Journal of Pioneering Medical Sciences
Received: January 06, 2025 | Accepted: February 07, 2025 | Published: July 05, 2025
Volume 14, Issue 06, Pages 45-52

*DOI https://doi.org/10.47310/jpms2025140607*

# Can Lifestyle Habits Predict Happiness? An Exploratory Machine Learning Study Using a Visual Data Mining Platform

**Anas Ali Alhur[1*], Muneef Alsahmmari[2] and Mohammed Ahmed Al-Khattab[3]**
[1-3]Department of Health Informatics, College of Public Health and Health Informatics, University of Hail, 81411 Hail, Saudi Arabia

Author Designation: [1,3]Lecturer, [2]Assistant Professor

*Corresponding author: Anas Ali Alhur (e-mail: anas.ali.alhur@gmail.com).

**Abstract** **Background:** The association between modifiable lifestyle behaviors and subjective well-being has gained increasing attention within health informatics and public health research. This study investigates whether self-reported lifestyle variables-including sleep duration, stress level, screen time and other related behaviors-can be used to predict levels of self-perceived happiness using supervised machine learning techniques. **Methods:** A publicly available dataset comprising responses from 3,000 individuals was analyzed using Orange, a visual, low-code machine learning platform. The outcome variable, happiness score, was discretized into three categories (Low, Moderate, High) to facilitate multi-class classification. Two supervised classifiers-logistic regression and random forest-were trained and evaluated using stratified 5-fold cross-validation. **Results:** Model performance was limited, with both classifiers achieving prediction accuracies of approximately 33% and F1 scores near 0.33. These outcomes suggest minimal discriminative power of the selected lifestyle features for categorizing subjective happiness levels. Exploratory visual analyses supported the absence of strong trends between individual predictors and the outcome variable. **Conclusion:** The results indicate that the selected lifestyle variables, when used in isolation, are insufficient for accurately predicting subjective well-being. Nonetheless, the study demonstrates the utility of visual machine learning platforms such as Orange for educational purposes and exploratory analyses in health informatics. Future research should consider incorporating psychological, social and environmental variables to improve model validity and predictive performance.

**Key Words** Happiness prediction, lifestyle behavior, supervised classification, health informatics, visual machine learning, orange platform, subjective well-being

## INTRODUCTION

Happiness, frequently conceptualized as a subjective state of well-being and overall life satisfaction, has emerged as a critical indicator of population health. Its significance extends beyond philosophical discourse, with empirical research linking happiness to numerous health-related outcomes, including improved physical functioning, reduced morbidity, greater productivity and increased life expectancy [1,2]. Consequently, understanding the determinants of happiness has become an interdisciplinary research priority, spanning psychology, behavioral science, public health and health informatics.

In recent years, the application of Artificial Intelligence (AI) and Machine Learning (ML)-a subset of AI that enables algorithms to learn patterns from data-has offered new methodologies for analyzing complex and high-dimensional datasets. Prior studies have successfully employed ML techniques to predict subjective well-being using a variety of inputs, such as personality traits, socio-economic indicators and digital behavior traces [3-6]. For instance, Hossain et al. demonstrated that deep learning models applied to behavioral data can achieve high predictive performance, reporting macro-averaged F1 scores exceeding 0.70 [5].

While these approaches have yielded valuable insights, most have relied on non-modifiable or externally sourced predictors, including online activity and demographic variables. By contrast, modifiable lifestyle factors-such as sleep duration, physical activity, diet, stress levels and screen time-are not only measurable through self-report but also directly actionable in everyday life. Although prior health studies have established correlations between individual lifestyle behaviors and mental health outcomes [7-9], limited research has examined the cumulative predictive value of these features when modeled using ML techniques.

Furthermore, most machine learning studies in this domain utilize code-based platforms such as Scikit-learn or TensorFlow, which may present accessibility barriers for learners and practitioners without technical programming backgrounds. Orange, an open-source visual ML platform, offers a graphical, drag-and-drop interface that simplifies the application of machine learning workflows. It is particularly suited to educational and prototyping contexts. However, there is a notable lack of empirical research evaluating Orange's practical application in modeling psychological outcomes, particularly within the field of health informatics.

