**Report**

**1. Data Loading and Exploration**

The goal of this project is to predict the sleep quality of the participants.

**2. Data Preprocessing**

Data Cleaning

• autoclean() and klib.datacleaning() were not used

**3. Data Visualisation** is also shown with barplot based on value counts of each categorical columns.

**4. Model Training**

Splitted the dataset into training and test sets using an 80-20 split

**5. Model Selection**

Since the target variable is discretized and split into bins {'Poor' : <=2 ,'Okay' : >2 and and <=5 ,'Good': >5 and <=7 , 'Excellent' : > 7 }, it is a classification project. Therefore, we chose models where we can predict categorical features. Below models are used:

1. Logistic Regression - to learn the pattern between features and output and predict the right quality of sleep
2. Decision Tree - since tree models can handle both classification and regression, we used Decision Tree Classifier for our project to predict more accurately
3. Random Forest Classifier - it is also used because of the relatively low accuracy score in Decision Tree Classifier model, assuming that multiple trees with different root nodes would better predict the target variable
4. Gradient Boosting Classifier - we used this model to combine trees (base learners) sequentially to create a stronger model by learning its errors to improve accuracy
5. XGBoost - in order to improve the model’s generalization ability and for more complex patterns between features and the output
6. LightGBM - another GB model used to to achieve a balance between efficiency and accuracy

**6. Evaluation**

The following metrics were to evaluate the model

• Accuracy score - used for each model to see how accurately models works

• Classification report - used to see the model’s precision in predicting values and how accurately model predicts the positive cases. Also, as our models are classifier models, f1 score is also important since it shows the percentage of the positive predictions that are correct.

**7. Improve** - through hyperparameter tuning, we tried to improve the performance of our models. For tuning, GridSearch is used since the dataset is relatively small with around 5k rows.