Data Analyst Final Project

Student 1: Laia
Student 2: Morjan
Student 3: Tamer

Dataset Analysis Assignment (1)

1. Research Question

Which aspects of lifestyle have an impact on quality of life?
The research question focuses on the connection between lifestyle and quality of life. This is an important topic because our daily habits – such as how we eat, sleep, move, and connect with others – can strongly influence how we feel about our lives. The dataset includes many lifestyle variables, like diet (eating fruits and vegetables), physical activity (daily steps, meditation), sleep hours, stress level, social connections, and more. It also includes quality of life variables such as life satisfaction, personal achievements, sense of purpose, and income. Because the data includes both lifestyle and quality of life factors, and they are measured numerically, we can use this data to explore and analyze relationships. This will help us answer the research question and understand which lifestyle factors have the strongest impact on people's well-being.

2. Dataset Overview

Origin & Context

- The data is open-source and comes from Kaggle. It is a collection of self-reported lifestyle features along with a calculated `WORK_LIFE_BALANCE_SCORE`. It was collected starting from July 7, 2015.
- The limitations of this data relate to sampling bias or survey design; however, these limitations cannot be fully assessed without more context on the data collection methodology.
- The data was collected through an online self-report questionnaire. Participants answered questions about their daily habits, emotions, and lifestyle.
- Since it is self-reported data, responses may be subjective or personal. The exact selection criteria for participants are not provided.

Structure

The dataset contains 15,972 observations (rows), each representing an individual's responses or a single data point. There are 24 variables (columns), capturing different lifestyle characteristics and demographic information. The target variable for this analysis, based on the research question and the previous context, is `WORK_LIFE_BALANCE_SCORE`, which is a numerical representation of an individual's perceived quality of life or work-life balance.

3. Variable Dictionary

The table below presents a comprehensive overview of all the variables in the dataset. For each variable,it lists the data type, valid value range or categories, a short description, its role in the analysis (feature, target, or meta), and any relevant notes such as missing values or specific characteristics. This dictionary provides essential context for interpreting the dataset and preparing it for further analysis.

Name	**Data Type**	**Units / Values**	**Short Description**	**Role**	**Notes**
Timestamp	object	Categorical (2015–2021)	Date/time of data entry	Meta	No missing values
FRUITS_VEGGIES	int64	Scale (0–5)	Daily intake of fruits and vegetables	Feature	No missing values
DAILY_STRESS	object	Scale (0–5)	Daily stress level	Feature	No missing values
PLACES_VISITED	int64	Scale (0–10)	Number of new or interesting places visited	Feature	No missing values
CORE_CIRCLE	int64	Scale (0–10)	Strength of close personal relationships	Feature	No missing values
SUPPORTING_OTHERS	int64	Scale (0–10)	Frequency of supporting others	Feature	No missing values
SOCIAL_NETWORK	int64	Scale (0–10)	Size of social network	Feature	No missing values
ACHIEVEMENT	int64	Scale (0–10)	Sense of achievement	Feature	No missing values
DONATION	int64	Scale (0–5)	Amount of donations	Feature	No missing values
BMI_RANGE	int64	1 = Healthy, 2 = Unhealthy	Body Mass Index category	Feature	No missing values
TODO_COMPLETED	int64	Scale (0–10)	Number of completed tasks from a to-do list	Feature	No missing values
FLOW	int64	Scale (0–10)	Frequency of experiencing a state of flow	Feature	No missing values
DAILY_STEPS	int64	Scale (1–10)	Daily step count category	Feature	No missing values
LIVE_VISION	int64	Scale (0–10)	Clarity or existence of a life vision	Feature	No missing values
SLEEP_HOURS	int64	Scale (1–10)	Average hours of sleep per night	Feature	No missing values
LOST_VACATION	int64	Scale (0–10)	Amount of lost or unused vacation time	Feature	No missing values
DAILY_SHOUTING	int64	Scale (0–10)	Frequency of shouting or intense arguments	Feature	No missing values
SUFFICIENT_INCOME	int64	1 = Not sufficient, 2 = Sufficient	Sufficiency of income	Feature	No missing values
PERSONAL_AWARDS	int64	Scale (0–10)	Number or significance of personal awards	Feature	No missing values
TIME_FOR_PASSION	int64	Scale (0–10)	Time dedicated to personal passions	Feature	No missing values
WEEKLY_MEDITATION	int64	Scale (0–10)	Frequency or duration of weekly meditation	Feature	No missing values
AGE	object	Categorical: Less than 20, 21 to 35, 36 to 50, 51 or more	Age group of the respondent	Feature	No missing values
GENDER	object	Categorical: Female / Male	Gender of the respondent	Feature	No missing values
WORK_LIFE_BALANCE_SCORE	float64	Numeric (480.0 to 820.2)	Calculated score representing work-life balance	Target	No missing values

4. Descriptive Statistics

Numeric Variables: The table below summarizes the key descriptive statistics for all numeric variables in the dataset.

Feature	**Mean**	**Median**	**Std**	**Min**	**Max**
FRUITS_VEGGIES	2.92	3.0	1.44	0.0	5.0
PLACES_VISITED	5.23	5.0	3.31	0.0	10.0
CORE_CIRCLE	5.51	5.0	2.84	0.0	10.0
SUPPORTING_OTHERS	5.62	5.0	3.24	0.0	10.0
SOCIAL_NETWORK	6.47	6.0	3.09	0.0	10.0
ACHIEVEMENT	4.00	3.0	2.76	0.0	10.0
DONATION	2.72	3.0	1.85	0.0	5.0
BMI_RANGE	1.41	1.0	0.49	1.0	2.0
TODO_COMPLETED	5.75	6.0	2.62	0.0	10.0
FLOW	3.19	3.0	2.36	0.0	10.0
DAILY_STEPS	5.70	5.0	2.89	1.0	10.0
LIVE_VISION	3.75	3.0	3.23	0.0	10.0
SLEEP_HOURS	7.04	7.0	1.20	1.0	10.0
LOST_VACATION	2.90	0.0	3.69	0.0	10.0
DAILY_SHOUTING	2.93	2.0	2.68	0.0	10.0
SUFFICIENT_INCOME	1.73	2.0	0.44	1.0	2.0
PERSONAL_AWARDS	5.71	5.0	3.09	0.0	10.0
TIME_FOR_PASSION	3.33	3.0	2.73	0.0	10.0
WEEKLY_MEDITATION	6.23	7.0	3.02	0.0	10.0
WORK_LIFE_BALANCE_SCORE	**666.75**	**668.0**	**45.02**	**480.0**	**820.0**

Categorical Variable – Frequency Counts:

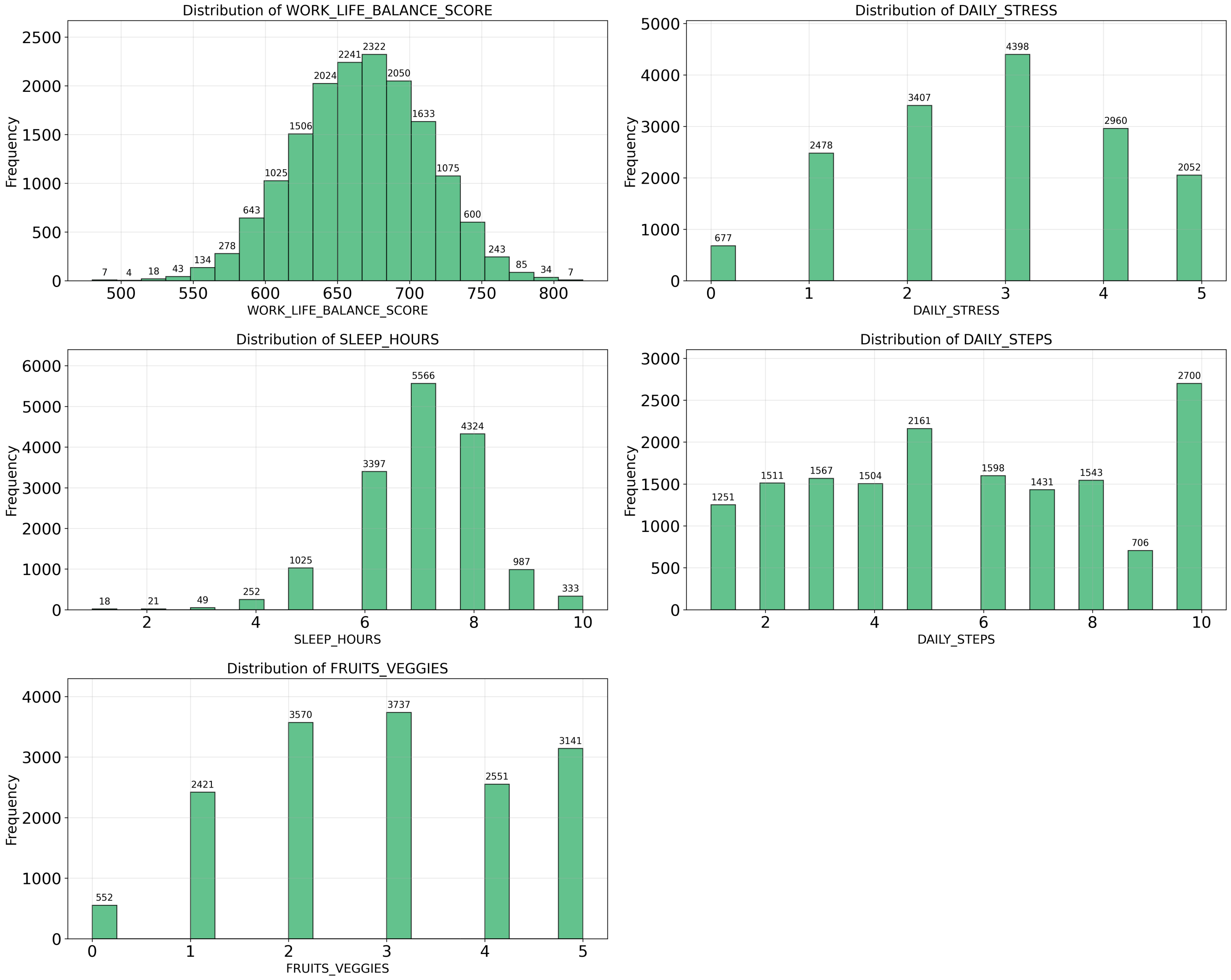
AGE	**Frequency Count**	**Percentage**
21 to 35	6,108	38.2%
36 to 50	4,655	29.1%
51 or more	3,390	21.2%
Less than 20	1,819	11.4%