## LITERATURE REVIEW

Machine learning models for predicting happiness have increasingly incorporated multi-source data, including social media behavior, psychological assessments and economic indicators. Studies have shown that such models can achieve substantial accuracy. For example, Zhang *et al*. [6] utilized population-level variables to predict subjective well-being, demonstrating the effectiveness of integrating diverse data inputs. Similarly, Nichols *et al*. [3] found that a combination of lifestyle behaviors and cognitive factors could predict mental health status with reasonable accuracy.

Nonetheless, existing research has shown inconsistent results when lifestyle data are examined in isolation. Some studies suggest modest associations between physical health behaviors and happiness, while others report null or non-significant findings, indicating that lifestyle habits alone may not sufficiently capture the multidimensional nature of subjective well-being [7-9]. This inconsistency underscores the need for additional investigation into the role of lifestyle variables as stand-alone predictors.

Moreover, few studies acknowledge or publish negative findings in this domain, contributing to a potential publication bias that inflates the perceived effectiveness of ML models. Additionally, although visual ML platforms such as Orange have gained popularity in academic training, their application in empirical mental health research remains limited. Little is known about the extent to which these platforms can support rigorous predictive modeling of subjective outcomes like happiness.

### Research Gap and Rationale

Despite the growth of machine learning applications in well-being research, two key gaps persist. First, there is insufficient evidence on the predictive utility of modifiable lifestyle factors-used independently-within machine learning frameworks for happiness classification. These behaviors, while theoretically and clinically relevant, have not been comprehensively examined as a primary feature set. Second, the practical role of visual, low-code ML platforms such as Orange in conducting such analyses is underexplored. While Orange is increasingly used for teaching, its empirical validation as a tool for applied health informatics remains limited.

The current study seeks to address these gaps by examining (1) Whether a selection of self-reported lifestyle behaviors can be used to predict happiness levels using supervised ML classifiers and (2) Whether Orange serves as a feasible and pedagogically valuable platform for developing such models in an accessible manner.

### Study Objectives

This study employs an exploratory cross-sectional research design to evaluate the feasibility of predicting subjective happiness levels using lifestyle-related features and supervised machine learning models developed through the Orange platform.

### Primary Objective

- To assess the predictive performance of two supervised classification algorithms-logistic regression and random forest-in categorizing individuals into Low, Moderate, or High happiness levels, based on self-reported modifiable lifestyle variables

### Secondary Objectives

- To explore associations between individual lifestyle features (e.g., sleep duration, stress level, screen time and social interaction) and categorized happiness levels using descriptive and visual analyses
- To evaluate the usability and educational value of Orange as a visual, low-code platform for implementing supervised machine learning in health informatics education and research

### Prediction Framework

The study adopts a multi-class classification approach, with the happiness score discretized into three ordinal categories using equal-frequency binning to ensure balanced class distribution and reduce training bias during model development.

## METHODS

### Study Design

This study employed a cross-sectional exploratory design to evaluate the feasibility of predicting self-reported happiness levels using supervised Machine Learning (ML) techniques. The investigation also assessed the educational and practical utility of a visual ML platform for health informatics research and instruction. The primary objective was to determine whether modifiable lifestyle behaviors could serve as reliable predictors of subjective well-being.

### Data Source and Ethical Considerations

The analysis was conducted using the publicly available Mental Health and Lifestyle Habits (2019-2024) dataset, retrieved from Kaggle. This dataset includes self-reported responses from 3,000 individuals on a range of lifestyle behaviors and psychological indicators, including a subjective happiness score measured on a 1-10 scale. The dataset was published under a Creative Commons Zero (CC0) license, allowing unrestricted use for research purposes.

As the data were anonymized and publicly accessible, no institutional ethical approval was required for this secondary analysis. Nevertheless, ethical principles were observed by maintaining data confidentiality, respecting the original context of data collection and acknowledging limitations related to the absence of consent documentation. The use of mental health-related responses was approached with caution to avoid overinterpretation or stigmatization.

