# 

# 

# Exploration and Prediction of Sleep Disorders

## By

Group NASA

### Chun Wang, Xi Du, Shangzhou Xia, Yichen Ding

Nowadays, about 50 to 70 million Americans have sleep disorders, and about 84 million people do not regularly get the recommended amount of uninterrupted sleep (Alnawwar et al.). Studies showed that sleep disorders impair attention and memory, negatively affecting physical, psychological, and social action, thus making early and accurate predictions crucial for effective intervention and treatment. In our project, we aim to address this pressing concern by exploring the effect of lifestyle on sleep conditions and determining the best predictive model for sleep disorders among four methods: logistic regression, classification tree, random forest, and neural network.

The Sleep Health and Lifestyle Dataset comprises 374 rows and 13 columns, including features related to demographics, sleep, and daily habits. We preprocess the data by converting all the categorical variables (occupation, gender, and BMI category) with the data type of string into dummy variables and changing the textual data (blood pressure) into a binary variable indicating the normality state. To ensure effective model building and evaluation, we split the data into 75% training data and 25% testing sets after cleaning.

The immediate step was to create visualizations to discover any obvious patterns, helping us grasp an idea of the data. To study factors that drive sleep disorders, we chose Matplot to create two separate bar charts to analyze the relationship between stress levels, quality of sleep, and sleep disorders (Exhibits 1 and 2). The output aligns with our intuition. Next, we leveraged the seaborn package to create an informative scatter plot involving prominent factors such as sleep disorder, quality of sleep, and age (Exhibit 3). The graph illustrates duplicates in different colors, adding another explanatory dimension. Furthermore, we incorporate demographic data into visualizations using Pandas native groupby function (Exhibit 4), which tells us how genders in different occupations contribute to sleep disorder. Last, we used a heatmap to depict the correlation matrix by specifying decimals and color in the seaborn package (Exhibit 5).

The dependent variable, *Disorder*, is binary, so we ran a logistic regression model on the training set. After calculating Variance Inflation Factor (VIF) values (Exhibit 6), we included only explanatory variables with VIF values below 10 in Model 1. The summary of Model 1 (Exhibit 7) showed that the variables *Overweight* and *high\_bp* are statistically significant in predicting sleep disorder (p-value < 0.05). We then built a second logistic regression, Model 2, using only the selected predictors. The summary of Model 2 (Exhibit 8) confirmed that these two variables remain statistically significant. Using Model 2 to predict *Disorder* on the testing set, we stored the actual sleep disorder values in the *Disorder* column and the predicted values in the *predicted\_raw* column of a data frame. Setting a threshold of 0.5, we created a *predicted\_binary* column, assigning 1 to predicted values greater than 0.5 and 0 otherwise (Exhibit 9).

Model 2 achieved 95.7% accuracy, but we remain cautious for two reasons: 1. The chosen model includes only two predictors and may not perform as well in capturing patterns as models with more variables when the data size increases. 2. The choice of threshold influences the prediction results. A cost matrix is needed to help us determine a better threshold value.

The second model we applied to predict sleep disorders using demographic and lifestyle variables is Classification Tree. While a shallow tree may oversimplify the relationship in the data and a deep tree risks capturing noise and overfitting the training set, to optimize the performance of the model, we experimented fitting the training data into four different classification trees with depths of one to four, due to the small sample size. By evaluating the performance of each tree, a classification tree with a max depth of 2 achieves the highest test accuracy (0.9681) while maintaining simplicity, reducing the risk of overfitting observed at greater depths (Exhibit 10).

The tree starts with an initial entropy of 0.984 at the root node. As the tree splits on health condition features, the entropy decreases at each level, showing the effectiveness of the split in reducing uncertainty (Exhibit 11). The model has an overall accuracy of 96.8%. The true positive rate (sensitivity) was 94.4%, and the true negative rate (specificity) was 98.3%, indicating that the model can correctly classify both positive and negative cases. The features "Overweight," "Heart Rate," and "High Blood Pressure" were identified as the most significant factors illustrating the combined effect on sleep disorder (Exhibit 12). The area under the ROC curve (0.96) further confirmed the model's effectiveness and reliability (Exhibit 13).

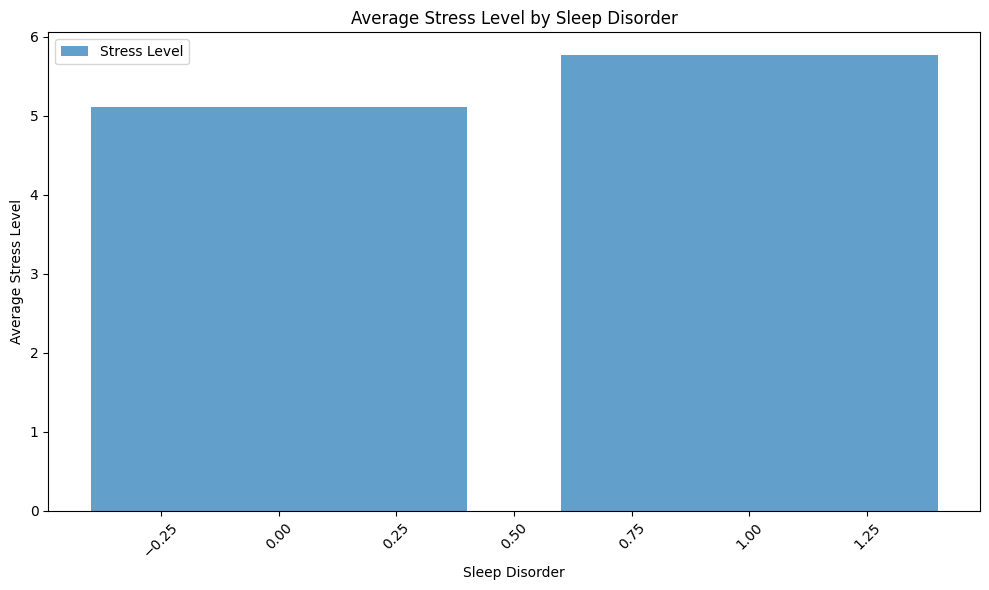
The third model being applied for prediction is Random Forest, which combines predictions from multiple classification trees to produce a final prediction. We initialized the random forest with ten trees and reported the model score, the confusion matrix, and feature importance (Exhibit 14). With this initial model, although it produced an accuracy of 93.6% on the test set, the sensitivity is merely 86.1%, with the top five important features: high\_blood pressure, Age, Sleep\_Duration, Daily\_Steps, and Occupation\_Doctor (Exhibit 15). Thus, we decided to refine the model further with 'GridSearchCV" to find the best combination of parameters. In the setup, we put many choices for the number of trees, the depth of trees, and the number of folds in the cross-validation process, etc. As a result, the best model contains 50 trees with a depth of 5 with an average validation accuracy of 93.2% (Exhibit 16). More importantly, it yielded a sensitivity of 94.4% and an overall accuracy of 96.8%, which is quite a great improvement (Exhibit 17&18). Moreover, the new top five important features of this best model turned out to be Overweight, Age, high\_bp, Sleep\_Duration, and Daily\_Steps (Exhibit 19).

