Lifestyle Habits and Psychoemotional Well-Being

A Data Mining Final Project

Faculty: SITE

Prepared by: Smanova Alua, Kabiyev Zhanbek

Assessed by: Adilet Yerkin

Almaty, 2025

1. Introduction

1.1 Background

In everyday life, many people experience a persistently elevated level of stress, reduced subjective well-being, and chronic fatigue. Unlike acute stress caused by isolated events, these conditions often develop gradually and are shaped by regular daily habits, such as sleep patterns, physical activity, dietary behavior, screen usage, and the consumption of stimulants.

Despite the widespread availability of recommendations promoting a “healthy lifestyle,” it remains unclear which everyday behavioral patterns have the strongest association with stress and subjective happiness, and which effects may be secondary or overestimated. This uncertainty complicates both individual self-regulation and the development of evidence-based preventive strategies at the level of educational institutions and workplaces.

Understanding how lifestyle habits are associated with psychoemotional states is therefore a relevant and practically important research problem, particularly in populations exposed to constant cognitive and emotional load, such as students and young professionals.

1.2 Problem Statement

The core problem addressed in this study is the limited understanding of the relationship between daily lifestyle habits and the dynamics of psychoemotional states over time.

Specifically:

- levels of stress, anxiety, and fatigue may fluctuate even in the absence of explicit external stressors;
- subjective well-being and life satisfaction are not always directly aligned with objective conditions such as work or academic demands;
- individual lifestyle factors (sleep, physical activity, screen time, nutrition) are often examined in isolation, without accounting for their combined effects.

Within the scope of this project, the problem is formulated as the task of identifying and analyzing stable associative relationships between everyday behavioral patterns and indicators of stress and subjective well-being using data mining methods.

1.3 Actuality and Practical Relevance

The practical relevance of this study lies in its potential applications across several contexts. For students and early-career professionals, the analysis provides insights into which daily habits are most strongly associated with elevated stress and reduced well-being under conditions of sustained workload. In the context of self-tracking and digital health

monitoring, the results may support more informed lifestyle adjustments based on data-driven evidence.

Additionally, the findings can serve as a foundation for preventive initiatives in educational institutions aimed at reducing chronic stress. For example, if late-night screen usage is consistently associated with lower sleep quality and higher stress levels, a practical recommendation may involve limiting screen exposure after 22:00 as part of stress-prevention programs.

1.4 Research Questions and Hypotheses

The main objective of this study is to explore and analyze associative relationships between daily lifestyle habits and psychoemotional well-being using data mining techniques applied to self-reported survey data.

The study focuses on identifying which everyday behaviors are most strongly associated with stress, anxiety, fatigue, and subjective well-being, as well as assessing whether these factors jointly form stable and interpretable patterns.

Specifically, the study aims to answer the following research questions:

- Which lifestyle factors (sleep, screen usage, physical activity, nutrition, stimulant consumption) are most strongly associated with stress, anxiety, and fatigue?
- How are sleep-related behaviors, including sleep quality and night awakenings, linked to subjective well-being and life satisfaction?
- Do lifestyle and psychoemotional variables jointly provide sufficient signal to distinguish individuals with high stress levels from others?
- Are psychoemotional indicators such as stress, anxiety, and fatigue stable across different temporal self-reports (today, last week, last month)?

Based on prior research and the conceptual framework of the study, the following hypotheses were formulated to guide the analysis:

H1: Lower sleep quality is associated with higher levels of stress, anxiety, and fatigue.

H2: Late-night screen usage is negatively associated with subjective sleep quality.

H3: Regular physical activity is positively associated with subjective well-being and life satisfaction.

H4: Higher levels of anxiety and fatigue increase the probability of experiencing high stress.

H5: Lifestyle habits and psychoemotional indicators jointly provide sufficient information to classify individuals into high-stress and non-high-stress groups using supervised learning methods.

2. Novelty and Originality of the Study

The novelty of this study lies in its integrative and interpretation-oriented analytical approach applied to self-reported lifestyle data. Instead of focusing on isolated variables or single-time measurements, the study introduces composite psychoemotional indicators by aggregating assessments collected over multiple temporal horizons (today, last week, last month). This aggregation strategy reduces random fluctuations inherent in self-reported data and allows the analysis to focus on more stable and meaningful psychoemotional patterns.