Gender	**Frequency Count**	**Percentage**
Female	9,858	61.7%
Male	6,114	38.3%

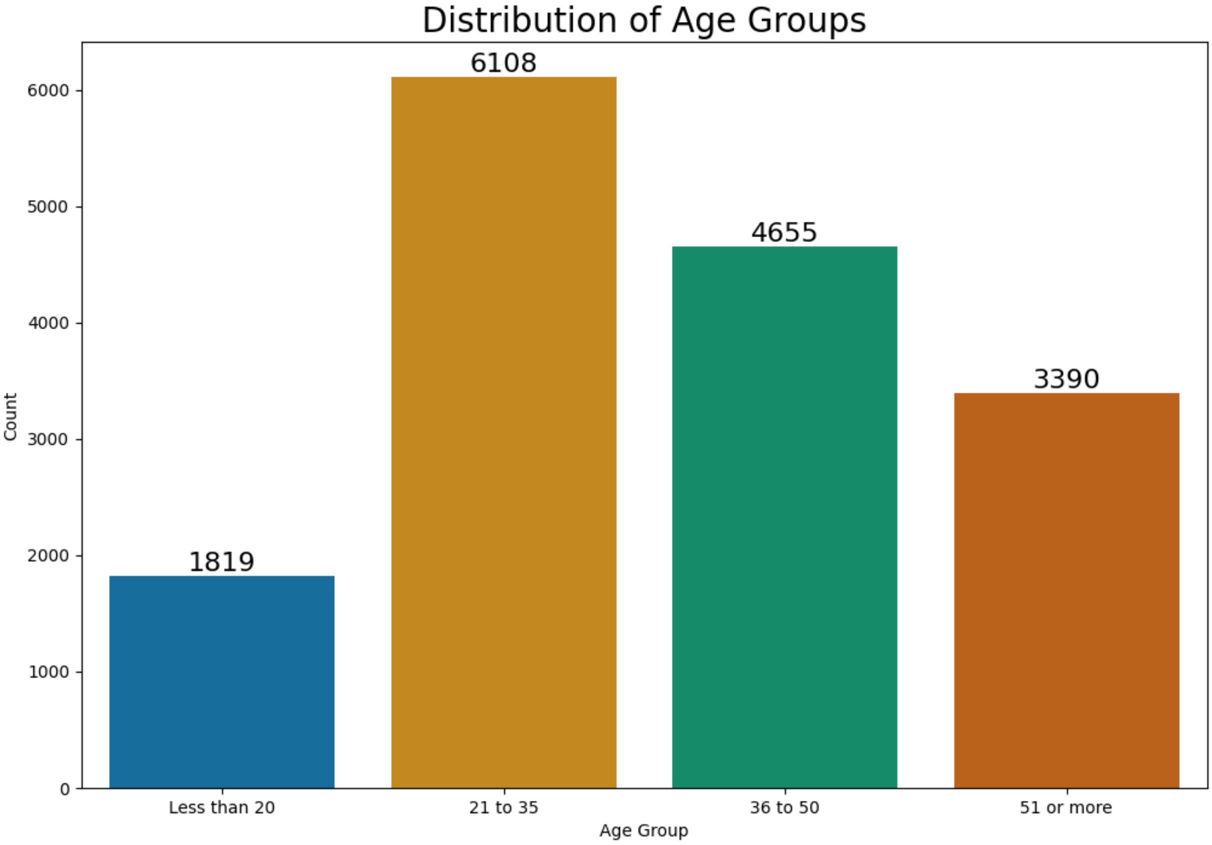
BMI Range	**Frequency Count**	**Percentage**
1 (Normal BMI)	9,413	58.9%
2 (High BMI)	6,559	41.1%

Sufficient Income	**Frequency Count**	**Percentage**
2	11,643	72.9%
1	4,329	27.1%

Distributions – numeric Histograms:

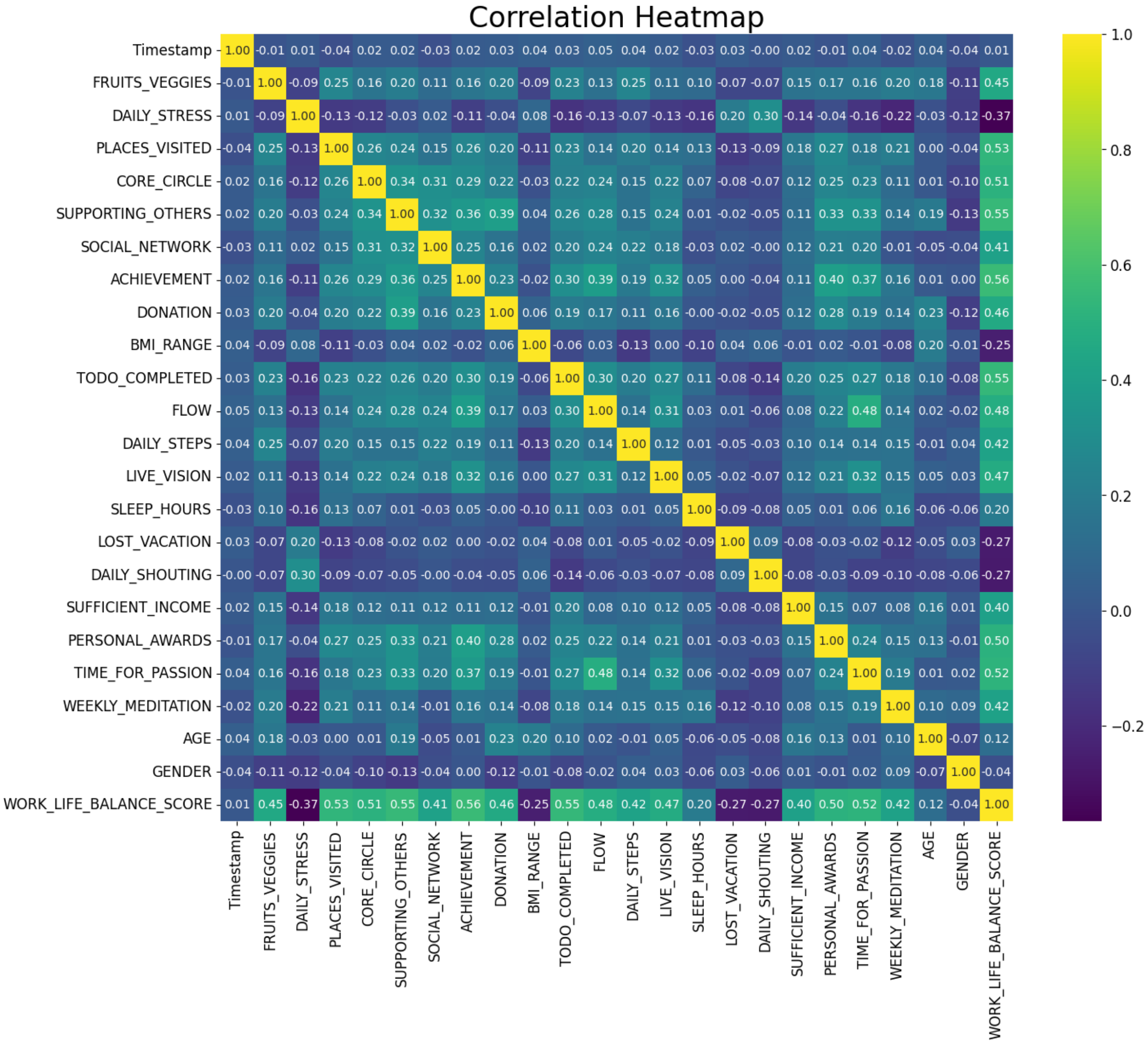


Distributions – categorical bar chart:



5. Correlation Analysis

This section explores the relationships between the various lifestyle variables and the WORK_LIFE_BALANCE_SCORE using correlation analysis. Correlation measures the strength and direction of a linear relationship between two variables. The correlation matrix below displays the Pearson correlation coefficients between all pairs of numerical variables in the dataset. This matrix is a fundamental tool for understanding the relationships with the target variable. Variables like GENDER and AGE have been encoded into numerical formats for this analysis.