The fourth model we explored is a Multilayer Perceptrons (MLP) Neural Network. Since our data is structured and cleaned, MLP is more suitable than other sequential data models like CNN, RNN, or Transformer. Classifying sleep disorder is a binary classification task, so the output layer will involve a sigmoid function. To select the best hyperparameters, we used GridSearchCV function to explore the learning rate, solver function, L2-Regularization weight, activation function, hidden layer size, and max-iteration. Eventually, our experiment indicated that 'adam' as the solver, 30 perceptrons for a single hidden layer, and 'relu' activation function with inv-scaling learning rate perform the best. The test result achieved roughly 81.9% of accuracy. We are aware that this data could be too little to underlie some complicated pattern for neural networks to fully reach its potential.

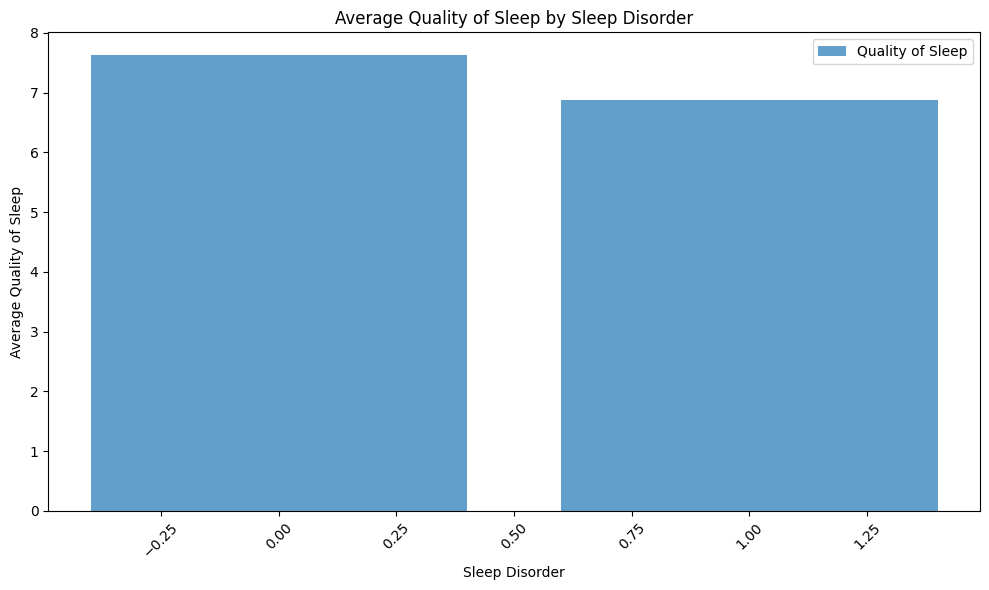
Logistic Regression, Classification Tree, and Random Forest demonstrate similarly high accuracies (Exhibit 20), ranging from 95% to 97%. Among these, we prefer Random Forest. The logistic regression model has two notable drawbacks: the low number of predictors and the choice of the threshold. The classification tree model has high variance in prediction results, meaning that small changes in parameter values can lead to significant changes in the model's structure. On the other hand, Random Forest aggregates multiple classification tree models, effectively addressing the issues of overfitting and high variance. Additionally, it can capture a broader range of patterns, which is especially beneficial as the data size increases.

Knowing that Random Forest is the best predictive model and uncovering those key factors, our findings offer valuable applications in the healthcare industry. For individuals, it assesses sleep disorder risk and suggests actions. For the public, it informs clinical decisions, public health policies, and advances in screening technologies.

