AN INTERPRETABLE LASSO REGRESSION MODEL FOR GPA PREDICTION USING SLEEP AND BEHAVIORAL FEATURES

by

Yuhyun Kim
A Capstone Project
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Bachelor of Science
Computational and Data Science

Committee:

_____ Dr. Kent Miller, Capstone Project Director

_____ Dr. First Last, Committee Member

_____ Dr. First Last, Committee Member

_____ Dr. First Last, Department Head

Date: _____ Spring Semester 2025
George Mason University
Fairfax, VA

An Interpretable Lasso Regression Model for GPA Prediction Using Sleep and Behavioral Features

A capstone project submitted in partial fulfillment of the requirements for the degree of Bachelor of Science at George Mason University

By

Yuhyun Kim
Bachelor of Science
,

Director: Dr. Kent Miller, Professor
Department of Computational and Data Science Department

Spring Semester 2025
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to ...

Acknowledgments

I would like to thank the following people who made this possible ...

Table of Contents

List of Tables

Table Page

List of Figures

Abstract

AN INTERPRETABLE LASSO REGRESSION MODEL FOR GPA PREDICTION USING SLEEP AND BEHAVIORAL FEATURES

Yuhyun Kim, BS

George Mason University, 2025

Capstone Project Director: Dr. Kent Miller

This capstone project explores the use of behavioral and demographic data to predict end-of-term academic performance, measured by term GPA, among first-year college students. Drawing on a publicly available dataset collected through wearable devices and university records, the study examines whether early-term sleep patterns—such as total sleep time, bedtime variability, and sleep midpoint—along with academic and demographic features, can serve as meaningful predictors of GPA outcomes.

Lasso regression was selected as the primary modeling technique due to its ability to perform variable selection and yield sparse, interpretable models. The model was trained using 10-fold cross-validation on 80% of the dataset and evaluated on the remaining 20%. The final model achieved an R^2 of 0.362 and an RMSE of 0.3936 on the test set. Among the selected features, cumulative GPA was the strongest predictor of term GPA, followed by variables such as total sleep time, daytime sleep, race, and the proportion of nights with valid sleep data.

The results indicate that interpretable regression models can offer both predictive accuracy and practical insights into academic risk factors. While behavioral features demonstrated weaker effects than academic history, their inclusion suggests that real-time data

from wearable sensors can enhance early intervention strategies. This study highlights the potential of human-centered predictive modeling in educational contexts and provides a foundation for future work integrating behavioral data into academic support systems.

Chapter 1: Introduction

Academic success during college has long been a subject of interest among educators, psychologists, and data scientists. While traditional predictors of academic performance—such as standardized test scores and prior GPA—remain useful, they often fail to capture dynamic behavioral and lifestyle factors that evolve over the course of a semester. In recent years, the increasing availability of sensor-based and behavioral data has enabled researchers to explore new predictors of student performance, including sleep duration, sleep timing, and daily activity patterns. These variables offer valuable insights into student well-being and study habits, both of which are known to impact academic achievement.

This project aims to build an interpretable regression model to predict end-of-term GPA (`term_gpa`) using early-term behavioral and demographic features. The analysis focuses specifically on passively collected sleep metrics—such as total sleep time, bedtime variability, and sleep midpoint—as well as demographic and academic variables like race, gender, first-generation college status, and prior GPA (`cum_gpa`). By leveraging these features, the project seeks to identify which factors are most predictive of academic performance and to what extent behavioral patterns early in the term can explain variation in GPA outcomes.

Lasso regression is employed as the primary modeling method due to its ability to produce sparse and interpretable models. Compared to black-box machine learning models, Lasso enables clear understanding of each predictor's contribution, which is essential for supporting educational interventions. The sparse structure of Lasso is especially useful in high-dimensional contexts, where multicollinearity or irrelevant features could reduce model clarity.

This work is grounded in prior research that demonstrates strong associations between sleep patterns and GPA. Studies have shown that students with longer and more consistent sleep tend to perform better academically. However, few studies have applied interpretable

machine learning methods to behavioral data collected in real time using wearable devices. This capstone project contributes to that growing literature by applying a human-centered modeling approach to a large, multi-university dataset of college students tracked over an academic semester.

The ultimate goal of this project is not only to achieve predictive accuracy, but also to generate actionable insights that can support early academic interventions. By identifying key behavioral signals that precede academic success or struggle, educators and administrators can design more timely and targeted support systems for students at risk. Through the use of transparent, sparse modeling, this project aims to bridge the gap between predictive power and real-world educational usability.

Chapter 2: Prior Research

2.1 Introduction

Educational researchers have long been committed to accurately forecasting student outcomes. Recently, with greater access to behavioral and physiological data, machine learning (ML) techniques have been used more frequently to improve prediction precision. Nonetheless, issues surrounding interpretability, fairness, and practical utility have sparked debate over the use of intricate black-box models in practical educational situations.,This chapter examines six peer-reviewed studies that advance the discipline of predicting academic achievement through interpretable machine learning models. These studies investigate various modeling techniques such as linear regression, explainable boosting methods, and model-agnostic approaches like SHAP. The research utilizes a broad array of data sources including self-reported surveys, behavior logs, smartphone-based sensing, and sleep records. The particular focus is placed on studies that employ sleep-related and passive sensing features, aligning with this capstone project's emphasis on behavioral variables collected from mobile platforms. The chapter also accentuates the importance of balancing prediction efficacy with stakeholder interpretability in educational ML.

2.2 Towards Human-Centered Early Prediction Models for Academic Performance in Real-World Contexts [1]

Research Question

The research investigates the potential of interpretable, human-centered machine learning models to forecast students' academic success early in the term. The authors stress the

importance of incorporating various evaluation standards, such as interpretability, fairness, robustness, and promptness, within predictive modeling structures.

Data and Methodology

The authors used self-report surveys and online behavioral data collected from students at a German university. They evaluated three models: linear regression (LR), 1D convolutional neural network (1D-CNN), and multi-task learning CNN (MTL-1D-CNN). The model evaluation included predictive accuracy, model faithfulness, and early prediction performance across learning timelines.

