Analyzing the relationship between sleep quality and daily stress by using machine learning

Guy ben ari 209490473 | Ben shirvi 205478308 | 01/05/2024

1. The subject of the study....1
1.1) The research question....1
1.2) Defining the problem....1
1.3) Defining the solution....1
2. Literature survey....1
2.1) List of articles....1
2.2) Description of the algorithms....2
2.3) Description of research outputs....2
3. Explanation of the input....3
3.1) Description of the data source....3
3.2) Description of the less and more significant data....3
4. Explanation of the output....4
4.1) Score/grouping, explanation....4
4.2) Description of the performance indicators of the results of the algorithms....4
5. Preparation of data....4
5.1) Explanation of feature extraction....4
5.2) PCA....5
5.3) Formatting and cleaning....5
6. Actions....5
6.1) Explanation of the creation of the code....5
6.2) Explanation of what we expect to be in the code and the results....5
6.3) Explanation of preliminary results of the code....6
6.4) Description of the selected algorithms....6
7. Data analysis....6
7.1) The effect of occupation on sleep disorders....6
7.2) Correlation and relations....7
8. Improving the results....7
8.1) Enhancing Data Quality and Quantity....7
8.2) Optimizing Algorithm Performance....7
8.3) Advancing Model Complexity and Robustness....7
9. Summary and conclusions....8
9.1) Discussion and comparison of the results to the articles....8
9.2) Limitations and conclusions....8
9.3) Suggestions for further work....8
10. Links....9
10.1) Link to the Kagle website with the data set....9
10.2) Link to the GitHub project and articles....9

1. The subject of the study

1.1) The research question

- How can machine learning algorithms unveil the intricate connection between sleep habits and daily stress levels?

1.2) Defining the problem

- Understanding stress and sleep is pivotal. Stress, a physiological response to demanding situations, impacts mental and physical well-being. Sleep, essential for cognitive function and emotional regulation, serves as a crucial indicator of overall health. These two domains are deeply intertwined, with disrupted sleep patterns often exacerbating stress levels, and vice versa. The dataset df_SPUML serves as the primary source, containing comprehensive information regarding sleep health and lifestyle parameters. It includes vital metrics such as snoring range, respiration rate, body temperature, limb movement rate, blood oxygen levels, eye movement, hours of sleep, heart rate, and stress levels.

1.3) Defining the solution

- The solution entails leveraging machine learning techniques to uncover the intricate relationship between sleep habits and daily stress levels. Key quality criteria encompass accuracy, cross-validation, and cross-entropy. The prediction task revolves around forecasting next-day stress levels based on an individual's sleep patterns from the preceding night. This predictive modeling aims to provide valuable insights into the dynamic interplay between sleep quality and stress, enabling proactive interventions to enhance overall well-being.

2. Literature survey

2.1) List of articles

- **"Stress Prediction Using Machine Learning and IoT"** - This article explores the application of machine learning techniques in predicting stress levels, leveraging data collected through the Internet of Things (IoT) devices. It delves into how interconnected devices can provide valuable insights into stress patterns, facilitating proactive stress management strategies.

- **"Human Stress Detection Based on Sleeping Habits Using Machine Learning Algorithms"** - This research paper focuses on detecting human stress levels based on sleeping habits utilizing various machine learning algorithms. It investigates how

sleep-related parameters can serve as reliable indicators of stress, paving the way for effective stress monitoring and intervention strategies.

2.2) Description of the algorithms

- **Random Forest (RF):** A versatile ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It's robust against overfitting and works well with high-dimensional datasets.

- **Decision Trees (DT):** These hierarchical tree structures divide the dataset into smaller subsets based on features that lead to the best classification or regression. They're intuitive, easy to interpret, and capable of handling both numerical and categorical data.

- **Support Vector Machine (SVM):** SVM aims to find the optimal hyperplane that separates different classes in the feature space while maximizing the margin between them. It's effective in high-dimensional spaces and suitable for both linear and nonlinear classification tasks.

- **Multilayer Perceptron (MLP):** A type of artificial neural network characterized by multiple layers of nodes (neurons), including input, hidden, and output layers. MLP can learn complex patterns and relationships in data but may require extensive tuning of hyperparameters.

- **K-Nearest Neighbors (KNN):** A non-parametric algorithm that classifies data points based on the majority class of their k-nearest neighbors in the feature space. It's simple to implement and suitable for datasets with well-defined clusters.

- **Logistic Regression (LR):** Despite its name, logistic regression is a linear model for binary classification that estimates the probability of a certain class using a logistic function. It's interpretable and efficient for large datasets.

- **Naïve Bayes (NB):** A probabilistic classifier based on Bayes' theorem with the "naïve" assumption of feature independence. It's simple, fast, and performs well on text classification tasks.