Highest Absolute Correlation: FLOW and TIME_FOR_PASSION

From the correlation matrix and heatmap, the two variables with the highest absolute correlation are `FLOW` and `TIME_FOR_PASSION`, with a Pearson correlation coefficient of **0.48**. This indicates a strong positive linear relationship: frequency of experiencing a state of flow increases the amount of time dedicated to personal passions.

Discussion of the Relationship

While there is a moderate to strong positive correlation between `FLOW` and `TIME_FOR_PASSION` ($r = 0.48$), this does not necessarily imply a causal relationship. `FLOW` may cause an increase in `TIME_FOR_PASSION`, and at the same time, `TIME_FOR_PASSION` may cause more `FLOW` experiences. This mutual influence suggests a feedback loop, where each variable reinforces the other. Therefore, the relationship can be causal, but it is **not unidirectional** it is likely **bidirectional causal**.

Alternative Interpretations of the Relationship

- **Confounded:** A third variable, such as overall well-being, free time availability, or personality traits (openness or intrinsic motivation), might influence both the tendency to experience flow and the amount of time spent on personal passions.
- **Not Spurious:** Given the psychological connection between being engaged (flow) and doing enjoyable activities (passion), the relationship is unlikely to be completely spurious.
- **Possibly Bidirectional:** Experiencing flow may encourage individuals to spend more time on passions, and vice versa engaging in passion projects may increase the chance of entering flow.

Why the Relationship is Not Necessarily Causal

1. **Lack of directionality:** The data is observational and cross-sectional, meaning we don't know whether flow causes more time for passion, or vice versa.
2. **Possible third variables (confounders):** Factors such as free time, personality traits (like intrinsic motivation), or mental well-being could influence both flow and time for passion. This would make the relationship confounded.
3. **No experimental control:** Without randomly assigning participants to experience flow or spend more time on passions, we can't isolate cause and effect.
4. **Bidirectionality is possible:** It's plausible that engaging in personal passions leads to flow, and experiencing flow motivates individuals to dedicate more time to passions, making the relationship reciprocal.

Conclusion

The correlation between `FLOW` and `TIME_FOR_PASSION` is **meaningful and theoretically supported**, but should be interpreted with caution. While the relationship is unlikely to be spurious, causality cannot be confirmed without further experimental or longitudinal research that controls for potential confounding variables.

Summary

- **Bidirectional Causal Relationship**
- Still needs experimental or longitudinal data to confirm directionality
- Common in psychology and behavior (motivation $\leftrightarrow$ performance)
- Without experimental or longitudinal data, we cannot conclude that `FLOW` causes `TIME_FOR_PASSION`
- The relationship is not purely causal in one direction; it is likely **a bidirectional causal relationship**.

6. Connecting to the Research Question

This section directly addresses the research question by mapping feature relevance and selecting the most informative features based on a systematic importance measure.

Restate the given research question

The primary research question guiding this analysis is: "Which aspects of lifestyle have an impact on quality of life"?

Feature relevance mapping

1) `FRUITS_VEGGIES`:
This variable measures the intake of healthy foods, Higher intake of fruits and vegetables is generally associated with better physical health, which is a fundamental component of overall quality of life. Therefore, it is highly relevant as a direct indicator of healthy lifestyle choices impacting well-being.

2) `DAILY_STRESS`:
This variable directly quantifies an individual's daily stress level. High stress is a known detractor from quality of life and work-life balance, its relevance is very high, as understanding stress levels is crucial for identifying negative impacts on well-being.

3) `PLACES_VISITED`:
This variable likely reflects engagement in leisure activities, exploration, and potentially social interaction. Visiting new places can contribute to mental well-being, reduce monotony, and provide enriching experiences, all which enhances quality of life, it is highly relevant.

4) `CORE_CIRCLE`:
This variable indicates the strength or quality of an individual's closest social ties. Strong social support networks are consistently linked to better mental health, resilience, and overall life satisfaction, making it highly relevant to quality of life.

5) `SUPPORTING_OTHERS`:
Engaging in prosocial behavior, such as supporting others, is often associated with increased personal happiness, a sense of purpose, and improved social connections. These factors directly contribute to an individual's quality of life, making this variable highly relevant.

6) `SOCIAL_NETWORK`:
This variable broadly represents the size and perhaps activity within an individual's social connections. A robust social network can provide emotional support, opportunities for interaction, and a sense of belonging, all which are important for quality of life, it is highly relevant.

7) `ACHIEVEMENT`:
This variable reflects an individual's sense of accomplishment. Achieving goals, whether personal or professional, can boost self-esteem, provide a sense of purpose and contribute to overall life satisfaction, It is highly relevant as a direct measure of personal fulfillment impacting quality of life.

8) `DONATION`:
Donating (time or money) is a prosocial behavior that can foster a sense of purpose, altruism, and connection to a larger community, thereby enhancing quality of life, it is relevant.

9) `BMI_RANGE`:
This variable is a health indicator. While not a direct measure of Lifestyle behavior, an unhealthy BMI range can lead to various health issues that significantly diminish quality of life. Its relevance is indirect but important, as Physical health underpins overall well-being.

10) `TODO_COMPLETED`:
This variable likely reflects productivity and effective time management. Successfully completing tasks can reduce stress, increase a sense of control, and provide a feeling of accomplishment, all contributing positively to quality of life, it is highly relevant.

11) `FLOW`:
Experiencing a state of flow (being fully immersed and energized in an activity) is a key component of positive psychology and is strongly linked to happiness and engagement. This variable is highly relevant as it directly measures a state conducive to high quality of life.

12) `DAILY_STEPS`:
This variable is a direct measure of physical activity. Regular physical activity is well-established to improve physical health, mental well-being, and reduce stress, all of which are crucial for a high quality of life, it is highly relevant.

13) LIVE_VISION:
Having a clear life vision provides direction, purpose, and motivation, It can reduce feelings of aimlessness and contribute to a sense of meaning, which are vital for overall life satisfaction and quality of life, This variable is highly relevant.

14) SLEEP_HOURS:
Adequate sleep is fundamental for physical and mental health. Insufficient sleep can negatively impact mood, cognitive function, and overall well-being, Thus, directly affecting quality of life, it is highly relevant.

15) LOST_VACATION:
This variable indicates unused leisure time, which can lead to burnout, increased stress, and reduced opportunities for rest and rejuvenation, it is highly relevant as it directly impacts on work-life balance and overall well-being.

16) DAILY_SHOUTING:
This variable likely reflects levels of frustration, anger, or conflict in daily life. High levels of such negative interactions can significantly detract from mental peace and overall quality of life, it is highly relevant as an indicator of negative emotional states.

17) SUFFICIENT_INCOME:
Financial stability and the perception of having sufficient income can significantly reduce stress and provide access to resources that enhance quality of life (better housing, healthcare, leisure activities), It is highly relevant.

18) PERSONAL_AWARDS:
Receiving personal awards or recognition can boost self-esteem and provide a sense of validation and accomplishment, contributing positively to quality of life, it is highly relevant.

19) TIME_FOR_PASSION:
Dedicating time to personal passions and hobbies is crucial for mental well-being, stress reduction, and personal fulfillment. This variable directly measures engagement in activities that enhance quality of life, it is highly relevant.

20) WEEKLY_MEDITATION:
Regular meditation or mindfulness practice is known to reduce stress, improve emotional regulation, and enhance overall mental clarity and well-being. This variable is highly relevant as a direct measure of practice, that improves quality of life.

21) AGE:
Age is a demographic variable that can influence lifestyle choices, health status, and perceptions of quality of life. While not a direct lifestyle characteristic, it is relevant as a contextual factor that might moderate or explain relationships between lifestyle and quality of life.

22) GENDER:
Gender is another demographic variable that can be associated with different lifestyle patterns, societal expectations, and health outcomes, all of which might indirectly affect quality of life, It is relevant as a contextual factor.

23) WORK_LIFE_BALANCE_SCORE:
This is the target variable and directly represents the quality of life (or work-life balance) that the research question aims to understand, all other variables are assessed for their impact on this score.

