UNIVERSITY OF
WOLVERHAMPTON

HERALD
COLLEGE
KATHMANDU

Academic Year	Module	Assessment Number	Assessment Type
2024	5CS037-Concept and Technology of AI.	**3-Coursework**	**Report Writing**

Classification Logistic Regression Report

Student Id	: 2408291
Student Name	: Rohit Ghimire
Section	: L5CG2
Module Leader	: Siman Giri
Tutor	: Sunita Parajuli
Submitted on	: 11-02-2025

Contents

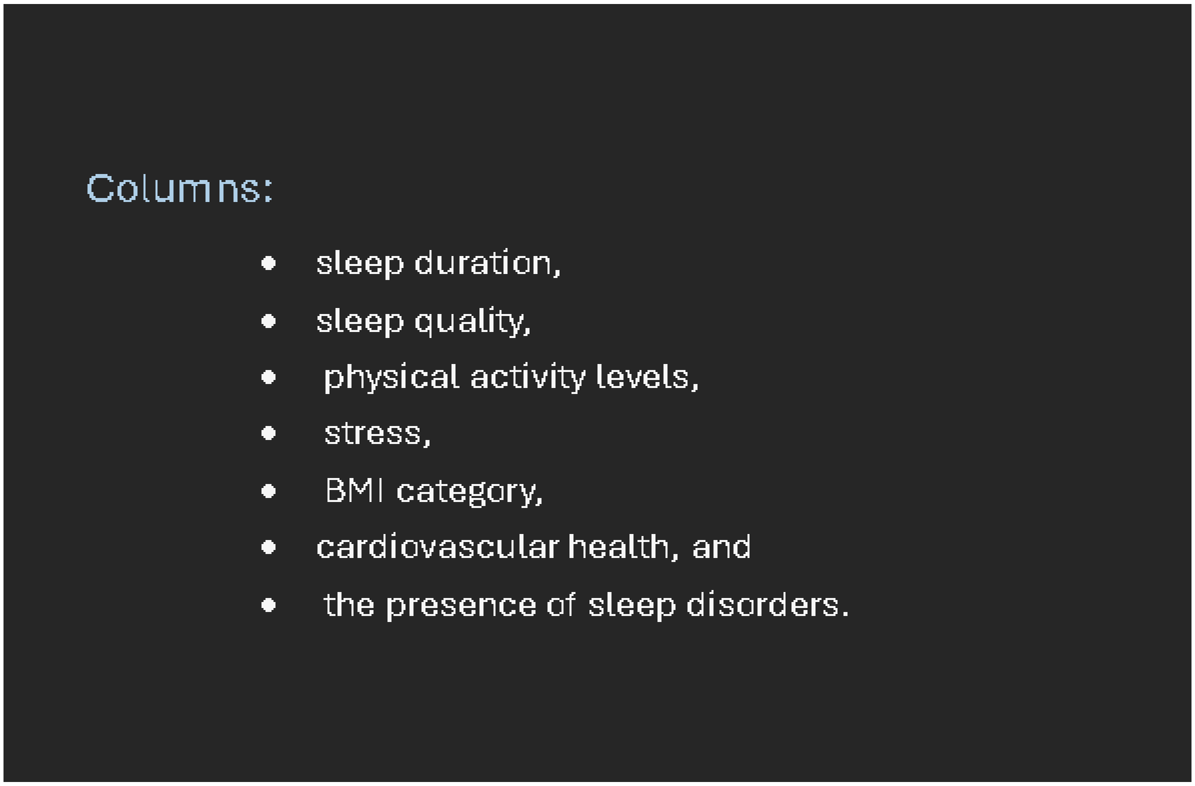
1. Introduction:

It aligns with the UNSDG since the quality sleep is not focused on by the current youths . They sleep late at night which may result in insomnia and other sleep disorders . According to the UNSDG goals number 3 (Good health and wellbeing). I chose sleep health lifestyle dataset to predict the sleep health of a group pf persons according to features. The dataset is extracted from Kaggle website, and the publisher of this dataset was Siamak Tahmasbi · Updated 2 months ago.

Dataset Information:

The Sleep Health and Lifestyle Dataset provides detailed insights into the sleep patterns, daily habits, and lifestyle factors of individuals. This synthetic dataset comprises 400 rows and 13 columns and the data on blood pressure and heart rate, essential for analyzing correlations with sleep and lifestyle.

