Predicting Sleep Disorders

Team Member Names: Ina Holtschmit

Project Title: Predicting Disordered Sleeping in Individuals

Problem Statement

Obstructive Sleep Apnea (OSA) and Insomnia are two out of the four most common sleep disorders that affect the general population. OSA manifests as an interruption of regular breathing during sleep, where a person often gasps or makes snorting noises. This repetitive disruption of oxygen flow will result in increased fatigue, and if comorbid with other health conditions, can pose a serious health risk (https://www.cdc.gov/sleep/about_sleep/key_disorders.html). Insomnia is a sleep disorder that prevents individuals from getting an adequate amount of sleep. This can be due to having problems falling asleep, being unable to stay asleep throughout the night, or waking up hours before intended with the inability to fall back asleep. Insomnia, like OSA, also leads to excessive daytime fatigue, and consequently, poor performance throughout the day.

From the perspective of a health care provider, being able to predict the presence of OSA or insomnia in a patient by using a small amount of information about their lifestyle, would be greatly beneficial to both the patient and provider. Early prediction is important to prevent later health complications that would otherwise arise due to a lack of diagnosis (reference 2). This could ultimately result in lower health costs, as well as potentially saving a patient's life. This report aims to predict whether an individual has a sleep disorder using information about their cardiovascular health (blood pressure and heart rate), sleep metrics (sleep duration and quality), and lifestyle factors (general physical activity level, stress levels, and BMI category). A variety of classification models will be trained to predict whether they are affected by a sleep disorder.

Data Source

The data set used to investigate and answer the problem statement is open sourced from Kaggle, with the title "Sleep Health and Lifestyle Dataset" and made available by Laksika Tharmalingam. This data set is synthetic, meaning that it is manufactured data that is modeled after original data, resulting in new data values with the same statistical properties and distributions of the original data. It is composed of 374 instances, with 11 predicting variables (Gender, Age, Occupation, Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, BMI category, Blood Pressure, Heart Rate, Daily Steps), and the one response variable being "Sleep Disorder" ("None", "Sleep Apnea", and "Insomnia").

Data Cleaning/Preprocessing

Of the predictor variables, BMI category originally held the values "Normal", "Overweight", and "Obese", of string type, so the BMI categories were changed to dummy variables. Similarly, the response variable "Sleep Disorder", held the values "None", "Sleep Apnea", and "Insomnia". These values were changed to 0, 1, and 2, respectively, for model compatibility. The attribute

"Occupation" was deleted, as only a handful of occupations were present, which is not reflective of the real world. "Blood Pressure", originally in the format "###/##", was split into two columns (systolic pressure and diastolic pressure). Of the 374 subjects/instances, 58% of them did not have a sleep disorder, 21% of them had sleep apnea, and 21% of them had insomnia. Of all the classification models used, only Logistic Regression required scaling of the data from 0 to 1, for which the StandardScaler function from sklearn preprocessing was used.

Methodology

The problem statement will be answered in two parts, however, both parts will still use the same classification models: Random forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gaussian Naïve-Bayes.

For part 1, the classification models will be trained to predict a binary response, whether an individual has no sleep disorder (0), or a sleep disorder (1). Due to the nature of the problem, the models will be evaluated with a greater weight on correct classification of the sleep disorder category, rather than overall model accuracy. For part 2, the classification models will be used to predict whether an individual has no sleep disorder (0), sleep apnea (1), or insomnia (2). Once again, the models will be evaluated with a greater weight on the accuracy of the "sleep apnea" and "insomnia" predictions rather than overall model accuracy.

During the initial exporatory data analysis, the distribution of independent variables were examined to confirm that the dataset lacked bias relative to a real world population (Karna, Bibek, et al). Additionally, the distribution of the lack of and presence of sleep disorders in the data was investigated to confirm it accurately reflected the distribution of the general population.

