered, while academic records coupled with demographics were the most considered predictors in performance. These methods help identify at-risk students and carry timely interventions.

3. Dataset Description

In this current section, we present the description of the dataset on which our study has been performed. We have divided the description of the dataset into three sub-sections. In Section 3.1, we present the data collection information; Section 3.2 gives a description of preprocessing operations performed to clean and transform the data; and Section 3.3 briefly presents the most significant attributes of the dataset, with a description of their function in the study.

3.1. Dataset Overview

The data set used in the present study is the Student Performance & Behavior Dataset, which is authentic data collected from 5,000 students by a private educational organization. The data set comprises a combination of academic, social, and demographic attributes needed in student performance analysis and forecasting. It provides an in-depth description of students' performance, socioeconomic status, and behavioral trends. Therefore, it can be utilized in educational data mining.

3.2. Features of the Dataset

The data set contains features that can be categorized into three categories: academic features, social features, and demographic features. Academic variables include attendance rate, midterm and final exam marks, assignment averages, and project marks. These are the variables that provide feedback of a student's class participation and performance. Social variables such as participation in extracurricular activities, study group members, social network use, and availability of internet resources at home as sources of study materials are the variables that establish the impact of social behavior on academic performance. Demographics such as gender, age, parents' educational level, family income group, level of stress perceived, and sleeping hours also allow external sources of impact on students' performance to be cross-checked. A summary of the key dataset properties is presented in Table 1.

Table 1: Summary of Key Attributes in the Dataset

Attribute	**Description**
Student ID	Unique identifier for each student
Attendance (%)	Percentage of classes attended
Midterm Score	Midterm exam performance (out of 100)
Final Score	Final exam performance (out of 100)
Assignments Avg	Average assignment scores (out of 100)
Quizzes Avg	Average quiz scores
Projects Score	Evaluation of submitted projects
Total Score	Cumulative academic performance
Grade	Letter grade (A, B, C, D, F)
Study Hours per Week	Weekly study time in hours
Extracurricular Activities	Participation in clubs, sports, competitions
Internet Access at Home	Whether the student has home internet access
Parental Education Level	Highest qualification of parents
Family Income Level	Socioeconomic status (Low, Medium, High)
Stress Level (1-10)	Self-reported stress level

Sleep Hours per Night	Average nightly sleep duration

3.3. Importance of the Dataset

The dataset can be specifically employed to forecast students' performance through data mining and machine learning models. With the integration of the academic, behavioral, and demographic features, prediction models can be trained to label students and allocate them to their corresponding classes. Moreover, the dataset facilitates early identification of at-risk students, allowing for timely interventions and personalized learning strategies. Researchers can also utilize this data to examine how family history, economic status, and extracurricular activity influence students' overall academic path.

To render the dataset suitable for predictive modeling, several preprocessing steps must be performed. Some of these include handling missing values, scaling quantitative attributes such as study time and grades, and the selection of suitable attributes to improve model efficiency. Another consideration is handling potential imbalance in the data to avoid over-representation by certain categories such as socio-economic status or gender to bias model prediction. Given the real-world origin of the dataset, ethical aspects must also be considered, and foremost among them are data privacy and the use of student information responsibly.

The Students Grading Dataset offers a balanced basis on which to conduct research on the performance of students using both academic and social information. Through application of machine learning techniques on this dataset, researchers can build predictive models that help in academic performance prediction, intervention design, and evidence-based policy-making in the field of education. The feature set heterogeneity of this dataset makes it a strong candidate for testing the various factors contributing towards student success and thereby making educational decision-making and analytics more productive.

4. Proposed Method

We focus on the importance of predicting student performance through data mining techniques cannot be overstressed the effort put into ensuring early intervention and a student's path toward academic success. The methodology, proposed in this research, combines academic, social, and behavioral data to build an intelligent system capable of identifying at-risk students and providing some recommendations for self-improvement.

