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

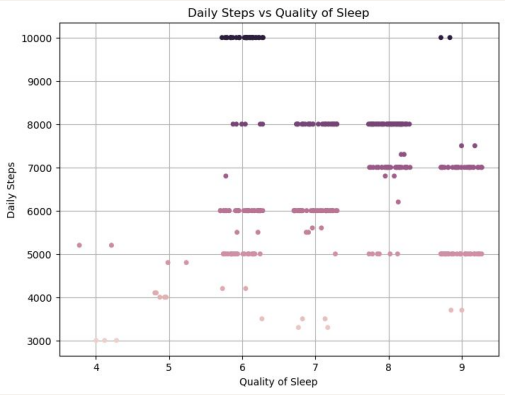
- 400 rows x 13 columns; 400 observations.
- https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data

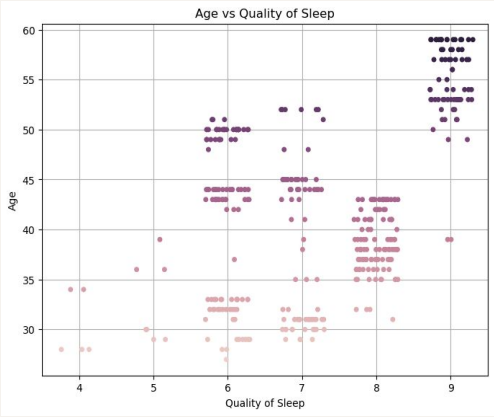
Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
1	Male	27	Software Engineer	6.1	0.0	42	6	Overweight	126/83	77	4200	None
2	Male	28	Doctor	6.2	0.0	60	8	Normal	125/80	75	10000	None
3	Male	28	Doctor	6.2	0.0	60	8	Normal	125/80	75	10000	None
4	Male	28	Sales Representative	5.9	0.0	30	8	Obese	140/90	85	3000	Sleep Apnea
5	Male	28	Sales 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

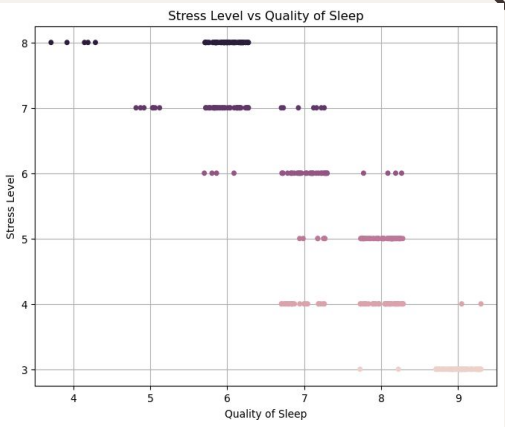
Questions of Interest?

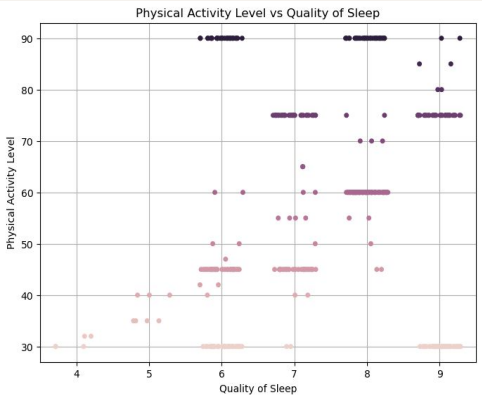
- Can we predict sleep quality based on other lifestyle factors?
- Which lifestyle factors are the most important for predicting sleep quality?
- Which models are the best predictors?

