INTRODUCTION TO MACHINE LEARNING COURSE

DREAMS & GRADES

PREPARED BY :
ZIED ELLOUZI, ASSIL TEKAYA, YASMINE ALAYA

Introduction

Motivation :

This study was motivated by a growing concern we observed among students: despite spending long hours studying, many experience difficulties maintaining concentration, managing stress, and achieving consistent academic performance. Through personal experiences, sleep-related issues, such as insufficient sleep, irregular sleep schedules, and poor sleep quality appeared frequently and were often associated with fatigue, reduced focus, and decreased motivation during the day.

"Sleep is not optional. It's a biological necessity that affects every aspect of our health, performance, and safety."
Dr. Charles A. Czeisler, Harvard Medical School.

These observations led us to question whether sleep could be a key factor influencing academic performance.

Ref: - The consequences of sleep deprivation on cognitive performance - PMC
- Role of Sleep in Memory Consolidation
- Sleep duration and subject-specific academic performance among adolescents in China | npj Science of Learning

Scientific Question :

How does sleep (duration and quality) influence academic performance among students? and how do factors such as (stress, social media use, caffeine consumption…) can affect it ?

This study aims to provide benefits at multiple levels:

- For students: helping them improve daily habits and lifestyle choices to achieve better academic performance and results.
- For professors: helping identify factors that affect students' concentration, learning abilities, and grades.
- For educational institutions: enabling early detection of academic risks and the implementation of timely support and alert systems

Data and Variables :

To address our scientific question, we first focused on identifying the key variables that are most relevant to both sleep and academic performance. Our variable selection was guided by two main considerations:

(1) findings from existing scientific literature, and
(2) real-life factors that are common and modifiable in students' daily routines.

Since academic performance is influenced by cognitive processes such as attention, memory, and decision-making, sleep naturally emerged as a central variable. However, sleep itself is not an isolated phenomenon; it is affected by multiple behavioral and psychological factors. Therefore, our approach was to identify the most influential and well-documented determinants of sleep among students.

Based on prior research, we selected **sleep duration** and **sleep quality** as the core sleep-related variables. Sleep duration captures the quantitative aspect of sleep, while sleep quality reflects how restorative and effective that sleep is. Together, these two variables provide a comprehensive representation of a student's sleep pattern.

To explain variations in sleep duration and sleep quality, we focused on three commonly cited factors in student populations: **screen time**, **stress level**, and **caffeine intake**. These variables were chosen because they are highly prevalent in academic environments, have strong theoretical and empirical links to sleep disruption, and represent behaviors that can potentially be modified or regulated.

- **Screen time** was selected due to the widespread use of digital devices among students and its known impact on circadian rhythms and sleep onset.
- **Stress level** was included because academic pressure and psychological stress are known to directly interfere with sleep regulation and recovery.
- **Caffeine intake** was chosen as it is commonly used by students to cope with fatigue, despite its stimulating effects on sleep.

Finally, **academic performance** was chosen as the target variable, as it represents a measurable outcome that reflects learning efficiency and cognitive functioning. By modeling academic performance as a continuous variable, we are able to study both linear and non-linear effects of sleep-related factors.

This structured selection of variables allows us to build a coherent model in which behavioral factors influence sleep, and sleep in turn affects academic performance.

Ref: - Why Gen Z can't sleep: The shocking link between social media, stress, and sleepless nights
- Factors Affecting Sleep - Internal & External Factor

Data Exploration :

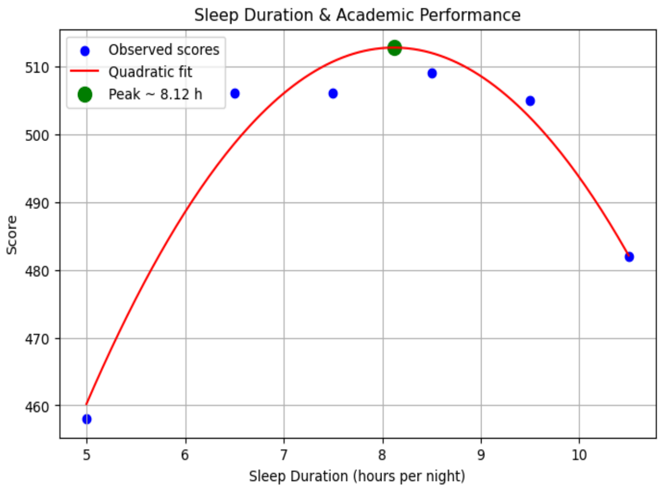
After defining the variables required to answer our scientific question, we conducted an initial data exploration phase. The objective of this step was not to build the final model, but rather to understand how these variables behave in real-world contexts and to verify whether the relationships assumed in our conceptual model are supported by existing data and scientific evidence.

To do this, we relied on two complementary sources:
(1) scientific articles, to understand large-scale and validated relationships, and
(2) public datasets, to explore concrete data patterns and empirical dependencies between variables

Exploration Based on Scientific Literature :

We first explored peer-reviewed articles from the National Library of Medicine to study the relationship between sleep and academic performance. In particular, we relied on a large-scale study involving 54,102 students, which examined the association between sleep duration and standardized mathematics scores.We extracted the results from the article and visualized them, which revealed a U-shaped relationship between sleep duration and academic performance. This visualization gave us a idea about the nature of the relationship. The extracted data are as follows:

Sleep Duration	Original Math Score
<6 h	458
6–7 h	506
7–8 h	506
8–9 h	509
9–10 h	505
>10 h	482



In addition, we reviewed other studies focusing on sleep quality and academic achievement. Based on findings from multiple theses, students were categorized using PSQI (Pittsburgh Sleep Quality Index) scores:

Good – PSQI ≤ 5 —Higher GPA
Moderate — PSQI 5–8 — Medium GPA
Poor — PSQI ≥ 8 — Lower GPA

The evidence indicates a positive relationship between sleep quality and academic performance: students with better sleep quality achieve higher academic outcomes

Exploration Based on Real Datasets :

To complement the literature review, we explored three publicly available datasets from Kaggle that contained information related to students' daily habits and sleep patterns:

500 sample

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Student_ID	Age	Gender	University_Ye	Sleep_Duratio	Study_Hours	Screen_Time	Caffeine_Intak	Physical_Acti	Sleep_Quality	Weekday_Slee	Weekend_Slee	Weekday_Slee	Weekend_Sleep_End	
2	1	24	Other	2nd Year	7.7	7.9	3.4	2	37	10	14.16	4.05	7.41	7.06	
3	2	21	Male	1st Year	6.3	6	1.9	5	74	2	8.73	7.1	8.21	10.21	
4	3	22	Male	4th Year	5.1	6.7	3.9	5	53	5	20	20.47	6.88	10.92	
5	4	24	Other	4th Year	6.3	8.6	2.8	4	55	9	19.82	4.08	6.69	9.42	
6	5	20	Male	4th Year	4.7	2.7	2.7	0	85	3	20.98	6.12	8.98	9.01	
7	6	25	Other	1st Year	4.9	12	3.2	3	96	9	9.8	18.83	5.04	10.51	
8	7	22	Female	2nd Year	6.5	11.7	3.4	1	99	6	13.05	20.96	8.58	10.81	
9	8	22	Male	2nd Year	6.1	7.8	3	1	108	4	10.49	10.85	5.6	10.02	
10	9	24	Female	1st Year	8.6	2.4	1.4	1	86	7	11.06	18.88	8.14	8.78	

