Final Report: Building a Predictive Model to Support Students in Need

By: Grace Brown, Nicholas Hughes, Zoya Masood, Emma Mills, Bora Ya Diul, Lauren Wisniewski, and Valyn Grebe

GitHub repository: https://github.com/laurenwisniewski/DS-3001-Project

Abstract

The COVID-19 pandemic resulted in extensive changes to many aspects of student life: academics, socialization, and health (both mental and physical). This study investigates the predictability of lifestyle factors in the variation of sleep duration - a critical indicator of physical and mental health - by age, social media use, meal frequency, and combined time spent on online classes and independent study, 'classAndStudy'. Identification of such associations will provide insight for targeted interventions to support students' well-being.

Using a data set containing observations for 1,182 students in the Delhi NCR region of India, a Principal Component Analysis (PCA) was done for reducing the dimensionality, and then linear regression analysis was applied to check the correlations. The PCA showed that the first three components explain more than 95% of the variance in the dataset. Linear regressions between lifestyle variables and sleep duration revealed small but significant correlations: R-squared value equals 0.09. Key findings included a negative correlation with age of -0.021, a negative correlation with 'classAndStudy' of -0.070, a very slight positive correlation with social media use of 0.009, and a stronger positive correlation with meal frequency of 0.122.

Our results show that although lifestyle factors alone explain only a small percentage in the variation in sleep, they provide the actionable targets for university health services seeking to identify students at risk. However, reliance on self-report data and a correlational design result in less strength for this study's interpretation. Using sleep as a proxy for overall health necessarily restricts the scope of possible implications, while cultural and geographic characteristics of the sample mean these findings must be generalized beyond the NCR with caution.

Future research should incorporate causal analysis, broader health indicators, and diverse datasets to further refine predictive models and extend their generalizability. In spite of these limitations, our study highlights the importance of embedding behavioral insights into health

interventions and lays the foundation for data-driven approaches toward improving student support systems.

Introduction

COVID-19 has fundamentally transformed the lives of students. From academic schedules to social interactions with friends, routines have become disrupted in nearly every aspect. Through these disruptions, the importance of lifestyle behavior in the consequences of health among students gained awareness. The rapid changes in the environment have created challenges for universities and schools to support students' well-being around the world. In this scenario, knowing the predictors of health, particularly those that are within the sphere of lifestyle choices, has become an important step toward the design of effective interventions for student support. Among the most available and valid health indicators is sleep duration, which is directly proportional to both physical and mental health.

Adequate sleep has been proven to enhance cognitive function, emotional regulation, and overall productivity (Chen, Wang, and Jeng 2006). In contrast, sleep deprivation has been linked to several adverse conditions such as increased levels of stress, poor academic performance, and heightened risks of developing mental disorders (Meldrum and Restivo 2014). Due to its importance, sleep is an appropriate indicator for assessing the welfare state of students. The current study seeks to establish the relationship between lifestyle factors like age, time spent on social media, number of meals per day, and time devoted to academic activities and sleep duration. These variables were selected based on their established connections to health outcomes in the literature.

The main goal is to determine the predictive power of these lifestyle variables regarding sleep duration and thus provide clues on behaviors that may warrant additional support from university health services. We will be analyzing the trend of sleep duration using PCA combined with linear regression modeling, using a dataset comprising responses from 1,182 students in the Delhi NCR, India, during the pandemic. The dataset includes self-reported data on students' age, academic commitments, social media usage, meal frequency, and sleep habits-a snapshot of how daily routines affect health in a period of great disruption in society. Initial data preprocessing

included the removal of incomplete responses and the consolidation of variables for analysis. Of particular note, we constructed a new variable, ‘classAndStudy’, combining time spent on online classes and independent study to account for the interconnected demands of academic life.
We used PCA on our data to reduce the dimensionality and to find major components that explain most of the variance.