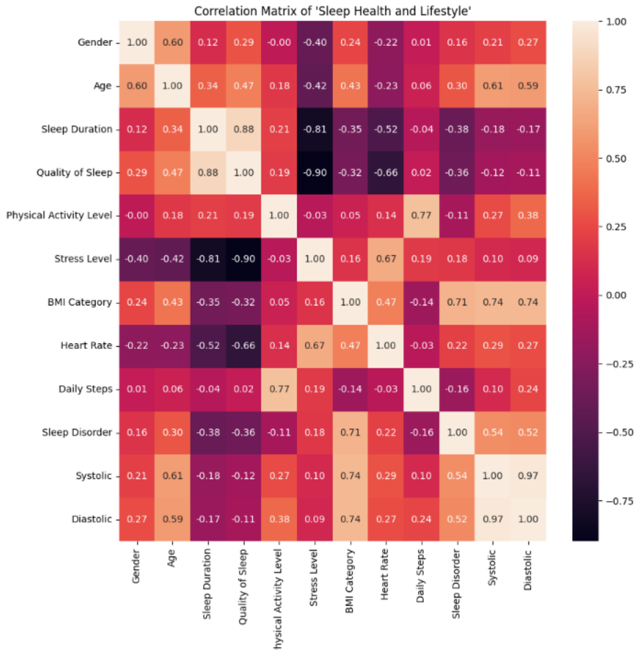
The dataset was split randomly to create the training and testing data sets, where the former is made up by 75% of the original data, and the latter by 25%. Using the train_test_split function from the sklearn library, the stratify parameter was used. This ensured that the train and test sets both contained instances that reflected the proportion found in the overall data. For example, since the original data contained 58% instances of class 0, 21% of class 1, and 21% of class 2, the Part A random train and test splits contained 58% instances of class 0 and 42% instances of class 1, and the part 2 splits contained the same ratio of classes as the original data.

The tuning of hyperparameters will be explored in each of the models. For the KNN model, the optimal number of neighbors will be explored. For the Random Forest model, the optimal number of trees relative to computational expense will be explored. For SVM, a range of C values will be explored for an optimal hyperparameter margin. For the Decision Tree model, the number of leaf nodes and maximum depth of the tree will be explored. Due to the size of the dataset not being large, cross validation measures will be used to get an average of each models respective results.

Python libraries such as pandas, numpy, matplotlib, seaborn, and sklearn, are used in the data wrangling and modeling process.

Evaluation and Final Results

Figure I and II.   EDA Correlation Matrix | Decision Tree with Max Leaf-Node of 10



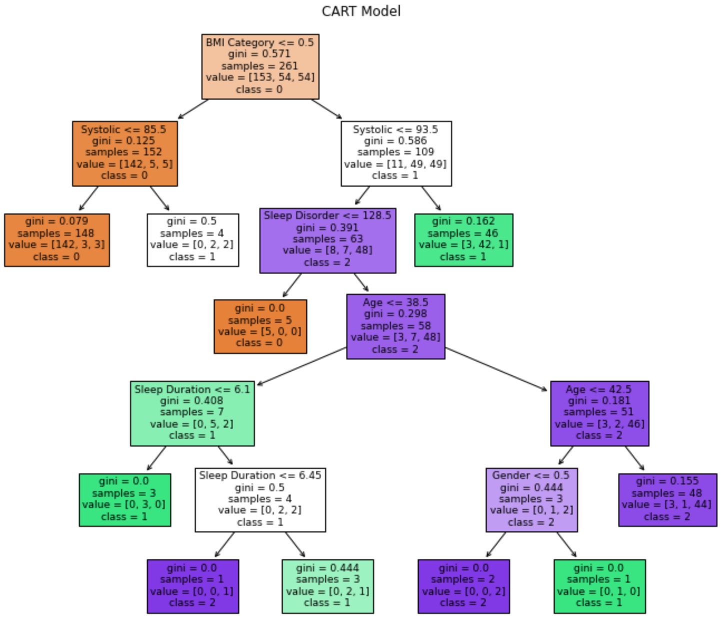
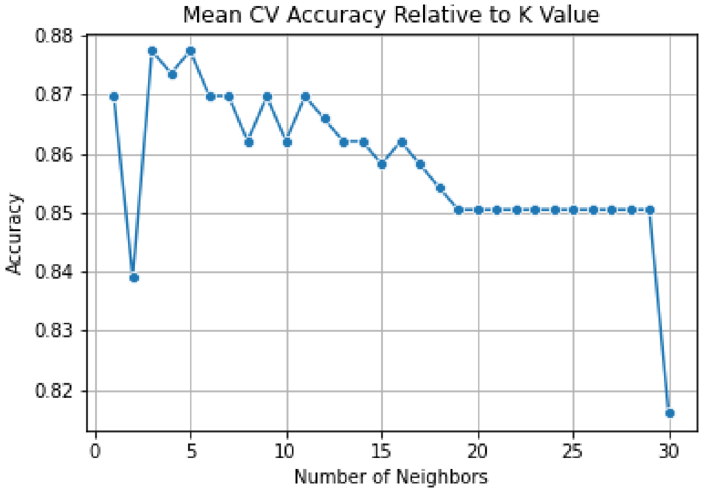