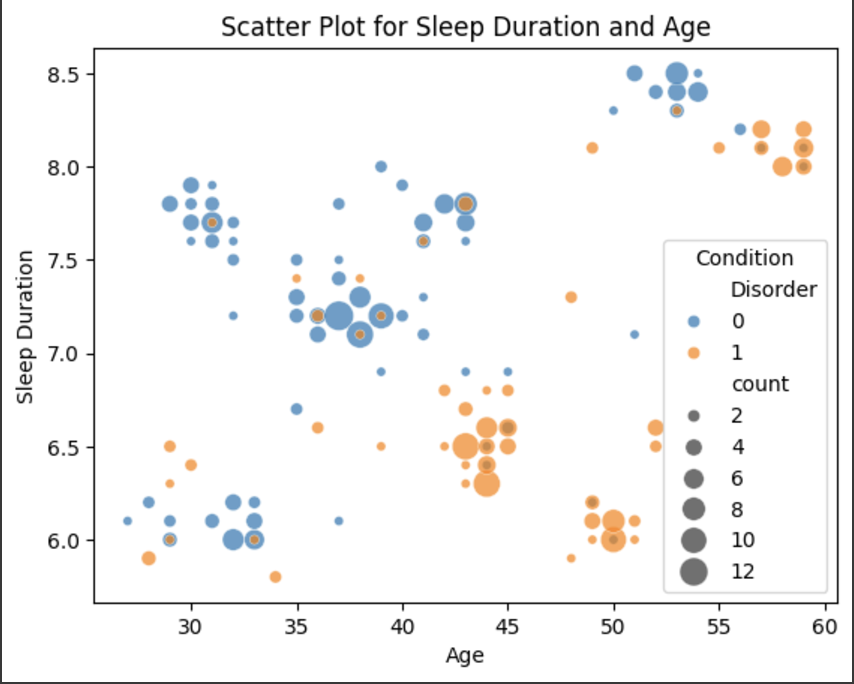
**Appendix**

****

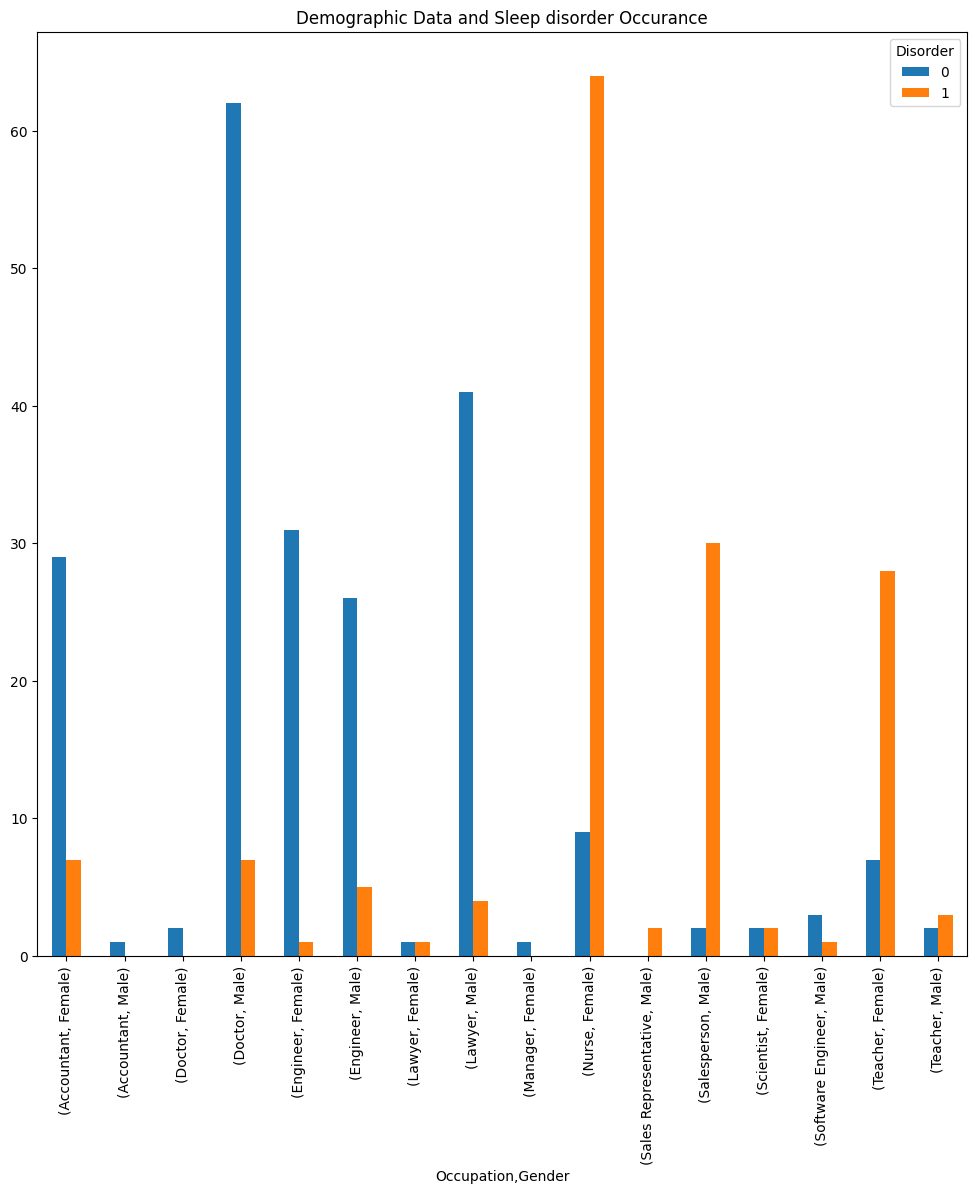
**Exhibit 1 (Sleep Disorder vs Stress Level Bar Chart)**