### Sample Characteristics and Limitations

The dataset did not include demographic variables such as age, gender, socioeconomic status, or geographic location. Consequently, it was not possible to assess the diversity or representativeness of the sample. The absence of such descriptors limits the generalizability of the findings and precludes subgroup analysis. Moreover, since the dataset was collected independently of the present study, the sampling method is unknown and sample selection bias cannot be ruled out.

### Outcome and Predictor Variables

The primary outcome variable was the happiness score, which participants rated on a scale from 1 to 10. To facilitate classification modeling, the continuous variable was transformed into a categorical variable comprising three classes: Low, Moderate and High happiness. This discretization was performed using equal-frequency binning to maintain class balance and reduce training bias. While practical for multi-class classification, this transformation may have reduced the granularity of the happiness measure.

Eight predictor variables were selected for modeling based on their theoretical relevance to subjective well-being and their completeness within the dataset. These included self-reported sleep duration, exercise frequency, diet type, perceived stress level, daily screen time, social interaction score, weekly work hours and mental health condition status. Features with low variability or high missingness were excluded during preprocessing.

### Data Preprocessing and Bias Control

Data preprocessing was performed using a visual machine learning environment. Missing values were handled through listwise deletion, as missingness for the selected variables remained below 5%. Categorical variables were encoded using one-hot encoding and continuous variables were standardized to improve comparability and model stability.

To mitigate model bias and enhance generalizability, stratified 5-fold cross-validation was implemented. This approach preserved the class distribution across all folds and ensured robust performance evaluation. By using stratified sampling, the risk of class imbalance was minimized and overfitting was controlled.

It is important to note that all predictor and outcome data were self-reported, which introduces the potential for recall bias, misclassification and social desirability effects. These limitations are addressed in detail in the discussion section.

### Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to assess the distribution and interrelationships of variables prior to modeling. Scatter plots and box plots were used to visualize associations between lifestyle behaviors and happiness categories. A heat map was also generated to display the distribution of numeric variables across the sample. These visualizations indicated a lack of strong univariate associations, supporting the need for multivariate classification techniques.

### Model Development and Evaluation

Two supervised ML algorithms were selected for this study. Logistic regression was used as a baseline linear model due to its interpretability, while random forest was selected as a non-linear ensemble model capable of capturing complex interactions. Both models were trained and validated using stratified 5-fold cross-validation.

Model performance was evaluated using multiple metrics, including classification accuracy, macro-averaged F1 score, area under the receiver operating characteristic curve (AUC), precision, recall and the Matthews Correlation Coefficient (MCC). These metrics were chosen to provide a comprehensive and balanced assessment, especially given the multi-class structure of the outcome variable.

### Visualization and Interpretation

In addition to quantitative metrics, model outputs were interpreted using visual aids. Bar charts compared performance across classifiers, while scatter and box plots illustrated the distribution of happiness scores across key predictors. A heat map summarized the range and concentration of numeric features. These visuals contributed to the interpretation of findings and supported the conclusion that lifestyle features alone are likely insufficient for reliably predicting happiness.

## RESULTS

### Workflow Summary

The data mining and model development workflow was executed entirely within the Orange platform and is illustrated in Figure 1. The process consisted of five sequential steps: (1) Importing the dataset, (2) Selecting relevant variables, (3) Discretizing the continuous happiness score into three equal-frequency categories (Low, Moderate, High), (4) Applying two supervised machine learning

