Final Assessment Report

5CS037: Concepts and Technologies of AI

Student name: Prajwal Limbu

University ID: 2408429

Group: L5CG19

Title: Regression Analysis Report

Module leader: Simon Giri

Tutor: Ronit Shrestha

Submitted on: February 11, 2025

Abstract

This report will be predicting the sleep duration with the help of regression models. The dataset being used is the Sleep Health and Lifestyle Dataset, which consists of some lifestyle and health determinants. The models being used here are Linear Regression and Random Forest. The processes include EDA, model construction, hyperparameter tuning, and feature selection. The results are being compared based on accuracy, precision, recall, F1-score for regression. Random Forest made the most accurate predictions overall, indicating that the data is most accurately described by non-linear models.

Contents

1. Introduction

1.1 Problem Statement

To predict sleep duration based on health and lifestyle factors such as stress levels, physical activity, and sleep quality.

1.2 Dataset

- Name**:** Sleep Health and Lifestyle Dataset
- Source**:** Kaggle
- Description**:** Contains 374 entries with features like age, occupation, physical activity level, stress level, BMI category, and sleep quality.
- Alignment to UNSDG**:** Aligns with Goal 3: Good Health and Well-being by analyzing factors affecting sleep health.

2. Methodology

2.1. Data Preprocessing

2.1 Data Preprocessing

The data that I ended up using was already fairly clean, with no missing values. In order to get the categorical variables ready for analysis, several encoding methods were used. The "Gender" column was coded as numbers by substituting "Male" with 0 and "Female" with 1. Likewise, the "Sleep Disorder" column that contained "None," "Insomnia," and "Sleep Apnea" as its values was encoded as 0, 1, and 2, respectively. The "BP" column, which was initially a string (e.g., "120/80"), was separated into two columns: Systolic_BP and Diastolic_BP, to allow analysis of the effect of blood pressure. One-Hot Encoding was also used on the "Occupation" and "BMI Category" columns to convert them into numeric features without losing their categorical information.

2.2. Exploratory Data Analysis (EDA)

Together, the narratives give a sleep duration and population profile of a study group. The sex ratio is roughly equal. Sleep duration clusters at 6, 7, and 8 hours. Age distribution is in the mid-40s, with an age range of late 20s to late 50s.

