Data Mining in Educational Learning Platforms: Using Machine Learning Algorithms for Predicting Student Academic Performance

Caber, Jexter Jhon S.

Caparas, Allen Jay

IV - ACDS

December 2024

Introduction

In recent years, educational technology platforms such as Khan Academy have revolutionized the way students engage with learning materials (Gros and García-Peñalvo 2023). These platforms provide learners with easy access to a wide range of courses, allowing them to learn at their own pace. However, despite these benefits, student success on such platforms can be highly variable, influenced by factors such as engagement levels, subscription choices, and social interactions like referrals (Zhang et al. 2020). Understanding and knowing the factors that contribute to student success is critical for enhancing platform design, improving personalized recommendations, and fostering long-term educational outcomes.

Predictive modeling has emerged as a key solution to these challenges, providing data-driven insights to enhance student outcomes. Using machine learning algorithms, these models analyze features like usage patterns, engagement metrics, and subscription behaviors to predict student performance. By identifying trends and segmenting students based on their likelihood of success, predictive modeling helps institutions proactively offer tailored support to those who need it most.

This study will focus on developing predictive models for knowing student success on an educational technology platform. Using a simulated dataset that reflects realistic student interactions with the platform—including metrics such as engagement levels, course completion, subscription types, and referral activities—this research aims to identify the most important factors influencing success and to create a model capable of making accurate predictions about student academic performance..

The key research questions guiding this study are:

1. Which features are the strongest characteristics of student success on the platform?
2. How do these metrics impact student outcomes?
3. Can a predictive model be developed to accurately group students based on these factors?

The objectives of this research are as follows:

- To analyze the simulated dataset for patterns and trends in student behavior and academic performance success.
- To identify the most significant predictors of student success through exploratory data analysis (EDA) and feature engineering.
- To develop and evaluate a machine learning model that can accurately predict students based on the available data.

This research will contribute to the growing body of knowledge on predictive analytics in education. By using simulated data that mimics real-world usage of educational platforms, the study will provide insights into how students interact with online learning tools and how those interactions can be leveraged to improve learning outcomes. The predictive model developed in this study has the potential to assist educators, platform designers, and administrators in making data-informed decisions to support student success.

Methods

The research design for this study will follow a quantitative exploratory approach, utilizing secondary data analysis of a simulated dataset to build a predictive model of student success on an educational technology platform. The use of machine learning models in educational data mining has proven to be an effective method for uncovering hidden patterns that can help improve educational services and personalize learning experiences (Romero and Ventura 2020).

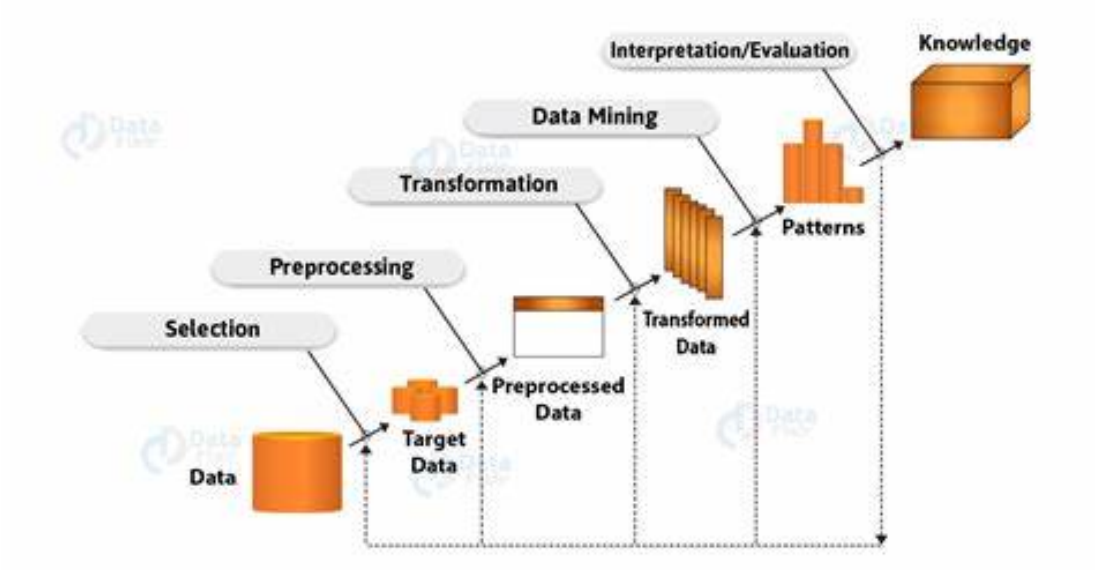

Figure 1. Knowledge Discovery in Databases Process (KDD)

As shown in Figure 1, the study will adopt the Knowledge Discovery in Databases (KDD) framework to guide the data mining process and model development. The KDD process involves a series of steps including data selection, preprocessing, transformation, data mining, and interpretation. This structured approach is critical in transforming raw, unstructured student interaction data into meaningful insights about their behaviors and success on the platform. The KDD framework is well-suited for this

study because it provides a comprehensive methodology for handling large datasets, ensuring that important patterns and correlations, such as the relationships between engagement metrics and academic performance, are effectively discovered and analyzed.