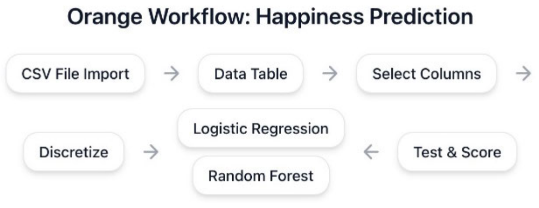

Figure 1: Workflow implemented in Orange for happiness prediction

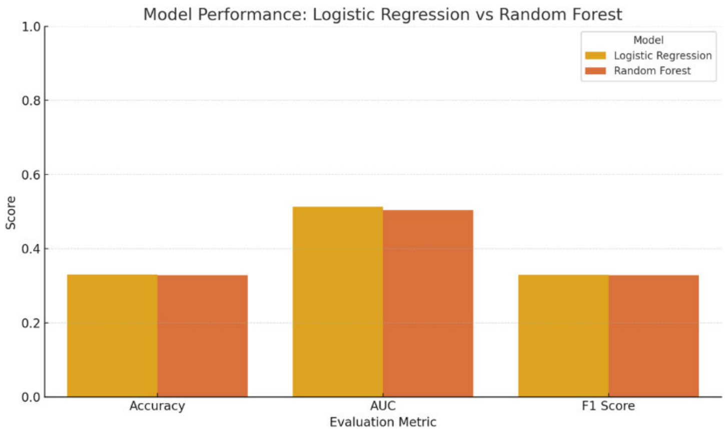

Figure 2: Comparative performance of Logistic Regression and Random Forest classifiers across accuracy, AUC, and F1 score

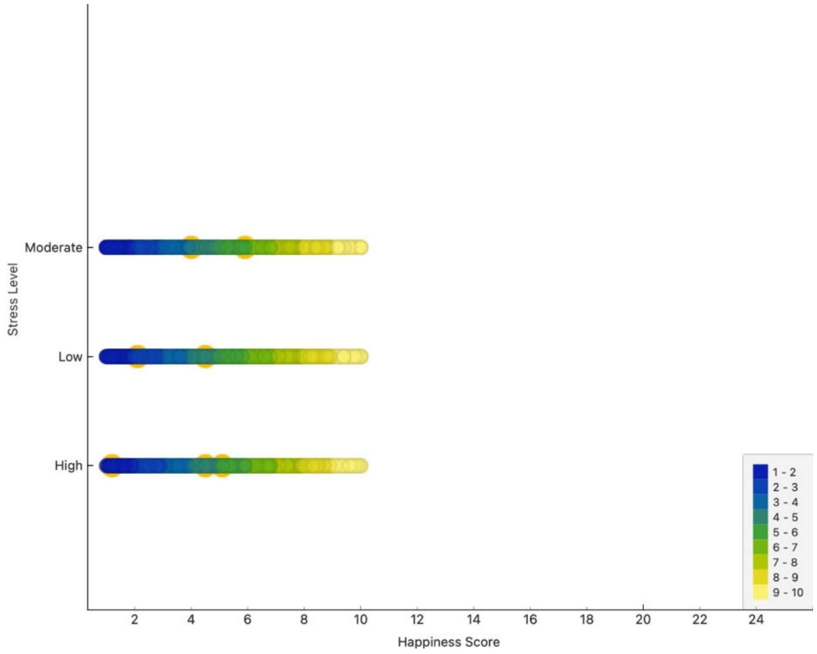

Figure 3: Scatter plot showing distribution of Happiness Score across Stress Level groups

models (logistic regression and random forest) and (5) Evaluating the models using 5-fold stratified cross-validation through the Test & Score widget.

## Model Performance Evaluation

The performance metrics of the two classifiers are summarized in Figure 2 and detailed below. Logistic regression achieved an overall classification accuracy of 33.0%, with an F1 score of 0.329 and an area under the receiver operating characteristic curve (AUC) of 0.513. Similarly, the random forest classifier yielded an accuracy of 32.8%, an F1 score of 0.328 and an AUC of 0.504.

These results indicate that both models performed only marginally better than random classification, which would yield an expected accuracy of approximately 33.3% in a balanced three-class setting. The low F1 scores further reflect poor model calibration and class-level precision, while AUC values close to 0.5 suggest that the models failed to meaningfully differentiate between happiness categories. These findings highlight the limited predictive power of the selected lifestyle variables when used independently within standard ML classifiers.

