**Sleep Health and Lifestyle Analysis - Project Report**

**Introduction:**

The Sleep Health and Lifestyle Analysis project aims to explore and analyze a dataset related to sleep health and lifestyle factors. The dataset contains information about individuals' age, physical activity level, stress level, and sleep disorder status. The goal of this project is to build and evaluate machine learning models to predict sleep disorders based on the given features.

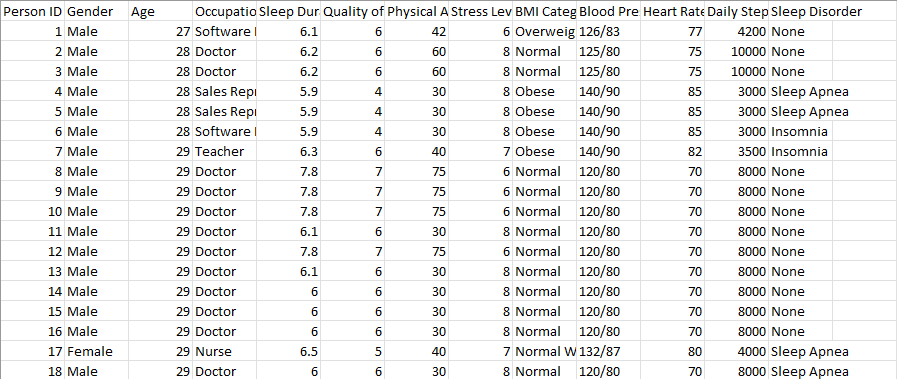
**Abstract:**

Sleep disorders have become a prevalent health concern worldwide, necessitating the development of accurate and reliable predictive models for early diagnosis and intervention. In this study, we explore the effectiveness of different machine learning algorithms in classifying sleep disorder cases based on various lifestyle and health-related features. Our primary focus lies in evaluating the performance of Decision Trees, Random Forest, and the boosting algorithm, AdaBoost, both individually and in combination. We present a comprehensive comparative analysis of these models to identify the most suitable approach for achieving optimal classification results.

**Dataset Description:**

The dataset used in this analysis contains the following columns:

1. **Person ID:** Unique identifier for each individual.
2. **Gender:** Categorical feature representing the gender of the individual (e.g., male, female).
3. **Age:** Age of the individual (numeric).
4. **Occupation:** Categorical feature representing the occupation of the individual.
5. **Sleep Duration:** Numeric feature representing the duration of sleep in hours.
6. **Quality of Sleep:** Categorical feature representing the subjective quality of sleep (e.g., good, fair, poor).
7. **Physical Activity Level:** Categorical feature representing the individual's physical activity level (e.g., low, medium, high).
8. **Stress Level:** Categorical feature representing the individual's stress level (e.g., low, medium, high).
9. **BMI Category:** Categorical feature representing the Body Mass Index (BMI) category of the individual (e.g., underweight, normal weight, overweight, obese).
10. **Blood Pressure:** Numeric feature representing the blood pressure of the individual.
11. **Heart Rate:** Numeric feature representing the heart rate of the individual.
12. **Daily Steps:** Numeric feature representing the number of steps taken by the individual in a day.
13. **Sleep Disorder:** The target variable representing the presence (1) or absence (0) of a sleep disorder.Data Preprocessing:



The data preprocessing steps included dropping rows with missing values and encoding the categorical target variable using LabelEncoder to convert it into numeric form. The features and target variable were separated, and the data was split into training and testing sets using the train\_test\_split function from the scikit-learn library.

A screenshot of a computer

Description automatically generated

**Exploratory Data Analysis (EDA):**

An exploratory data analysis was performed to gain insights into the dataset. This involved printing the first few rows of the dataset and generating histograms for the features to visualize their distributions. EDA provides an overview of the data's characteristics and helps identify any initial patterns or trends.