This Dataset contains : - Sleep Duration
- Screen Time
- Caffeine
- Sleep Quality

But doesn't contain : - Stress Level
-Exam scores

374 sample

Person ID	Gender	Age	Occupation	Sleep Duratio	Quality of Slee	Physical Acti	Stress Level	BMI Category	Blood Pressur	Heart Rate	Daily Steps	Sleep Disorder
1	Male	27	Software Eng	6.1	6	42	6	Overweight	126/83	77	4200	None
2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
4	Male	28	Sales Repres	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
5	Male	28	Sales Repres	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
6	Male	28	Software Eng	5.9	4	30	8	Obese	140/90	85	3000	Insomnia
7	Male	29	Teacher	6.3	6	40	7	Obese	140/90	82	3500	Insomnia
8	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
9	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None

This Dataset contains : - Sleep Duration
- Sleep Quality
- Stress Level

But doesn't contain : - Caffeine
- Screen Time
- Exam score

1000 sample

student_id	age	gender	study_hours_p	social_media_	netflix_hours	part_time_job	attendance_p	sleep_hours	diet_quality	exercise_freq	parental_educ	internet_quali	mental_health	extracurricula	exam_score
S1000	23	Female	0	1.2	1.1	No	85	8	Fair	6	Master	Average	8	Yes	56.2
S1001	20	Female	6.9	2.8	2.3	No	97.3	4.6	Good	6	High School	Average	8	No	100
S1002	21	Male	1.4	3.1	1.3	No	94.8	8	Poor	1	High School	Poor	1	No	34.3
S1003	23	Female	1	3.9	1	No	71	9.2	Poor	4	Master	Good	1	Yes	26.8
S1004	19	Female	5	4.4	0.5	No	90.9	4.9	Fair	3	Master	Good	1	No	66.4
S1005	24	Male	7.2	1.3	0	No	82.9	7.4	Fair	1	Master	Average	4	No	100
S1006	21	Female	5.6	1.5	1.4	Yes	85.8	6.5	Good	2	Master	Poor	4	No	89.8
S1007	21	Female	4.3	1	2	Yes	77.7	4.6	Fair	0	Bachelor	Average	8	No	72.6
S1008	23	Female	4.4	2.2	1.7	No	100	7.1	Good	3	Bachelor	Good	1	No	78.9

This Dataset contains : - Sleep Duration
- Screen Time
- Exam scores

But doesn't contain : - Caffeine
- Sleep Quality
- Stress Level

=> The scientific articles and real datasets explored previously helped us identify the main correlations and dependencies between the variables. However, these sources could not be directly used for our analysis since the datasets were separate, had limited sample sizes, and each contained only a subset of the required variables. None of them included all the information needed to represent our complete model.

For this reason, we decided to simulate our own dataset, using the relationships and patterns observed in the literature and exploratory analyses, in order to build a coherent and complete dataset suitable for our study

Data generation :

Throughout the data generation process, we consulted multiple scientific references and online sources to inform our decisions. In several cases, we directly adopted values supported by specific research articles. In other instances, where the literature provided a range of values, we synthesized insights from various studies to determine realistic and balanced parameter choices for our model.

Independent variables :

Screen time

```
# Screen time (lognormal → skewed distribution)
screen_time = np.random.lognormal(mean=1.5, sigma=0.5, size=N)
screen_time = np.clip(screen_time, 0, 10)   # clamp for realism
```

Choice of distribution (Log-normal) : We modeled daily screen time using a log-normal distribution. Indeed, usage durations typically show strong asymmetry: many people have moderate use, while a minority of “heavy users” accumulate very high durations, which produces a long right tail A log-normal distribution reflects this reality well, because it generates a mean higher than the median, reproducing a small number of individuals concentrating most of the screen time (researchgate.net).

a -The distribution of screen time per person fits a log-normal... | Download Scientific Diagram

Chosen parameters : Screen time is generated with a log-normal mean μ = 1.5 and σ = 0.5 (on the log scale) and then bounded between 0 and 10 hours. These parameters produce a median around ~4.5 hours and a mean ~5 hours per day, consistent with current data (for example, the global average screen time is about 6 hours 40 minutes per day Alarming Average Screen Time Statistics (2025)). Bounding at 10 hours ensures we remain within realistic limits: although some populations (notably Gen Z adolescents) can reach an average of ~9 hours per day, exceeding 10 hours remains exceptional. This upper threshold also prevents the long tail of the log-normal from generating unrealistic extreme values. In summary, the log-normal distribution with these parameters makes it possible to obtain a moderate average screen time while reflecting a wide variation (a few individuals spending very long durations in front of screens, in line with real-world observations).

Caffeine intake :

```
# Caffeine intake (Poisson distribution)
caffeine_intake = np.random.poisson(lam=2, size=N)
caffeine_intake = np.clip(caffeine_intake, 0, 6)
```

Choice of distribution (Poisson) : Daily caffeine consumption (expressed as the number of coffees or caffeinated beverages) is modeled as a discrete variable following a Poisson distribution (λ = 2). This choice is explained by the fact that it is a count variable (number of cups per day): the Poisson distribution is a classic model for events counted over a fixed period, assuming they occur relatively independently. : With λ = 2, we obtain an average of 2 cups per day, which corresponds to a realistic order of magnitude for example, American coffee drinkers consume on average ~3 cups per day(More Americans Drink Coffee Each Day Than Any Other Beverage, Bottled Water Back in Second Place), so if we include those who do not drink coffee, an overall population average around 2 cups is plausible. The shape of the Poisson distribution (many people at 0, 1, or 2 coffees, and increasingly rare beyond that) matches reality, where the majority consumes moderately, while a minority consumes large quantities.

Parameters and bounds : After generation, consumption is bounded between 0 and 6 per day. This means we limit the maximum to 6 units of caffeine. This choice is motivated by biological plausibility considerations: although extreme cases exist, it is very rare to see individuals regularly consume more than 5 or 6 caffeinated beverages per day. Studies show that approximately 95% of coffee drinkers consume fewer than 5 cups per day (Coffee Drinking Is Widespread in the United States, but Usual Intake Varies by Key Demographic and Lifestyle Factors - PMC). Setting the upper bound at 6 therefore makes it possible to cover the highest cases without exceeding commonly observed behaviors. In summary, the Poisson distribution with $\lambda = 2$ provides a realistic distribution of caffeine consumption (many values between 0–2 cups, some cases at 3–4, and very rarely 5–6), and the threshold at 6 avoids outlier values while encompassing the vast majority of consumers

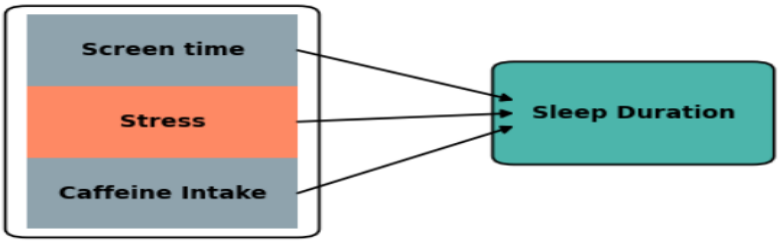
Stress level :

```
# Stress level (normal distribution then clamped from 1 to 10)
stress_level = np.random.normal(loc=5, scale=2, size=N)
stress_level = np.clip(stress_level, 1, 10)
```