This dataset follows the UNSDG, promoting well-being of an individual, by providing valuable insights into the factors that affect sleep quality.



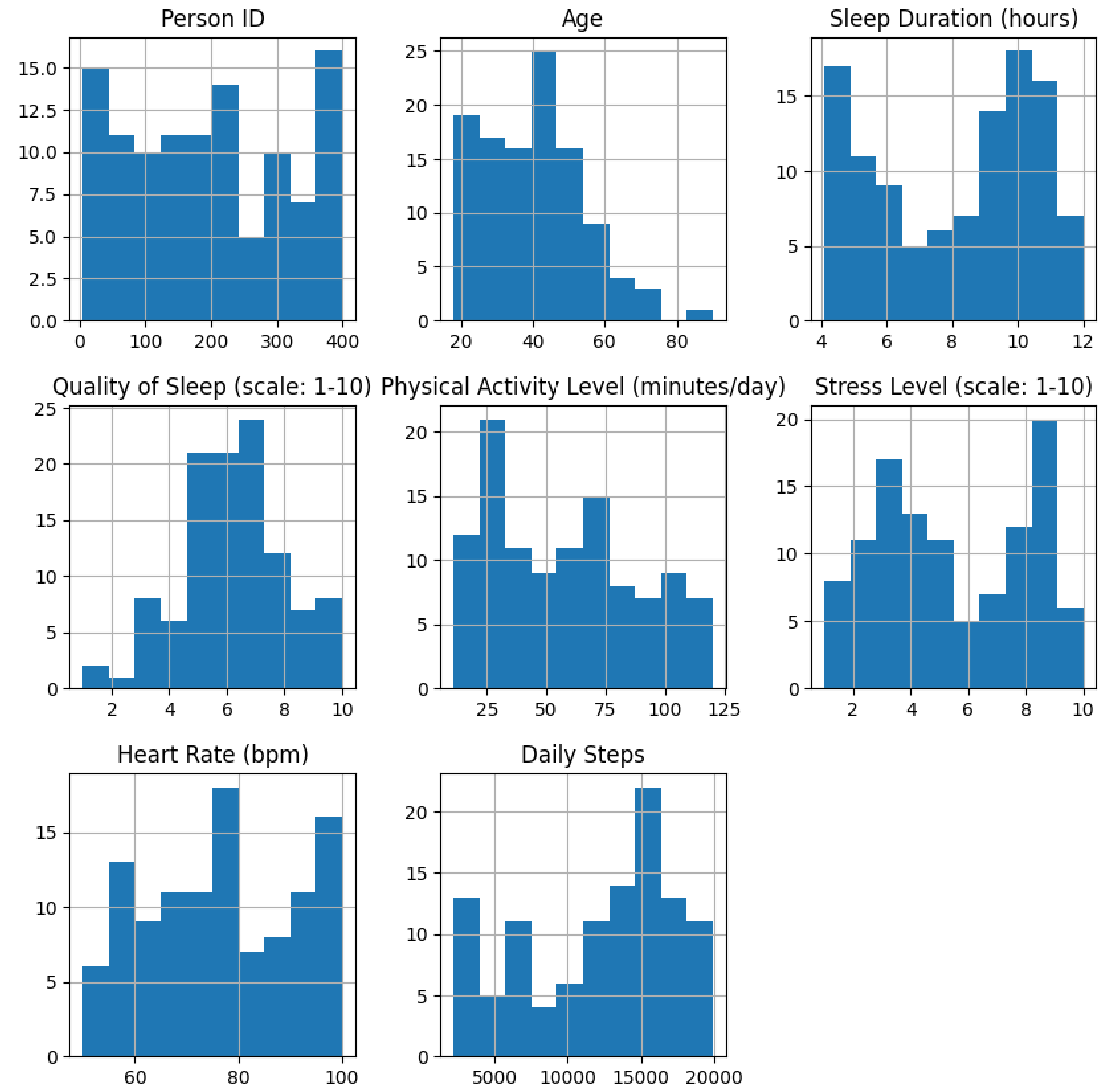
Columns:

- sleep duration,
- sleep quality,
- physical activity levels,
- stress,
- BMI category,
- cardiovascular health, and
- the presence of sleep disorders.

