ABSTRACT

Sleep quality plays a pivotal role in determining academic success, mental health, and overall well-being, particularly among university students. This study aims to predict sleep quality by analyzing various lifestyle and behavioral factors, including sleep duration, study hours, screen time, caffeine intake, physical activity, and sleep timing patterns. To ensure the dataset was ready for model training, comprehensive data preprocessing was carried out, including handling missing values, encoding categorical variables, and scaling numerical features. The ultimate goal is to develop a reliable predictive model that offers actionable insights and personalized recommendations to improve students' sleep hygiene.

Several machine learning algorithms were explored to build the prediction models, including Random Forest, XGBoost, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Each model was evaluated using key performance metrics such as accuracy, precision, recall, and F1-score. The results demonstrated that XGBoost and Random Forest outperformed the other models, achieving 99% accuracy and high precision and recall scores, making them the most suitable models for sleep quality prediction. KNN and SVM also performed well, with an accuracy of 96%, while Logistic Regression had a comparatively lower accuracy of 83%.

The study found that sleep duration and screen time were the most influential factors in determining sleep quality, followed by physical activity and study hours. By analyzing feature importance, it was evident that maintaining optimal screen time and increasing physical activity positively impact sleep patterns, highlighting the potential for targeted lifestyle interventions. The models achieved high performance due to the quality and richness of the dataset. Feature correlation analysis further validated the importance of these factors, emphasizing the need to focus on reducing excessive screen time and maintaining a balanced study-sleep routine.

Future work will focus on integrating real-time data from wearable devices such as Fitbit and Apple Health to further enhance the predictive accuracy of the models. Additionally, expanding the dataset to include a more diverse demographic can improve the model's generalizability and applicability across different populations. The deployment of the model as a mobile application can empower students to monitor their sleep patterns, receive real-time feedback, and adopt healthier sleep habits. This project highlights the transformative potential of machine learning in analyzing and predicting sleep quality, providing valuable insights that can help university students make informed decisions about their sleep and lifestyle.

CHAPTER I

INTRODUCTION

Sleep quality plays a critical role in maintaining the overall health, academic performance, and emotional well-being of university students. Adequate and consistent sleep is essential for memory consolidation, cognitive processing, and emotional regulation, which are necessary for academic success and personal growth. However, modern-day challenges, including academic pressure, increased screen time, irregular schedules, and unhealthy lifestyle habits, have led to a significant decline in sleep quality among university students. Poor sleep patterns can result in fatigue, lack of focus, irritability, and long-term health issues such as obesity, anxiety, and cardiovascular problems. Understanding and addressing these sleep-related challenges can lead to improved student outcomes and healthier lifestyles.

This project focuses on predicting sleep quality among university students by analyzing key lifestyle and behavioral factors that influence sleep patterns. The study leverages a comprehensive dataset that includes attributes such as Student_ID, Age, Gender, University Year, Sleep Duration, Study Hours, Screen Time, Caffeine Intake, Physical Activity, Sleep Quality, Weekday and Weekend Sleep Start and End Times. By analyzing these attributes, the project aims to identify patterns, correlations, and key contributors that affect the sleep quality of students.

Importance of Sleep Quality Among University Students

University students often face demanding academic schedules, part-time jobs, extracurricular activities, and social commitments, which contribute to inconsistent sleep patterns. Sleep deprivation and poor sleep quality have been linked to a decline in academic performance, increased stress levels, and the risk of developing long-term health complications. Inconsistent sleep patterns can disrupt the body's circadian rhythm, leading to reduced alertness, impaired decision-making, and heightened emotional reactivity. Chronic sleep deprivation can also increase susceptibility to mental health conditions such as anxiety, depression, and emotional instability.

Students who maintain regular sleep schedules tend to perform better academically and experience improved emotional resilience. However, those who engage in irregular sleep routines, excessive screen time, and unhealthy lifestyle habits often encounter diminished academic outcomes and compromised well-being. By identifying these factors and addressing

their impact on sleep, students can make informed decisions to improve their sleep quality and overall health.

Challenges in sleep quality prediction

Predicting sleep quality among university students presents several challenges due to the complexity and variability of factors influencing sleep patterns. One major challenge is the inconsistency and inaccuracy of self-reported data, which often includes recall bias and incomplete responses, affecting the overall data quality. Additionally, individual differences in sleep behavior make it difficult to establish uniform patterns, as factors such as stress, mental health, and lifestyle choices vary widely across students. The variability between weekday and weekend sleep patterns further complicates modeling, making it challenging to capture consistent trends. Defining and labeling sleep quality into categories such as Poor, Moderate, Good, and Excellent introduces classification complexity, as it oversimplifies a multidimensional phenomenon. Moreover, dynamic lifestyle changes, exam pressures, and seasonal variations affect sleep habits unpredictably, which makes it difficult for models to account for these variations.

Another challenge lies in handling multiple interacting factors such as study hours, screen time, caffeine intake, and physical activity, which may have a compounded effect on sleep quality. Model performance and generalization also present issues, as overfitting and underfitting can occur, and selecting the most relevant features for prediction requires careful consideration. Integrating real-world data from wearable devices like Fitbit or Apple Watch can enhance accuracy but introduces complexities in data synchronization and merging subjective and objective data. Ethical concerns surrounding data privacy, security, and informed consent must also be addressed, especially when dealing with sensitive information related to student behavior. Developing personalized recommendations based on sleep quality predictions is another hurdle, as individual preferences and responses to lifestyle changes vary. Furthermore, temporal analysis and capturing long-term patterns require robust models capable of identifying delayed behavioral impacts on sleep, while complex models such as Random Forest and XGBoost often lack interpretability, necessitating the use of Explainable AI (XAI) techniques like SHAP or LIME. Addressing these challenges is essential for improving the accuracy, relevance, and trustworthiness

Sleep Health Labels

In this project, Sleep Health Labels are categorized to classify and predict sleep patterns among university students based on sleep duration. The labels include Poor Sleep for students sleeping less than 6 hours, which is linked to fatigue, reduced concentration, and health risks. Healthy Sleep applies to those sleeping between 6 to 8 hours, indicating an optimal duration that supports cognitive and emotional well-being. Unhealthy Sleep is assigned to students sleeping more than 8 hours, which may suggest irregular sleep patterns or underlying health concerns. These categories allow for accurate classification and personalized recommendations to improve sleep quality.

Approaches to improve Sleep prediction

The approach taken in this project follows a systematic, data-driven methodology to predict sleep quality using machine learning techniques. It begins with data collection and preprocessing, where a dataset containing attributes such as Student_ID, Age, Gender, University Year, Sleep Duration, Study Hours, Screen Time, Caffeine Intake, Physical Activity, Sleep Quality, and Weekday/Weekend Sleep Start and End Times is cleaned, transformed, and prepared for analysis. Handling of missing values, encoding categorical variables, and normalizing numeric features ensure that the data is ready for model training. Feature engineering and labeling play a critical role in extracting meaningful patterns and trends. Sleep labels are assigned to categorize sleep quality, and relevant features such as study hours, screen time, and physical activity are analyzed to identify their impact on sleep health. Following this, model selection and training involves training multiple machine learning models, including Random Forest, XGBoost, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) to determine the best-performing algorithm. These models are fine-tuned using hyperparameter optimization and evaluated using cross-validation to ensure reliable predictions.

Model evaluation and comparison are performed using metrics such as accuracy, precision, recall, F1-score to compare the performance of different algorithms and select the most effective model for predicting sleep quality. Insights gained from feature importance analysis help identify the most influential factors affecting sleep quality, guiding the development of targeted interventions. The project also emphasizes the real-world application of the model by generating personalized recommendations for students based on their predicted sleep quality. These recommendations provide actionable insights that encourage students to

adopt healthier sleep habits, ultimately enhancing their academic performance and mental health.

Implications of Results

The results of this project hold significant implications for improving the sleep quality and overall well-being of university students. Accurate prediction of sleep quality enables early identification of poor sleep patterns, allowing for timely intervention and personalized guidance. By identifying high-risk individuals prone to poor sleep quality, universities can implement targeted wellness initiatives, promote awareness of sleep hygiene, and encourage healthier sleep behaviors. The insights gained from the predictive model also highlight the impact of lifestyle factors such as study hours, screen time, and caffeine consumption on sleep quality, allowing students to make informed decisions about their daily habits. Personalized recommendations derived from the model offer practical advice tailored to individual sleep patterns, empowering students to improve their sleep duration and quality. Moreover, integrating these findings into real-time monitoring systems can enhance the effectiveness of sleep management programs by providing continuous feedback and personalized interventions.

The implications extend beyond individual benefits, as the findings contribute to a deeper understanding of the relationship between sleep quality and academic success. Universities can use this knowledge to design comprehensive wellness programs that address the unique challenges faced by their student populations. Additionally, the ability to predict and analyze sleep quality opens the door for future research in developing AI-powered tools that monitor sleep patterns and provide real-time feedback, ensuring long-term behavioral improvements. By fostering healthier sleep habits, this project aims to enhance cognitive performance, emotional resilience, and overall quality of life among university students, paving the way for a more supportive and health-conscious academic environment.

CHAPTER II

LITERATURE REVIEW

TITLE : "Predicting sleep based on physical activity, light exposure, and Heart rate . variability data using wearable devices." [1]

AUTHORS : Kyung Mee Park, Sang Eun Lee, Changhee Lee, Hyun Duck Hwang, Do Hoon Yoon

- This paper explored the relationship between sleep quality and physiological data such as physical activity, light exposure, and heart rate variability (HRV) using data collected from wearable devices. The study aimed to predict sleep patterns by analyzing data obtained from participants over a specific period. The dataset included information on various factors influencing sleep quality, such as the duration and intensity of physical activity, exposure to different light levels, and fluctuations in heart rate variability. Data preprocessing was carried out to clean and standardize the data, ensuring consistency and accuracy. Exploratory Data Analysis (EDA) was conducted to identify trends and correlations between these factors and sleep quality. Machine learning models were applied to predict sleep patterns based on the preprocessed data, with performance evaluation carried out using accuracy, precision, and other relevant metrics. The results highlighted that increased physical activity and controlled light exposure positively impacted sleep quality, while irregular heart rate variability was associated with poor sleep. The findings underscored the potential of using wearable devices to monitor and predict sleep patterns effectively, offering valuable insights into improving sleep hygiene. The study emphasized the importance of incorporating real-time physiological data to develop personalized recommendations for enhancing sleep quality. Overall, this paper demonstrated the feasibility of leveraging machine learning and wearable device data to gain a deeper understanding of sleep behavior and health outcomes. The study also explored the potential of integrating additional physiological metrics, such as skin temperature and sleep stages, to further refine the prediction model. Future work could focus on expanding the dataset to include a larger and more diverse population to improve the generalizability of the results. Overall, the integration of wearable devices and machine learning offers a promising avenue for real-time, individualized sleep improvement interventions.