CART Model
BMI Category <= 0.5 gini = 0.571 samples = 261 value = [153, 54, 54] class = 0
Systolic <= 85.5 gini = 0.125 samples = 152 value = [142, 5, 5] class = 0
Systolic <= 93.5 gini = 0.586 samples = 109 value = [11, 49, 49] class = 1
gini = 0.079 samples = 148 value = [142, 3, 3] class = 0
gini = 0.5 samples = 4 value = [0, 2, 2] class = 1
Sleep Disorder <= 128.5 gini = 0.391 samples = 63 value = [8, 7, 48] class = 2
gini = 0.162 samples = 46 value = [3, 42, 1] class = 1
gini = 0.0 samples = 5 value = [5, 0, 0] class = 0
Age <= 38.5 gini = 0.298 samples = 58 value = [3, 7, 48] class = 2
Sleep Duration <= 6.1 gini = 0.408 samples = 7 value = [0, 5, 2] class = 1
Age <= 42.5 gini = 0.181 samples = 51 value = [3, 2, 46] class = 2
gini = 0.0 samples = 3 value = [0, 3, 0] class = 1
Sleep Duration <= 6.45 gini = 0.5 samples = 4 value = [0, 2, 2] class = 1
Gender <= 0.5 gini = 0.444 samples = 3 value = [0, 1, 2] class = 2
gini = 0.155 samples = 48 value = [3, 1, 44] class = 2
gini = 0.0 samples = 1 value = [0, 0, 1] class = 2
gini = 0.444 samples = 3 value = [0, 2, 1] class = 1
gini = 0.0 samples = 2 value = [0, 0, 2] class = 2
gini = 0.0 samples = 1 value = [0, 1, 0] class = 1

Fig III and IV. KNN K-Value Hyperparameter Tuning | Random Forest Number of Trees



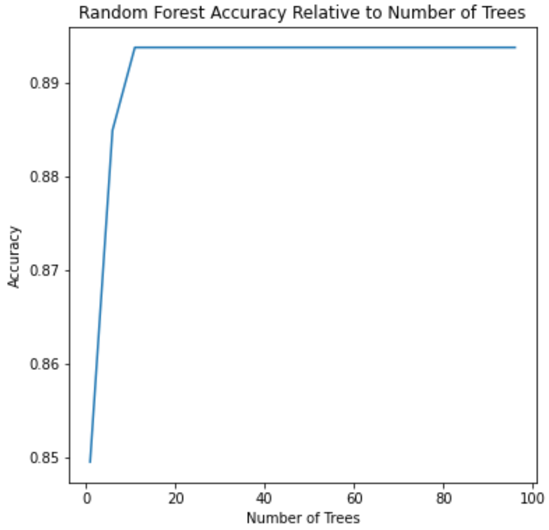


Table I. Multi-Classification Model Accuracies

Model	**No Disorder**	**Sleep Apnea**	**Insomnia**
Random Forest	0.9848	0.8333	0.8261
Decision Tree	0.9848	0.8333	0.6957
SVM	0.9848	0.9167	0.7826
KNN	0.9545	0.7917	0.7826
Logistic Regression	0.9848	0.8333	0.7826
Gaussian Naïve-Bayes	0.9697	0.8333	0.7826