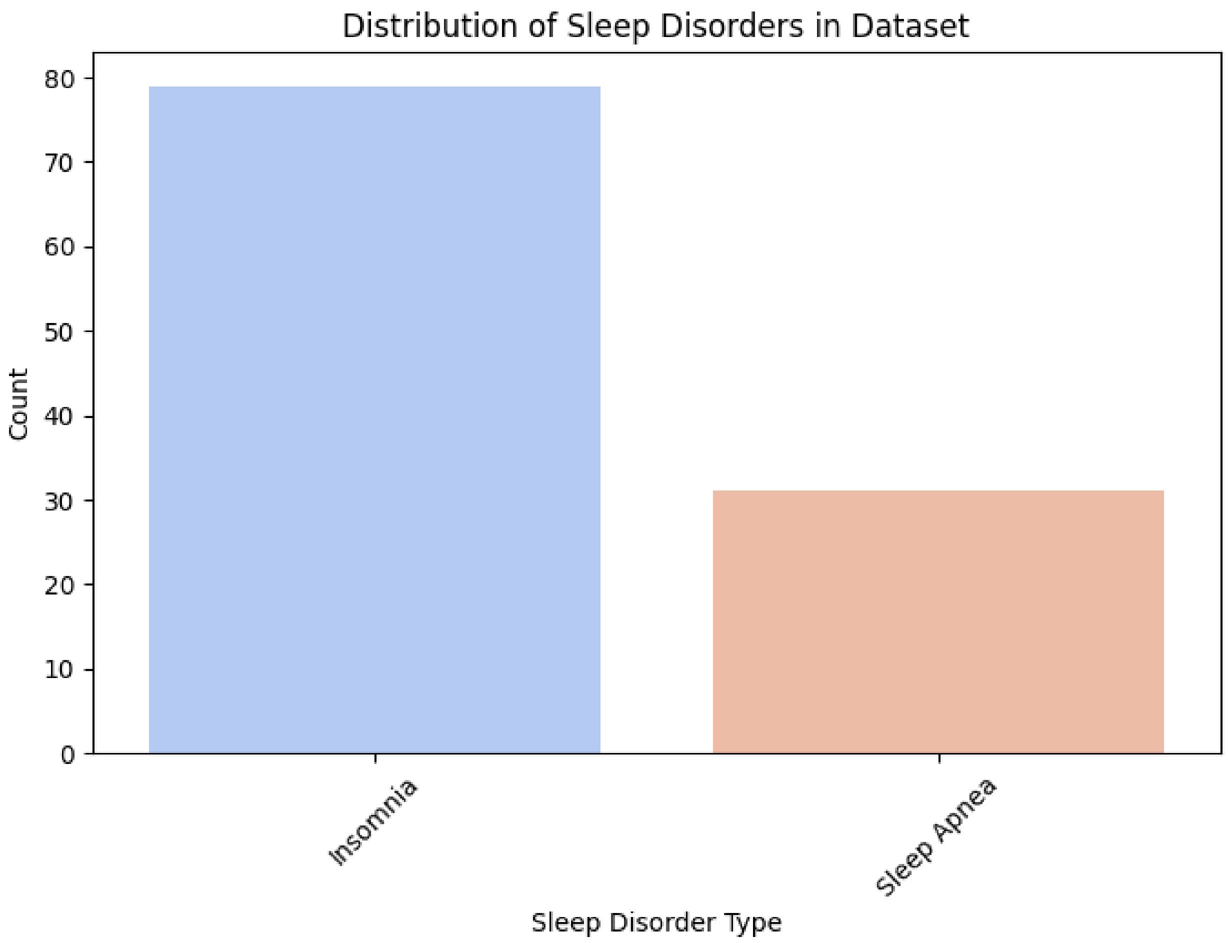
This report is done to predict a sleep disorder using a logistic regression algorithm from a scratch and for primary model scikit and other random-forest classifier are used .This analysis aims to predict the likelihood of a person's sleep quality or a specific sleep-related outcome (e.g., whether a person experiences good or poor sleep) based on their lifestyle habits and sleep patterns.