TITLE : "Adolescents' sleep patterns and psychological functioning"[2]

AUTHORS : Lemola, S., Perkinson-Floor, N., Brand, S., Dewald-Kaufmann, J. F., & Grob, A.

- The study by Lemola et al. (2015) delves into the vital connection between adolescents' sleep patterns and their psychological functioning, focusing on how factors like sleep duration, timing, and quality can deeply impact mental health. During adolescence, biological changes, such as shifts in circadian rhythms, often cause delayed sleep onset, which conflicts with societal expectations, including early school start times. This misalignment can result in sleep deprivation, negatively affecting cognitive performance, emotional regulation, and overall well-being. The paper explores how poor sleep is frequently associated with mental health issues, including depression, anxiety, and increased stress. It highlights the bidirectional relationship between sleep and mental health, where insufficient sleep exacerbates psychological problems, and pre-existing mental health conditions can further disrupt sleep.Adolescents are particularly vulnerable to the effects of poor sleep due to the developmental changes occurring during this life stage, which influence both biological sleep patterns and psychological outcomes. These sleep disturbances can lead to difficulties in academic performance, interpersonal relationships, and emotional stability. Inadequate sleep also impairs decision-making, attention span, and memory, further compounding the challenges adolescents face during this developmental period. Given the complexity of these issues, Lemola et al. stress the importance of understanding the dynamic interplay between sleep and psychological health in adolescents.The authors also emphasize the significance of sleep hygiene in promoting healthy sleep and improving mental well-being. Sleep hygiene includes practices like maintaining a consistent sleep schedule, creating a conducive sleep environment, and reducing screen time before bed to help regulate sleep patterns. Adolescents are increasingly exposed to electronic devices, which can interfere with sleep by delaying sleep onset and disrupting circadian rhythms due to the blue light emitted from screens. The paper suggests that these sleep disturbances can be mitigated by adopting better sleep hygiene practices, particularly through limiting screen use in the evening.

TITLE : "Sleep Quality Prediction from Wearable Device Data."[3]

AUTHORS : Chanda, H. N. S.

- The study by Chanda (2024) focuses on predicting sleep quality using data from wearable devices, offering a modern approach to understanding sleep patterns. The research explores how physiological metrics such as heart rate, physical activity, and sleep stages collected from wearable devices can be used to predict sleep quality with high accuracy. Chanda emphasizes the growing role of wearable technology in monitoring sleep, as it allows for continuous, real-time data collection, which is more detailed and accurate than traditional sleep studies. By using machine learning algorithms, the study aims to develop a model that can analyze the large volumes of data generated by wearable devices and predict sleep quality effectively. The paper discusses various machine learning techniques, including decision trees and neural networks, for analyzing the data. It also highlights the importance of preprocessing the data to handle missing values, normalize inputs, and extract relevant features before training models. The author demonstrates the potential of these technologies in providing personalized insights into sleep behavior and health. This approach is particularly beneficial in identifying factors influencing poor sleep quality, such as physical inactivity, irregular sleep patterns, or disruptions in heart rate variability.One of the key findings of the study is the correlation between physical activity levels and improved sleep quality, with more active individuals generally experiencing better sleep. The research also emphasizes the potential for integrating wearable device data with other lifestyle factors, such as diet and stress levels, to create more holistic models of sleep prediction. However, the study also points out the challenges of achieving accurate predictions, particularly when dealing with noisy data or individual variations in sleep needs.Chanda's work ultimately underscores the promise of wearable technology in the field of sleep research, with the potential to create real-time, personalized recommendations for improving sleep quality. However, the paper calls for further refinement of machine learning models and the integration of additional data sources to enhance predictive accuracy.

TITLE: "Predicting Sleep Quality through Biofeedback: A Machine Learning Approach Using Heart Rate Variability and Skin Temperature"[4]

AUTHORS: Di Credico, A., Perpetuini, D., Izzicupo, P., La Malva, P., Gaggi, G., Mammarella, N., & Cardone, D.

- The study by Di Credico et al. (2024) investigates the potential of using biofeedback data, specifically heart rate variability (HRV) and skin temperature, to predict sleep quality through machine learning models. The authors emphasize the growing interest in biofeedback as a tool to monitor and improve sleep, focusing on physiological markers that can provide valuable insights into sleep patterns and overall sleep health. The paper highlights how HRV, which reflects autonomic nervous system activity, and skin temperature, a key indicator of circadian rhythms, are instrumental in assessing sleep quality. By applying machine learning algorithms, such as support vector machines and random forests, the study demonstrates how these biofeedback metrics can be used to predict sleep outcomes with high accuracy. The research underscores the importance of preprocessing biofeedback data, including normalization and feature extraction, to enhance the performance of predictive models. One of the key findings is the strong correlation between HRV and sleep quality, as higher HRV values were associated with better sleep outcomes, while lower HRV indicated poor sleep quality. Similarly, the study identifies skin temperature fluctuations as another significant predictor, with stable and favorable temperature patterns contributing to improved sleep. Moreover, the authors discuss the advantages of using biofeedback data, as it provides non-invasive, continuous monitoring, making it a valuable tool for real-time, personalized sleep quality predictions. The research further explores the challenges of handling noisy biofeedback data and the need for sophisticated algorithms to refine predictions and account for individual variations. The paper calls for additional research to integrate more biofeedback signals and external factors such as environmental conditions and lifestyle behaviors to improve prediction accuracy. In conclusion, the work by Di Credico et al. (2024) highlights the promising role of biofeedback in sleep quality prediction and advocates for further exploration of its potential in personalized sleep interventions. The study emphasizes the importance of developing robust machine learning models that can leverage biofeedback data to provide real-time, actionable recommendations for improving sleep health.

TITLE: "Research on Sleep Health Prediction and Algorithms Based on Big Data"[5]
AUTHORS: Li, Mo

- Li (2023) investigates the potential of big data and machine learning algorithms to predict sleep health outcomes. The study emphasizes the growing importance of utilizing large datasets to understand sleep patterns and identify key factors that influence sleep health. By leveraging big data, including variables like sleep duration, quality, and lifestyle behaviors, the study aims to develop predictive models that can accurately assess sleep health. Li explores various machine learning techniques such as regression analysis and neural networks to process vast amounts of data. The research also highlights the challenges of handling and analyzing big data, such as data privacy concerns and standardization issues. One of the key findings of the study is the identification of factors such as screen time, sleep consistency, and physical activity, which significantly affect sleep quality. The paper suggests that integrating data from wearable devices, sleep diaries, and environmental sensors could provide a more comprehensive understanding of sleep patterns. Furthermore, Li discusses how real-time predictions could offer personalized recommendations to improve sleep health. The study also points out the limitations of current algorithms and advocates for further research to refine prediction models and enhance their accuracy. Li calls for the integration of diverse data sources and collaboration with healthcare professionals to create effective, real-world applications for sleep health monitoring. In addition, the author stresses the need for developing models that can scale across diverse populations and environments, making them more applicable in real-world scenarios. This paper offers valuable insights into the future potential of big data for improving sleep health outcomes on a global scale. The author further explores the potential of integrating environmental data, such as noise levels and light exposure, into predictive models, which could improve sleep health predictions. By incorporating these external factors, the model could better account for the diverse environments that influence sleep. Additionally, the paper discusses how these predictive models can be used in personalized sleep interventions, providing actionable insights for individuals to optimize their sleep health. The study also highlights the importance of collaboration between data scientists and healthcare professionals to translate these predictions into practical tools for improving sleep hygiene across different populations.

TITLE: "Sleep Loss, Learning Capacity, and Academic Performance"[6] AUTHORS: Curcio, G., Ferrara, M., & De Gennaro, L.

- Curcio et al. (2006) examine the impact of sleep deprivation on learning and academic performance, particularly focusing on the cognitive and psychological effects of insufficient sleep. The study emphasizes that sleep plays a crucial role in memory consolidation, attention, and problem-solving, all of which are essential for academic success. The authors discuss various mechanisms by which sleep deprivation impairs cognitive performance, such as reduced attention span, slower processing speeds, and difficulty in retaining new information. They highlight the bidirectional relationship between sleep loss and academic performance, where poor sleep leads to diminished learning capacity, and academic stress or performance demands further disrupt sleep patterns. The paper reviews multiple studies and provides compelling evidence showing that sleep loss negatively affects students' ability to perform complex tasks, solve problems, and concentrate. The authors advocate for the importance of sleep hygiene and suggest interventions, such as improving sleep schedules and reducing academic stress, to mitigate the adverse effects of sleep deprivation. The study also explores the broader consequences of chronic sleep deprivation, including mental health issues such as depression and anxiety. Curcio and colleagues call for schools and parents to educate students on the importance of sleep and implement policies that promote healthier sleep habits. Overall, the paper underscores the need for better sleep management to improve cognitive function and academic outcomes. Additionally, it emphasizes the societal need to rethink educational systems to prioritize sleep and mental health, advocating for later school start times and reduced academic pressures to support optimal sleep for students. The study further explores the long-term consequences of chronic sleep deprivation on academic performance, noting that sleep loss over extended periods can lead to persistent cognitive deficits. It also addresses how sleep deprivation impacts students' ability to retain and recall information, particularly in the context of high-stakes exams. The paper suggests that improved sleep quality, along with proper time management and stress reduction techniques, could help students better manage academic pressures.

TITLE: "Understanding Adolescents' Sleep Patterns and School Performance: A Critical Appraisal"[7]

AUTHORS: Wolfson, A. R., & Carskadon, M. A.

