March 26, 2025

Capstone: Sleep Efficiency

Dennis Michael Noel III

012055681

C964: Computer Science Capstone

Task 2 - Parts A, B, C and D

Part A: Letter of Transmittal & Project Proposal

Letter of Transmittal

March 26th, 2025

Mike Noel
Software Engineer
Sleep Solutions
100 Sleepy Sheep Dr
Raleigh, NC 27712

Jake Reynolds
Chief Technology Officer
Sleep Solutions
100 Sleepy Sheep Dr
Raleigh, NC 27712

Subject: Proposal for Machine Learning-Based Sleep Efficiency Application

Dear Mr. Reynolds,

I am pleased to submit this proposal to develop an advanced machine-learning application to enhance Sleep Solutions' services. This innovative solution will integrate data-driven decision-making into your sleep counseling programs, offering personalized recommendations and dynamic results tailored to individual users.

This proposal outlines how our application utilizes predictive analytics to improve sleep efficiency assessments, providing a valuable tool for clients and sleep specialists. Additionally, we have included a detailed breakdown of the methodology, project timeline, funding requirements, and potential impact on stakeholders. Ethical considerations regarding data privacy and security are also addressed to ensure compliance with industry standards.

My team and I are confident that this technology will significantly enhance Sleep Solutions' ability to serve clients more precisely and effectively. We look forward to discussing this proposal further and exploring how to bring this vision to life.

Sincerely,

Mike Noel

Software Engineer

Project Proposal

Summary of the Problem

Sleep disorders and inefficient sleep patterns affect millions (National Heart, Lung, and Blood Institute, 2022). Traditional sleep counseling relies heavily on self-reported data and generalized recommendations, which may not effectively address individual needs. The absence of real-time, data-driven analysis limits the precision of current counseling methods, often leading to suboptimal sleep improvement strategies for clients.

Benefits to the Customer and Decision-Making Process

The proposed machine-learning application will enhance Sleep Solutions' services by offering predictive analytics and personalized sleep recommendations. By utilizing data-driven metrics, customers will receive tailored guidance based on their unique sleep patterns and lifestyle habits. Sleep specialists will benefit from more precise data to assist clients effectively, improving decision-making and overall service quality.

Outline of the Data Product

This machine-learning application will utilize predictive models to assess sleep efficiency and provide actionable insights. The tool will feature an interactive dashboard that allows users to input lifestyle factors and receive personalized sleep recommendations. Over time, the model will improve as more data is collected, refining predictions and enhancing accuracy.

Description of the Data Used

The application will use a dataset containing key sleep-related variables such as age, gender, sleep duration, number of awakenings, alcohol consumption, smoking status, and exercise frequency. This data will be processed and analyzed to develop accurate sleep efficiency predictions.

Objectives and Hypotheses

The primary objective of this project is to improve sleep counseling services by integrating machine

learning to deliver personalized recommendations. We hypothesize that data-driven insights will improve sleep efficiency and customer satisfaction. The model aims to identify trends and correlations between lifestyle habits and sleep quality.

Project Methodology

The project will be developed through an iterative Agile approach and follow a structured methodology, beginning with data preprocessing, exploratory data analysis, and feature engineering. A regression-based machine-learning model will be developed, trained, and validated using historical sleep data. The final model will be deployed within an interactive dashboard, allowing users to engage with real-time sleep efficiency predictions.

Funding Requirements

The estimated budget for this project includes development costs of approximately $40,000, covering machine learning model development, UI/UX design, and engineering. Hosting and maintenance costs are projected to be between $6,000 and $7,500 annually. Additionally, opportunity costs—potentially lost business without this technology—range from $20,000 to $50,000 annually.

Impact on Stakeholders

This solution will directly benefit Sleep Solutions by enhancing its service offerings and providing valuable insights for customers. Sleep specialists will have access to more precise data, improving their ability to guide clients. Customers will better understand their sleep patterns and receive tailored recommendations, leading to better sleep quality.

Ethical and Legal Considerations

Overseeing sensitive sleep-related data requires strict adherence to ethical and legal guidelines. The application will comply with HIPAA and GDPR regulations to ensure data privacy and security. All user data will be encrypted, and explicit consent will be obtained before collection. Transparency in data usage and protection measures will be prioritized to maintain user trust.

Expertise Relevant to the Solution

Our team brings experience in machine learning, data analytics, and application development. With expertise in predictive modeling and software engineering, we are well-equipped to create a robust, secure, and user-friendly solution. Our machine learning and UI/UX design knowledge ensures a seamless experience for customers and sleep specialists. This project will incorporate modern security protocols and scalable technologies to maintain reliability and performance.

Part B: Executive Summary

Sleep Solutions aims to enhance its sleep counseling services with a machine-learning application that provides data-driven insights and personalized recommendations. The primary objective is to address the limitations of traditional sleep assessments, which rely on self-reported data and generalized recommendations. By leveraging predictive analytics, this application will offer more accurate sleep efficiency assessments, improving decision-making for both users and sleep specialists.

The target customers for this product are individuals seeking to improve their sleep quality and specialists at Sleep Solutions who provide counseling services. Users will benefit from data-backed recommendations tailored to their sleep patterns, lifestyle, and habits. The interactive insights generated by the application will facilitate improved engagement and enable specialists to make more precise, data-driven decisions.

