Project Title: ‘Sleep Disorder Prediction using Supervised Machine Learning’

Introduction

In this project, we used a Supervised Machine Learning algorithm (Random Forest) to predict and identify sleep disorders such as sleep apnea and insomnia. The dataset we used was from Kaggle and contained information about different lifestyle factors affecting/causing sleep disorders. We believe that this project can help many healthcare professionals better evaluate whether patients have sleep disorders and advise them accordingly.

Background

The main AI concept we used in this project is Supervised Machine Learning. Specifically, we decided to use the Random Forest algorithm. Random Forest is an ensemble learning method. It is derived from the idea that it is an ensemble of decision trees, and each tree is constructed using a random subset of data and features.

Additionally, we modeled the tabular data through three different models and the main concept present in this was the splitting of data into training and testing. By splitting the data, we were able to train and test each model and determine how accurate the model is.

Methodology

The AI technologies used in this project were based on the concept of Supervised Machine Learning.

The Supervised Machine Learning algorithm that was implemented was Random Forest. Random Forest is a common Machine Learning algorithm that combines the output of many decision trees to reach one result.

The Random Forest algorithm works as follows:

1. Preparing the Data: Random Forest is usually used for classification and regression problems. In the case of our project, we are using Random Forest for classification of sleep disorders. We start with our labeled Kaggle dataset where each data point has a set of features and a target variable. Our target variable was sleep disorder.

2. Bootstrapping: Multiple subsets of the original dataset are created through bootstrapping. In bootstrapping, the algorithm randomly selects data points with replacement to create new datasets of the same size as the original dataset. This means that some data points could appear multiple times in these subsets and some not at all.

3. Random Feature Selection: For each subset of data the algorithm creates, it also randomly selects a subset of features. This allows the introduction of diversity and reduced risk of overfitting.

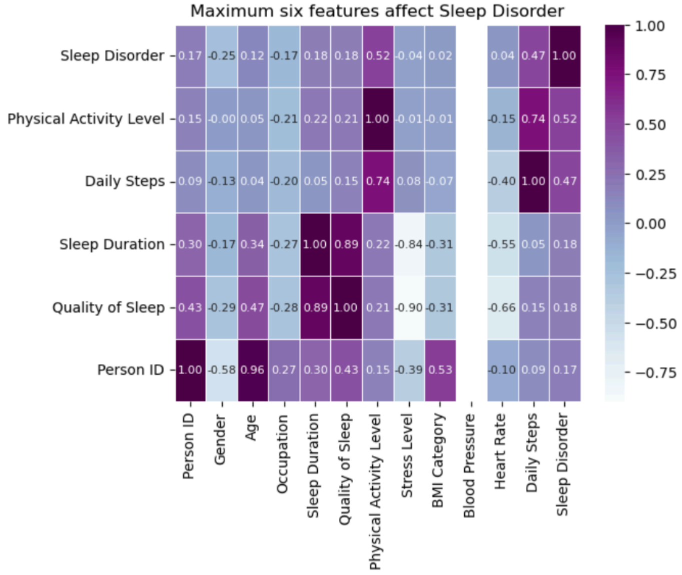
4. ‘Growing’ Decision Trees: For each bootstrapped dataset, a decision tree is grown. These trees are called “weak learners” because of their limited predictive power. The decision trees are built by selecting the best split at each node of the tree based on criteria. The tree continues to grow until the criteria is met.

5. Averaging: Once the forest of decision trees has been created, each tree in the forest predicts the target variable. In the case of our project, the class that receives the most votes from the individual trees is considered the final prediction.

6. Feature Importance: Random Forest can provide information about the importance of each feature in making predictions. This helped us understand what lifestyle factors caused people to be more prone to sleep disorders.

Additionally, we implemented three models to determine the best model for the dataset. The three models were:

1. Gradient Boosting Classifier: The GBC Model is meant to adjust varying weights of individual, weak learners, to increase classification accuracy through re-sampling when implementing something like a Random Forest Algorithm. When implementing this model, we received 92% predictive testing accuracy.
2. Linear Regression: The Regression Model is designed to determine the strength of associations such as those between sleep disorder and lifestyle features. From this model, we received an 89.33% predictive testing accuracy. The correlations between the lifestyle features and sleep disorder are below.



3. Support Vector Classifier (SVM): The SVM is a Supervised Learning Model that is designed specifically for classification problems (such as our project). They are known for their effectiveness in handling linear and nonlinear classification. We received 97%

testing accuracy. To implement this, the trained ML model was loaded, then the prediction data was prepared using .predict from sci-kit learn. Using the fitted SVC model, we predicted whether a patient had a sleep disorder - insomnia or sleep apnea. A few rows of the dataset were then used to predict sleep disorders.