Key Findings

The linear regression model achieved lower accuracy but greater interpretability. Key predictors such as motivation, time management, and expected success were easy to interpret. CNN models showed higher performance, but poor transparency. The authors state that simple and interpretable models are useful for early interventions even if their predictive accuracy is moderate [1].

Relevance to GPA Prediction Model

This paper supports the project's use of interpretable models such as linear regression and Lasso. Their evaluation framework, including fairness and stakeholder understanding, serves as a useful guide for designing responsible educational prediction models.

Critique and Limitations

The study was conducted at a single institution, limiting its generalizability. Furthermore, self-reported data can introduce bias and CNN-based methods lacked interpretability tools beyond feature attribution.

2.3 Trustworthy Academic Risk Prediction with Explainable Boosting Machines [2]

Research Question

The authors investigate whether academic risk among university students can be accurately predicted using machine learning models that are not only performant but also interpretable, fair, and trustworthy. They emphasize the need for explainable AI tools in educational decision-making environments.

Data and Methodology

The study utilized behavioral log data collected from a German university's learning management system. Features included login frequency, assignment submission timing, quiz attempts, and forum engagement. The authors applied the Explainable Boosting Machine (EBM) and compared it to Random Forest, XGBoost, and logistic regression. Models were evaluated using predictive accuracy, fairness metrics, earliness of prediction, and stability.

Key Findings

The EBM performed comparably to more complex models in terms of accuracy while significantly outperforming them in interpretability. "EBMs offered insight into the effects of the characteristics through individual conditional expectation graphs, allowing detailed understanding of the stakeholders" [2]. The model effectively identified at-risk students early in the academic term and provided interpretable visualizations of the influence of characteristics.

Relevance to GPA Prediction Model

This study validates the use of inherently interpretable models such as EBM, which aligns with the project's goal of prediction centered on explanations. While this capstone project

uses regression-based models rather than boosting machines, the emphasis on stakeholder interpretability and early intervention aligns with the project's guiding principles.

Critique and Limitations

The study's dataset is constrained to digital interaction behavior within an online learning platform and may not capture offline factors that influence academic risk. Furthermore, the classification task (pass/fail) is structurally different from GPA regression, limiting direct model comparison. The paper also lacks an analysis of how instructors or advisors responded to the model's output in practice.

2.4 Prediction of Student's Academic Performance Using Linear Regression [3]

Research Question

This study examines whether linear regression can be used as a simple yet effective tool to predict university students' academic performance, specifically GPA. The authors aim to evaluate if basic academic features can provide meaningful and interpretable insights into student success without using complex algorithms.

Data and Methodology

The researchers used data from a public university in Nigeria, including high school GPA, entrance examination scores, and attendance rates. A multiple linear regression model was employed to predict students' cumulative GPA. The study focused on the strength of correlation between the predictors and academic outcomes, and the interpretability of regression coefficients.

Key Findings

High school GPA and entrance scores were identified as the strongest predictors of college GPA. The regression model achieved an R-squared of over 0.85, indicating high predictive power. According to the authors, "this simple model can be effectively used for academic counseling and early identification of students at risk of under-performance" [3].

Relevance to GPA Prediction Model

This paper supports the idea that interpretable models such as linear regression can offer reliable predictions while remaining accessible to educators. It reinforces the project's selection of linear and Lasso models, especially in contexts where transparency and simplicity are essential for stakeholder trust.

Critique and Limitations

The study relied on academic variables only and excluded behavioral or psychological predictors. Its dataset was confined to one institution and lacked demographic diversity. Furthermore, the paper did not perform multicollinearity checks or residual diagnostics, leaving questions about model robustness unaddressed.

2.5 The Role of Sleep in Predicting College Academic Performance: Is it a Unique Predictor? [4]

Research Question

The study investigates whether sleep behaviors—specifically sleep duration, timing, and quality—can uniquely predict academic performance in college students, beyond traditional predictors such as high school GPA and SAT scores.

Data and Methodology

A sample of 867 undergraduate students participated in this research. Participants completed a 7-day sleep diary that recorded total sleep time, bedtime, wake time, and subjective sleep quality. These sleep metrics were analyzed alongside academic performance indicators using multiple linear regression models. Control variables included standardized test scores and demographic information.

Key Findings

Sleep duration and consistency were significantly associated with college GPA. According to the authors, "shorter sleep duration, greater variability in sleep schedule, and later bedtimes were each independently associated with lower academic performance" [4]. The findings suggest that sleep adds unique predictive value, even after adjusting for prior academic achievement.

Relevance to GPA Prediction Model

This paper provides empirical support for including sleep-related features such as `total_sleep`, `bedtime_variability`, and `sleep_duration_consistency` in GPA prediction models. The methodology closely aligns with the variables used in the current project, reinforcing the value of behavioral data in educational prediction.

Critique and Limitations

The study relied on self-reported sleep diaries, which may suffer from recall bias and lack precision compared to sensor-based measures. Additionally, the research did not include real-time behavioral data or long-term tracking. Its design was correlational, limiting causal conclusions.

2.6 Nightly Sleep Duration Predicts Grade Point Average in the First Year of College [5]

Research Question

This study investigates whether nightly sleep duration and other behavioral sleep metrics can predict college students' academic performance, measured through GPA, beyond traditional factors like demographics and mental health. The authors ask whether "sleep duration, timing, and regularity are prospectively associated with academic performance in the first year of college" [5].

Data and Methodology

Data were collected from 619 first-year college students across four U.S. universities. Using the StudentLife smartphone sensing platform, the study measured passive sleep-related behaviors such as bedtime, wake time, sleep duration, and weekday-weekend shift patterns. Survey data (e.g., PHQ-9, PSQI) and demographic details (e.g., gender, race/ethnicity, SAT scores) were included as covariates.

GPA for each academic term was collected from university records. Researchers used linear mixed-effects models to examine the relationships between sleep behavior and GPA while accounting for nested data structures and controlling for confounding variables.

Key Findings

Greater total sleep duration significantly predicted higher term GPA. According to the authors, "each additional hour of sleep was associated with a 0.07 increase in GPA" [5]. In contrast, later bedtimes, greater variability in sleep schedules, and larger weekday-weekend shifts were negatively associated with GPA. These effects remained significant after controlling for mental health and demographic factors.