Table 1. Dataset Features and Definitions

Feature	Definition
User_ID	Unique identifier for each student
Age_in_Months	Age of the student in months
Gender	Gender of the student (e.g., Male, Female, Other)
Location	Geographic location of the student
Grade	Current academic grade of the student (e.g., 8th Grade)
Logins_per_Month	Number of logins per month on the platform
Days_Completed_Activity	Number of days where the student completed learning activities
Excersices_Started	Total number of exercises started by the student
Total_Time_Spent_in_Minutes	Total time spent on the platform in minutes
Course_Name	Name of the course the student is enrolled in

Course_Category	Category of the course (e.g., Science, Programming)
Completion_Rate	Percentage of the course completed by the student
Average_Score	Average score achieved by the student in assessments
Course_Rating	Rating of the course by the student
Recommendation_Likelihood	Likelihood of the student recommending the course (1-5)
Excercises_Complete	Number of exercises completed by the student
Points_Earned	Total points earned by the student
Subscription_Tier	Type of subscription (e.g., Free, Basic, Premium)
Subscription_Cost	Cost of the subscription plan in USD
Subscription_Length_in_Months	Length of the subscription in months
Renewal_Status	Whether the subscription was renewed (Yes/No)
Tutoring	Whether the student used tutoring services (Yes/No)
Referrals	Number of referrals made by the student

Academic_Grade	Final academic grade achieved by the student (A-F)

Table 1 presents the features and definitions of the simulated dataset related to various student behaviors and engagement metrics on the educational technology platform. The features include variables such as age, gender, academic grade, number of logins per month, total time spent on the platform, and subscription type. Additionally, the dataset tracks learning outcomes such as course completion rate, average assessment scores, and the student's final academic grade. These features are simulated to reflect realistic patterns of student interaction and performance, providing a comprehensive overview of how different behaviors and engagement levels contribute to academic success on the platform.

Selection

The study will focus on the following key variables to build the predictive model:

- **Logins_per_Month**: Frequency of logins as a measure of student engagement.
- **Days_Completed_Activity**: Number of days in which the student completed any learning activity.
- **Exercises_Started**: Count of exercises the student began.
- **Exercises_Completed:** Number of exercises completed by the student.
- **Total_Time_Spent_in_Minutes**: Total time the student spent learning on the platform.
- **Points_Earned:** Total points earned by the student.
- **Completion_Rate:** Percentage of the course completed by the student.

- **Average_Score**: Average score achieved by the student in assessments.
- **Academic_Grade**: This will serve as the target variable, representing student success.

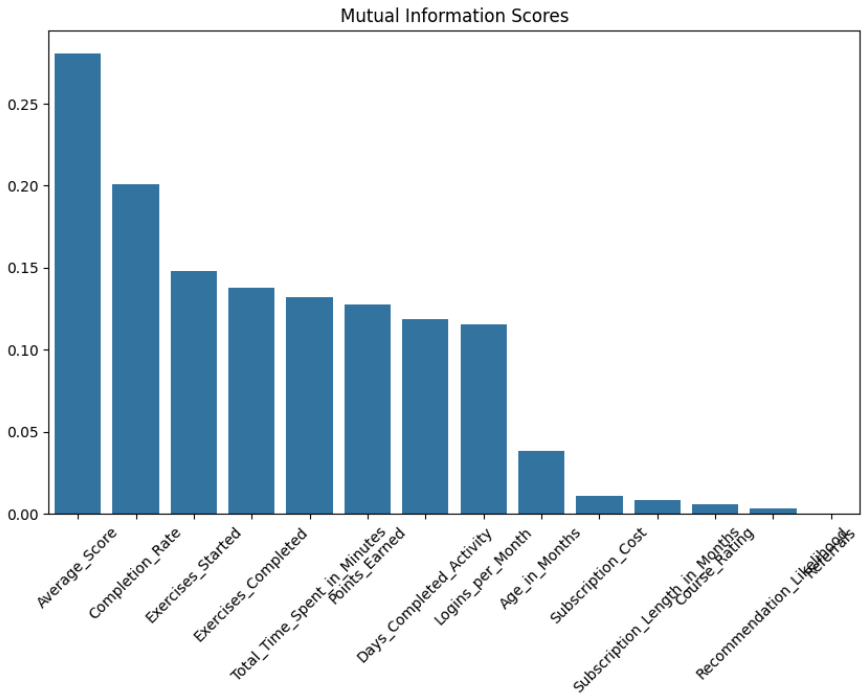

Figure 2. Mutual Information Score Bar Graph.