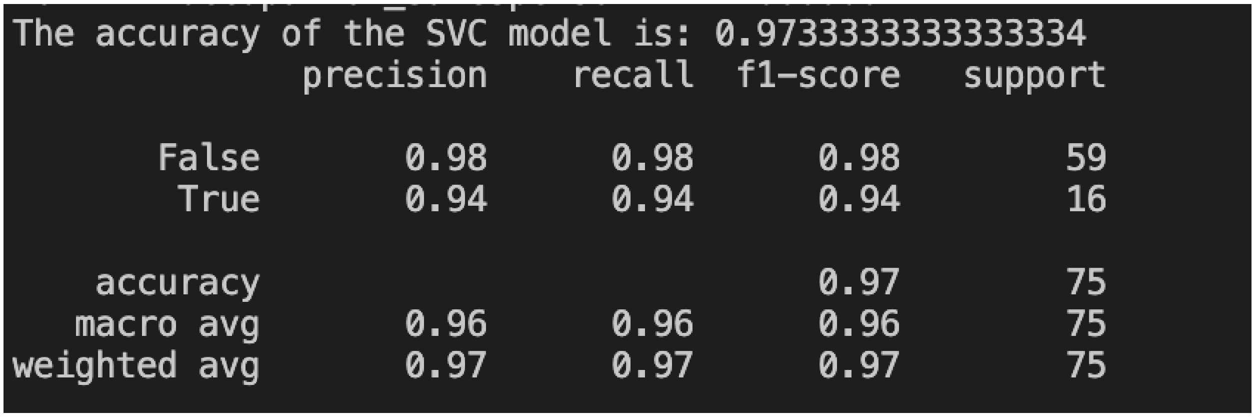
Results

We were able to achieve our project goal of predicting sleep disorders in patients. Our initial idea was to solely focus on one model but we realized that it was important to consider other models too.

While fitting three different models on the dataset, we made sure to analyze the models based on other metrics in addition to accuracy. The Support Vector Classifier (SVM) model had the best accuracy of about 0.97. We also developed a comparison report to evaluate our models with class imbalances. It was important to choose metrics that aligned with the goals of our project and the relative importance of different types of errors (accuracy, sensitivity, interpretability tradeoffs). The following results were used to weigh the model that was chosen.

1. **Precision:** It measures the accuracy of positive predictions made by the model (ratio of true positives to total number of positive - true positives + false positives).
2. **Recall (Sensitivity):** model's ability to identify all relevant instances from dataset (how many were correctly predicted)
3. **F1-Score:** harmonic mean of precision and recall - provides balance between these
4. **Support:** number of instances in dataset that belong to particular class - informative value
5. **Macro Average:** computes unweighted average of precision, recall and F1-score across all classes
6. **Weighted Average:** computes average of precision, recall and F1-score taking into account the number of instances in each class

The classification report for the SVC model is as follows:



```
The accuracy of the SVC model is: 0.9733333333333334
              precision    recall  f1-score   support

       False       0.98      0.98      0.98        59
        True       0.94      0.94      0.94        16

    accuracy                           0.97        75
   macro avg       0.96      0.96      0.96        75
weighted avg       0.97      0.97      0.97        75
```

The metrics in the classification include precision, recall, f1-score and support. These are important factors to determine how well the model performed on the dataset.

After much tuning, the SVC model performed better than the other two models, we were able to better predict whether the person had a sleep disorder or not. This fulfilled our project goal for accuracy. Since sleep data can be quite complex and non-linear, SVM's are adequate when trying to effectively capture these relationships due to the kernel functions which uncover hidden patterns. Additionally, there are outliers in sleep data and SVM is a valuable tool to detect these. Lastly, its margin-based classification abilities allows us to find the support vectors closest to a set decision boundary that is used to determine decisions the model makes when classifying data.