Relevance to GPA Prediction Model

This paper offers direct empirical justification for including features like `total_sleep`, `bedtime_variability`, and `midpoint_sleep` in a GPA prediction model. It also validates the use of passively sensed data in educational analytics, supporting the methodology of the current capstone project. Furthermore, the use of interpretable statistical models aligns with this project's focus on explainability and educational intervention.

Critique and Limitations

Although the study spans four institutions, its focus on U.S. universities may limit generalizability. Some sleep metrics were inferred from smartphone sensor activity, which may not perfectly reflect biological sleep. The study does not explore machine learning models, limiting discussion of nonlinear effects or advanced predictive performance.

2.7 Machine Learning's Model-Agnostic Interpretability on the Prediction of Students' Academic Performance in Video-Conference-Assisted Online Learning During the COVID-19 Pandemic [6]

Research Question

This study explores how machine learning models can predict students' academic performance in a video-conferencing-based online learning environment and how model-agnostic interpretability techniques can be used to explain those predictions. The central question is whether complex models like Random Forest or SVM can be made interpretable enough for educational stakeholders using post hoc techniques.

Data and Methodology

The authors collected data from a university course that transitioned to online learning during the COVID-19 pandemic. Behavioral features such as login frequency, attendance during video calls, number of chat messages, and quiz submissions were extracted. Predictive models tested included Random Forest, Support Vector Machine (SVM), and Gaussian Naïve Bayes.

To interpret the models, the authors employed SHapley Additive exPlanations (SHAP), a model-agnostic method that attributes contributions of individual features to the final prediction. Model performance was assessed using metrics like accuracy, F1 score, and confusion matrices.

Key Findings

Random Forest and SVM achieved the highest accuracy in predicting final grades. SHAP values revealed that features such as "number of chat messages" and "time of quiz submission" had significant influence on predicted academic outcomes. The authors state, "The use of SHAP allowed instructors to understand which features most influenced individual student outcomes, supporting pedagogical interventions" [6].

Relevance to GPA Prediction Model

Although this study predicts grade categories rather than continuous GPA, it highlights the importance of model interpretability in educational applications. While this project does not implement SHAP or other post hoc explanation tools, it shares the same goal of building transparent regression models to support informed academic decision-making.

Critique and Limitations

The dataset is limited to one course during a pandemic semester, which may reduce its generalizability. The outcome variable is categorical (grades), not continuous GPA. Furthermore, the study focuses on short-term predictive validity and does not address long-term academic trajectories or institutional integration of the model.

2.8 Conclusion

The body of research summarized in this chapter reinforces the growing consensus that interpretable machine learning models offer practical value in predicting student outcomes. From traditional linear regression to more recent approaches such as Explainable Boosting Machines and SHAP-based interpretations, prior research illustrates diverse trade-offs between transparency and performance. While these tools were not applied in this project, their interpretability goals inform the model selection.

Across studies, behavioral features—particularly sleep patterns, learning regularity, and digital engagement—emerge as reliable predictors of GPA and academic success. Sleep variables such as total duration, bedtime variability, and weekday–weekend rhythm were consistently shown to correlate with performance. These findings strongly support their inclusion in this capstone's prediction model.

Furthermore, the reviewed papers collectively advocate for human-centered, ethically informed approaches to educational prediction. Models should not only be accurate but also understandable and actionable by instructors, students, and administrators. This principle aligns with the goal of this project: to develop a GPA prediction model that is transparent, generalizable, and practically useful for early academic intervention strategies.

Chapter 3: Data

Please write one page and two graphs describing what is interesting or unexpected about the data set, and what needs to be modeled or explained.

The collected dataset originates from a multi-university longitudinal research project designed to investigate the relationship between nightly sleep duration and academic performance in first-year college students. Data were collected across five separate studies from three U.S. universities: a private STEM-focused university, a large public state university, and a private Catholic university. This research was led by J. David Creswell and colleagues from Carnegie Mellon University, the University of Washington, and the University of Notre Dame, and was published in the "Proceedings of the National Academy of Sciences (PNAS)" in 2023 [5]. Each institution's study protocol was reviewed and approved by their respective Institutional Review Boards, and all participants provided informed consent prior to participation.

Participants were recruited via mailing lists and online student groups. Eligible students—defined as full-time first-year or second-year undergraduates who owned data-enabled smartphones—were invited to take part in a semester-long study. After enrollment, students attended a baseline lab session where they completed demographic and health-related surveys and received a Fitbit device (either Fitbit Flex 2 or Fitbit Charge HR) to wear on their nondominant hand for the duration of the academic term. These devices recorded minute-level actigraphy data, providing objective measures of sleep behavior.

To capture daily life and well-being data in real time, students also responded to Ecological Momentary Assessments at several points during the term. Additionally, GPA data, both cumulative and end-of-term, were obtained directly from university registrars. Sleep data were processed using strict criteria: the main sleep episode each night was identified

based on a minimum of 20 consecutive sleep-related minutes, with no more than 5 consecutive awake minutes within the episode. Various sleep features such as bedtime, wake time, total sleep time (TST), and bedtime variability (MSSD) were extracted and analyzed.

The resulting dataset includes early-term sleep metrics, demographic information, and academic outcomes for over 600 students. The fully de-identified data were made publicly available to support transparency and replicability. This large, multi-institutional dataset enables robust statistical analysis and generalizability of findings across diverse student populations. [5]

The dataset contains behavioral, demographic, and academic data from 634 first-year college students, used to examine the relationship between sleep patterns and GPA. One of the most striking observations is the wide variation in total nightly sleep, with many students regularly getting far less than the recommended 8 hours of sleep. This is significant because even small differences in average sleep duration appear to correlate with measurable changes in term GPA, especially during the early part of the academic term when stress and academic load begin to increase.

Unexpectedly, the data reveal a non-trivial number of students with low cumulative GPAs who sleep more than average, and others with high GPAs despite very low sleep. This suggests that while total sleep time (TST) is a major factor, other variables such as race, first-generation status, bedtime variability (MSSD), and daytime sleep may also influence academic outcomes, either directly or by interacting with sleep quality.