- Wolfson and Carskadon (2003) critically assess the relationship between sleep patterns and academic performance in adolescents, focusing on how biological and societal factors influence sleep behavior. The study points out that adolescents undergo biological changes during puberty that lead to a natural shift in sleep patterns, with a tendency for delayed sleep onset. This shift often conflicts with early school start times, contributing to sleep deprivation. The authors explore how insufficient sleep negatively affects cognitive performance, including memory, attention, and learning capacity, which are all vital for academic success. The paper emphasizes the need for schools to consider the biological sleep needs of adolescents by possibly adjusting school start times. The research also discusses the consequences of sleep deprivation on mental health, such as increased risk of anxiety, depression, and irritability. Wolfson and Carskadon advocate for a shift in societal attitudes towards sleep, urging schools, parents, and healthcare providers to recognize the importance of adequate sleep for students' academic success and mental well-being. The authors suggest that interventions aimed at improving sleep hygiene, such as later school start times and reducing evening screen time, could help mitigate the impact of sleep deprivation. The study calls for further research to explore the long-term effects of sleep deprivation on adolescent development and academic achievement. Additionally, Wolfson and Carskadon emphasize the need for comprehensive strategies, including educational campaigns and policy changes, to help adolescents manage sleep effectively. This paper provides critical insights into how sleep patterns affect not only academic success but also emotional and social well-being, calling for broader systemic changes to support better sleep health.

TITLE: "Cognitive Performance, Sleepiness, and Mood in Partially Sleep-Deprived Adolescents: The Need for Sleep Study"[8]

AUTHORS: Lo, J. C., Ong, J. L., Leong, R. L., Gooley, J. J., & Chee, M. W. (2016)

- Lo et al. (2016) explore the cognitive, mood, and performance effects of partial sleep deprivation in adolescents. The study examines how even modest sleep restriction can impair cognitive abilities such as attention, memory, and decision-making, all of which are crucial for academic and social functioning. The authors focus on the impact of sleepiness and mood disturbances, including irritability and anxiety, which are common outcomes of inadequate sleep. Using experimental studies, the research demonstrates that partially sleep-deprived adolescents struggle to concentrate, retain information, and perform complex tasks. The study also reveals that mood disturbances, such as increased irritability and negative emotions, are exacerbated by sleep loss, further affecting cognitive function. Lo and colleagues emphasize the cumulative effects of partial sleep deprivation, which can impair adolescents' ability to perform well in both academic and social settings. The authors suggest that interventions aimed at improving sleep hygiene, such as consistent sleep schedules and reducing evening screen time, could help mitigate the effects of sleep deprivation. The research advocates for increased awareness and education about the importance of sleep for cognitive and emotional well-being. The paper calls for schools to reconsider early start times and homework schedules to help ensure that adolescents receive adequate rest. Overall, Lo et al. stress the need for policy changes and further research to better understand the relationship between sleep deprivation, cognitive performance, and mental health in adolescents. The study concludes by emphasizing the importance of public health initiatives to improve sleep education, which could potentially enhance both academic outcomes and overall adolescent development.

CHAPTER III

DATASET DESCRIPTION

The dataset used in this project captures a comprehensive range of information on the relationship between sleep patterns, academic performance, and lifestyle factors among university students. It includes attributes that describe students' sleep habits, daily routines, study behavior, and health-related metrics. Each record in the dataset represents an individual student, providing a holistic view of how various factors interact to influence sleep quality and academic outcomes.

Attributes and Their Descriptions

The dataset contains 10000 rows (students) and 14 columns. Each row represents a unique student, and the columns describe their attributes and sleep-related information.

Student_ID	Unique identifier for each student
Age	Age of the student in years
Gender	Gender of the student (categorical: 'Male', 'Female', 'Other').
University_Year	The student's current year of university (categorical: '1st Year', '2nd Year', '3rd Year', '4th Year').
Sleep_Duration	Total hours of sleep per night
Study_Hours	Average number of hours spent studying per day
Screen_Time	Average number of hours spent on screens (excluding studying) per day
Caffeine_Intake	Average number of caffeinated beverages consumed per day
Physical_Activity	Average minutes spent on physical activity per day
Sleep_Quality	Subjective rating of sleep quality on a scale of 1 to 10 (1 being the worst, 10 being the best)
Weekday_Sleep_Start	Time the student typically goes to sleep on weekdays
Weekend_Sleep_Start	Time the student typically goes to sleep on weekends
Weekday_Sleep_End	Time the student typically wakes up on weekdays
Weekend_Sleep_End	Time the student typically wakes up on weekends

Student Information:

The dataset includes basic demographic information such as Student_ID, which serves as a unique identifier for each student, ensuring that records remain distinct. Age records the age of the student in years, providing a basis for analyzing how sleep patterns and academic performance vary across different age groups. Gender captures the student's gender (e.g., Male, Female, or Other), facilitating an analysis of gender-based differences in sleep quality and study habits. University_Year specifies the academic year of the student, allowing for an exploration of how academic demands across different years influence sleep behavior.

Sleep-Related Attributes

The dataset includes several attributes that provide a detailed view of students' sleep patterns. Sleep_Duration records the total number of hours a student sleeps per night, making it a key factor in assessing sleep quality. Sleep_Quality is a subjective measure where students rate the quality of their sleep (e.g., Poor, Healthy and Unhealthy), serving as the target variable for prediction tasks. Additionally, Weekday_Sleep_Start and Weekday_Sleep_End record the time students go to bed and wake up on weekdays, allowing for the analysis of weekday sleep schedules. Similarly, Weekend_Sleep_Start and Weekend_Sleep_End capture sleep patterns on weekends, highlighting differences in sleep schedules between weekdays and weekends

Study and Lifestyle Attributes

To analyze the impact of study habits and lifestyle choices on sleep quality, the dataset includes attributes such as Study_Hours, which indicates the average number of hours a student spends studying daily. This attribute helps assess how academic workload affects sleep patterns. Screen_Time measures the total daily screen time in hours, providing insights into how prolonged exposure to electronic devices may disrupt sleep onset and quality. Caffeine_Intake records the daily consumption of caffeine (measured in cups), which is a known factor that can delay sleep onset and reduce sleep quality. Physical_Activity reflects the student’s level of physical activity, categorized as Low, Moderate, or High, offering insights into the correlation between exercise and improved sleep patterns.

Derived Features for Model Training

Several derived features have been included to enhance the dataset’s predictive power and capture more granular insights into students' sleep patterns. One such feature is Sleep Difference, which quantifies the variance between weekday and weekend sleep durations. This

metric helps identify irregularities in sleep schedules that may contribute to poor sleep quality and fatigue. Weekend Sleep Duration and Weekday Sleep Duration are also calculated to provide a clearer understanding of how students' sleep duration varies across the week. Furthermore, Sleep Health Labels categorize sleep health into distinct classes:

- **Poor Sleep:** Less than 6 hours of sleep.
- **Healthy Sleep:** Between 6 to 8 hours.
- **Unhealthy Sleep:** More than 8 hours.

These derived features allow for a more detailed exploration of how variations in sleep duration across different days influence sleep quality and psychological functioning.

Significance of the dataset

The dataset used in this project is highly significant as it provides a comprehensive view of the relationship between sleep patterns, lifestyle habits, and academic performance among university students. It includes key attributes such as Sleep Duration, Study Hours, Screen Time, Caffeine Intake, Physical Activity, Sleep Quality, and Sleep Start and End Times (Weekdays and Weekends), enabling the development of accurate machine learning models to predict sleep health. By identifying critical factors such as excessive screen time, irregular sleep patterns, and low physical activity, the dataset facilitates targeted interventions to improve sleep quality. It also supports the generation of personalized recommendations to enhance sleep hygiene and academic outcomes. The dataset's richness allows for real-world applications, with the potential to integrate real-time data from wearable devices like Fitbit and Apple Health to enhance predictive accuracy. Additionally, it promotes interdisciplinary research by combining insights from sleep science, psychology, and data science, contributing to improved student well-being and long-term health.

CHAPTER IV

METHODOLOGY

The methodology followed in this project involved several key stages to ensure accurate prediction of sleep quality and identification of influencing factors among university students. The process included data collection, preprocessing, exploratory data analysis (EDA), feature engineering, model selection, and performance evaluation, followed by future recommendations for model improvement and real-world deployment. Data preprocessing involved handling missing values, encoding categorical variables, and scaling numerical features to ensure data consistency. EDA helped uncover correlations between features such as sleep duration, screen time, and study hours, providing valuable insights into student sleep patterns. Feature engineering involved creating new variables, such as sleep duration differences between weekdays and weekends, to enhance model performance. Finally, model evaluation was performed using multiple metrics to ensure the selected models could effectively predict sleep quality.

Data Preparation:

- **Data Cleaning:** Handled missing values using appropriate imputation techniques to prevent inconsistencies and biases. Applied median or mode imputation for numerical and categorical variables respectively to preserve data integrity.
- **Encoding Categorical Variables:** Transformed categorical features such as gender and university year using one-hot or label encoding for model compatibility. One-hot encoding was applied to prevent introducing ordinal relationships where none exist.
- **Scaling Numerical Features:** Standardized or normalized numerical features (e.g., sleep duration, study hours, and screen time) to ensure uniform scale and prevent feature dominance. Min-max scaling was used to bring all features within a 0-1 range, improving model efficiency.
- **Outlier Detection and Treatment:** Identified and treated outliers to minimize their impact on model performance. Z-score and IQR (Interquartile Range) techniques were used to detect and handle outliers.
- **Feature Engineering:** Created derived features, such as the difference between weekday and weekend sleep duration, to capture variations in sleep patterns. Added

new features like average sleep duration and consistency of sleep patterns to enhance prediction.

- **Data Splitting:** Split the dataset into training and testing sets to evaluate model performance on unseen data and ensure generalizability. Stratified splitting was used to maintain class distribution across training and test sets.

Approaches to improve performance:

- **Data Exploration and Preprocessing:** Addressed missing values, encoded categorical features, and scaled numerical variables for consistency. Analyzed the distribution of each variable to identify inconsistencies and patterns.
- **Exploratory Data Analysis (EDA):** Identified correlations between key variables such as sleep duration, study hours, screen time, and physical activity. Visualized data through heatmaps and pair plots to assess feature relationships and trends.
- **Feature Engineering:** Derived additional features, including the difference between weekday and weekend sleep durations, to capture variations in sleep patterns. Created features representing sleep irregularity and weekend catch-up to capture nuanced insights.
- **Model Implementation:** Applied various machine learning models, including Random Forest, XGBoost, Logistic Regression, KNN, and SVM. Implemented cross-validation to assess model robustness and prevent overfitting.
- **Performance Evaluation:** Assessed models using accuracy, precision, recall, and F1-score to compare performance. Generated confusion matrices and ROC-AUC curves to evaluate classification effectiveness.
- **Hyperparameter Tuning:** Optimized models through hyperparameter tuning to enhance their accuracy and effectiveness. Used grid search and random search to find the best parameter combinations.
- **Model Selection and Integration:** Selected the best-performing models for future integration with real-time data from wearable devices to ensure real-world applicability. Ensured scalability and efficiency of the models for deployment in real-time systems.