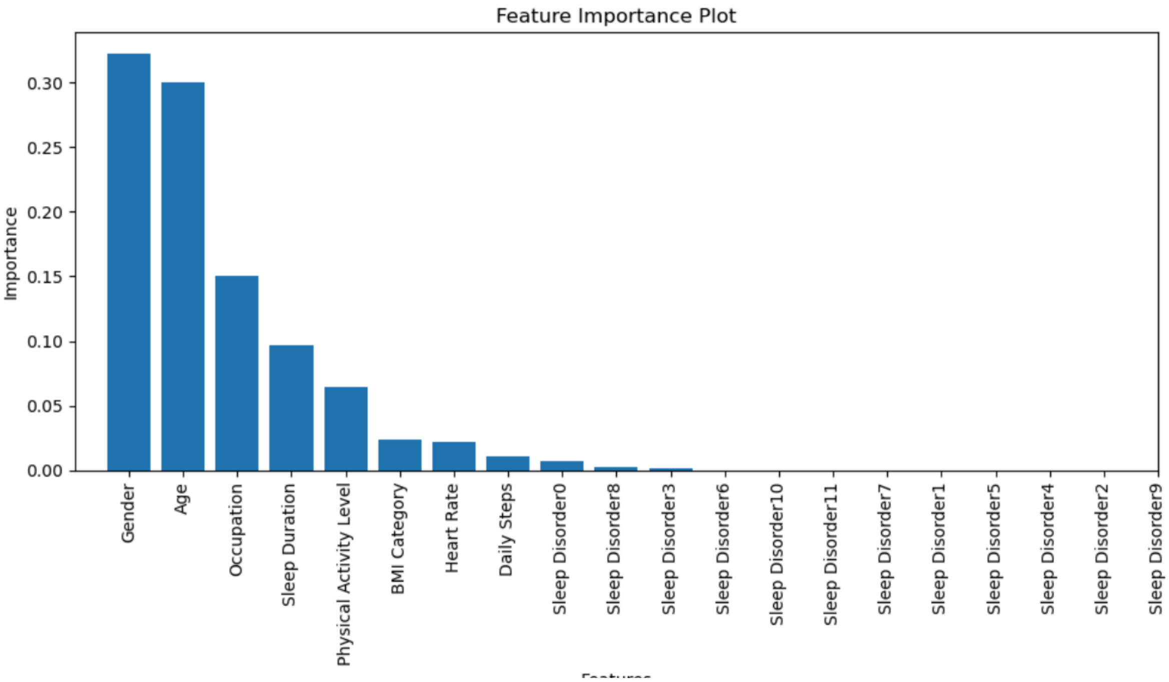
While limited in interpretability, it has stronger predictive powers. To make it more interpretable, we could have used post-processing techniques such as SHAP or LIME so that we had more information on feature contributions and individual predictions.

Our predicted results and analysis tools such as feature importance, etc. were then compared to current scientific sleep data. Different variables are commonly known to impact sleep quality in individuals. We found that Gender, Occupation, Sleep Duration, and Age (Older Adults) were important indicators of a sleep disorder among each model and current science. According to our feature map from the GBC model, gender is one of the most important dictators of sleep disorder. Science shows that insomnia, sleep apnea, and restless legs syndrome (RLS), affect gender at different rates. Women are more likely than men to be diagnosed with insomnia (40% higher). Additionally, major risk factors for sleep apnea are obesity and family history of sleep apnea in males. Further, older adults are more likely to suffer from a sleep disorder according to our GBC model and SVC model. In addition to changes in sleep architecture, circadian rhythm-related changes are also present in older adults, with many having an advanced phase rhythm resulting in an early bedtime and rise time