Existing solutions in sleep counseling primarily rely on generic sleep tracking tools or manual assessments, which do not provide dynamic, individualized reports. This new application will bridge the gap by integrating machine-learning models trained on real-world sleep data to deliver tailored recommendations, addressing the shortcomings of current solutions.

The development of this data product will utilize available sleep datasets, including information on sleep duration, awakenings, alcohol consumption, smoking status, and exercise frequency. These variables will be analyzed through a supervised learning model to predict sleep efficiency. Additional data will be collected to refine the model further, ensuring long-term adaptability and performance improvements.

The methodology for this project follows an iterative, structured approach through an Agile implementation methodology. This process will include data collection, preprocessing, exploratory analysis, feature engineering, model training, validation, and deployment. The model will employ

regression techniques to predict sleep efficiency and will be continuously optimized to improve accuracy. Security measures will be implemented to ensure compliance with industry data privacy regulations.

Key deliverables for this project include a functional machine-learning model, an interactive dashboard for visualizing sleep trends, and a backend infrastructure supporting secure data storage and processing. Additional deliverables will include documentation detailing the data product's design, implementation, and user instructions.

Scheduled Timeline

The project is expected to span five months, beginning with a three-week research and planning phase. Model development will take approximately six weeks, followed by four-week backend and frontend development phases. Testing and optimization will occur over a four-week period, leading to deployment and ongoing maintenance. The structured approach ensures smooth integration into Sleep Solutions' existing services, improving sleep assessments and user outcomes while maintaining operational efficiency and compliance with industry standards.

Milestone or deliverable	Duration (Weekdays)	Projected start date	Anticipated end date
Research & Planning	15 days	4/14/25	5/2/25
Model Development	30 days	5/5/25	6/13/25
Backend Development	20 days	6/16/25	7/11/25
Frontend Development	20 days	7/14/25	8/8/25
Testing & Optimization	20 days	8/11/25	9/5/25
Deployment & Maintenance	Ongoing	9/8/25	Continuous

Evaluation Plan

The effectiveness and reliability of the data product will be assessed through a structured evaluation plan. During the data preparation stage, integrity checks and exploratory data analysis will be conducted to ensure high data quality. As the model is developed, performance metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) will validate accuracy. After deployment, user and A/B testing will gauge engagement levels and assess the real-world accuracy of sleep efficiency predictions. Continuous post-implementation monitoring will track model performance, allowing for retraining and refinement based on user feedback.

Resources and Costs

The estimated project costs include software, labor, and maintenance expenses. Cloud-based machine learning services such as AWS or Google Cloud are expected to cost approximately $2,500 annually, with additional computational resources for model training estimated at $5,000. The development process will require a team of two data scientists, one machine learning engineer, and one UI/UX developer, with total labor costs reaching approximately $67,500. Deployment and ongoing maintenance, including hosting and model retraining, will require an additional $12,500 annually. The total estimated first-year cost for this project is $87,500.

Part C: Application

The submitted application includes the following components:

- **Google Colab Notebook**: The Jupyter Notebook is hosted on GitHub Gist and executed in Google Colab, enabling users to run the model directly in the browser without requiring any local setup. The notebook includes full implementation for data preprocessing, model training, evaluation, and an interactive form for generating sleep efficiency predictions.
- **Source Code**: All Python scripts for data analysis, machine learning model training, and visualizations are contained within the notebook.
- **Dataset**: The raw and cleaned dataset is sourced from the linked repository and used to train the model.
- **Visualizations**: Descriptive visualizations are embedded in the notebook to enhance understanding of the dataset and model performance.

To access and run the application, please follow the instructions in the **User Guide** section.

Google Colab Launch Link:

https://colab.research.google.com/gist/mike4278/3b2ad895ec8a1ef5528e0a7147a4b3db/sleep_efficiency.ipynb

Part D: Post-implementation Report

Solution Summary

Sleep deprivation and poor sleep efficiency impact millions of individuals, leading to negative health outcomes and decreased productivity (National Heart, Lung, and Blood Institute, 2022. Traditional sleep counseling relies heavily on subjective self-reports, which can lead to inconsistencies in recommendations. The machine learning application developed in this project addresses these challenges by integrating data-driven insights to provide personalized sleep recommendations.

The solution uses predictive analytics to assess sleep efficiency based on various lifestyle factors, including age, sleep duration, awakenings, alcohol consumption, smoking status, and exercise frequency. By utilizing machine learning models, the application identifies patterns in user behavior and generates actionable insights for improving sleep quality. Sleep specialists can use this tool to offer personalized recommendations and enhance the effectiveness of counseling sessions.

This application streamlines data collection, analysis, and visualization, ensuring that users and sleep consultants can make informed decisions backed by empirical evidence. Additionally, the solution is designed with scalability and security in mind, ensuring compliance with industry data privacy and protection standards.

Data Summary

The dataset for this project, Sleep_Efficiency.csv, was sourced from publicly available sleep studies and includes key factors influencing sleep quality. Attributes such as age, gender, sleep duration, number of awakenings, alcohol consumption, smoking status, and exercise frequency were selected to predict sleep efficiency.

The data underwent preprocessing steps to ensure accuracy and consistency, including handling missing values, encoding categorical variables, and normalizing numerical features. Smoking status was converted

into numerical form using label encoding, while sleep duration and other numerical features were standardized for improved model training. The dataset was then split into training and testing subsets to evaluate model performance effectively.

