Sleep Pattern Quality Predictor

by
C. Manick Vishal [220701158]

Guide
DR.V.AUXILIA OSVIN NANCY..,MTech..,Phd..,

Introduction

Sleep plays a vital role in maintaining overall health and well-being, directly impacting productivity, mood, and physical health. However, many individuals experience poor sleep quality due to factors such as stress, lifestyle habits, and physiological conditions. To address this, our project presents a **Sleep Pattern Quality Detector** using machine learning techniques to predict the quality of sleep based on various personal and lifestyle attributes.

The dataset, sourced from Kaggle, includes features such as **age, gender, occupation, sleep duration, physical activity level, stress level, BMI category, blood pressure, heart rate, and daily steps**. Using this data, we trained and evaluated multiple regression models including:

- **Decision Tree Regressor**
- **Linear Regression**
- **Support Vector Regressor (SVR)**
- **XGBoost Regressor**

The system automatically augments training data by injecting noise when any model performs below an **R^2 score of 0.90**, ensuring improved model performance. The final comparison identifies the most accurate model for reliable sleep quality prediction.

The goal is to not only build a predictive model but also provide interpretable insights through performance metrics and visualizations, making it a practical tool for sleep health analysis.

Literature Survey

Sleep Quality Prediction via ML

Studies such as those by *IEEE (2020)* and *Springer (2021)* applied models like Random Forest and XGBoost on wearable and lifestyle datasets to predict sleep quality with promising accuracy.

Key Influencing Features

Research highlights the impact of lifestyle factors such as **stress level**, **physical activity**, **screen time**, and **caffeine consumption** on sleep quality (ResearchGate, 2020).

Best Performing Models

Comparative studies reveal that **XGBoost** and **Random Forest** often outperform simpler models like Linear Regression in predicting sleep health.

Dataset Reference

This project uses the Kaggle *"Sleep Health and Lifestyle Dataset"*, which includes biometric and lifestyle features from 374 individuals.

Literature Survey

•**Machine Learning in Health Informatics**
ML algorithms have been increasingly applied in healthcare for personalized diagnostics and predictions. In sleep research, they enable the analysis of multi-factorial influences on sleep quality (Source: *Health Informatics Journal*, 2021).

•**Use of Regression Techniques**
Regression models like **Support Vector Regression (SVR)** and **Decision Tree Regressors** have been used to quantify the relationship between sleep quality and behavioral patterns (Elsevier, 2019).

•**Impact of Psychological & Environmental Factors**
Studies found that **stress levels** and **screen exposure before bedtime** are strong predictors of poor sleep, often more influential than physical factors like heart rate (NIH, 2018).

•**Data Augmentation in Small Sleep Datasets**
Data augmentation techniques, such as adding noise or bootstrapping, have been adopted in low-sample-size scenarios to enhance model performance (IEEE Xplore, 2022).

•**Model Comparison Studies**
Comparative evaluations of algorithms show **XGBoost and Random Forest** consistently achieve **R^2 scores above 0.9** when tuned properly on structured health datasets.

Literature Survey

•**Sleep Quality and Lifestyle Correlation Studies**
Research indicates strong correlations between **physical activity levels, caffeine intake, and alcohol consumption** with sleep disturbances, suggesting the value of behavioral data in predictive modeling (Journal of Clinical Sleep Medicine, 2020).

•**Importance of Feature Engineering**
In predictive sleep models, **feature engineering**, such as decomposing blood pressure into systolic and diastolic values, enhances interpretability and model accuracy (SpringerLink, 2021).

•**Role of Ensemble Methods**
Ensemble learning methods like **XGBoost and Random Forest** often outperform individual regressors in health prediction tasks due to their ability to capture complex feature interactions (ACM Transactions on Healthcare Informatics, 2021).