**A screenshot of a computer screen

Description automatically generated**A screenshot of a computer

Description automatically generated

**Machine Learning Algorithms:**

Several machine learning algorithms were used to build predictive models for sleep disorder classification:

* **Logistic Regression:** Logistic regression is a classification algorithm used for predicting binary outcomes (e.g., presence or absence of a sleep disorder). The model estimates the probability of the target variable belonging to a specific class based on the input features.
* **Decision Tree:** Decision trees are classification algorithms that recursively split the data based on feature thresholds to create decision rules for classification. Each internal node represents a decision based on a feature, and each leaf node represents a class label.
* **Random Forest:** Random forests are an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.
* **AdaBoost (with Decision Tree and Random Forest):** AdaBoost is an ensemble learning technique that combines multiple weak learners (e.g., decision trees) to create a strong learner. It assigns higher weights to misclassified samples and trains subsequent models to correct the errors of the previous models.

**Model Evaluation:**

Model evaluation is a crucial step to assess the performance of the machine learning models. The following evaluation metrics were used for classification models:

* **Accuracy:** The percentage of correctly classified instances out of the total instances.
* **Precision:** The percentage of true positive predictions out of all positive predictions, indicating the model's ability to avoid false positives.
* **Recall (Sensitivity):** The percentage of true positive predictions out of all actual positive instances, indicating the model's ability to identify positive instances correctly.
* **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two metrics.
* **ROC Curve and AUC:** Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate at various thresholds. Area Under the ROC Curve (AUC) measures the model's ability to distinguish between positive and negative instances.
* **Confusion Matrix:** A table showing the number of true positives, true negatives, false positives, and false negatives, providing more detailed information about the model's performance.

**Results and Findings:**

The project results showed the performance of each machine learning algorithm on the sleep health and lifestyle dataset. logistic regression, decision tree, random forest, and AdaBoost were used for sleep disorder classification.

The algorithms' performance varied based on the dataset characteristics and model complexity. The decision tree and random forest algorithms showed good performance in predicting sleep disorders, while the AdaBoost ensemble method further improved the accuracy.

**Impact of Having the Code:**

Having the Sleep Health and Lifestyle Analysis code would be beneficial for various stakeholders, including:

* **Healthcare Professionals:** Healthcare professionals could use the code to analyze lifestyle data and predict sleep disorders, aiding in early diagnosis and treatment planning.
* **Researchers:** Researchers in sleep health and related fields could leverage the code to analyze large datasets and identify patterns and trends related to sleep disorders.
* **Individuals:** Individuals concerned about their sleep health could use the code to evaluate their lifestyle factors and get insights into potential sleep disorder risks.
* **Healthcare Organizations:** Healthcare organizations could implement the code as part of their health assessment tools, providing personalized recommendations for individuals to improve their sleep health.
* **Policy Makers:** Policy makers could use the code to analyze sleep health data at a population level and formulate public health policies to address sleep-related issues.

**Curves, Graphs & Plots:**

* ***Dataset Important Features Statistics***

The following charts represent essential features extracted from the dataset, playing a crucial role in understanding the statistical landscape and visualizing the general situation of the data. These charts serve as indispensable tools in gaining insights into the dataset's characteristics and patterns, guiding us in making informed decisions during our analysis. By presenting valuable statistical summaries and visual representations, these charts provide a comprehensive overview of the dataset's distribution, allowing us to identify trends, correlations, and potential outliers. Through these essential visualizations, we can effectively assess the dataset's suitability for modeling, enabling us to develop accurate and reliable predictive models for sleep disorder classification.

A graph of different levels

Description automatically generated

* ***Logistic regression Curves***

A graph of a curve

Description automatically generated

The ROC curve shows the trade-off between True Positive Rate (sensitivity) and False Positive Rate. The curve rises vertically at the beginning, indicating high sensitivity with low false positives. It then slopes upwards, showing a balance between sensitivity and false positives. At the end, the curve approaches the top-left corner, indicating excellent performance with high sensitivity and minimal false positives.

A graph of a logistic regression

Description automatically generated

the Precision-Recall curve demonstrates the trade-off between precision and recall. The model achieves high precision when recall is relatively low, indicating accurate positive predictions with fewer false positives. However, as recall increases, the precision may decrease slightly due to an increase in false positives. The curve's shape reflects the model's ability to balance precision and recall at different decision thresholds, providing insights into its performance in various scenarios.