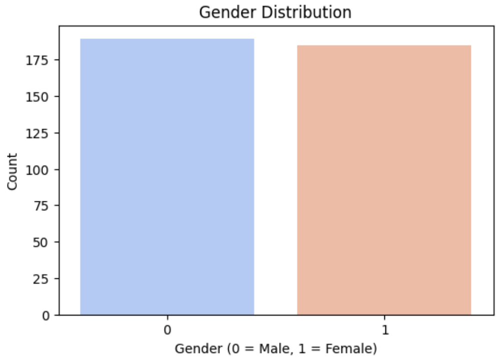


Figure 1: Gender Distribution

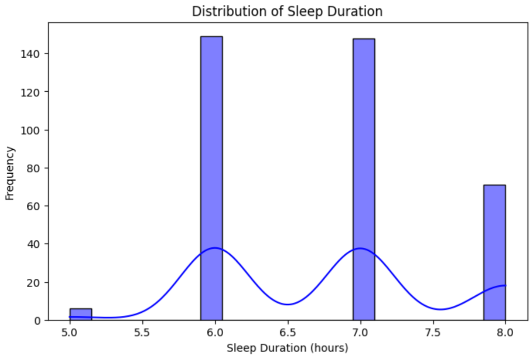


Figure 2: Distribution of Sleep Duration

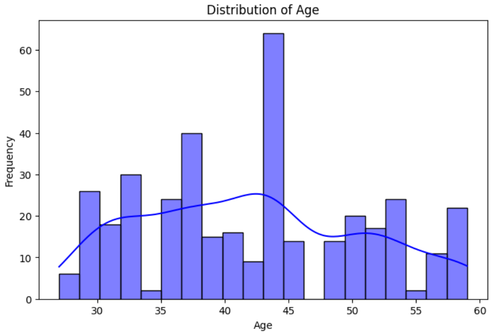


Figure 3: Distribution of Age

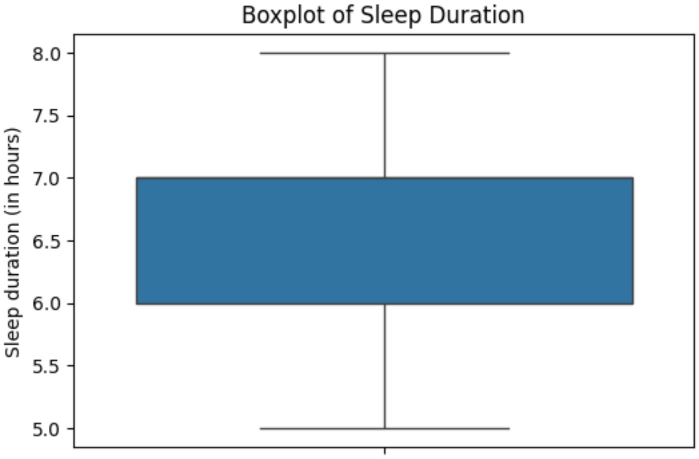
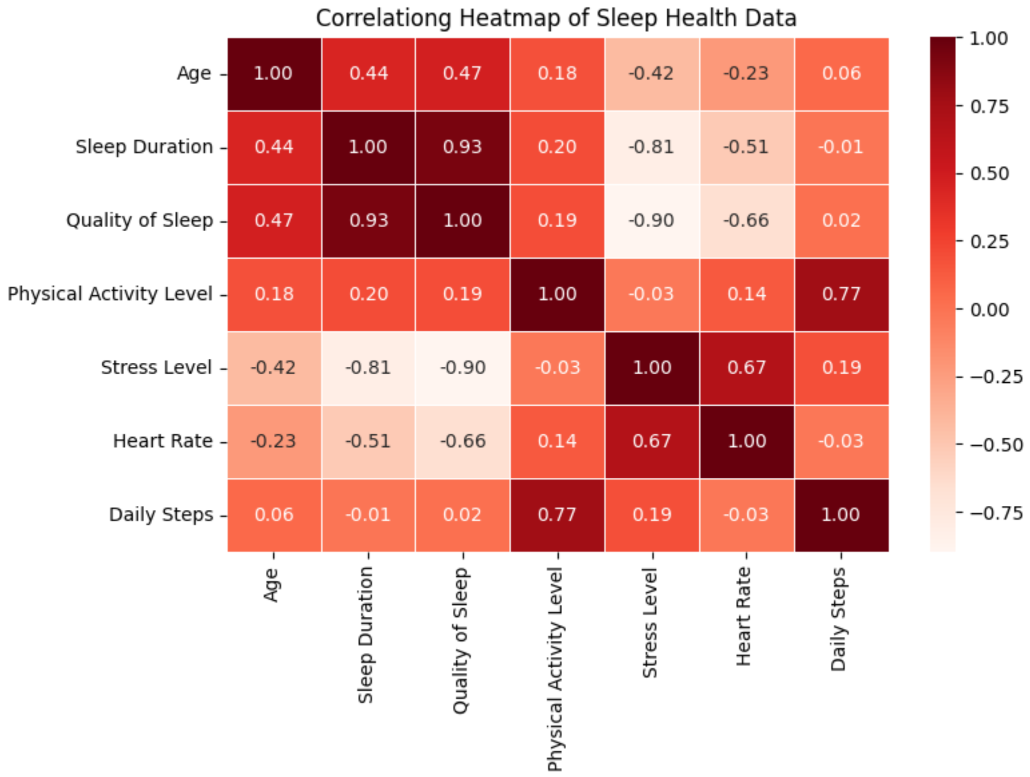


Fig 4: Boxplot of Sleep Duration

This box plot is a statistical graph of the sleep time in hours, displaying important details such as spread, central tendency, and outliers. The box is the IQR with central tendency indicated within the box. The whiskers stretch to cover the min and max within 1.5 times the IQR. The lack of single points beyond indicates no outliers in the data. The plot illustrates that the majority of individuals sleep between roughly 6 to 7 hours, with a range of roughly 5 to 8 hours.