Choice of distribution (Truncated normal) : Stress level (on a scale from 1 to 10) is modeled using a normal distribution with mean 5 and standard deviation 2, then truncated to the interval [1, 10]. This choice reflects the assumption that, within a large student population, perceived stress is distributed around a moderate average value, with fewer individuals at extremely low or extremely high levels. Psychological surveys confirm that, on average, the adult population is around 5 out of 10 on stress scales (for example, the American Psychological Association reported an average stress level of ~5/10(Stress in America 2022: Concerned for the future, beset by inflation" On a scale of 1 to 10, where 1 means you have "little or no stress" and 10 means you have "a great deal of stress," the average reported level of stress during the past month among all adults was 5.0, which has held steady since 2020")). A normal distribution centered at 5 with $\sigma = 2$ implies that approximately 68% of students have a stress level between 3 and 7 (i.e., "moderate" stress), which aligns well with general observations for instance, a study among medical students found that nearly half of participants had "normal/moderate" stress and only ~12% experienced severe stress (Distribution of stress levels. | Download Scientific Diagram). This shape, both symmetric and concentrated around the mean, reflects a situation where most students experience moderate stress, while only a few experience very low or very high stress, in line with field statistics.

Scale bounds (1 to 10): Stress is naturally limited to values from 1 (minimal) to 10 (extreme), since this is the measurement scale. Truncating the normal distribution to these bounds ensures that no out-of-scale values are generated. This avoids, for example, random draws producing stress levels of "0" or "11," which are impossible under the definition of the score. Moreover, values near the bounds remain rare (a truncated N(5,2) distribution has very few points above 10 or below 1), which matches the idea that it is uncommon to have absolutely no stress or maximal stress, only a small number of individuals reach the extremes of the scale.

Dependent variables:



Sleep Duration :

```
# Base sleep duration (normally distributed)
sleep_duration = np.random.normal(loc=7.5, scale=0.7)
```

We modeled the baseline nightly sleep duration as a normal distribution with mean 7.5 hours and standard deviation 0.7 hours to reflect typical adult sleep patterns. First, a mean around 7.5 h is justified by population sleep statistics: young adults generally average about 7–8 hours of sleep per night ("In terms of total sleep time, normal adult sleep appears to be around 7.5 hrs of sleep at age 20 declining to 6.5 hrs of sleep at age 60. After age 60 things seem to level off" What Is Normal Sleep For An Adult?). Second, a standard deviation of 0.7 h (≈42 minutes) represents realistic variability in individual sleep times. Large-scale surveys find that while the global average is ~7 h, the typical variation is on the order of 1 hour or less (Reported sleep duration reveals segmentation of the adult life-course into three phases - PMC). By choosing 0.7 h, we capture a moderate spread (most individuals ±1 hour of the mean) suitable for a relatively homogeneous group (e.g. students). This yields a plausible baseline dataset where most nights are near 7–8 hours, with fewer extreme short or long sleep values.

Influences of independent variables on Sleep Duration:

Stress Level effect:

```
# Adjust for stress level
if stress_level >= 8:
    sleep_duration -= np.random.uniform(0.8, 1.5)
elif stress_level <= 3:
    sleep_duration += np.random.uniform(0.5, 1.0)
```

Psychological stress has a well-documented impact on sleep duration. High stress levels (e.g. stress score > 7 on a 10-point scale) are associated with difficulty sleeping and reduced total sleep. Under chronic high stress, the body's cortisol-driven arousal can lead to insomnia or fragmented sleep, often cutting sleep by ~1–2 hours in extreme cases. Studies confirm that as stress increases, sleep duration tends to decrease markedly Is Stress Affecting My Sleep? - Baptist Health. Conversely, low stress (score < 3) is linked to more relaxed, restorative sleep. Individuals with minimal stress often sleep longer, we allow an increase of ~0.5–1.0 h because they experience fewer worry-induced awakenings or sleep disruptions How Stress Affects Sleep | Sleep.ai . This rule reflects the observed continuum where lower stress correlates with longer sleep, and high stress can severely curtail sleep time.

Caffeine Intake effect:

```
# Adjust for caffeine intake
if caffeine_intake >= 4:
    sleep_duration -= np.random.uniform(0.5, 1.0)
```

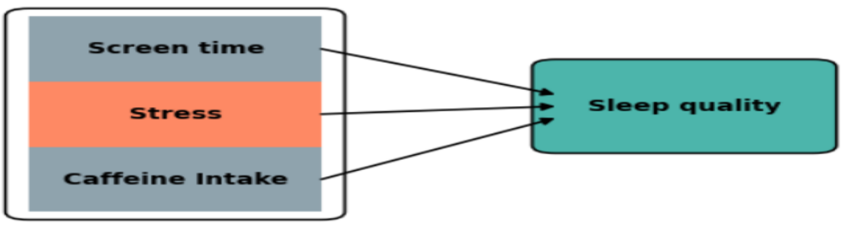
Caffeine is a stimulant that delays sleep onset and reduces sleep duration, especially at higher doses. We impose a penalty of −0.5 to −1.5 hours if daily caffeine intake is heavy (≥ 3 caffeinated drinks). This is supported by research showing that consuming caffeine late in the day or in large quantities significantly shortens sleep. For example, a meta-analysis found that caffeine use can cut total sleep time by roughly 45 minutes on average ("The study authors reported that caffeine consumption was associated with approximately 45.3 minutes less of total sleep time when compared to the control group. There was a significant influence of timing of consumption and final dose of caffeine on total sleep time, according to the investigators" Study: Coffee Consumption Should Occur 9 Hours Before Sleep, Reducing Disturbances | Pharmacy Times), even when consumed 6 hours before bed. Three or more cups of coffee (≈300+ mg of caffeine) is likely to have an even larger effect, in line with our 0.5–1.5 h reduction range.

Screen Time effect:

```
# Adjust for screen time
if screen_time > 6:
    sleep_duration -= np.random.uniform(0.3, 1.0)
```

Prolonged screen exposure (particularly in the evening) is widely tied to shorter sleep duration. We therefore reduce sleep by 0.5 to 1.0 hours for individuals with very high screen time (> 6 hours per day). The rationale is that excessive screen use (TV, smartphones, etc.) before or at bedtime can delay sleep onset and displace time that would otherwise be spent sleeping. Empirical studies have found that adolescents using screens ≥6 h per day slept significantly less on the order of tens of minutes shorter – than those with minimal screen time Screen use and sleep duration and quality at 15 years old: Cohort study - PubMed. One large survey of young adults even showed that each additional hour of screen use at bedtime can shorten sleep by ~24 minutes on average Just One Hour of Bedtime Screen Use Increases Insomnia Risk by 59%. Our chosen decrement of 0.5–1.0 h for >6 h screen time is consistent with these findings, allowing that heavy device use (especially late-night social media, gaming, or videos) can easily rob someone of around half an hour to an hour of sleep.

