SmartStudy

The Study Architects

Heraa Muqri Elorie Bernard-Lacroix Najma Sultani

Context (Problem)

Students often feel overwhelmed with work, deadlines, and extracurriculars. The stress of managing these responsibilities can make it difficult for students to establish effective study habits. In this proposal, we discuss the development of a machine learning model designed to recommend personalized lifestyle changes, such as the optimal weekly study time, the need for tutoring, and increased lecture attendance for students. Overall, this tool aims to increase organization and time-management among students and ultimately help them succeed academically.

Conditions of Success (ML task)

To measure the success of our model, we need to determine the increase in the quality of a student's studying and the utility of the model in helping a student improve their academic performance. A student's satisfaction with the model can be obtained through ratings on a feedback survey. The survey will collect the student's rating of the model performance based on how long it takes the model to recommend lifestyle changes such as the amount of weekly study time, tutoring, and class attendance rate. The usability of the model will also be measured based on how much time is spent using the model. The student's engagement with the model can also be examined through the time they spent using the platform with metrics such as cursor times and how often they use it. The difference in the student's grades before and after using the tool can also be calculated to determine the utility of the model in increasing a student's academic success.

Data

EdNet Dataset: The EdNet dataset contains various features of student actions such as which learning material they have consumed, their response, and how much time they have spent solving a given question or reading through an expert's commentary.

Students Performance Dataset: This dataset contains information on 2,392 high school students, detailing their demographics, study habits, parental involvement, extracurricular activities, and academic performance (target variable, GradeClass, more). This makes this dataset, a robust one for educational research, predictive modeling, and statistical analysis.

We will focus on the Students Performance Dataset, as it provides insights into how study habits (e.g studyTimeWeekly, extracurricular, …) correlate with academic performance. By identifying patterns in effective study behaviors, we will recommend changes to improve a student's academic outcome.

Example: Below is an example of datasets rows and columns that will be used in our model.

StudentID	Age	Gender	Ethnicity	ParentalEducation	StudyTimeWeekly	Absences	Tutoring	ParentalSupport	Extracurricular	Sports	Music	Volunteering	GPA	GradeClass
1001	17	1	0	2	19.833722807854700	7	1	2	0	0	1	0	2.929195591667680	2.0
1002	18	0	0	1	15.40875605584670	0	0	1	0	0	0	0	3.042914833436380	1.0

Inputs: Our input data will be the students' demographic details and study habits (including extracurriculars, tutoring, parental involvement, absence, study time and more as seen in picture above). This user input data can be treated as text to our model.

Outputs: "Study Time Recommendations" column suggesting an increase/decrease # of hours studying per week. "Behavioral Recommendations" column suggesting reducing absence, seeking tutoring or parental support, limiting or increasing extracurricular activities, volunteering, music sport activities to certain hours.

"Predicted Outcome" column contains a prediction of GPA improvement to certain points by doing changes.

Modeling

Proposal(s)

We will train a supervised learning model (trained on the inputs and outputs specified above) that can take information from users including their demographic details, extracurricular activities (level of sports, music, and volunteering), study habits (weekly study time, absences, and tutoring), parental involvement, and academic performance, and suggest changes to some of these categories (like study habits and extracurricular activities) to optimize their academic performance. In other words, rather than predicting a student's GPA, it is predicting how changing various factors that are in their control (like sports, music, volunteering, tutoring) will impact their GPA and which combination of changes is predicted to yield the optimal GPA.

Baseline Model

Many students can do their own research and manually look through studies to find trends in students' behaviors and their resulting academic performance. So, the most basic solution is a model that outputs similar generic advice. Users can then, apply it to their own life, and by trial and error, find a routine that suits them the best.

The addition of a simple supervised learning model can allow us to make more customized recommendations with the user's unique information. Therefore, instead of starting with very generic advice, they can quickly move forward with recommendations that have worked for others who are more similar to them.

State of the Art (SOTA)

Below are a few existing models that look at student information to predict academic performance.

- Student Grade Prediction: This model uses the Student Performance Dataset and maps demographic data, study habits, and extracurriculars to a particular grade. Although our final goal is slightly different, we could inspire ourselves from the methods used in this notebook since it works with the same dataset. There are also several other code examples associated with the Kaggle dataset.
- A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques : This provides a more general review of student performance models that exist, the data they used, and what models/model combinations they choose. We can refer to any of the ones outlined here to guide us in training our model.

Evaluation

We will track standard regression and metrics such as MAE to measure the difference between the predicted (improved grade) and the actual improved grade. Kendall Tau method will be used to compare the order of two ranks (generated by model vs info already in the table) [1], besides that we will also track the percentage of recommendation's output from our model within acceptable error threshold. R^2, or the coefficient of determination, can also be used to measure how well a model fits data, and how well it can predict future outcomes such as variance in GPA or study effectiveness [2]. The closer the R-squared value is to one, the better your model fits the data [2]. We will evaluate models on several splits: random (to ensure model evaluation is general), performance-based (to ensure model works for both high and low performers) and cluster (to ensure recommendations outputs for different types of students).

Additional Experiments / Stretch goals

- Web demo for tool
- Generate personalized study schedule PDF
- Integrate with Notion to incorporate tool into daily planning activities

References

[1]. Xebia, "Using Kendall's Tau to Compare Recommendations," *Xebia Blog*. Accessed Jan. 21, 2025. [Online]. Available: https://xebia.com/blog/using-kendalls-tau-to-compare-recommendations

[2] Arize AI, "R-Squared: Understanding the Coefficient of Determination," *Arize AI Blog*, Dec. 20, 2023. [Online]. Available: https://arize.com/blog-course/r-squared-understanding-the-coefficient-of-determination/. [Accessed: Jan. 30, 2025].