2.3) Description of research outputs

- In the study "Stress Prediction Using Machine Learning and IoT," the research outputs primarily revolve around predictive models capable of forecasting stress levels based on data collected from Internet of Things (IoT) devices. These models not only predict stress but also highlight significant features contributing to stress prediction, offering valuable insights into stress management. The chosen algorithm, Random Forest, proves effective in handling complex and high-dimensional datasets commonly encountered in IoT

applications, ensuring robust stress prediction capabilities.
Similarly, in "Human Stress Detection Based on Sleeping Habits Using Machine Learning Algorithms" by J. G. Jayawickrama and Rahm Rupasingha, the research outputs include predictive models and insights into the relationship between sleep patterns and stress levels. Leveraging machine learning algorithms such as Naïve Bayes and Random Forest, the study accurately predicts stress levels based on sleep-related parameters. These algorithms are selected for their ability to handle diverse features and generate reliable predictions, thereby contributing to effective stress detection and management strategies.

3. Explanation of the input

3.1) Description of the data source

- In this study, two distinct datasets were utilized to explore the relationship between sleep habits and daily stress levels. The primary dataset, df_SPUML (Sleep Health and Lifestyle Dataset), was chosen for its comprehensive collection of relevant variables directly related to sleep patterns and lifestyle factors. This dataset encompasses a wide range of parameters such as sleep duration, quality of sleep, physical activity level, heart rate, and more, making it an ideal choice for investigating the correlation between sleep habits and stress.
  Additionally, the secondary dataset, df_HSDBOSH (Human Stress Detection in and through Sleep Dataset), was selected based on its association with the research article "Human Stress Detection Based on Sleeping Habits Using Machine Learning Algorithms" authored by J. G. Jayawickrama and Rahm Rupasingha. Through communication with the authors via email, it was confirmed that df_HSDBOSH was the dataset utilized in their study. This dataset contains variables pertinent to sleep-related parameters and stress levels, aligning closely with the objectives of this research.

3.2) Description of the less and more significant data

- **More Significant Data**: The significant data for research work includes parameters such as sleep duration, sleep quality, physical activity level, heart rate, and other physiological indicators. These features are crucial as they directly influence an individual's sleep patterns and stress levels, forming the basis for predictive modeling.

- **Less Significant Data**: While all data is valuable, some parameters may have less direct impact on stress prediction. These may include demographic information like gender and age, which, although relevant, might have less weight in determining stress levels compared to factors directly related to sleep quality and physiological measurements. Nonetheless, they still contribute to the overall understanding of the subject matter.

4. Explanation of the output

4.1) Score/grouping, explanation

- The output of this study primarily focuses on predicting stress levels based on sleep habits using machine learning algorithms. The target variable, stress level, is classified into discrete categories or scores representing different levels of stress experienced by individuals. The prediction task involves utilizing features extracted from the datasets to predict the stress level of individuals for the subsequent day. This predictive modeling approach allows for the assessment of how various sleep-related factors contribute to the level of stress experienced by individuals, facilitating early intervention or management strategies.

4.2) Description of the performance indicators of the results of the algorithms

- The performance of the machine learning algorithms is evaluated using several performance indicators, including accuracy, cross-validation (CV), and classification error (CE). Accuracy measures the proportion of correctly predicted stress levels out of the total predictions made by the model. Cross-validation is employed to assess the generalization ability of the models by partitioning the dataset into multiple subsets and training the model on different combinations of these subsets. Classification error quantifies the discrepancy between predicted and actual stress levels, providing insights into the model's predictive accuracy.

5. Preparation of data

5.1) Explanation of feature extraction

- Feature extraction is a crucial step in the data preparation process, where relevant information is extracted from the raw datasets to serve as input variables for the machine learning models. In this study, various sleep-related features are extracted from the datasets, including sleep duration, quality of sleep, physical activity level, heart rate, and other lifestyle factors. These features provide valuable insights into the sleep patterns and habits of individuals, which are hypothesized to influence their daily stress levels. Feature extraction involves processing and transforming the raw data into a structured format suitable for input into machine learning algorithms, ensuring that the selected features adequately capture the underlying relationships between sleep habits and stress levels.

5.2) PCA

- Principal Component Analysis (PCA) is employed as a dimensionality reduction technique to further refine the feature space and enhance the predictive capabilities of the machine learning models. By identifying the principal components that explain the maximum variance in the data, PCA reduces the dimensionality of the feature space while retaining most of the relevant information. This enables the models to focus on the most significant patterns and relationships within the data, leading to improved model performance and interpretability. In this study, PCA is applied to both datasets to select the most informative features and streamline the input variables for the machine learning algorithms.