Because the first three components alone expressed over 95% of the dataset's variability, they are a good backbone for further regression analysis. Small-size but statistically significant associations of sleep duration were obtained in the framework of linear regression modeling, together with chosen lifestyle variables. For example, there is a negative relationship between sleep duration and age: one might therefore conjecture that older students are those who sleep shorter compared to their younger colleagues. In a similar vein, time used for academic activities was inversely related to sleep, suggesting that heavier academic loads may be associated with less time for sleeping. On the other hand, meal frequency was positively related to sleep, reflecting the significance of regular eating for a healthy lifestyle.

Social media use showed a weak positive correlation, which is a finding that deserves further investigation. Although the R-squared value of our regression model was relatively low at 0.09, the results are actionable. They suggest areas where university health services can direct their efforts, such as promoting regular meal schedules and encouraging balanced academic workloads to help students improve their sleep and overall health. These findings further suggest that sleep duration, while influenced by a multitude of factors, can be partly predicted by routine behaviors. However, these findings should be interpreted with caution. The data in this study are self-reported, which may introduce certain biases or inaccuracies. Furthermore, the correlational design of the study precludes any firm conclusions about causality. The cultural and geographical specificity of the sample further limits generalizability of the findings. Despite these limitations, this study lays the groundwork for data-driven approaches to student health and highlights the potential of behavioral interventions in supporting well-being. In the following sections, we present a detailed description of the dataset and preprocessing, analytical methods used, presentation of results, and discussion. We also discuss the limitations of the study and suggest some future research avenues to refine and expand on these findings. By linking data analysis to application, this study intends to add to the growing body of knowledge in informing targeted interventions that will improve students' health and resilience in the face of ongoing challenges.

Dataset

Our dataset contains information about COVID-19 and its impact on students attending educational institutions in the Delhi National Capital Region (NCR) in India. The data were gathered through a cross-sectional survey which was completed by 1,182 students of different age groups in the NCR area in 2020. The original dataset contained rows for each respondent and their answers to the survey. The columns (and their data type) in the original dataset were as follows: region of residence (str), age of subject (int), time spent on online class (float), rating of online class experience (str), medium for online class (str), time spent on self study (float), time spent on fitness (float), time spent on sleep (float), time spent on social media (float), preferred social media platform (str), time spent on TV (str), number of meals per day (int), change in your weight (str), health issue during lockdown (str), stress busters (str), time utilized (str), do you find yourself more connected with your family, close friends, and relatives ? (str), what you miss the most (str).

For our cleaning process, we dropped 92 responses that had a missing value in one or more of the columns. We also dropped all the columns that were not related to our selected numerical variables, leaving us with the columns “age of subject”, “time spent on online class”, “time spent on self study”, “time spent on fitness”, “time spent on sleep”, “time spent on social media”, “time spent on tv”, and “number of meals per day”. We also created a column called “classAndStudy” to create an interaction between time spent in online class and time spent on self study. This dataset is useful because it includes many lifestyle and health variables that are important for students’ wellbeing.

We immediately noticed the majority of the data were from students aged 13-23 years old. In India, depending on the program, university can last between 3-5 years. Considering this, we grouped students by level of education: university students (ages 18-23) and high school (secondary) students (ages 13-17). Additionally, since we do not have a variable that explicitly measures mental health, we used the proxy variable “time spent on sleep”. We chose this variable because it reflects what could be an important physical effect of mental health.

In addition to the challenge of identifying predictive variables, we anticipate difficulties in applying our findings to other geographical regions and school environments outside of the NCR in India. Since, the survey performed an in-depth questionnaire of the virtual learning experience for NCR students, our predictive model will be able to accurately predict how educational level affects mental health for NCR students. However, there will be significant challenges in applying our insights to different learning environments and cultural contexts.

Methods

Our main question was to see whether we could use lifestyle variables to predict an important indicator of health, which could help identify which students are more likely to need more support at school. The lifestyle variables we considered were the age of students, time spent on social media, number of meals per day, and the combined time spent on online class as well as independent study (the interaction variable classAndStudy). The health variable we used was time spent on sleep. We believe that a predictive model such as ours (which uses lifestyle variables to predict an important measure of health in students) could be useful to university health clinics who may want to identify individuals at a potentially higher risk for struggling with their physical or mental health. As will be noted in our criticisms and concerns section, however, our model is based on correlations, not causations and should be used with caution.

