IOWA STATE UNIVERSITY
Department of Computer Science

# Sleep Habits and Lifestyle Factors

Nina De Grandis and Lucas Sorge

December 10th, 2025

# Background

- What is the background of your task?
- Kaggle dataset from a few years ago.
- Machine learning can help address this problem by analyzing the data we have so far to predict whether a person is likely to have a sleep disorder or not, and their predicted quality of sleep.
- What kind of data you have?
- Demographics, lifestyle factors, physiological, and current sleep metrics.
- Both categorical and numerical
- What problem you want to solve?
- Gain a deeper understanding of how lifestyle choices impact sleep and to highlight potential ways people can improve their sleep habits.
- This problem is relevant to everyone.

# Datasets

- What is the dataset and what features?
- **Demographics:** Gender, Age, Occupation
- **Lifestyle factors:** Sleep Duration, Physical Activity, Daily Steps
- **Physiological measures:** Resting Heart Rate, Stress Level, Blood Pressure
- **Outcomes:** Sleep Disorder (categorical), Quality of Sleep (numeric)
- A critical characteristic of the dataset is that **only 155 individuals have Sleep Disorder labels**, almost perfectly balanced between Insomnia (77) and Sleep Apnea (78). The remaining 219 unlabeled rows are still usable for regression but must be dropped for classification.
- What is the expected output by your machine learning models.
- Classification models: likelihood of sleep disorder (Sleep Apnea, Insomnia)
- Regression models: predict quality of sleep score, from 1-10

# Related work

- Logistic Regression
  - Referenced papers that provided insight on how logistic regression models are analyzed and reported
  - Helpful for interpreting classification outputs and understanding which factor contribute most to sleep disorders
- Random Forest
  - Explained why random forest often outperforms linear models on nonlinear problems
- Sleep and Lifestyle Existing Research
  - Found similar studies using machine learning to study sleep health

# Method

## Intuition

- Use machine learning to predict:
  - **Sleep Disorder** (Insomnia vs. Sleep Apnea)
  - **Sleep Quality Score** (1–10)
- Idea: lifestyle + physiological variables should reveal patterns in sleep health.

## Data Processing

- Dataset: 374 adults
- Preprocessing:
  - Median / mode imputation
  - Standardization of numeric features
  - Drop non-informative ID
- Train/test split: **80/20**
  - Classification: stratified (155 labeled cases)
  - Regression: random split (299/75)

# Method

## Algorithm Details

**Classification:**

- Logistic Regression (baseline)
- Random Forest Classifier (nonlinear)

**Regression:**

- Linear Regression (baseline)
- Ridge & Lasso
- Random Forest Regressor

# Experimental setups

## Features Used

- **Demographics:** Gender, Age, Occupation
- **Lifestyle:** Sleep Duration, Physical Activity, Daily Steps
- **Physiological:** Heart Rate, Stress Level, Blood Pressure
- Targets:
  - Sleep Disorder (categorical)
  - Sleep Quality (numeric)

## Training Process

- 5-fold cross-validation on training set
- Standardized features for linear models
- Models trained on CPU (dataset small → fast training)

# Experimental results

## Classification Results

| Model | Accuracy | F1 |
|---|---|---|
| Logistic Regression | 0.839 | 0.839 |
| Reg. Logistic (C=0.01) | 0.774 | 0.774 |
| Random Forest Classifier | 0.903 | 0.903 |

## Insights

- Random Forest captured nonlinear interactions → best performance.
- Sleep Duration, Stress, Heart Rate, and Blood Pressure were most important predictors.

# Experimental results

## Regression Results

| Model | $R^2$ | RMSE | MAE |
|---|---|---|---|
| Linear Regression | 0.967 | 0.225 | 0.119 |
| Ridge Regression | 0.964 | 0.233 | 0.167 |
| Lasso Regression | 0.955 | 0.261 | 0.181 |
| Random Forest Regressor | 0.982 | 0.164 | 0.046 |

## Insights

- Linear relationships are strong → baseline linear reg already good.
- Random Forest Regressor shows sleep quality has **nonlinear thresholds** (stress, HR).
- Results matched expectations:
  - Higher stress / heart rate / Blood Pressure → worse sleep
  - More activity / steps → higher sleep quality

# Conclusion and Future Work

- Machine learning is an effective tool for predicting sleep disorders based on lifestyles and health factors
- Confirmed our hypothesis
- Nonlinear random forest models performed the best, in both classification and regression tasks
- Future work
  - Expanding the dataset
  - Adding more detailed health data
  - Testing additional ML models