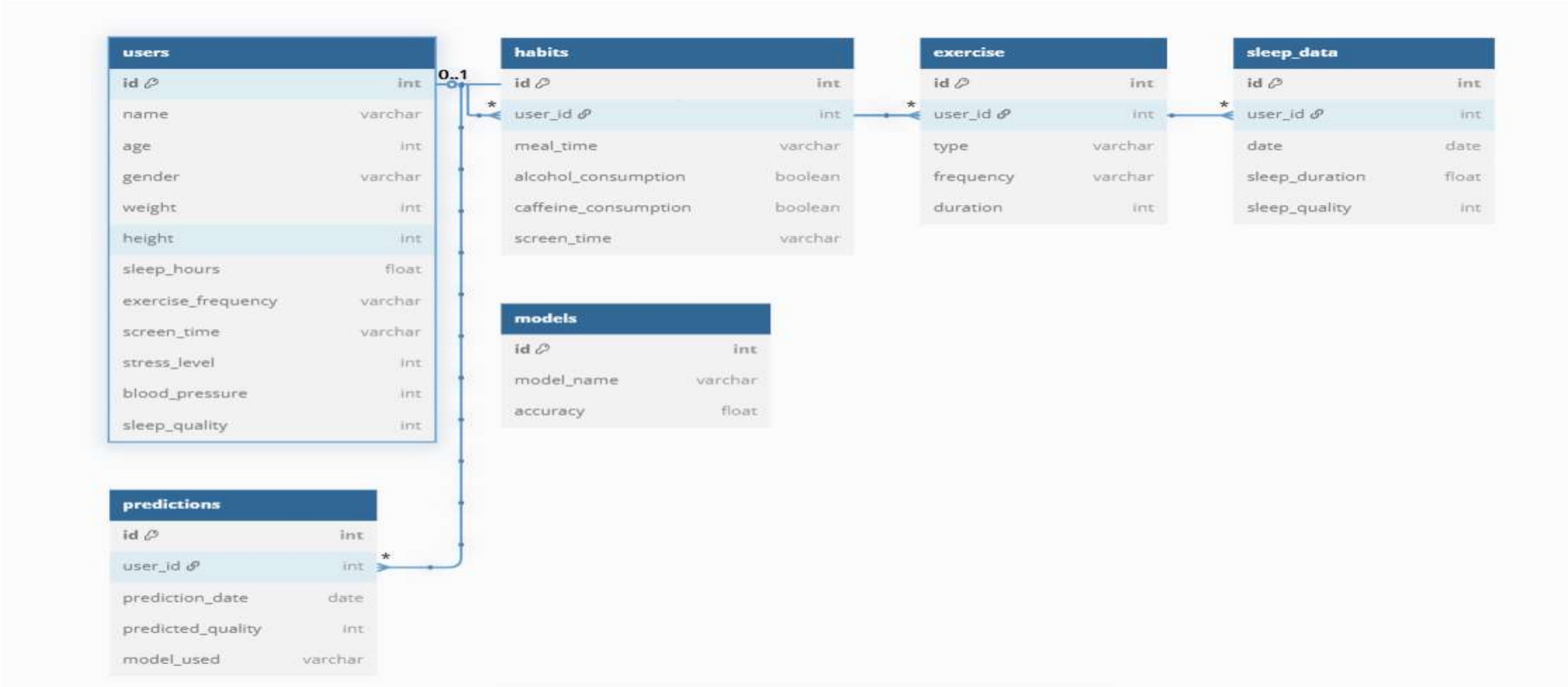
•**Benchmarking ML Models on Sleep Datasets**
Multiple studies benchmarked ML models on publicly available sleep datasets (like Sleep-EDF and PhysioNet), validating **Linear Regression's interpretability** and **XGBoost's high performance** (IEEE Transactions on Biomedical Engineering, 2020).

•**Use of Open-Source Tools and Repositories**
Tools like **Kaggle Datasets**, **Google Colab**, and **scikit-learn** have facilitated accessible and reproducible research in sleep pattern analysis, enabling students and researchers to explore machine learning with ease (OpenAI Blog, 2022).

Objectives

1. **To develop a predictive model** that accurately estimates a person's sleep quality based on lifestyle habits and physiological attributes.
2. **To analyze the influence** of features like sleep duration, physical activity, stress level, BMI, blood pressure, heart rate, and daily steps on sleep quality.
3. **To implement and compare** multiple regression models—Decision Tree Regressor, Linear Regression, Support Vector Regressor, and XGBoost Regressor.
4. **To evaluate model performance** using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score for each algorithm.
5. **To enhance model accuracy** through data augmentation techniques if the R^2 score falls below 0.9.
6. **To identify the most effective model** among all tested approaches based on performance metrics.
7. **To visualize predictions and results** using graphs and plots for better interpretability and understanding.