An observation in our study represents an individual student, who has self-reported their activity (such as attending class, eating, exercising, and sleeping). Using proxy variables to represent health will help us predict what types of students may be most at risk without explicitly asking all students to classify their own wellbeing. For example, we expect that students who tend not to exercise or go to class will have poorer outcomes than other students.

To address our prediction question, we first did a PCA decomposition of the lifestyle variables to find that the first three principal components explained almost all of the variation in our dataset. Then, we ran linear regression (supervised learning) using these three components to find correlation coefficients that could be used to predict hours of sleep (a representation of overall student wellbeing in our study) from our lifestyle variables. We determined that our approach would “work” and have success if our model produced a significant R-squared value and useful correlation coefficients for predicting hours of sleep.

Since we anticipated that there might be some overlap in the effects that our lifestyle variables have, we created an interaction variable. Considering that more challenging classes tend to be more demanding of students’ time spent in both class and in self study, we used a transformation to create the classAndStudy interaction variable to include the sum of these times. When we ran our regression, we used this interaction variable to represent the combination of these two variables. We also used PCA to figure out which variables truly explained most of the variation in our data as well as to avoid multicollinearity.

Results

As stated previously, our main question was to see whether we could use lifestyle variables to predict an important indicator of health, which could help identify which students are more likely to need more support at school. The lifestyle variables we considered were the age of students, time spent on social media, number of meals per day, and the combined time spent on online class as well as independent study (the interaction variable classAndStudy). The health variable we used was time spent on sleep. We believe that a predictive model such as ours (which uses lifestyle variables to predict an important measure of health in students) could be useful to university health clinics who may want to identify individuals at a potentially higher risk for struggling with their physical or mental health. As will be noted in our criticisms and concerns section, however, our model is based on correlations, not causations and should be used with caution.

To address our prediction question, we first did a PCA decomposition of the lifestyle variables to find that the first three principal components explained almost all of the variation in our dataset. Then, we ran linear regression using these three components to find small but significant correlations between the lifestyle variables and our health variable. Our R-squared value was 0.09 and the correlation coefficients for our lifestyle variables were -0.02091545 (for age), -0.07014782 (for the classAndStudy variable), 0.00909555 (for time spent on social media), and 0.12185532 (for number of meals per day). These correlation coefficients can be used to predict hours of sleep (a representation of overall student wellbeing in our study) from our lifestyle variables.

Figure 1:

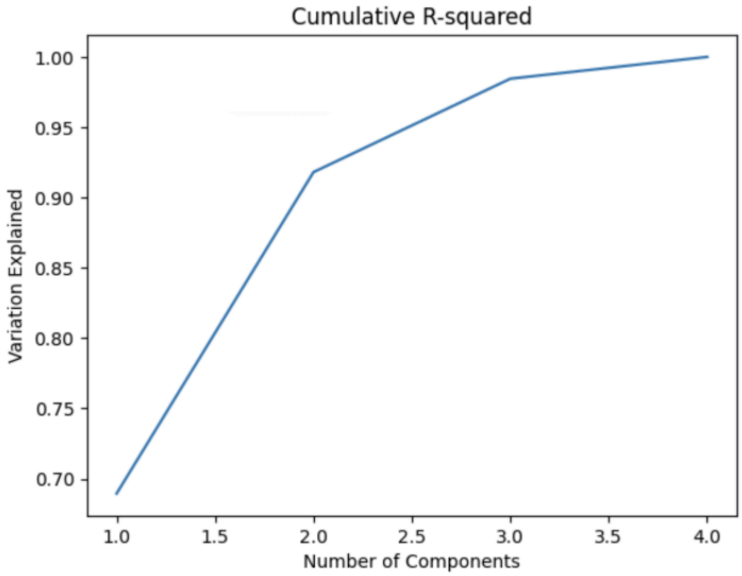
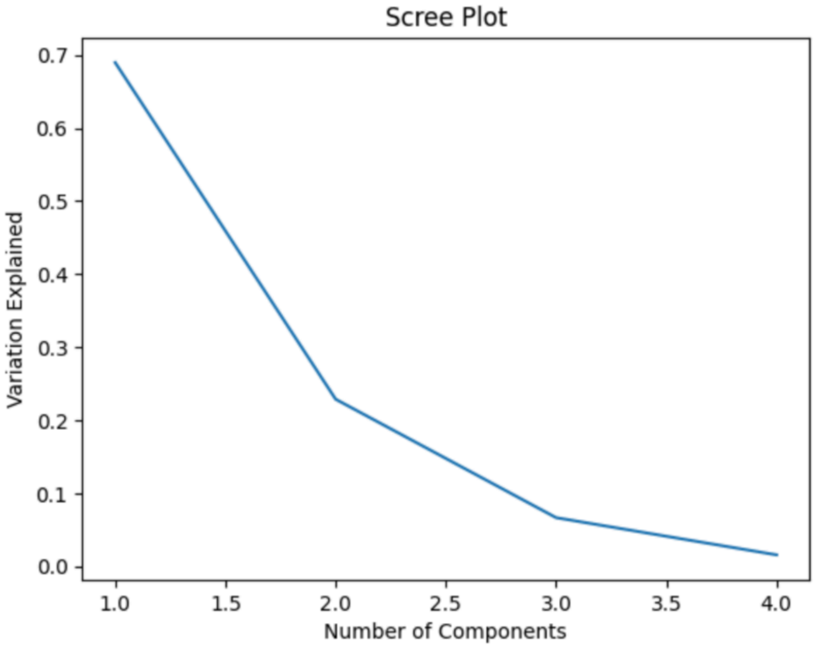


Figure 1 illustrates the cumulative R-squared as a function of the number of components, showing the percentage of variation explained as additional components are added. There is a significant increase in cumulative R-squared from one to two components, with explained variation rising from approximately 70% to over 90%. Adding a third component increases the explained variation to over 95%, while the fourth component brings it close to 100%. Using only one component in our PCA would likely result in underfitting, as it fails to capture sufficient variance in the data. Conversely, using all four components may lead to overfitting, where the model becomes overly complex and less generalizable. Selecting two components provides an adequate balance, capturing around 90% of the variance. However, we wanted a higher threshold of explained variance closer to 95%, so incorporating a third component was preferable. This graph was helpful determining the optimal number of components to use in the principal component analysis, ensuring a model that balances simplicity with explanatory power.

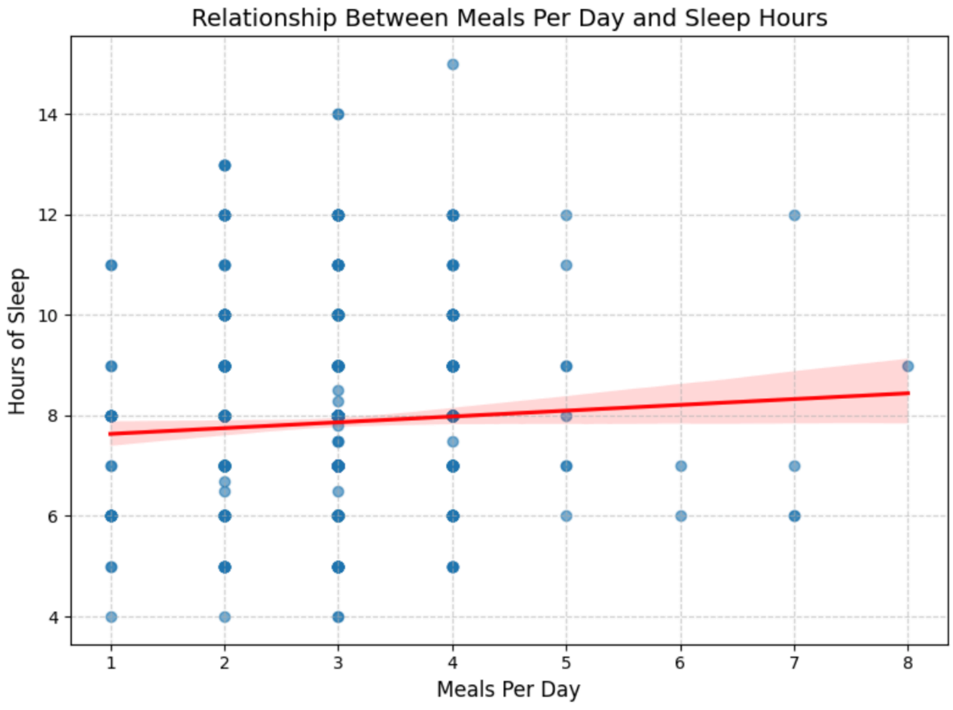
Figure 2:



The Scree Plot, *Figure 2,* demonstrates the eigenvalues of the PCA components within our analysis and looks at how much of the variance can be explained by a certain number of components. When looking at this graph, there is a large drop from one to two components before the decline is a bit more gradual from two components to four components. At one component, the corresponding eigenvalue of 0.689 before steeply dropping off to an eigenvalue of 0.229 when a second component is added. With a third component, the amount of variance explained decreases to a value of .066 and the addition of a fourth component drops this amount to 0.0156. The 'elbow' point of this graph, or where the variance explained by the number of components begins to level off, appears to occur when there are three components included. This is significant as it indicates that including one to two components within our PCA would likely result in an optimal output for our analysis. There is the initial drop in the amount of variance explained from one to two components, however, this drop off continues until three components where the eigenvalues begin to level off. Therefore, utilizing only one or two components will not optimally explain the variance whereas using three components will. Further, using a number of components past three will result in overfitting the data, as it is past the optimal level.

Because our dataset is a cross-sectional survey of students taken at a single point in time, we do not recommend drawing causal relationships from our linear regression model. For example, we found that the number of meals per day has a positive relationship with our proxy variable for health, time spent on sleep. This does suggest that students who eat a lesser number of meals per day are more likely to have worse mental health statuses, however it does not imply that increasing the number of these students' meals would automatically lead to better health outcomes.

Figure 3:



The scatterplot, *Figure 3,* examining the relationship between meals per day and hours of sleep reveals a slight positive trend, where individuals consuming more meals tend to sleep marginally longer. However, the weak slope of the regression line and the wide confidence interval suggest a low correlation, indicating that meals per day alone do not strongly predict sleep patterns. Most observations cluster around 1-3 meals per day, with sleep hours ranging widely within this group. This finding highlights the limited role of meal frequency in explaining sleep duration and suggests that additional variables, such as stress levels, time spent online, or physical activity, may also account for variations in sleep behavior. This visualization underscores the importance of considering multiple factors when modeling behaviors influenced by complex health and lifestyle interactions.

Conclusion

Over the course of the semester, we conducted an in-depth investigation into the relationship between students' lifestyle variables and proxy variables representing mental health during the COVID-19 pandemic in Delhi, India, specifically the NCR region. Our study aimed to understand how various aspects of students' daily routines, such as time spent on online classes, self-study, fitness, social media, and meals per day, influenced their wellbeing during the unprecedented challenges of lockdown. By focusing on time spent on sleep as a proxy for mental health, we sought to uncover patterns that could help identify which students may be at a higher risk of experiencing stress or other challenges to their mental and physical health.

Our model had an R-squared value of 0.09, indicating its limited explanatory power. This highlights the first key limitation of our dataset: the difficulty in implying causation given the nature of the data and the context of COVID-19. Numerous factors simultaneously impacted students' daily lives during the lockdown, all of which likely influenced their wellbeing. As a result, it is challenging to pinpoint any single factor as the definitive cause. Additionally, since our model is correlation-based, it does not imply causation. Given the significant changes experienced during the lockdown, it is more reasonable to assume that a combination of factors collectively explains changes in students' mental health, rather than attributing these changes to one specific factor.