****

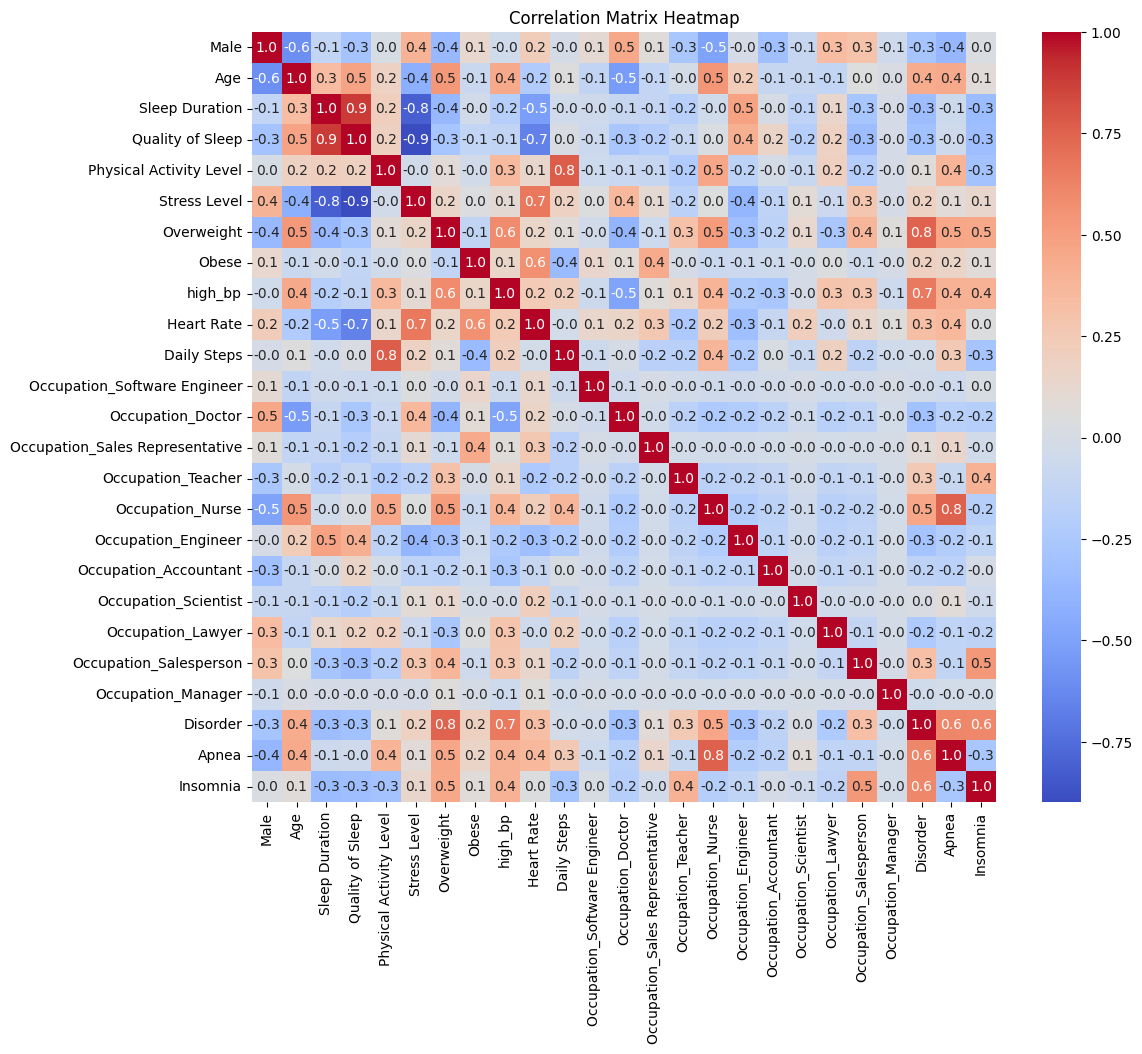
**Exhibit 2 (Sleep Disorder vs Sleep Quality Bar Chart)**

****

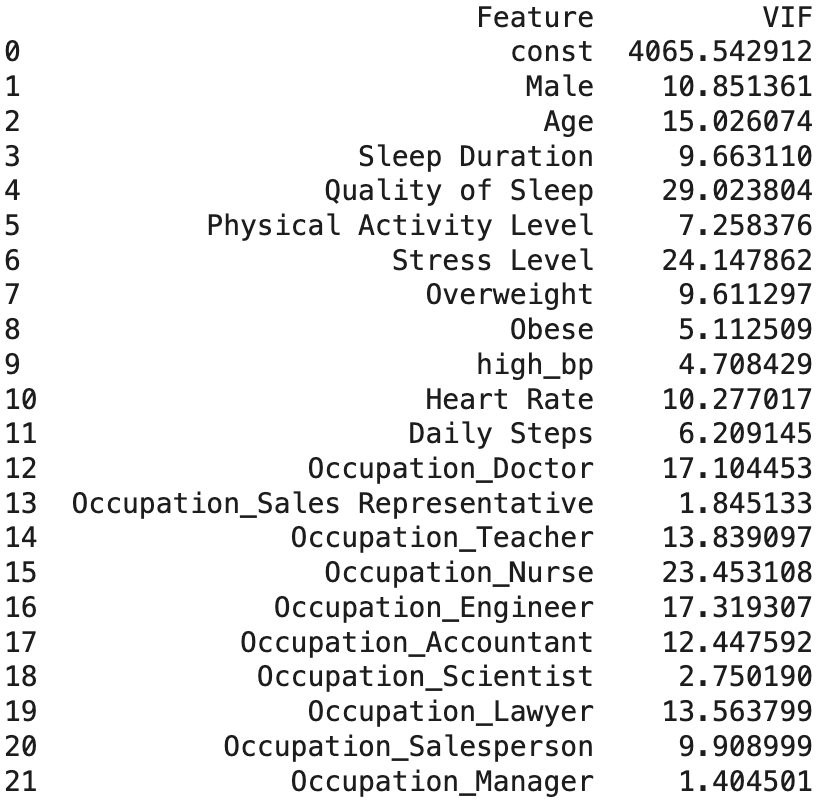
**Exhibit 3 (Sleep Disorder vs Duration vs Age Scatter Plot)**