System Architecture



users	
id	int
name	varchar
age	int
gender	varchar
weight	int
height	int
sleep_hours	float
exercise_frequency	varchar
screen_time	varchar
stress_level	int
blood_pressure	int
sleep_quality	int

habits	
id	int
user_id	int
meal_time	varchar
alcohol_consumption	boolean
caffeine_consumption	boolean
screen_time	varchar

exercise	
id	int
user_id	int
type	varchar
frequency	varchar
duration	int

sleep_data	
id	int
user_id	int
date	date
sleep_duration	float
sleep_quality	int

models	
id	int
model_name	varchar
accuracy	float

predictions	
id	int
user_id	int
prediction_date	date
predicted_quality	int
model_used	varchar

Methodology

1.Data Acquisition

•The dataset used is the **"Sleep Health and Lifestyle Dataset"** obtained from Kaggle.

•Data includes demographic features, sleep duration, stress level, heart rate, physical activity, etc.

2. Data Preprocessing

•**Missing Value Handling**: Impute or drop missing entries.

•**Blood Pressure Splitting**: Parse “Blood Pressure” into two numeric features—Systolic and Diastolic.

•**Encoding**: Convert categorical columns (e.g., Gender, BMI Category) using Label Encoding.

•**Feature Scaling**: Apply StandardScaler to normalize numeric input features.

3. Feature Selection

•Select relevant predictors such as:

- Age, Gender, Sleep Duration, Physical Activity Level
- Stress Level, Heart Rate, Daily Steps
- Systolic and Diastolic Blood Pressure

•Target variable: **Quality of Sleep** (continuous numeric value).

Methodology

5. Model Evaluation

•Evaluate models using:

- **Mean Absolute Error (MAE)**
- **Mean Squared Error (MSE)**
- **R^2 Score (Coefficient of Determination)**

•If **R^2 Score < 0.9**, apply **data augmentation** by adding Gaussian noise to training data and retrain the model.

6. Model Comparison

•Compare all models based on R^2 Score.