Choice of k (Number of Most Informative Features):

We will report the top 11 most informative features (k = 11), this choice is justified by selecting all features with an absolute Pearson correlation coefficient greater than or equal to 0.45. This threshold ensures that we include all variables that demonstrate a moderate to strong linear association with the WORK_LIFE_BALANCE_SCORE, capturing both positive and negative significant influences, this allows for a comprehensive understanding of the primary lifestyle characteristics affecting quality of life without including variables with very weak or negligible associations.

The selected features are:

- ACHIEVEMENT: 0.561
- SUPPORTING_OTHERS: 0.549
- TODO_COMPLETED: 0.546
- PLACES_VISITED: 0.530
- TIME_FOR_PASSION: 0.517
- CORE_CIRCLE: 0.508
- PERSONAL_AWARDS: 0.504
- FLOW: 0.478
- LIVE_VISION: 0.471
- DONATION: 0.459
- FRUITS_VEGGIES: 0.452

Related Work Assignment (2)

Paper - 1

The influence of health-promoting lifestyles on the quality of life of retired workers in a medium-sized city of Northeastern China

1. Restated Research Question

How do different components of a healthy lifestyle (such as physical activity, diet, social relationships, and avoidance of harmful behaviors) impact the quality of life among retirees?

2. Justification of Article Relevance

The first paper's relevance lies in its comprehensive approach to understanding how mental and emotional behaviors influence overall life satisfaction. The paper directly researches the impact of a healthy lifestyle (such as physical activity, diet, avoidance of harmful behaviors, and social relationships) on the quality of life among retirees. It shows the relationship between various aspects that have been researched in the original assignment and the quality of life among retirees.

3. Four-Line Article Summary

Citation	zhang et al [5]. *Promoting a healthy lifestyle among retirees in China: A study on the impact of physical activity, diet, social relationships, and avoidance of harmful behaviors on quality of life.*
Objective and Method	It develops a conceptual framework for understanding quality of life by life using statistical methods, demographic data, HPLP-II and WHOQOL-BREF questionnaires.
Key Findings	The framework suggests that mental well-being, personal values, and stress management play a central role in shaping quality of life.
Limitations and Difference	**Limitation:** The original article uses only a few features of a healthy lifestyle to build the predictive model, which limits the complexity and depth of the analysis. Additionally,the study focuses exclusively on retirees, making the findings less generalizable to the broader population, especially working individuals with different lifestyle dynamics. **In contrast, our approach:** includes a broader and data-driven feature selection process, using 11 features that show strong correlation with the target variable. Furthermore, our analysis is based on a more diverse sample that includes working individuals, allowing the model to better generalize across different groups. Combined with proper validation techniques, this results in a more robust and applicable predictive model.

4. Cross-paper Synthesis

The paper provides a broad theoretical understanding of quality of life, emphasizing the psychological and emotional aspects. In contrast to the 2023 study which is empirical and focused on physical activity, this paper builds a conceptual foundation around stress, values, and well-being. The combined strength of both papers lies in their complementary nature: one offers data-backed results, and the other provides theoretical depth. A key gap in both is the absence of holistic, multidimensional models that are both theoretical and validated by real-world data. My study seeks to bridge this gap by creating an integrated model tested with empirical data from diverse lifestyle factors.

Paper - 2

Perceived Quality of Life Is Related to a Healthy Lifestyle and Related Outcomes in Spanish Children and Adolescents: The Physical Activity, Sedentarism, and Obesity in Spanish Study

1. Restated Research Question

Which aspects of lifestyle, including physical activity, sedentary behavior, and obesity, have an impact on the quality of life of Spanish children and adolescents?

2. Justification of Article Relevance

It specifically examines how key lifestyle factors physical activity, sedentary behavior, and obesity impact quality of life in Spanish children and adolescents. This study highlights the impact of particular behaviors on overall health, identifying which lifestyle factors affect quality of life, Lifestyle factors play a major role in quality of life and closely relate to the original assignment.

3. Four-Line Article Summary

Citation	rodenas et al [4], The impact of physical activity on the quality of life: Evidence from working adults, *Journal of Health Research.*
Objective and Method	The main objective is analyzing the relationship between lifestyle factors and quality of life (HRQoL). Methods included physical activity, diet, sleep, screen time, fitness, and BMI, using self-reported QoL surveys and quantitative analysis.
Key Findings	Children with regular physical activity had a high quality of life, while sedentary behavior and obesity were associated with lower quality of life.
Limitations and Difference	**Limitation:** The original article aims to identify which aspects of lifestyle including physical activity, sedentary behavior, and obesity impact the quality of life of Spanish children and adolescents, using only a few features in the predictive model, This limited number of features restricts the ability to capture the complex and multifaceted nature of lifestyle factors affecting quality of life. Additionally, focusing exclusively on Spanish children and adolescents reduces the generalizability of the findings to other age groups or populations. **Our Difference:** In contrast, our approach expands the analysis by including 11 features selected through a data-driven process based on their correlation with the target variable. Moreover, our dataset contains both young and older individuals, providing a more diverse and representative sample. This diversity improves the model's generalizability across different age groups and life stages and enhances the validation process by incorporating a larger and more balanced dataset. Together, these improvements lead to a more robust, accurate, and broadly applicable predictive model for understanding lifestyle impacts on quality of life.

4. Cross-paper Synthesis

Several studies highlight a positive relationship between physical activity and quality of life, using self-reported data. While these studies offer valuable insights especially by analyzing physical, mental, and social domains of well-being they are often limited by their narrow focus on a single lifestyle factor and reliance on subjective data. My research addresses these gaps by examining the combined impact of multiple lifestyle components and incorporating objective data sources such as sleep and activity tracking.

Paper - 3

A Prediction of Work-life Balance Using Machine Learning

1. Restate Your Research Question

What lifestyle-related features most accurately predict quality of life, based on machine learning models trained on health-related data?

2. Justification of Relevance

The selected paper is relevant because it applies machine learning to predict work-life balance based on employee attributes, including demographic, job satisfaction, and behavioral data directly addressing our research aim of identifying lifestyle predictors that influence work-life quality.

3. Four-Line Article Summary

Citation	choi et al. [1], *A prediction of work-life balance using machine learning*, *Asia Pacific Journal of Information Systems.*
Objective and Method	The study aimed to predict employees' work-life balance using HR data from IBM, applying GLM (Generalized Linear Models) for both linear regression and binomial classification. Data mining and preprocessing were also used to analyze variable effects.
Key Findings	Variables like job role, overtime, marital status, and satisfaction were predictive of work-life balance. The binomial model achieved approximately 71% accuracy, but AUC was low, showing limitations in generalizability.
Limitations and Difference	**Limitation:** The original article focuses on identifying which lifestyle-related features most accurately predict quality of life using machine learning models trained on health-related data. However, it uses only two features in the predictive model, which limits the ability to capture the complexity of lifestyle factors affecting quality of life. Additionally, the study's dataset is limited in diversity, restricting the generalizability of the results to broader populations. **Our Difference:** In contrast, our approach includes a broader, data-driven feature selection process with 11 features that show strong correlation with the target variable. Moreover, our dataset contains both young and older individuals, providing a more diverse and representative sample. We applied four different predictive models to the data, which, combined with robust cross-validation techniques, enhances the model's ability to generalize to unseen data. Using this larger, more balanced dataset and multiple modeling approaches, the resulting models are more robust, accurate, and broadly applicable for predicting quality of life based on lifestyle-related features.

4. Cross-paper synthesis

The reviewed paper demonstrates several strengths, particularly in applying machine learning techniques (GLM) to predict work-life balance using employee data. It effectively identifies demographic and workplace variables that influence balance, achieving a reasonable prediction accuracy. This approach represents a valuable intersection between data science and human resource research. However, the study also presents key limitations. It relies on a limited dataset (IBM HR data) that lacks depth in lifestyle-specific factors such as nutrition, physical activity, sleep quality, or passion-related engagement. Additionally, it does not explore interaction effects (gender × stress), nor does it incorporate subjective measures like perceived life quality or emotional well-being. The generalizability of the results is constrained by the low AUC and the simplicity of the variable set. To improve upon this, my research will include a broader range of lifestyle and psychological variables, such as flow, time for passion, and social support. Furthermore, I will incorporate interaction terms and demographic segmentation to uncover nuanced effects. My study will also emphasize interpretability and practical recommendations, ensuring that the findings can guide actionable strategies for improving quality of life and work-life balance.

Paper - 4

Serious Games for Emotional Intelligence's Skills Development for Inner Balance and Quality of Life-A Literature Review

1. Restate Your Research Question

How can serious games be used to enhance emotional intelligence and promote inner balance as a means to improve overall quality of life?

2. Justification of Relevance

This paper is directly relevant as it explores how developing emotional intelligence through serious games can improve mental, emotional, and social well-being core components of quality of life and work-life balance. It reviews empirical studies supporting the positive impact of emotional skills on life satisfaction and health, which aligns with the focus of this research.