## Exploratory Visualization of Key Predictors

To explore the explanatory contribution of individual predictors, a series of visualizations were produced. As shown in Figure 3, happiness scores were plotted against self-reported stress levels (High, Moderate, Low). The

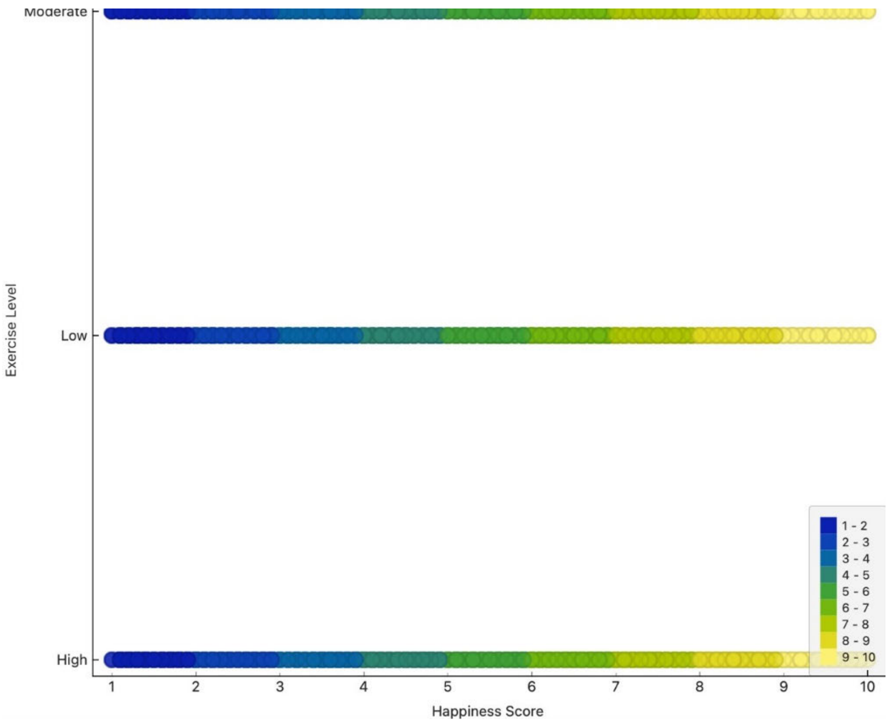

Figure 4: Scatter plot showing distribution of Happiness Score across Exercise Level groups

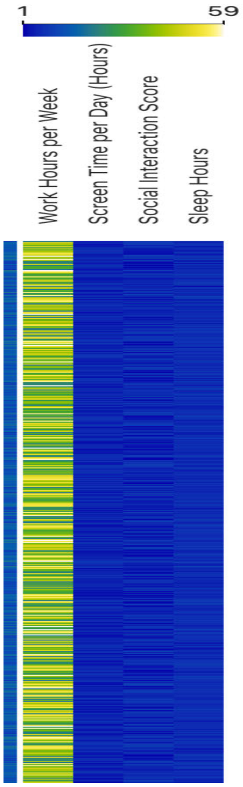

Figure 5: Heat map of numeric lifestyle features across 3,000 participants. Brighter colors indicate higher values

distribution revealed no discernible trend; participants in each stress category exhibited a wide range of happiness scores. This suggests a weak univariate relationship between perceived stress and subjective happiness.

A similar pattern was observed in Figure 4, which visualizes happiness scores across varying levels of exercise frequency. Although some participants in each group reported high happiness, the distribution did not indicate a clear upward trend, again reflecting a lack of linear association.

Figure 5 provides a heat map of four continuous lifestyle variables: work hours per week, daily screen time, social interaction score and sleep duration. While sleep and social interaction values appeared relatively uniform across the sample, work hours and screen time showed wider variability. However, these patterns did not translate into meaningful differences in happiness classifications during modeling.

## Group Comparisons Using Statistical Testing

Chi-square ($\chi^2$) tests were conducted to assess the relationship between categorical lifestyle and mental health variables.

In Figure 6, a box plot compares stress levels across five self-reported mental health conditions. The chi-square analysis produced a non-significant result ($\chi^2 = 5.86$, p = 0.663), suggesting no statistically meaningful association between reported mental health status and perceived stress level.