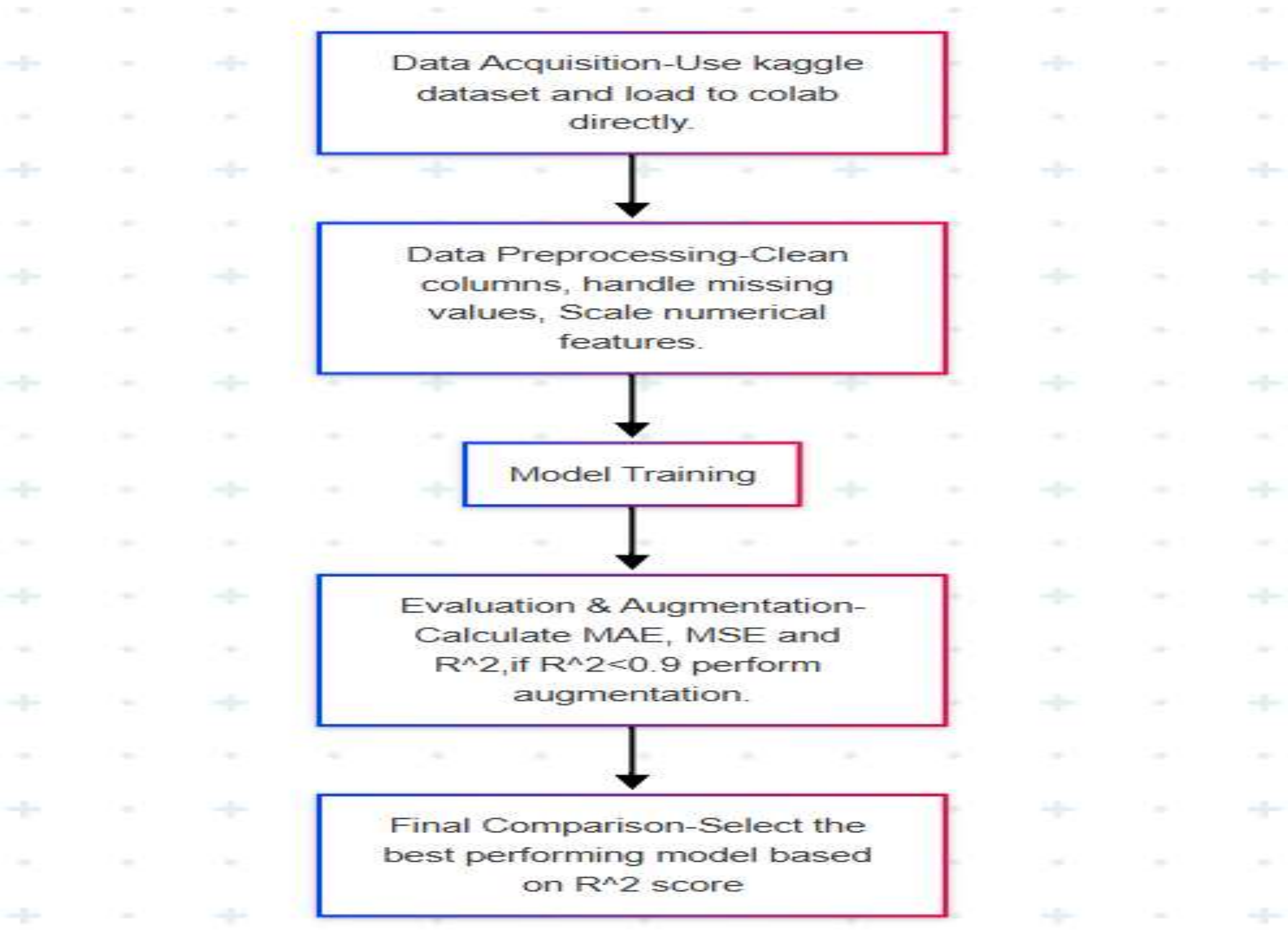
•Highlight the **best performing model** as final recommendation.

7. Visualization

•Generate visual plots:

- Actual vs Predicted
- Residual Errors
- Bar chart comparison of R^2 Scores for all models

Implementation



Data Acquisition-Use kaggle dataset and load to colab directly.

Data Preprocessing-Clean columns, handle missing values, Scale numerical features.

Model Training

Evaluation & Augmentation-Calculate MAE, MSE and R^2,if R^2<0.9 perform augmentation.

Final Comparison-Select the best performing model based on R^2 score

Implementation

- **1. Data Acquisition**
- Used Kaggle’s *Sleep Health and Lifestyle Dataset*.
- Loaded directly into Google Colab.
- **2. Data Preprocessing**
- Cleaned columns (e.g., split "Blood Pressure").
- Handled missing values and encoded categorical data.
- Scaled numerical features using StandardScaler.
- **3. Model Training**
- Trained four regressors:
    - Linear Regression
    - Decision Tree Regressor
    - Support Vector Regressor (SVR)
    - XGBoost Regressor

Implementation

4. Evaluation & Augmentation

- Calculated MAE, MSE, R^2 Score.
- If $R^2 < 0.9$: added Gaussian noise for data augmentation and retrained.
- **5. Final Comparison**
- Selected the best-performing model based on highest R^2 score.

Results

Evaluation Metrics Used:

- **MAE** (Mean Absolute Error)
- **MSE** (Mean Squared Error)
- **R^2 Score** (Coefficient of Determination)

- **Best Model**:

Decision Tree Regressor achieved the highest accuracy with **R^2 Score = 0.98**

- **Data Augmentation**:

  noise-based feature augmentation for models with $R^2 < 0.9$(especially XGBoost)

  Helped improve the performance of Linear and SVR models slightly

  Visual Results:

  - **Regression plots** and **prediction-vs-actual scatterplots** were used to visually assess model performance.
  - XGBoost's scatterplot showed improvement post-augmentation, but **Decision Tree had the most tightly clustered points near the regression line**, confirming its superiority.