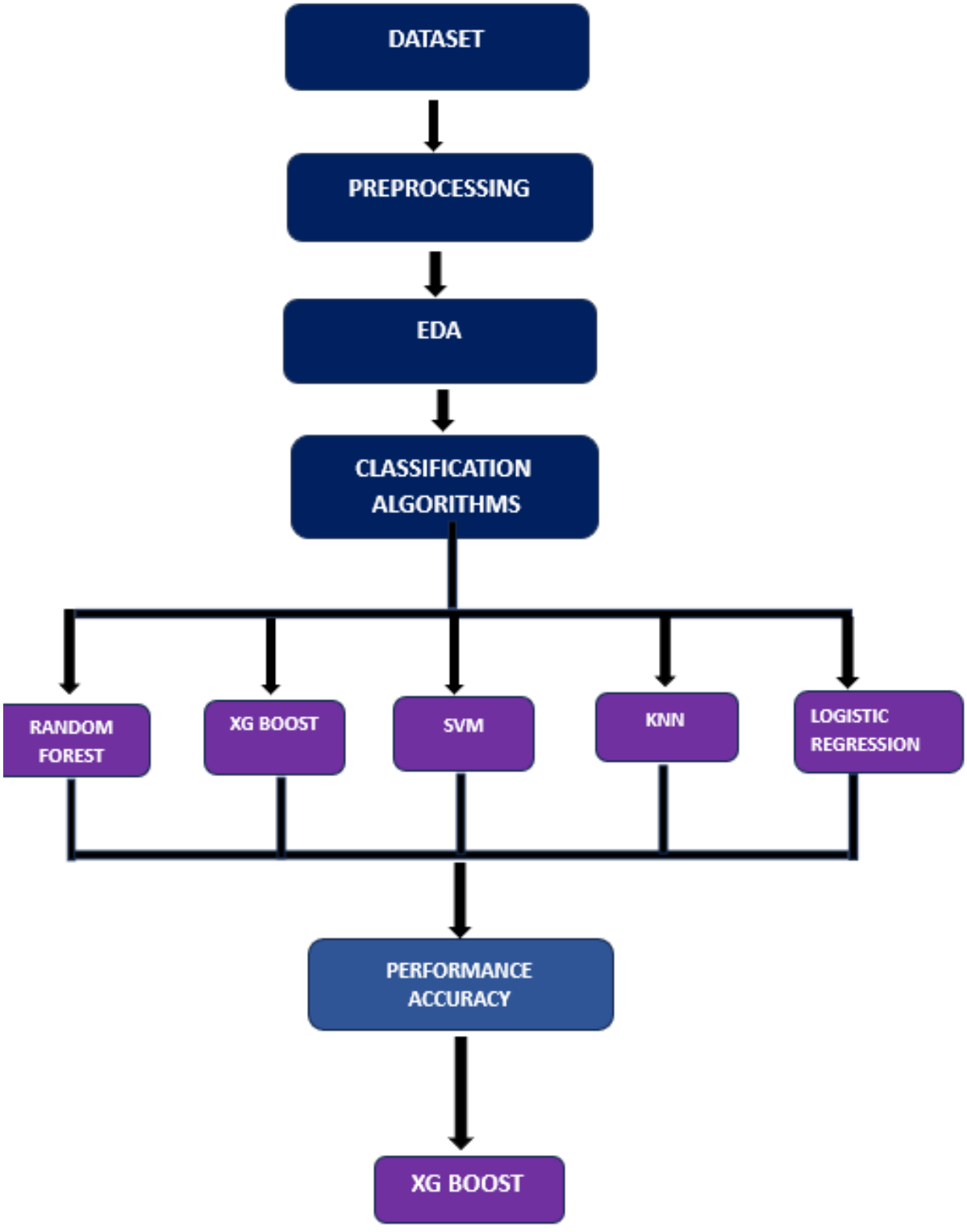


Figure 1. Architecture of working methodology

Models Implemented:

- **Random Forest:**

  An ensemble algorithm that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. Hyperparameter tuning was performed to optimize estimators and tree depth.

- **XGBoost (Extreme Gradient Boosting):**

  A gradient-boosting algorithm that constructs decision trees sequentially to minimize errors. Early stopping and fine-tuning of learning rates were applied to enhance model generalization.

- **Logistic Regression:**

  A statistical model that predicts class probabilities by establishing a linear relationship between features and the target variable. Regularization techniques (L1 and L2) were used to prevent overfitting.

- **K-Nearest Neighbors (KNN):**

  A classification algorithm that assigns class labels based on the majority vote of k-nearest neighbors. Optimal k-values were selected using cross-validation.

- **Support Vector Machine (SVM):**

  A classification model that finds the optimal hyperplane to separate classes with maximum margin. Various kernel functions (linear, polynomial, RBF) were explored to improve performance.

Dataset Extraction:

The dataset used in this project was generated synthetically to simulate real-world data reflecting the lifestyle and behavioral factors affecting sleep quality among university students. The synthetic data was designed to closely resemble patterns observed in actual survey-based studies, capturing diverse attributes relevant to sleep quality predictions.

Synthetic Data Generation:

The data was generated using statistical techniques to mimic realistic distributions and correlations between features. Attributes such as sleep duration, study hours, screen time,

caffeine intake, physical activity, and sleep quality were modeled to reflect typical variations observed in a student population.

Feature Design:

- **Continuous Variables:** Sleep duration, study hours, and screen time were generated using Gaussian distributions to capture normal variations.
- **Categorical Variables:** Features such as gender, university year, and BMI category were created using random sampling to ensure realistic proportions.

Data Consistency and Validation:

To maintain consistency, feature relationships such as the correlation between sleep duration and physical activity or the impact of excessive screen time on sleep quality were preserved in the synthetic data. The generated data was validated by comparing statistical properties, including mean, standard deviation, and feature distributions, to ensure realistic representation.

Derived Features:

Additional features, such as the difference between weekday and weekend sleep duration, were created to capture variations in sleep patterns, adding depth to the dataset and improving predictive accuracy.

The synthetic dataset provided a well-balanced combination of categorical and numerical variables, enabling the development of accurate machine learning models and facilitating insights into the factors influencing sleep health among students.

Preprocessing:

The preprocessing stage was crucial in ensuring that the dataset was clean, consistent, and ready for effective model training. It involved several steps, including handling missing values, encoding categorical variables, scaling numerical features, detecting and treating outliers, and creating derived features to capture deeper insights.

1. Handling Missing Values:

To maintain data integrity, missing values were identified and treated using appropriate imputation techniques. For numerical features like sleep duration and study hours, mean or median imputation was applied, while mode imputation was used for categorical variables such

as gender and university year. This step minimized data loss and ensured the completeness of the dataset.

2. Encoding Categorical Variables:

Categorical variables were transformed into numerical representations to make them compatible with machine learning models.

- **One-Hot Encoding:** Applied to nominal variables such as gender and university year, creating binary columns for each category.
- **Label Encoding:** Used for ordinal variables like sleep quality labels to retain ordinal relationships between categories.

3. Scaling and Normalization:

Numerical features such as sleep duration, study hours, and screen time were standardized to ensure all features operated on a similar scale, preventing any one feature from dominating the model.

- **Standardization:** Applied using z-score normalization to bring the data to a mean of 0 and a standard deviation of 1.
- **Min-Max Scaling:** In some cases, Min-Max scaling was explored to rescale values between 0 and 1 for distance-based models like KNN.

4. Outlier Detection and Treatment:

Outliers were identified using techniques such as the **Interquartile Range (IQR)** and **Z-score analysis**. Extreme values that could distort model performance were treated by either capping or removing them, depending on their impact on the dataset.

5. Feature Engineering and Derived Features:

To enhance model performance, new features were derived from the existing data.

- **Sleep Difference:** Calculated as the difference between weekday and weekend sleep duration to capture inconsistencies in sleep patterns.
- **Study-Sleep Ratio:** Created to analyze the balance between study hours and sleep duration.

These derived features added valuable insights that helped improve the predictive power of the models.

Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) was performed to gain a deeper understanding of the dataset, identify relationships between key variables, and uncover patterns influencing sleep quality among university students. EDA included various techniques such as data visualization, correlation analysis, and statistical summaries, which helped refine the dataset and guide feature selection for the machine learning models.

1. Data Overview and Statistical Summary:

A detailed examination of the dataset was conducted to summarize the central tendencies and distributions of features.

- **Descriptive Statistics:** Measures such as mean, median, standard deviation, and percentiles were calculated to assess the range and variability of numerical features such as sleep duration, study hours, screen time, and physical activity.
- **Distribution Analysis:** Histograms and box plots were used to visualize the distribution of continuous variables and detect potential outliers or skewed data.

2. Correlation Analysis:

Correlation analysis was performed to identify relationships between different variables and assess their impact on sleep quality.

- **Correlation Matrix:** A heatmap was generated to visualize the correlation between numerical features, revealing strong positive or negative relationships.
- **Key Findings:**
    - Sleep duration was negatively correlated with screen time, indicating that increased screen time reduced sleep duration.
    - Physical activity showed a positive correlation with sleep quality, highlighting its role in maintaining healthy sleep patterns.
    - Study hours had a moderate negative correlation with sleep duration, reflecting the trade-off between study time and sleep.

3. Visualization of Sleep Patterns:

Visual analysis was conducted to explore differences in sleep patterns across various categories.

- **Bar Charts:** Showed the distribution of sleep quality across different university years and gender groups.
- **Box Plots:** Illustrated variations in sleep duration, study hours, and screen time across sleep quality levels.
- **Scatter Plots:** Examined relationships between study hours, screen time, and sleep quality to highlight trends and anomalies.

4. Feature Distributions and Trends:

Visualizations provided insights into the distribution of individual features and their influence on sleep quality.

- **Weekday vs. Weekend Sleep Duration:** Box plots and histograms revealed significant differences between weekday and weekend sleep durations, emphasizing irregular sleep patterns.
- **Study-Sleep Balance:** Line graphs demonstrated how excessive study hours contributed to reduced sleep duration, affecting overall sleep quality.

5. Analysis of Derived Features:

EDA also explored the impact of derived features such as:

- **Sleep Difference:** Highlighted variations between weekday and weekend sleep patterns, indicating irregular sleep habits.
- **Study-Sleep Ratio:** Analyzed how the balance between study hours and sleep duration affected overall sleep quality.