2. Methodologies

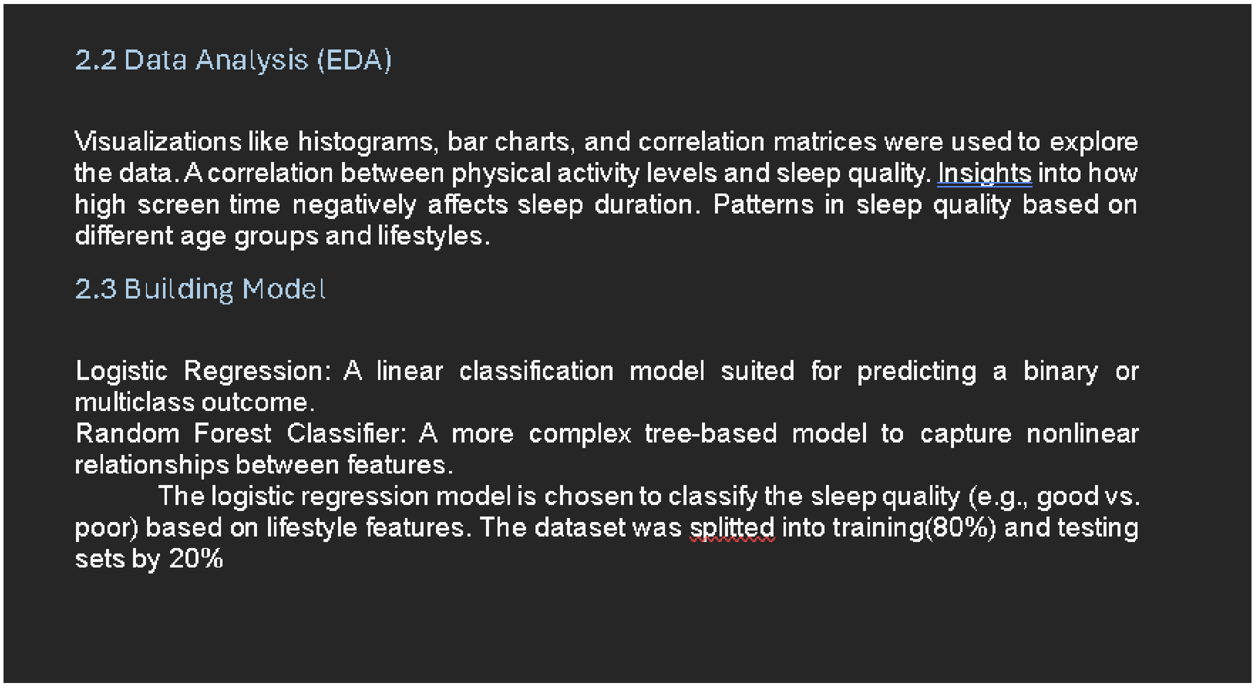
Before the model was developed it gone through handling Missing Values, Missing data points were addressed using imputation or by removing rows with excessive missing values.



Histogram of all features.