Another important aspect of originality is the simultaneous consideration of multiple lifestyle domains within a single analytical framework. While many existing studies examine sleep, physical activity, screen usage, or nutrition separately, this project analyzes their combined associations with stress and subjective well-being. Such a holistic perspective better reflects real-world conditions, where daily habits interact rather than operate independently.

From a methodological standpoint, the study is novel in its use of machine learning models not primarily as predictive tools, but as analytical instruments for structure discovery and hypothesis validation. Supervised learning is employed to assess whether lifestyle and psychoemotional variables jointly provide sufficient signal to distinguish individuals with high stress levels from others. This risk-based framing (“high stress” vs. “non-high stress”) shifts the focus from predicting exact numeric scores to identifying potentially vulnerable groups, which is both more interpretable and more robust for small datasets.

In contrast to purely performance-driven modeling, particular emphasis is placed on interpretability and methodological transparency. The inclusion of a simple baseline model (Logistic Regression) alongside a more flexible model (Random Forest) allows for comparison between linear and non-linear patterns, highlighting that strong results can be achieved even with interpretable models. This reinforces the idea that the detected associations reflect meaningful structure rather than model complexity.

An additional layer of originality is introduced through the planned use of model explanation techniques, specifically SHAP (SHapley Additive exPlanations). SHAP is incorporated not as an auxiliary visualization, but as a conceptual extension of the analytical pipeline, aimed at explaining why certain lifestyle factors contribute to higher stress levels. By enabling both global and individual-level interpretations, SHAP bridges the gap between statistical modeling and practical understanding, allowing for the exploration of “what-if” scenarios and actionable insights.

Overall, the originality of this study does not stem from the use of novel algorithms, but from the coherent integration of data preparation, exploratory analysis, supervised learning, and interpretability within a single, logically structured pipeline. This approach transforms standard data mining techniques into a tool for understanding complex lifestyle–well-being relationships rather than merely predicting outcomes.

3. Related Work

A substantial body of research has established a strong relationship between sleep and psychoemotional functioning. Poor subjective sleep quality and insufficient sleep duration are

consistently associated with elevated perceived stress, increased fatigue, impaired cognitive performance, and reduced emotional regulation. Reviews and conceptual frameworks emphasize that sleep plays a central role in physiological recovery and emotional resilience, particularly in populations exposed to sustained cognitive and emotional demands, such as students and working adults (Åkerstedt et al., 2014; Killgore, 2010). In addition, fragmented sleep characterized by frequent nocturnal awakenings has been linked to heightened physiological arousal and reduced stress tolerance, further reinforcing the importance of sleep continuity for psychological well-being.

Closely related research examines the impact of digital media use on sleep and mental health. Numerous studies report that late-night screen exposure is associated with delayed sleep onset, reduced sleep duration, and poorer subjective sleep quality. These effects are commonly attributed to increased cognitive stimulation, emotional engagement, and exposure to artificial light, which interferes with circadian regulation and melatonin secretion (Cain & Gradisar, 2010; Hale & Guan, 2015). As a result, excessive evening screen use is frequently associated with next-day fatigue, reduced alertness, and lower psychological well-being.

Beyond sleep and screen-related behaviors, lifestyle factors such as physical activity, diet, and stimulant consumption have also been widely studied in relation to stress and subjective well-being. Regular physical activity is consistently linked to improved mood, lower stress levels, and enhanced emotional resilience in non-clinical populations (Rebar et al., 2015; World Health Organization, 2022). Mechanistic studies further suggest that physical activity influences neurobiological pathways related to stress regulation and emotional functioning (Kandola et al., 2019). In contrast, poorer diet quality and irregular eating patterns have been associated with lower life satisfaction and a higher prevalence of common mental health symptoms (Jacka et al., 2014).

Stimulant consumption, particularly caffeine intake later in the day, represents another important behavioral factor. Experimental and epidemiological studies demonstrate that caffeine consumed in the evening can significantly disrupt sleep architecture, delay sleep onset, and increase next-day fatigue and irritability (Drake et al., 2013; Clark & Landolt, 2017). Through its indirect effects on sleep quality, caffeine use may therefore contribute to elevated stress and reduced emotional well-being.