Classification:

Classification is a supervised machine learning technique used to predict categorical outcomes by assigning input data points to predefined classes. In this project, classification was applied to predict sleep quality among university students based on various features such as sleep duration, study hours, screen time, and physical activity. The target variable, sleep quality, was categorized into classes like Poor, Moderate, Good, and Excellent. Popular classification algorithms, including Logistic Regression, Decision Tree, Random Forest, XGBoost, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), were used to build the models. The models were trained on labeled data and evaluated using key performance metrics such as accuracy, precision, recall, and F1-score to ensure robustness and reliability. By identifying patterns and relationships within the dataset, these classification models provided valuable insights into the factors influencing students' sleep quality, enabling the development of effective prediction and intervention strategies.

Logistic Regression

Logistic Regression is a supervised machine learning algorithm used for binary and multiclass classification tasks by predicting the probability that a given input belongs to a particular class. In this project, Logistic Regression was implemented to classify sleep quality based on various features such as sleep duration, study hours, screen time, and physical activity. It models the relationship between the independent variables and the target variable using a logistic (sigmoid) function, which transforms the output into a probability value between 0 and 1.

The formula for **Logistic Regression** is given by:

$p=1/(1+ \text{eps} -(b0+b1*x1+b2*x2+\ldots+bn*xn))$

- $\beta 0$ are the intercept term.
- $\beta 1,\beta 2,\ldots,\beta n$ are the coefficients of the respective features.
- $X1,X2,\ldots,Xn$ are the input features or independent variables.
- e is the base of the natural logarithm, approximately equal to 2.718.

The sigmoid (logistic) function transforms the linear combination of the features into a probability value between 0 and 1, which is then used to classify the input based on a decision threshold (usually 0.5).

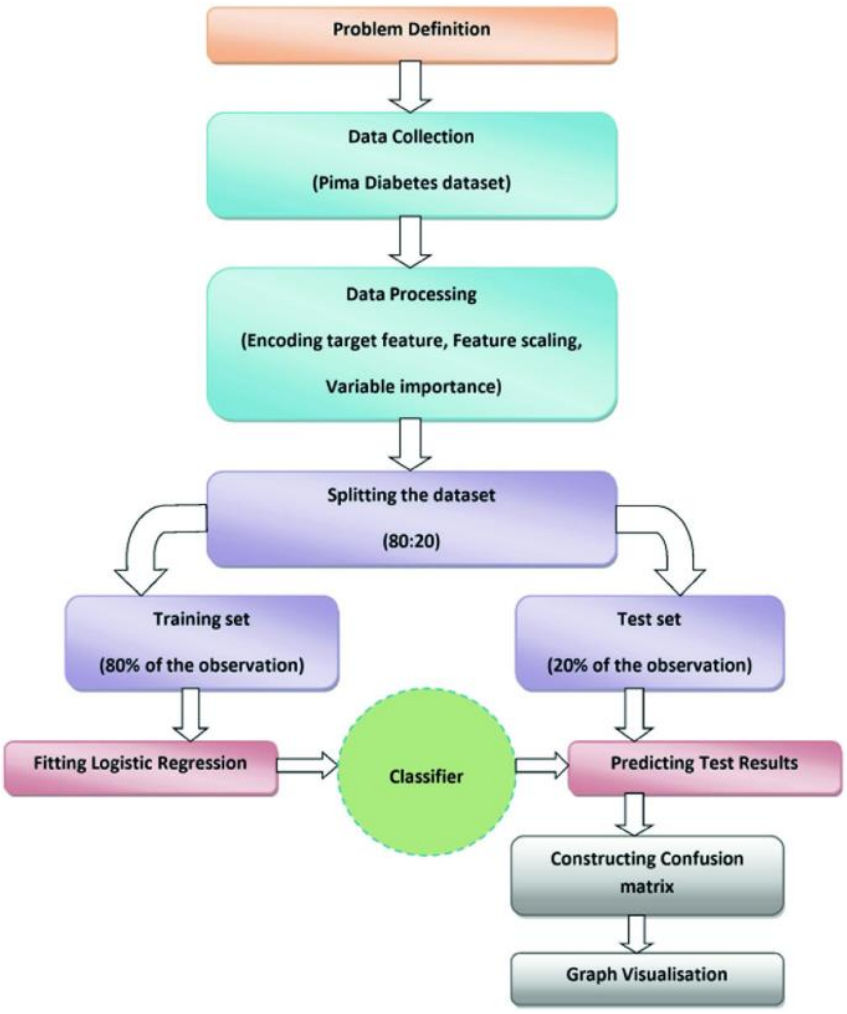


Figure 2. Work flow of Logestic Regression

Pseudocode for Logistic Regression Classification

Input: Student sleep pattern dataset

Output: Model Accuracy and Execution Time

Step 1: Import necessary libraries and read the dataset.

Step 2: Preprocess the dataset by handling missing values and encoding categorical variables.

Step 3: Split the dataset into training (70%) and testing (30%).

Step 4: Generate the logistic regression classifier with penalty as a parameter.

Step 5: Analyze the dataset by varying dependent and independent variables.

Step 6: Logistic regression predicts sleep quality as a categorical variable.

Step 7: Finally predict the probability of sleep quality using the sigmoid function.

Step 8: Print or return the accuracy of the predictions for the model.

Random Forest

Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to enhance classification or regression performance by aggregating their predictions. It works by creating random subsets of the training data through bootstrapping and selecting a random subset of features for each split, ensuring diversity and robustness among the trees. During prediction, classification tasks use majority voting, while regression tasks take the average of the tree outputs. Random Forest reduces the risk of overfitting and improves accuracy by distributing learning across multiple trees. It also highlights feature importance, making it useful for identifying key factors influencing outcomes. However, it can be computationally intensive and less interpretable compared to individual decision trees, making it less suitable for applications requiring high explainability.

In this project, Random Forest was employed to predict sleep quality based on various lifestyle factors such as sleep duration, screen time, and physical activity. By analyzing feature importance, the algorithm was able to identify the most influential factors impacting students' sleep patterns. The model achieved high accuracy and performed well in distinguishing between different sleep quality levels, making it an effective choice for this application. Additionally, hyperparameter tuning was applied to optimize the number of trees and maximum tree depth, further enhancing the model's predictive capability and ensuring robust performance across diverse data inputs.

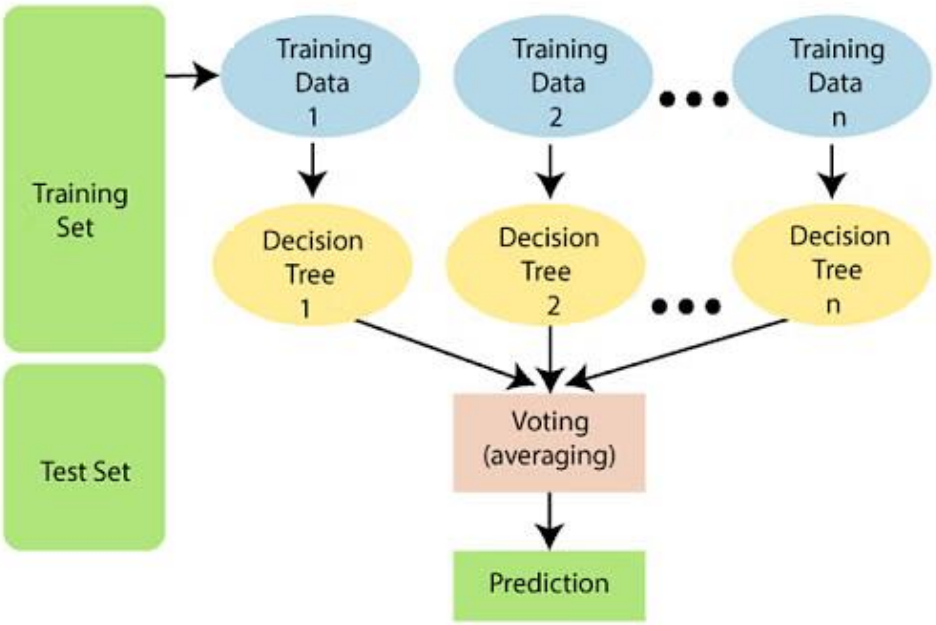


Figure 3. Work flow of Random forest

Pseudocode for Random Forest Classification

Input: Student sleep pattern dataset

Output: Model Accuracy and Execution Time

Step 1: Import necessary libraries and read the dataset.

Step 2: Preprocess the dataset by handling missing values and encoding categorical variables.

Step 3: Split the dataset into training (70%) and testing (30%).

Step 4: Generate the Random Forest classifier with the number of trees as a parameter.

Step 5: Train the model using the training data.

Step 6: Predict sleep quality for the test data.

Step 7: Evaluate the model by calculating the accuracy and classification report.

Step 8: Print or return the accuracy of the predictions for the model.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple and intuitive machine learning algorithm used for classification and regression tasks. It works by identifying the 'k' nearest data points to a given input and assigning the majority class (for classification) or averaging the values (for regression). The distance between the data points is typically measured using Euclidean distance, although other distance metrics like Manhattan or Minkowski can also be used. KNN is a non-parametric algorithm, meaning it does not make assumptions about the underlying data distribution, making it highly versatile. However, it can be computationally expensive for large datasets, as it requires calculating distances between all data points during prediction.

In this project, KNN was applied to predict sleep quality based on lifestyle attributes such as study hours, screen time, and physical activity. The optimal value of 'k' was determined using cross-validation to minimize classification errors and enhance model performance. KNN provided competitive accuracy, demonstrating its ability to capture patterns in the data effectively. However, since KNN is sensitive to the choice of 'k' and the scale of features, preprocessing steps such as feature scaling were applied to ensure consistent distance measurements and improve classification accuracy.

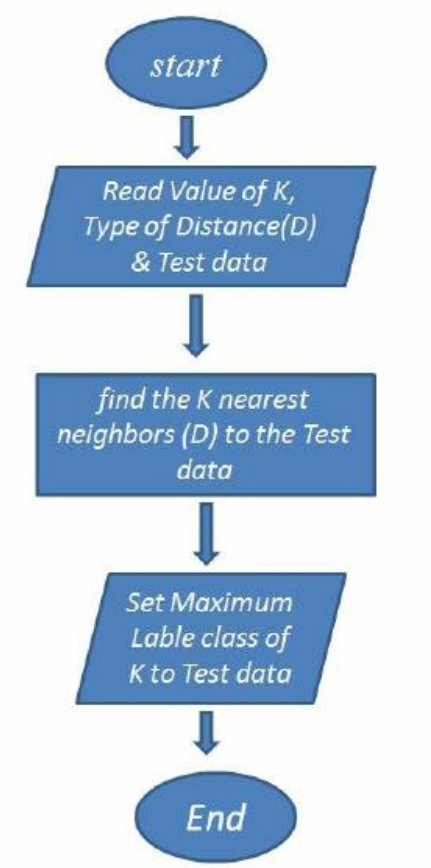


Figure 4. Work flow of KNN

Pseudocode for KNN Classification

Input: Student sleep pattern dataset

Output: Model Accuracy and Execution Time

Step 1: Import necessary libraries and read the dataset.

