## **Predicting Sleep Quality Based on Lifestyle Factors**

By: Leah Schneck, Christina Jurotich, Libby Abts, Niyati Vijay & Tanisha Anand Raaj STAT 451: Project 20

#### **Sleep Health and Lifestyle Dataset**

- 400 rows x 13 columns; 400 observations.
- <u>https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data</u>

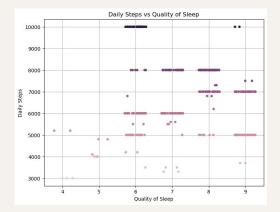
|              | Gender | Age | Occupation              | Sleep<br>Duration | Quality of<br>Sleep | Physical Activity<br>Level | Stress<br>Level | BMI<br>Category | Blood<br>Pressure | Heart<br>Rate | Daily<br>Steps | Sleep<br>Disorder |
|--------------|--------|-----|-------------------------|-------------------|---------------------|----------------------------|-----------------|-----------------|-------------------|---------------|----------------|-------------------|
| Person<br>ID |        |     |                         |                   |                     |                            |                 |                 |                   |               |                |                   |
| 1            | Male   | 27  | Software Engineer       | 6.1               | 0.0                 | 42                         | 6               | Overweight      | 126/83            | 77            | 4200           | None              |
| 2            | Male   | 28  | Doctor                  | 6.2               | 0.0                 | 60                         | 8               | Normal          | 125/80            | 75            | 10000          | None              |
| 3            | Male   | 28  | Doctor                  | 6.2               | 0.0                 | 60                         | 8               | Normal          | 125/80            | 75            | 10000          | None              |
| 4            | Male   | 28  | Sales<br>Representative | 5.9               | 0.0                 | 30                         | 8               | Obese           | 140/90            | 85            | 3000           | Sleep Apnea       |
| 5            | Male   | 28  | Sales<br>Representative | 5.9               | 0.0                 | 30                         | 8               | Obese           | 140/90            | 85            | 3000           | Sleep Apnea       |
| 6            | Male   | 28  | Software Engineer       | 5.9               | 0.0                 | 30                         | 8               | Obese           | 140/90            | 85            | 3000           | Insomnia          |
| 7            | Male   | 29  | Teacher                 | 6.3               | 0.0                 | 40                         | 7               | Obese           | 140/90            | 82            | 3500           | Insomnia          |
| 8            | Male   | 29  | Doctor                  | 7.8               | 0.0                 | 75                         | 6               | Normal          | 120/80            | 70            | 8000           | None              |
| 9            | Male   | 29  | Doctor                  | 7.8               | 0.0                 | 75                         | 6               | Normal          | 120/80            | 70            | 8000           | None              |
| 10           | Male   | 29  | Doctor                  | 7.8               | 0.0                 | 75                         | 6               | Normal          | 120/80            | 70            | 8000           | None              |

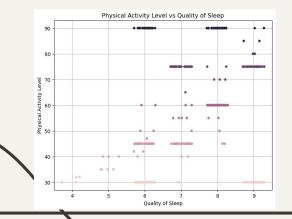
#### **Questions of Interest?**

- Can we predict sleep quality based on other lifestyle factors?
- Which lifestyle factors are the most important for predicting sleep quality?
- Which models are the best predictors?

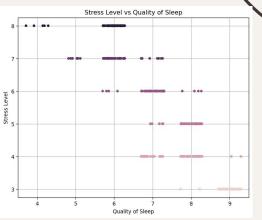


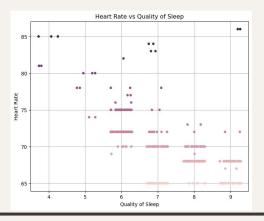
#### **Data Visualization – Numerical Variables**



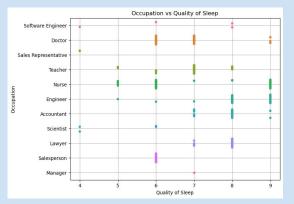


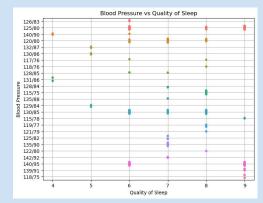


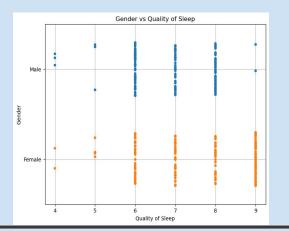


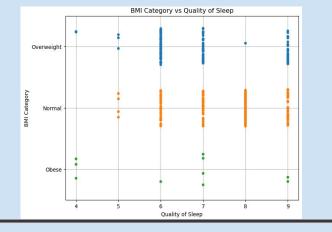


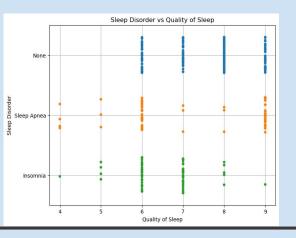
#### **Data Visualization – Categorical Variables**