Figure 7 explores exercise frequency across the same mental health categories. Again, the result was non-significant ($\chi^2 = 5.22$, p = 0.733), indicating no apparent link between mental health status and engagement in physical activity.

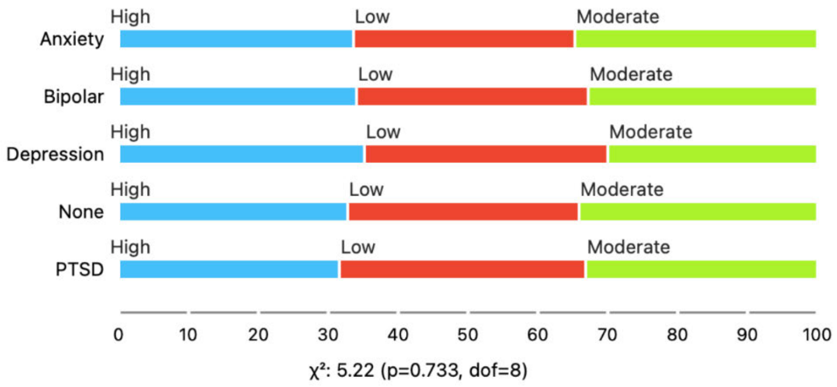

Figure 6: Distribution of stress levels across mental health conditions

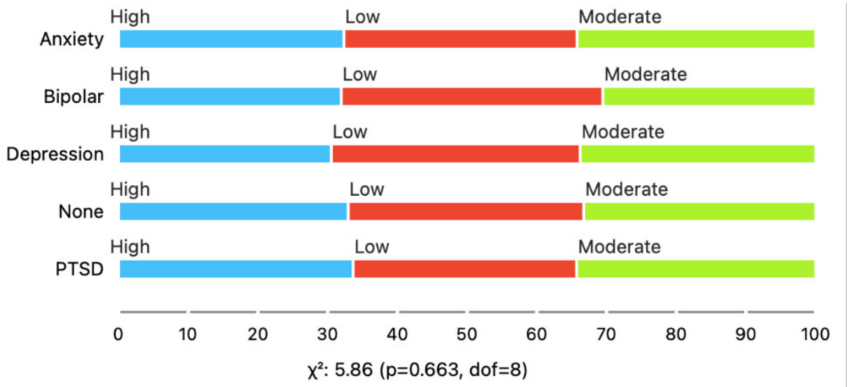

Figure 7: Distribution of exercise levels across mental health conditions

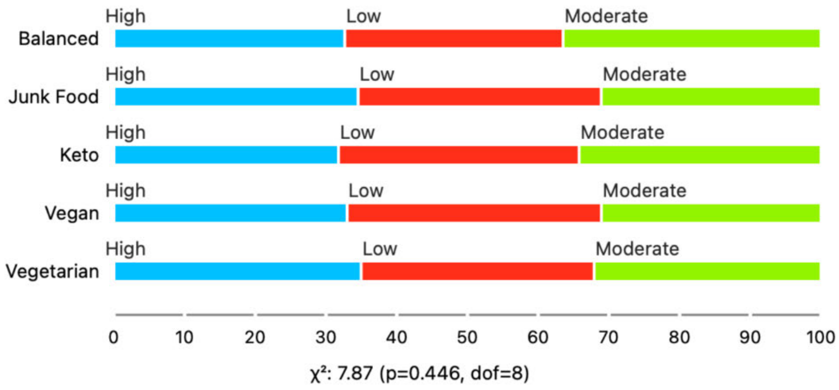

Figure 8: Stress level distributions across different dietary patterns

Lastly, Figure 8 presents stress levels across various diet types (e.g., balanced, junk food, vegetarian, vegan, keto). The chi-square test yielded χ² = 7.87 with a p-value of 0.446, confirming no statistically significant difference in stress levels among dietary groups.