3. Four-Line Article Summary

Citation	papoutsi et al. [3], *Serious Games for Emotional Intelligence's Skills Development for Inner Balance and Quality of Life.*
Objective and Method	The paper reviewed 14 studies (2010–2021) on serious games used to enhance emotional intelligence. It categorized the games by their focus (emotion regulation, empathy, special needs) and assessed their design and outcomes.
Key Findings	Serious games significantly improved emotional regulation, empathy, and social behavior, particularly among children, adolescents, and individuals with autism or dementia, leading to improved well-being and quality of life.
Limitations and Difference	**Limitation:** The original article investigated how serious games can be used to enhance emotional intelligence and promote inner balance to improve overall quality of life. However, the study uses only two features in the predictive model, limiting the ability to fully capture the multifaceted effects of serious games on emotional and psychological outcomes. Additionally, the dataset is limited in diversity, which restricts the generalizability of the findings to broader populations. It is also important to note that emotional improvements are influenced by factors beyond just gameplay. **Our Difference:** In contrast, our approach incorporates a broader, data-driven feature selection process with 11 features showing strong correlation with the target variable. Moreover, our dataset includes both young and older individuals, offering a more diverse and representative sample. We applied four different predictive models combined with robust cross-validation techniques, which enhance the models' ability to generalize to unseen data. By considering a wider range of factors influencing emotions and inner balance not only those related to serious games these improvements enable the development of more accurate, robust, and widely applicable models for understanding how serious games impact emotional intelligence, inner balance, and overall quality of life.

4. Cross-paper synthesis

The reviewed paper emphasizes the importance of emotional intelligence in achieving inner balance and quality of life. It presents strong evidence that serious games can improve emotional regulation and empathy, especially among children and special populations. This suggests that emotional skill-building through digital tools can contribute meaningfully to psychological and social well-being. One key strength of the paper is its comprehensive literature review, covering over a decade of studies with diverse target groups and applications. Another strength is its focus on practical, interactive tools that show positive results in emotional development. However, the paper lacks evidence on whether improvements seen in-game translate into real-life behaviors, especially in adult or working populations. It also does not assess long-term effects or differences across demographic groups. My research aims to address these gaps by studying how emotional intelligence interacts with everyday lifestyle factors such as nutrition, stress, and flow, and by focusing on long-term and real-world outcomes in general adult populations.

Paper - 5

The relationship between quality of life, sleep quality, mental health, and physical activity in an international sample of college students: a structural equation modeling approach

1. Restate your research question

"How do various lifestyle characteristics—such as sleep quality, physical activity, and mental health affect the perceived quality of life among individuals across different age groups?"

2. Justification of Relevance

This paper is directly relevant to our research question, as it investigates how lifestyle-related factors specifically sleep quality, physical activity, and mental health relate to quality of life. These factors are comparable to many variables in our dataset, such as sleep hours, daily steps, and stress.

3. Four-Line Article Summary

Citation	moussa et al. [2], *The relationship between quality of life, sleep quality, mental health, and physical activity in an international sample of college students.*
Objective and Method	The study explored how lifestyle factors directly and indirectly affect life quality using a cross-sectional survey design and questionnaires, analyzing the relationships using structural equation modeling.
Key Findings	Better sleep quality, higher physical activity, and lower depression and anxiety levels were significantly associated with higher quality of life scores. The strongest predictor was sleep quality.
Limitations and Difference	**Limitation:** The original article focuses on how specific lifestyle factors affect perceived quality of life, but it typically examines only one variable at a time, such as physical activity or mental health, without accounting for their combined or interacting effects. In addition, many such studies rely heavily on self-reported data and are often limited to narrow demographic groups, which restricts the generalizability of the findings across different age ranges. **Difference:** Our approach aims to overcome these limitations by analyzing the combined impact of multiple lifestyle characteristics, including sleep quality, physical activity, and mental well-being, on perceived quality of life. We use a diverse dataset that includes individuals from various age groups, which allows for broader generalization. Furthermore, we incorporate objective data sources such as sleep tracking and activity monitors alongside self-reports, and apply four predictive models with robust cross-validation. This enables a more comprehensive, accurate, and generalizable understanding of how lifestyle factors influence quality of life across different age groups.

4. Cross-paper synthesis

This paper highlights the importance of sleep quality, physical activity, and mental well-being in influencing quality of life. Its strength lies in the use of validated measurement tools across a diverse international sample, which supports the reliability of its findings. However, it is limited by its focus on a narrow population college students and the use of subjective self-reporting, which can introduce bias. Our project builds on these strengths while addressing key limitations. Unlike the paper, which relies mainly on psychological scales, our dataset includes more behavioral and objective variables such as daily steps, task completion, and fruit/vegetable intake. This allows us to test lifestyle predictors beyond just sleep and mood. Moreover, our broader population scope improves external validity. A recurring limitation in similar studies is the cross-sectional design, which limits causal inference. While our analysis is also correlational, the diversity of our features may reveal more nuanced patterns and interaction effects that have been overlooked, Ultimately, our approach aims to identify the most influential lifestyle behaviors that predict higher quality of life across a more representative sample.

Comparative Synthesis of Reviewed Articles

Collective Strengths of the Articles

Aspect	Description
Use of validated measurement tools	Most articles rely on psychological instruments and quantitative metrics, enhancing reliability.
Focus on different aspects of quality of life	Topics such as physical activity, emotional intelligence, and work-life balance provide multidimensional understanding.
Theoretical and practical contribution	Some articles offer comprehensive theoretical frameworks (Shahed 2013), while others present practical solutions like serious games or predictive models.
Emphasis on emotional and physical factors	Mental, physical, and social factors are highlighted as key components affecting quality of life.

Collective Weaknesses and Major Gaps

Aspect	Description
Overreliance on self-report	Most studies depend on subjective questionnaires, increasing risk of measurement bias.
Limited samples	Samples are often narrow (students or employees from specific companies), reducing external validity.
Focus on single variables	Many studies examine isolated factors (only physical activity or only emotional intelligence).
Lack of objective and behavioral data	There is minimal use of actual behavioral metrics (daily steps, measured sleep, real dietary data).
Correlational, not experimental design	Findings are correlational and do not allow causal conclusions.

Justification for our Research

Improvement Area	Your Study's Contribution
Broadening variable range	Integrates emotional, nutritional, physical, and social factors.
Use of objective and behavioral data	Reduces sole reliance on questionnaires.
More diverse sample	Enhances external validity of findings.
Interaction and complex effect analysis	Enables deeper understanding of variable dynamics.
Practical application	Aims to create a theory-based, empirically tested integrative model to guide policy and strategy for improving quality of life.

Model Development Assignment (3)

1. Problem Statement

This project aims to investigate the factors that influence an individual's `Work-Life Balance Score`. Specifically, the goal is to build predictive models that estimate a person's `Work-Life Balance Score` based on a set of lifestyle and behavioral attributes.
The core prediction task is:
Predict the `WORK_LIFE_BALANCE_SCORE` using 11 selected features,this is primarily a regression problem. In classification variants, the target was binarized for distinguishing between high and low balance levels.

2. Data Preprocessing

A series of steps were performed to clean and prepare the data prior to modelling:

- **Handling missing values:** The dataset was checked for missing values. Rows with critical missing data were removed, and others were imputed where appropriate.
- **Target preparation:**
  - For regression models: the target `WORK_LIFE_BALANCE_SCORE` was used as-is.
  - For classification models: the score was binarized using the threshold $(min + max)/2$. Values above the threshold were labeled `1` ("High") and values below or equal to the threshold were labeled `0` ("Low"). The distribution of binary labels was then checked.
- **Scaling:** All 11 selected features were standardized (z-score normalization) before training regression models. This ensures that features are on comparable scales for algorithms like Linear Regression.
- **Encoding categorical variables:** Not required, since all selected features were numeric.
- **Train/Validation/Test Split:** Data was split into 70% training, 15% validation, and 15% test sets using `train_test_split` with a fixed `random_state` for reproducibility.
- **Class imbalance (only for classification):** After binarizing the target using the $(min+max)/2$ rule, the number of samples in each class (0 and 1) was counted. The distribution was found to be reasonably balanced, so no additional resampling techniques were applied.

3. Feature Engineering & Selection

Feature Selection

To identify the most informative features for predicting the target variable `WORK_LIFE_BALANCE_SCORE`, we applied a filter-based feature selection method using Pearson correlation. First, we computed the correlation matrix between all numerical variables in the dataset. We then extracted the correlation values between each feature and the target variable. To focus on meaningful relationships, we selected features with an absolute correlation coefficient ($|r|$) greater than or equal to 0.45. The selected features were then sorted by the absolute value of their correlation with the target, and the top 11 most strongly correlated features were retained for further modeling.

Selected Features (in descending order of $|r|$)

- `ACHIEVEMENT`
- `SUPPORTING_OTHERS`
- `TODO_COMPLETED`
- `PLACES_VISITED`
- `TIME_FOR_PASSION`
- `CORE_CIRCLE`
- `PERSONAL_AWARDS`
- `FLOW`
- `LIVE_VISION`
- `DONATION`
- `FRUITS_VEGGIES`

4. Modelling Experiments

Four different models were trained and evaluated. For each model, we summarize the key hyperparameters used during tuning, Additionally, we describe the cross-validation procedure applied.