****

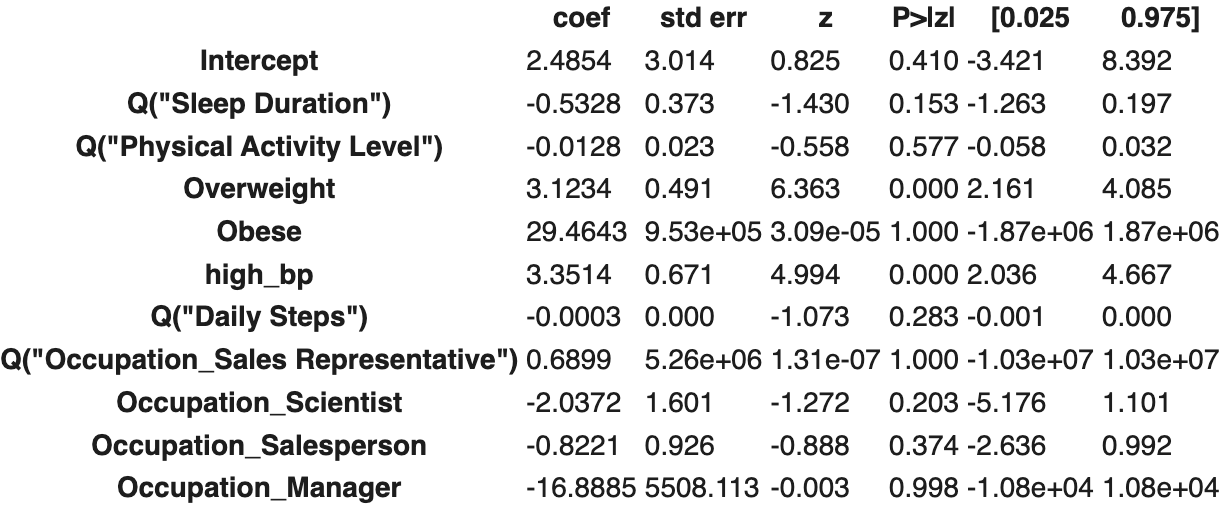
**Exhibit 4 (Sleep Disorder vs Demographic Data Bar Chart)**