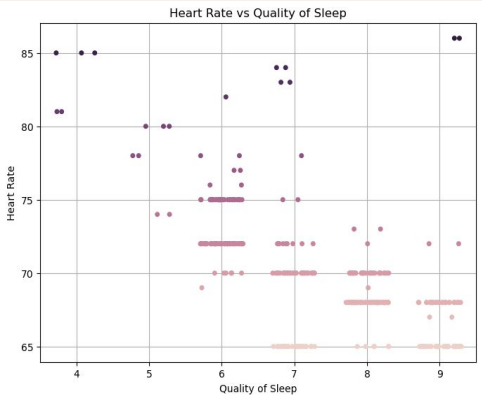
Data Visualization – Numerical Variables



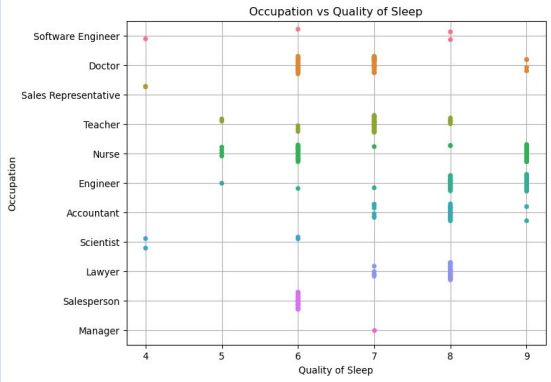


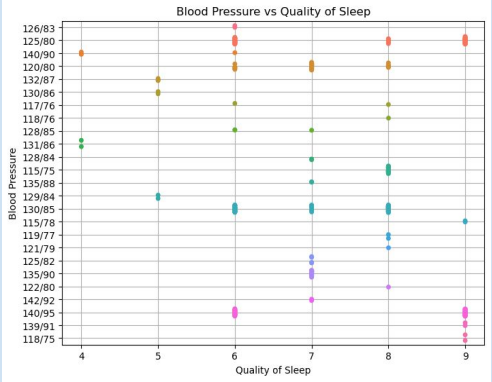


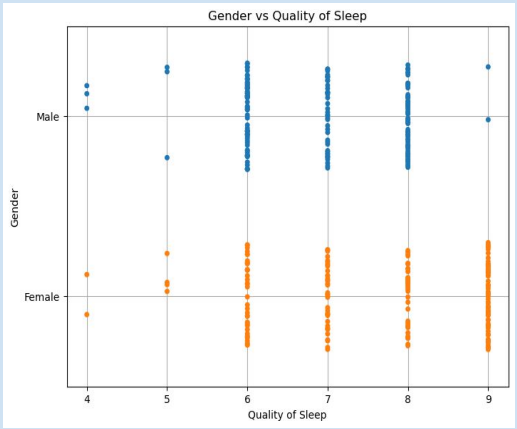


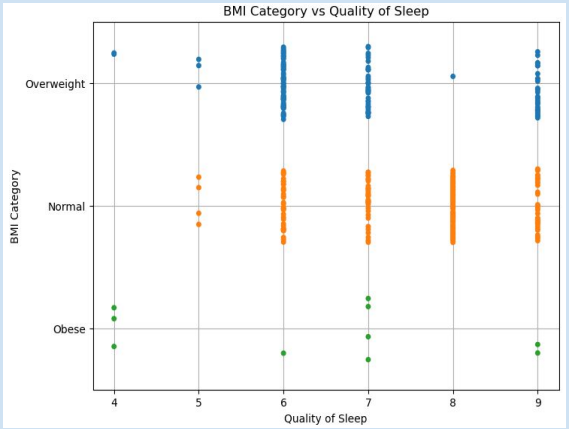


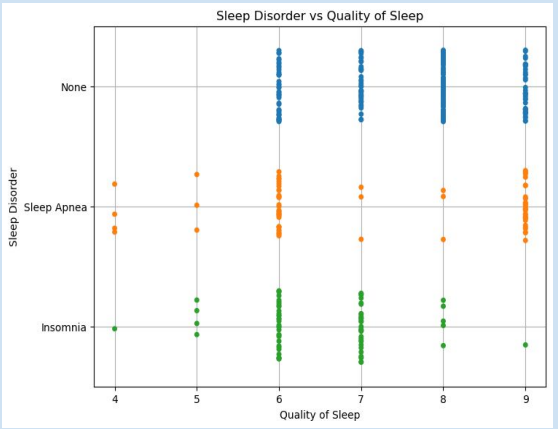
Data Visualization – Categorical Variables