We can see that insomnia has the high number of count

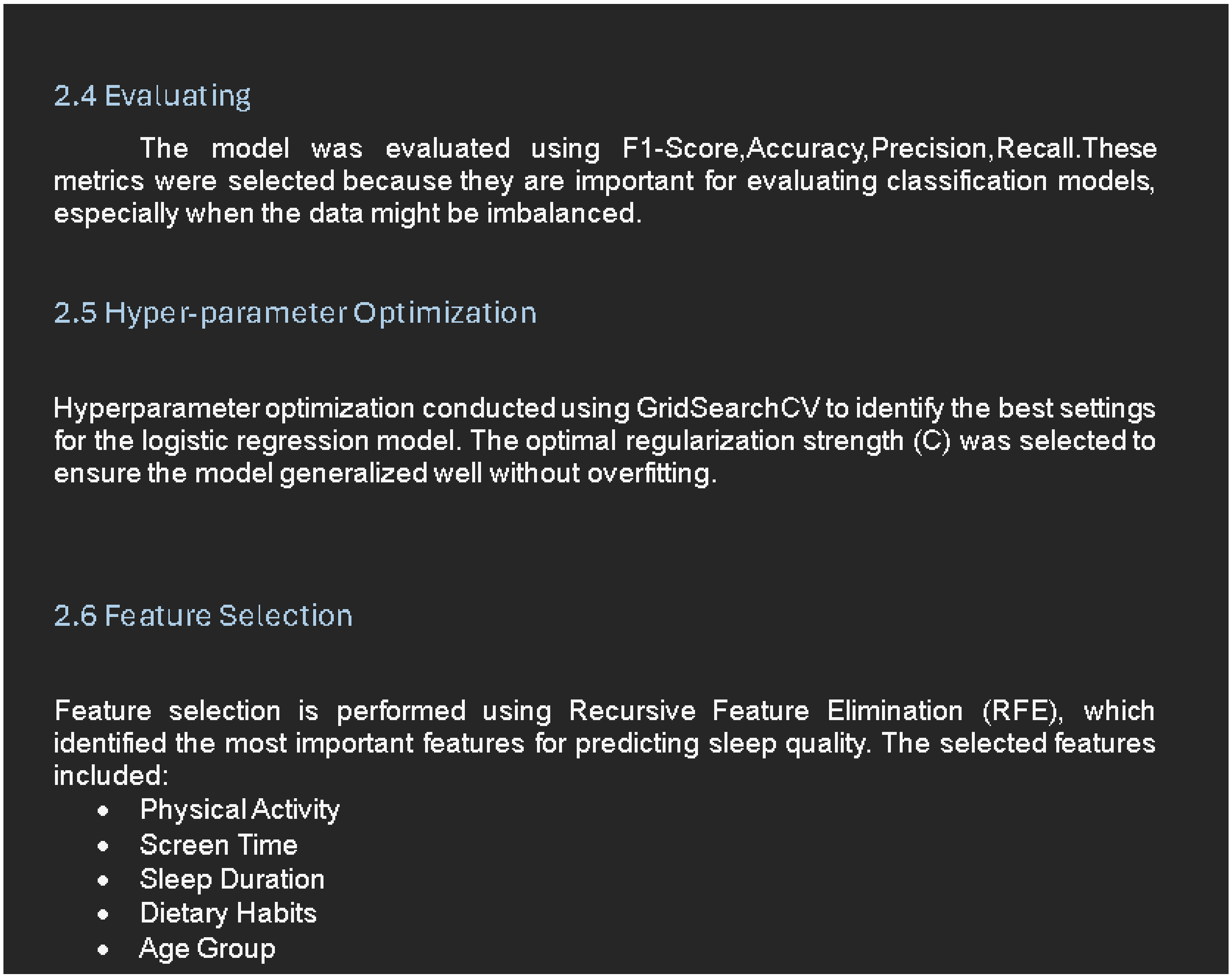


2.2 Data Analysis (EDA)

Visualizations like histograms, bar charts, and correlation matrices were used to explore the data. A correlation between physical activity levels and sleep quality. Insights into how high screen time negatively affects sleep duration. Patterns in sleep quality based on different age groups and lifestyles.

2.3 Building Model

Logistic Regression: A linear classification model suited for predicting a binary or multiclass outcome.
Random Forest Classifier: A more complex tree-based model to capture nonlinear relationships between features.
The logistic regression model is chosen to classify the sleep quality (e.g., good vs. poor) based on lifestyle features. The dataset was splitted into training(80%) and testing sets by 20%



2.4 Evaluating

The model was evaluated using F1-Score,Accuracy,Precision,Recall.These metrics were selected because they are important for evaluating classification models, especially when the data might be imbalanced.

2.5 Hyper-parameter Optimization

Hyperparameter optimization conducted using GridSearchCV to identify the best settings for the logistic regression model. The optimal regularization strength (C) was selected to ensure the model generalized well without overfitting.

2.6 Feature Selection

Feature selection is performed using Recursive Feature Elimination (RFE), which identified the most important features for predicting sleep quality. The selected features included:

- Physical Activity
- Screen Time
- Sleep Duration
- Dietary Habits
- Age Group

3. Conclusion

3.1 Findings

Logistic regression model performed well while predicting sleep quality. The key findings include:

The model achieved an accuracy of 77% on the test dataset.
Features like physical activity and sleep duration had the highest impact on the model's performance.
The model showed high recall for predicting poor sleep quality, indicating it was effective at identifying those with sleep issues.