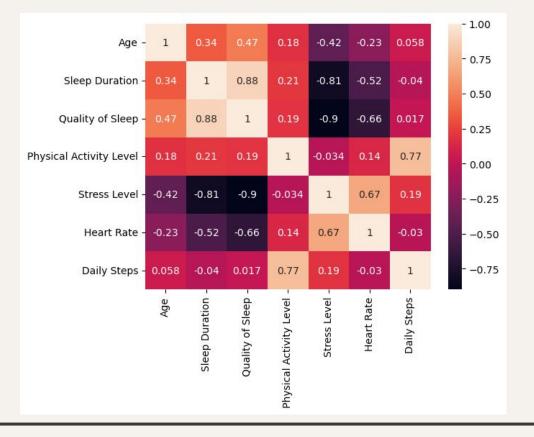


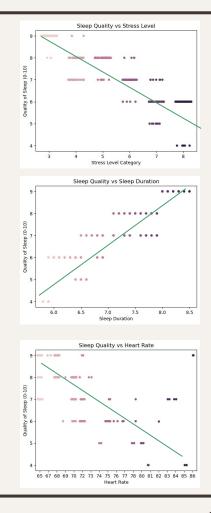




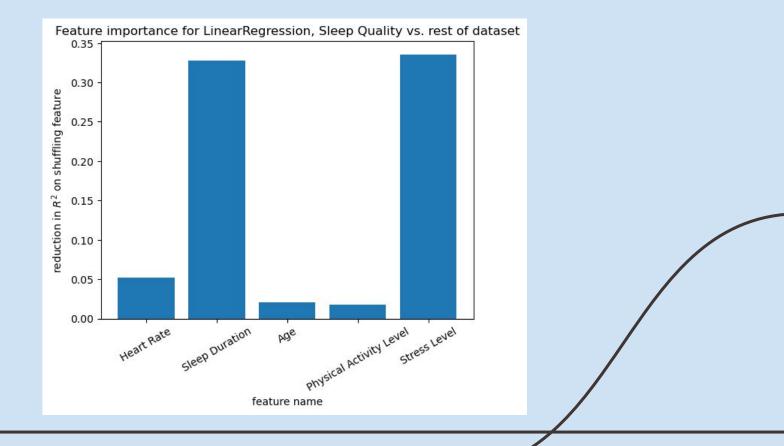


#### **Proposed Model: Linear Regression**



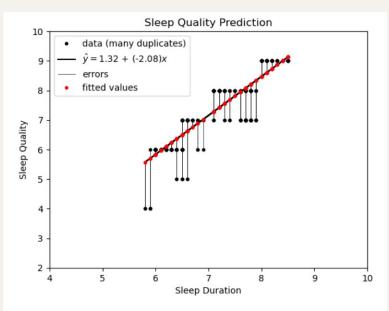


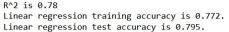
#### **Linear Regression: Feature Permutation Importance**



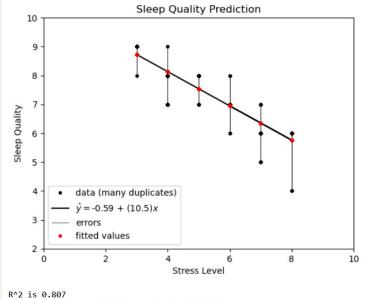
#### **Simple Linear Regression**

**Sleep Quality vs Sleep Duration** 





#### **Sleep Quality vs Stress Level**



Linear regression training accuracy is 0.791. Linear regression test accuracy is 0.84.

#### **Multiple Linear Regression**

Predicting 'Sleep Quality' based on factors 'Stress Level' and 'Sleep Duration'.

Linear regression training accuracy is 0.874. Linear regression test accuracy is 0.884. Linear regresstion slope is -0.346

#### **Linear Regression Conclusions**

#### • Simple linear regression

- Accuracy of 0.795 when using Sleep Duration to predict sleep quality.
- Accuracy of 0.84 when using Stress Level to predict sleep quality.