Importantly, prior research also highlights several methodological challenges that are directly relevant to the present study. Psychoemotional constructs such as stress, anxiety, fatigue, and life satisfaction are most commonly measured using self-reported Likert-type scales. While such instruments are practical and scalable, they are sensitive to recall bias, transient mood states, and individual differences in scale interpretation (Podsakoff et al., 2003). Single-time assessments may therefore overrepresent short-term fluctuations rather than stable behavioral patterns. To address this limitation, previous studies recommend collecting responses across multiple temporal horizons (e.g., daily, weekly, monthly) or using aggregated indicators to stabilize measurement noise and improve reliability.

In recent years, machine learning methods have increasingly been applied to mental health and well-being data to identify multivariate patterns that may not be detectable through traditional univariate or regression-based analyses. However, in health-adjacent domains, predictive accuracy alone is insufficient, as interpretability and transparency are essential for trust and practical applicability. Consequently, explainable machine learning approaches—such as SHAP (Shapley Additive Explanations)—have gained prominence as

theoretically grounded tools for attributing model predictions to individual features in a human-interpretable manner (Lundberg & Lee, 2017).

Despite these advances, a notable gap remains in the literature. Many existing studies focus on isolated lifestyle factors, analyze multiple behaviors without a clearly structured variable framework, or report associations without evaluating whether a stable multivariate signal can reliably distinguish individuals with elevated stress levels. The present study addresses this gap by analyzing multiple lifestyle dimensions simultaneously within a unified dataset, applying transparent variable mapping, and combining exploratory data analysis with supervised learning as a structured test of signal strength. Importantly, the objective is not to infer causality, but to identify stable associative patterns that can inform future longitudinal research and data-driven prevention hypotheses.

4. Methods

This section describes the methodological framework of the study, including data scope, preprocessing, exploratory analysis, and the supervised learning setup. The goal of the methods is not only to obtain predictive performance, but to ensure transparency, interpretability, and reproducibility of the analytical process.

4.1 Scope and Limitations

This study is subject to several methodological constraints.
First, all data used in the analysis are **self-reported**, which introduces the potential for recall bias, subjective distortion, and short-term mood effects. Respondents may overestimate or underestimate both lifestyle habits and psychoemotional states.

Second, the survey design is **cross-sectional** with retrospective elements (self-evaluations for today, the past week, and the past month). As a result, the study does not allow for causal inference or temporal modeling of behavioral changes over time.

Third, the analysis is conducted at the **group level** and does not aim to provide individual clinical assessments, diagnoses, or medical recommendations. External life events such as examinations, deadlines, personal crises, or health conditions are not explicitly captured and remain outside the scope of the project.

Accordingly, all findings are interpreted as **associative patterns**, not as evidence of cause-and-effect relationships.

4.2 Data Overview

The dataset used in this study consists of **105 valid survey responses**. The survey captures information across multiple domains relevant to lifestyle behavior and psychoemotional well-being, including:

- socio-demographic characteristics;
- sleep habits and daytime rest;
- physical activity and dietary behavior;
- screen usage and stimulant consumption;
- subjective assessments of productivity, stress, anxiety, fatigue, and life satisfaction;
- temporal self-evaluations of psychoemotional states (today, last week, last month).

The diversity of variables enables both exploratory analysis and the application of supervised learning methods. The dataset size is sufficient to demonstrate methodological principles and structured pattern discovery, while requiring careful model evaluation due to limited sample size.

4.3 Data Mapping and Variable Structure

To ensure clarity, consistency, and interpretability, all variables were systematically renamed and grouped according to their semantic meaning. Each variable was assigned a standardized name, data type, and category.

Variables were organized into three main groups:

1. **Demographic features** (e.g., age group, gender, status);
2. **Lifestyle-related features**, including:
    - sleep and rest behavior,
    - physical activity,
    - dietary patterns,
    - screen usage,
    - stimulant consumption;
3. **Outcome variables** reflecting psychoemotional state and subjective well-being, such as stress, anxiety, fatigue, productivity, and life satisfaction.

This structured mapping facilitated downstream analysis, reduced ambiguity in feature interpretation, and supported transparent model explanation.

4.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the structure, variability, and internal consistency of the data before applying machine learning models.

The EDA focused on the following objectives:

- examining distributions of key psychoemotional variables (stress, anxiety, fatigue, life satisfaction);
- analyzing lifestyle behavior patterns (sleep quality, screen time, physical activity);
- assessing correlations between lifestyle features and psychoemotional outcomes;
- evaluating temporal consistency of self-reported states across different time horizons.