Sleep quality :



```python
# sleep quality
sleep_quality = np.random.normal(loc=7.0, scale=1.0, size=N)
sleep_quality = np.clip(sleep_quality, 1, 10)
```

Choice of distribution and parameters : Baseline subjective sleep quality is modeled using a normal distribution with mean 7.0 and standard deviation 1.0 on a 1–10 scale. This choice is motivated by both statistical and empirical considerations: sleep quality results from the aggregation of multiple independent factors (e.g., stress, environment, health, daily habits), and according to the Central Limit Theorem, such combined influences tend to produce approximately normally distributed outcomes. Empirical studies on student and young adult populations consistently report average self-rated sleep quality values around 6–7 with standard deviations close to 1–1.5, supporting the selection of $\mu \approx 7$ and $\sigma \approx 1$ for a typical 18–25-year-old cohort (Analysis of Medical Student Sleep Metrics using Wrist Actigraphy - OSU Center for Health Sciences Research Profiles). A mean of 7 reflects "moderately good" sleep, consistent with the fact that many students obtain around the recommended 7 hours of sleep per night. The standard deviation of 1.0 implies that most observations fall between 6 and 8, capturing realistic night-to-night variability while limiting extreme values. Because sleep quality is inherently measured on a bounded 1–10 scale, values are clipped to this interval to enforce realistic limits and prevent impossible ratings. This results in a truncated normal distribution that preserves natural variability while respecting the scale's floor and ceiling constraints.

Stress effect:

```python
#  Stress effect
if stress_level[i] >= 7:
    sleep_quality[i] -= np.random.uniform(1.0, 2.0)
elif stress_level[i] <= 3:
    sleep_quality[i] += np.random.uniform(0.2, 0.8)
```

Stress level is incorporated as a major determinant of sleep quality, with asymmetric effects for high and low stress. When stress reaches high levels (≥7 on a 1–10 scale), sleep quality is penalized by −1.0 to −2.0 points, reflecting extensive evidence that elevated perceived stress strongly degrades sleep among students. Empirical studies show that stress can account for nearly one quarter of the variance in sleep quality and that highly stressed students are more than twice as likely to experience poor sleep, due to stress-induced hyperarousal and increased cortisol levels that disrupt sleep onset and continuity (PDF) Associations between Stress, Anxiety, Depression and Sleep Quality among Healthcare Students. The threshold of 7 corresponds to the upper third of stress distributions, commonly classified as "high stress" in psychological scales. Conversely, very low stress levels (≤3) are associated with a modest improvement in sleep quality (+0.2 to +0.8), as reduced cognitive and emotional arousal facilitates more restorative sleep. This positive effect is intentionally smaller, acknowledging that while low stress supports better sleep, improvements are limited by physiological constraints. Overall, these thresholds and adjustment magnitudes reflect a realistic, evidence-based, and graded relationship between stress and subjective sleep quality.

"Stress not only significantly predicts sleep quality but also significantly affects sleep quality through rumination, emotion-focused coping strategies, and smartphone dependence as independent mediators. Additionally, stress influences sleep quality through both dual-mediation

and triple-mediation paths." **[Frontiers | The impact of stress on sleep quality: a mediation analysis based on longitudinal data]**

screen time effect:

```
# Screen time effect
if screen_time[i] >= 6:
    sleep_quality[i] -= np.random.uniform(0.8, 1.8)
elif screen_time[i] >= 4:
    sleep_quality[i] -= np.random.uniform(0.3, 0.8)
```

Screen time is modeled as a graded negative influence on sleep quality, with stronger penalties at higher exposure levels. When daily screen use reaches 6 hours or more, sleep quality is reduced by −0.8 to −1.8 points, reflecting strong evidence that excessive screen exposure particularly in the evening suppresses melatonin secretion, increases cognitive arousal, and significantly disrupts sleep. Empirical studies in student populations identify ~6 hours per day as a critical threshold, showing markedly poorer sleep quality scores and a sharp rise in insomnia prevalence beyond this level. For moderately high screen time (4–5 hours per day), a smaller penalty of −0.3 to −0.8 is applied, acknowledging that even intermediate exposure is associated with delayed sleep onset and reduced sleep duration, though with less severe effects than extreme use. This stepped penalty structure reflects observed nonlinear relationships between screen time and sleep, where sleep quality declines gradually at moderate levels but deteriorates more rapidly once usage becomes excessive. Overall, these thresholds and penalty magnitudes align with sleep research recommendations to limit daily screen exposure to approximately 3–4 hours to preserve sleep quality.

[Impact of screen time on sleep quality: A Cross-sectional study among medical students in south India – DOAJ]

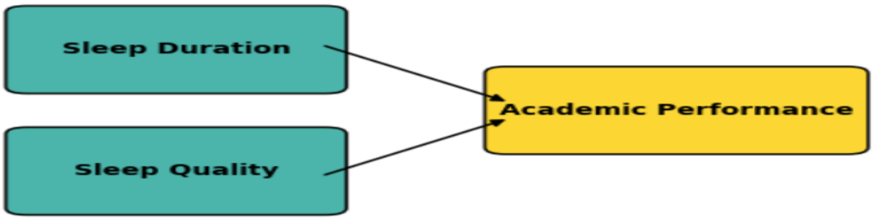
Caffeine Intake effect:

```
# Caffeine effect
if caffeine_intake[i] >= 3:
    sleep_quality[i] -= np.random.uniform(0.4, 1.0)
elif caffeine_intake[i] == 2:
    sleep_quality[i] -= np.random.uniform(0.1, 0.4)
```

High Caffeine (≥ 3 cups/day) Moderate Negative Impact: For students consuming three or more caffeinated beverages per day, we apply a −0.4 to −1.0 point penalty to sleep quality. This is consistent with literature showing that high daily caffeine intake can significantly disrupt sleep. Caffeine is a stimulant that blocks adenosine receptors, delaying sleepiness and reducing deep sleep([Caffeine and Sleep Problems]). Importantly, 3 cups of coffee (roughly 300+ mg of caffeine) are often cited as a level where sleep problems become pronounced. Sleep experts generally recommend capping daily caffeine around 300–400 mg (about 3–4 cups) and avoiding it later in the day to minimize sleep disruption.

Moderate Caffeine (2 cups/day) Slight Negative Impact: If a student has a moderate caffeine habit (about 2 cups a day), we impose a −0.1 to −0.4 penalty, a relatively minor downgrade in sleep quality. Two cups (≈200 mg caffeine) are generally considered a moderate dose that might not severely disrupt sleep for everyone, especially if taken in the morning or early afternoon. However, even this amount can have some effect, particularly for sensitive individuals or if consumed late. Research indicates that any caffeine can reduce sleep efficiency and satisfaction to a degree. A systematic review reported that on average, caffeine use (compared to none) reduced sleep efficiency by ~7% and increased sleep latency by about 9 minutes while cutting total sleep by ~45 minutes(Caffeine and Sleep Problems).

academic performance:




```
academic_performance = np.random.normal(loc=70, scale=10, size=N)
```

Academic Performance Distribution: The code begins by modeling a baseline academic performance as a normally distributed random variable with mean 70 and standard deviation 10. This choice is sensible because in large student populations, exam scores often cluster around an average (here ~70%) with fewer students scoring extremely high or low, approximating a bell-curve (normal) distribution Understanding test results | Educational Psychology. In practical terms, a mean of 70 (out of 100) reflects a typical average grade in many courses, and a standard deviation of 10 indicates moderate variability i.e. most students' baseline scores will fall between about 60 and 80, which is a realistic spread for class test results. For example, one study of medical students reported an average exam score around 69.8% (with SD ≈7.6) in a well-rested group Association Between Sleep Duration and Academic Performance in Medical Students: A Cross-Sectional Analytical Study | Journal of Heart Valve Disease, aligning with the code's assumption that ~70 is a representative mean performance before accounting for sleep effects.

