Department of Computer Science and Engineering

National Institute of Technology, Jamshedpur

Branch: MCA (4thsem.)
Course: Artificial Intelligence
Course Code: CS3402
Submission Date: 16/05/25

Project: *Sleep Quality Prediction Using Smartphone*

Submitted to:
Dr. Gopa Bhaumik

Submitted by:
Shubham Singh
2023PGCSCA026

Objective

The goal of this project is to predict a person's sleep quality (1 to 5) based on their smartphone usage patterns, such as:
- Screen time before sleep
- Blue light exposure
- Sleep duration
- Bedtime

We used Logistic Regression as our machine learning model, along with Cross Validation to improve accuracy and generalization.

Problem Statement

Smartphone usage, especially at night, impacts the quality of sleep. This project aims to analyze how different phone habits affect sleep and build a model to predict sleep quality using a few key features.

Dataset

We used a dummy dataset created for this project with the following features:
- screen_time_min: Total screen time before sleep (in minutes)
- blue_light_exposure_min: Duration exposed to blue light (in minutes)
- sleep_duration_hr: Total sleep duration (in hours)
- bedtime_24h: Bedtime in 24-hour format (e.g., 22 = 10 PM)
- sleep_quality_1_5: Target column (Sleep quality on scale 1–5)

Selected Features

We selected the following 4 features as input for our model:

- screen_time_min
- blue_light_exposure_min
- sleep_duration_hr
- bedtime_24h

Tools and Libraries Used

- Python
- Pandas
- Scikit-learn (sklearn)
- Logistic Regression (from sklearn)
- Cross Validation (cross_val_score)
- StandardScaler (for feature scaling)

Methodology

Step 1: Data Preparation
- Load the dataset.
- Check for missing values and handle them if needed.
- Select relevant features for model training.

Step 2: Feature Scaling
- Standardized the data using StandardScaler to ensure all features contribute equally.

Step 3: Model Selection
- First tried using Random Forest, but it gave only 49.9% accuracy.
- Switched to Logistic Regression, which gave better accuracy (62.5%).

Step 4: Cross Validation
- Used 5-Fold Cross Validation to ensure the model performs well on unseen data and to avoid overfitting.

Final Model – Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import cross_val_score

model = LogisticRegression(max_iter=1000, random_state=42)
scores = cross_val_score(model, X_scaled, y, cv=5)

print("Cross-validated accuracy scores:", scores)
print("Average accuracy:", scores.mean())
```

Result

Model	Accuracy
Random Forest	49.9%
Logistic Regression	62.5% ✓

The Logistic Regression model worked better because it’s simple and fits well with our small dataset and limited features.

Conclusion

- Logistic Regression is a good model for this type of small and clean dataset.
- Sleep quality is affected by screen time, blue light exposure, and bedtime.
- With more data and better features (like caffeine intake or stress levels), this model can be further improved.

Future Improvements

- Collect real-world data using surveys or tracking apps.

- Add more features: physical activity, water intake, caffeine, etc.
- Try other models like Decision Trees, XGBoost, or Neural Networks.