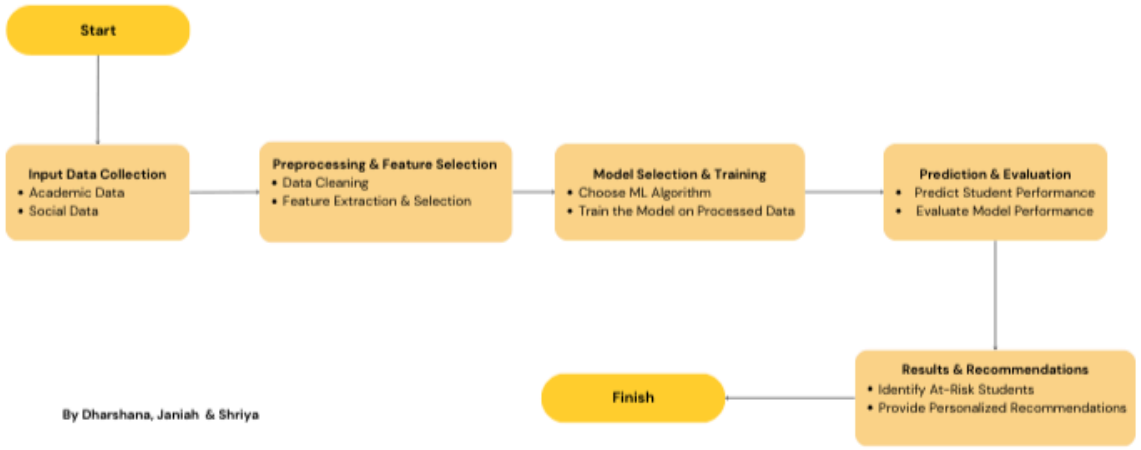
Through the cleansing and transformation methods, Step 2 will take care of the data inconsistencies and make the datasets ready for modeling. Data cleansing will deal with the missing values through either mean imputation or the regression-based method. Further, outlier detection and removal will be achieved through some statistical techniques, for such, the IQR method. For feature selection, PCA is acknowledged for redundant feature reduction, chi-square for categorical feature importance, and information gain for strongly influencing student performance predictions.

Step 3 deals with model selection and training using the cleaned dataset to predict student performance. Some of the selected models include decision trees (#interpretable rule-based predictions#), random forests (for handling diverse data types), and neural networks (complex relationships in large datasets). The training procedure will include splitting the dataset into training (80%) and testing (20%) sets, followed by hyperparameter tuning through algo/Gridsearches and k-fold cross-validation in order to ensure robustness.

In step 4, the model is trained and makes a prediction of student performance, which is rigorously evaluated. The predictions will classify the performance levels as low, average, or high; to identify concluding whether a student is at risk based on threshold values. Evaluation of the models is based on metrics of the following types: accuracy (how well the model classifies students), precision and recall (ensuring early detection of students struggling), and F1 score (a measure balancing false positives and false negatives).

Step 5 is results analysis and personalized recommendations, after the predictions have been made, and an extra set of insights is provided to the educators, students, and parents to promote their success.

Figure 1: Architecture Diagram



5. Conclusion and Future Work

By integrating data from academic indicators, social indicators, and behavioral indicators, this approach provides data-driven insights for the following goal-driven initiatives assisting institutions in improving student success rates. Future improvements will include these focuses: extending the dataset for multi-institutional analyses; development of deep learning models to examine more finely predictions; implementation of real-time student dashboards to allow for constant surveillance.

References

[1] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Educational Information Technology*, vol. 28, no. 1, pp. 905–971, 2023, doi: 10.1007/s10639-022-11152-y.

[2] S. A. Alwarthan, N. Aslam, and I. U. Khan, "Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, Article ID 8924028, 26 pages, 2022, doi: 10.1155/2022/8924028.

[3] A. Namoun and A. Alshanqiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review," *Appl. Sci.*, vol. 11, no. 1, p. 237, Dec. 2020, doi: 10.3390/app11010237.

[4] Khan, A., Ghosh, S.K. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Educ Inf Technol* 26, 205–240 (2021). https://doi.org/10.1007/s10639-020-10230-3.