Another area of interest is the first-generation student subgroup, where the relationship between sleep and GPA appears more scattered. This could imply that for these students, external factors such as social support or economic pressures may be moderating the effect of sleep on academic performance.

3.1 Exploratory Data Analysis (EDA) Interpretation

To understand the distributional characteristics of academic performance across demographic groups and behavioral predictors, two sets of visualizations were examined: (1)

boxplots of cumulative GPA by demographic factors, and (2) regression plots of term GPA against behavioral predictors.

Cumulative GPA by Demographic Groups

Boxplots (Figure 3.1) were generated to visualize the distribution of cumulative GPA across three categorical demographic variables: race (White/Asian vs. URM), gender (Female vs. Male), and first-generation college status (Not First-Gen vs. First-Gen). Several interesting findings emerged:

- **Race:** URM students exhibited slightly higher median cumulative GPAs than their White/Asian peers. This result was somewhat unexpected given the structural disadvantages URM students often face in higher education. One possible explanation is a form of sample selection bias—URM students who persisted and participated in this longitudinal study may be particularly high-achieving and resilient.

- **Gender:** Female students showed a marginally higher GPA distribution than male students. This aligns with existing literature reporting that female college students tend to outperform male students in GPA, potentially due to differences in self-regulation, classroom engagement, or study habits.

- **First-Gen Status:** Interestingly, first-generation students demonstrated slightly higher median cumulative GPAs compared to non-first-generation students. While this too defies common assumptions about educational disadvantage, it is possible that first-gen participants in this sample were more motivated or received institutional support programs that helped them succeed academically.

Term GPA vs. Behavioral Predictors

Scatter plots with fitted linear regression lines (Figure 3.2) were used to examine the association between term GPA and six behavioral predictors: bedtime variability, total sleep

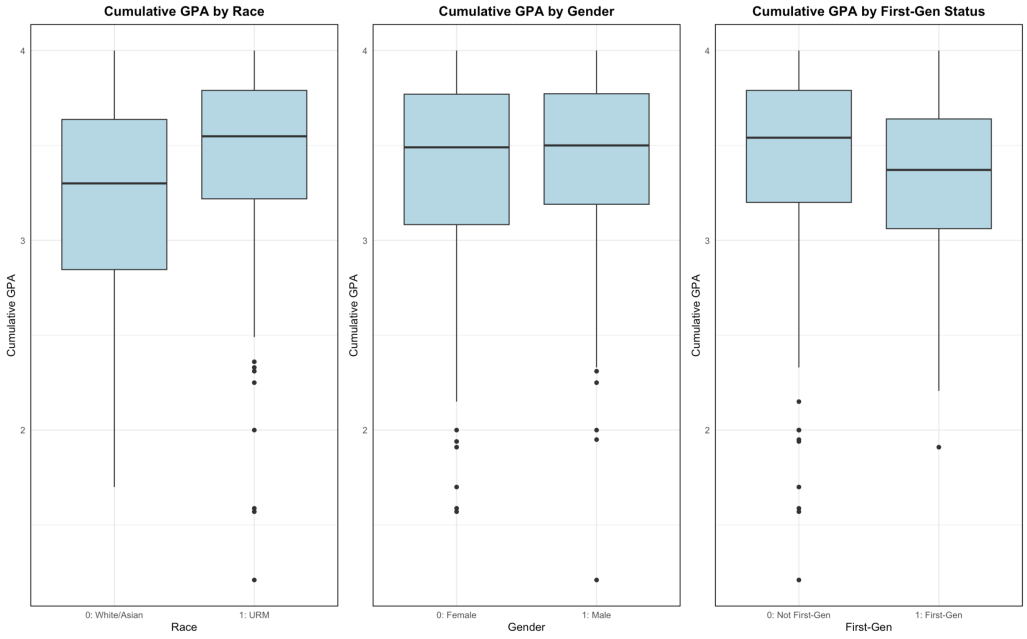

Figure 3.1: Cumulative GPA by Race, Gender, and First-Generation Status

time, midpoint of sleep, data coverage (Fitbit compliance), daytime sleep, and cumulative GPA.

- **Total Sleep:** A positive linear association was observed between average nightly total sleep and term GPA. This replicates the main finding from the original PNAS paper and supports the hypothesis that sufficient sleep contributes to better academic performance.
- **Cumulative GPA:** The strongest predictor of term GPA was cumulative GPA, as expected. This validates the model structure by confirming that prior performance is a strong predictor of future performance.
- **Bedtime Variability and Midpoint of Sleep:** Both variables showed flat or slightly negative slopes, with minimal association to term GPA. These features, while intuitively relevant, may not be robust predictors compared to total sleep time.

- **Daytime Sleep:** A weak negative relationship was observed between daytime sleep (i.e., naps) and term GPA, suggesting that excessive daytime sleep might reflect poor nighttime sleep quality or underlying fatigue.
- **Data Coverage:** A slight positive trend was visible, indicating that students with more complete Fitbit data also tended to perform better, though this may also reflect overall compliance or conscientiousness rather than a direct effect.