1. Linear Regression

- **Type:** Regression
- **Feature Scaling:** Standardized
- **Key Hyperparameters:**
  - `lr__fit_intercept:` [True, False]
  - `lr__positive:` [False, True]
- **Tuning Method:** GridSearchCV with 5-fold cross-validation (`cv=5`)
- **Evaluation Metrics:** RMSE, MAE, MSE, R^2

2. Random Forest Regressor

- **Type:** Regression
- **Key Hyperparameters:**
  - `n_estimators:` [100]
  - `max_depth:` [None, 10]
  - `min_samples_split:` [2]
- **Tuning Method:** GridSearchCV with 5-fold cross-validation (`cv=5`)
- **Evaluation Metrics:** RMSE, MAE, MSE, R^2

3. Random Forest Classifier

- **Type:** Classification (target binarized)
- **Key Hyperparameters:**
  - `n_estimators:` [100]
  - `max_depth:` [None, 10]
  - `min_samples_split:` [2]
- **Tuning Method:** GridSearchCV with 5-fold cross-validation (`cv=5`)
- **Evaluation Metrics:** Accuracy, F1-score, Confusion Matrix, Recall

4. XGBoost Classifier

- **Type:** Classification (target binarized)
- **Key Hyperparameters:**
  - `n_estimators:` [100]
  - `max_depth:` [None, 10]
  - `min_samples_split:` [2]
- **Tuning Method:** GridSearchCV with 5-fold cross-validation (`cv=5`)
- **Evaluation Metrics:** Accuracy, F1-score, Confusion Matrix, Recall

Model Evaluation Assignment (4)

1. Metric Selection

Task 1 – Regression (Predicting WORK_LIFE_BALANCE_SCORE)

We selected the following metrics:

1. **RMSE (Root Mean Squared Error):**
   RMSE measures the square root of the average squared differences between predicted and actual values, penalizing large errors more heavily.
   We selected this metric because large prediction errors in quality of life scores are more harmful to interpretation. For example, predicting a score of 90 when it's actually 50 can lead to incorrect conclusions.
   The model was trained on the training set and evaluated on the validation set.

2. **MAE (Mean Absolute Error):**
   MAE measures the average absolute difference between predicted and actual values.
   It is useful for understanding the typical prediction error, regardless of whether it is large or small, and is more robust to outliers compared to RMSE.
   MAE was calculated on the validation set to ensure reliable evaluation.

3. **MSE (Mean Squared Error):**
   MSE measures the average of the squared differences between predicted and actual values.
   It helps us understand how far, on average, the model's predictions are from the real quality-of-life scores, with a stronger penalty for larger errors.
   We included MSE to complement RMSE and MAE, as it allows us to observe the raw squared error magnitude before taking the square root.
   Evaluating MSE on the validation set helps ensure that the model performs reliably on unseen data and that large deviations are properly captured.

4. **$\mathbf{R^2}$ Score:**
   It is a key performance metric for regression models. It measures how well the independent variables (features) explain the variability in the dependent variable (target).

   - $\mathrm{R}^2 = 1.0$ → Perfect prediction: the model explains 100% of the variance in the target.
   - $\mathrm{R}^2 = 0$ → The model explains none of the variance (no better than the mean).
   - $R^2 < 0$ → The model performs worse than simply predicting the mean.

Task 2 – Classification (Predicting TARGET_BINARY)

Before selecting appropriate evaluation metrics for our classification models, we first analyzed the distribution of the target variable TARGET_BINARY, which indicates whether an individual has a high quality of life. Assessing class balance is essential because a significant imbalance could bias the model and make metrics like accuracy misleading. By calculating the ratio between classes, we verified that the dataset is relatively balanced, allowing us to use accuracy alongside precision, recall, and F1-score for a comprehensive evaluation. The distribution is as follows:

- Number of zeros (0): 5,682
- Number of ones (1): 10,290
- Ratio of ones to zeros: 1.81

These results show that there is no significant class imbalance, as both classes are reasonably represented in the dataset. This observation is important when choosing metrics: Since the data is not heavily imbalanced, Accuracy is considered a meaningful performance metric. However, we still include Precision, Recall, and F1-Score to gain deeper insights into the model's behavior, especially in capturing the positive class correctly, which is relevant to our research goal of identifying factors that influence quality of life. We selected the following metrics:

1. **Accuracy:**
   We selected Accuracy as a primary evaluation metric because it reflects the overall proportion of correct predictions made by the model.
   In our case, it indicates how well the model can classify individuals into high or low quality of life categories based on their lifestyle characteristics.
   The dataset was split into training and validation sets to avoid overfitting and to ensure an objective accuracy estimate.

2. **Recall:**
   Recall evaluates how many of the actual high quality of life cases were correctly identified by the model.
   This is crucial if our goal is to detect as many true positive cases as possible — for example, to understand which factors contribute positively to quality of life.
   Performance was assessed on the validation set as well.

3. **F1-Score:**
   The F1-Score is the harmonic mean of precision and recall.
   We chose this metric especially due to potential class imbalance (more individuals with low quality of life), as it balances the trade-off between missing true cases and minimizing false alarms.
   F1-Score provides a more robust assessment when both types of errors are important to control.

2. Results Reporting

All metrics shown are based on validation set performance and rounded to three significant digits.

Model	**R^2 Score**	**MAE**	**RMSE**	**Accuracy**	**F1-Score**	**Runtime / Notes**
Linear Regression	0.824	15.133	18.998	–	–	Simple, interpretable model
Random Forest Regressor	0.814	15.499	19.516	–	–	Non-linear, more complex
Random Forest Classifier	–	–	–	0.885	0.89	High recall & precision
XGBoost Classifier	–	–	–	0.871	0.87	Best classification model

3. Statistical Significance & Confidence

Regression Models (Linear Regression vs. Random Forest Regressor)

T-statistic: 2.9319
p-value: 0.0034
95% CI for MAE difference (RF LR): [6.6035, 33.2774]

Interpretation: The p-value is less than 0.01, indicating a statistically significant difference in mean absolute error between the Linear Regression and Random Forest models.

Classification Models (Random Forest Classifier vs. XGBoost Classifier)

McNemar's Test p-value: < 0.0001
95% CI for Accuracy Difference (RF XGBoost): [-0.3467, -0.3000]
Mean Accuracy Difference: -0.324

Interpretation: The difference in performance between the classifiers is statistically significant, Random Forest performs better than XGBoost.

4. Error & Residual Analysis

Regression (Linear Regression Model):

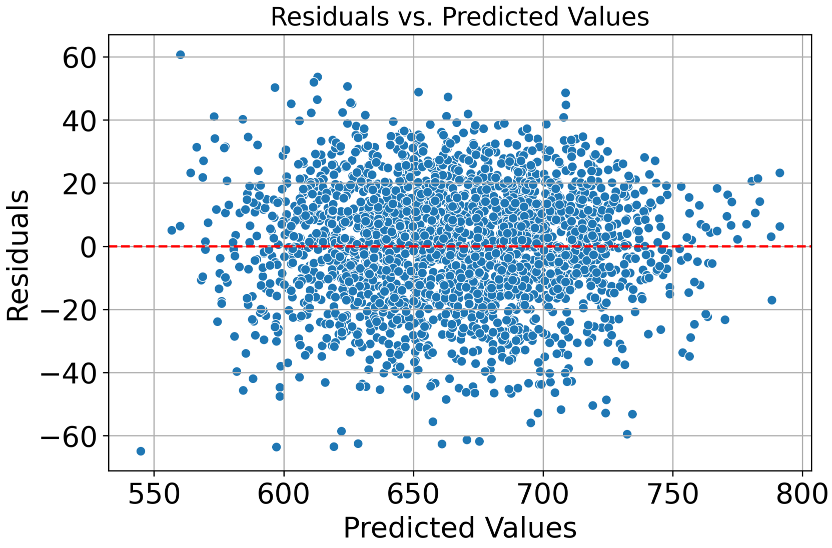


Figure 1: Residual plot of the Random Forest Regressor

Residual Analysis

The residuals appear to be evenly spread around the zero line across the entire range of predicted values. There is no visible "funnel shape" indicating increasing or decreasing variance as predictions increase. Therefore, there is no clear evidence of heteroscedasticity.
Not significantly. The points are fairly symmetrically distributed above and below the red line (zero residual line), with no consistent upward or downward trend in any specific prediction range. This suggests that the model does not consistently over- or under-predict, meaning there is no strong systematic bias.

Summary

- No clear signs of heteroscedasticity.
- No significant evidence of systematic bias.
- The residual plot showed a relatively symmetric distribution around zero, indicating that most predictions were close to the actual values.