Step 2: Preprocess the dataset by handling missing values and encoding categorical variables.

Step 3: Split the dataset into training (70%) and testing (30%).

Step 4: Define the KNN classifier and set the number of neighbors (k).

Step 5: Fit the model using the training data.

Step 6: Predict sleep quality for the test data.

Step 7: Evaluate the model by calculating the accuracy and classification report.

Step 8: Print or return the accuracy of the predictions for the model.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful and versatile machine learning algorithm used for classification and regression tasks. It works by identifying a hyperplane that best separates data points of different classes while maximizing the margin between them. SVM aims to find the optimal decision boundary that minimizes classification errors and maximizes the distance between the closest data points (support vectors) from each class. It can handle both linear and non-linear classification by using kernel functions such as linear, polynomial, and radial basis function (RBF), allowing it to adapt to complex data structures. SVM is highly effective for high-dimensional data but can be computationally expensive, especially with large datasets.

In this project, SVM was utilized to predict sleep quality by analyzing various lifestyle factors, including sleep duration, study hours, and screen time. Different kernel functions were tested to determine the best-performing model for the dataset. SVM demonstrated good classification performance, particularly with the RBF kernel, which captured non-linear relationships in the data. To further enhance model efficiency, hyperparameter tuning was conducted to optimize parameters like the regularization parameter (C) and kernel coefficient (gamma), ensuring high predictive accuracy and robustness

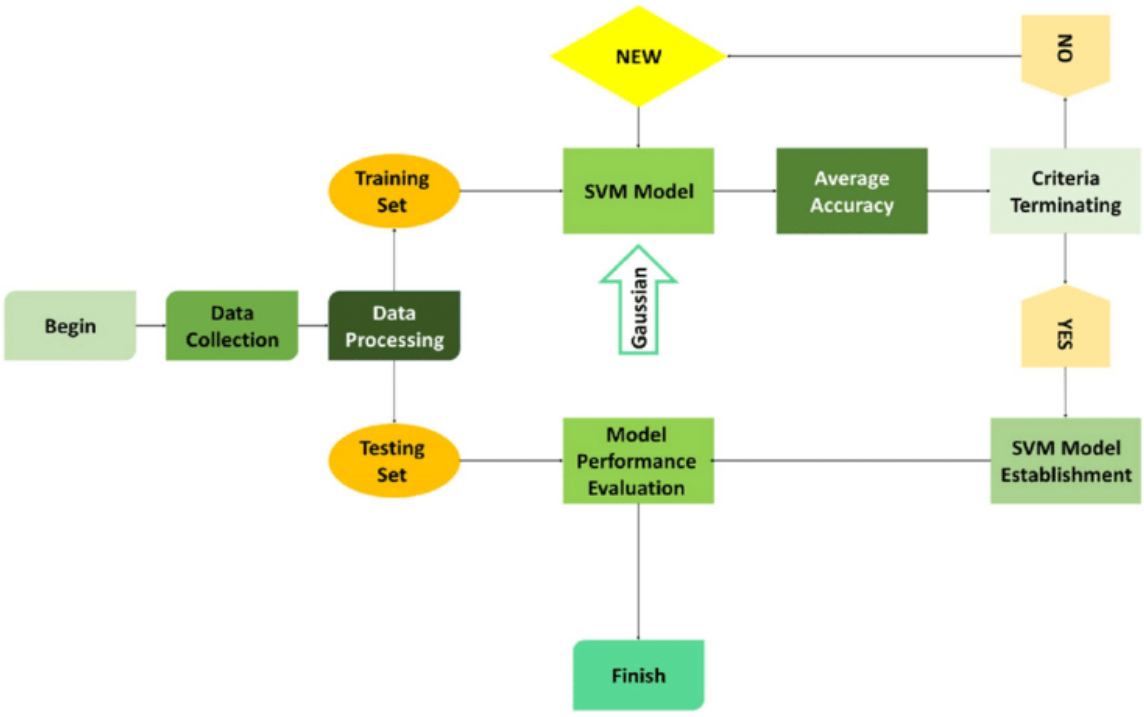


Figure 5. Work flow of SVM

Pseudocode for SVM Classification

Input: Student sleep pattern dataset

Output: Model Accuracy and Execution Time

Step 1: Import necessary libraries and read the dataset.

Step 2: Preprocess the dataset by handling missing values and encoding categorical variables.

Step 3: Split the dataset into training (70%) and testing (30%).

Step 4: Define the SVM classifier with appropriate kernel parameters.

Step 5: Fit the model using the training data.

Step 6: Predict sleep quality for the test data.

Step 7: Evaluate the model by calculating the accuracy and classification report.

Step 8: Print or return the accuracy of the predictions for the model.

XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting) is a highly efficient and powerful machine learning algorithm that builds decision trees sequentially to minimize errors and improve prediction accuracy. It employs gradient boosting, where each new tree corrects the errors made by the previous one by optimizing a loss function. XGBoost is known for its speed and scalability due to parallel processing, tree pruning, and regularization techniques that prevent overfitting. It supports various objective functions, making it suitable for both classification and regression tasks. Additionally, XGBoost handles missing values effectively and provides feature importance, making it easier to interpret model results.

In this project, XGBoost was applied to predict sleep quality by analyzing various factors such as sleep duration, screen time, and study hours. Early stopping was implemented to prevent overfitting, and hyperparameter tuning was conducted to optimize the learning rate, maximum tree depth, and number of estimators. XGBoost demonstrated exceptional performance, achieving high accuracy and robustness in identifying the key factors affecting students' sleep patterns. Its ability to handle large datasets and model complex relationships made it one of the top-performing models in this study.

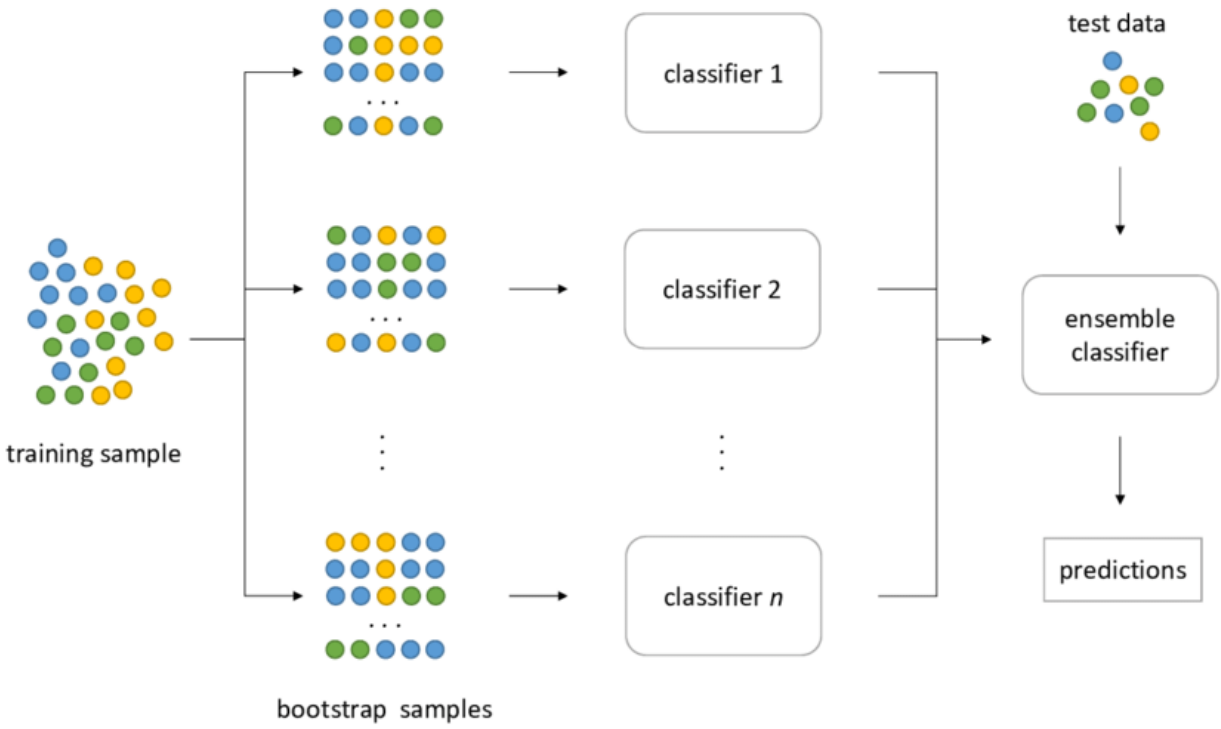


Figure 6. Work flow of XGBoost

Pseudocode for XGBoost Classification

Input: Student sleep pattern dataset

Output: Model Accuracy and Execution Time

Step 1: Import necessary libraries and read the dataset.

Step 2: Preprocess the dataset by handling missing values and encoding categorical variables.

Step 3: Split the dataset into training (70%) and testing (30%).

Step 4: Define the XGBoost classifier with hyperparameters.

Step 5: Fit the model using the training data.

Step 6: Predict sleep quality for the test data.

Step 7: Evaluate the model by calculating the accuracy and classification report.

Step 8: Print or return the accuracy of the predictions for the model.

CHAPTER V

EXPERIMENTAL RESULTS

The experimental results of the Sleep Quality Prediction Model highlight the effectiveness of various machine learning algorithms in predicting sleep quality among university students. The dataset used for this project contained features such as sleep duration, study hours, screen time, caffeine intake, and physical activity, along with derived features like Sleep_Duration_Diff, Sleep_Start_Diff, and Sleep_End_Diff to capture variations in weekday and weekend sleep patterns.

Dataset Overview

The dataset used for this project contained critical attributes that captured the lifestyle and sleep habits of university students. These attributes include:

- **Demographics:** Age, Gender, and University Year.
- **Lifestyle Factors:** Study Hours, Screen Time, Caffeine Intake, and Physical Activity.
- **Sleep Patterns:** Weekday and Weekend Sleep Start and End Times.
- **Derived Features:** Additional features such as differences in sleep duration, start time, and end time between weekdays and weekends.