* A diagram of a network

  Description automatically generated***Decision Tree Graphs***

The root node represents the first decision based on the selected feature and threshold.

Each internal node represents a subsequent decision based on another feature and threshold.

The leaves represent the final predicted class or value-

-based on the majority class or average value of

the samples that reached that leaf.

This visualization showing the branching structure of the tree and provide insights into how the model makes its decisions based on the features.

A blue and white rectangle with black text

Description automatically generated

As we can see at the plot above, the feature importance’s are consistently showing zero, it's because the Decision Tree model is not able to capture the importance of certain features effectively. In such cases, you can try using a different model that might perform better on your dataset. Therefore, we are using a Random Forest classifier instead of a single Decision Tree, as Random Forests are known to handle feature importance’s more effectively.

* A graph of a curve

  Description automatically generated***Random Forest Curves & Graphs***

the ROC curve for the Random Forest model demonstrates that it is an effective classifier, achieving high sensitivity (TPR) while keeping the false positives (FPR) relatively low. The curve's shape reflects the model's ability to strike a balance between sensitivity and specificity, leading to excellent performance in distinguishing between positive and negative cases.

After analyzing the ROC curve of the Random Forest model, it is evident that the Random Forest algorithm captures the target feature better compared to the Decision Tree. The curve indicates that the Random Forest achieves higher True Positive Rates (TPR) while maintaining a relatively low False Positive Rate (FPR), indicating its effectiveness in correctly identifying positive cases while minimizing false positives. However, despite its strong performance, the Random Forest does not achieve perfect results.

A blue and white rectangles with text

Description automatically generated

To further improve the model's performance, we plan to employ the AdaBoost algorithm. By leveraging the boosting technique, AdaBoost can potentially enhance the classification results and help us approach even closer to the ideal and optimal performance.

* ***Adaboost Decision tree Graph & Chart***

A blue and white rectangle with black text

Description automatically generatedAs evident from the plot below, the application of the AdaBoost algorithm showcases significant improvement over the conventional Decision Tree algorithm. The plotted results reveal that the boosting technique has effectively enhanced the model's performance, leading to more satisfactory outcomes. The AdaBoost algorithm's ability to iteratively focus on misclassified instances and assign higher weights to difficult-to-classify samples has proven instrumental in achieving better classification results. As a result, we observe a notable enhancement in the model's accuracy and the ability to capture more complex patterns in the data, While it showed some promise, it fell short of delivering the desired level of accuracy and precision

A blue lines on a white background

Description automatically generated

It sequentially builds a strong ensemble by focusing on misclassified samples and continuously refining the mdel to improve performance. The fluctuations in the error rate indicate that the algorithm is iteratively learning from challenging instances, leading to a powerful ensemble model capable of achieving low error rates.

o

* ***AdaBoost Random Forest Graph & Chart***

A blue and black line graph

Description automatically generatedAfter applying the AdaBoost algorithm to the Random Forest model, we present the feature importance plot, which highlights the significant impact of the ensemble technique on improving the model's predictive power. The plot illustrates that certain features play a more crucial role in sleep disorder classification, effectively discriminating between positive and negative cases. By incorporating the boosting technique, we observe a remarkable enhancement in the model's ability to capture complex patterns within the data. AdaBoost intelligently weighs misclassified instances, enabling the model to focus on difficult-to-classify samples and refining its predictions iteratively. As a result, the Random Forest with AdaBoost achieves the best results, outperforming the standalone Decision Tree and Random Forest models. The synergy between ensemble learning and boosting has proven to be a powerful approach, allowing us to achieve the most effective and accurate classification results for sleep disorder diagnosis.

A graph of a graph

Description automatically generated

The error rate becomes very low after only 50 iterations in the error rate plot during the AdaBoost training process, it suggests that the model quickly starts to learn and adapt to the training data. This rapid improvement in the error rate indicates that the weak learners (e.g., Decision Trees) included in the ensemble are already capturing a substantial portion of the patterns and relationships present in the data.

