Pre-Analysis Plan

Title
Can Sleep Patterns Help Predict Productivity?

1. Introduction
We often hear that "sleep is important," but how much can we actually say about its connection to productivity? In this project, we want to explore whether sleep-related behaviors—such as how long someone sleeps, when they go to bed, or how consistent their sleep schedule is—can help us predict how productive they'll be the next day.

In simple terms, we're asking:
Can we predict a person's productivity using their sleep data?

We're approaching this as a machine learning problem, where sleep variables will serve as input features, and productivity will be the target we try to predict.

2. What's an Observation?
Each row in our dataset represents one day of data for one individual. It includes their sleep behavior from the previous night and their productivity outcome for that day.

3. What Type of Learning Are We Doing?
This is a supervised learning task, since we're trying to predict a known outcome.

- If productivity is measured as a number (like a rating from 0 to 10), we'll use regression.
- If it's recorded as categories (like low, medium, high), we'll treat it as a classification problem.

4. Which Models Will We Use?
All of the models we plan to use come from our course materials and the GitHub repositories provided by our instructor. We're starting simple and building up:

- Linear regression: A good baseline if productivity is numeric.
- Logistic regression: If we're predicting categories.
- k-Nearest Neighbors (kNN): A model that classifies by looking at similar examples.
- Decision Trees (CART): Helps us see how different variables split the data.

Depending on how things go, we may also try:
- Artificial Neural Networks: If we need a more flexible, powerful model.
- LASSO or PCA: To reduce the number of features or handle correlated variables.

We'll be using Python and scikit-learn, following the same tools and workflows from class.

5. Feature Engineering
To prepare the data for modeling, we'll take these steps:

- Create new features, like total hours of sleep, bedtime hour, and sleep consistency.
- One-hot encodes categorical variables (like day of the week).
- Standardize numerical values, especially for distance-based models like kNN.
- Check for correlations between features—if we find strong multicollinearity, we'll use PCA or LASSO to address it.

So far, our early exploration shows that sleep duration alone might not be a strong predictor. That means we'll need to look at combinations of features and non-linear patterns.

6. How Will We Measure Success?

For regression:
- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)
- R^2 (how well our model explains the variability in productivity)

For classification:
- Accuracy
- F1 Score
- Precision and Recall
- Confusion Matrix

We'll also split the data into training and testing sets, and possibly use cross-validation to get more reliable estimates.

7. What Does "Success" Look Like?

We're not expecting to perfectly predict productivity—there's a lot going on in people's lives that we can't see from sleep data alone. But if our models perform better than a baseline (like always predicting the average), and if we can understand which sleep behaviors are most predictive, we'll consider that a success.

Also, if our results are consistent across validation sets, that will be a good sign that we're not just overfitting.

8. What Could Go Wrong?

We're aware of a few potential issues:

- Sleep might not matter that much: Maybe other factors (like stress, meetings, or caffeine) play a bigger role in productivity.
- Small dataset: If we don't have much data, some models might not work well.
- Messy target variable: Productivity is hard to measure and might be subjective.
- Correlated features: Variables like bedtime and wake-up time might be too closely related.

If things go badly, we're prepared to shift directions. For example, we might use clustering to group people by their sleep habits and see whether those groups differ in average productivity. That would move us into unsupervised learning, which is also covered in our course.

9. Sharing the Results

We'll present our findings with:
- Plots showing how predicted and actual productivity compare.
- Confusion matrices for classification models.
- Tables that compare different models' performance.
- Feature importance charts or regression coefficients to show what variables mattered most.

We'll aim for clear visualizations and straightforward explanations.

10. Conclusion

This plan sets the foundation for a thoughtful and structured analysis using techniques we've learned in class. Whether or not sleep turns out to be a strong predictor of productivity, we'll gain experience applying real data science tools to a real-world question.

We're ready to test a few ideas, learn from what works (or doesn't), and avoid falling into the trap of endlessly tweaking our models. The goal is to build something meaningful, not just something that gets a high accuracy score.