Effects of Sleep Duration on Performance

```
if sleep_duration[i] < 6:
    academic_performance[i] -= np.random.uniform(10, 20)
if 7 <= sleep_duration[i] <= 8:
    academic_performance[i] += np.random.uniform(5, 15)
if sleep_duration[i] > 9:
    academic_performance[i] -= np.random.uniform(8, 18)
```

Short Sleep (<6 hours): If a student sleeps under 6 hours, the model subtracts a sizable 10–20 points from their performance. This heavy penalty is justified by research showing that very short sleep is detrimental to academic functioning. Students regularly getting less than ~6 hours per night experience a pronounced decline in academic performance "The researchers found that students who receive less than six hours of sleep experienced a pronounced decline in academic performance" (Nightly Sleep Is Key to Student Success - News - Carnegie Mellon

University). Thus, the subtraction of 10–20 points captures the substantial negative impact that chronic insufficient sleep can have on a student's grades.

Optimal Sleep (7–8 hours): If sleep is in the 7–8 hour range (often considered the optimal duration for young adults), the code adds 5–15 points to academic performance. This boost reflects the well-documented benefits of adequate sleep on learning and memory. Research consistently finds that students who obtain roughly 7–8 hours of nightly sleep achieve higher academic scores on average than those with either shorter or longer sleep "These findings suggest that optimal academic performance was associated with an average sleep duration of 6–8 hours, while both shorter and longer sleep durations were linked to lower scores (Table 2)" Association Between Sleep Duration and Academic Performance in Medical Students: A Cross-Sectional Analytical Study | Journal of Heart Valve Disease. The +5–15 point adjustment in the code acknowledges that a well-rested student can outperform their baseline ability for instance, by roughly a half to one and a half letter grade compared to if they were sleep deprived.

Oversleeping (>9 hours): If a student sleeps more than 9 hours, the model subtracts 8–18 points. While getting enough sleep is crucial, excessive sleep (beyond ~9 hours for adults) has been associated with slight performance declines. Research suggests a U-shaped relationship between sleep length and academic outcomes: both too little and too much sleep can be disadvantageous "Interestingly, our results also indicated that excessive sleep (>8 hours) was associated with a decline in academic scores" Association Between Sleep Duration and Academic Performance in Medical Students: A Cross-Sectional Analytical Study | Journal of Heart Valve Disease

Effects of Sleep Quality on Performance

```
if sleep_quality[i] < 5:
    academic_performance[i] -= np.random.uniform(5, 15)
if sleep_quality[i] > 8:
    academic_performance[i] += np.random.uniform(5, 12)
```

Poor Sleep Quality (<5): If sleep quality is very low (e.g. frequent awakenings or restless sleep), the code subtracts 5–15 points. Even if a student spends enough hours in bed, poor-quality sleep means the brain isn't getting the deep rest needed for learning. Scientific studies show that persistently poor sleepers experience more daytime fatigue, sleepiness, and worse cognitive function compared to those who sleep well Sleep quality, duration, and consistency are associated with better academic performance in college students | npj Science of Learning. By subtracting up to 15 points, the model acknowledges that a student who sleeps poorly may underperform academically for example, struggling to focus on class or recall material, leading to test scores noticeably below their rested potential.

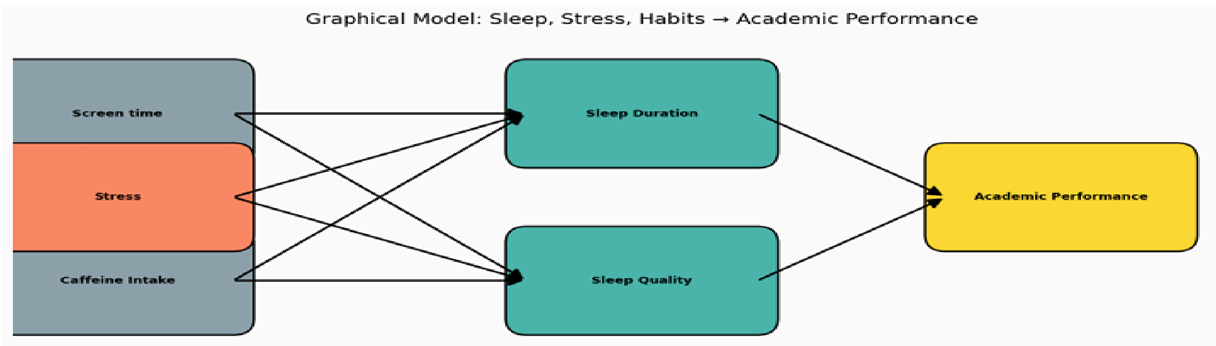
High Sleep Quality (>8): If sleep quality is excellent, the model adds 5–12 points to performance. Good sleep quality (deep, uninterrupted sleep) is known to bolster learning during quality sleep, the brain actively consolidates memories from the day, which is critical for studying retention. Research supports this boost: students who enjoy better sleep quality tend to have better academic outcomes, including higher grades Sleep quality, duration, and consistency are associated with better academic performance in college students | npj

Science of Learning. One study using wearable sleep trackers found a significant positive correlation between sleep quality scores and course performance in college students. Essentially, when a student sleeps soundly, they wake up more refreshed and mentally sharp, allowing them to absorb lessons and solve problems more effectively. The code's +5–12-point adjustment reflects this improved cognitive efficiency. Even a well-prepared student can perform a bit above their usual baseline if they are well-rested and alert, thanks to high-quality sleep facilitating optimal brain function.

Capping Performance between 0 and 100

```
# Ensure values between 0 and 100
academic_performance = np.clip(academic_performance, 0, 100)
```

Finally, after all adjustments, the code clips academic performance between 0 and 100. This simply ensures the simulated grades remain within a realistic range of 0% to 100%. In real grading systems a score cannot go below 0 (no credit) or above 100 (perfect score), so this step keeps the model's outputs plausible. It prevents any extreme adjustments from producing impossible values (for instance, a very sleep-deprived student's score will bottom out at 0 rather than going negative, and an excellent student with optimal sleep won't exceed 100). This maintains the academic realism of the simulation.



Analysis :

Objective of the Machine Learning Analysis

In this analysis, our goal is to model the relationships between lifestyle habits, sleep behavior, and academic performance using regression techniques. We explore how three behavioral factors screen time, stress level, and caffeine intake affect two key aspects of sleep: duration and quality. In turn, we examine how these two sleep-related variables influence students' academic performance.

To do this, we train three regression models:

- The first two models predict sleep duration and sleep quality based on behavioral inputs.
- The third model predicts academic performance using sleep indicators.