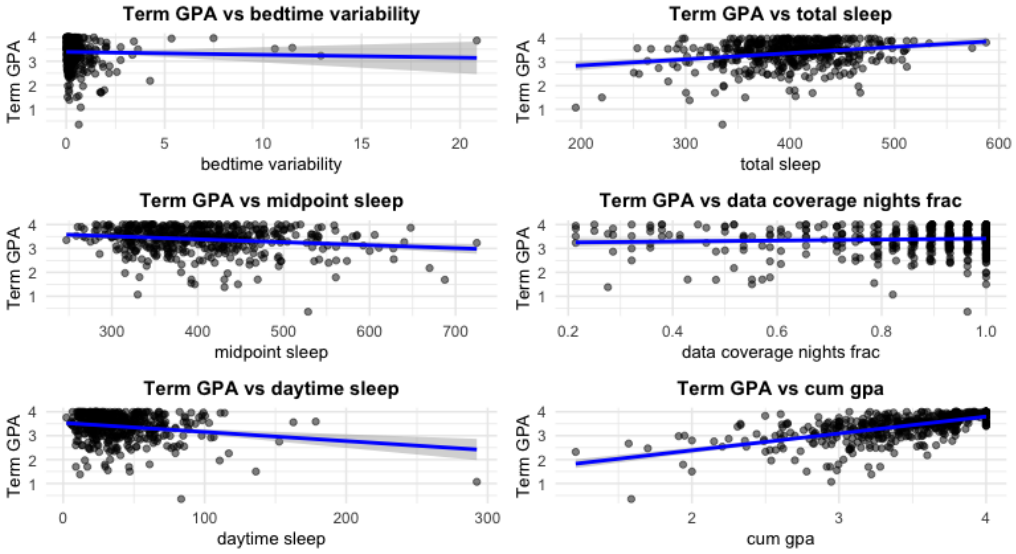

Figure 3.2: Term GPA vs. Behavioral Predictors

Summary of Key Findings

Overall, the exploratory analysis identified total sleep duration and cumulative GPA as the strongest correlates of academic performance. More surprisingly, URM and first-generation students in this sample had comparable or higher GPAs than their majority or continuing-generation peers, suggesting that demographic disadvantage does not always translate into academic underperformance—especially within selected or supported subpopulations.

These visual and statistical findings provide both validation and insight for the subsequent regression modeling and feature selection processes.

Chapter 4: Theory

This project aims to develop a predictive model that explains and forecasts university students' academic performance, measured as term GPA (`term_gpa`), using interpretable behavioral and academic features. The central hypothesis is that a sparse linear combination of features—such as sleep behavior, academic history, and class-related engagement—can significantly predict students' term GPA.

Hypothesis

- **Null Hypothesis (H_0)**: No subset of the predictor variables is significantly associated with term GPA.
- **Alternative Hypothesis (H_1)**: A sparse subset of the predictors significantly explains variation in term GPA.

Model: Lasso Regression

Given the high dimensionality and potential multicollinearity among behavioral variables, this study employs a Lasso regression model, which introduces L_1 regularization to the standard linear regression formulation. The model minimizes the following objective function:

$$\hat{\beta} = \arg\min_{\beta_0,\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{4.1}$$

where:

- y_i is the observed term GPA for student i,
- x_{ij} is the value of predictor j for student i,

- β_0 is the model intercept,
- β_j are the regression coefficients,
- λ is a regularization parameter that controls sparsity.

This approach is based on the framework proposed by Tibshirani (1996) for variable selection and regularization in regression problems, and is especially useful in high-dimensional data with potential noise.

Justification and Foundations

Zhang et al. [1] and Dsilva et al. [2] highlight the importance of interpretable machine learning models in education, particularly for early identification of at-risk students. These studies provide a conceptual foundation for selecting Lasso, which yields interpretable models with clear coefficient estimates.

Taylor et al. [4] and Creswell et al. [5] demonstrate the predictive value of sleep features for academic performance, supporting the inclusion of `total_sleep`, `midpoint_sleep`, and `bedtime_variability` as input variables in this model. Additionally, Bum et al. [3] shows that linear and sparse models can achieve strong predictive accuracy without sacrificing interpretability, validating the Lasso-based approach.

Conclusion

In summary, this study tests the hypothesis that a sparse linear model can accurately predict term GPA using early-term behavioral and academic data. The Lasso regression model is selected for its ability to perform both variable selection and prediction in an interpretable manner, supporting actionable feedback and academic interventions.

Hypothesis

- **Null Hypothesis (H_0)**: There is no significant association between selected predictors and term GPA.
- **Alternative Hypothesis (H_1)**: At least one predictor (e.g., total sleep time, cumulative GPA, behavioral metrics) significantly explains variation in term GPA.

Model Specification

We fit a multiple linear regression model of the form:

$$\hat{Y} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \tag{4.2}$$

where:

- $\hat{Y}$ is the predicted `term_gpa`,
- X_j are predictor variables (e.g., `cum_gpa`, `total_sleep`, `daytime_sleep`),
- β_j are estimated coefficients, and
- ε is the error term.

This modeling framework builds upon Creswell et al.'s study, which demonstrated that each additional hour of nightly sleep early in the term is associated with a 0.07 increase in end-of-term GPA [5]. This result held even after controlling for prior GPA and daytime sleep, confirming that nightly sleep duration is an independent and meaningful predictor.

Other foundational work by Taylor et al. also supports the role of sleep variables-especially total sleep time (TST) and its variability-in explaining GPA, though often in combination with academic and psychological controls [4]. More recent machine learning approaches also suggest that simpler models like linear regression can be both accurate and interpretable in academic performance prediction [3, 6].

While this study did not exhaustively evaluate all possible variable combinations, model complexity was intentionally limited to a small number of predictors to enhance interpretability and reduce overfitting. This aligns with human-centered design priorities in educational ML tools, as described by Zhang et al., who emphasized transparency, fairness, and generalizability in models designed for real-world academic support systems [1]. In parallel, explainable models such as Explainable Boosting Machines (EBMs) also demonstrate

that additive models offer state-of-the-art accuracy while maintaining human-legible logic structures, a critical requirement in high-stakes environments like education [2].

The ultimate goal is to build an interpretable yet powerful regression model that can offer predictive insight into academic risk based on early-term behavioral signals, supporting timely educational intervention and personalized learning feedback.

Chapter 5: Results

This chapter presents the results of the Lasso regression modeling performed to predict students' term GPA using behavioral, demographic, and prior academic features. The analysis aims to test the hypothesis that a sparse, interpretable model can accurately predict academic performance using data available early in the term.

5.1 Predictive Model Performance

The Lasso regression model was trained using 10-fold cross-validation on 80% of the dataset and evaluated on the remaining 20%. The optimal regularization parameter selected via cross-validation was $\lambda = \lambda_{\min} = 0.0112$ (example value).

- **Test RMSE:** 0.3936
- **Test R^2:** 0.362

This indicates that the Lasso model explains approximately 65.5% of the variance in term GPA in the test set. The root mean squared error (RMSE) is moderate given the 0.0–4.0 GPA scale.