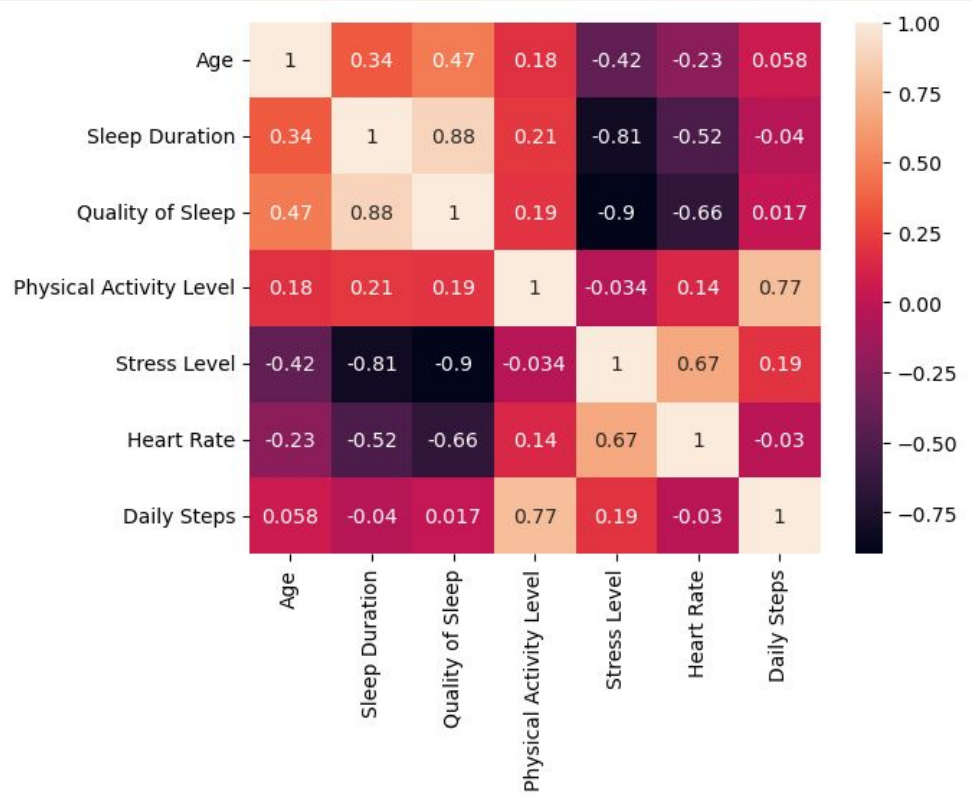


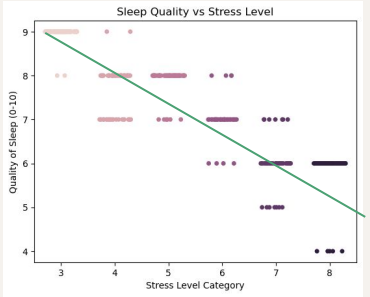


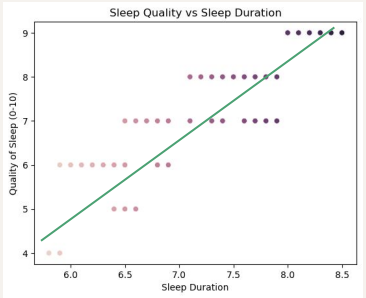


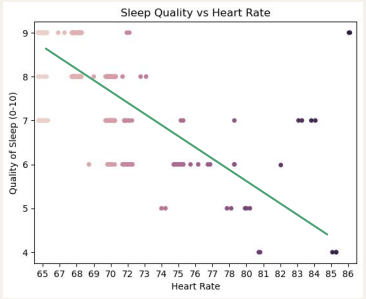