## DISCUSSION

### Interpretation of Findings

This study aimed to determine whether modifiable lifestyle behaviors could be used to classify self-reported happiness levels using supervised machine learning models. The analysis found that both logistic regression and random forest classifiers performed close to the baseline expected from random classification, with accuracies around 33% and F1 scores below 0.33. These findings indicate that, in the context of this dataset, the selected lifestyle features lacked sufficient discriminative power to support reliable prediction of subjective well-being.

While initially surprising, this result aligns with the broader understanding that happiness is a multidimensional construct influenced by a complex interplay of psychological, emotional, social and contextual factors. Variables such as personality traits, affective states, perceived control and interpersonal relationships are known to play significant roles in shaping subjective well-being-none of which were included in the present feature set.

### Comparison with Previous Studies

The poor predictive performance observed contrasts with findings from earlier studies that reported strong results using ML approaches. For instance, Hossain et al. demonstrated that deep learning models trained on multimodal behavioral data could predict well-being with macro F1 scores above 0.70 [1]. Similarly, Saha et al. used socio-economic and digital footprint data to develop robust well-being prediction models [2]. In both cases, the inclusion of psychologically rich or digitally derived features likely contributed to model success.

In contrast, the current study relied exclusively on self-reported lifestyle behaviors-data that, while modifiable and accessible, may not capture the internal emotional and cognitive dimensions essential to modeling happiness. These findings are consistent with literature suggesting that lifestyle factors, while important for overall health, have limited predictive strength for psychological outcomes when modeled in isolation [3-5].

### Possible Explanations for Model Failure

Several methodological and conceptual factors may help explain the weak model performance observed. First, the discretization of the happiness score into three categories, while intended to facilitate classification and ensure class balance, may have obscured subtle distinctions between individuals' well-being levels. Future studies might consider modeling happiness as a continuous outcome or exploring ordinal regression frameworks to preserve more information. Second, the use of self-reported data introduces potential for error due to recall bias, misreporting, or social desirability effects. Features such as screen time or diet type may be inconsistently interpreted or under-reported by participants, thereby diluting their association with psychological outcomes.

Third, the models employed in this study-logistic regression and random forest-were not optimized through hyperparameter tuning and no feature engineering techniques were applied. More advanced models, such as gradient boosting machines or deep neural networks, may be better suited to uncovering weak or non-linear relationships, particularly if paired with a richer set of features.

Finally, it is possible that the relationship between lifestyle behaviors and happiness is mediated by unobserved variables such as coping mechanisms, cognitive appraisal, or resilience. In this case, traditional ML classifiers applied to raw behavioral inputs would be insufficient without incorporating these underlying psychological constructs.

### Reflecting on Educational and Methodological Value

Despite the limitations in predictive performance, this study illustrates the educational utility of accessible machine learning platforms such as Orange. The platform allowed for the transparent construction of classification models and the integration of visual analytics without requiring programming skills. As such, it holds value as a teaching tool for health informatics students and early-career researchers learning to apply ML techniques in real-world data contexts.

However, it is important to emphasize that ease of use must not come at the expense of methodological rigor. Simplified tools may help learners engage with data science concepts but should be accompanied by critical reflection on model assumptions, data limitations and ethical considerations.

## CONCLUSIONS

This study investigated whether self-reported lifestyle behaviors could be used to classify individuals into distinct levels of happiness using supervised machine learning models. The results demonstrated that both logistic regression and random forest classifiers performed only marginally above chance, indicating that lifestyle variables-when used in isolation-are insufficient for accurately predicting subjective well-being.

These findings underscore a critical insight: while behaviors such as sleep, exercise and screen time are often associated with well-being, they may not provide adequate explanatory power without the inclusion of psychological, emotional and contextual variables. This outcome aligns with previous research that emphasizes the multidimensional nature of happiness and supports the need for more holistic modeling frameworks.

The study also highlights the practical value of visual machine learning platforms like Orange for educational and exploratory purposes. While the tool enabled an accessible and transparent modeling process, its simplicity does not compensate for the limitations inherent in the input data. Therefore, emphasis should remain on the quality and scope of data rather than the platform itself.