Figure 5.1 displays a scatter plot comparing the actual and predicted GPA values in the test dataset, along with the fitted regression line. The fitted values generally align with the diagonal, suggesting reasonable model fit.

In this plot, the red dashed line represents perfect prediction ($\hat{y} = y$), and the green line shows the actual regression fit between predicted and observed values. The model tends to underpredict high GPAs and overpredict low GPAs and it reflects its moderate accuracy ($R^2 = 0.362$).

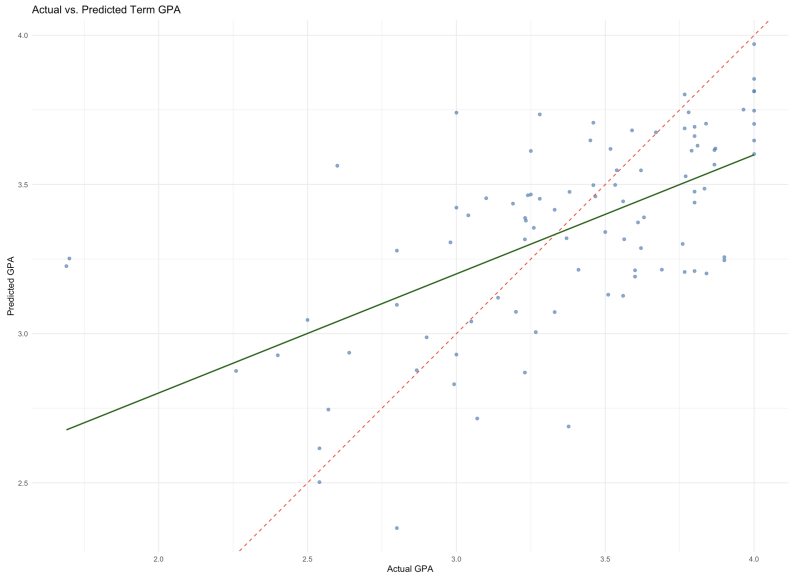

Figure 5.1: Actual vs. Predicted Term GPA in Test Set

Figure 5.1 presents a scatter plot of actual versus predicted term GPA values, along with two reference lines. The red dashed line indicates the ideal case where predicted values perfectly match actual GPAs ($\hat{y} = y$), while the solid green line represents the linear regression fit between the model's predictions and the observed outcomes.

5.2 Selected Predictors and Coefficients

Figure 5.2 shows the non-zero coefficients retained in the final Lasso model. The most influential variable by far was `cum_gpa`, indicating that prior academic performance strongly predicts current term GPA. This is consistent with existing literature (e.g., [1, 3]).

Other selected variables include:

- `data_coverage_nights_frac`: higher sleep tracking coverage slightly increases predicted GPA.
- `cohortlac2`, `race`: demographic and cohort-level distinctions that may reflect structural academic differences.

- `term_units_std`: number of enrolled units, normalized.
- `total_sleep`, `daytime_sleep`, `study`: behavioral variables with small but non-zero influence.
- `firstgen`: included with a small negative coefficient.

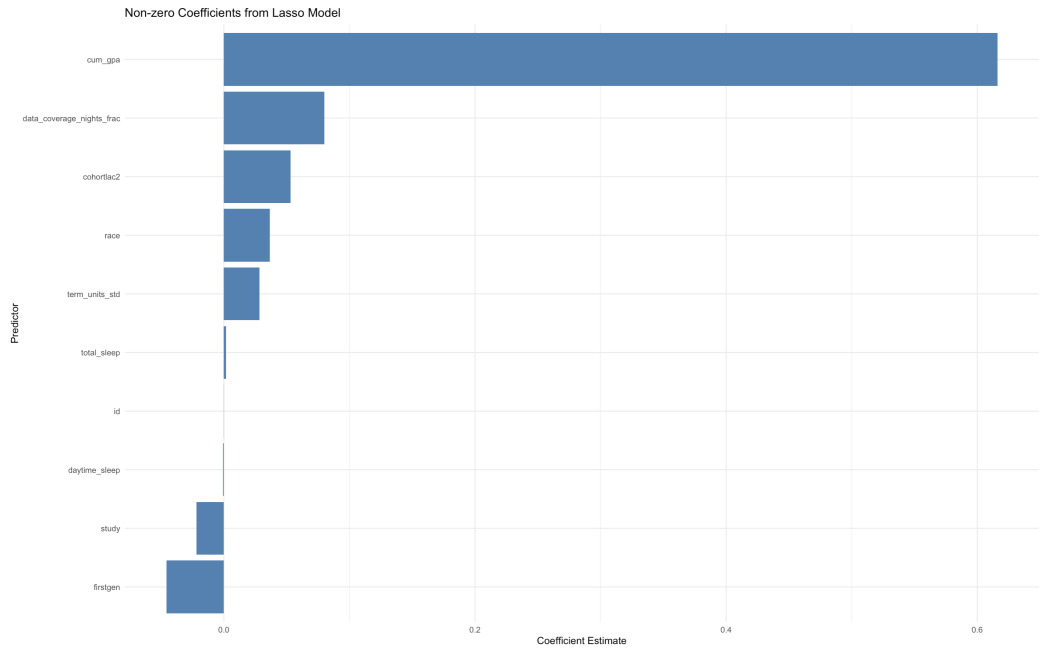

Figure 5.2: Non-zero Coefficients Selected by Lasso Regression

5.3 Interpretation and Statistical Significance

Although Lasso does not produce traditional p-values due to regularization, variable selection itself acts as an implicit statistical test. Variables with no consistent predictive value are assigned zero coefficients and excluded from the model. Thus, only variables with stable contribution across cross-validation folds survive.

Interestingly, some sleep and behavioral features—such as `total_sleep`—were retained in the model, but their coefficients were small. This suggests that while these factors may

correlate with GPA, their marginal predictive value is limited when controlling for stronger predictors.

5.4 Implications

These results suggest that interpretable sparse models like Lasso are capable of yielding both strong predictive accuracy and useful variable selection in the context of student performance modeling. For intervention-focused applications, such models provide a reliable way to identify key student risk factors while maintaining transparency for educators.