5.3) Formatting and cleaning

- Before proceeding with feature extraction and PCA, the datasets undergo formatting and cleaning to ensure data integrity and consistency. This involves handling missing values, encoding categorical variables, and addressing any outliers or inconsistencies in the data. By standardizing the data format and removing noise or irrelevant information, the datasets are prepared for further analysis and model training. Overall, data preparation lays the foundation for robust and reliable predictive modeling, enabling meaningful insights to be derived from the datasets.

6. Actions

6.1) Explanation of the creation of the code

- The code development process involves implementing various machine learning algorithms and techniques to address the research question regarding the relationship between sleep habits and daily stress levels. Each step of the code is meticulously crafted to ensure accuracy, efficiency, and reproducibility of results.

6.2) Explanation of what we expect to be in the code and the results

- In the code, we expect to see the performance of various machine learning algorithms assessed based on their ability to accurately predict stress levels from sleep habit data. We anticipate that algorithms like Random Forest, Decision Trees, and Support Vector Machine may perform well due to their capacity to handle complex relationships and high-dimensional data effectively. However, considering the nature of the data and potential limitations such as noise and imbalance, algorithms like Logistic Regression and Naïve Bayes, which are known for their robustness to noise and simplicity, may also yield competitive results. It's crucial to carefully analyze the trade-offs between model complexity and interpretability, considering the practical application of the predictions. Furthermore, we anticipate that feature selection and dimensionality reduction techniques like PCA will play

a role in enhancing model performance by identifying the most informative features and reducing overfitting. Overall, we expect the code to provide valuable insights into the strengths and weaknesses of different machine learning algorithms in predicting stress levels based on sleep habits, guiding the selection of the most suitable models for real-world applications.

6.3) Explanation of preliminary results of the code

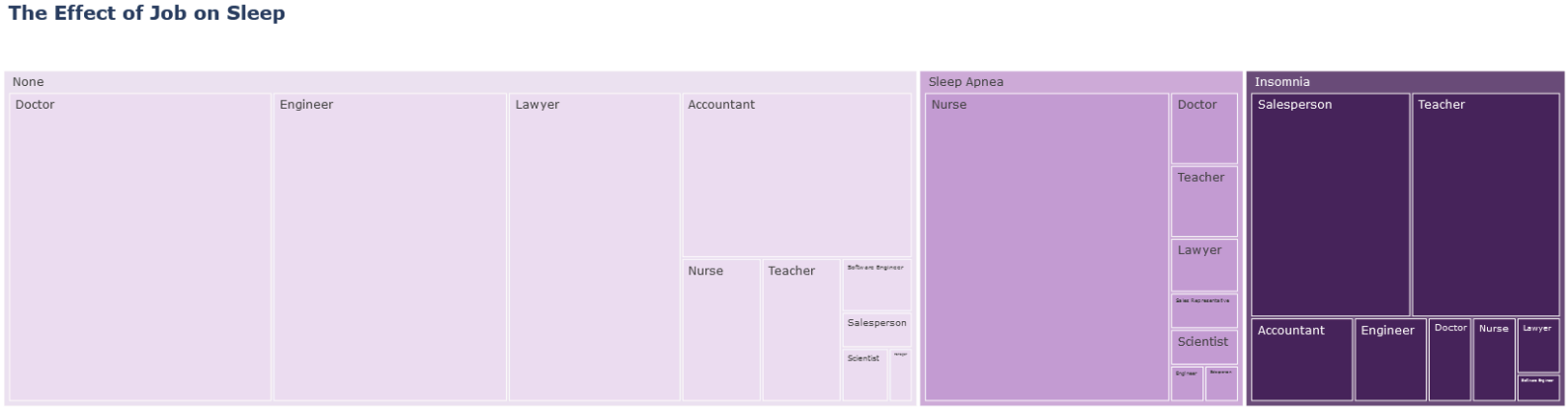
- Preliminary results of the code showcase the performance of the machine learning models in predicting stress levels based on sleep habits and other lifestyle factors. These results provide valuable insights into the predictive capabilities of the models and highlight the significance of different features in determining stress levels. Additionally, the preliminary results help identify areas for further refinement and optimization in the modeling process.

6.4) Description of the selected algorithms

- The selected algorithms for this study include Random Forest, Decision Trees, Support Vector Machine, Multilayer Perceptron, K-Nearest Neighbors, Logistic Regression, and Naïve Bayes. Each algorithm offers unique strengths and capabilities for modeling complex relationships in the data. Random Forest and Decision Trees excel in handling nonlinear relationships and feature interactions, while Support Vector Machine and Multilayer Perceptron are well-suited for high-dimensional data and nonlinear classification tasks. K-Nearest Neighbors is effective for local pattern recognition, Logistic Regression provides probabilistic interpretations of predictions, and Naïve Bayes offers simplicity and scalability for large datasets. By leveraging a diverse set of algorithms, we aim to explore different modeling approaches and identify the most effective strategies for predicting stress levels based on sleep habits.