The target variable, **Sleep_Label**, classified sleep quality into three categories:

- **0 - Poor Sleep**
- **1 - Healthy Sleep**
- **2 - Unhealthy Sleep**

Encoding Categorical Variables

- Gender was encoded with numerical values as:
    - **Male = 0, Female = 1, Other = 2**
- University_Year was mapped to numerical values as:
    - **1st Year = 1, 2nd Year = 2, 3rd Year = 3, 4th Year = 4**

Handling Negative Values

- Negative values in columns such as Study_Hours and Physical_Activity were corrected by converting them to positive using the **absolute value method (.abs())**.

Feature Engineering

New features were created to capture variations in sleep patterns between weekdays and weekends:

- **Sleep_Duration_Diff:** Difference between weekday and weekend sleep duration.
- **Sleep_Start_Diff:** Difference between weekday and weekend sleep start times.
- **Sleep_End_Diff:** Difference between weekday and weekend sleep end times.

Exploratory Data Analysis (EDA)

Categorical Analysis

- Count Plots were used to analyze the distribution of categorical variables such as Gender and University_Year.
- These visualizations provided insights into how demographic factors may influence sleep patterns and lifestyle habits**.**

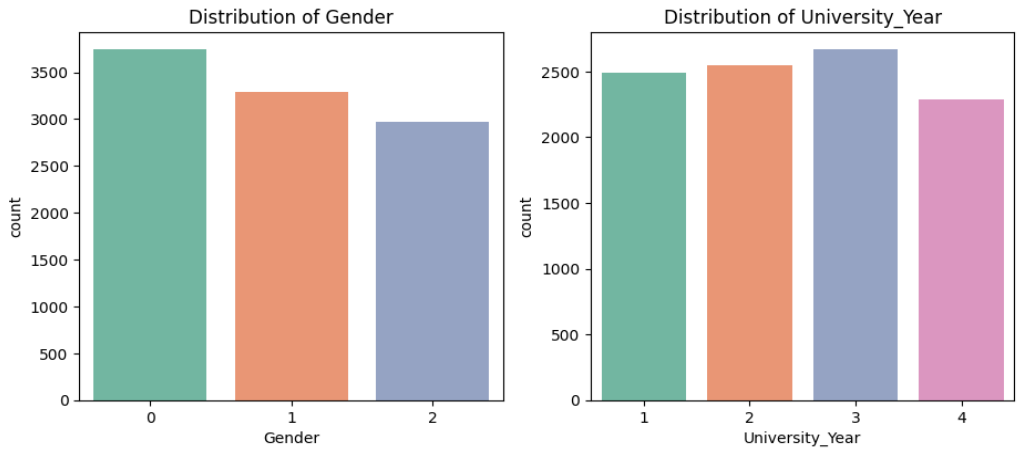


Figure 7. Distribution of gender and university

Distribution Analysis

- Histograms and Box Plots were used to visualize the distribution of numerical features, highlighting trends in variables like sleep duration, study hours, and screen time.

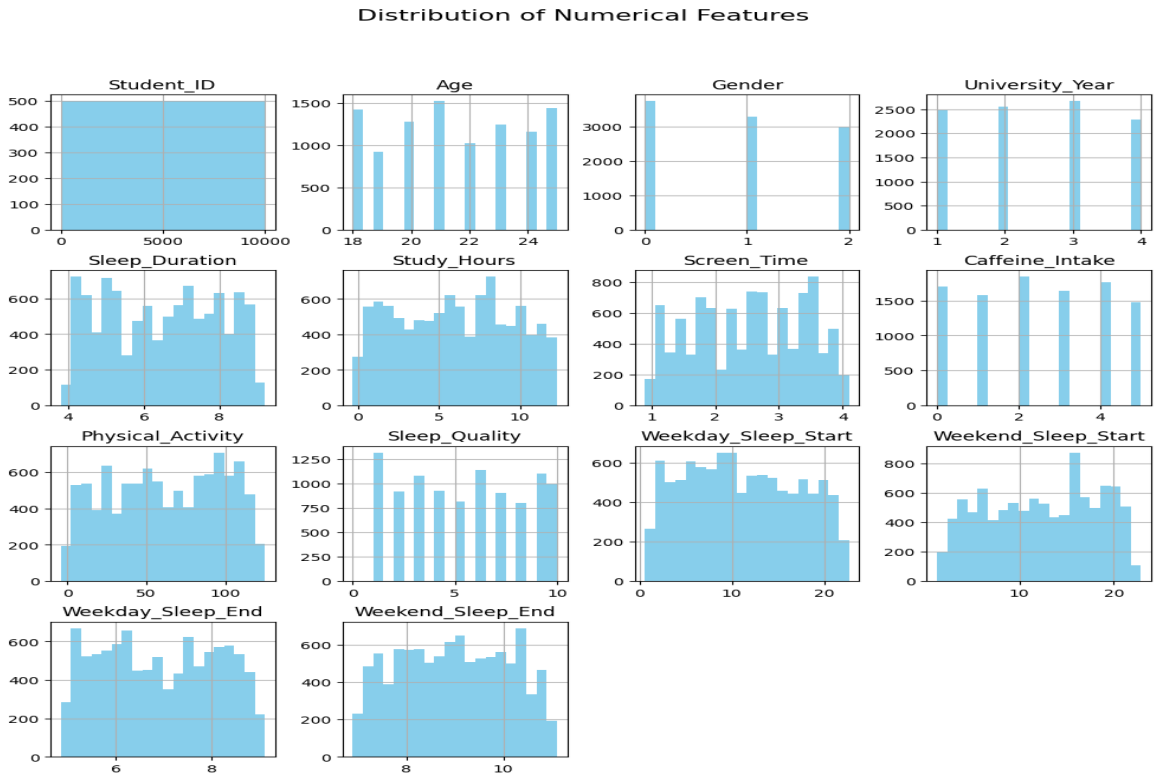


Figure 8. Distribution of numerical features

Correlation Matrix : A **correlation heatmap** was generated to examine relationships between numerical variables.

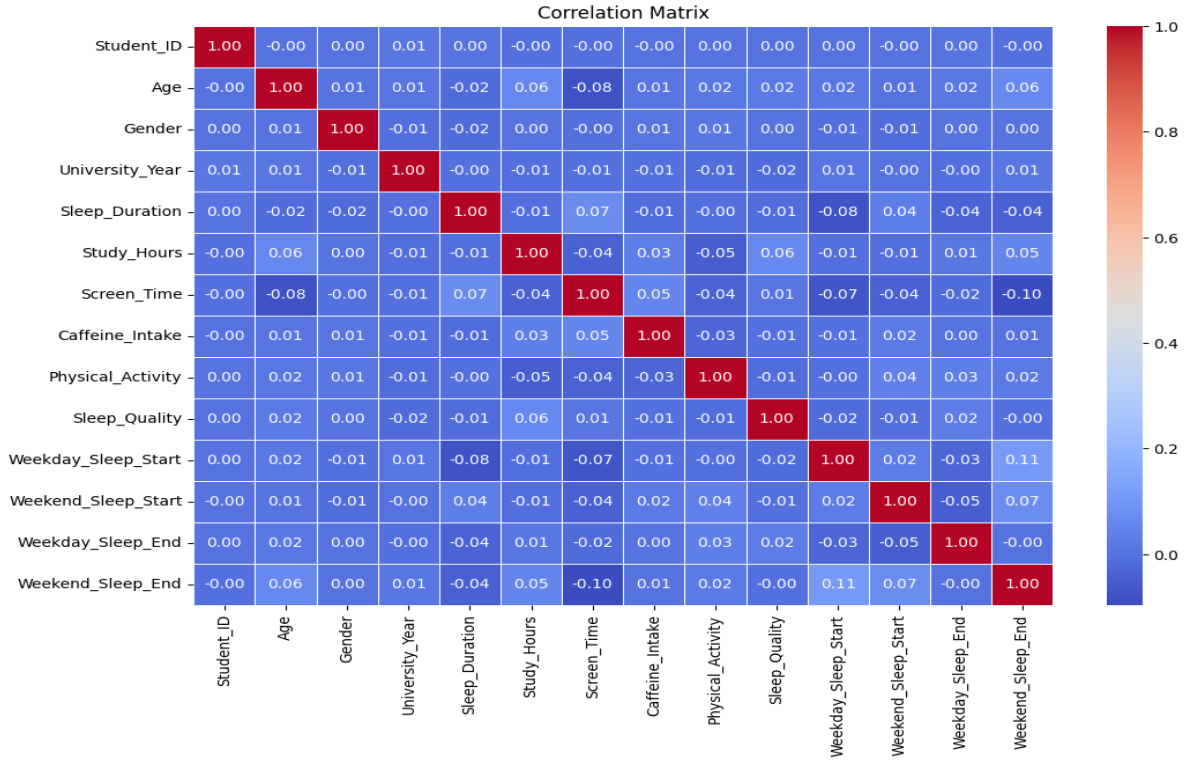


Figure 9. Correlation matrix

Feature Importance

Feature Importance quantifies how much each feature contributes to the prediction.