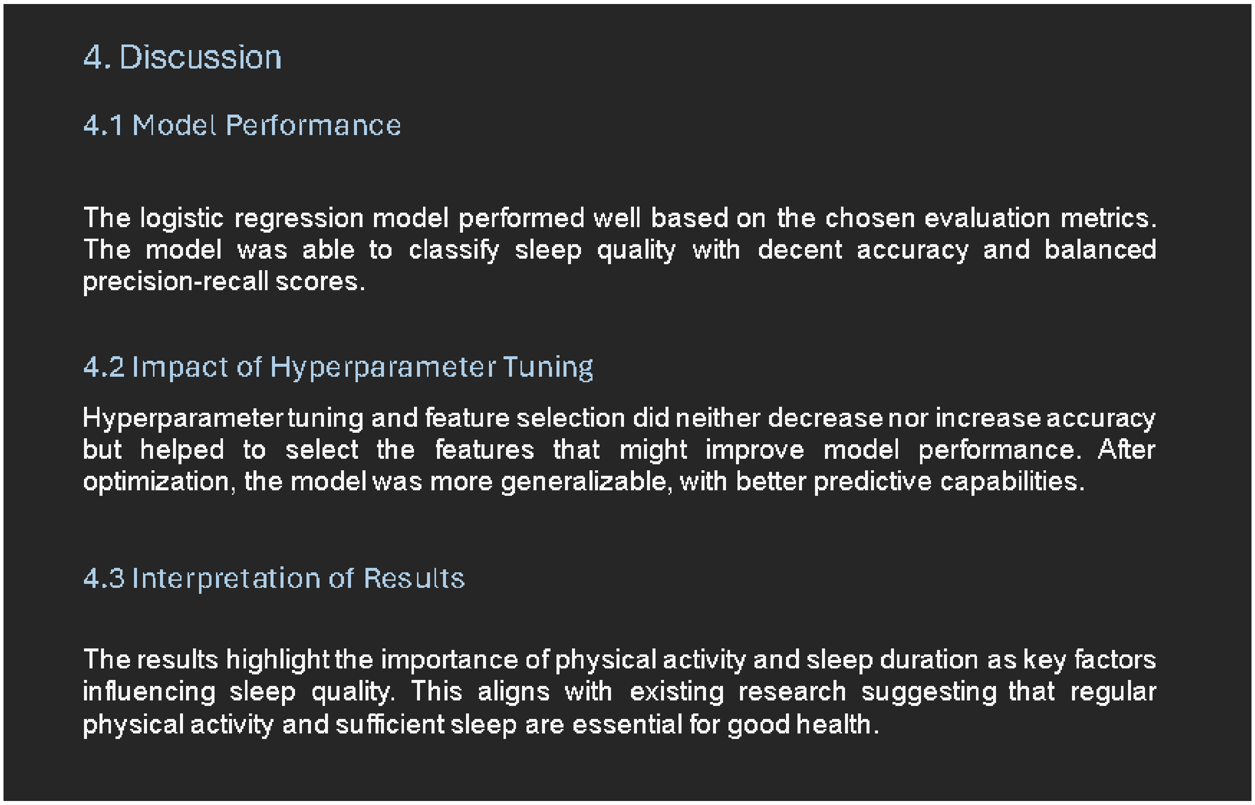
3.2 Final Model

The final model selected was Logistic Regression, which achieved the best results for classifying sleep quality based on the lifestyle data. The model's accuracy was 77%, and the F1-score was 0.67, demonstrating strong performance.

3.3 Challenges

Challenges encountered included:

Data Quality Issues: Missing or inconsistent data points in lifestyle factors such as dietary habits.
Feature Selection: Identifying the most important features was challenging, given the large number of lifestyle-related variables.



4. Discussion

4.1 Model Performance

The logistic regression model performed well based on the chosen evaluation metrics. The model was able to classify sleep quality with decent accuracy and balanced precision-recall scores.

4.2 Impact of Hyperparameter Tuning

Hyperparameter tuning and feature selection did neither decrease nor increase accuracy but helped to select the features that might improve model performance. After optimization, the model was more generalizable, with better predictive capabilities.

4.3 Interpretation of Results

The results highlight the importance of physical activity and sleep duration as key factors influencing sleep quality. This aligns with existing research suggesting that regular physical activity and sufficient sleep are essential for good health.

4.4 Limitations

Some limitations included:

Limited Data: The dataset may not represent the full spectrum of lifestyle habits and sleep patterns across different populations.

Linear Assumptions in Logistic Regression: Logistic regression assumes a linear relationship between features and the target, which may not always be appropriate.

4.5 Future Research

Alternative Models: Trying neural networks or ensemble methods to capture more complex relationships.
Broader Data Collection: Expanding the dataset to include additional lifestyle factors or more diverse participant groups.
Incorporating Time Series Data: Analyzing changes in lifestyle over time and their effect on sleep quality.