Table II. Classification Model Accuracies

Model	**No Disorder**	**Disorder**
Random Forest	0.9565	0.8621
Decision Tree	0.9565	0.8621
SVM	0.9565	0.9310
KNN	0.9130	0.8276
Logistic Regression	0.9565	0.9310
Gaussian Naïve-Bayes	0.9565	0.9310

Figure V. Multi-Classification Model Confusion Matrices of No Disorder, Apnea, or Insomnia

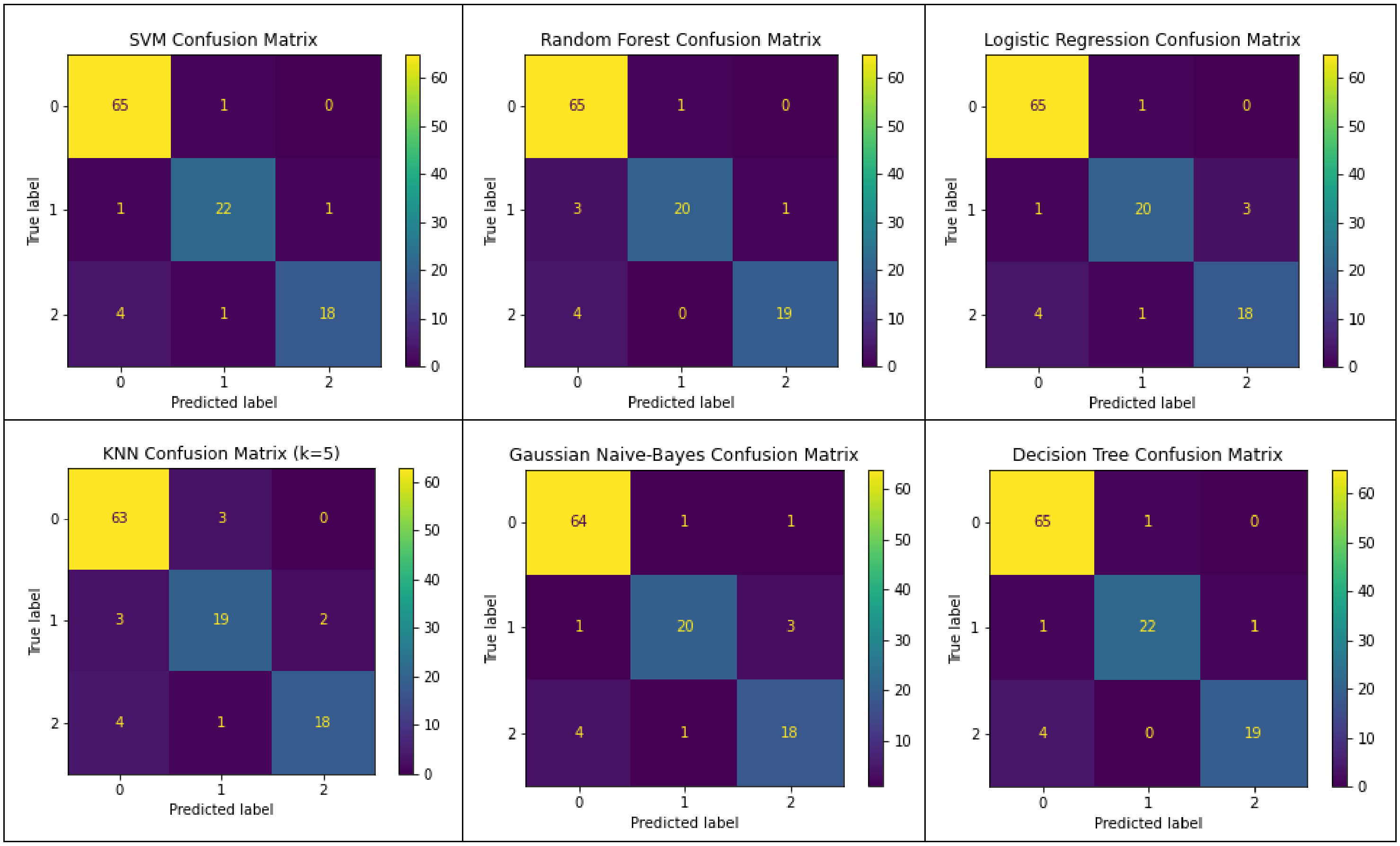


Figure VI. Classification Model Confusion Matrices