**Predicting Sleep Quality: Model Selection, Training, and Evaluation**

# **Algorithm Exploration**

We will explore the following three algorithms due to their capacity to manage multiple input features and their adaptability to non-linear data:

1. Random Forest (RF): An ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. Also, RF is known for its high accuracy, and ability to handle large data sets with higher dimensionality.
2. Gradient Boosting Machines (GBM): An ensemble learning method that builds models sequentially, with each new model attempting to correct the errors of the previous one. We choose GMBs due to their effectiveness in handling various types of data structures and distributions, including non-linear data. In addition, GBMs are also capable of handling missing data and providing importance scores for different features.
3. Support Vector Machines (SVM): A powerful classifier that works well in high-dimensional spaces and is suitable for cases where the number of dimensions exceeds the number of samples. In addition, SVM is useful for non-linear data thanks to its use of kernel functions, allowing it to adapt its decision boundary.

# **Training and Evaluation**

The models will be trained on a split dataset, using a portion of the data: 70% for training and 30% for testing. Cross-validation techniques, specifically K-fold cross\_validation, will be used to ensure that our model generalizes successfully to unseen data and to mitigate the risk of overfitting.

Also, to evaluate the model’s performance, we will use a combination of metrics including accuracy, precision, recall, and F1 score. Given the non-linear nature of our data, we will also examine the area under the Receiver Operating Characteristic (ROC) curve for classification tasks, which provides a comprehensive measure of performance across all classification thresholds.

**Hyper-Parameter Tuning**

For each model, specific parameters will be tuned using grid search to find the optimal combination that yields the best performance:

* RF: Number of trees, max depth of trees, min samples split, and min samples leaf
* GBM: Learning rate, number of estimators, max depth of trees, min samples split, and min samples leaf
* SVM: C, Gamma, and Kernel

# **Accuracy Determination**

The model’s accuracy will be determined primarily by its ability to predict outcomes correctly on the test set. Accuracy is a suitable metric for datasets where the classes are roughly balanced. However, to ensure a comprehensive evaluation, especially in the presence of imbalanced classes, we will also consider precision, recall, and the F1 score as supplementary metrics. These additional metrics will help us understand the balance between the model’s sensitivity and specificity.