Proposed Model: Linear Regression

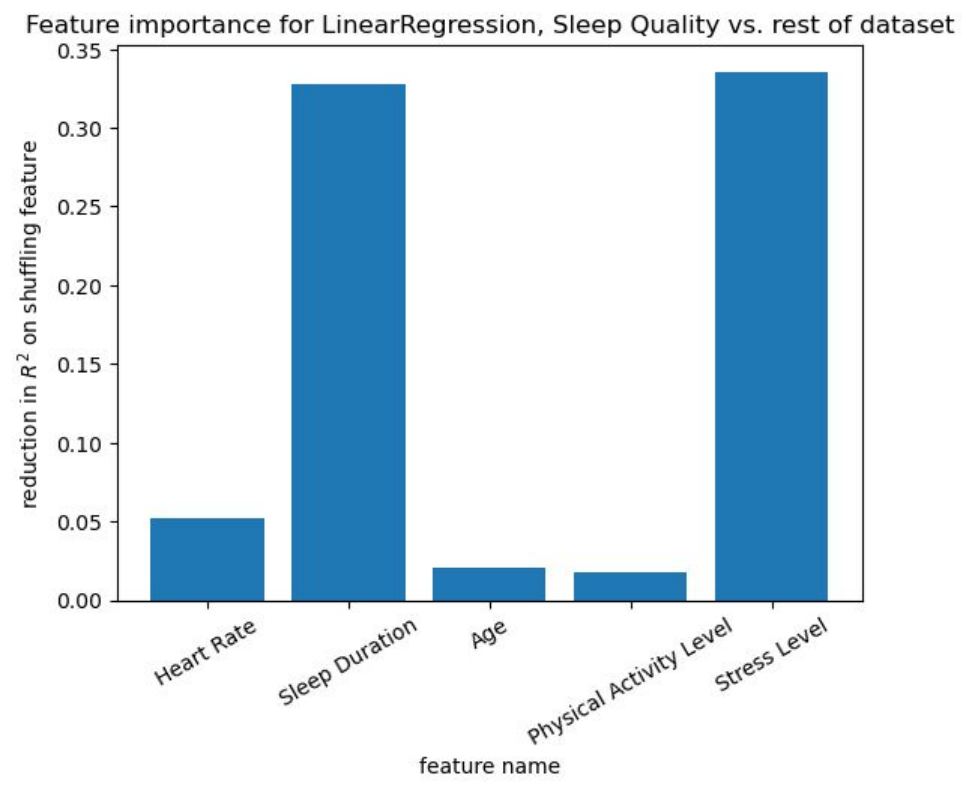






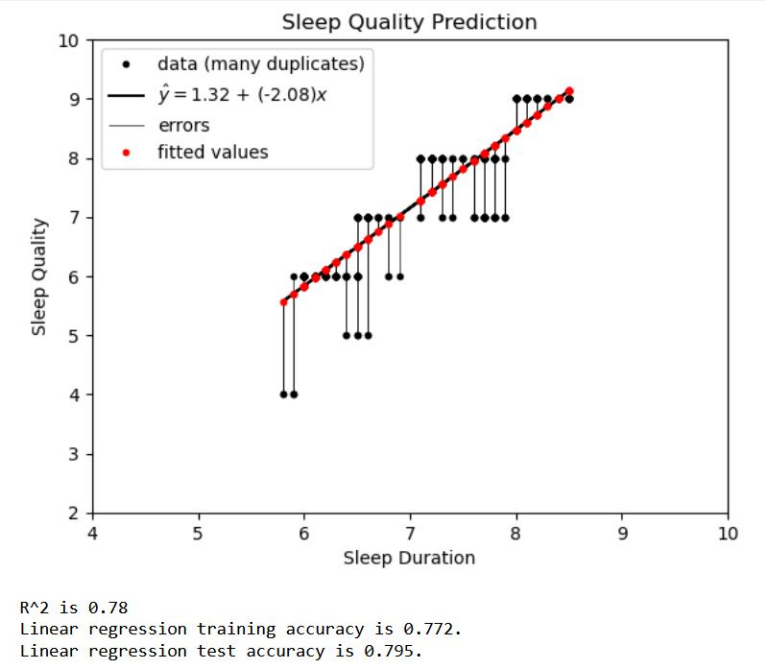


Linear Regression: Feature Permutation Importance