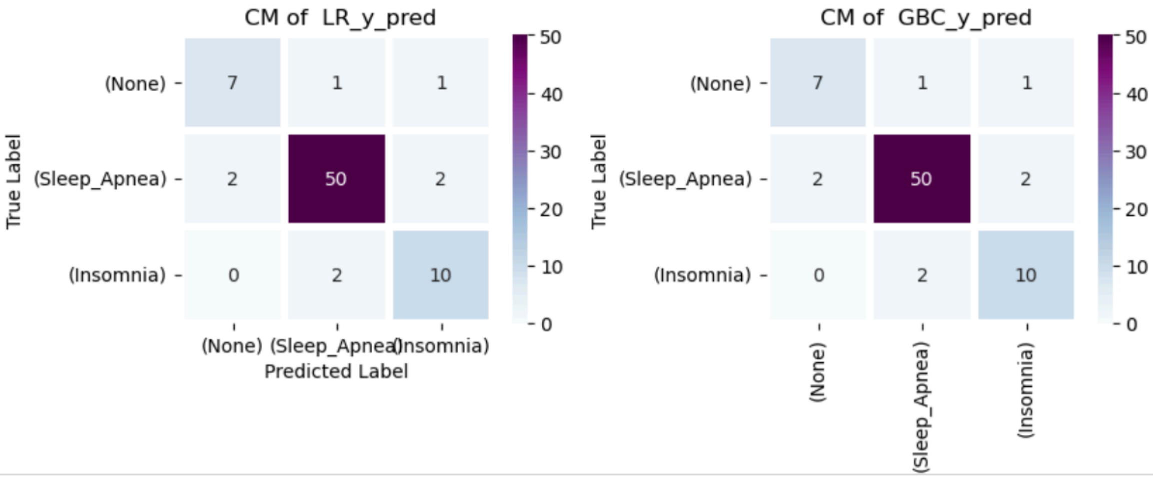
Discussion

To establish which model we would use as our predictive algorithm we had to consider several factors. Interpretability of a model refers to the ability to understand and explain how the model makes predictions. It's a crucial aspect of machine learning, especially in scenarios where decisions have real-world consequences, such as healthcare, finance, and legal applications. Interpretable models provide insights into the relationships between input features and output predictions, allowing humans to comprehend and trust the model's decisions. Here are some considerations related to model interpretability:

- **Feature Importance:** Understand which features have the most influence on the model's predictions. Techniques like feature importances from tree-based models and permutation importance can help identify significant features.



- **Coefficient Analysis:** In linear models, coefficients can provide insights into the direction and magnitude of the relationship between each feature and the target variable.
- **Sensitivity Analysis:** Understand how changes in feature values affect the model's predictions. Sensitivity analysis helps identify which features are most sensitive to variations. We used this as our validation method for the GBC and LR models.



- **Ethical and Fairness Considerations:** Interpretability can help identify bias and discrimination in model predictions, ensuring that the model's behavior aligns with ethical and fairness standards.

We needed a balance between complexity and interpretability to reach predictive accuracy and interpretability which is why we compared results to existing scientific data on sleep. Complex models might offer higher predictive accuracy but can be challenging to interpret. Simpler models tend to be more interpretable but might sacrifice some predictive power.

Conclusion

In conclusion, this project provides a simple and accurate way of stream-lining sleep disorder diagnosis. It is not only an accessible option for those in the medical field, it is also a strong method for doctors. Our implementation was simple, yet effective. We were able to adequately draw from interpretability and accuracy tradeoffs when determining the model that was chosen. This also proved to be an important aspect of validation, which either supported our models strength, or negated it. By using existing scientific knowledge and data surrounding sleep and sleep studies, we developed a more well-rounded algorithm that can pose a drastic benefit to the general public. The negative impacts of impaired sleep on lifestyle and health has become more prevalent over the years. One in every 15 Americans, or 6.62% of the total American population have a case of sleep disorders. Sleep is vital to health and daily performance (Houston Sleep Solutions). Knowledge of a tool that everyday consumers can use to identify issues with sleep can vastly contribute to improving their lives. Becoming knowledgeable of one's sleep habits, and finding the general patterns correlated with their sleep disorders (such as a certain profession and activity level, etc.) may promote change in their lives. This may be the solution for individuals finally resolving their lethargy and tiredness on a daily basis by exposing them to their sleep deficiency and can promote the need for sleep-aids, medical resolutions, EEG recordings, and more. It is a tool that can be useful to the average person and a sleep doctor.

Breakdown of Responsibilities

Created the GitHub for the project and handles all commits to the repository so no new code is lost due to any hardware/software failure. Downloaded the dataset from Kaggle and used matplotlib to visualize the data and provide an outline for preprocessing. Read data from the dataset and used the train test split provided by Sci-kit learn for the algorithm. Implemented the Support Vector Classifier model which generated an accuracy of 96%. Used the classification report in Sci-kit Learn to evaluate the model based on factors other than accuracy. Predicted whether a patient has sleep apnea or insomnia using the model. Added code from other team members onto Github. Did a lot of research about each model and how to properly evaluate the models using different factors and parameters. Worked on the peer feedback, proposal and project check in with other team members. Helped other team members navigate Visual Studio Code. Created the models for GBC and Linear Regression. Completed the validation method using a confusion matrix as well as the features importance plot. Also developed the heat map for the correlations between the six features affecting sleep. Helped other team members with their VS code and submitted code that I would work on to Ria who handled the GitHub. Produced one version of the pre-processing. Produced detailed work on the reflections, proposal, and project-check in. Pre-processed all data and created the layout plan document to track out work and items that needed to be completed. Fit the model to training data to look at accuracy and establish the hyperparameters. Implemented Cross - Validation. Worked on the models and the validation method we are using. Conducted research on how sensitivity would fit into our project and helped group members along the way in the various tasks we had. Worked on the peer feedback, proposal and project check in with other team members.