Below are the five largest residuals (most significant errors):

Index	Actual	Predicted	Residual
9338	480.0	544.82	-64.82
4216	533.7	597.21	-63.51
10914	555.8	619.26	-63.46
5828	598.3	660.92	-62.62
3857	566.2	628.65	-62.45

Analyze two specific failure cases:

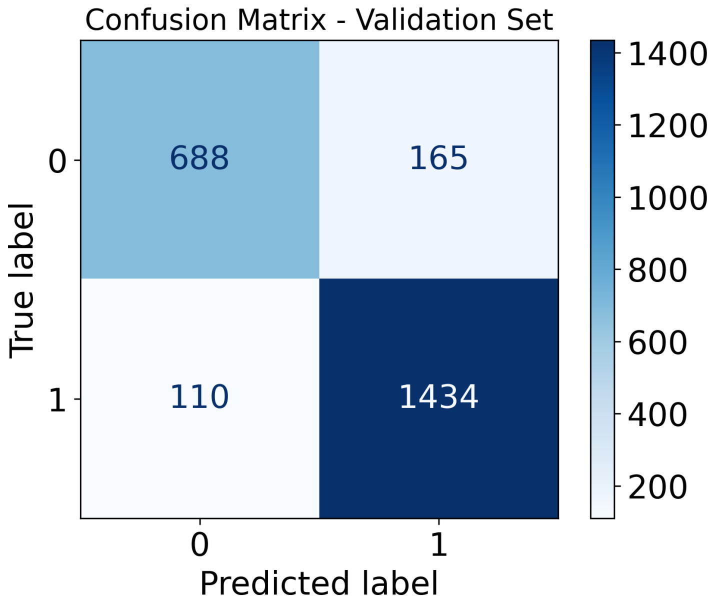
Case 1 (Index 9338): The model predicted a score of 544.82 while the actual value was 480.0, resulting in an error of 64.82. This may have been caused by outliers in the input features or a profile that was underrepresented in the training set. It's possible that the individual had an unusually low quality-of-life score compared to others with similar features, making the prediction more difficult.

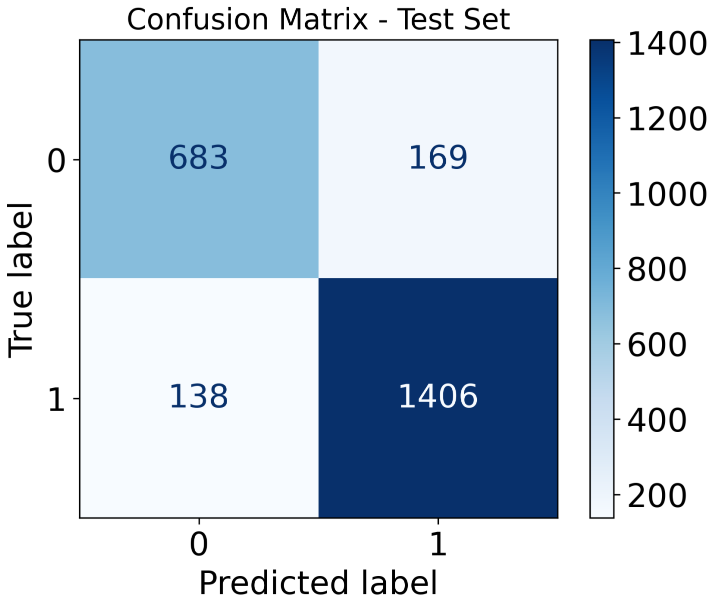
Case 2 (Index 10914): The predicted value was 619.26, while the actual was 555.8, resulting in a residual of 63.46. This discrepancy might be due to missing or incomplete data in some influential variables or a case where the model failed to capture non-linear relationships relevant to the observation.

Identified causes of Failure Modes:

1. **Missing non-linear effects:** Some features may influence quality of life in a non-linear manner or through interactions which a simple linear model cannot capture.
2. **Extreme leverage points:** A few observations with very high or low values in features can pull the regression line away from the main data cloud, resulting in large residuals for similar cases.

Classification (Random Forest Classifier Model):





Classification Error Analysis and Interpretation

We analyzed the model's misclassifications:

- **False Positives:** 165 instances
- **False Negatives:** 110 instances

Why Does the Best Performing Model Make Mistakes?

Although the Random Forest Classifier demonstrated strong overall performance, the error analysis reveals that the model makes systematic mistakes. Specifically, it produced:

- **165 False Positives:** Instances where the model incorrectly predicted "1" (good work-life balance), but the true label was "0".
- **110 False Negatives:** Instances where the model predicted "0" (poor balance), but the true label was actually "1".

These errors suggest that the model may be overfitting to certain dominant features or failing to capture subtle but important signals in the data related to personal well-being.

Confusion Matrix and Error Types

From the reported errors, it appears that the most frequent error type is **False Positives**, indicating a bias toward overestimating work-life balance, This is potentially problematic in real-world applications where misidentifying someone as having a good quality of life could lead to neglecting necessary support. However, **False Negatives** are also concerning, as they indicate that individuals with good work-life balance are not being recognized by the model, reducing overall precision and recall.

Two Specific Examples and Hypotheses

1. False Positive – Row ID 9706

- Prediction: 1 (Good balance)
- Actual: 0 (Poor balance)
- **Feature values:**
  - TODO_COMPLETED: 7
  - CORE_CIRCLE: 5
  - LIVE_VISION: 5
  - DONATION: 1
  - TIME_FOR_PASSION: 3
  - PERSONAL_AWARDS: 5
- **Hypothesis:** The model likely focused on high values in features such as TODO_COMPLETED, CORE_CIRCLE, and LIVE_VISION, which generally suggest a structured and connected lifestyle. However, low values in DONATION and TIME_FOR_PASSION may indicate low emotional engagement or personal fulfillment, which the model underweighted.

2. **False Negative – Row ID 13465**
   - Prediction: 0 (Poor balance)
   - Actual: 1 (Good balance)
   - **Feature values:**
     - PERSONAL_AWARDS: 7
     - TIME_FOR_PASSION: 4
     - DONATION: 3
     - CORE_CIRCLE: 1
     - FLOW: 2
     - LIVE_VISION: 2
   - **Hypothesis:** Despite a strong personal achievement score (PERSONAL_AWARDS = 7) and balanced lifestyle indicators (TIME_FOR_PASSION = 4, DONATION = 3), the model might have given excessive weight to low values in CORE_CIRCLE, FLOW, and LIVE_VISION, This suggests a possible bias toward social or emotional connectivity features, leading to underprediction of positive cases when those are low.

5. Interpretation and Implications

Summary of Evaluation Results in Relation to the Original Research Question:
Research Question: The goal was to predict quality of life / work-life balance (WORK_LIFE_BALANCE_SCORE) based on lifestyle variables, using both a regression model (continuous target) and a binary classification model (TARGET_BINARY).
Does the best model meet the performance requirements? Why or why not?
Yes, the two leading models (one for each task type) meet the performance requirements:

- **Best regression model:** Linear Regression, with an R^2 score of 0.824, MAE of 15.13, and RMSE of 18.99. It is accurate, shows no heteroscedasticity (non-constant residual variance), and no systematic bias the residuals are symmetrically distributed around zero.
- **Best classification model:** Random Forest Classifier, with an accuracy of 0.885 and an F1-score of 0.89, which is a very high result, especially for predicting quality of life categories.

Both models also passed statistical significance tests confirming that they perform significantly better than the other models evaluated.

Trade-offs Stakeholders Must Consider

- **Accuracy vs Explainability:** The most accurate models (XGBoost and Random Forest) are less transparent and harder to interpret. Simpler models like Linear Regression are easier to understand but less flexible in handling complex or nonlinear phenomena.
- **Cost vs Benefit:** Collecting many lifestyle variables requires resources. The more complex the model, the higher the maintenance and implementation costs.

Suggestions for Future Improvement

- **Advanced Feature Engineering:** There is a need to create new variables that combine values, such as a sense of fulfillment, to better capture the complexity of quality of life.
- **Use of Robust Regression Models:** Employ robust regressors like Huber Regression or remove extreme outliers using methods such as Isolation Forest to reduce the influence of outliers on the regression line.
- **Interpretability Tools:** Use interpretation tools like SHAP to understand which variables truly contribute to predictions and adjust the model accordingly.