Particular attention was given to the relationship between sleep characteristics and stress levels. *Figure 1* illustrates the association between sleep duration and the composite stress index, which aggregates self-reported stress levels across multiple time periods. The scatter plot reveals a general tendency toward higher stress levels among individuals reporting shorter sleep durations, while longer sleep durations are more frequently associated with moderate stress levels. At the same time, a substantial spread of values is observed within each sleep category, indicating notable inter-individual variability and suggesting that sleep represents an important, but not exclusive, factor influencing stress.

To further contextualize this relationship, *Figure 2* presents the temporal dynamics of stress, anxiety, and fatigue based on self-reports for today, one week ago, and one month ago. The line plot demonstrates a high degree of temporal consistency across all three psychoemotional indicators, with stress, anxiety, and fatigue following similar patterns over time. This temporal stability supports the reliability of the self-reported measures and suggests that the observed associations between sleep characteristics and stress are not driven solely by short-term fluctuations, but reflect more persistent psychoemotional states.

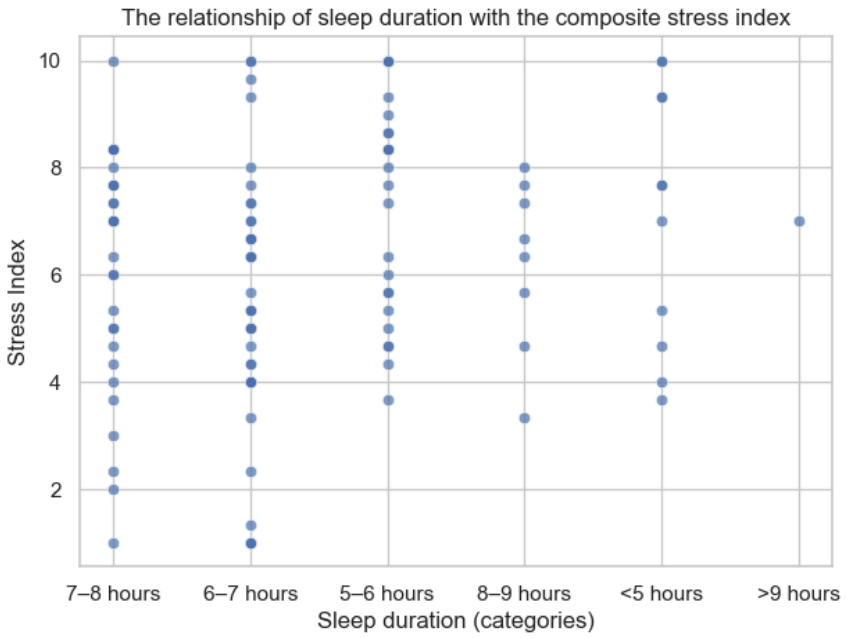


Figure 1. Relationship between sleep duration and the composite stress index

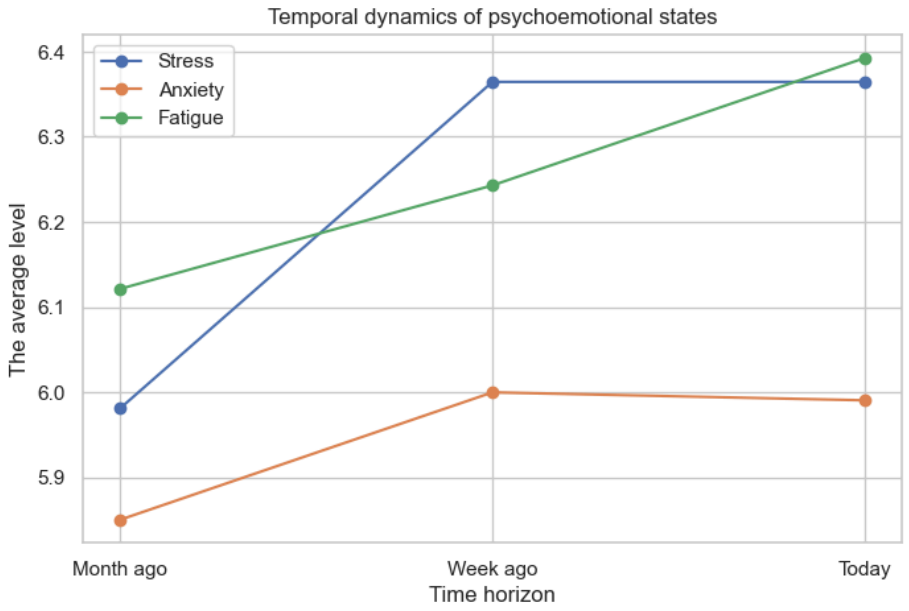


Figure 2. The temporal dynamics of stress, anxiety, and fatigue based on self-reports for today, one week ago, and one month ago

Standard visualization techniques were applied, including histograms, boxplots, scatter plots, line plots, and correlation heatmaps.

At this stage, the analysis was descriptive and diagnostic in nature, serving to validate data quality, detect anomalies, and inform feature selection for modeling. Detailed interpretation of observed patterns is presented in the Results and Discussion sections.

4.5 Supervised Learning Setup

To demonstrate supervised learning, the study formulates a **binary classification task** aimed at identifying respondents with elevated stress levels.

The original stress score (stress_current) is measured on a scale from 1 to 10. This variable was transformed into a binary target:

- **high_stress = 1**, if stress_current ≥ 7
- **high_stress = 0**, if stress_current < 7

The threshold of 7 was selected to represent the upper segment of the scale and to align with a risk-based analytical perspective. This formulation simplifies interpretation and is suitable for classification on small datasets.

The feature set includes lifestyle-related variables (sleep quality, screen behavior, physical activity, diet, stimulant use) and selected psychoemotional indicators (anxiety and fatigue). The original stress variable was excluded from the feature set to prevent data leakage.

4.6 Models and Evaluation Protocol

Two supervised learning models were applied:

- **Logistic Regression**, serving as an interpretable baseline model;
- **Random Forest Classifier**, used to capture potential non-linear relationships.

Categorical variables were transformed using one-hot encoding.
For Logistic Regression, feature scaling was applied using standardization. Random Forest was trained on unscaled features, as tree-based models are not sensitive to feature magnitude.

The dataset was split into training and test sets using a **stratified split** to preserve class proportions. Due to the limited dataset size, class imbalance was addressed using **class weighting** rather than synthetic oversampling methods.

Model performance was evaluated using the following metrics:

- Accuracy;
- Precision;
- Recall;
- F1-score;
- ROC-AUC;
- PR-AUC.

Confusion matrices and ROC curves were used for visual evaluation. The evaluation protocol prioritizes robustness and interpretability over maximal predictive performance.

5. Results

5.1 Exploratory Data Analysis Results

Exploratory data analysis revealed that the sample is characterized by a moderately elevated psychoemotional load. Most respondents reported medium to above-average levels of stress, anxiety, and fatigue, with values distributed across the full scale. This indicates the presence of both relatively resilient individuals and more vulnerable subgroups within the dataset.

Despite elevated levels of stress and fatigue, overall life satisfaction remained moderate to high for a substantial portion of respondents. This suggests that subjective well-being is not

solely determined by current stress levels and may be shaped by a combination of behavioral and contextual factors.

Correlation analysis demonstrated strong positive relationships between stress, anxiety, and fatigue. These indicators were also temporally stable: current, weekly, and monthly self-reports of the same psychoemotional state showed high internal correlations. This temporal consistency supports the reliability of the collected self-reported measures.

Sleep-related variables emerged as central factors in the observed associations. Higher sleep quality was consistently associated with lower levels of stress, anxiety, and fatigue, and positively correlated with life satisfaction and emotional well-being. In contrast, frequent night awakenings showed positive associations with stress and anxiety and negative associations with sleep quality and well-being indicators.

Screen-related behaviors exhibited meaningful patterns. Higher levels of late-night screen usage were associated with lower subjective sleep quality, supporting the assumption that evening screen exposure may interfere with rest and recovery. Physical activity showed a generally positive association with subjective well-being and life satisfaction, although notable variability was observed within activity groups.

Overall, the EDA results indicate that the dataset contains coherent and interpretable patterns, sufficient variability for modeling, and no critical anomalies. These findings justify the application of supervised learning methods to test whether the observed associations form a stable multivariate signal.

5.2 Supervised Learning Results

The supervised learning phase aimed to evaluate whether lifestyle and psychoemotional features jointly allow for reliable classification of individuals with high stress levels.

5.2.1 Classification Performance

Two binary classification models were evaluated: Logistic Regression and Random Forest.

The Logistic Regression model demonstrated strong and balanced performance on the test set. The confusion matrix showed only two misclassifications out of 27 test observations, indicating high overall accuracy. Precision, recall, and F1-score values were all high and well-balanced, suggesting that the model was able to detect high-stress cases without excessive false positives or false negatives.