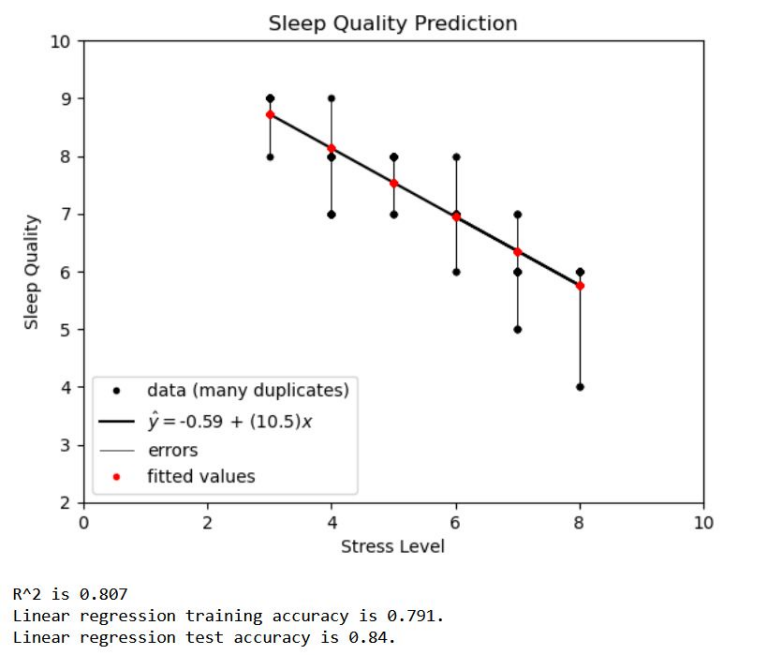
Simple Linear Regression

Sleep Quality vs Sleep Duration



R^2 is 0.78
Linear regression training accuracy is 0.772.
Linear regression test accuracy is 0.795.