Throughout the development lifecycle, data management followed structured phases. During the design phase, key predictors were analyzed to assess the dataset's suitability for machine learning. The development phase involved feature engineering and data transformation to optimize predictive capabilities. In the maintenance phase, the data pipeline was designed to allow for future updates and continuous model retraining, ensuring long-term adaptability and improved accuracy. These structured practices ensure reliable performance, scalability, and data integrity.

Machine Learning

A linear regression model was employed to predict sleep efficiency based on key lifestyle factors. This supervised learning approach was selected for its interpretability and ability to model relationships between multiple independent variables and a continuous outcome. By analyzing patterns in sleep-related data, the model assigns weights to factors such as age, gender, sleep duration, awakenings, alcohol consumption, smoking status, and exercise frequency to generate predictions.

The model was developed using Python within Google Colab Jupyter Notebooks, utilizing the scikit-learn library for implementation. The development process included data preprocessing, where categorical variables were encoded, and numerical features were standardized to ensure consistency. Feature selection was performed using correlation analysis to identify key predictors, and the dataset was split into training (80%) and testing (20%) subsets. The model was then trained using the training data and evaluated based on performance metrics such as Mean Absolute Error (MAE) and R-squared (R^2).

Linear regression was chosen due to its balance between simplicity and effectiveness. Its interpretability allows specialists to understand how individual factors influence sleep efficiency, providing actionable insights for users. Additionally, its efficiency in training and inference makes it suitable for deployment in

a scalable application. Over time, the model can further refine its accuracy with continuous updates and new data integration, enhancing personalized sleep recommendations.

Validation

The machine learning model's performance was evaluated using two key regression metrics: Mean Absolute Error (MAE) and R^2 Score. These metrics provide insight into the model's accuracy and ability to explain sleep efficiency variations.

MAE measures the average difference between predicted and actual sleep efficiency values, offering a straightforward interpretation of prediction accuracy. Lower MAE values indicate better performance. Meanwhile, the R^2 Score quantifies how well the independent variables explain the variance in sleep efficiency, with values closer to 1.0 signifying a stronger model fit.

Following model training, evaluation was conducted using a separate test set to ensure unbiased performance assessment. The results showed that the model achieved an MAE of 0.06 and an R^2 score of 0.70, demonstrating its capability to generate reasonably accurate predictions.

Visualizations

Visualizations played a crucial role in understanding data relationships, evaluating model performance, and identifying areas for improvement. These insights guided feature selection and refinement of the predictive model.

A scatter plot comparing predicted vs. actual sleep efficiency values was created to assess model accuracy. Ideally, data points align closely along the $y = x$ line, indicating strong predictive performance. A histogram is generated to display the frequency of various sleep efficiency results. A bar plot is included to indicate which factors most influence the sleep efficiency score, allowing specialists and users to identify habits that most impact the patient's sleep efficiency. Several boxplots are included to show

the relationship between identified important features and sleep efficiency, including awakenings, alcohol consumption, smoking status, and exercise frequency.

These visualizations provide a comprehensive understanding of how key factors influence sleep efficiency while highlighting the model's strengths and potential areas for refinement. Screenshots and graphical outputs from the code execution are included below to support these findings.