7. Data analysis

7.1) The effect of occupation on sleep disorders



7.2) Correlation and relations

- Understanding the link between occupation and stress levels is crucial, as occupations shape daily routines and stress exposure. While our project reveals a weak correlation (0.02) between occupation and stress level, it underscores the complexity of stress determinants. Exploring factors like personal habits and environmental influences is vital for a comprehensive understanding. Investigating stress's connection to caffeine consumption and technology usage sheds light on lifestyle factors affecting well-being. Excessive caffeine intake, often tied to stress management, can worsen stress symptoms. Similarly, technology use before bedtime can disrupt sleep patterns, highlighting the need for healthy routines. By examining these relationships, our study aims to provide actionable insights for better stress management and sleep quality.

8. Improving the results

8.1) Enhancing Data Quality and Quantity

- Enhancing the dataset can be achieved through the collection of additional data manually or through augmentation techniques. Manual data collection may involve gathering complementary datasets related to physiological signals or environmental factors, providing richer contextual information for stress prediction. Augmentation techniques, such as synthetic data generation or data perturbation, can also diversify the dataset and improve the models' robustness against variations in input conditions.

8.2) Optimizing Algorithm Performance

- Algorithm refinement is crucial, especially considering the high-dimensional nature of the data and potential feature redundancies. Fine-tuning and optimizing hyperparameters are essential steps in improving algorithm performance. Techniques like grid search or random search can help in systematically exploring the hyperparameter space and identifying the configuration that yields the best results.

8.3) Advancing Model Complexity and Robustness

- Ensemble learning methods, such as bagging, boosting, or even convolutional neural networks (CNNs), present opportunities for improving prediction accuracy and model robustness. Ensemble models combine predictions from multiple base models, effectively mitigating individual biases and variances. CNNs, known for their ability to capture spatial and temporal dependencies, can be particularly effective in capturing complex patterns in the sleep habit dataset. By carefully considering these options and experimenting with different approaches, we aim to enhance the overall performance and reliability of our stress prediction model.

9. Summary and conclusions

9.1) Discussion and comparison of the results to the articles

- Upon analyzing the results obtained from our machine learning experiments, we will engage in a comprehensive discussion and comparison with the findings presented in the selected articles. We will assess the performance of our models in predicting stress levels based on sleep habits and compare them to the results reported in "Stress Prediction Using Machine Learning and IoT" and "Human Stress Detection Based on Sleeping Habits Using Machine Learning Algorithms."
  By evaluating metrics such as accuracy and cross-validation error, we aim to gauge the effectiveness and generalization capabilities of our models relative to those proposed in the literature. Notably, our study demonstrates the robust performance of the Random Forest Classifier (RFC) across various feature sets. The RFC consistently achieves high accuracy, especially when utilizing principal component analysis (PCA) for feature selection. This aligns with the findings in the literature, where ensemble methods like RFC have shown promise in stress prediction tasks. Additionally, we will scrutinize the similarities and differences in the methodologies, feature engineering techniques, and model architectures employed in our study and the referenced articles, shedding light on potential areas of improvement and avenues for future research.

9.2) Limitations and conclusions

- In our study, we will critically evaluate the limitations and constraints encountered during the experimental process, acknowledging factors that may have influenced the reliability and applicability of our findings. These limitations may encompass dataset biases, model assumptions, feature engineering challenges, and inherent complexities of stress prediction from sleep habit data. Through a transparent and objective assessment, we will provide insights into the boundaries of our study and the implications for its real-world implementation. Subsequently, we will draw conclusive remarks on the feasibility and effectiveness of utilizing machine learning algorithms for stress prediction based on sleep habits, highlighting key takeaways and lessons learned from our investigation.
  Moreover, we have chosen to focus solely on the df_SPUML dataset due to its relevance and comprehensiveness in capturing sleep habits and stress levels. While other datasets may offer additional variables or larger sample sizes, they might lack the specificity or granularity required to investigate the intricate relationship between sleep and stress adequately. By prioritizing depth over breadth, we ensure a more targeted analysis, maximizing the insights derived from our study.

9.3) Suggestions for further work

- Beyond the enhancements proposed in improving algorithms and data quality, we advocate for the exploration of cutting-edge technologies such as deep learning architectures,

time-series analysis techniques, and multimodal data fusion methods. Additionally, delving into the influence of contextual factors such as socio-economic status, environmental conditions, and lifestyle choices could provide valuable insights into stress-sleep dynamics. By embracing these future-oriented avenues, we aim to unlock novel insights, develop more robust predictive models, and foster transformative advancements in personalized healthcare solutions.

10. Links

10.1) Link to the Kagle website with the data set

- https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset
- https://www.kaggle.com/datasets/laavanya/human-stress-detection-in-and-through-sleep

10.2) Link to the GitHub project and articles

- https://github.com/guybenari1/Seminar-in-ML