of No Disorder vs Disorder

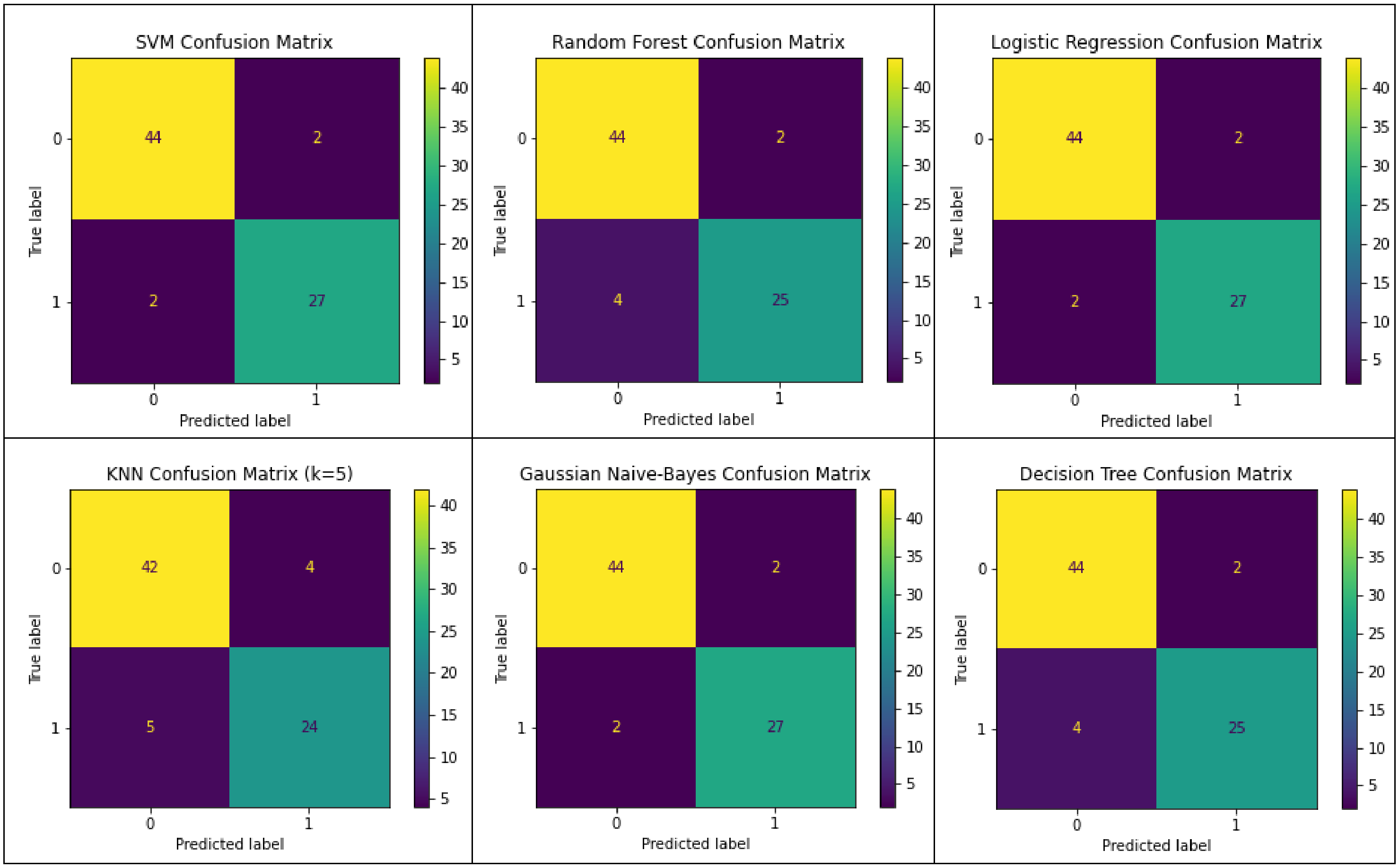


Figure VII. Final Results of Classification Models

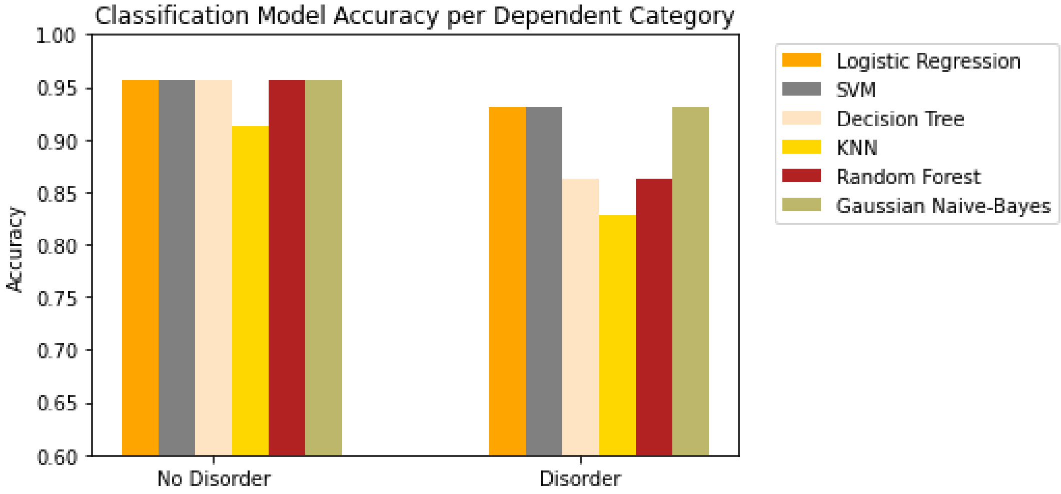
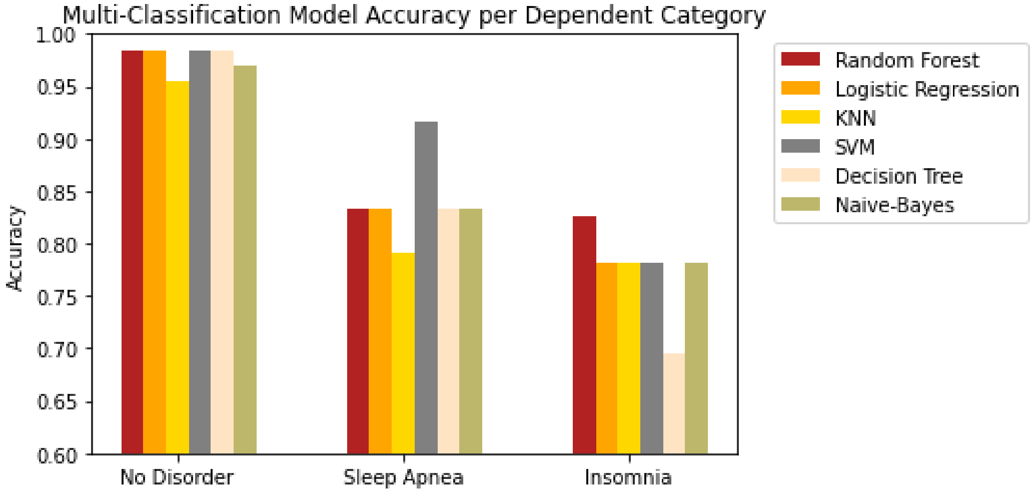