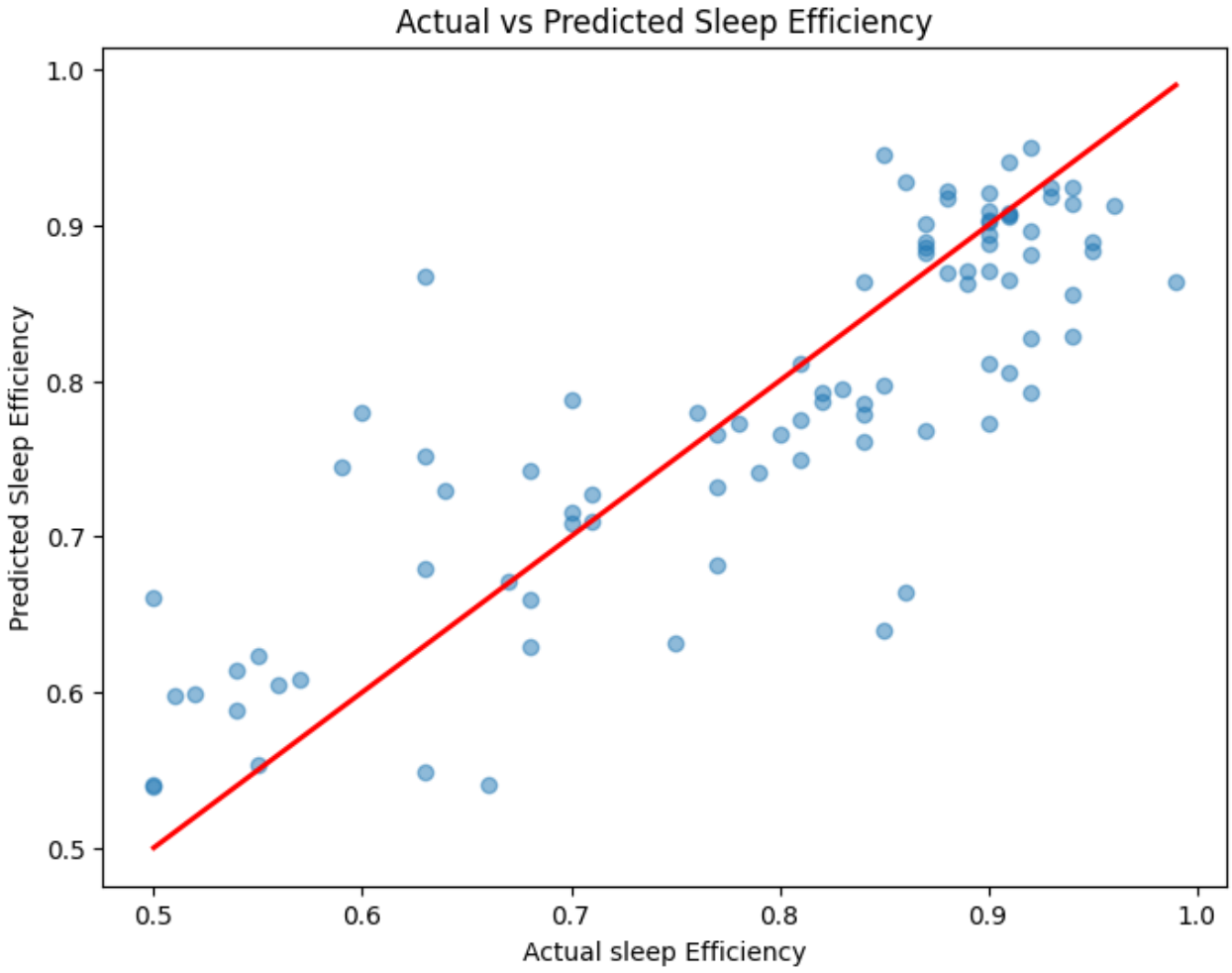

Figure 1 - Scatter Plot of Actual vs. Predicted Results

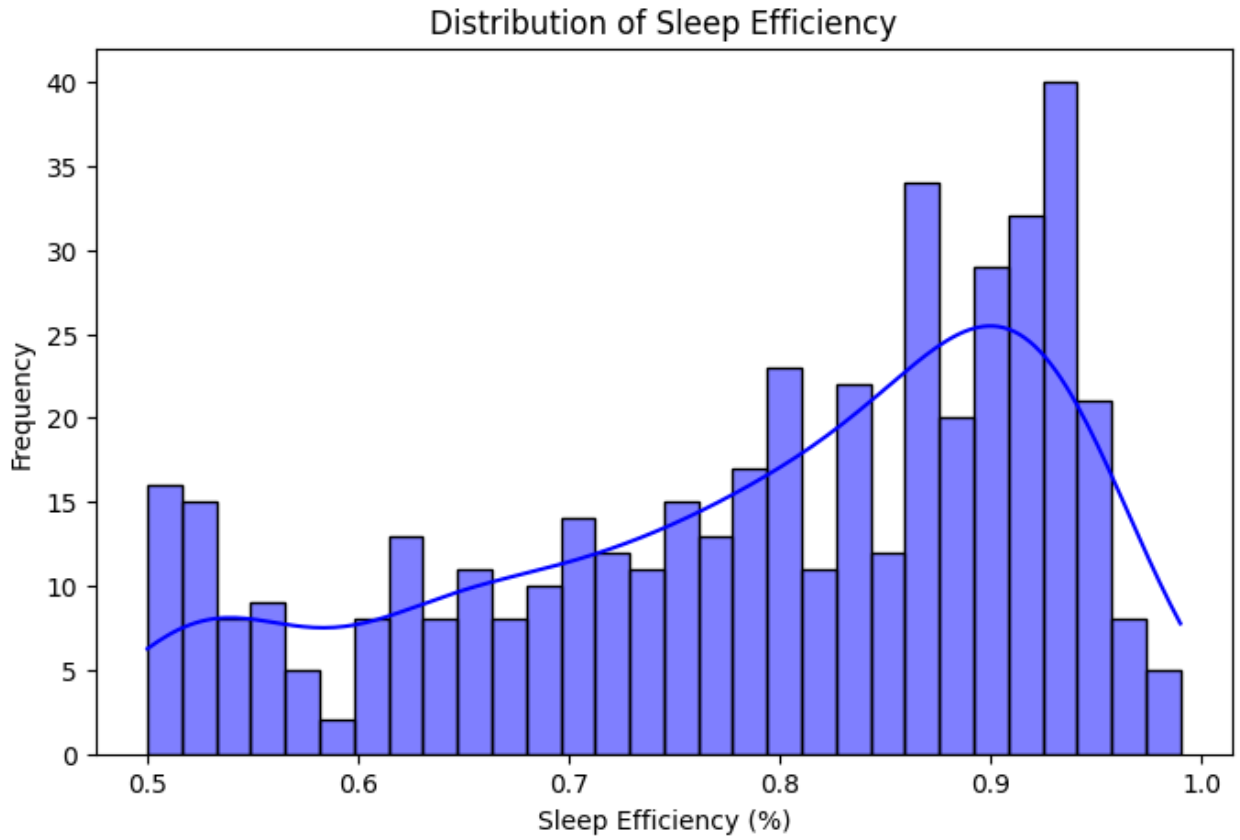

Figure 2 - Histogram of Sleep Efficiency Distribution

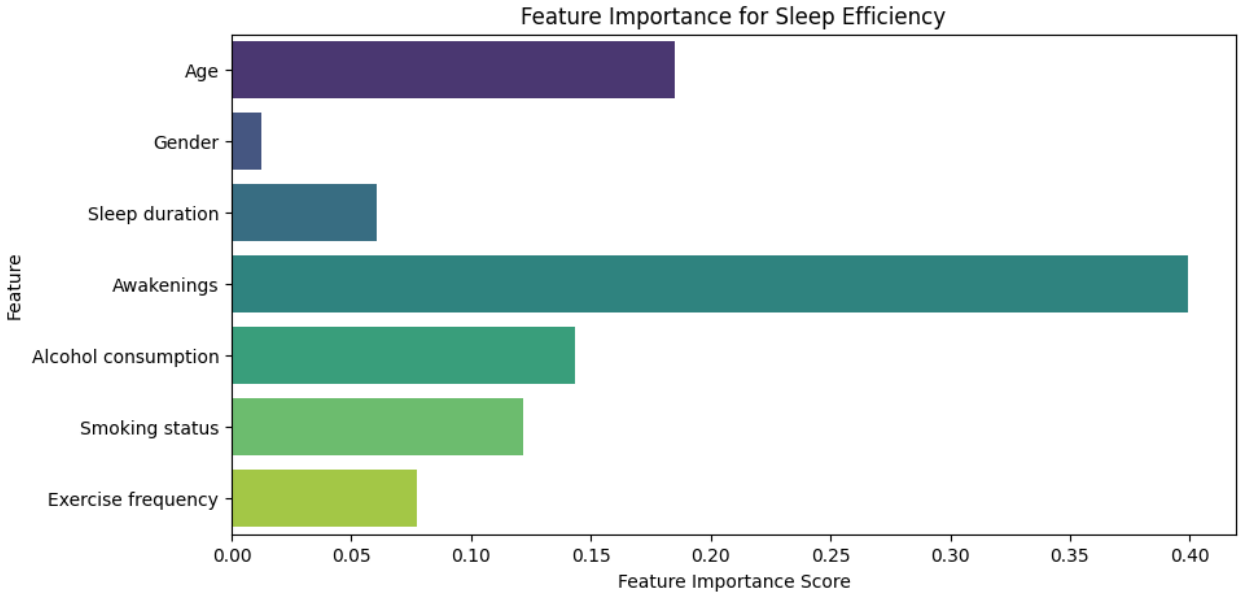

Figure 3 - Bar Plot of Important Features

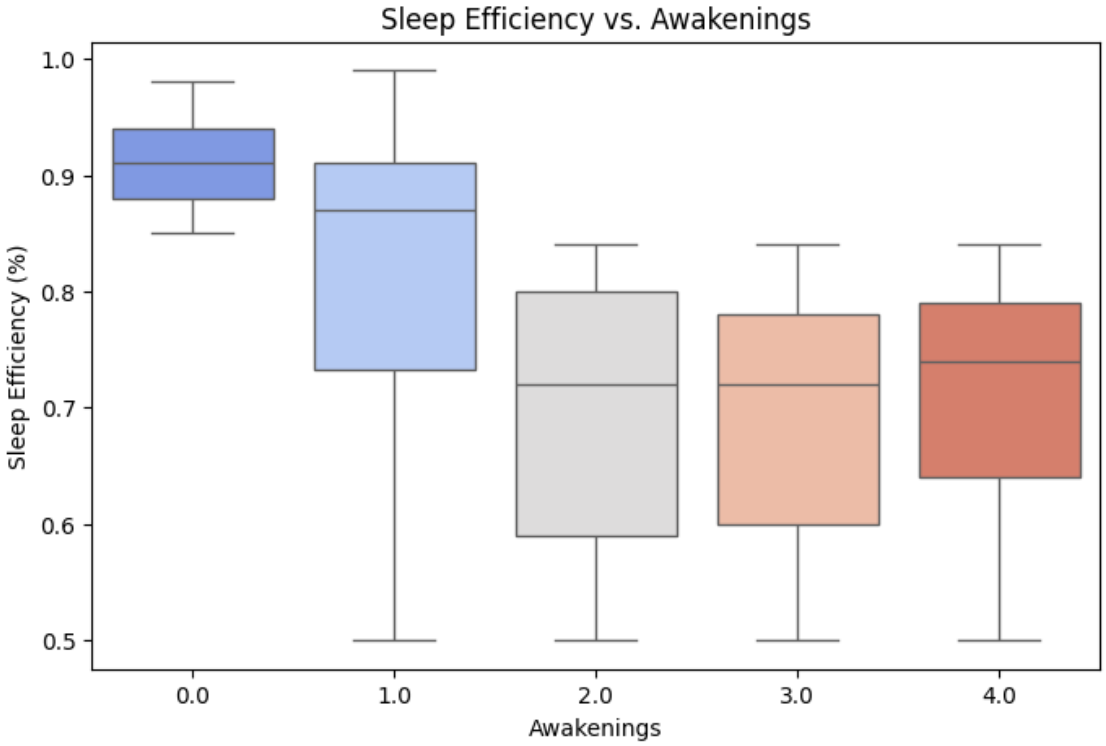

Figure 4 - Boxplot of Sleep Efficiency vs Awakenings