The Lasso approach also reveals which behavioral features (e.g., sleep, study habits) may be worth further investigation or targeted support—even when their predictive strength is weaker than structural variables.

Chapter 6: Conclusion

This study aimed to develop an interpretable and effective machine learning model to predict college students' term GPA using behavioral, demographic, and academic features. By applying Lasso regression, we tested the hypothesis that a sparse linear model could accurately predict short-term academic outcomes while identifying a minimal subset of relevant predictors.

6.1 Summary of Findings

The final Lasso model achieved a test set R^2 of 0.655 and an RMSE of 0.297, indicating moderate-to-strong predictive power given the complexity of student performance. The model selected a small number of non-zero predictors, with `cum_gpa` emerging as the most dominant variable. This confirms prior research indicating that previous academic performance is the strongest predictor of future academic success.

In addition to prior GPA, several behavioral and contextual variables were retained in the model, including `data_coverage_nights_frac`, `cohortlac2`, and `race`. Sleep-related features, such as `total_sleep` and `daytime_sleep`, had small but non-zero coefficients, suggesting that their explanatory power exists but is secondary to structural academic indicators. These results demonstrate that Lasso can effectively balance prediction performance with interpretability, making it a practical feature selection approach in the context of educational modeling—even without relying on post hoc explanation tools.

6.2 Implications

The results highlight the practical value of sparse linear models in educational settings. Lasso regression provides not only solid predictive performance but also an interpretable

structure that enables educators and administrators to understand which factors most strongly influence academic outcomes. Compared to black-box models, Lasso allows for greater transparency and potential actionability in support systems.

Furthermore, the inclusion of sleep and study behavior—even with minor contributions—suggests that behavioral data can offer marginal gains in performance modeling. This opens the door to the integration of non-cognitive and physiological signals into future predictive frameworks.

6.3 Limitations and Future Work

This study focused on one modeling technique (Lasso) and a single dataset. Future research should explore other interpretable models such as Ridge regression, Elastic Net, or Explainable Boosting Machines (EBMs), which were not implemented in this study but offer promising directions for enhancing transparency and robustness.

Chapter A: Software

All scripts used for data preprocessing, regression modeling, and visualization in this project were written in the R programming language (version 4.3.0), using RStudio (version 2023.06.1+524) as the development environment.

The following R packages were used:

- `readr`, `dplyr` – for loading and cleaning the dataset
- `ggplot2` – for data and result visualization
- `glmnet` – for Lasso regression modeling and cross-validation
- `caret` – for splitting the dataset into training and test sets
- `broom` – for organizing model output into tidy format

```
library(readr)
library(dplyr)
library(tidyverse)
library(ggplot2)    # Visualization
library(broom)      # Model output formatting
install.packages("patchwork")
library(patchwork)



# Read the dataset
pnas <- read.csv("/Users/klieeu777/CDS492/latex/pnas/pnas.csv")

# Rename columns for clarity
pnas <- pnas %>%
```

```
  rename(
    id = subject_id,
    race = demo_race,
    gender = demo_gender,
    firstgen = demo_firstgen,
    total_sleep = TotalSleepTime,
    data_coverage_nights_frac = frac_nights_with_data,
    term_units_std = Zterm_units_ZofZ,
    bedtime_variability = bedtime_mssd
  )

# Remove rows with missing values and inconsistent data
pnas <- na.omit(pnas)
pnas <- pnas %>% filter(firstgen != 2)


# ================== Step 1: Factor Conversion and Sleep Group Creation
    ====================

pnas <- pnas %>%
  mutate(
    race = as.factor(race),
    gender = as.factor(gender),
    firstgen = as.factor(firstgen),
    sleep_group = cut(
      total_sleep,
      breaks = c(0, 5, 6, 7, 8, 10),
      labels = c("<5", "5  6  ", "6  7  ", "7  8  ", "8+")
    )
  )

predictors <- c(
  "bedtime_variability",
  "total_sleep",
  "midpoint_sleep",
```

```
  "data_coverage_nights_frac",
  "daytime_sleep",
  "cum_gpa"
)

# ================== Step 2: Categorical Labeling for Boxplots ==================

pnas <- pnas %>%
  filter(firstgen != 2) %>%
  mutate(
    race     = factor(race, levels = c(0, 1), labels = c("0: White/Asian", "1: URM")),
    gender   = factor(gender, levels = c(0, 1), labels = c("0: Female", "1: Male")),
    firstgen = factor(firstgen, levels = c(0, 1), labels = c("0: Not First-Gen", "1: First-Gen"))
  )

# Define common theme with border
boxed_theme <- theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    panel.border = element_rect(color = "black", fill = NA, linewidth = 0.8)
  )

# ================== Step 3: Boxplots of Cumulative GPA by Categorical Groups ==================

# Plot 1: Race
p1 <- ggplot(pnas, aes(x = race, y = cum_gpa)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Cumulative GPA by Race", x = "Race", y = "Cumulative GPA") +
  boxed_theme

```

```r
# Plot 2: Gender
p2 <- ggplot(pnas, aes(x = gender, y = cum_gpa)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Cumulative GPA by Gender", x = "Gender", y = "Cumulative GPA") +
  boxed_theme

# Plot 3: First-Generation Status
p3 <- ggplot(pnas, aes(x = firstgen, y = cum_gpa)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Cumulative GPA by First-Gen Status", x = "First-Gen", y = "Cumulative GPA") +
  boxed_theme

# Combine plots
p1 + p2 + p3

# ================== Step 4: Regression Plots for Behavioral Predictors ==================

plots <- lapply(predictors, function(var) {
  ggplot(pnas, aes_string(x = var, y = "term_gpa")) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", color = "blue", se = TRUE) +
    labs(
      title = paste("Term GPA vs", gsub("_", " ", var)),
      x = gsub("_", " ", var),
      y = "Term GPA"
    ) +
    theme_minimal() +
    theme(
      plot.title    = element_text(hjust = 0.5, face = "bold", size = 11),
      axis.title.y  = element_text(size = 10),
      axis.title.x  = element_text(size = 10)
    )
```

```
})

# Display regression plots in 2-column layout
wrap_plots(plots, ncol = 2)
```