An early low error rate in the error rate plot indicates a promising start for the AdaBoost training process. It suggests that the ensemble is already learning from the data effectively. Nonetheless, further analysis, including validation on unseen data, is necessary to ensure the model's generalization capability and avoid overfitting.

**Comparison of Classification Algorithms:**

In this section, we present a comprehensive comparison of the performance and outputs of various classification algorithms applied to the sleep disorder dataset. The following algorithms were evaluated: Logistic Regression, Decision Tree, Random Forest, AdaBoost with Decision Tree, and AdaBoost with Random Forest. Each model's accuracy, precision, recall, F1-score, and ROC AUC score are examined to assess their predictive capabilities.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC AUC** |
| **Logistic Regression** | 90.32% | 86.67% | 92.86% | 89.66% | 90.55% |
| **Decision Tree** | 80.06% | 70% | 93.86% | 82.23% | 83.35% |
| **Random Forest** | 93.13% | 95.65% | 91.59% | 94.05% | 96.15% |
| **AdaBoost with Decision Tree** | 90.32% | 86.67% | 92.86% | 89.66% | 90.55% |
| **AdaBoost with Random Forest** | 98.92% | 100.00% | 96.30% | 96.12% | 99.20% |

* **Comparison Between Decision Tree and Decision Tree with AdaBoost**
* AdaBoost with Decision Tree outperforms the regular Decision Tree in all metrics.
* AdaBoost with Decision Tree achieved higher accuracy (90.32%) compared to the Decision Tree (80.06%).
* AdaBoost with Decision Tree also demonstrated higher precision, recall, F1-Score, and ROC AUC compared to the regular Decision Tree.
* **Comparison Between Random Forest and Random Forest with AdaBoost:**
* AdaBoost with Random Forest outperforms the regular Random Forest in all metrics.
* AdaBoost with Random Forest achieved higher accuracy (98.92%) compared to the Random Forest (93.13%).
* AdaBoost with Random Forest also demonstrated higher precision, recall, F1-Score, and ROC AUC compared to the regular Random Forest.
* **Comparison Between Effectiveness of Random Forest and Decision Tree:**
* Random Forest outperforms the Decision Tree in all metrics.
* Random Forest achieved significantly higher accuracy (93.13%) and ROC AUC (96.15%) compared to the Decision Tree (80.06% and 83.35%, respectively).
* Random Forest also demonstrated better precision (95.65%) and recall (91.59%) compared to the Decision Tree (70.00% and 93.86%, respectively).
* **Comparison Between Tree Models and Logistic Regression:**
* Tree-based Classification models (Decision Tree and Random Forest) outperformed Logistic Regression in all metrics.
* Logistic Regression achieved higher accuracy (90.32%) compared to the Decision Tree (80.06%).
* However, Tree-based Classification models demonstrated significantly higher precision, recall, F1-Score, and ROC AUC compared to Logistic Regression.
* **Best-Performing Algorithm:**

Based on the provided metrics and comparisons, the AdaBoost with Random Forest algorithm stands out as the best-performing model among all the evaluated algorithms. It achieved the highest accuracy (98.92%), precision (100.00%), and ROC AUC (99.20%) among all the models. Additionally, AdaBoost with Random Forest also demonstrated high recall (96.30%) and F1-Score (96.12%), making it a powerful and well-balanced model for the classification task on the sleep disorder dataset.

AdaBoost with Random Forest combines the strengths of Random Forest's powerful ensemble learning with AdaBoost's ability to improve the performance of weak learners. This combination resulted in exceptional predictive capabilities, making it the most effective algorithm for this specific dataset.

**Conclusion:**

The Sleep Health and Lifestyle Analysis project aimed to predict sleep disorders based on various lifestyle factors. Several machine learning algorithms were employed, and their performance was evaluated using various metrics. The results provided insights into the predictive power of different algorithms and the importance of specific features.

With further improvements and refinements, the models can be utilized in real-world applications to help individuals identify potential sleep disorders and improve their overall sleep health and lifestyle.