This stepwise modeling strategy allows us to simulate how changes in daily habits could impact academic outcomes through their effects on sleep. Special attention is given to potential non-linear effects, particularly in the relationship between sleep duration and performance. By combining interpretable regression techniques, we aim to provide actionable insights that link lifestyle, sleep, and academic success.

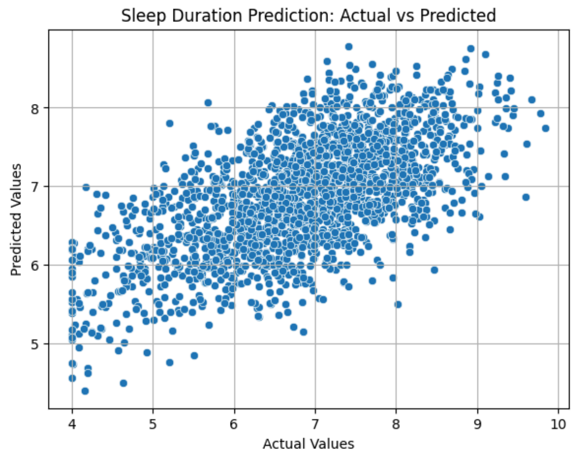
Preparation of Training and Testing Sets

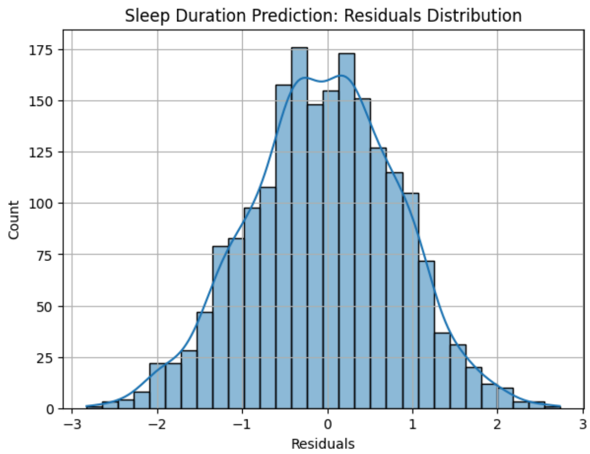
To evaluate the predictive performance of our machine learning models, we split the dataset into training and testing subsets using an 80/20 ratio. This split was applied separately for each regression task: predicting sleep duration, sleep quality, and academic performance. We used the `train_test_split` function from scikit-learn with a fixed `random_state=42` to ensure reproducibility. For the academic performance model based on sleep duration, a polynomial transformation (degree 2) was applied before splitting to capture the non-linear relationship between sleep time and grades.

Sleep duration & sleep quality

To predict both sleep duration and sleep quality based on lifestyle-related factors, we employed **Bayesian Ridge Regression** for its ability to handle multicollinearity and reduce overfitting. Since predictors such as stress level, screen time, and caffeine intake may be interrelated, Bayesian Ridge offers an effective way to regularize model parameters by incorporating prior distributions. This approach is particularly suitable for behavioral data, which tends to be noisy and prone to limited sample sizes. Using the same regression technique across both models ensures methodological consistency and improves interpretability.

Sleep Duration Model : Evaluation and Interpretation





SleepDuration_i = 9.38 − 0.27·Caffeine_i − 0.12·ScreenTime_i − 0.28·Stress_i + u_i

Model Performance

The regression model predicting sleep duration based on caffeine intake, screen time, and stress level produced the following metrics on the test set:

- **Mean Squared Error (MSE):** 0.75
- **Root Mean Squared Error (RMSE):** 0.86
- **R^2 (Coefficient of Determination):** 0.39

These values indicate that the model captures a significant portion of the variance in sleep duration, though not all. An R^2 of ~0.39 suggests that approximately 39% of the variation in sleep duration is explained by the three lifestyle-related features.
The **cross-validation results** confirm the model's consistency:

- **Average RMSE:** 0.87
- **Average R^2:** 0.40

These close values between test and cross-validation errors suggest low overfitting and good generalizability.

Coefficients Interpretation

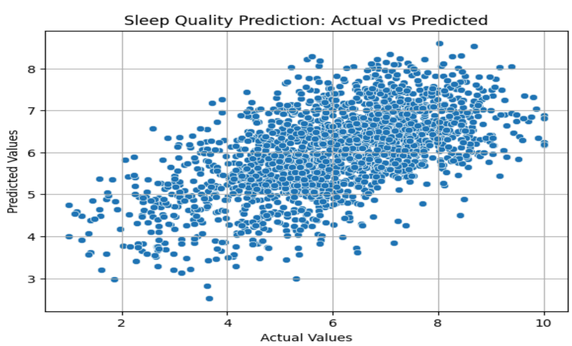
- **Caffeine Intake (−0.275):** Higher caffeine consumption is associated with shorter sleep duration, which aligns with established literature.
- **Screen Time (−0.120):** More screen time modestly reduces sleep, likely due to delayed sleep onset from blue light exposure.
- **Stress Level (−0.273):** Increased stress significantly reduces sleep duration, consistent with psychological findings.
- **Intercept (9.38):** This represents the predicted sleep duration for a student with zero stress, caffeine, and screen time—an idealized baseline.

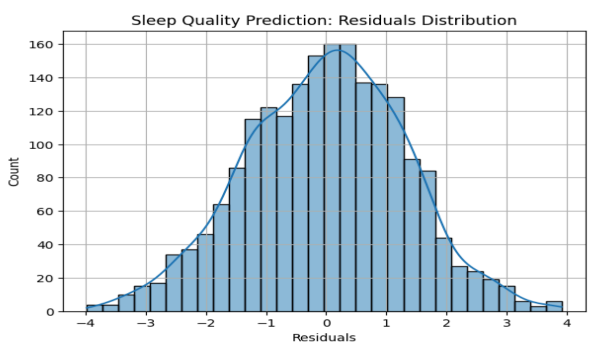
Visual Evaluation

- **Actual vs. Predicted Plot:** The scatter plot shows a reasonably strong linear trend along the diagonal, indicating a good fit, though some spread remains.
- **Residuals Histogram:** The residuals follow a bell-shaped, approximately normal distribution centered at zero, which supports the assumptions of the regression model.

→ The model performs reasonably well, demonstrating that caffeine, screen use, and stress are meaningful predictors of sleep duration among students. While there is room for improvement, the metrics and visualizations suggest that the model is reliable and valid for capturing broad trends in sleep behavior.

Sleep Quality Model : Evaluation and Interpretation





$$SleepQuality\_i = 9.58 - 0.13 \cdot Caffeine\_i - 0.32 \cdot ScreenTime\_i - 0.34 \cdot Stress\_i + u\_i$$

Model Performance

The regression model predicting sleep quality based on caffeine intake, screen time, and stress level produced the following metrics on the test set:

- **Mean Squared Error (MSE):** 1.75
- **Root Mean Squared Error (RMSE):** 1.32
- **R^2 (Coefficient of Determination):** 0.35

These results suggest a moderately strong model. The R^2 score indicates that around 35% of the variability in sleep quality is explained by the selected predictors.

The cross-validation results are in close agreement:

- **Average RMSE:** 1.33
- **Average R^2:** 0.36

This consistency suggests that the model generalizes well and does not overfit.