Sleep Quality vs Stress Level



R^2 is 0.807
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
```

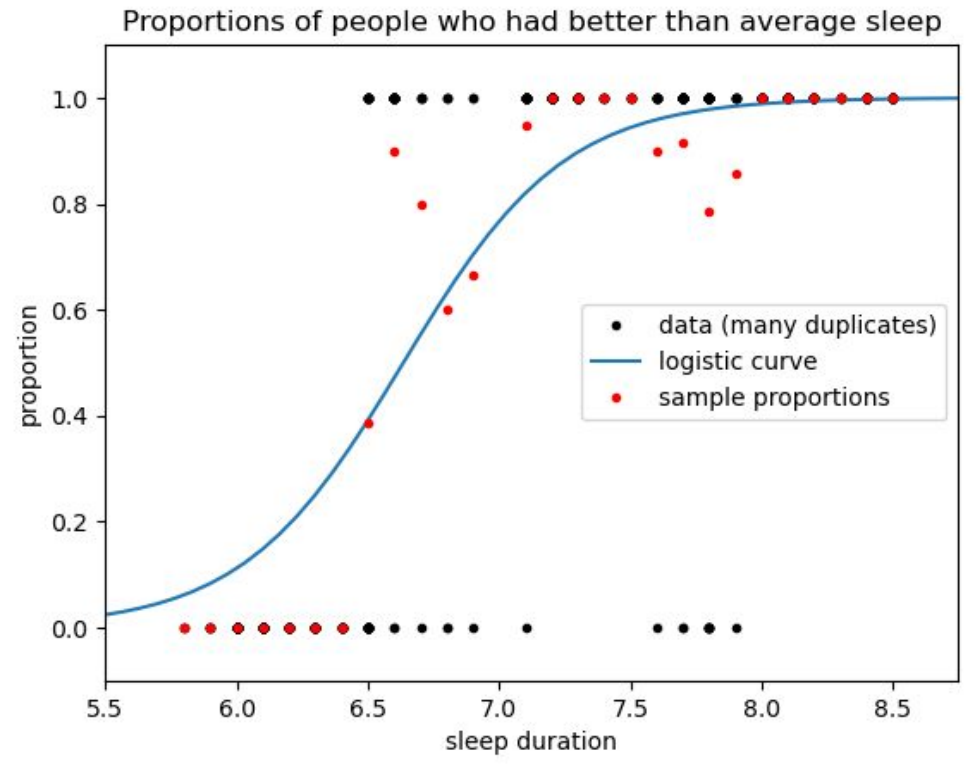
Linear Regression Conclusions

- **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)

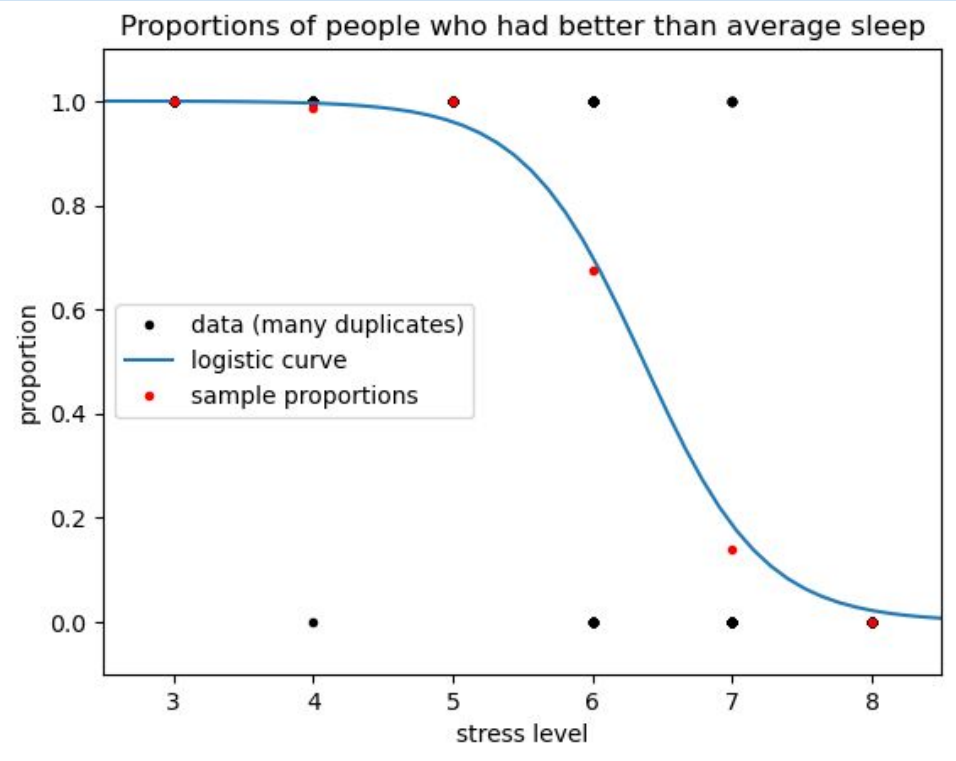
Logistic Regression: Sleep Duration vs Quality of Sleep




