REPORT

**Dataset Description**

The dataset contains information about students' sleep patterns and associated lifestyle factors.

* It includes 500 entries with 14 columns.
* Each row represents an individual student, providing details about their demographics, habits, and sleep-related data.
* The columns include: Student\_ID, Age, Gender, University\_Year, Sleep\_Duration, Study\_Hours, Screen\_Time, Caffeine\_Intake, Physical\_Activity, Sleep\_Quality, Weekday\_Sleep\_Start, Weekend\_Sleep\_Start, Weekday\_Sleep\_End, and Weekend\_Sleep\_End. 1

**Models and Performance**

1. **K-Nearest Neighbors (KNN)**
   * The model's performance metrics indicate poor performance: RMSE (3.29) and R² Score (-0.13).
   * A negative R² score means the model performs worse than a horizontal line (mean value), showing it fails to explain the variance in the data.
   * The model likely suffers from underfitting and requires better features, tuning, or a different algorithm.
2. **Decision Tree**
   * The decision tree predicts whether a person will get enough sleep based on their sleep status and sleep duration.
   * If the person's sleep status is "yes", the tree predicts they will get enough sleep.
   * If the sleep status is "no", the tree checks the sleep duration.
   * If the sleep duration is less than or equal to 5.6 hours, the tree predicts they will not get enough sleep.
   * If the sleep duration exceeds 5.6 hours, the tree predicts they will get enough sleep.
   * The tree's accuracy is shown by the Gini values, with lower values indicating higher confidence in the prediction.
3. **Simple Linear Regression**
   * The linear regression model was used to predict sleep duration based on study hours.
   * The intercept of the model is 6.5255, and the slope is -0.0104.
   * This suggests a weak negative relationship between study hours and sleep duration.
   * The model's predictions tend to underestimate or overestimate the actual values.
4. **K-Means Clustering**
   * K-Means clustering was implemented to identify clusters in the dataset.
   * The Elbow Method and Silhouette Scores were used to determine the optimal number of clusters.
   * The analysis suggests that two clusters provide the best separation in the data.