From a research perspective, future studies should aim to integrate behavioral data with validated psychological constructs and demographic information to improve model performance. Methodologically, continuous outcome modeling or longitudinal data collection may offer richer insights into the dynamics of happiness over time.

From a policy and practice standpoint, the findings caution against over-reliance on behavioral indicators in digital mental health tools without accounting for the broader psychosocial context. For educational institutions and health informatics programs, this study illustrates both the potential and limitations of using low-code platforms to teach foundational ML concepts while encouraging critical reflection on model validity.

## REFERENCES

[1] Nan, Jason *et al.* "Personalized machine learning-based prediction of wellbeing and empathy in healthcare professionals." *Sensors,* vol. 24, no. 8, April 2025. https://www.mdpi.com/1424-8220/24/8/2640.

[2] Galkin, Fedor *et al.* "Optimizing future well-being with artificial intelligence: self-organizing maps (SOMs) for the identification of islands of emotional stability." *Aging (Albany NY),* vol. 14, no. 12, June 2022, pp. 4935-4958. https://pmc.ncbi.nlm.nih.gov/articles/PMC9271294/.

[3] Nichols, Emily S. *et al.* "A design for life: Predicting cognitive performance from lifestyle choices." *Plos One,* vol. 19, no. 4, April 2024. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0298899.

[4] Alhur, Anas, "Redefining Healthcare With Artificial Intelligence (AI): The Contributions of ChatGPT, Gemini and Co-pilot." *Cureus,* vol. 16, no. 4, April 2024. https://pubmed.ncbi.nlm.nih.gov/38721180/.

[5] Khan, Alif Elham *et al.* "Predicting life satisfaction using machine learning and explainable AI." *Heliyon,* vol. 10, no. 10, May 2024. https://www.sciencedirect.com/science/article/pii/S2405844024071895.

[6] Zhang, Naixin *et al.* "Prediction of adolescent subjective well-being: A machine learning approach." *General Psychiatry,* vol. 32, no. 5, September 2019. https://pmc.ncbi.nlm.nih.gov/articles/PMC6738679/.

[7] Pelt, Dirk H.M. *et al.* "Building machine learning prediction models for well-being using predictors from the exposome and genome in a population cohort." *Nature Mental Health,* vol. 2, August 2024, pp. 1217-1230. https://www.nature.com/articles/s44220-024-00294-2.

[8] Alhur, Anas, "Curricular Analysis of Digital Health and Health Informatics in Medical Colleges Across Saudi Arabia." *Cureus,* vol. 16, no. 8, August 2024. https://pubmed.ncbi.nlm.nih.gov/39280399/.

[9] Sulaiman, Sazan Kamal *et al.* "Statistical data mining methods in predicting happiness and habits." *ITM Web of Conferences,* vol. 64, 2024. https://www.itm-conferences.org/articles/itmconf/abs/2024/07/itmconf_icacs2024_01019/itmconf_icacs2024_01019.html.

[10] Sailaja, N. Venkata *et al.* "Happiness index prediction of students using machine learning." *Proceedings of the Fourth International Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023), Springer Nature,* vol. 18, 2023.

[11] Naveen and Bhatia, Anupam. *Need of machine learning to predict happiness: A systematic review* Edumania: An International Multidisciplinary Journal 2023, https://icertpublication.com/index.php/edu-mania/edumania-vol-1issue-2/need-of-machine-learning-to-predict-happiness-a-systematic-review/.

[12] Alhur, Anas. "The role of informatics in advancing emergency medicine: A comprehensive review." *Cureus,* vol. 16, no. 7, July 2024. https://www.cureus.com/articles/269532-the-role-of-informatics-in-advancing-emergency-medicine-a-comprehensive-review.pdf.

[13] Chaipornkaew, Piyanuch and Takorn Prexawanprasut. *A prediction model for human happiness using machine learning techniques* 2019 5th International Conference on Science in Information Technology (ICSITech). IEEE, 2019. 10.1109/ICSITech46713.2019.8987513, https://ieeexplore.ieee.org/abstract/document/8987513/.