#### • Multiple Linear Regression

• Sleep Duration + Stress Level = 0.88

#### Boundless model

- Can predict sleep qualities outside of the specified range (0-10).
- Sleep quality for stress level 2 and sleep duration 13hr is 12.6.
  Sleep quality for stress level 3 and sleep duration 18hr is 15.8.
  Sleep quality for stress level 7 and sleep duration 2hr is 3.15.
- To avoid this, we will try modeling sleep quality with logistic regression.

#### **Updated Model: Logistic Regression**

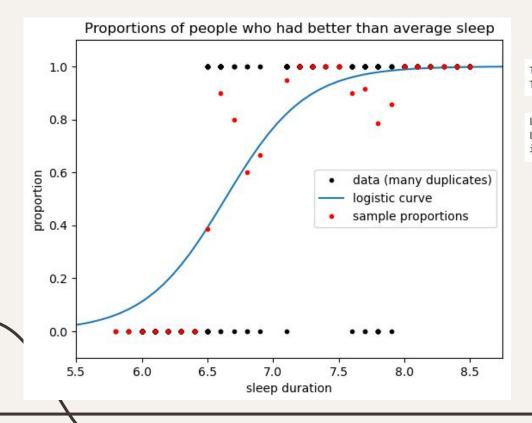
#### • Selected features:

- Sleep duration
- Stress level

#### • Methods:

- Grid search to determine best parameters
- Train\_test\_split to train on training data and test accuracy on testing data using the same parameters from linear regression
- Changing quality of sleep rating to 1 (better than average quality of sleep) or 0 (worse than average quality of sleep)

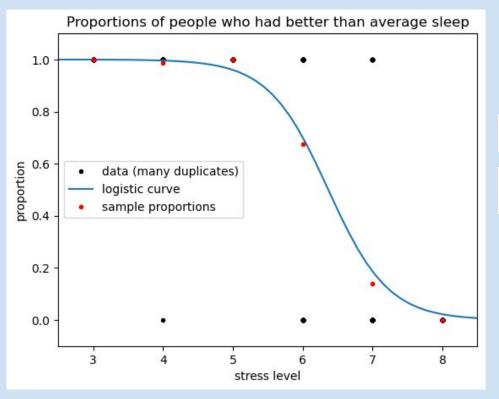
#### **Logistic Regression: Sleep Duration vs Quality of Sleep**



The best accuracy score on the validation data is: 0.912 This means that the best parameter is {'C': 1, 'max\_iter': 5000}

Logistic regression training accuracy for sleep duration is 0.874. Logistic regression test accuracy for sleep duration is 0.912. intercept=[-21.65839429], slope=[3.26427132]

#### Logistic Regression: Stress Level vs Quality of Sleep



The best accuracy score on the validation data is: 0.947 This means that the best parameter is {'C': 1, 'max iter': 5000}

Logistic regression training accuracy for stress level is 0.935. Logistic regression test accuracy for stress level is 0.947. intercept=[14.81876522], slope=[-2.32815118]

### **Logistic Regression Conclusions**

- Logistic regression improved prediction accuracy compared to linear regression
  - 0.88 → 0.95!
- Using stress level to predict sleep quality yielded a better model than using sleep duration.
  - Accuracy 0.912 → 0.947
- Can we improve this accuracy further with a decision tree model?