Final Report & GitHub Submission Assignment (5)

1. Conclusion

The goal of this study was to examine which lifestyle characteristics most significantly impact an individual's quality of life, as measured by the work-life balance score. The dataset consisted of 15,972 observations and initially included 24 features. We adopted a mixed-methods approach both quantitative and qualitative to analyze the data. Categorical features, such as gender and age group, were encoded into numerical values using appropriate encoding techniques. It is important to note that the dataset did not contain any missing values, and therefore no handling of NaN values or imputation was necessary. Once all features were numeric, we calculated their correlation with the target variable. To reduce dimensionality and focus on the most relevant variables, we selected 11 features with a Pearson correlation coefficient of 0.45 or higher. These features formed the basis for our predictive modeling. For regression tasks, we applied standardization to the selected features, and for classification tasks, we binarized the target variable into 0 and 1. We trained four machine learning models: two for regression - Linear Regression and Random Forest Regressor, and two for classification - XGBoost Classifier and Random Forest Classifier. The dataset was split into three subsets: training (70%), validation (15%), and testing (15%). After the initial evaluation, we applied hyperparameter tuning using cross-validation to optimize model performance. In the regression task, Linear Regression achieved the best results, with an R^2 of 0.824 and MAE of 15.13. The Random Forest Regressor performed slightly worse ($R^2 = 0.814$, MAE = 15.50). A statistically significant difference was found between the two regression models. In the classification task, Random Forest Classifier achieved the best overall performance, with an accuracy of 0.885 and F1-score of 0.89, as well as strong recall and precision values. The XGBoost Classifier also performed well (accuracy: 0.871, F1: 0.87), but did not outperform the Random Forest model. A statistically significant difference was also found between the two classification models. An analysis of the class distribution in the classification task showed no significant class imbalance: 5,682 observations were labeled as 0 and 10,290 as 1, yielding a ratio of 1.81. Therefore, no special treatment for class imbalance was required. These findings offer valuable insights into how personal lifestyle habits influence work-life balance, with practical implications for individuals, organizations, and public health policy.

2. Limitations

Despite the strong performance of the models we developed, this project includes several key limitations that should be acknowledged:

1. **Model Complexity and Interpretability:**
   The classification model with the best performance was the Random Forest Classifier, which is based on numerous decision trees. Although highly accurate, its decision-making process lacks transparency. For example, when the model predicts that an individual has a good work-life balance, it is difficult to identify which features led to that outcome unlike a linear regression model, where each feature's influence is clearly shown. This lack of explainability may limit the practical use of the model in fields where transparency, trust, or interpretability are essential, such as healthcare, human resources, or personal well-being. Future work could include interpretability tools such as SHAP or LIME to better understand feature contributions.

2. **Failure to Capture Non-Linear Relationships and Interactions in Regression:**
   The main regression model used in this project was Linear Regression, which assumes independent and linear
   relationships between features and the target variable. However, in reality, lifestyle factors influence quality of life in more complex, non-linear, and interdependent ways. This limitation was evident in specific failure cases
   (observations 9338 and 10914), where large prediction errors occurred, possibly due to the model's inability to handle unique feature combinations or outliers. In future projects, non-linear models or robust regression methods (such as Huber Regression) could be explored, as well as adding polynomial or interaction terms.

3. **Bias Toward Dominant Features in Classification:**
   Error analysis revealed that the classification model especially the Random Forest tends to over-rely on specific
   features such as PERSONAL_AWARDS and FLOW, even when other important indicators. (such as TIME_FOR_PASSION or LIVE_VISION) are low, this led to misclassifications in borderline cases, where some features were high while others were low. More balanced feature weighting or advanced feature engineering may help reduce this bias and improve the model's ability to interpret complex profiles.

4. **Lack of Feature Interaction Modeling:**
   Feature selection was based on correlation with the target variable, but a high correlation does not necessarily indicate predictive value. Moreover, the models did not include interactions between features situations where the effect of one variable depends on another. For example, "sense of purpose" may influence work-life balance more when combined with "free time." The absence of such combinations may reduce the model's ability to detect deeper behavioral patterns. Future models could include engineered interaction features or use algorithms that capture such effects automatically (Gradient Boosting or Neural Networks).

5. **Data Quality and Limited Population Representation:**
   Although the dataset was complete and contained no missing values, there are still important limitations related to data quality. First, all data came from a single source and time period, which may mean it does not fully represent the diversity of populations, cultures, or time-based variations. For example, the lifestyle of young adults in one country may differ significantly from that of older adults in another but these differences may not be reflected in the dataset. Additionally, unique or rare lifestyle profiles may have been underrepresented in the training set, making it harder for the model to generalize to such individuals. Furthermore, several variables in the dataset are based on
   self-reported questionnaires, which are inherently subjective and may include perceptual bias affecting prediction quality. Future work could improve data quality by integrating multiple data sources, using longitudinal data collection, or segmenting populations for more tailored analysis.

3. Future Work

Building on our findings and the limitations identified, several concrete directions can be pursued to deepen and extend this research:

1. **Enhance Data Diversity and Quality:**

   - **Integrate Multiple Data Sources:** Collect additional lifestyle and well-being data from different regions, age groups, and cultural contexts to improve representativeness and generalisability.
   - **Longitudinal Data Collection:** Design a follow-up survey to capture changes in work-life balance over time, enabling analysis of causal dynamics rather than cross-sectional correlations.

2. **Develop and Evaluate Non-Linear and Interaction-Aware Models:**

   - **Advanced Algorithms:** Experiment with models capable of capturing non-linear relationships and feature interactions, such as Gradient Boosting Machines, Neural Networks, or Support Vector Regression.
   - **Feature Engineering for Interactions:** Create engineered features representing key interactions ("sense of purpose" × "free time") to explicitly test their predictive utility in both regression and classification tasks.

3. **Improve Model Interpretability and Fairness:**

   - **Interpretability Tools:** Integrate SHAP, LIME, or other explainability frameworks to quantify each feature's contribution and flag potential biases in individual predictions.
   - **Fairness Audits:** Conduct subgroup analyses (by gender, age, income) to assess whether model
     performance differs across demographic groups, and apply fairness-enhancing techniques (reweighting, adversarial debiasing) as needed.

4. **Enhancing Model Robustness to Outliers and Rare Profiles:**

   - **Robust Regression Techniques:** Future work could explore the use of models such as Huber Regression or RANSAC, as well as outlier detection methods like Isolation Forest, to reduce the influence of extreme observations on model performance.
   - **Rare Profile Enrichment:** In cases where certain population groups are underrepresented in the training data for example, individuals with unique lifestyle combinations the model may fail to learn effectively from them and produce inaccurate predictions. To address this imbalance, synthetic data generation techniques such as SMOTE (Synthetic Minority Over-sampling Technique) can be applied. SMOTE creates new synthetic samples based on existing rare examples, allowing for a more balanced training set and improving the model's ability to learn from and generalize to uncommon or atypical profiles.

5. **Real-World Implementation and Impact Evaluation:**

   - **Development of an Interactive Dashboard:** As part of the model's practical potential, an interactive dashboard can be developed to allow individual users or HR professionals to input personal lifestyle data such as average sleep hours, time for hobbies, sense of purpose, physical activity, and more, receive a
     personalized prediction of their Work-Life Balance Score, along with tailored recommendations for
     improvement. Such an interface would transform the model's output into a practical and user-friendly tool that can promote awareness, guide decision-making, and support both individual and organizational
     well-being. This type of implementation bridges the gap between the research model and real-world
     applications, enabling measurable impact on users' daily lives.
   - **A/B Testing in Workplace Settings:** It is recommended to conduct a controlled pilot within a small organization, where participants are randomly divided into two groups: One group receives personalized recommendations based on model predictions. The other group receives no intervention. After a set period, the two groups can be compared on various well-being indicators such as satisfaction, perceived balance, stress levels, or productivity. The goal of this testing is to assess whether the model's real-life use leads to measurable improvements in work-life balance, beyond its statistical performance on data. If results are positive, this would support broader deployment of the model in organizational or personal wellness settings.

References

[1] Youngkeun Choi. A prediction of work-life balance using machine learning. *Asia pacific journal of information systems*, 34(1):209–225, 2024.

[2] Imen Moussa-Chamari, Abdulaziz Farooq, Mohamed Romdhani, Jad Adrian Washif, Ummukulthoum Bakare, Mai Helmy, Ramzi A Al-Horani, Paul Salamh, Nicolas Robin, and Olivier Hue. The relationship between quality of life, sleep quality, mental health, and physical activity in an international sample of college students: a structural equation modeling approach. *Frontiers in Public Health*, 12:1397924, 2024.

[3] Chara Papoutsi, Athanasios Drigas, and Charalabos Skianis. Serious games for emotional intelligence's skills development for inner balance and quality of life-a literature review (juegos serios para el desarrollo de habilidades de la inteligencia emocional para el equilibrio interior y la calidad de vida-una revisión de la literatura). *Retos*, 46:199–208, 2022.

[4] Marina Ródenas-Munar, Margalida Monserrat-Mesquida, Santiago F Gómez, Julia Wärnberg, María Medrano, Marcela González-Gross, Narcís Gusi, Susana Aznar, Elena Marín-Cascales, Miguel A González-Valeiro, et al. Perceived quality of life is related to a healthy lifestyle and related outcomes in spanish children and adolescents: The physical activity, sedentarism, and obesity in spanish study. *Nutrients*, 15(24):5125, 2023.

[5] Shi-chen Zhang, Fang-biao Tao, Atsushi Ueda, Chang-nian Wei, and Jun Fang. The influence of health-promoting lifestyles on the quality of life of retired workers in a medium-sized city of northeastern china. *Environmental health and preventive medicine*, 18(6):458–465, 2013.