Using mutual information score, we were able to gather the features that would be useful for our target variable. Mutual information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable (Brownlee, 2020). It helps identify features that have a strong statistical relationship with the target variable, which can improve the performance of predictive models. The researchers picked out the features with more than a .10 score for the predictive model.

Preprocessing

The dataset will undergo several preprocessing steps:

Categorical Encoding: The Academic variable will be encoded using one-hot encoding, converting categories (B, C, D, F) into binary variables.

Standardization: We will identify and scale the numerical values namely, the Average_Score', 'Completion_Rate', 'Exercises_Started', 'Exercises_Completed', 'Total_Time_Spent_in_Minutes', 'Points_Earned', 'Days_Completed_Activity', 'Logins_per_Month' features using MinMaxScaler to ensure they are standardized, improving performance of the predictive model.

Data Mining

The study will use predictive models to predict students based on their statistical data. Specifically, Naive Bayes, SVM, Logistic Regression, KNN, Random Forest and Neural Network will be employed and compared to predict students based on feature variables. These predictions will help identify distinct student engagement patterns that correlate with success or failure.

Interpretation/Evaluation

To evaluate the performance of the prediction model and assess its ability to reveal meaningful insights about student engagement and success, several evaluation metrics will be employed:

Where:

- **TP (True Positives)**: Correctly predicted positive cases.
- **TN (True Negatives)**: Correctly predicted negative cases.
- **FP (False Positives)**: Incorrectly predicted as positive.
- **FN (False Negatives)**: Incorrectly predicted as negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 1. Accuracy Score

$$Precision = \frac{TP}{TP + FP}$$

Equation 2. Precision Score

$$Recall = \frac{TP}{TP + FN}$$

Equation 3. Recall Score

$$F1\ Score = 2 * (\frac{Precision * Recall}{Precision + Recall})$$

Equation 4. F1 Score

These are the metrics that will be used to gather the accuracy of a predictive model. Accuracy represents the number of correctly classified data instances over the total number of data instances. Precision, Recall and F1 Score are also used in interpretation (Harikrishnan, 2019).

$$TPR = \frac{TP}{TP + FN}$$

Equation 5. True positive rate (Recall)

$$FPR = \frac{FP}{FP + TN}$$

Equation 6. False positive rate

The AUC-ROC curve is used to evaluate the performance of a classification problem. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for different threshold values. The AUC represents the area under the ROC curve, which summarizes the performance of the classifier across all thresholds. The ideal value for AUC is 1 (Bobbitt, 2021).

Results

The results section presents the findings from the analysis and the effectiveness of the predictive models on the success on student's education utilization of educational platforms based on metrics such as AUC, Accuracy, F1, Precision and Recall.

Table 2. AUC, CA, F1, precision and recall values of the models

Model	(AUC)	Accuracy	F1	Precision	Recall
Naive Bayes	0.86	0.74	0.80	0.80	0.80
SVM	0.89	.078	0.82	0.82	0.83
Logistic Regression	0.91	0.78	0.82	0.83	0.84
KNN	0.86	0.74	0.79	0.79	0.80
Random Forest	0.89	0.77	0.82	0.81	0.82
Neural Network	0.91	0.79	0.83	0.81	0.85

The AUC value of NB, SVM, LR, KNN, RF, and NN algorithms were 0.86, 0.89, 0.91, 0.86, 0.89, and 0.91 respectively. The accuracy of the NB, SVM, LR, KNN, RF, and NN algorithms were also 0.74, 0.78, 0.78, 0.74, 0.77, and 0,79 respectively. According to these findings, for example, the NN algorithm was able to achieve 79% accuracy. In other words, there was a very high-level correlation between the data predicted and the actual data. As a result, 79% of the samples were classified correctly.

Implications for the Educational Platform

1. **Student Engagement Monitoring**
    - Models with high AUC and accuracy can identify patterns in features such as "Days Completed Activity" and "Total Time Spent in Minutes," which are part of the tutoring dataset.
2. **Personalized Learning Recommendations**
    - The NN and LR models' superior performance can support personalized interventions by predicting students who might benefit from tailored exercises or additional tutoring.
3. **Resource Allocation**
    - Insights from these models could guide educators or administrators in allocating resources effectively, focusing on students with higher dropout risks or low engagement.

These findings provide actionable insights to refine platform design and promote equitable success among diverse student populations.