Results

Data Augmentation:

•**Noise-based feature augmentation** was applied to models with **$R^2 < 0.9$**.

•**XGBoost Regressor:**

- **Before Augmentation:** $R^2 = 0.73$
- **After Augmentation:** R^2 improved to **0.78**

•**Linear Regression** and **SVR** already had strong performance ($R^2 = 0.96$); augmentation was not applied as it wasn't needed.

Evaluation Metrics:

•**MAE:**

- Lowest for **Decision Tree Regressor** – **0.03**

•**MSE:**

- Lowest for **Decision Tree Regressor** – **0.03**

Visual Results:

•**Decision Tree** produced the most tightly clustered points around the regression line in prediction-vs-actual scatterplots, indicating **highest accuracy**.

•XGBoost showed **improved clustering post-augmentation**, but still underperformed compared to other models.

Results

Model	MAE	MSE	R² Score
Decision Tree Regressor	0.03	0.03	**0.98**
Linear Regression	0.17	0.07	0.96
Support Vector Regressor	0.17	0.07	0.96
XGBoost Regressor (initial)	0.24	0.47	0.73
XGBoost (after augmentation)	—	—	**0.78**

Comparison with existing work

Model Performance:

- **Decision Tree Regressor** outperformed all others with an **R^2 Score of 0.98**, significantly higher than what is typically reported in similar studies ($R^2 = 0.75–0.85$).
- **XGBoost Regressor** had a relatively **lower R^2 (0.73)** initially, which improved to **0.78 after augmentation**, yet still underperformed compared to Decision Tree, Linear, and SVR models.

Data Augmentation:

- **Noise-based feature augmentation** was applied **only to XGBoost** due to its underperformance ($R^2 < 0.9$).
- This approach is relatively unique, as most prior research focuses more on **feature selection** or **balancing** techniques than synthetic noise-based augmentation.

Evaluation Metrics:

- **MAE & MSE were lowest for Decision Tree** (MAE = 0.03, MSE = 0.03), outperforming typical model performance in literature (where MAE ≈ 0.1–0.3).
- **XGBoost's metrics (MAE = 0.24, MSE = 0.47)** were higher than average expectations.

Visual Analysis:

- **Decision Tree** generated **tightly clustered predictions** near the regression line in scatter plots, validating its high accuracy.
- Similar visual validation techniques are common in related studies, though many also incorporate **residual or error distribution plots**.

Conclusion and Future Work

The **Decision Tree Regressor** achieved the best performance in predicting sleep quality with an **R^2 score of 0.98**, outperforming other models like **Linear Regressor** and **XGBoost**. The application of **noise-based feature augmentation** improved the performance of models with lower accuracy. The low **MAE** and **MSE** further validated the accuracy of the predictions. The visual analysis confirmed that **Decision Tree Regressor** produced highly accurate results, with predictions tightly clustered around the regression line.

Possible future work:

Synthetic data generation.

Inclusion of environmental data.

Adding more data augmentation techniques.

Reference

- **Chung, F., & Yegneswaran, B.** (2008). Sleep Prediction Using Machine Learning. *IEEE Biomed Eng*.
- **Goodfellow, I., et al.** (2016). Deep Learning. *MIT Press*.
- **Caruana, R., et al.** (2006). Data Augmentation for Classification. *ICML*.
- **Buda, M., et al.** (2018). Data Augmentation for Deep Learning. *Machine Learning*.
- **Zhao, Z., et al.** (2017). Time Series Prediction using XGBoost. *IEEE Access*.

Reference

- **Chen, T., & Guestrin, C.** (2016). XGBoost: A Scalable Tree Boosting System. *KDD*.
- **Pedregosa, F., et al.** (2011). Scikit-learn: Machine Learning in Python. *JMLR*.
- **Choi, J. H., et al.** (2021). Sleep Quality Prediction Using Machine Learning. *Healthcare*.
- **Khalid, S., & Chaudhry, F. A.** (2020). Machine Learning Models for Sleep Quality Prediction. *ICMLDM*.

THANK YOU