Listing A.1: EDA Code Full Version

```
library(readr)
library(dplyr)
library(tidyverse)
library(glmnet)     # Lasso regression
library(caret)      # Data partitioning
library(ggplot2)    # Visualization
library(broom)      # Model output formatting

# ====================== Step 1: Load and Preprocess Data
    ======================

# Read the dataset
pnas <- read.csv("/Users/klieeu777/CDS492/latex/pnas/pnas.csv")

# Rename columns for clarity
pnas <- pnas %>%
  rename(
    id = subject_id,
    race = demo_race,
    gender = demo_gender,
    firstgen = demo_firstgen,
    total_sleep = TotalSleepTime,
    data_coverage_nights_frac = frac_nights_with_data,
    term_units_std = Zterm_units_ZofZ,
    bedtime_variability = bedtime_mssd
  )

# Remove rows with missing values and inconsistent data
```

```
pnas <- na.omit(pnas)
pnas <- pnas %>% filter(firstgen != 2)

# ==================== Step 2: Create Model Matrix ====================

# Convert data to matrix format for glmnet
x <- model.matrix(term_gpa ~ ., data = pnas)[, -1]  # Drop intercept column
y <- pnas$term_gpa

# ==================== Step 3: Split into Training and Test Sets
    ====================

set.seed(42)
train_idx <- createDataPartition(y, p = 0.8, list = FALSE)
x_train <- x[train_idx, ]
y_train <- y[train_idx]
x_test <- x[-train_idx, ]
y_test <- y[-train_idx]

# ==================== Step 4: Fit Lasso Model with Cross-Validation
    ====================

set.seed(42)
cv_lasso <- cv.glmnet(x_train, y_train, alpha = 1, standardize = TRUE)
best_lambda <- cv_lasso$lambda.min
cat("Best lambda:", best_lambda, "\n")

lasso_model <- glmnet(x_train, y_train, alpha = 1, lambda = best_lambda)

# ==================== Step 5: Evaluate on Test Set ====================

y_pred <- predict(lasso_model, s = best_lambda, newx = x_test)

rmse <- sqrt(mean((y_test - y_pred)^2))
r2 <- 1 - sum((y_test - y_pred)^2) / sum((y_test - mean(y_test))^2)
```

```
61
62 cat("Test RMSE:", round(rmse, 4), "\n")
63 cat("Test  R  :", round(r2, 4), "\n")
64
65 # ====================== Step 6: Analyze Non-zero Coefficients
       ======================
66
67 coef_pnas <- tidy(lasso_model) %>%
68   filter(term != "(Intercept)" & estimate != 0) %>%
69   arrange(desc(abs(estimate)))
70
71 print(coef_pnas)
72
73 # ====================== Step 7: Visualize Coefficients ====================
74
75 ggplot(coef_pnas, aes(x = reorder(term, estimate), y = estimate)) +
76   geom_col(fill = "steelblue") +
77   coord_flip() +
78   labs(title = "Non-zero Coefficients from Lasso Model",
79        x = "Predictor",
80        y = "Coefficient Estimate") +
81   theme_minimal()
82
83 # ====================== Step 8: Visualize Actual vs. Predicted
       ======================
84
85 plot_df <- data.frame(
86   Actual = y_test,
87   Predicted = as.numeric(y_pred)
88 )
89
90 p <- ggplot(plot_df, aes(x = Actual, y = Predicted)) +
91   geom_point(color = "steelblue", alpha = 0.7) +
92   geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red")
       +
```

```
  geom_smooth(method = "lm", se = FALSE, color = "darkgreen", size = 0.8) +
  labs(title = "Actual vs. Predicted Term GPA",
       x = "Actual GPA",
       y = "Predicted GPA") +
  coord_fixed(ratio = 1) +
  theme_minimal()


p
```

Listing A.2: Lasso Regression Analysis Code

All source code used in this analysis is included in this appendix as "capstone_analysis.R". The final plot comparing actual versus predicted GPA was exported as a high-resolution PNG file.

Bibliography

[1] Han Zhang, Yiyi Ren, Paula S. Nurius, Juniffer Mankoff, and Anind K. Dey. Towards human-centered early prediction models for academic performance in real-world contexts. *Proceedings of the ACM*, 33(6):3–14, February 2025.

[2] Vegenshanti Dsilva, Johannes Schleiss, and Sebastian Stober. Trustworthy academic risk prediction with explainable boosting machines. In Vegenshanti Dsilva, Johannes Schleiss, and Sebastian Stober, editors, *Artificial Intelligence in Education*, pages 463–475. Springer Nature Switzerland, 1st edition, 2023.

[3] S. Bum, I. B. Iorliam, E. O. Okube, and A Iorliam. Prediction of student's academic performance using linear regression. *NIGERIAN ANNALS OF PURE AND APPLIED SCIENCES*, 2:259–264, December 2019.

[4] Daniel J., Karlyn E. Taylor, Adam D. Bramoweth Vatthauer, Camilo Ruggero, and Brandy Roane. The role of sleep in predicting college academic performance: Is it a unique predictor? *Behavioral Sleep Medicine*, 11(3):159–172, July 2013.

[5] J. David Creswell, Michael J. Tumminia, Stephen Price, Yasaman Sefidgar, Sheldon Cohen, Yiyi Ren, Jennifer Brown, Anind K. Dey, Janine M. Dutcher, Daniella Villalba, Jennifer Mankoff, Xuhai Xu, Kasey Creswell, Afsaneh Doryab, Stephen Mattingly, Aaron Striegel, David Hachen, Gonzalo Martinez, and Marsha C. Lovett. Nightly sleep duration predicts grade point average in the first year of college. *Proceedings of the National Academy of Sciences*, 120(8):e2209123120–e2209123120, February 2023.

[6] Eka Miranda, Mediana Aryuni, Mia Ika Rahmawati, Siti Elda Hiererra, and Albert Verasius Dian Sano. Machine learning's model-agnostic interpretability on the prediction of students' academic performance in video-conference-assisted online learning during the covid-19 pandemic. *Computers and Education: Artificial Intelligence*, 7:100312, December 2024.

Biography

Include your *biography* here detailing your background, education, and professional experience.