Figure VIII. Final Results of Multi-Classification Models



Discussion

Unsurprisingly, true positive accuracy rates where much higher in the traditional binary classification models (Part A), where prediction was based on the no disorder (0) or disorder (1) class labels, compared to the true positive accuracy rates in the multi-classification models, where prediction was based on the no disorder (0), sleep apnea (1), or insomnia (2) class labels (Part B). For model evaluation, overall accuracy it not being considered. The reason for this, is that there is no benefit for health professionals to accurately predict whether a person does not have a disorder. On the other hand, there is a benefit to being able to predict whether a person may have a sleeping disorder simply from looking at their vitals and general lifestyle questions. The goal is for the model to “trigger” whether an individual is classified as having disordered sleeping, and consequently informing a health professional to ask questions inquiring about possible symptoms to correctly diagnose the problem.

Another consideration that needs to be taken into account, is that insomnia is very easily self-diagnosed by an individual, whereas sleep apnea is not due to the individual being unconscious when the symptoms are present. Due to this, we are more interested in accuracy of sleep apnea predictions than those of insomnia in Part B.

For Part A, all models had a “no disorder” accuracy prediction of 96%, with the exception of KNN, which had an accuracy of 91%. For “disorder” accuracy prediction, Logistic Regression, SVM, and Gaussian Naïve-Bayes all had the same accuracy score of 93%, with the remaining models having scores below 90%. For Part B, all models had very high accuracy for the “no disorder” label, varying from 95% to 98%. However, the accuracies for “sleep apnea” predictions dropped to about 83%, with SVM being the outlier with a high score of 92%. The accuracies for “insomnia” dropped even further, to a range of 83 to 78%, with the Decision Tree model being the outlier with a low score of 70%.

Suprisingly, tuning hyperparameters did not have much of an effect, if any at all, in nearly all of the models. After cross-validation in KNN models, k = 5 was the chosen number of neighbors to be used in the final version of the model. For Random Forest, only a small amount of trees were needed to reach the optimal accuracy that could be found in the model. For the Decision Tree model, the maximum depth allowed and the number of leaf nodes was explored, but neither had an effect on true postive accuracy scores. Lastly, for SVM, a range of C values were used to train the model in cross-validation to explore whether increasing the penalization parameter would increase true positive accuracy scores, but it did not.

Some limitations of this study are that sleep apnea and insomnia can be comorbid disorders, meaning that both can be present in a patient at once. This partially explains why the results of the multilabel classification model are not as promising as those found in Part A.

For a final case of model evaluation, in Part A, Logistic Regression, SVM, and Gaussian Naïve-Bayes can be used interchangeably in a real-life implementation as they all returned the same accuracies. In the case of Part B, SVM was by far the best choice. Even though the Random Forest model had a slightly higher accuracy for "insomnia" classification, we are more interested in being able to diagnose sleep apnea. Taking that into account, SVM outperformed all other models by 9%.

The implications of these findings are that data modeling could be potentially used to flag hospital patients for which the hospital already posseses very basic lifestyle records of and vitals from previous visits. If a patient is flagged, a healthcare provider could reach out and inquire about symptoms, potentially resolving a sleep disorder that could be negatively affecting their quality of life without them knowing.

References

[1] Karna, Bibek, et al. "Sleep Disorder." *National Library of Medicine*, NCBI Bookshelf, www.ncbi.nlm.nih.gov/books/NBK560720/.

[2] "Key Sleep Disorders - Sleep and Sleep Disorders." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 14 Dec. 2022, www.cdc.gov/sleep/about_sleep/key_disorders.html.

[3] Tharmalingam, Laksika. "Sleep Health and Lifestyle Dataset." *Kaggle*, 18 Sept. 2023, www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset.