DOI: https://doi.org/10.26565/2304-6201-2026-69-01

УДК (UDC) 004.85:004.056

Bazilevych Kseniia *Associate Professor of Department of Mathematical Modeling and Artificial Intelligence; National Aerospace University "Kharkiv Aviation Institute", Vadym Manko St., 17, Kharkiv, Ukraine 61070*
e-mail: k.bazilevych@khai.edu
https://orcid.org/0000-0001-5332-9545

Parfeniuk Yurii *senior lecturer of Department of Theoretical and Applied Informatics; Karazin Kharkiv National University, Svobody Sq. 4, Kharkiv, Ukraine,61022*
e-mail: parfeniuk@karazin.ua
https://orcid.org/0000-0001-5357-1868

Application of Exploratory Data Analysis for Investigating Factors Influencing Sleep Quality

Relevance. The research of the multifactorial nature of sleep quality requires the analysis of large datasets, which is impossible without the use of exploratory data analysis (EDA) methods to identify hidden patterns. In this regard, the development of approaches for the intelligent analysis of factors influencing sleep is a relevant scientific and technical task. **Goal**. To examine and identify the relationships between physiological, behavioral, and environmental factors and sleep quality using exploratory data analysis methods. **Research methods**. The research was based on exploratory data analysis (EDA) methods, primarily aimed at examining the presence of correlations between sleep quality and variables such as sleep duration, stress level, and physical activity. The subsequent construction of a heatmap was necessary to identify latent relationships and to extract the most relevant features. In addition, a linear regression model, a decision tree model, and a logistic regression model were employed to investigate the factors influencing human sleep quality. **The results**. The results obtained using the developed software application with a graphical user interface for analyzing factors influencing human sleep quality are presented. The software application enables data loading, exploratory data analysis, model construction, and result visualization in a user-friendly format. It supports the application of both classification and regression algorithms, allowing it to be adapted to a wide range of analytical tasks. An analysis of the obtained results was conducted, and models with the highest accuracy, adaptability to complex relationships, and interpretability were identified. **Conclusions**. The obtained results confirm the versatility of decision tree methods for the analysis of sleep-related factors. Their accuracy and algorithmic transparency make this approach optimal for modeling complex interrelationships within the scope of the study. Overall, the analysis of factors influencing sleep using EDA methods enables the transformation of complex data into meaningful analytical models, which represents a relevant task for digital medicine.

Keywords: *Sleep, sleep quality, machine learning, regression, classification, logistic regression, decision tree, eda, python, sleep health dataset*

Як цитувати: Bazilevych K., Parfeniuk Y. Application of Exploratory Data Analysis for Investigating Factors Influencing Sleep Quality. *Вісник Харківського національного університету імені В. Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління*. 2025. вип. 69. С.6-19. https://doi.org/10.26565/2304-6201-2026-69-01

How to quote: K. Bazilevych, and Y. Parfeniuk, “Application of Exploratory Data Analysis for Investigating Factors Influencing Sleep Quality”, *Bulletin of V. N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 69, pp. 6-19, 2025. https://doi.org/10.26565/2304-6201-2026-69-01

1. Introduction

Under contemporary conditions of increased psychological and emotional stress, an unstable socio-economic environment, and prolonged exposure to stressors, the issue of sleep quality has become particularly significant. Chronic sleep deprivation and disturbances in sleep architecture adversely affect an individual’s physical, mental, and cognitive functioning, increasing the risk of cardiovascular, endocrine, and mental disorders. Moreover, reduced sleep quality is associated with decreased productivity, impaired attention and memory, as well as elevated levels of anxiety. [1]. Sleep quality is regarded as a complex multifactorial characteristic shaped by the influence of a range of physiological,

psychological, and behavioral factors. These factors include age, body mass index, level of physical activity, stress level, blood pressure, sleep duration, as well as the presence of chronic diseases or sleep disorders such as insomnia or sleep apnea [2]. In countries where societies are exposed to unusual stressful conditions, there is an observed increase in the incidence of insomnia, nocturnal awakenings, difficulties falling asleep, and elevated anxiety levels, which negatively affect the overall health of the population [3]. In Ukraine, according to sociological surveys, nearly half of the population reports a decline in sleep quality since the onset of the full-scale invasion, which is associated with elevated levels of anxiety, forced displacement, and an unstable living environment [4, 5].

Sleep quality is traditionally assessed using questionnaires, clinical instruments, or wearable trackers. However, these methods have certain limitations, ranging from subjectivity to the high cost of equipment, which reduces the accessibility of comprehensive diagnostics for a wide range of users. Consequently, in recent years, there has been growing interest in the application of analytical approaches and machine learning algorithms to study and predict sleep quality based on available physiological and behavioral parameters.

Regression and classification methods allow for the identification of key factors affecting sleep, as well as the construction of predictive models capable of detecting potential disorders or forecasting sleep quality levels [6]. Their application provides flexibility, adaptability to various types of data, and high accuracy, provided that proper approaches to input data processing are employed. Research findings indicate the effectiveness of such models in addressing tasks related both to sleep quality prediction and insomnia diagnosis [7, 8]. The relevance of this research is driven by the need to develop tools that enable the effective analysis of factors influencing sleep using open data and mathematical models. This is particularly valuable in the context of psychoprophylaxis, early risk detection, and the promotion of population mental health.

2. Objective of the study and research tasks

The primary objective of this study is to investigate and identify the relationships between physiological, behavioral, and environmental factors and sleep quality using exploratory data analysis (EDA).

To achieve this objective, the following research tasks are defined:

1. To analyze the characteristics of studying factors influencing human sleep quality.
2. To conduct an analytical review of machine learning methods relevant to achieving the study objectives.
3. To perform exploratory data analysis on the datasets considered in the study.
4. To develop algorithmic models for investigating factors affecting human sleep quality using machine learning techniques.
5. To implement a software application for conducting the study, including visualization of the obtained results.
6. To assess the obtained results.

Object of the Study: The process of investigating factors influencing human sleep quality.

Subject of the Study: Applying Machine Learning Techniques to the Analysis of Factors Affecting Sleep Quality.

The analysis of factors influencing sleep quality requires the completion of the following tasks:

Prediction of the numerical value of a respondent's sleep quality score (ranging from 1 to 10) based on parameters such as age, sleep duration, stress level, and physical activity;

Determination of the presence or absence of a sleep disorder in an individual based on their physiological and behavioral characteristics. Classification is performed by dividing the data into two categories: 'sleep disorder present' and 'sleep disorder absent.

3. Research methods

3.1 Exploratory data analysis

Exploratory data analysis (EDA) is a fundamental stage in the data analytics lifecycle, aimed at thoroughly familiarizing oneself with the available dataset prior to the construction of machine learning models or statistical hypotheses. The concept of EDA was first formulated by the American mathematician and statistician John Tukey in the 1970s. In his seminal work, he emphasized the importance of studying data in their "raw" form to uncover hidden patterns, rather than merely confirming pre-established assumptions [9]. In practice, EDA serves as an intermediate bridge between the data

acquisition stage and data preparation for modeling. Its primary objective is to examine the data structure, identify outliers, missing values, anomalies, potential relationships between variables, and possibly erroneous or incorrect observations. The quality of EDA directly influences both the accuracy of the constructed models and the validity of the analytical decisions made [10].

Within a typical Data Science lifecycle, EDA is usually conducted after the data cleaning stage and prior to the construction of predictive or classification models. Its outcomes can significantly influence the selection of variables, determine the appropriateness of transformations, or reveal new relationships that were not initially apparent. In particular, identifying strong correlations between predictors or visualizing their distributions enables the formulation of valid hypotheses regarding causal relationships [11].

Structurally, EDA encompasses a range of approaches: descriptive statistics (mean, median, variance), graphical visualization (boxplots, histograms, scatter plots), and basic tools for identifying relationships (correlation matrices, cluster analysis). In most modern approaches, EDA results serve not merely as an auxiliary tool but as a full-fledged analytical component that informs the subsequent strategy for model development or decision-making. Thus, EDA is not simply a preparatory stage but a comprehensive analytical practice that enables a researcher to interact with data in an informed manner. Its systematic application helps to avoid critical errors in subsequent stages, improves model quality, and fosters a deeper understanding of the subject domain.

In modern data science, the process of information analysis is implemented as a cycle comprising several sequential stages: data collection, cleaning and preparation, exploratory data analysis, model construction, evaluation of results, implementation, and subsequent monitoring [12]. Within this sequence, the EDA stage functions as a bridge between preliminary data processing and formal modeling, allowing for a deeper understanding of the nature of the data, as well as an assessment of its quality, structure, and statistical patterns.

EDA enables the identification of trends, anomalies, missing values, and multicollinearity, as well as the formulation of hypotheses regarding potential relationships between variables. For this reason, performing EDA is a necessary prerequisite for making informed decisions about the choice of an appropriate machine learning algorithm, the method of data normalization, or the feature engineering strategy [13].

3.2 Methods for studying factors affecting human sleep quality

The investigation of factors influencing human sleep quality is a complex interdisciplinary task that integrates medical, psychological, and analytical aspects. Sleep quality is shaped by a set of interrelated variables: physiological (age, body mass index, heart rate, blood pressure), psychological (stress level, anxiety, depression), and behavioral (physical activity, number of steps, sleep duration and latency, daily routine). In real-world conditions, these variables can exert both independent and combined effects, which significantly complicates the construction of a definitive analytical model.

The complexity of the analysis is further heightened by the high variability of individual characteristics: for example, age and sex may modify the impact of stress on sleep, while physical activity can either improve sleep or worsen it in the case of excessive exertion. Moreover, a significant portion of the variables in sleep quality studies are latent (i.e., not directly observable) and require indirect assessment methods, such as questionnaires, biometric sensors, or psychophysiological testing [2].

Traditionally, scientific practice employs descriptive statistics, correlation analysis, regression, classification, and factor analysis methods to study such complex systems. For example, research on the relationship between BMI and the frequency of nighttime awakenings typically begins with describing mean values within groups and formulating hypotheses regarding their dependence. Subsequently, multivariate analysis methods are applied, allowing the simultaneous consideration of the effects of multiple variables on sleep quality [6].

In contemporary conditions, with the availability of large volumes of data from sleep trackers, questionnaires, and medical devices, machine learning algorithms are increasingly used, allowing for the consideration of nonlinear relationships between parameters. This enables researchers not only to confirm the influence of individual factors but also to develop predictive models of sleep quality for specific population groups. Thus, the study of sleep quality factors requires an integrated approach that combines classical statistical methods with modern data-driven algorithms. This approach allows for the consideration of variable interdependencies, improves diagnostic accuracy, and opens prospects for personalized interventions in the field of sleep medicine.

One of the key directions in the study of factors affecting sleep quality is regression analysis—a classical statistical approach that enables modeling relationships between variables in the form of functional dependencies. In the context of sleep research, this approach makes it possible to predict quantitative characteristics, such as sleep quality scores or sleep duration, based on the influence of predictors including age, stress level, physical activity, body mass index (BMI), and heart rate.

The simplest method is linear regression, which assumes a linear relationship between independent variables and the target variable. In sleep studies, linear regression is often used to assess the extent to which a change in one factor (e.g., the number of daily steps) is associated with an improvement or deterioration in sleep [8]. However, when relationships are more complex or involve interactions among multiple factors, multivariate regression is applied, allowing the simultaneous modeling of the effects of several variables. Furthermore, in the field of behavioral sleep medicine, regression models are used not only for prediction but also to evaluate the importance of individual variables. For example, in the study by Lundgren O., Moneta G. B. (2011) the relationships between depression, anxiety, and physical symptoms (headache, somatic complaints) and subjective sleep quality were analyzed. It was found that depressive symptoms and anxiety were the strongest predictors of poor sleep quality [14]. In the study by Lemma et al. (2012), predictors of poor sleep quality were identified, including stress, anxiety, depression, excessive use of electronic devices, and poor sleep hygiene [15]. The study demonstrates a strong association between psychological state and subjective sleep quality. Such approaches enable a more targeted strategy for addressing factors that contribute to sleep disturbances.

Another popular method is regression trees (decision tree regressors), which offer advantages in interpretability and the ability to account for nonlinear relationships. Unlike linear regression, trees split the data into subgroups based on specific features (e.g., age > 45), allowing for more precise identification of patterns within different respondent groups. The application of regression tree methods in sleep research has practical significance: the resulting models enable the development of automated systems for predicting sleep quality and timely risk detection. When combined with EDA methods, this approach enhances the accuracy and adaptability of solutions in the field of medical technologies.

When the objective of a study is to identify the presence or absence of a sleep disorder, classification methods are employed—a machine learning approach that allows the dataset to be divided into discrete categories. In this case, the target variable is binary or categorical (e.g., 0 - no disorder, 1 - presence of a sleep disorder), while the independent variables consist of physiological, behavioral, or psycho-emotional factors.

One of the fundamental tools of classification is logistic regression—a mathematical model that estimates the probability of belonging to a particular class based on the logistic function. In sleep quality research, logistic regression is used to identify the factors that most strongly influence the likelihood of disorders such as insomnia or sleep apnea [9]. This method also allows for the calculation of the weight of each factor, enabling not only prediction but also the interpretation of the contribution of each feature.

Another popular approach is decision tree classification, which hierarchically splits the dataset based on features, forming rules such as: “if stress level > 6 and sleep duration < 5 hours, then the probability of insomnia is high.” The advantage of this method lies in its interpretability—the researcher can easily trace which factors were decisive in assigning an observation to a particular class. [16].

For tasks with imbalanced classes (for example, when the majority of respondents do not have disorders and only a small portion do), more advanced algorithms such as Random Forest or Support Vector Machines can be applied. These methods improve classification accuracy by simultaneously accounting for multiple factors [17].

Classification models are indispensable in the field of sleep research, as they enable the automated and highly accurate identification of at-risk groups. This is particularly relevant in large population studies or in the development of personalized health monitoring systems.

In contemporary sleep quality research, combining exploratory data analysis (EDA) with the development of predictive or classification models is considered an effective approach. This methodology not only allows for the description of the structure of the available data but also facilitates a deep understanding of the relationships between variables and the identification of hidden patterns, which can subsequently serve as the foundation for machine learning models.

EDA serves as the initial stage—visualizing the distributions of sleep duration, stress level, BMI, heart rate, and other factors allows for the formulation of hypotheses regarding their influence on sleep quality. Using histograms, boxplots, or heatmaps, one can identify, for example, a negative correlation between stress level and sleep quality, or excessive variability in physical activity across different age groups.

Based on such insights, predictive models are developed, including regression models for estimating numerical indicators (e.g., sleep quality scores from 1 to 10) and classification models for determining the likelihood of disorders (insomnia, sleep apnea, etc.). Thus, models are not constructed "blindly" but rely on clearly identified relevant variables revealed through EDA. The integration of EDA and machine learning enhances the interpretability of results and provides a better understanding of causal relationships, which is particularly important in sensitive domains such as sleep research, where an imperfect model may lead to incorrect interpretations of medical risks [18].

Thus, the combination of descriptive, visual, and modeling methods creates an integrated analytical framework that not only enables predictions to be made but also allows them to be interpreted within an applied context.

One of the fundamental tools is linear regression, which models the relationship between predictors (age, sleep duration, stress level) and the sleep quality score. Such a model is easily interpretable, enables the evaluation of each variable's contribution, and helps identify the main factors affecting nighttime rest. At the same time, in cases where nonlinear relationships or a high degree of interdependence among predictors (multicollinearity) are observed, regression trees (Decision Tree Regressors) are preferred. This method constructs the model as a sequence of conditions, allowing for the identification of complex interdependencies between variables without the need for prior transformation.

In the context of sleep research, regression models are often used to estimate the average level of sleep quality based on physiological and behavioral characteristics, such as the frequency of physical activity, the presence of daytime stress, or the average duration of awakenings. Additionally, regression models are employed to assess the effectiveness of corrective interventions—such as changes in sleep hygiene, physical activity, or reduction of stressors. In this way, regression serves not only as an analytical tool but also as a means of monitoring quality of life, particularly within clinical diagnostics or studies of psycho-emotional state.

Another effective method is the decision tree classifier, which constructs a model as a sequence of branching conditions based on predictor values. This structure allows researchers not only to classify subjects but also to trace the logic of decision-making, which is especially valuable in medical or psychological contexts. In more complex cases, where data contain a large number of variables or exhibit high levels of noise, ensemble methods such as Random Forest or Gradient Boosting are applied. These algorithms combine the advantages of multiple decision trees, enhancing result stability and reducing the risk of overfitting. Studies demonstrate that such methods allow for highly accurate classification of sleep disorders even when only a limited number of variables are available [19-22].

4. Research Results

4.1 Initial Data

The study utilized a structured dataset, referred to as the 'Sleep Health and Lifestyle Dataset' [23], which comprises information on individuals' physiological, behavioral, and social characteristics potentially affecting sleep quality. This dataset serves as a well-structured source of input information for the developed software application. The data were stored in a tabular format (.csv), ensuring ease of processing and compatibility with the selected tools in the Python programming language.

The table contains 374 rows, each representing an individual respondent, and twelve variables that are important for modeling purposes. Below is a list of the main columns included in the input data:

1. Person ID - a unique identifier for each study participant. This variable serves an administrative purpose and is not considered in subsequent analysis.
2. Occupation - type of professional employment. A categorical variable reflecting social status and lifestyle.
3. Gender - the respondent's sex (Male or Female). A nominal variable.
4. Age - age of the respondent (quantitative variable).
5. Sleep Duration - number of hours of sleep per day (float).
6. Quality of Sleep - subjective assessment of sleep quality on a scale from 1 to 10. This variable serves as the target in the regression task.
7. Physical Activity Level - level of physical activity, represented as a quantitative value from 0 to 100. Reflects the respondent's overall daily movement, with higher values corresponding to more active individuals. Used as a quantitative predictor in modeling.
8. Stress Level - the respondent's stress level on a scale from 1 to 10 (self-reported).
9. BMI Category - body mass index category (Underweight, Normal, Overweight, Obese).

10. Blood Pressure - arterial blood pressure, presented in the “systolic/diastolic” format, e.g., “120/80.”
11. Heart Rate - heart rate (beats per minute).
12. Daily Steps - average number of steps per day.
13. Sleep Disorder - type of sleep disorder, if present. Possible values: “None,” “Insomnia,” or “Apnea.” This variable serves as the target in the classification task.

For the purpose of further processing, these variables were classified into quantitative, categorical, and target types. Specifically, the variable “Quality of Sleep” serves as the target in the regression task, while “Sleep Disorder” serves as the target in the classification task. The remaining parameters are used as predictors.

During the initial analysis stage, a verification of missing values, variable types, and a visual assessment of proper formatting was conducted. Before proceeding to model construction, each variable was checked for missing or anomalous values, adherence to realistic ranges within the context of the study, and the presence of excessive repetition. Basic visualization methods—histograms, boxplots, and scatter plots—were used to better analyze the distribution of values and to identify potential anomalies. Additionally, the degree of differentiation among variables was assessed to avoid situations in which two features are nearly identical, which could affect model stability. This analysis ensured that the selected factors genuinely influence sleep quality and can be useful for training algorithms.

4.2 Original Dataset

After completing the preliminary data processing, the software automatically generates modeling results, which are presented in an interpretable format—numerical, graphical, or combined. The type of output depends on the chosen task: regression or classification.

When the selected task involves predicting sleep quality, i.e., a numerical indicator representing the respondent’s subjective assessment of sleep, the model produces a point estimate on a scale from 1 to 10. This prediction is based on a combination of input variables, including age, sleep duration, stress level, and physical activity. Such an approach enables analytical forecasting of individual sleep quality prior to direct medical or clinical evaluation. If the user selects a classification task, i.e., determining whether an individual has a sleep disorder, the model analyzes the input features and outputs a categorical result as one of two classes: “sleep disorder present” or “sleep disorder absent.” This approach enables the early identification of potential sleep disorder risks and can be used as an auxiliary tool for preliminary screening.

To evaluate the performance of each model, the software automatically calculates a set of relevant quality metrics. For regression models, MAE (Mean Absolute Error) is used to reflect the average deviation of predicted values from actual values; MSE (Mean Squared Error) assigns greater weight to larger errors; and R^2 (coefficient of determination) indicates the degree of agreement between the model and the actual data. For classification tasks, performance is assessed using metrics such as Accuracy, which measures the overall classification correctness.,

Precision - the accuracy of predicting the positive class; Recall - the completeness of identifying the positive class; F1-score - a combined metric representing the harmonic mean of Precision and Recall.

4.3 Exploratory Data Analysis and Data Preprocessing

The developed application allows EDA to be performed directly through the interface, which implements the construction of four main types of plots. By clicking the “EDA (Data Analysis)” button, the user is presented with graphical visualizations that help form an initial understanding of the data.

The first plot—a boxplot of BMI by sleep disorder type (Figure 1)—illustrates the distribution of body mass index across the groups None, Sleep Apnea, and Insomnia. It can be observed that the mean BMI values differ slightly between groups. For example, respondents with sleep apnea tend to have higher BMI values. The presence of outliers in the plot indicates individual atypical values, which could potentially influence model performance.

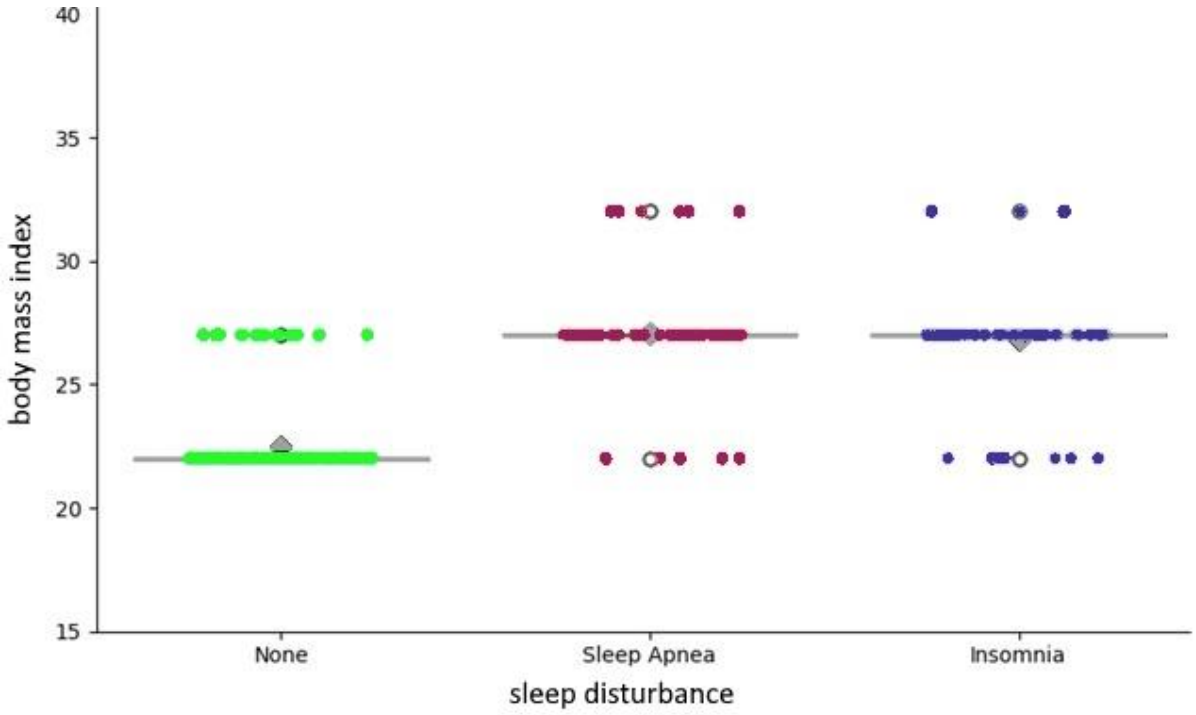

Fig. 1. Boxplot of Body Mass Index (BMI) distribution by sleep disorder type
Рис. 1. Boxplot розподілу індексу маси тіла (BMI) за типами порушень сну

The second plot—a scatterplot showing the relationship between age and sleep duration (Figure 2)—reveals a slight trend: as age increases, the average sleep duration tends to decrease slightly. It is also evident that respondents with different sleep disorders are concentrated in specific ranges. For example, most individuals with insomnia fall within the 30-45-year age range and have sleep durations of less than 7 hours.

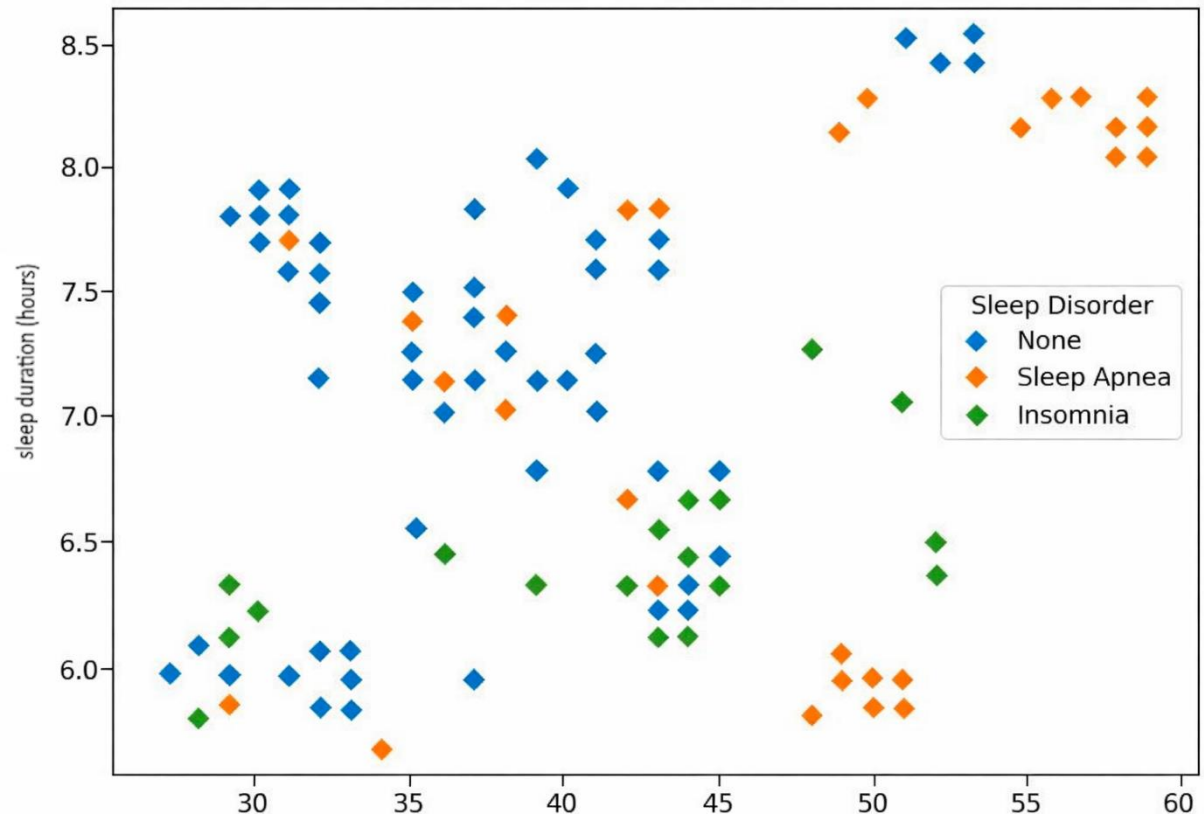

Fig. 2. Scatter plot: relationship between age and sleep duration
Рис. 2. Діаграма розсіювання: взаємозв'язок між віком та тривалістю сну

The third plot—the distribution of sleep disorder types (Figure 3)—illustrates class imbalance. The largest number of respondents belongs to the “None” group, indicating no sleep disorders, whereas the “Sleep Apnea” and “Insomnia” groups are represented by significantly fewer individuals. This distribution should be taken into account when constructing classification models, as the imbalance may lead to bias toward the majority class.

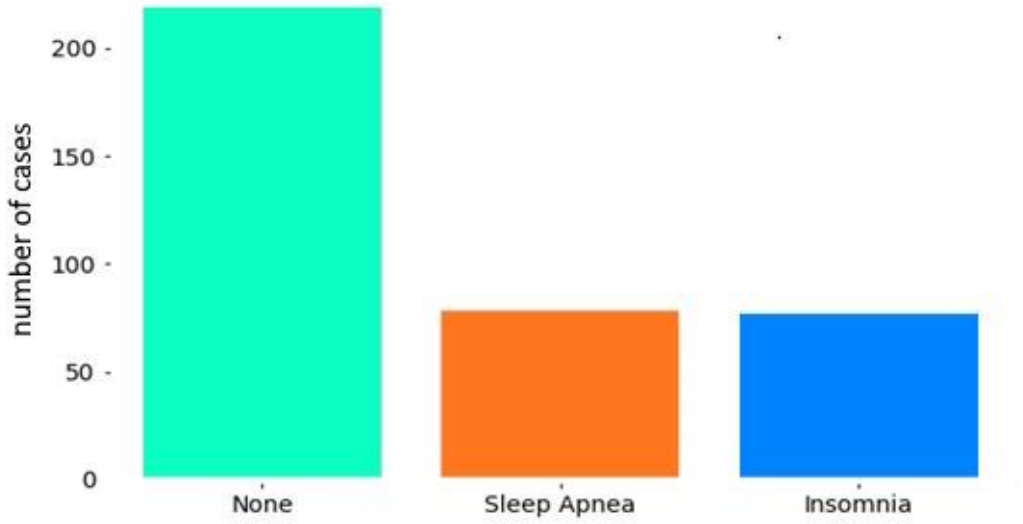

Fig. 3. Histogram of respondents by sleep disorder type
Рис. 3. Гістограма розподілу респондентів за типами порушень сну

The fourth plot - a correlation heatmap (Figure 4) - visually illustrates the relationships between numerical features. The strongest positive correlation is observed between sleep duration and sleep quality (coefficient ≈ 0.88), whereas stress level shows a negative correlation with sleep quality (≈ -0.90). This supports the hypothesis that chronic stress significantly reduces sleep quality. A strong correlation is also observed between daily steps and physical activity level (≈ 0.77), which is consistent with the nature of these indicators. The provided heatmap illustrates how various factors (such as age, sleep, stress, and activity) are interrelated. Red colors indicate that as one variable increases, the other tends to increase as well (positive correlation), while blue colors indicate that as one variable increases, the other tends to decrease (negative correlation). The intensity of the color reflects the strength of the correlation.

Considering the variable “target” (the dependent variable), it shows a moderate positive correlation with Age (coefficient 0.43) and Person ID (0.45). This suggests that, in general, the target value slightly increases with age. The Person ID is almost perfectly correlated with Age (0.99), indicating that older individuals have higher IDs in this dataset. In contrast, Sleep Duration (-0.34) and Sleep Quality (-0.31) exhibit a weak negative correlation with the target, meaning that longer and higher-quality sleep is somewhat associated with lower target values. Heart Rate shows a weak positive correlation with the target (0.33). Meanwhile, Physical Activity Level, Stress Level, and Daily Steps display very weak or nearly nonexistent linear relationships with the target, with coefficients close to zero.

Among other factors, several very strong relationships stand out. The most notable is the extremely strong negative correlation between Sleep Quality and Stress Level (-0.90), indicating that higher stress levels are associated with a marked decrease in sleep quality. Sleep Duration also decreases significantly with increasing stress levels (-0.81). It is logical that Sleep Duration and Sleep Quality are strongly positively correlated (0.88), meaning that longer sleep is generally of higher quality. Additionally, higher Stress Levels are noticeably correlated with higher Heart Rate (0.67), whereas better Sleep Quality and longer Sleep Duration are associated with lower Heart Rate (-0.66 and -0.52, respectively). Physical Activity Level is also strongly correlated with Daily Steps (0.77), which is expected.

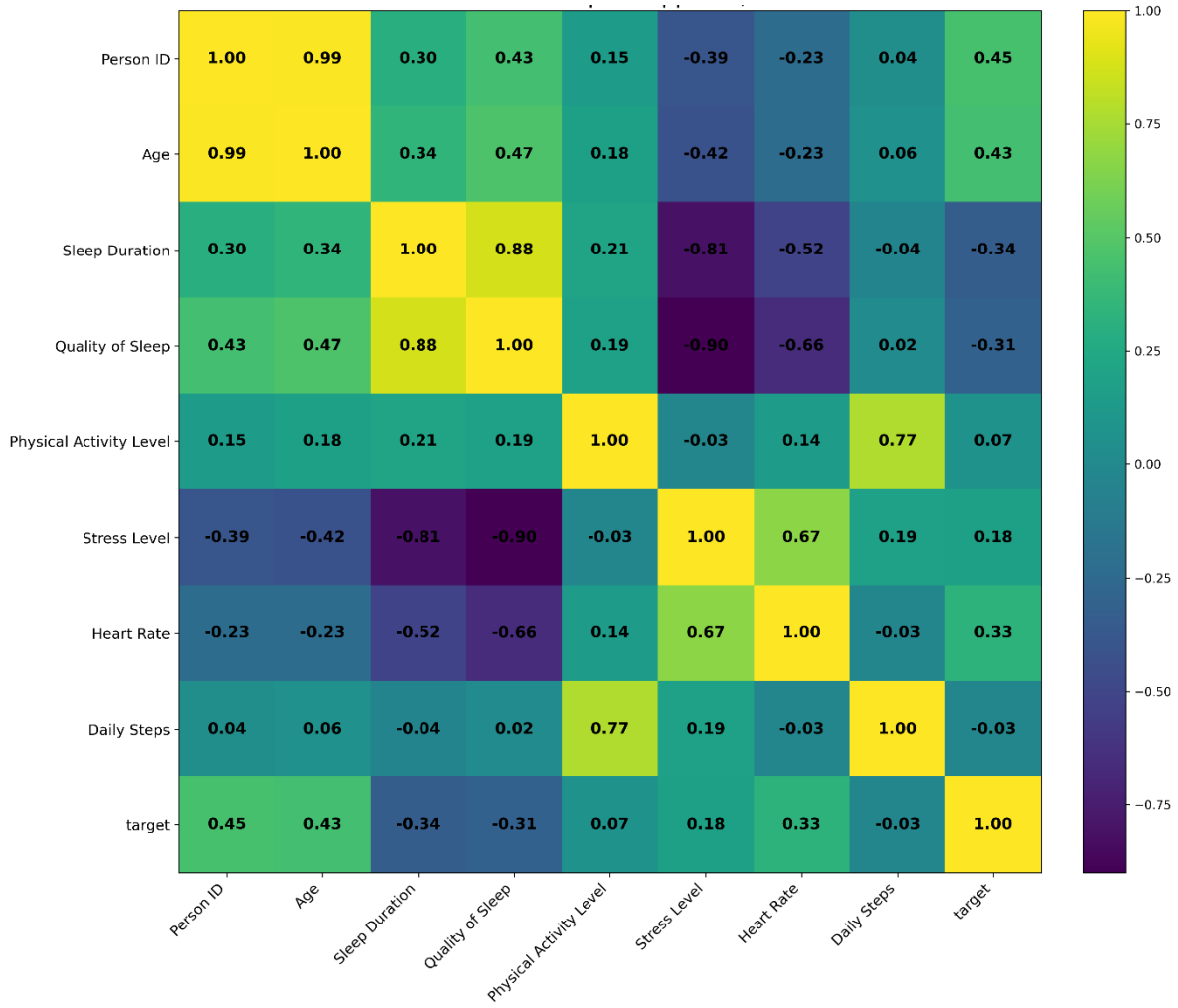

Fig. 4. Теплова карта кореляцій між числовими ознаками
Рис. 4. Теплова карта кореляцій між числовими ознаками

In summary, performing EDA allowed for verification of the dataset quality, identification of key relationships, and formulation of hypotheses regarding significant factors. These findings subsequently serve as the foundation for constructing regression and classification models.

4.4 Analysis of the results obtained

As a result of running the linear regression model, predictions of sleep quality were made based on the variables age, sleep duration, stress level, and physical activity level. The obtained numerical indicators indicate high model quality. The Mean Absolute Error (MAE) is 0.32, reflecting a small average deviation of predicted values from the actual values. The Mean Squared Error (MSE) is 0.18, demonstrating model stability with a low incidence of large errors. The coefficient of determination (R^2) is 0.88, indicating that 88% of the variance in the target variable is explained by the selected predictors.

For a visual representation of the model's accuracy, a scatter plot was constructed (Figure 5), with the true sleep quality values on the X-axis and the predicted values on the Y-axis.

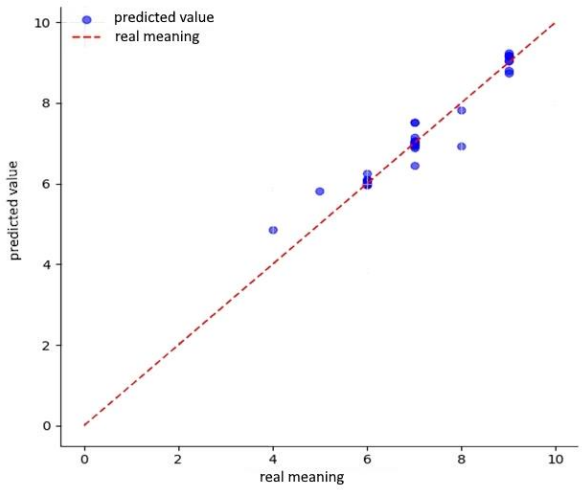

Fig. 5. Scatter plot for the Linear Regression model
Рис. 5. Діаграма розсіювання для моделі Linear Regression

The ideal line is represented by a dashed line and corresponds to a perfect match between predicted and actual values. Analysis of the scatter plot shows that most points are clustered along this line, confirming the adequacy of the model. No visible systematic deviations or outliers are observed.

Thus, the linear regression model demonstrated a high capacity to predict the target variable based on the available parameters, making it a suitable baseline tool for predictive analysis in tasks related to sleep quality assessment.

The next stage involved using a Decision Tree algorithm for the regression task, which allows the construction of an interpretable model based on a hierarchical splitting of the dataset according to feature values. After training the model, the resulting metrics indicate very high prediction accuracy. The MAE is 0.03, reflecting an almost negligible average error. The MSE is also 0.03, indicating minimal deviations in the predictions. The highest performance is demonstrated by the coefficient of determination, R^2 = 0.98, meaning that 98% of the variance in the target variable is explained by the selected features.

Such an R^2 value indicates an almost complete correspondence between predicted and actual values. At the same time, this very high accuracy may suggest a risk of overfitting, especially given the limited size of the training dataset, which necessitates additional validation on an external dataset. It should be noted that the structure of the decision tree allows interpretation of the decision-making process, which can be useful when applying the model in medical or social research.

In the scatter plot (Figure 6), almost all points lie on or very close to the ideal prediction line. Deviations are minimal, so the graph practically demonstrates a match between predicted and actual values. This confirms the high accuracy of the model as well as its potential for further use in sleep quality prediction tasks.

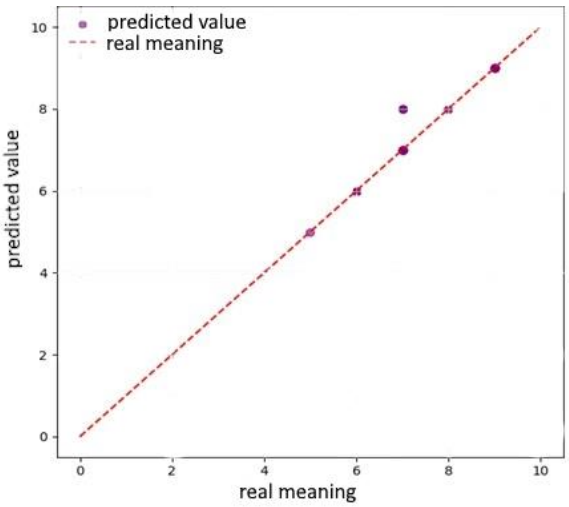

Fig. 6. Scatter plot for the Decision Tree Regressor model
Рис. 6. Діаграма розсіювання для моделі Decision Tree Regressor

Thus, the decision tree model demonstrates high predictive accuracy, making it suitable for tasks where minimizing error, ensuring interpretable decision logic, and adapting to new input data are important. For the classification task, a logistic regression model was used to predict the probability of each respondent belonging to one of two classes: absence of sleep disorders or presence of a sleep disorder. Features included age, stress level, and physical activity. The target variable was formulated as a binary indicator: class “0” - normal sleep, class “1” - presence of a disorder (insomnia or apnea).

The obtained results indicate moderate performance of the model. Accuracy is 0.72, meaning that 72% of predictions were correct. Precision = 0.68 indicates that nearly 7 out of 10 predictions regarding the presence of a sleep disorder were correct. Recall = 0.61 shows that the model identified 61% of all actual sleep disorder cases. F1-score = 0.64 reflects the balance between precision and recall and characterizes the overall classification quality.

Additionally, a histogram of predicted classes was constructed (Figure 7), illustrating the distribution of respondents with and without sleep disorders.

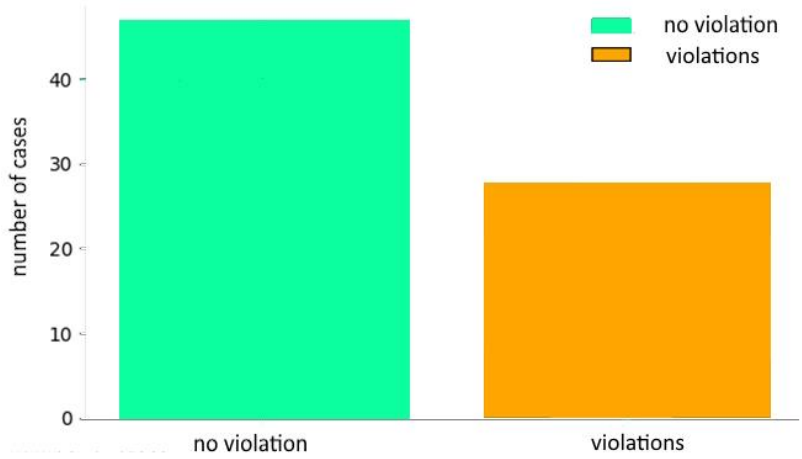

Fig. 7. Predicted class distribution for the Logistic Regression model
Рис. 7. Розподіл передбачених класів для моделі Logistic Regression

The plot shows that the model slightly favors predicting the “no disorder” class, which may be related to the class imbalance in the dataset. However, the overall shape of the distribution corresponds to the actual data structure, indicating that the algorithm behaves adequately.

Thus, the logistic regression model demonstrated satisfactory accuracy and balance in classifying the binary variable. Considering its simplicity, interpretability, and stability, the model can be applied for preliminary screening of sleep disorder risks. Within the binary classification task, a decision tree model was also employed, allowing the creation of a branched classification logic based on the values of the input variables. The input features included age, stress level, and physical activity, while the target variable indicated the presence or absence of a sleep disorder. Unlike logistic regression, the decision tree allows modeling nonlinear relationships and can adapt to more complex data structures.

The obtained numerical metrics indicate a high quality of classification. An Accuracy of 0.89 shows that the model correctly classified 89% of the examples. Precision = 0.87 means that predictions of the “sleep disorder” class were correct in 87% of cases, while Recall = 0.87 indicates that 87% of all true cases of sleep disorders were detected. The F1-score = 0.87 reflects a high level of model balance and its ability to handle moderately imbalanced datasets (Figure 4.16).

The histogram (Figure 7) illustrates the distribution of predicted classes after applying the decision tree model for the classification task. The target variable was “Sleep Disorder”, which was previously encoded in binary format. The task was to determine whether a respondent had a sleep disorder based on features such as age, stress level, and physical activity. The plot allows for assessing the balance of predictions between class 0 (no disorder) and class 1 (disorder present) and confirms that the model does not favor one class over the other. The histogram shows that the classification is relatively uniform, without significant bias, which is particularly important when the positive class is underrepresented.

In conclusion, the decision tree model demonstrated the best performance among all classification approaches, making it suitable for practical applications in detecting sleep disorder risks, considering its accuracy, stability, and interpretability. A comparative summary of all models is presented in Table 1.

Table 1. Comparative Performance of Developed Models
Таблиця 1. Порівняльна характеристика результатів побудованих моделей

Модель	Тип задачі	MAE	MSE	R^2	Accuracy	Precision	Recall	F1-score
Linear Regression	Prediction of Subjective Sleep Quality	0.32	0.18	0.88	-	-	-	-
Decision Tree Regressor	Prediction of Subjective Sleep Quality	0.03	0.03	0.98	-	-	-	-
Logistic Regression	Detection of the Presence or Absence of Sleep Disorders	-	-	-	0.72	0.68	0.61	0.64
Decision Tree Classifier	Detection of the Presence or Absence of Sleep Disorders	-	-	-	0.89	0.87	0.87	0.87

5. Conclusions and prospects for further research

As a result of the practical application of the developed software tool, a complete data analysis cycle was carried out - from preliminary examination of the variables to the construction of machine learning models and evaluation of their performance. Each implemented model demonstrated a different level of accuracy, allowing for a comparative assessment of their advantages in the context of the posed tasks. The results of the exploratory data analysis (EDA) confirmed the presence of strong correlations between sleep quality and variables such as sleep duration, stress level, and physical activity. The construction of a heatmap enabled the identification of hidden dependencies and the determination of the most relevant features for further modeling. Among the regression models, the decision tree yielded the best performance, achieving the lowest error values (MAE = 0.03; MSE = 0.03) and the highest coefficient of determination ($R^2 = 0.98$), indicating the model's strong ability to capture patterns in the data. The linear regression model also produced satisfactory results, though it was outperformed by the decision tree in terms of predictive accuracy.

Among the classification models, the decision tree proved to be the most effective, achieving high values across all key metrics (Accuracy, Precision, Recall, F1-score = 0.87). Logistic regression showed moderate performance, which may be attributed to the linear nature of the algorithm and the potential complexity of the data. In summary, the results indicate that the decision tree algorithm is the most suitable for both regression and classification tasks within this subject area. Its high accuracy, adaptability to complex relationships, and interpretability make it an optimal choice for further applications in analyzing factors affecting sleep quality.

Prospects for further research lie in scaling the developed methodology to significantly larger and more heterogeneous datasets, including those collected in real time via Internet of Things (IoT) sensors and wearable devices. This will enable a transition from identifying general patterns to building high-precision models for personalized monitoring, capable of adapting to individual user characteristics. Thanks to the high interpretability of the decision tree algorithm, the obtained results can serve as a foundation for developing intelligent clinical decision support systems in digital healthcare. In the future, this approach will not only allow for the prediction of sleep disorder risks but also facilitate the automatic generation of scientifically grounded lifestyle recommendations to improve the overall psychophysiological well-being of the population. Further application of this methodology to large-scale datasets will contribute to additional model validation and assessment of its robustness against variability in clinical research samples.