[5] L. Cagliero, L. Canale, L. Farinetti, E. Baralis, and E. Venuto, "Predicting Student Academic Performance by Means of Associative Classification," *Appl. Sci.*, vol. 11, no. 4, p. 1420, 2021, doi: 10.3390/app11041420.

[6] Y. M. I. Hassan, A. Elkorany, and K. Wassif, "Utilizing Social Clustering-Based Regression Model for Predicting Students GPA," *IEEE Access*, vol. 10, pp. 12345–12356, 2022, doi: 10.1109/ACCESS.2022.3172438.

[7] S. Amjad, M. Younas, M. Anwar, Q. Shaheen, M. Shiraz, and A. Gani, "Data Mining Techniques to Analyze the Impact of Social Media on Academic Performance of High School Students," *Wireless Commun. Mobile Comput.*, vol. 2022, Article ID 9299115, 2022, doi: 10.1155/2022/9299115.

[8] N. Eleyan, M. Al Akasheh, E. F. Malik and O. Hujran, "Predicting Student Performance Using Educational Data Mining," *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Milan, Italy, 2022, pp. 1-7, doi: 10.1109/SNAMS58071.2022.10062500.

[9] J. López-Zambrano, J. A. Lara Torralbo, and C. Romero, "Early Prediction of Student Learning Performance Through Data Mining: A Systematic Review," *Psicothema*, vol. 33, no. 3, pp. 456-465, Aug. 2021, doi: 10.7334/psicothema2021.62.

[10] Alyahyan, E., Düştegör, D. “Predicting academic success in higher education: literature review and best practices”. *Int J Educ Technol High Educ* 17, 3 (2020). https://doi.org/10.1186/s41239-020-0177-7

[11] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K. U. Sarker, and M. U. Sattar, "Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques," *Appl. Sci.*, vol. 10, no. 11, p. 3894, Nov. 2020, doi: 10.3390/app10113894.

[12] A. I. Al-Alawi and N. M. A. Alsubaiee, "Predicting Student’s Academic Performance Using Data Mining Methods: Review Paper," *2023 International Conference On Cyber Management And Engineering (CyMaEn)*, Bangkok, Thailand, 2023, pp. 18-23, doi: 10.1109/CyMaEn57228.2023.10050962.

[13] Yağcı, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* 9, 11 (2022). https://doi.org/10.1186/s40561-022-00192-z

[14] Khairy, D., Alharbi, N., Amasha, M.A. *et al.* Prediction of student exam performance using data mining classification algorithms. *Educ Inf Technol* 29, 21621–21645 (2024). https://doi.org/10.1007/s10639-024-12619-w

[15] Y. Fan, "SCB-Dataset: A Dataset for Detecting Student Classroom Behavior," *arXiv*, vol. 2304.02488v3 [cs.CV], Nov. 2024, [Online]. Available: https://arxiv.org/abs/2304.02488.

[16] D. Thakur and N. Kapoor, "Predicting Student's Performance using Data Mining Algorithm," 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), Coimbatore, India, 2022, pp. 1-5, doi: 10.1109/ICACTA54488.2022.9753265.

[17] H. Chen, G. Zhou, and H. Jiang, "Student Behavior Detection in the Classroom Based on Improved YOLOv8," *Sensors*, vol. 23, no. 20, p. 8385, 2023, doi: 10.3390/s23208385.

[18] B. H. Dagogo, "The student behavior and its relationship on academic achievement: A Study of Nigeria High Schools," *Int. J. Soc. Sci. Humanit. Res.*, vol. 8, no. 4, pp. 93-107, Oct.-Dec. 2020. Available: www.researchpublish.com.

[19] B. Alnasyan, M. Basheri, and M. Alassafi, "The power of Deep Learning techniques for predicting student performance in Virtual Learning Environments: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100231, 2024. doi: 10.1016/j.caeai.2024.100231.

[20] Kord, A., Aboelfetouh, A. & Shohieb, S.M. Academic course planning recommendation and students' performance prediction multi-modal based on educational data mining techniques. *J Comput High Educ* (2025). https://doi.org/10.1007/s12528-024-09426-0