The Random Forest model achieved perfect classification on the test set, with no observed misclassifications. Correspondingly, all evaluated metrics reached their maximum values. However, this result was interpreted with caution due to the relatively small size of the test sample, as such performance may indicate potential overfitting rather than true generalization.

5.2.2 ROC and Probability-Based Metrics

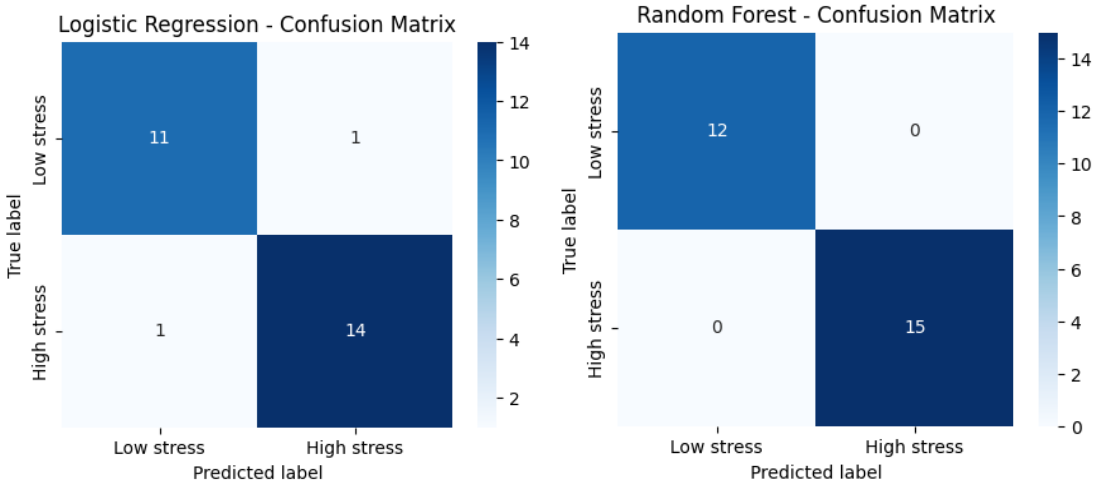
Receiver Operating Characteristic (ROC) analysis further confirmed the strength of the signal in the data. The Logistic Regression model achieved a ROC-AUC value close to 0.97, indicating excellent discrimination between high-stress and non-high-stress individuals. The Random Forest model reached a ROC-AUC of 1.0 on the test set, reflecting perfect separation under the given evaluation conditions.

Precision–Recall (PR) curves showed similarly strong results, particularly for the detection of the high-stress class. This is especially relevant in a risk-oriented framing, where correctly identifying high-stress individuals is more important than overall accuracy alone.

5.2.3 Summary of Supervised Learning Results

The supervised learning results confirm that the selected lifestyle and psychoemotional features contain a strong and structured signal related to stress levels. Even a simple linear model was able to achieve high classification performance, suggesting that the observed relationships are systematic rather than driven by random noise.

To further contextualize these results and provide insight into the nature of classification errors, confusion matrices were constructed for both the logistic regression and random forest models. These visualizations allow for a more granular assessment of model behavior by illustrating the distribution of true positives, true negatives, false positives, and false negatives, thereby complementing the aggregate performance metrics reported above.



At the same time, the results highlight the importance of cautious interpretation. While the models demonstrate strong predictive capability within the given dataset, their primary role in this study is analytical rather than diagnostic. The findings support the use of machine learning as a structured tool for hypothesis validation and pattern discovery, rather than as a standalone predictive system.

6. Discussion

The findings of this study highlight the central role of sleep-related behaviors in the relationship between lifestyle habits and psychoemotional well-being. Across both exploratory analysis and supervised modeling, poor sleep quality and frequent night awakenings consistently emerged as key factors associated with elevated stress, anxiety, and fatigue.

Late-night screen usage appears to be an important behavioral contributor, primarily through its association with reduced sleep quality. This aligns with existing literature emphasizing the disruptive effects of evening screen exposure on circadian rhythms and sleep onset.

Physical activity showed a generally positive association with subjective well-being and life satisfaction, although considerable variability existed within activity levels. This suggests that while physical activity is beneficial on average, its relationship with stress may depend on additional contextual or individual factors.