REFERENCES

1. Hobson J. A. Sleep is of the brain, by the brain and for the brain. *Nature*. 2005;437(7063):1254–1256. https://doi.org/10.1038/nature04283.
2. Irish L. A., Kline C. E., Gunn H. E., Hall M. H., Buysse D. J. The role of sleep hygiene in promoting public health: a review of empirical evidence. *Sleep Medicine Reviews*. 2015;22:23–36. https://doi.org/10.1016/j.smrv.2014.10.001.
3. Deng Z., Xie L., Wang Y., Li Y., Huang X., Sun L. Application of logistic regression in diagnosis of OSA severity. *Sleep and Breathing*. 2020;24(4):1379–1387. https://doi.org/10.1007/s11325-020-02029-x.
4. Korost Ya. V., Shkvarok A. K. Assessment of sleep quality of the population of Ukraine during martial law and the risk of cardiovascular complaints associated with clinically expressed insomnia. *Clinical and Preventive Medicine*. 2023;(7):68–73. https://doi.org/10.31612/2616-4868.7.2023.09
5. Ukrinform. Half of Ukrainians have sleep problems. URL: https://www.ukrinform.ua/rubric-society/3958411-polovina-ukrainciv-maut-problemi-zi-snom.html (accessed 12.12.2025).).
6. Carskadon M. A., Dement W. C. Normal human sleep: an overview. In: Kryger M. H., Roth T., Dement W. C., editors. *Principles and Practice of Sleep Medicine*. 6th ed. Philadelphia: Elsevier; 2017. p. 15–24.
7. Freeman D., Sheaves B., Waite F., Harvey A. G. Sleep disturbance and psychiatric disorders: a review of meta-analyses. *The Lancet Psychiatry*. 2017;4(8):684–698. https://doi.org/10.1016/S2215-0366(17)30150-9.
8. Khalid M., Klerman E. B., McHill A. W., Phillips A. J. K., Sano A. SleepNet: attention-enhanced robust sleep prediction using dynamic social networks. *arXiv preprint*. 2024;arXiv:2401.11113. Available from: https://arxiv.org/abs/2401.11113 (accessed 12 Dec 2025).
9. Tukey J. W. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977. 688 p.
10. Kelleher J. D., Tierney B. *Data Science: An Introduction*. Cambridge, MA: MIT Press; 2018. 280 p.
11. Behrens J. T., Yu C. H. Exploratory data analysis. In: Schinka J. A., Velicer W. F., editors. *Handbook of Psychology*. Vol. 2. Thousand Oaks, CA: SAGE; 2003. p. 33–48.
12. Provost F., Fawcett T. *Data Science for Business*. Sebastopol, CA: O'Reilly Media; 2013. 414 p.
13. Dasgupta A. *Practical Data Analysis*. Birmingham: Packt Publishing; 2014. 342 p.
14. Lundgren O., Moneta G. B. Associations of subjective sleep quality with depression, anxiety, and physical symptoms. *Scandinavian Journal of Psychology*. 2011;52(6):544–550. https://doi.org/10.1111/j.1467-9450.2011.00910.x
15. Lemma S., Gelaye B., Berhane Y., Worku A., Williams M. A. Sleep quality and its psychological correlates among university students in Ethiopia: a cross-sectional study. *BMC Psychiatry*. 2012;12:237. https://doi.org/10.1186/1471-244X-12-237
16. Trujillano J., Gil-Sánchez D., Párraga-Martínez I., Flores-Mateo G. Methodological review of classification trees for risk stratification in health research // *Nutrients*. – 2025. – Vol. 17, № 11. – Article 1903. – doi: 10.3390/nu17111903.
17. Rezvani S., Pourpanah F., Lim C. P., Wu Q. M. J. Methods for class-imbalanced learning with support vector machines: a review and an empirical evaluation. *Soft Computing*. 2024;28:11873–11894. https://doi.org/10.1007/s00500-024-09931-5
18. Fu W. Exploratory data analysis and machine learning models for stroke prediction. In: *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*; 2024. p. 211–217. https://doi.org/10.5220/0012783300003885.
19. Permana K. E., Iramina K. Enhancing sleep stage classification with single-channel EEG: feature extraction and Random Forest–XGBoost model. *IEEE Access*. 2025;13:149554–149566. https://doi.org/10.1109/ACCESS.2025.3599828.
20. Gao Q., Wu K. Automatic sleep staging based on power spectral density and random forest. *Journal of Biomedical Engineering*. 2023;40(2):280–285, 294. https://doi.org/10.7507/1001-5515.202207047.

21. Wang Y., Ye S., Xu Z., Chu Y., Zhang J., Yu W. Research on sleep staging based on support vector machine and extreme gradient boosting algorithm. *Nature and Science of Sleep*. 2024;16:1827–1847. https://doi.org/10.2147/NSS.S467111.
22. Xu X., Zhang B., Xu T., Tang J. An effective and interpretable sleep stage classification approach using multi-domain EEG and EOG features. *Bioengineering*. 2025;12(3):286. https://doi.org/10.3390/bioengineering12030286.
23. Sleep Health and Lifestyle Dataset. Kaggle. Available from: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset (accessed 12 Dec 2025).

Базілевич Ксенія Олексіївна *к.т.н., доцент, доцент кафедри математичного моделювання та штучного інтелекту Національного аерокосмічного університету "Харківський авіаційний інститут", вул. Вадима Манька, 17, 61070, Харків*
e-mail: k.bazilevych@khai.edu
https://orcid.org/0000-0001-5332-9545

Парфенюк Юрій Леонідович *PhD, старший викладач, кафедри теоретичної та прикладної інформатики*
Харківський національний університет імені В. Н. Каразіна, майдан Свободи 4, 61022, Харків
e-mail: parfeniuk@karazin.ua
https://orcid.org/0000-0001-5357-1868

Застосування розвідувального аналізу даних для дослідження факторів, що впливають на якість сну

Актуальність. Дослідження багатофакторної природи якості сну потребує аналізу великих масивів даних, що неможливо без застосування методів розвідувального аналізу (EDA) для виявлення прихованих закономірностей. У зв'язку з цим, розробка підходів до інтелектуального дослідження чинників впливу на сон є актуальною науково-технічною задачею **Мета**. Дослідити та виявити взаємозв'язки між наборами фізіологічних, поведінкових та зовнішніх факторів та якістю сну за допомогою exploratory data analysis. **Методи дослідження**. Дослідження базувалось на методах розвідувального аналізу даних (EDA) для того, щоб в першу чергу дослідити наявність кореляцій між якістю сну та такими змінними, як тривалість сну, рівень стресу й фізична активність. Подальша побудова теплової карти необхідна була для виявлення прихованих залежності та отримання найбільш релевантних ознак. Також було використано модель лінійної регресії, модель дерева рішень, модель логістичної регресії для дослідження факторів, що впливають на якість сну людини. **Результати**. Представлено результати, які отримані за допомогою розробленого програмного додатку з графічним інтерфейсом для дослідження факторів, що впливають на якість сну людини. Програмний додаток дозволяє виконувати завантаження даних, проводити розвідувальний аналіз, будувати моделі та виводити результати у зручному форматі, підтримує застосування як класифікаційних, так і регресійних алгоритмів, дозволяючи адаптувати її до різних агналітичних завдань. Було проведено аналіз отриманих результатів та виявлено моделі з найбільшою точністю, адаптивністю до складних зв'язків і пояснюваністю. **Висновки**. Отримані результати підтверджують універсальність методу дерев рішень для аналізу факторів сну. Його точність і прозорість алгоритмів роблять цей підхід оптимальним для моделювання складних взаємозв'язків у межах дослідження. В цілому, аналіз чинників впливу на сон за допомогою методів EDA дозволяє трансформувати складні дані у змістовні аналітичні моделі, що є актуальним завданням для цифрової медицини.

***Ключові слова**: сон, якість сну, машинне навчання, регресія, класифікація, логістична регресія, дерево рішень, eda, python, sleep health dataset.*

Надійшла у першій редакції 11.02.2026; рекомендована до друку 19.03.2026; опублікована 03.04.2026.
The first version has been received on 11.02.2026; was recommended for printing on 19.03.2026; published 03.04.2026.