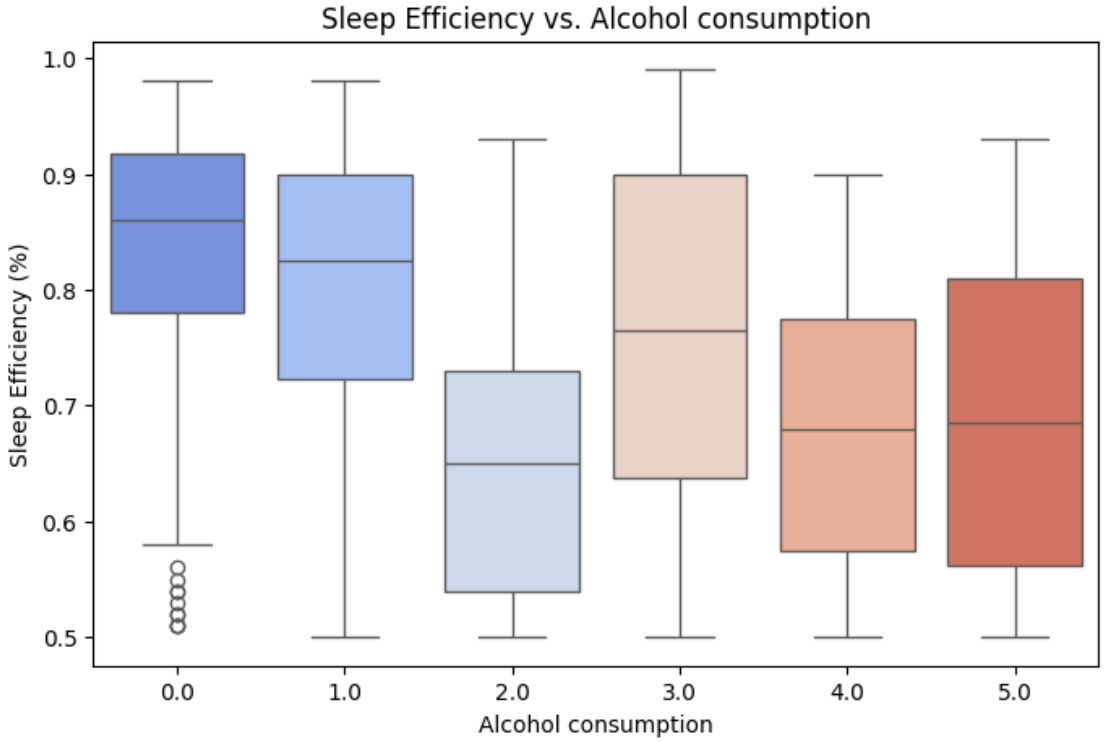

Figure 5 - Boxplot of Sleep Efficiency vs Alcohol Consumption

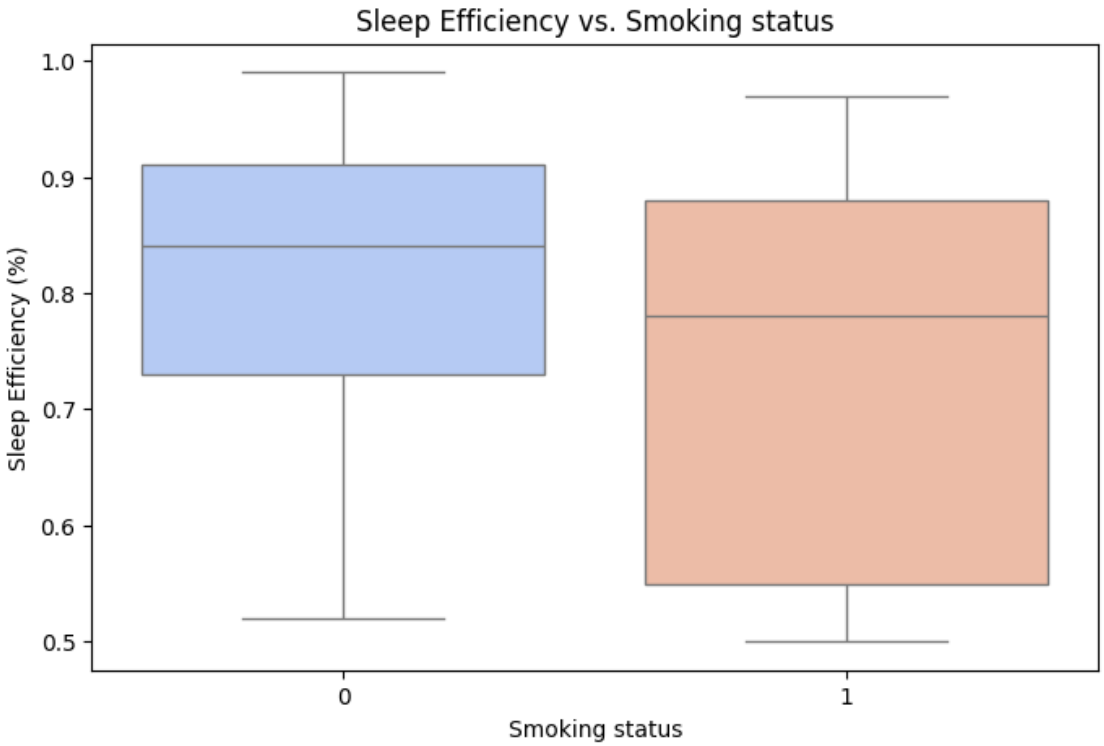

Figure 6 - Boxplot of Sleep Efficiency vs Smoking Status

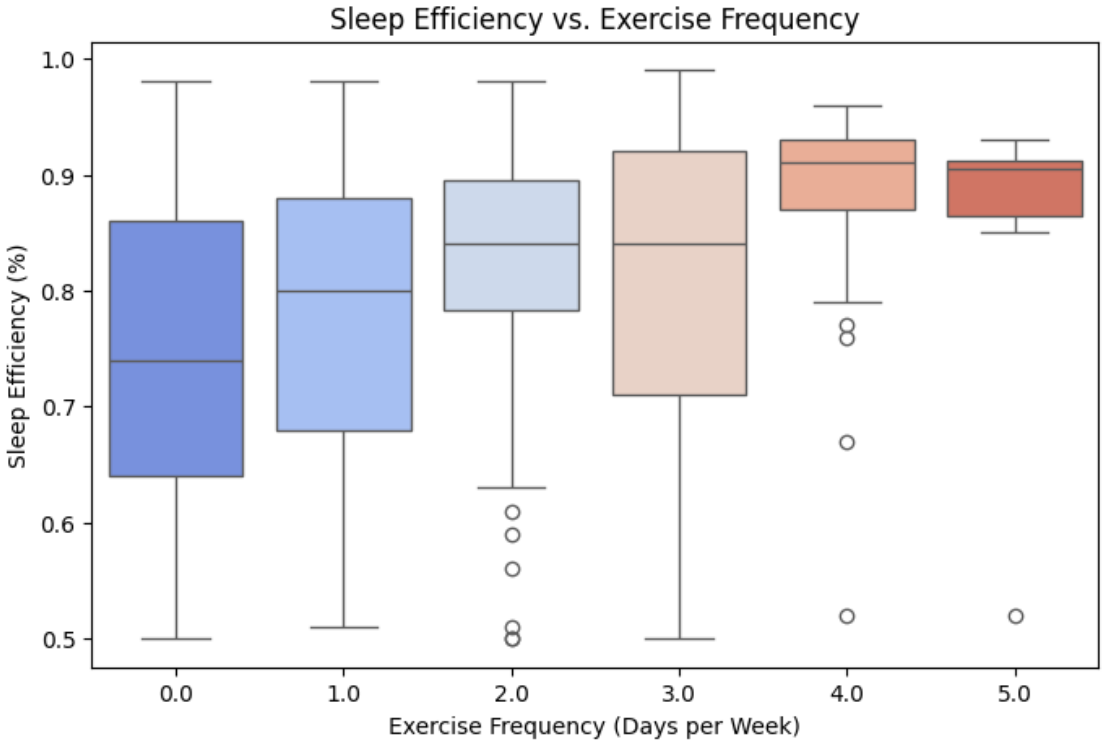

Figure 7 - Boxplot of Sleep Efficiency vs Exercise Frequency

User Guide

This guide explains how to use the Sleep Efficiency Prediction model through