```
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



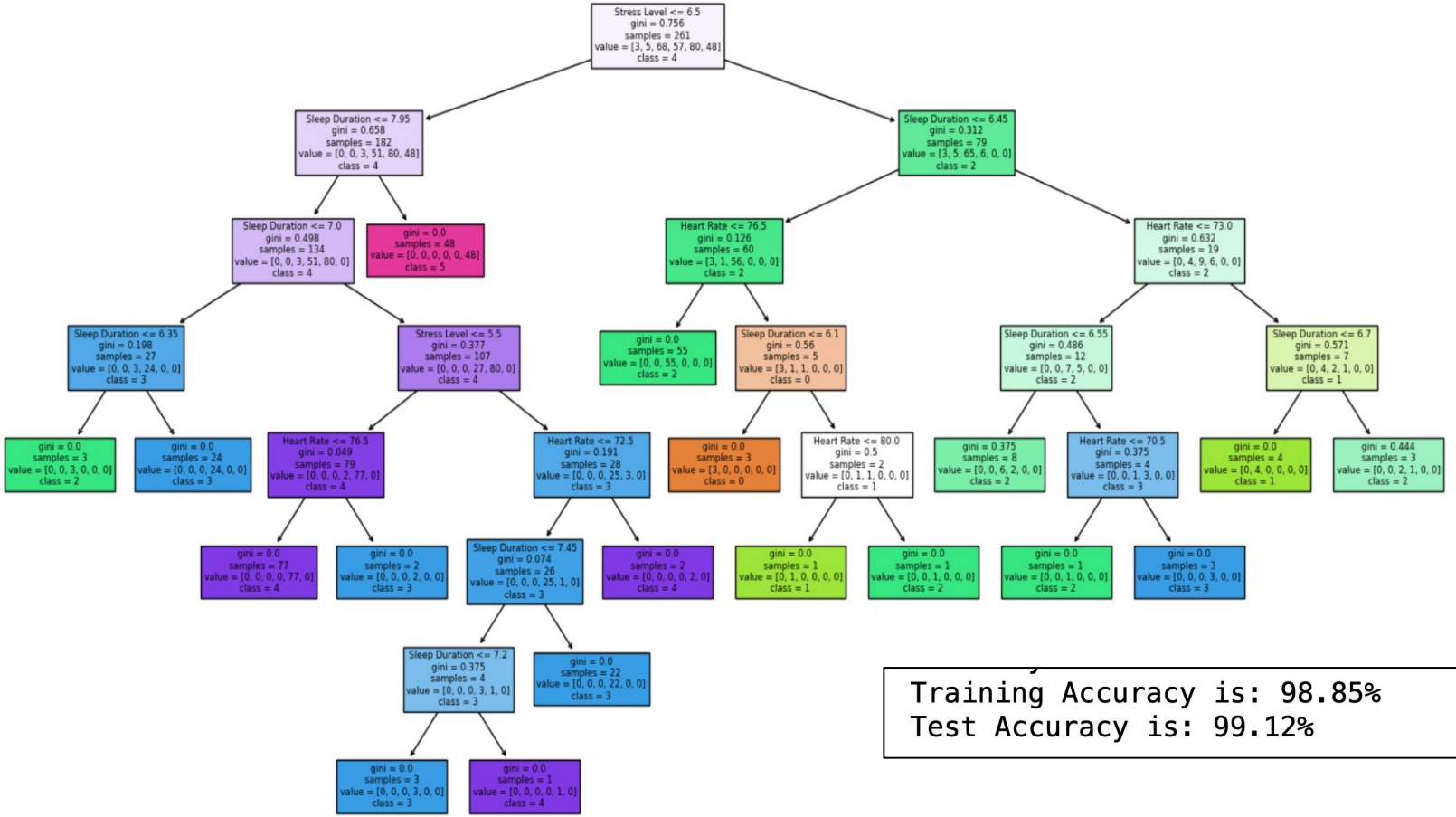

```
The best accuracy score on the validation data is: 0.947
This means that the best parameter is {'C': 1, 'max_iter': 5000}
```

```
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




```
Training Accuracy is: 98.85%
Test Accuracy is: 99.12%
```

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 Stress 3, duration 18 = 15.8 Stress 7, duration 2 = 3.15	Sleep 13 = 1; proba = 0.9 Sleep 18 = 1; proba = 1.0 Sleep 2 = 0; proba = 2e-7 Stress 2 = 1; proba = 0.9 Stress 3 = 1; proba = 0.9 Stress 7 = 0; proba = 0.18	Stress 2, duration 13 = class 5 Stress 3, duration 18 = class 5 Stress 7, duration 2 = class 2

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?