Interpretation of Results

The predictive analysis provided valuable insights into student behaviors and their impact on academic success. By examining

1. **Performance Comparison of Models**
    - AUC Values: The Neural Network (NN) and Logistic Regression (LR) showed the highest AUC scores (0.91), indicating superior performance in distinguishing between classes. These models may be better suited for tasks requiring accurate differentiation between student performance levels.
    - Accuracy Scores: NN achieved the highest accuracy (79%), suggesting it correctly classified the highest proportion of samples among all models. This makes it a strong candidate for deployment in this context.
2. **Practical Implications for Educational Insights**
    - **NN's High Predictive Accuracy**: The NN's performance suggests a robust alignment between the predicted outcomes (e.g., student progress or success rates) and actual data. This reliability could be leveraged to provide actionable insights, such as predicting which students are at risk of underperforming.
    - **Correlation with Real Outcomes**: The high AUC values indicate the models' strong capability to handle complex patterns in the dataset. This is particularly useful for educational platforms where student behaviors are often nonlinear and influenced by multiple factors.

3. **Model Suitability for Different Use Cases**
    - **Logistic Regression**: Its competitive performance (AUC 0.91 and accuracy 78%) makes it a simpler alternative to NN for scenarios where interpretability is crucial (e.g., understanding which features contribute most to outcomes).
    - **Random Forest (RF) and Support Vector Machine (SVM)**: These models, with decent AUC and accuracy scores, could be considered for ensemble methods or as baseline models for comparison.

Discussion

Interpretation of Results

- **Focus on Interpretability**: While NN performed the best, it is often a black-box model. For an educational setting, stakeholders might require more interpretable models like LR or RF.

Limitations

- **Generalization to Broader Contexts**: Further validation on diverse datasets is necessary to ensure the model's applicability across different groups of students or educational platforms.

Future Work

- Validate findings using real-world datasets to improve generalizability.
- Incorporate additional features, such as socio-economic data or student feedback, for more holistic modeling.
- Explore other machine learning models to enhance accuracy and predictive power.

Conclusion

The evaluation of machine learning models on data from the educational tutoring website reveals that the Neural Network (NN) and Logistic Regression (LR) are the most effective algorithms, achieving the highest AUC scores (0.91) and commendable accuracy rates (79% and 78%, respectively). These results highlight their potential for accurately predicting student performance and engagement, making them valuable tools for identifying at-risk students and personalizing learning interventions. While NN offers the highest predictive power, its complexity may pose challenges in interpretability, suggesting that LR could serve as a simpler alternative for insights requiring transparency. Overall, these models demonstrate strong alignment with actual outcomes, indicating their potential to drive data-informed decision-making within educational platforms.

References

Bobbitt, Z. (2021, September 9). *What is Considered a Good AUC Score?* Statology. https://www.statology.org/what-is-a-good-auc-score/

Brownlee, J. (2020, December 10). *Information Gain and Mutual Information for Machine Learning - MachineLearningMastery.com*. Machine Learning Mastery. Retrieved December 9, 2024, from https://machinelearningmastery.com/information-gain-and-mutual-information/

Future Machine Learning. (2023). *Understanding Silhouette Score: A Key Metric for Clustering*. Future Machine Learning. Retrieved October, 2024, from https://futuremachinelearning.org/understanding-silhouette-score-a-key-metric-for-clustering/

Gros, B., & García-Peñalvo, F. J. (2023). *Future trends in the design strategies and technological affordances of E-Learning*. Retrieved October, 2024, from https://link.springer.com/referenceworkentry/10.1007/978-3-319-17461-7_67

Harikrishnan, N. B. (2019, December 11). *Confusion Matrix, Accuracy, Precision, Recall, F1 Score*. Medium. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

Jain, S. (2024, July 8). *Understanding the Confusion Matrix in Machine Learning*. GeeksforGeeks. Retrieved October 19, 2024, from https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

Permetrics. (2021). *Dunn Index (DI)*. Permetrics. Retrieved October, 2024, from https://permetrics.readthedocs.io/en/stable/pages/clustering/DI.html

Permetrics. (2021). *Purity Score (PuS)*. Permetrics. Retrieved October, 2024, from https://permetrics.readthedocs.io/en/stable/pages/clustering/PuS.html

Romero, C., & Ventura, S. (618). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601. Retrieved October, 2024, from https://ieeexplore.ieee.org/document/5524021/authors#authors

Sampaio, C. (2023, November 17). *Definitive Guide to K-Means Clustering with Scikit-Learn*. Stack Abuse. Retrieved October 19, 2024, from https://stackabuse.com/k-means-clustering-with-scikit-learn/

Zhang, Z., Cao, T., Shu, J., & Liu, H. (2020). Identifying key factors affecting college students' adoption of the e-learning system in mandatory blended learning environments. *Interactive Learning Environments*, *30*(8), 1388-1401. Retrieved October, 2024, from https://doi.org/10.1080/10494820.2020.1723113