#### **Updated Model : Decision Tree Classification** Stress Level <= 6.5 gini = 0.756 samples = 261 value = [3, 5, 68, 57, 80, 48] class = 4leep Duration <= 6.45 Sleep Duration <= 7.95 gini = 0.658 gini = 0.312 samples = 182 samples = 79 value = [0, 0, 3, 51, 80, 48] value = [3, 5, 65, 6, 0, 0] class = 4class = 2 Sleep Duration <= 7.0 Heart Rate <= 73.0 Heart Rate <= 76.5 gini = 0.0 gini = 0.498 gini = 0.632 gini = 0.126 samples = 48 lue = [0, 0, 0, 0, 0, 48 samples = 134 samples = 60 samples = 19 value = [0, 0, 3, 51, 80, 0] alue = [3, 1, 56, 0, 0, 0 alue = [0, 4, 9, 6, 0, 0] class = 5 class = 4class = 2class = 2 Sleep Duration <= 6.55 Sleep Duration <= 6.7 Stress Level <= 5.5 Sleep Duration $\leq = 6.1$ gini = 0.377 gini = 0.0 gini = 0.198 gini = 0.56 gini = 0.486 gini = 0.571 samples = 55 samples = 27 samples = 107 samples = 5samples = 12 samples = 7alue = [0, 0, 55, 0, 0, 1 alue = [0, 0, 3, 24, 0, 0] alue = [0, 0, 0, 27, 80, 0] value = [3, 1, 1, 0, 0, 0] value = [0, 0, 7, 5, 0, value = [0, 4, 2, 1, 0, 0 class = 2class = 3 class = 4class = 0class = 2class = 1Heart Rate <= 72.5 Heart Rate <= 80.0 Heart Rate <= 70.5 gini = 0.375 gini = 0.0 gini = 0.444 gini = 0.0 gini = 0.0 gini = 0.0 gini = 0.049 gini = 0.5 gini = 0.375 gini = 0.191 samples = 3 samples = 24samples = 3samples = 8 samples = 4samples = 3samples = 2 samples = 28 samples = 4alue = [0, 0, 3, 0, 0, lue = [0, 0, 0, 24, 0, 0 alue = [3, 0, 0, 0, 0, 0] value = [0, 0, 6, 2, 0, 0] alue = [0, 4, 0, 0, 0, 0] value = [0, 0, 2, 1, 0, 0] ue = [0, 0, 0, 2, 77, 0] alue = [0, 0, 0, 25, 3, 0] value = [0, 1, 1, 0, 0, value = [0, 0, 1, 3, 0, class = 2class = 3class = 0class = 2class = 1class = 2 class = 4class = 3 class = 1class = 3 ep Duration <= 7.45 gini = 0.0 gini = 0.0gini = 0.0 qini = 0.1qini = 0.0gini = 0.074samples = 2 samples = 1 samples = 1 samples = 1 samples = 3 alue = [0, 0, 0, 3, 0, 0 samples = 26 alue = [0, 0, 0, 2, 0, 0 value = [0, 1, 0, 0, 0, 0] alue = [0, 0, 1, 0, 0, 0] ue = [0, 0, 1, 0, 0, 0 alue = [0, 0, 0, 25, 1, 0 class = 4class = 3 class = 4class = 1class = 2class = 2class = 3 class = 3 gini = 0.0 gini = 0.375 samples = 22 samples = 4lue = [0, 0, 0, 22, 0, 0 Training Accuracy is: 98.85% alue = [0, 0, 0, 3, 1, 0] class = 3 class = 3Test Accuracy is: 99.12% gini = 0.0 gini = 0.0 samples = 3 alue = 10.0.0.3.0.0 class = 3 class = 4

#### **Decision Tree Classification Reasons**

- Unlike Linear and logistic regression, decision tree can map more complex relationships
- Less sensitive to outliers since splitting is based on distribution of data not the magnitude
- Naturally accounts for interactions between variables through branching

#### **Decision Tree Conclusions**

- Decision tree improved prediction accuracy compared to logistic regression
  - 0.95 → 0.99
- Relatively unstable compared to linear and logistic regression since small changes in data can alter the tree structure significantly
- So what actually is the best final model?

#### **Comparison Between Models**

|             | Multiple Linear Regression  | Logistic Regression   | Decision Tree   |  |  |
|-------------|---|---|---|--|--|
| Accuracy    | 0.88  | 0.95  | 0.99  |  |  |
| Predictions | Stress 2, duration 13 = 12.6<br>Stress 3, duration 18 = 15.8<br>Stress 7, duration 2 = 3.15 | Sleep 13 = 1; proba = 0.9<br>Sleep 18 = 1; proba = 1.0<br>Sleep 2 = 0; proba = 2e-7<br>Stress 2 = 1; proba = 0.9<br>Stress 3 = 1; proba = 0.9 | Stress 2, duration 13 = class<br>5<br>Stress 3, duration 18 = class<br>5<br>Stress 7, duration 2 = class<br>2 |  |  |
|             |   | Stress 7 = 0; proba = 0.18  |   |  |  |

## Conclusion

- Best model option
  - Decision Tree
- Important features
  - Linear Regression
    - Stress Level
  - Logistic Regression
    - Stress Level
  - Decision Tree
    - Stress Level
- Ways to improve the model in the future
  - Cross-Validation
  - Feature Engineering

# Thank You

**Questions?**