****

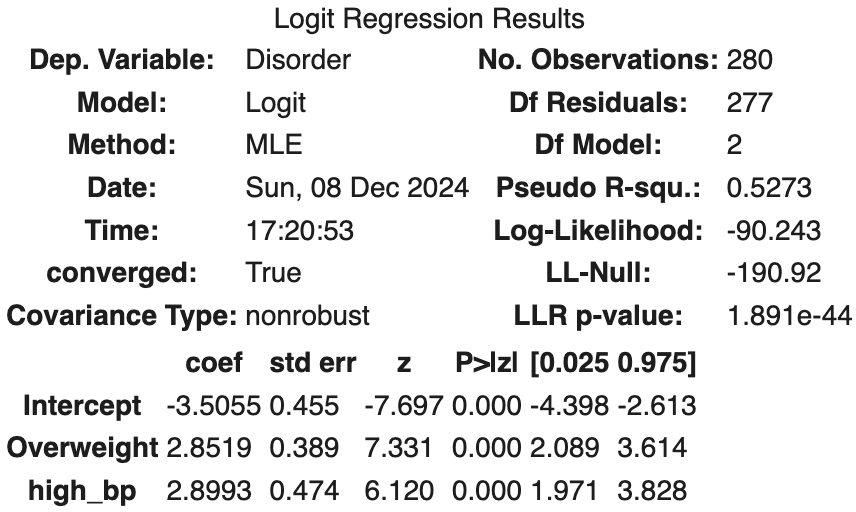
**Exhibit 5 (Correlation Matrix )**

****

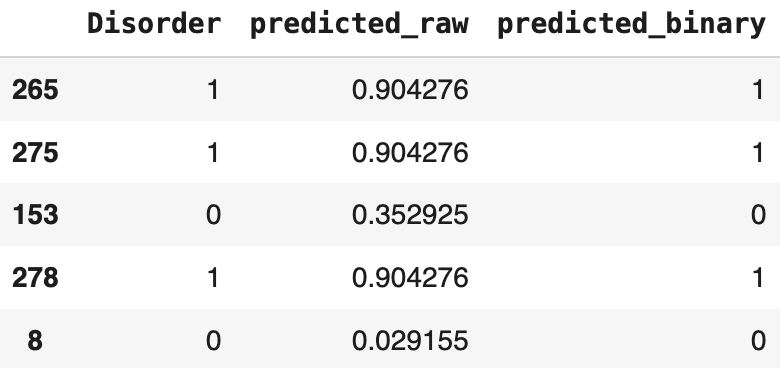
**Exhibit 6 (VIF Value Table)**

****

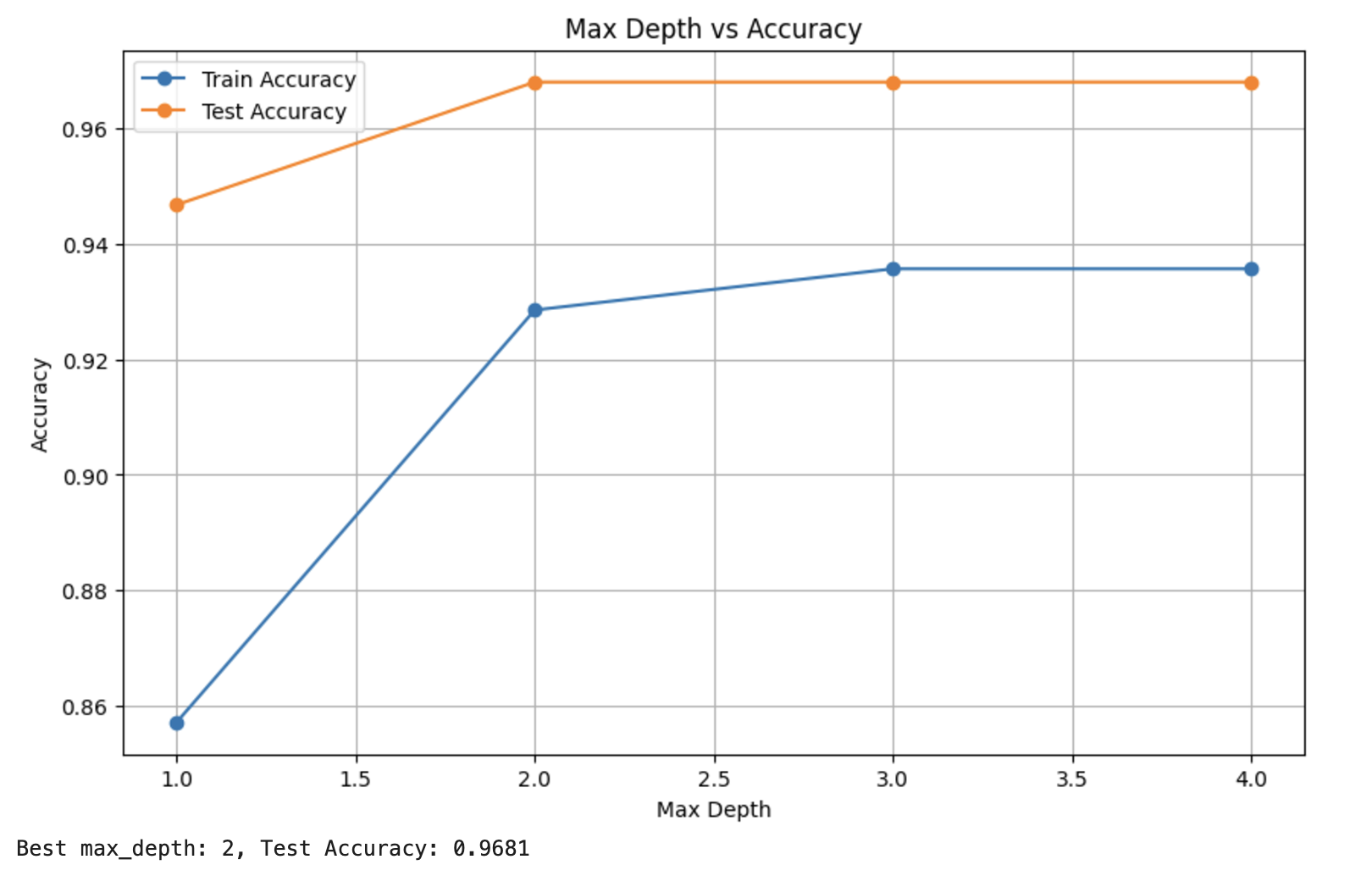
**Exhibit 7 (Model 1 Summary Table)**

****

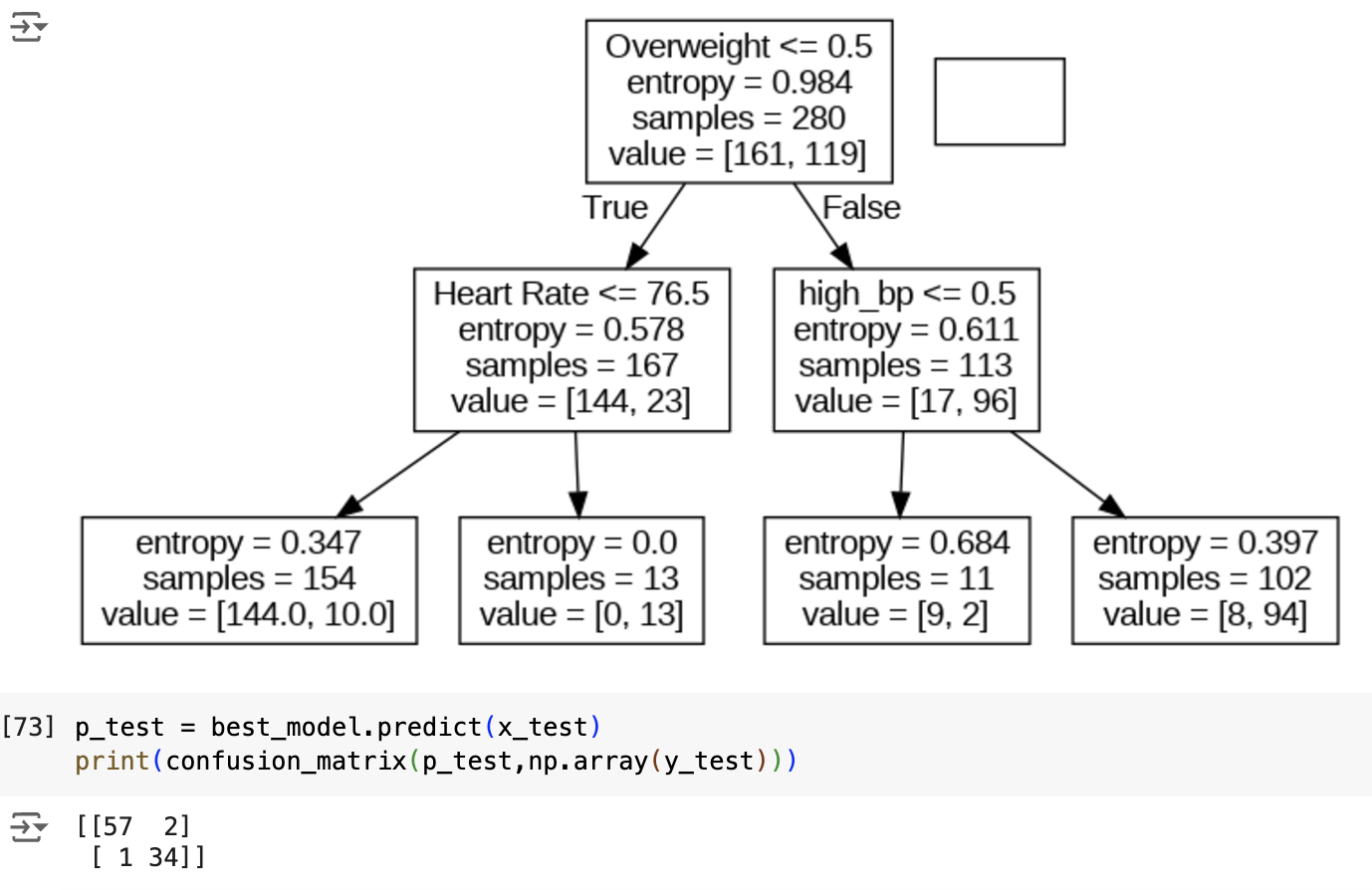
**Exhibit 8 (Model 2 Summary Table)**

****

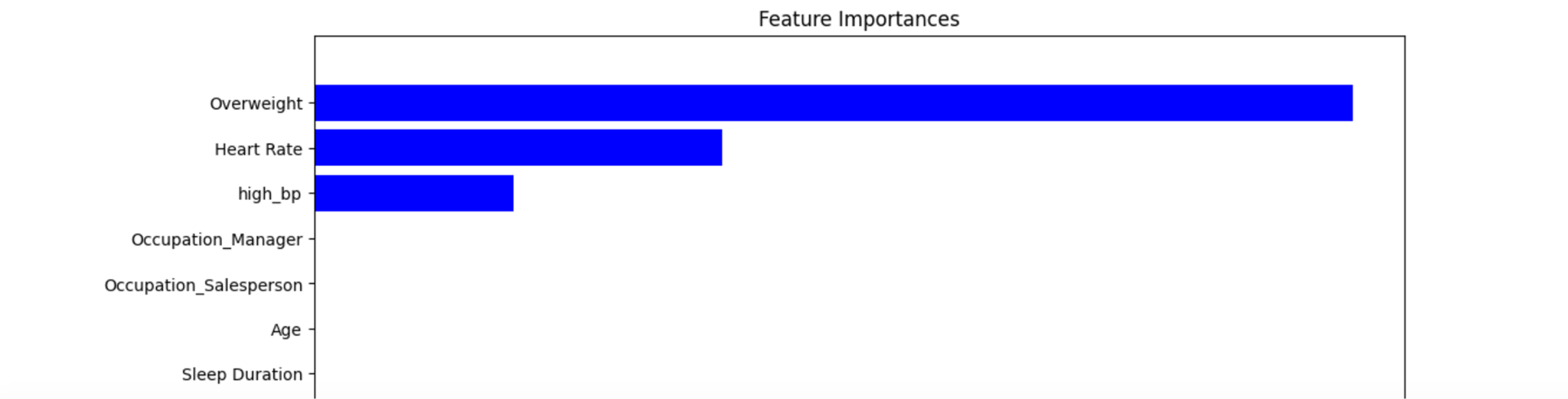
**Exhibit 9 (Prediction Result Table)**

****

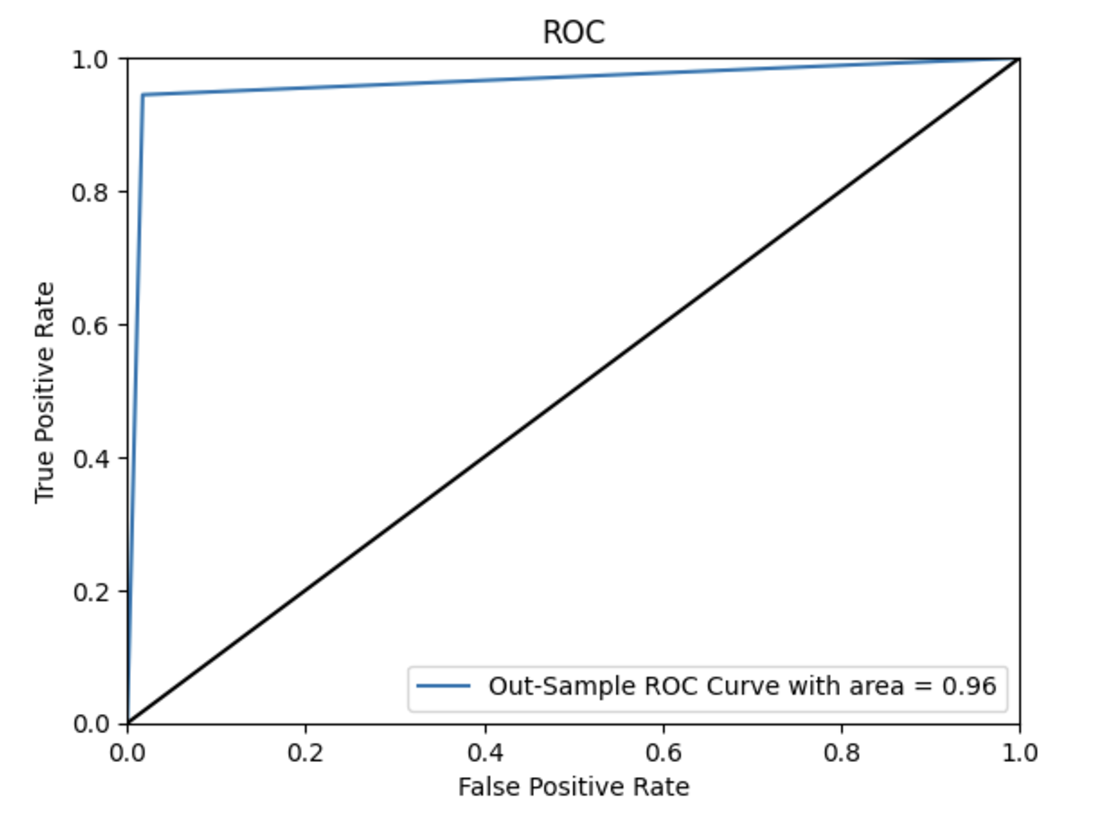
**Exhibit 10 (Best max\_depth)**

****

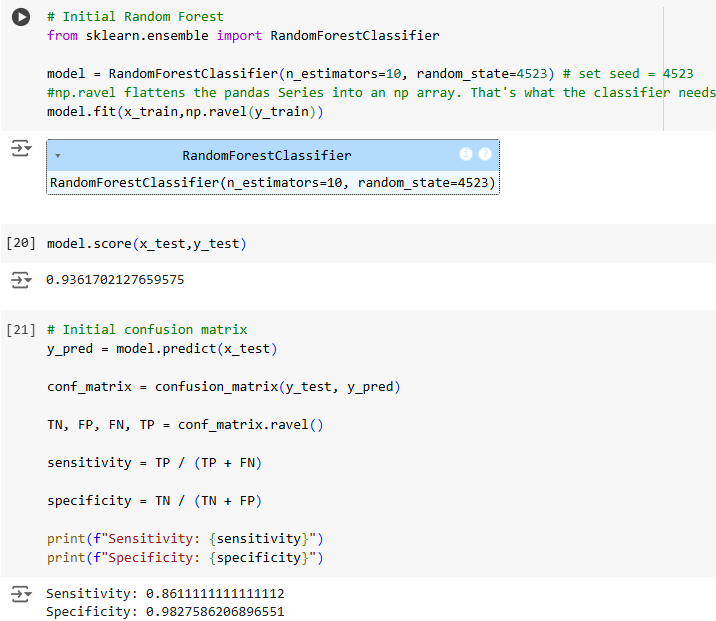
**Exhibit 11 (Classification Tree)**

****

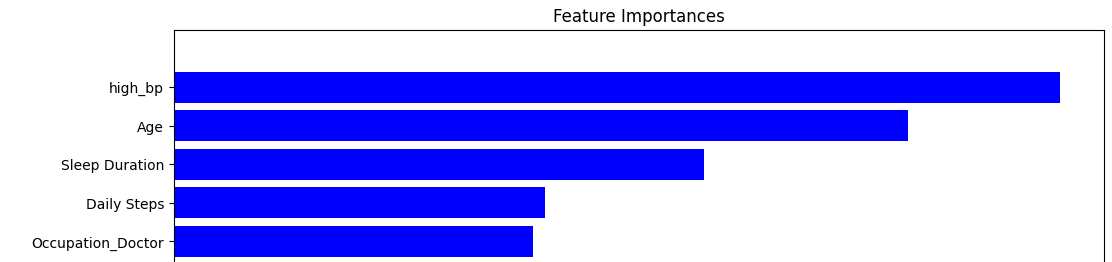
**Exhibit 12 (Feature Importance)**

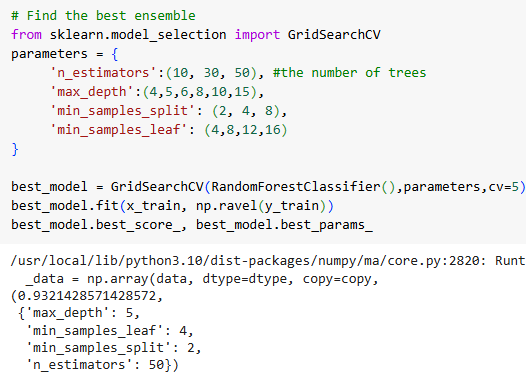
****

**Exhibit 13 (Classification Tree ROC)**

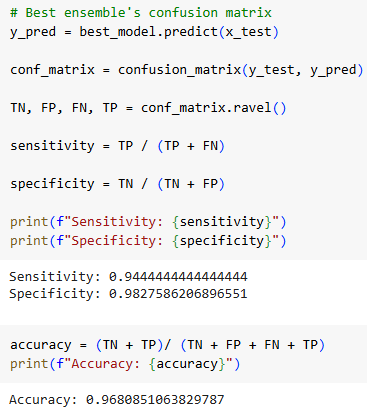
****

**Exhibit 14 (Initial Random Forest)**

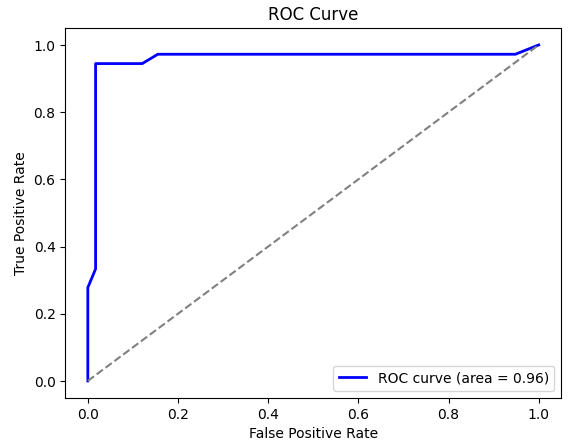
**Exhibit 15 (Initial Feature Importance)**

****

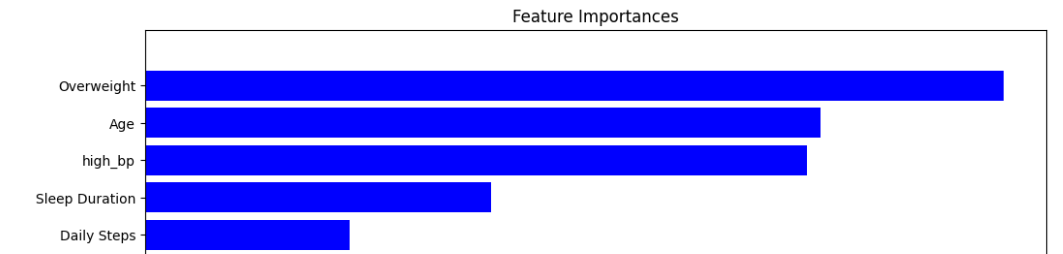
**Exhibit 16 (Best Ensemble)**

****

**Exhibit 17 (Best Ensemble’s Confusion Matrix)**

****

**Exhibit 18 (Best Ensemble ROC)**

****

**Exhibit 19 (Best Ensemble Feature Importance)**

|  | **Model Accuracy** |
| --- | --- |
| **Logistic Regression** | **95.7%** |
| **Classification Tree** | **96.8%** |
| **Random Forest** | **96.8%** |
| **Neural Network** | **81.9%** |

**Exhibit 20 (Model Comparison Table)**

**Source of Data**:

<https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data>

**Work Cited**:

1. Alnawwar, Majd A., et al. "The Effect of Physical Activity on Sleep Quality and Sleep Disorder: A Systematic Review." \*Cureus\*, vol. 15, no. 8, 16 Aug. 2023, e43595. National Center for Biotechnology Information, https://pmc.ncbi.nlm.nih.gov/articles/PMC10503965/.