Google Colab without requiring any software installation.

1. Accessing the Application

To begin, open the hosted notebook using the following link:

https://colab.research.google.com/gist/mike4278/3b2ad895ec8a1ef5528e0a7147a4b3db/sleep_efficiency.ipynb

If prompted, select **"Open with Google Colab."**

2. Running the Model

Once the notebook is open, follow these steps:

- Click **“Runtime” → “Run all”** to execute the entire notebook.

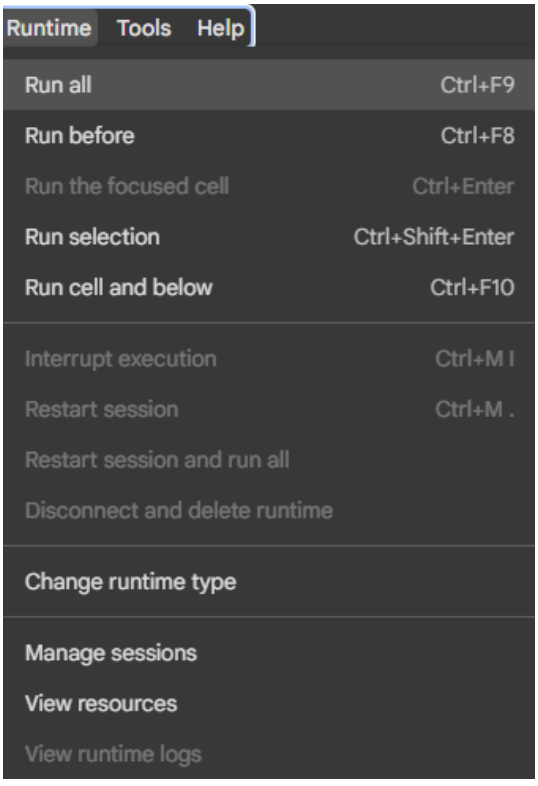

Figure 8 - Run all cells from Runtime menu

- A warning will be presented – “Warning: This notebook was not authorized by Google”
    - This is a standard security notice shown for all notebooks hosted outside of Google Drive, such as those on GitHub
    - Click “Run anyway” to proceed