In bivariate analysis, a heatmap is displayed to illustrate sleep health data feature correlations and relationships. Sleep duration and sleep quality are strongly positively correlated (0.93), meaning more sleep duration corresponds to higher sleep quality. Physical activity level is weakly positively correlated with daily steps (0.77), meaning more steps are weakly related to the physical activity level. Conversely, stress intensity is inversely related to sleep duration (-0.81) and sleep quality (-0.90), indicating that increased stress is connected to poor sleep.


Correlationg Heatmap of Sleep Health Data

2.3. Model Building

In this part, two models were used to forecast sleep duration. The data was divided into 80% train and 20% test for evaluation purposes. Linear Regression and Random Forest was used because it could easily detect non-linear interactions and feature interactions.

2.4. Model Evaluation

For model evaluation, I used 3 metrics: MAE, RMSE, and R^2. The Linear Regression model had an MAE of 0.095, RMSE of 0.125, and an R^2 score of 0.974, reflecting good predictive power. Nevertheless, the Random Forest Regressor was significantly better, with an MAE of 0.018, RMSE of 0.072, and an R^2 score of 0.991.

2.5. Hyper-parameter Tuning

For additional model performance enhancement, I used GridSearchCV and RandomizedSearchCV. The best hyperparameters for the Random Forest model were n_estimators = 100, max_depth = 30, min_samples_split = 2, and min_samples_leaf = 1.

2.6. Feature Selection

This was conducted to determine the best predictors of sleep duration. On the basis of feature importance analysis of the Random Forest model, Quality of Sleep, Systolic Blood Pressure, Stress Level, Physical Activity Level, and BMI Category were determined to be the most contributory features. These features were the largest contributors to the prediction of sleep duration, enabling a more parsimonious and interpretable model with no loss of performance.

3. Conclusion

3.1. Key Findings

The model's performance was measured using on MAE, RMSE, and R-squared (R^2). The Random Forest Regressor performed better than the Linear Regression model with 0.991 R^2 and 0.072 RMSE versus 0.974 R^2 and 0.125 RMSE for Linear Regression. This indicates that Random Forest did a better job of capturing more intricate relationships in the data. The most predictive of sleep duration were Quality of Sleep, Systolic Blood Pressure, Stress Level, Physical Activity Level, and BMI Category.

3.2. Final Model

The model used was the Random Forest Regressor since it always ended up being stronger and more precise than Linear Regression. When hyperparameter tuning was carried out, the model received an R^2 of 0.991, which means 99.1% variance in sleep duration. The tuned Random Forest model managed to learn patterns from the data set and was therefore the most appropriate in predicting sleep duration.

3.3. Challenges

Some of the issues experienced in the project involved handling outliers. Removal of outliers would have amounted to an extreme decrease in dataset size, and therefore, would have affected model performance. Encoding the categorical variables was also challenging in a way that numerical transformations did not collapse meaningful relationships in data. Striking a balance between feature selection and model interpretability was another challenge because selecting too many features could lead to unnecessary complication, while removing significant ones would affect accuracy.

4. Discussion

4.1. Model Performance

Random Forest was the best-performing model among those attempted. With the lowest RMSE (0.072) and high R^2 score (0.991), we can conclude the model performed very well in predicting sleep duration. Compared to Linear Regression with an R^2 of 0.974, Random Forest was able to better understand feature interactions and thus Random Forest is a more suitable model to be utilized here.

4.2. Effect of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning was an important factor in enhancing model performance. By adjusting parameters like n_estimators, max_depth, min_samples_split, and min_samples_leaf, the Random Forest model was enhanced and generalized to new data more effectively. Feature selection was able to pinpoint the best predictors of sleep duration so that the model could concentrate on the most relevant factors and eliminate noise. This benefited both accuracy and interpretability.

4.3 Interpretation of Results

The findings are as anticipated and show that Quality of Sleep and Stress Level were the highest predictive of sleep duration. People who had more stress had less sleep duration, whereas those who had better quality sleep had more sleep duration. The fact that Systolic Blood Pressure and Physical Activity Level are important implies that cardiovascular wellness and an active life could also influence sleep duration.

4.4 Limitations

While the model was successful, there are a few weaknesses. The dataset used was comparatively small (374 observations), which can affect the ability of the model to be generalized to a larger population. Also, while the features used were adequate, there could be some other parameters not used that can affect sleep duration, i.e., dietary data or medical conditions.