Importantly, the supervised learning results reinforce the conclusions drawn from EDA. The ability of both Logistic Regression and Random Forest models to accurately classify high-stress individuals confirms that the observed associations form a coherent multivariate pattern rather than isolated correlations. The strong performance of an interpretable baseline model supports the robustness and practical relevance of these relationships.

Machine learning models in this study were not used as deterministic predictors, but rather as analytical tools to validate the presence of structure in the data. The results support a data-driven understanding of how everyday behaviors collectively relate to psychoemotional state, while avoiding claims of causality.

Overall, the study demonstrates that combining exploratory analysis with supervised learning provides a powerful and interpretable framework for analyzing self-reported lifestyle and well-being data. These findings lay the groundwork for future extensions, including explainability analysis and longitudinal research designs.

7. Limitations and Future Work

Despite the structured analytical approach and consistent results, this study is subject to several limitations that should be acknowledged when interpreting the findings.

First, the dataset is based entirely on self-reported survey responses. While self-report instruments are commonly used in well-being research, they are inherently vulnerable to recall bias, subjective perception, and social desirability effects. Respondents may overestimate or underestimate both lifestyle behaviors and psychoemotional states, which introduces additional noise into the data.

Second, the survey design is cross-sectional with retrospective elements (assessments for today, the past week, and the past month). Although aggregated temporal indices were used to improve stability, the data do not support causal inference. The identified relationships should therefore be interpreted strictly as associative rather than causal.

Third, the sample size is relatively small (105 respondents), which limits the statistical generalizability of the results. While the dataset contains sufficient variability for exploratory analysis and supervised modeling, small test sets may lead to optimistic performance estimates, particularly for more flexible models such as Random Forest. As a result, model evaluation metrics should be interpreted with appropriate caution.

Fourth, external contextual factors — such as academic deadlines, examinations, personal life events, or health conditions — were not explicitly captured in the survey. These factors may have influenced stress and well-being levels but remain outside the scope of the current analysis.

Finally, the analysis was conducted at the group level and does not aim to provide individual-level clinical assessment or medical recommendations. The results are intended to support analytical understanding rather than diagnostic decision-making.

7.1 Future Work

Several directions for future research naturally follow from this study. First, collecting a larger and more diverse sample would improve robustness and generalizability. Second, longitudinal data collection would allow the investigation of temporal dynamics and potential causal pathways between lifestyle behaviors and psychoemotional outcomes.

Third, integrating objective measurements — such as wearable device data for sleep and activity — could complement self-reported information and reduce measurement bias. Fourth, extending the analysis with model interpretability techniques (e.g., SHAP) would enable detailed explanation of individual predictions and support “what-if” scenario analysis, improving practical applicability.

Finally, unsupervised learning methods could be further explored to identify typical lifestyle–well-being profiles, which may support personalized prevention strategies and future intervention design.

8. Conclusion

This study examined the relationship between everyday lifestyle habits and psychoemotional well-being using self-reported survey data and data mining techniques. The project addressed the problem of limited understanding of how regular behaviors—such as sleep patterns, screen usage, physical activity, and stimulant consumption—are associated with stress, fatigue, anxiety, and subjective well-being.

By combining exploratory data analysis with supervised learning, the study demonstrated that the collected data contain structured and meaningful patterns rather than random noise. Correlation analysis revealed strong and consistent associations between stress, anxiety, and fatigue, as well as temporal stability of these indicators across different self-report horizons. Sleep-related variables, particularly sleep quality and night awakenings, emerged as central factors linking lifestyle habits with psychoemotional outcomes.

The supervised learning phase further confirmed the presence of a strong multivariate signal in the data. Both Logistic Regression and Random Forest models were able to distinguish

respondents with high stress levels based on lifestyle and psychoemotional features, with high performance across multiple evaluation metrics. The fact that even a simple linear model achieved strong results suggests that the observed relationships are robust and interpretable.

Importantly, the goal of this project was not to provide deterministic predictions or clinical assessments, but to use modeling as a structured analytical tool to validate hypotheses and support understanding. The findings highlight the value of integrating multiple lifestyle dimensions within a single analytical framework and demonstrate how data-driven approaches can complement existing qualitative recommendations related to stress and well-being.

Overall, the study provides a transparent and interpretable perspective on the association between daily habits and psychoemotional states, laying a solid foundation for future research and practical applications in preventive and educational contexts.