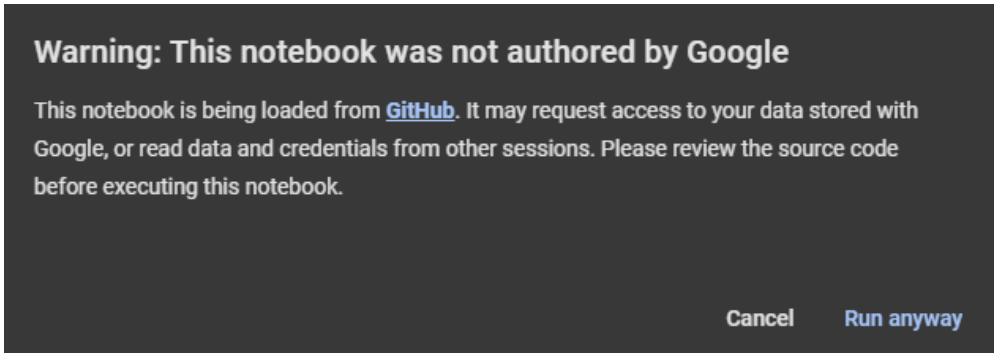

Figure 9 - Warning for projects hosted outside of Google Drive

- After confirming, the notebook will automatically:
    - Load and preprocess the sleep dataset.
    - Train the machine learning model.
    - Evaluate performance using key metrics (MAE, R^2 Score).
    - Generate and display relevant visualizations.

3. Making Predictions

To generate a personalized Sleep Efficiency prediction:

- Scroll to the **"Interactive Input"** section to access the user input form.
- Enter values for Age, Gender, Sleep Duration, Awakenings, Alcohol Consumption, Smoking Status, and Exercise Frequency.
- Click **"Submit"** to generate and display the predicted Sleep Efficiency Score.

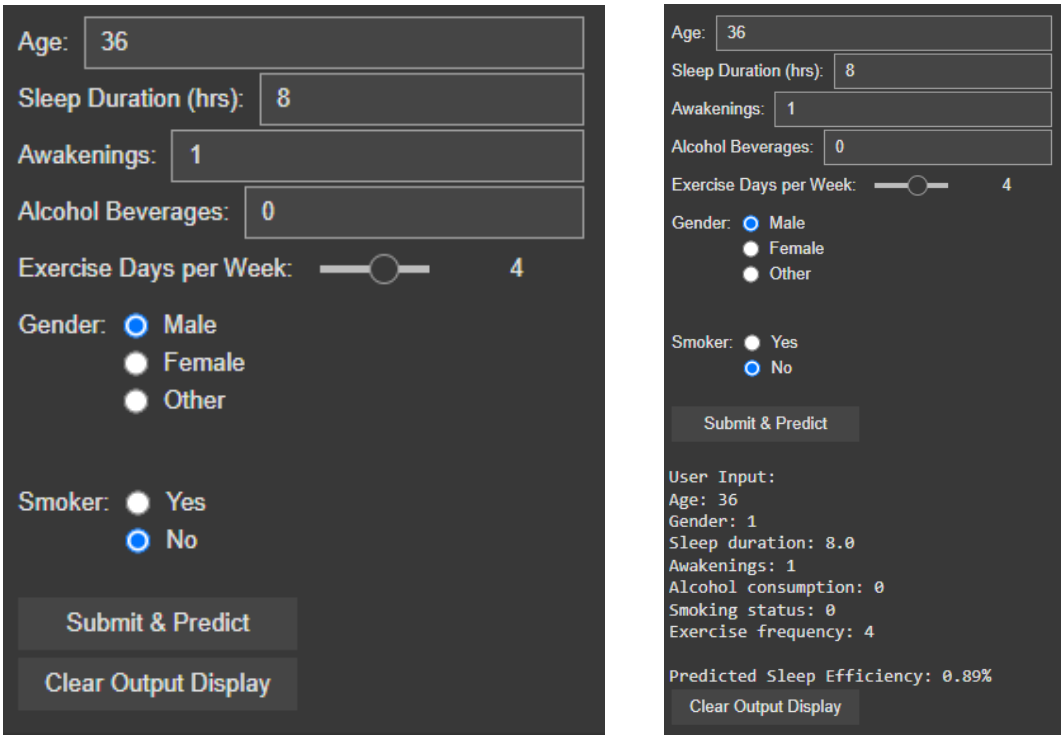

Figure 10 - User Input Form

- Click "Clear Output Display" at any time to remove currently displayed results.
    - If not cleared, predictions will remain visible in the order they were generated, allowing for easy comparison between multiple entries.

4. Interpret the Output

- The tool will generate a predicted Sleep Efficiency score between 0 and 1. A higher score indicates better efficiency. Users can analyze how different habits—such as reducing alcohol or increasing exercise—might improve their predicted sleep outcomes.

5. Example Use Cases

- A sleep specialist can enter client data to assess predicted sleep efficiency and offer personalized sleep recommendations.
- A researcher can use live model predictions to explore how lifestyle factors impact sleep quality.

6. Troubleshooting & Support

- If the notebook fails to execute, go to **“Runtime” → “Restart session”** then rerun all cells.
- If the dataset fails to load, verify that the file path or GitHub link is still valid and accessible.
- For additional assistance, contact **support@sleepsolutions.com**.

Reference Page

National Heart, Lung, and Blood Institute. (2022). *Sleep deprivation and deficiency*. U.S. Department of Health & Human Services. Retrieved from https://www.nhlbi.nih.gov/health/sleep-deprivation