Coefficients Interpretation

- **Caffeine Intake (−0.13):** A modest negative effect, consistent with literature showing that caffeine disrupts sleep architecture and latency.
- **Screen Time (−0.32):** Stronger negative effect, likely due to blue light exposure and overstimulation reducing perceived sleep quality.
- **Stress Level (−0.34):** Highest absolute impact among features; elevated stress is a well-known factor reducing sleep quality.
- **Intercept (9.58):** The predicted baseline quality for someone with zero caffeine, screen use, and stress—an ideal scenario.

Visual Evaluation

- **Actual vs. Predicted Plot:** The data shows a clear upward trend, though some spread is evident, especially at the extremes. This is expected in behavioral models.
- **Residuals Histogram:** The residuals distribution is bell-shaped and centered near zero, supporting the model's assumptions and suggesting no major bias.

→ The Bayesian Ridge model provides a meaningful and interpretable way to estimate sleep quality from daily habits. The results confirm that stress, screen time, and caffeine all negatively impact sleep quality, and that the model is reasonably accurate and stable across different validation folds.

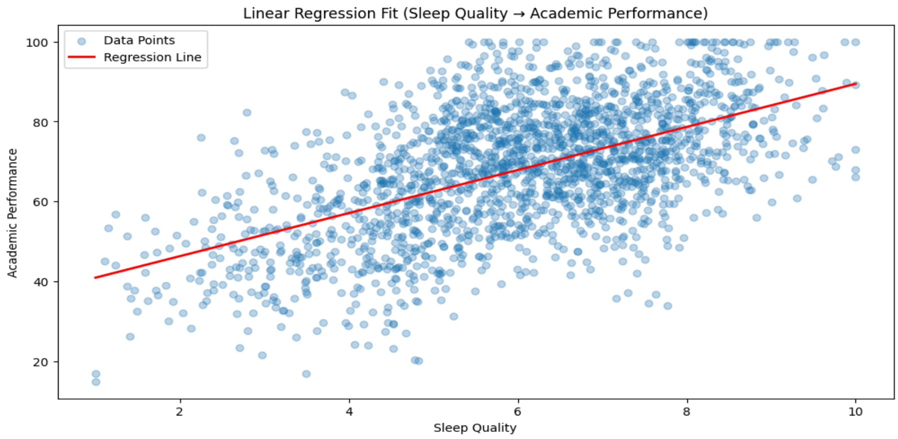
Academic Performance

Initially, we intended to use a single multiple linear regression model to predict academic performance based on both sleep duration and sleep quality. However, during exploratory analysis, we observed that the two predictors influence academic performance in distinct ways. Sleep quality tends to have a direct, approximately linear relationship with performance:

better sleep quality generally correlates with higher scores. On the other hand, sleep duration shows a non-linear (inverted U-shaped) effect. Academic performance improves with adequate sleep (around 7–8 hours) but decreases when sleep is either insufficient or excessive. To reflect these differing dynamics, we opted to use two separate models: a simple linear regression for sleep quality and a polynomial regression (degree 2) for sleep duration. This approach allows us to capture each variable's effect more accurately.

Linear Regression Model: Sleep Quality → Academic Performance

To explore how sleep quality influences academic performance, a simple linear regression model was implemented. The assumption behind this choice is based on the consistent linear trend observed between these two variables: as sleep quality improves, academic performance tends to increase proportionally. This allows us to estimate the direct impact of sleep quality without introducing higher model complexity.



AcademicPerformance_i = 35.50 + 5.39 · SleepQuality_i + u_i

Model Results:

- **MSE (Mean Squared Error)**: 173.61
- **RMSE (Root Mean Squared Error)**: 13.18
- **R^2 Score**: 0.317

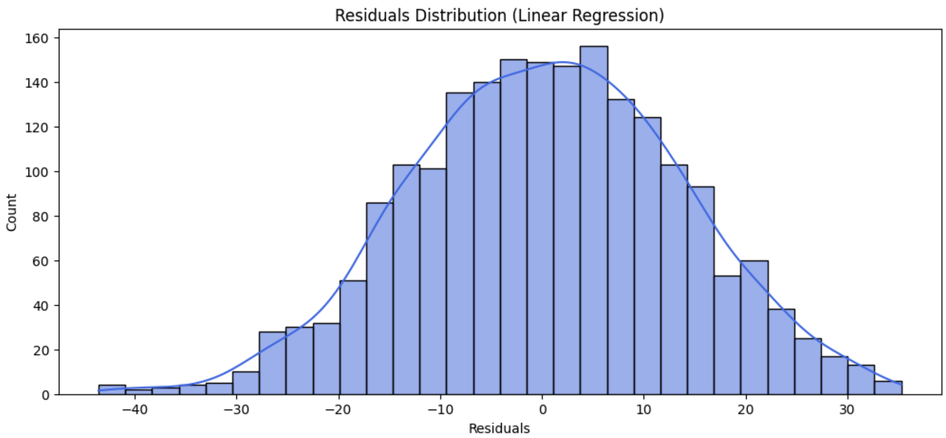
These metrics indicate that the model explains approximately 31.7% of the variance in academic performance, which is acceptable for behavioral data that often involve high variability and unobserved influences.

Model Interpretation:

- **Intercept**: 35.50
- **Coefficient (sleep_quality)**: 5.39

The coefficient suggests that for every 1-point increase in sleep quality (on a scale of 1–10), academic performance is predicted to increase by approximately 5.39 points (on a scale of 0–100), holding all else constant.

Residuals Interpretation :



The residual distribution closely resembles a normal distribution centered around zero, indicating that the model's errors are randomly distributed, which satisfies one of the key assumptions of linear regression. There is no obvious skew or pattern suggesting bias.

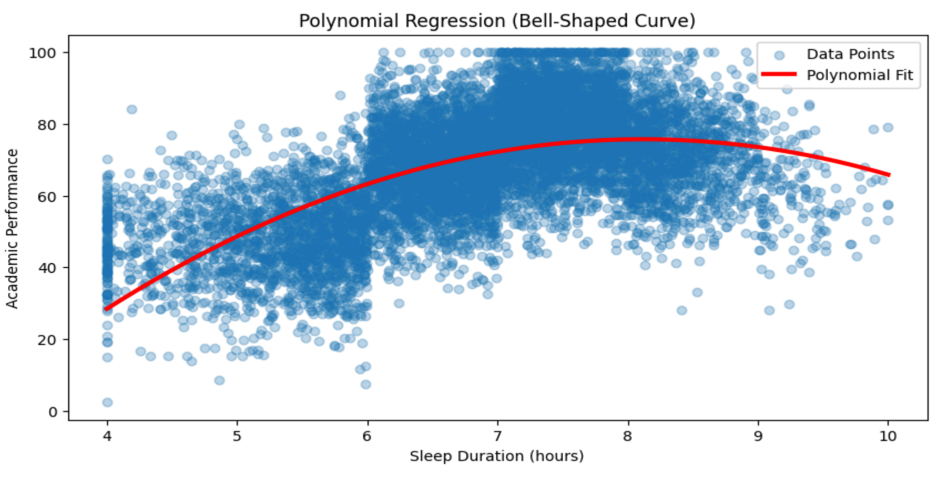
Cross-Validation:

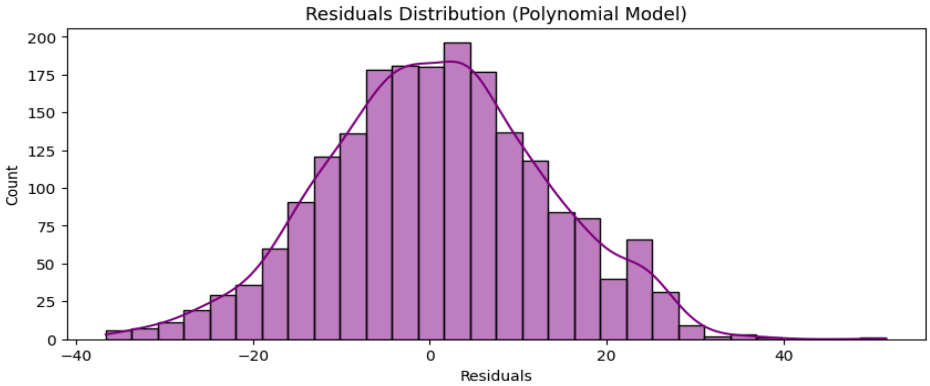
- **Average MSE**: 181.66
- **Average RMSE**: 13.48
  Average R²: 0.309

Cross-validation results are consistent with the initial model metrics, indicating that the model generalizes well across different data splits.

⟶ The linear regression model confirms a moderate but significant positive relationship between sleep quality and academic performance. While the model does not explain all variability (as expected in real-world behavioral data), the results are statistically meaningful and support the hypothesis that improving sleep quality can contribute to better academic outcomes

Polynomial Regression Analysis: Sleep Duration and Academic Performance.





$AcademicPerformance_i = -108.16 + 45.30 \cdot SleepDuration_i - 2.79 \cdot SleepDuration_i^2 + u_i$

To capture the non-linear relationship between sleep duration and academic performance, a polynomial regression model of degree 2 was implemented. This choice was motivated by prior evidence suggesting that both insufficient and excessive sleep may negatively impact academic outcomes, resulting in a U-shaped (or bell-shaped) curve. The model includes a squared term for sleep duration to capture this curvature effectively.

Model Performance Metrics:

- **MSE:** 156.99
- **RMSE:** 12.53
- **R^2:** 0.3827

Cross-Validation Results (5 folds):

- **Average MSE:** 163.34
- **Average RMSE:** 12.78
- **Average R^2:** 0.3781

These results indicate that the polynomial model explains approximately 38% of the variance in academic performance. While not exceptionally high, this R^2 is slightly stronger than that of the linear model with sleep quality, confirming the relevance of a non-linear fit for sleep duration.

Model Coefficients:

- **Sleep Duration Coefficient:** 45.30
- **Sleep Duration² Coefficient:** –2.79
- **Intercept:** –108.16

The positive coefficient on sleep duration and negative coefficient on its square validate the hypothesized bell-shaped effect. Academic performance tends to increase with sleep duration up to a certain optimal point, beyond which it begins to decline.

Visual Interpretation:

- The scatter plot with the red regression curve illustrates the curved fit, highlighting the peak performance zone around moderate sleep durations.

- The residuals distribution is approximately normal, indicating a reasonably well-fitted model with minimal bias.

→ This model supports the hypothesis of an optimal sleep duration range for academic success. The quadratic structure improves predictive power compared to a linear approach and offers valuable insights for understanding how lifestyle habits like sleep quantity can affect student outcomes.

What the Model Teaches Us

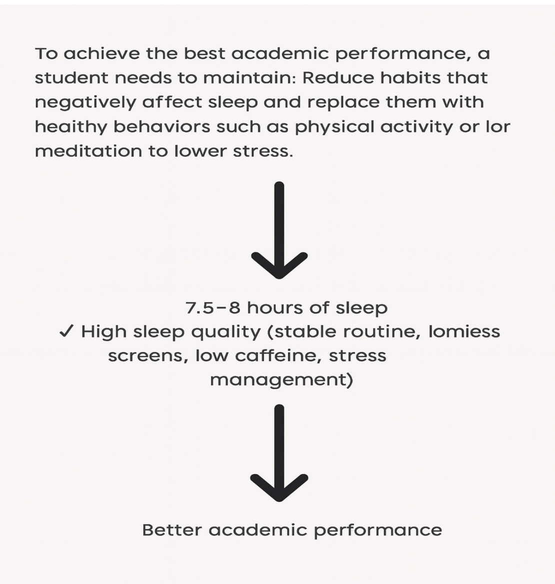
Our models tell a simple but powerful story: **sleep is the secret weapon** of academic performance.If you're a student aiming for top scores, it's not just about studying harder it's about sleeping smarter. According to our analysis, the sweet spot lies in getting **7.5 to 8 hours of sleep**, not too little and not too much. But it's not just quantity, quality matters just as much. Deep, restful sleep amplifies your cognitive functions and helps translate your effort into results.

So how do you get there?
The models uncovered **three key levers** that shape your sleep:

- **Caffeine intake** : keep it minimal. Even moderate caffeine levels can disrupt sleep cycles.
- **Screen time** : too much exposure to screens before bed delays sleep onset and reduces sleep quality. Stay under ~2 hours if you can.
- **Stress levels** : the quiet saboteur. Stress not only shortens your sleep but also makes it less restorative.

```
 ranges for 7.5–8 hours of sleep:
Caffeine intake: 0.0 → 0.5527638190954774
Screen time     : 0.0 → 1.9698492462311559
Stress level    : 0.0 → 2.713567783919598
```

By keeping these factors low (within the model-suggested thresholds), students can align themselves with the biological conditions that support peak performance.
In essence, the model draws a roadmap:



Limitations:

Although the simulation was grounded in scientific literature and real datasets, it cannot fully capture the complexity and variability of real human behavior.

Academic performance is influenced by a wide range of additional factors, such as teaching quality, learning environment, socioeconomic background, and individual motivation, which were not included in this analysis.

Similarly, sleep duration and sleep quality are themselves affected by many other variables, including physical health, mental well-being, environmental conditions, and genetic predispositions, which were beyond the scope of this study.

However, for the purpose of this project, the focus was deliberately placed on the most influential and well-documented factors in the literature, particularly those related to sleep and lifestyle, in order to provide a clear and scientifically supported understanding of their impact on academic performance.

Conclusion

Sleep plays a crucial role in students' academic performance through both its quality and duration.

The relationship between sleep quality and academic achievement is linear and proportional, with higher sleep quality being associated with better grades, confirming that students who sleep well tend to perform better academically.

In contrast, the relationship between sleep duration and academic performance is non-linear and follows a bell-shaped pattern, with optimal performance observed at around 7.5 to 8 hours of sleep per night. Insufficient sleep leads to fatigue, reduced attention, and poorer cognitive functioning, while excessive sleep is also associated with decreased academic performance.

Moreover, several behavioral and psychological factors negatively affect sleep. Higher stress levels reduce both sleep duration and sleep quality, increased screen time, particularly from social media use, worsens sleep quality and delays sleep onset, and higher caffeine consumption shortens sleep duration and degrades sleep quality.

Together, these factors indirectly impair academic performance by disrupting students' sleep.