Despite this, our model still depicted several weak correlations between student’s lifestyle indicators and time spent on sleep (proxy variable for mental health). First, it was shown that as age increases there is a 0.021 drop in hours spent on sleep. This slight negative correlation indicates that sleep was slightly reduced for student’s during COVID-19, however, not substantially different from hours spent on sleep prior to COVID-19. Second, there was a slight negative correlation, -0.07, between time spent on classes and studying and time spent on sleep. This indicates that during the pandemic time was spent on virtual learning and self-study is associated with slightly less sleep. Third, the correlation between time spent on social media and time spent on sleep is 0.009. This minimal positive correlation implies there is essentially no impact on time spent on social media and the amount of sleep students are receiving. Lastly, the model demonstrates a weak positive correlation, 0.122, between students' meals per day and time spent on sleep per night. This implies that students who tend to eat more throughout the day, tend to sleep slightly more. Overall, our model indicated several small correlations between sleep and lifestyle factors that impact students' lives during the pandemic, suggesting that none of our variables are strong predictors of time spent on sleep. This addresses the second key limitation of our dataset: the lack of variables to comprehensively explain variations in sleep or, more broadly, to capture impacts on students' wellbeing and mental health. The dataset lacked clear indicators specifically tied to mental health and did not include data collected over an extended period of time to track behavioral changes, providing only a snapshot of student experiences. However, our dataset serves as an important baseline for answering our research question. With additional research and more robust data, we could incorporate stronger variables to establish clearer and more distinct relationships.

The primary goal of our research was to develop a predictive model that could guide schools and universities in providing targeted support to students. By analyzing lifestyle data, we hoped to offer insights into how behaviors like study habits and meal frequency might interact with students' overall mental health. Given the current scope of our dataset, we could provide NCR school districts with actionable insights from our research to help improve students' wellbeing. For instance, our findings highlight potential correlations between lifestyle factors such as the number of meals per day, time spent on academic activities, and sleep duration, which serves as a proxy for mental health. By sharing this information, school districts could implement targeted programs that encourage healthier habits, such as promoting balanced meal plans, structuring academic workloads to reduce stress, and educating students on the importance of sleep for mental and physical health.

Furthermore, school districts and universities could use our research as a foundation for conducting further studies tailored to their specific needs. The educational system at all grade levels can conduct further research on incorporating more detailed data and data collected across time to accurately track behavioral trends and identify students with high risk for serious mental health concerns. The NCR school district and regional universities can develop early intervention strategies and programs to ensure the success and mental wellbeing of their students, setting the stage for other school districts in India and across the world to follow suit. This approach addresses the third limitation of our dataset: the results are only applicable to NCR students in Delhi, India. By scaling up the study to include diverse geographical regions and cultural contexts, the findings could be made more generalizable, enabling educators and policymakers worldwide to better understand the impact of lifestyle factors on student wellbeing. Expanding the dataset would provide insights into how different school environments, socio-economic conditions, and cultural practices influence mental health and academic performance. This could guide the development of globally relevant policies and programs that promote healthier habits, reduce academic stress, and create supportive learning environments for students in various settings.

To conclude, while our study has limitations in providing definitive insights and is constrained in its broader applicability, it establishes a strong foundation for understanding how lifestyle changes impact students' wellbeing. Although our initial focus was on changes prompted by COVID-19, our findings underscore the importance of examining these factors over time to uncover deeper patterns and relationships. This research serves as a starting point for future investigations into students' wellbeing in educational settings and highlights the potential for early intervention strategies to promote both mental health and academic success. By building on these insights, educators and policymakers can create more supportive environments that cater to the ever changing needs of students.

References

Chaturvedi, K. (2020, December 7). *Covid-19 and Its Impact on Students*. Kaggle. https://www.kaggle.com/datasets/kunal28chaturvedi/covid19-and-its-impact-on-students

Chen, M. Y., Wang, E. K., & Jeng, Y. J. (2006). Adequate sleep among adolescents is positively associated with health status and health-related behaviors. *BMC public health*, *6*, 1-8.

Meldrum, R. C., & Restivo, E. (2014). The behavioral and health consequences of sleep deprivation among US high school students: relative deprivation matters. *Preventive medicine*, *63*, 24-28.