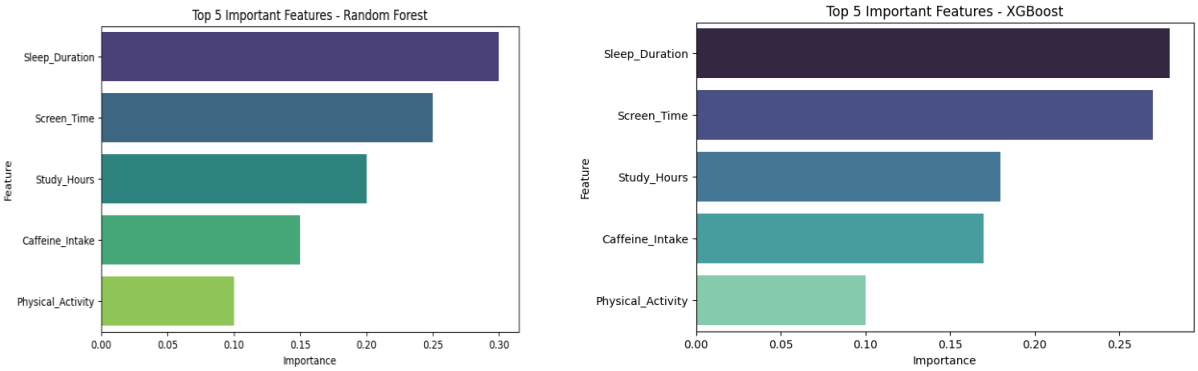


Figure 10. Importance features of the models

Random Forest Classifier

The **Random Forest Classifier** demonstrated excellent performance with an accuracy of **99.65%**. It effectively identified critical features, such as **Sleep_Duration**, **Screen_Time**, **Study_Hours**, **Caffeine_Intake**, and **Physical_Activity**, contributing significantly to predicting sleep quality. The ensemble nature of this model allowed it to capture complex patterns in the dataset, making it one of the most reliable models.

```
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       488
           1       0.98      1.00      0.99       285
           2       1.00      1.00      1.00      1227

    accuracy                           1.00      2000
   macro avg       0.99      1.00      1.00      2000
weighted avg       1.00      1.00      1.00      2000

Accuracy for Random Forest: 0.9965
```

Figure 11. Random forest classification report

XGBoost Classifier

The **XGBoost Classifier** emerged as the best-performing model with an accuracy of **99.95%**. It efficiently handled large datasets and identified feature importance with high precision. Similar to Random Forest, XGBoost emphasized **Sleep_Duration** and

Screen_Time as the most influential predictors. The model's superior performance was due to its gradient boosting mechanism, optimizing the learning process and reducing errors.

```
XGBoost Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       488
           1       1.00      1.00      1.00       285
           2       1.00      1.00      1.00      1227

    accuracy                           1.00      2000
   macro avg       1.00      1.00      1.00      2000
weighted avg       1.00      1.00      1.00      2000

Accuracy for XGBoost: 0.9995
```

Figure 12. XGBoost classification report

Logistic Regression

The **Logistic Regression** model, after scaling the data, achieved an accuracy of **83.65%**. Though it performed reasonably well, it did not capture complex relationships in the dataset as effectively as the tree-based models. Due to its linear nature, Logistic Regression was less effective in predicting subtle variations in sleep quality and did not emphasize feature importance as prominently.

```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.90      0.89       488
           1       0.70      0.38      0.49       285
           2       0.83      0.92      0.87      1227

    accuracy                           0.84      2000
   macro avg       0.81      0.73      0.75      2000
weighted avg       0.83      0.84      0.82      2000

Accuracy for Logistic Regression: 0.8365
```

Figure 13. Logestic Regression classification report

Support Vector Machine (SVM)

The **Support Vector Machine (SVM)** with an RBF kernel showed an accuracy of **96.25%** when trained on scaled data. It demonstrated better classification capability compared to Logistic Regression, but it still lagged behind Random Forest and XGBoost. Despite its capacity to map data into higher dimensions, SVM required more computational resources and was less efficient for large datasets.

```
Support Vector Machine (SVM) Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.97      0.96       488
           1       0.96      0.92      0.94       285
           2       0.97      0.97      0.97      1227

    accuracy                           0.96      2000
   macro avg       0.96      0.95      0.95      2000
weighted avg       0.96      0.96      0.96      2000

Accuracy for Support Vector Machine (SVM): 0.9625
```

Figure 14. SVM classification report

K-Nearest Neighbors (KNN)

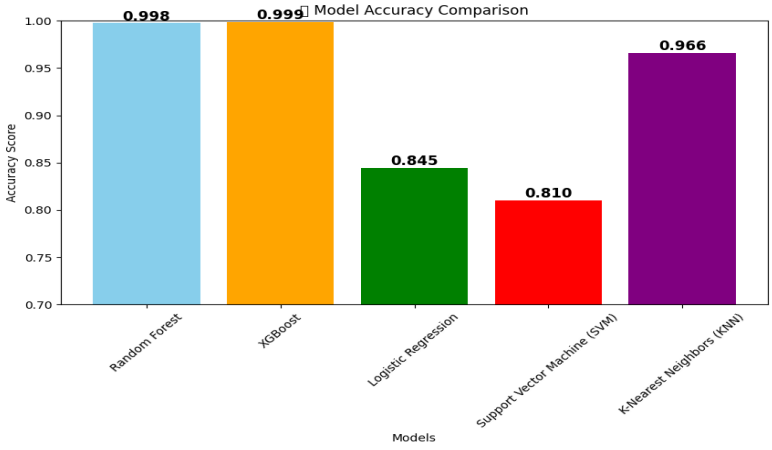
The **K-Nearest Neighbors (KNN)** model achieved an initial accuracy of **96.50%**. The model was evaluated using the default parameter settings, where the value of **K=5** was used. Despite its simplicity, KNN was less effective in capturing complex relationships in the data compared to other models.

```
K-Nearest Neighbors (KNN) Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       488
           1       0.93      0.94      0.94       285
           2       0.97      0.97      0.97      1227

    accuracy                           0.96      2000
   macro avg       0.96      0.96      0.96      2000
weighted avg       0.97      0.96      0.97      2000

Accuracy for K-Nearest Neighbors (KNN): 0.9650
```

Figure 15. KNN classification report

Model Comparison and Insights

The accuracies of the models are plotted in the bar chart.



DISCUSSION OF FINDINGS

Table 1. Model performance on different scenarios

MODELS	STRENGTHS	WEAKNESS
Random Forest	• Handles large datasets and high-dimensional data effectively. • Reduces overfitting by averaging predictions of multiple decision trees. • Provides high accuracy and stability. • Highlights feature importance, aiding in the identification of critical factors.	• Computationally intensive and slow, especially with large datasets. • Complex and less interpretable compared to individual decision trees. • Difficult to explain model decisions to non-technical users.
XGBoost	• Known for speed, scalability, and superior predictive performance. • Handles missing values efficiently. • Incorporates regularization to minimize overfitting. • Ideal for complex classification problems with large datasets.	• Requires careful hyperparameter tuning to prevent overfitting. • Computationally demanding and resource-intensive. • Complex and harder to interpret, limiting transparency.
Logistic Regression	• Simple, easy to implement, and interpretable. • Ideal for binary classification tasks. • Computationally efficient	• Struggles with capturing non-linear relationships in the data. • Performs poorly with complex patterns. • Sensitive to outliers

K-Nearest Neighbors (KNN)	• Intuitive and easy to understand. • Performs well with smaller datasets. • No model training required; adapts quickly to data changes. • Effective for multi-class classification problems.	• Computationally intensive with large datasets, leading to slower predictions. • Highly sensitive to noisy data, irrelevant features, and outliers. • Determining the optimal value of "k" is challenging and can impact accuracy.
Support Vector Machine (SVM)	• Performs well in high-dimensional spaces. • Effectively separates classes using a hyperplane with maximum margin. • Versatile due to various kernel functions for linear and non-linear classification.Suitable for complex datasets with clear class boundaries.	• Computationally intensive, especially with large datasets. • Requires significant resources and time for training. • Sensitive to parameter tuning; choosing the right kernel and hyperparameters is challenging. • Less interpretable, making it difficult to explain decision boundaries to end-users.

CHAPTER VI

CONCLUSION

The Sleep Quality Prediction project explored the relationship between various lifestyle factors and sleep quality among university students, using a dataset that included Sleep Duration, Study Hours, Screen Time, Caffeine Intake, and Physical Activity. The goal was to predict sleep quality using multiple machine learning models, which required thorough data preprocessing, such as handling missing values, encoding categorical variables, and feature engineering, ensuring the data was well-prepared for analysis. Several machine learning models, including Random Forest, XGBoost, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), were evaluated for their predictive performance. XGBoost and Random Forest emerged as the most accurate models, demonstrating the effectiveness of decision-tree-based algorithms in predicting sleep quality. Feature importance analysis showed that Sleep Duration and Screen Time were the most significant factors impacting sleep quality. While the project did not result in a fully functional tool, it laid the foundation for understanding key factors influencing sleep health. The findings highlighted that lifestyle changes, like reducing screen time and maintaining regular sleep schedules, could improve sleep quality. Future work could involve integrating real-time data from wearable devices, incorporating more lifestyle factors, and expanding the dataset to include more diverse demographics. These enhancements would increase the model's accuracy, generalizability, and robustness. User feedback and real-world data integration could further refine the prediction model and help provide personalized recommendations. The project serves as an initial step in creating a comprehensive sleep monitoring system that empowers individuals to make informed decisions about their sleep and overall health.

REFERENCE

[1] Park, K. M., Lee, S. E., Lee, C., Hwang, H. D., & Yoon, D. H. (2024). Predicting sleep based on physical activity, light exposure, and heart rate variability data using wearable devices.

[2] Lemola, S., Perkinson-Floor, N., Brand, S., Dewald-Kaufmann, J. F., & Grob, A. (2015). Adolescents' sleep patterns and psychological functioning

[3] Chanda, H. N. S. (2024). Sleep Quality Prediction from Wearable Device Data.

[4] Di Credico, A., Perpetuini, D., Izzicupo, P., La Malva, P., Gaggi, G., Mammarella, N., & Cardone, D. (2024). Predicting Sleep Quality through Biofeedback: A Machine Learning Approach Using Heart Rate Variability and Skin Temperature.

[5] Li.Mo, (2023). Research on Sleep Health Prediction and Algorithms Based on Big Data.

[6] Curcio, G., Ferrara, M., & De Gennaro, L. (2006). Sleep loss, learning capacity, and academic performance.

[7] Wolfson, A. R., & Carskadon, M. A. (2003). Understanding adolescents' sleep patterns and school performance: a critical appraisal.

[8] Lo, J. C., Ong, J. L., Leong, R. L., Gooley, J. J., & Chee, M. W. (2016). Cognitive performance, sleepiness, and mood in partially sleep-deprived adolescents: The need for sleep study.

[9] Garett, R., Liu, S., & Young, S. D. (2018). The relationship between social media use and sleep quality among undergraduate students.

[10] Galambos, N. L., Vargas Lascano, D. I., Howard, A. L., & Maggs, J. L. (2013). Who sleeps best? Longitudinal patterns and covariates of change in sleep quantity, quality, and timing across four college years.

[11] Hershner, S. D., & Chervin, R. D. (2014). Causes and consequences of sleepiness among college students.

[12] Beattie, L., Kyle, S. D., Espie, C. A., & Biello, S. M. (2015). Social interactions, emotion, and sleep: A systematic review and research agenda.

[13] Xu, Y., Wang, L., & Chen, H. (2021). A Deep Learning-Based Framework for Predicting Sleep Quality from Wearable Sensor Data.