NOVA IMS Information Management School

MACHINE LEARNING

WHO-LIFE

A GLOBAL INITIATIVE TO UNDERSTAND LIFESTYLES WORLDWIDE

GROUP 7

António Santos 20221966
Ashool Lakhani 20221971
Francisco Oliveira 20222049
Tara Kouros 20221970

ABSTRACT

This report is part of the WHO-LIFE initiative, which seeks to understand how people live across different countries by analyzing individual-level data. The main objective of our project was to determine whether machine learning techniques can help identify and predict lifestyle categories based on health, financial, and behavioral features.

To achieve this, we began with thorough data exploration and cleaning, handling missing values, transforming skewed distributions, and engineering meaningful features. For deeper insight, we also integrated external indicators such as GDP per capita and population size to contextualize entertainment and wellness patterns across regions.

We approached the problem in two main stages: descriptive and predictive modeling. In the descriptive phase, clustering was applied using tailored profiles (health, financial, and lifestyle), each defined by relevant features. The lifestyle profile, combining both health and financial traits, produced the most balanced and informative segments, highlighting distinct behaviors such as wellness-oriented individuals, digitally engaged users, and fitness-driven lifestyles.

For predictive modeling, we trained multiple classification models using a rigorous pipeline that included feature selection, dummy variable creation, and 10-fold stratified cross-validation. Hyperparameters were optimized using GridSearchCV. Among the tested models, Gradient Boosting achieved the best performance with minimal overfitting, making it our final choice for predicting lifestyle clusters.

Our findings demonstrate that machine learning can successfully segment and predict lifestyle patterns with high reliability. These insights not only contribute to global health understanding but also support WHO in designing targeted interventions and policies.

KEYWORDS

Machine Learning; Data Preprocessing; Lifestyle Segmentation; Descriptive Modeling; Predictive Modeling

Table of Contents

INTRODUCTION

This project is part of the WHO-LIFE initiative, which aims to understand how people live across different parts of the world. The dataset we worked with contains information about individual citizens, including health behaviors, financial habits, environmental awareness, and overall well-being. By analyzing this data, we try to understand what defines different lifestyles and to apply data science and machine learning techniques not just to explore the data, but also to surface actionable insights that reflect real-world behaviors.

The main question we aim to answer is "*Can machine learning techniques help us identify and predict different lifestyle categories based on individual-level features?"* To address this, we have structured our project into several key phases:

1. **Data Exploration and Understanding:**
   We begin with exploratory data analysis (EDA) to get an overview of the dataset, detect any patterns or anomalies, and better understand the relationships between variables.
2. **Data Preprocessing:**
   This phase involves cleaning the data, handling missing values, standardizing features, and encoding categorical/numerical variables to prepare the dataset for modeling.
3. **Additional Insights:**
   We investigate specific behavioral questions, such as whether eco-conscious individuals are more financially conservative, or if well-being is more strongly linked to exercise or stress management. We also integrate external data (e.g., GDP per capita) to add context.
4. **Descriptive Modeling (Clustering):**
   Using unsupervised learning methods like KMeans, we group individuals into lifestyle segments based on similar characteristics.
5. **Predictive Modeling (Classification):**
   Finally, we use supervised learning to predict lifestyle categories for new citizens based on their features, testing various classification algorithms, and evaluating their performance.

DATA EXPLORATION

We began our analysis by exploring the structure and distribution of the dataset. The data contains 22 columns and 8,327 rows, covering a range of features related to health, finances, lifestyle habits, and well-being. As shown in Figure 1, several columns contain missing values such as in *avg_weekly_exercise_hours*, *investment_portfolio_value*, and *stress_management_score*. These gaps highlighted the need for appropriate imputation strategies.

```
RangeIndex: 8327 entries, 0 to 8326
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   avg_monthly_entertainment_expenses  8077 non-null   object
 1   avg_weekly_exercise_hours           8119 non-null   object
 2   citizen_id                          8327 non-null   int64
 3   city                                8327 non-null   object
 4   country                             8327 non-null   object
 5   date_of_birth                       8327 non-null   object
 6   eco_consciousness_score             8119 non-null   object
 7   education_level                     8036 non-null   object
 8   environmental_awareness_rating      8294 non-null   object
 9   financial_wellness_index            8078 non-null   object
 10  hapiness_level                      8327 non-null   object
 11  health_consciousness_rating         8077 non-null   object
 12  investment_portfolio_value          8202 non-null   object
 13  investments_risk_appetite           8077 non-null   object
 14  investments_risk_tolerance          8160 non-null   object
 15  last_year_charity_donations         8202 non-null   object
 16  marital_status                      8327 non-null   object
 17  name                                8327 non-null   object
 18  social_media_influence_score        8327 non-null   object
 19  stress_management_score             8034 non-null   object
 20  title                               8327 non-null   object
 21  well_being_level                    8202 non-null   object
dtypes: int64(1), object(21)
```

Figure 1 – Basic Dataset Information

Additionally, Figure 2 shows that some columns contain negative values (e.g., in exercise) and many features exhibit skewed distributions, especially variables like *avg_monthly_entertainment_expenses* and *environmental_awareness_rating*.

	avg_monthly_entertainment_expenses	avg_weekly_exercise_hours	citizen_id	city	country	date_of_birth	eco_consciousness_score	education_level	environmental_awareness_rating	f
count	8077.000000	8119.000000	8327.000000	8327	8327	8327	8119.000000	8036.000000	8294.000000	
unique	NaN	NaN	NaN	3733	22	5711	NaN	NaN	NaN	
top	NaN	NaN	NaN	Tauranga	Spain	1996-02-18	NaN	NaN	NaN	
freq	NaN	NaN	NaN	24	446	6	NaN	NaN	NaN	
mean	21.590060	0.707644	4164.000000	NaN	NaN	NaN	0.131725	6.282893	0.827204	
std	17.409831	1.186189	2403.942179	NaN	NaN	NaN	0.358699	2.918705	0.650630	
min	0.000000	-5.579400	1.000000	NaN	NaN	NaN	-1.416700	0.000000	0.037600	
25%	8.128300	0.022400	2082.500000	NaN	NaN	NaN	0.000000	4.031225	0.540925	
50%	11.983100	0.109300	4164.000000	NaN	NaN	NaN	0.007800	6.344650	0.631200	
75%	38.757300	1.023750	6245.500000	NaN	NaN	NaN	0.024450	8.559725	0.774475	
max	135.420000	8.701700	8327.000000	NaN	NaN	NaN	3.247500	14.244000	7.402500	

Figure 2 – Dataset Summary Statistics

Lastly as shown in Figure 3**,** we grouped the dataset features into numerical, categorical, datetime, and identifier types. This separation allowed us to apply appropriate preprocessing techniques for each type, such as scaling for numerical features and encoding for categorical ones.

```
numerical_features=["last_year_charity_donations","financial_wellness_index","investment_portfolio_value","social_media_influence_score","investments_risk_appetite","investments_risk_tolerance",
                    "avg_monthly_entertainment_expenses","avg_weekly_exercise_hours","stress_management_score","eco_consciousness_score","well_being_level","environmental_awareness_rating","health_consciousness_rating","education_level"]
categorical_features=["name","title","city","country","hapiness_level","marital_status"]
datetime_features=["date_of_birth"]
id_features=["citizen_id"]
```

Figure 3 – Feature Type Classification

These observations led us to define our preprocessing steps to handle missing values through imputation, address outliers using robust techniques, and apply appropriate scaling methods to normalize skewed features.

DATA PREPROCESSING

After exploring the dataset and identifying key issues, we applied preprocessing in two parts. One for numerical features and other for categorical. Then our first step was identify any duplicates and removing them. This was to ensure the data was clean, consistent, and ready for modeling.

CATEGORICAL FEATURES

For the *country* column, we first handled missing values by checking for blank entries. We then created a mapping dictionary to assign countries to any remaining cities with missing or inconsistent values. To enhance regional analysis, we introduced a new *continent* column and correctly assigned each country to its respective continent, which provided a clearer geographic structure for our additional analysis.

As part of feature engineering, we extracted *age* from the *date_of_birth* column to turn it into a usable numerical feature. We also derived a *gender* variable from the *title* column and cleaned the *marital_status* field by grouping similar categories together. These features provided clearer and more consistent information, which helped improve both the clustering and predictive modeling phases.

Then we moved to *education_level* and *financial_wellness_index*, which we initially classified as numerical features. However, to improve interpretability in our clustering and analysis, we created categorical versions of both. For education_level, first we fill the missing value using medians and then we grouped the values into broader categories like low, medium, and high education (refer Table 1).

Education Level (Numeric)	Category
≤ 1	No formal education
≤ 4	Primary School
≤ 8	Middle School
≤ 10	High School
≤ 12	Bachelor's
≤ 15	Master's or Higher

Table 1 – Education Level Classification

Similarly, we filled the missing values for *financial_wellness_index* with random sample from the distribution to avoid unusual spikes and then we binned the column into levels such low, medium, and high (refer Table 2).

Financial Wellness Index (Numeric)	Category
≤ 70	Low
71–130	Medium
> 130	High

Table 2 – Financial Wellness Index Classification

Lastly, as seen in Figure 4 and careful examination, we decided to drop the *name*, *city*, and *hapiness_level* columns in our preprocessing, as these features contained either too many unique values or lacked meaningful patterns for our analysis.

	count	unique	top	freq
name	6659	3105	Emma	23
title	6659	4	Mr.	3222
city	6659	3237	Whangarei	20
country	6659	22	Spain	355
hapiness_level	6659	1	medium	6659
marital_status	6659	1		6659

Figure 4 – Categorical Features Statistics

NUMERICAL FEATURES

We knew from our initial exploration that many numerical features had missing values, were highly skewed, and had outlier issues (refer Figure 5).

	count	mean	std	min	25%	50%	75%	max
last_year_charity_donations	8048.0	0.381591	0.765184	0.0000	0.000000	0.01040	0.310425	5.5991
financial_wellness_index	8172.0	106.234129	64.189972	0.0000	87.132400	99.81570	113.754750	458.4510
investment_portfolio_value	8050.0	12.420477	21.231746	0.0000	3.294850	8.60450	13.264575	294.9500
social_media_influence_score	8172.0	6.099268	4.325992	0.0000	3.166325	4.40060	8.204975	42.1768
investments_risk_appetite	7925.0	4.377007	2.646153	0.0000	1.853800	4.81390	6.874600	8.6995
investments_risk_tolerance	8012.0	7.308825	3.679396	1.3937	4.796950	6.22060	8.543125	29.5132
avg_monthly_entertainment_expenses	7927.0	21.596337	17.430907	0.0000	8.128350	11.98960	38.777350	135.4200
avg_weekly_exercise_hours	7965.0	0.705633	1.182336	-5.5794	0.022400	0.10950	1.023500	8.7017
stress_management_score	7887.0	3.318718	1.114644	0.3192	2.560100	3.15010	3.979250	8.7123
eco_consciousness_score	7966.0	0.131752	0.358680	-1.4167	0.000000	0.00780	0.024475	3.2475
well_being_level	8050.0	4.764525	1.934727	1.1376	3.413150	4.31915	5.614075	14.1143
environmental_awareness_rating	8140.0	0.827895	0.651972	0.0376	0.540800	0.63120	0.774850	7.4025
health_consciousness_rating	7933.0	1.451597	1.395930	0.0000	0.685800	0.80620	1.761200	10.0530
education_level	8172.0	6.291799	2.868611	0.0000	4.122275	6.35960	8.474700	14.2440

Figure 5 – Numerical features statistics

To better understand the relationships between features and guide our imputation strategy, we first created a Spearman correlation heatmap. This allowed us to capture both linear and non-linear associations between variables.

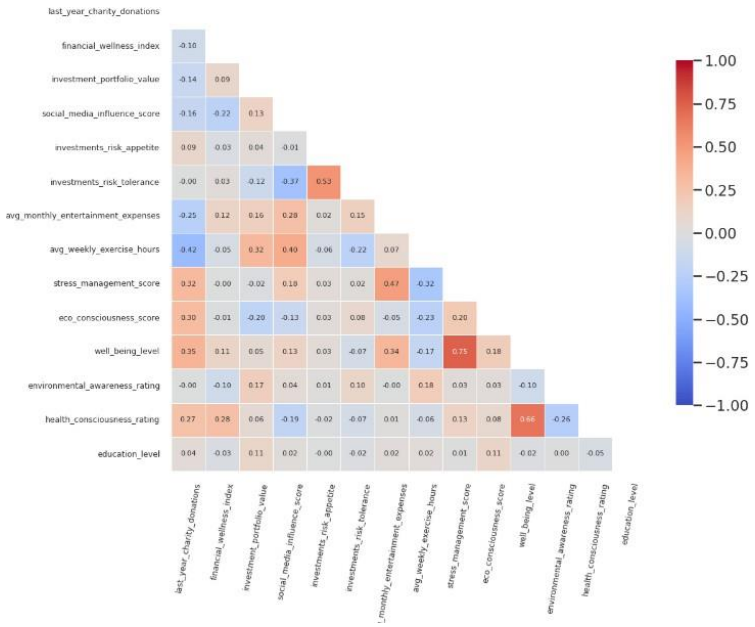

Figure 6 – Spearman Correlation Heatmap

Based on our results we saw from the heatmap there is strong correlation between *well_being_level* vs *stress_management_score*, *well_being_level* vs *health_consciousness_rating*, and *investments_risk_appetite* vs *risk_tolerance*. So, we plot them as scatterplot to further understand any potential trends, clusters, and outliers.

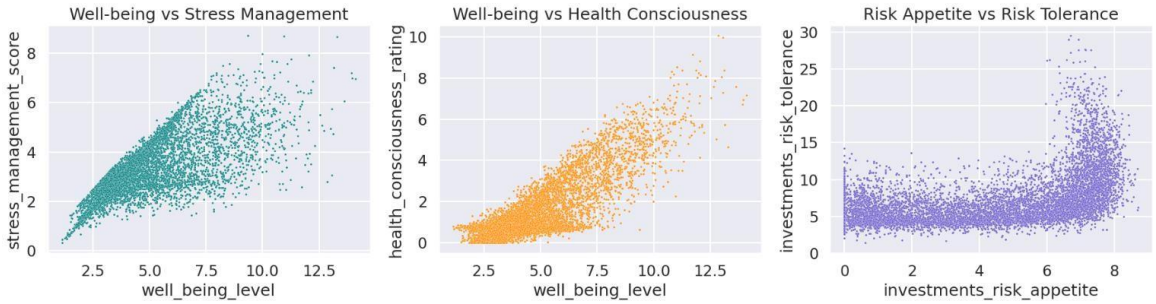

Figure 7 – Scatter Plots of Correlated Features

Based on the relationships observed in the Spearman correlation heatmap (Figure 6) and the scatter plots (Figure 7), we identified strong positive correlations between several key features. For example, *stress_management_score* showed a clear relationship with *well_being_level*, so we filled its missing values using the median, grouped by *well_being_level*. A similar approach was applied to *investments_risk_tolerance*, which was imputed based on *investments_risk_appetite*. While, for the *well_being_level* column, we applied a KNN imputer, using features like *stress_management_score* and *health_consciousness_rating* as reference points. Since these features were strongly correlated, this method provided more accurate estimates.

The remaining numerical columns with missing values were filled using their respective median values, as they did not show strong enough correlations to justify more complex imputation methods.

Finally, to address skewed distributions and reduce the impact of outliers, we applied a combination of RobustScaler, log transformation (log1p), and quantile clipping on the numerical features. These techniques helped normalize the data without removing valuable observations and the results can be seen in Figure 8.

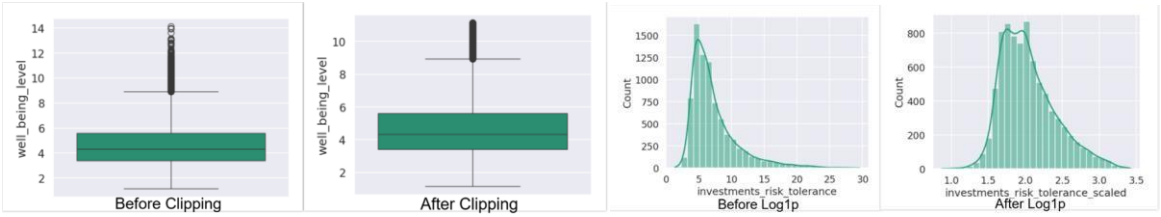

Figure 8 – Handling Skewness and Outliers in Numerical Features

In the final preprocessing step, we dropped *eco_consciousness_score* and *last_year_charity_donations*.These features contained mostly zero values, some unexpected negatives, and showed little correlation with other features, making them unhelpful for our analysis or modeling.

METHODOLOGY

This section outlines the procedures and decisions made across all phases of the project. Our approach is divided into three key components: Additional Insights, Descriptive Modelling, and Predictive Modelling.

ADDITIONAL INSIGHTS

We conducted analysis beyond the data exploration phase to better understand behavioral patterns. As an initial step, we compared various lifestyle trends across continents (refer to Figure 9). While this approach did reveal certain patterns, such as North America having the highest Social Media Influence, it ultimately proved to be less informative, as the overall differences were minimal, and the dataset lacked sufficient regional detail to support deeper conclusions.

Lifestyle Averages by Continent:

continent	well_being_level	avg_weekly_exercise_hours	financial_wellness_index	stress_management_score	social_media_influence_score	investments_risk_appetite	investments_risk_tolerance	health_consciousness_rating
Asia	4.84	0.72	107.40	3.33	5.92	4.51	7.32	1.50
Europe	4.73	0.69	106.30	3.30	6.13	4.33	7.26	1.41
North America	4.78	0.70	104.97	3.33	6.20	4.46	7.32	1.44
Oceania	4.80	0.68	105.65	3.31	6.04	4.35	7.21	1.46
South America	4.82	0.82	108.79	3.36	6.10	4.52	7.47	1.44

Figure 9 – Lifestyle Feature Averages by Continent

Afterward, we conducted additional correlation tests, like those performed before preprocessing, to check for any new trends or relationships in the cleaned dataset. We also decided to plot our two engineered categorical features, *financial_wellness* and *education_level*, against other key metrics to explore how these groupings impact different aspects of lifestyle behavior. This was useful in showing how our created categories relate to other metrics. As seen in Figure 10, higher education levels are associated with better stress management scores.

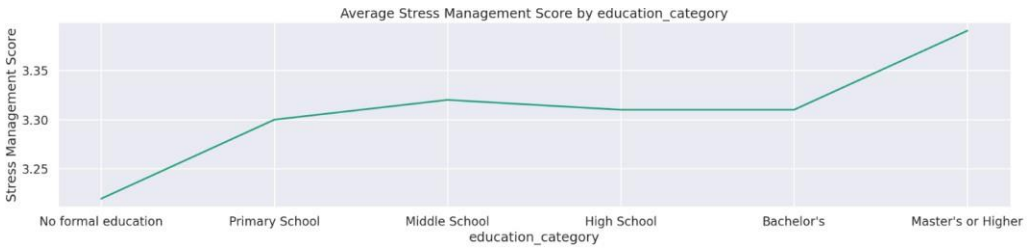

Figure 10 – Average Stress Management Score by Education Level

Our last step was adding extra data, like GDP per capita and country population, to help us get more insights from the analysis. This was useful because it helped us better understand how entertainment spending relates to both GDP per capita and population size across countries. As seen

in Figure 11 (right side), it clearly shows that higher GDP does not always correspond to higher spending.

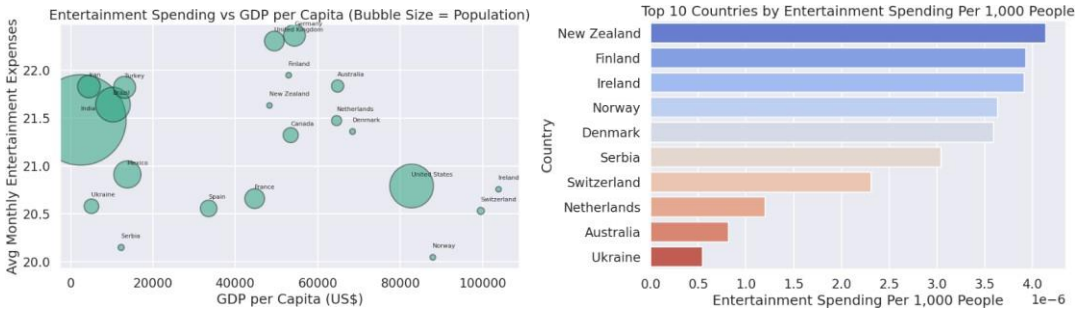

Figure 11 – Entertainment Spending vs. GDP per Capita

We also decide to normalize it abit as in Figure 11 (left side) so we could account the population difference between the countries. This comparison was also helpful, as it suggested that higher per-capita countries like New Zealand and Finland prioritize leisure and entertainment spending, despite having relatively small overall populations.

DESCRIPTIVE MODELING

After completing all preprocessing and exploratory analysis, we decided to use unsupervised learning for our descriptive modeling by adopting a profile-based clustering approach to uncover meaningful lifestyle segments.

Each profile was first assigned specific features, as outlined in Table 3, to ensure that clustering captured relevant behavioral traits. These features were then scaled using MinMaxScaler so that all variables contributed equally to the clustering process. We applied the KMeans algorithm, experimenting with different values of k and using both the elbow method and silhouette score to guide the final selection of the optimal number of clusters.

Profile	Features
Health	environmental_awareness_rating_scaled health_consciousness_rating_scaled avg_weekly_exercise_hours_scaled stress_management_score well_being_level
Financial	avg_monthly_entertainment_expenses investment_portfolio_value_scaled social_media_influence_score_scaled investments_risk_appetite investments_risk_tolerance_scaled

Table 3 – Features for Health and Financial Behavior

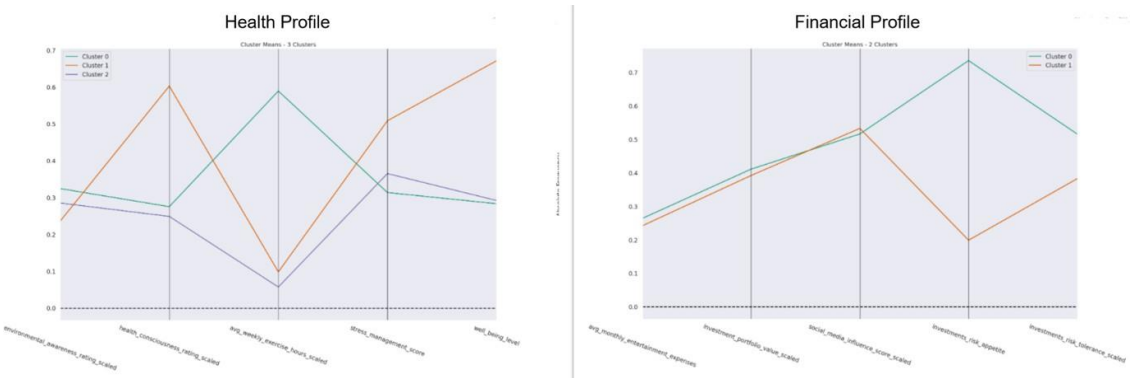

Figure 12 – Cluster Profiles for Health and Financial Behavior

The health profile revealed three distinct clusters. One group showed strong physical and mental wellness, with high exercise levels, well-being, and stress management. Another group was digitally engaged but scored low on wellness and exercise, suggesting a more entertainment- and screen-focused lifestyle. The third cluster had moderate well-being but stood out for consistent exercise and relatively balanced scores across health-related features. The financial profile identified two main groups. One cluster exhibited high investment activity, strong risk appetite and tolerance, and higher engagement in social media and spending, representing financially active individuals. The other cluster was more conservative, with lower risk scores, smaller portfolio values, and reduced entertainment and digital engagement. These differences reflect varied financial behaviors.

We initially focused on two primary profiles: health and financial. These profiles were selected based on their relevance to user behavior and the richness of available data. However, after exploring various feature combinations and evaluating we decided to introduce a third profile: lifestyle. This profile represents a hybrid category, combining the most insightful and impactful features from both the health and financial profiles (refer Table 4 for which features we used).

Profile	Features
Lifestyle	health_consciousness_rating_scaled avg_weekly_exercise_hours_scaled well_being_level social_media_influence_score_scaled

Table 4 – Features for Lifestyle profile.

PREDICTIVE MODELING

In the final phase, we aimed to predict an individual's lifestyle cluster using supervised learning techniques. After scaling and preparing the data, we split it into training and test sets (80/20 split). To prepare the data for classification, we first converted categorical variables like continent, marital_status, and gender into dummy variables, allowing machine learning models to interpret them numerically. Next, we applied SelectKBest with ANOVA F-score to rank feature importance for predicting the lifestyle clusters. This helped us retain only the most relevant variables, improving model focus and performance. To ensure all numerical features contributed equally, we applied MinMax Scaling to the final dataset. This normalized values to a [0, 1] range, which is especially important for distance-based models like KNN, helping improve consistency and model performance.

Then we applied GridSearchCV to optimize hyperparameters for multiple classification models, including Logistic Regression, K-Nearest Neighbors, Random Forest, Gradient Boosting, and Linear SVC. Each model was evaluated using 5-fold cross-validation, with the best configuration selected based on macro-averaged F1 score. This ensured we chose the most effective algorithm with optimal settings for accurate and generalizable predictions.

To ensure robust evaluation and reduce overfitting, we used Stratified K-Fold Cross-Validation with 10 folds, maintaining class balance across all splits. During each fold, the entire preprocessing pipeline was applied using only the training partition to prevent data leakage. This pipeline included:

- Imputation using training fold medians or random sampling (for financial variables),
- Outlier clipping at 0 for skewed features to reduce variance,
- Log + Robust Scaling for variables like health_consciousness_rating and environmental_awareness_rating,
- Feature engineering, creating scaled variants for key predictors like investment_portfolio_value and social_media_influence_score,
- Dropping irrelevant or duplicate columns, such as identifiers and unscaled versions of features,
- And final normalization using MinMaxScaler, fitted only on the training set and applied to both splits.

Performance was evaluated using metrics such as accuracy, macro precision, macro recall, macro F1-score, and runtime per fold, providing a thorough comparison of model effectiveness and efficiency.

Our final step was to preprocess the unseen test set using the same transformations applied during training (including cleaning the dataset, dropping irrelevant columns, and applying the same scaling). We then used our best performing model to generate predictions

RESULTS

Our analysis led to the identification of clear lifestyle patterns through both clustering and classification methods.

In the descriptive modeling phase, we created meaningful clusters based on health and financial behaviours. However, after evaluating the clustering outcomes, we combined the most informative features from both into a third, more balanced profile “lifestyle”. This cluster provided the most comprehensive segmentation, capturing patterns in health, financial habits, and social behaviour.

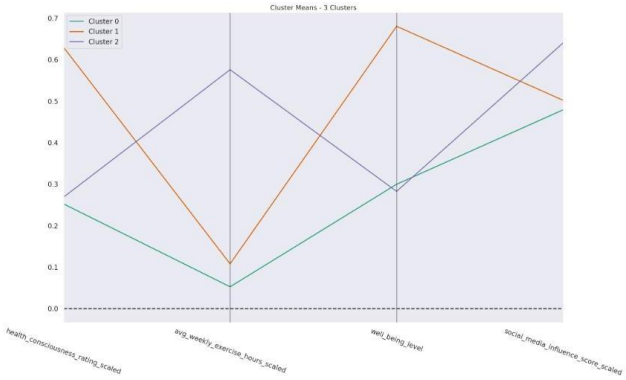

Figure 13 – Clusters for lifestyle profile

The lifestyle clustering revealed three meaningful groups that combined health and social behavior patterns. One cluster represented the less active individuals with high social media presence. Another group showed high health consciousness and well-being despite low physical activity, while the third combined high exercise levels and strong digital influence, reflecting a fitness-driven lifestyle. These patterns provided a more balanced view of real-world behaviors, which is why we chose this as our final solution.

In the predictive modeling phase, we trained and evaluated several classifiers to predict lifestyle groupings. After applying feature selection and hyperparameter tuning, the model we selected as the final model was **Gradient Boosting**. It offered a strong balance between performance and generalization, showing the lowest overfitting gap while maintaining a fairly high validation F1-score (refer Figure 14). This made it the most reliable choice for predicting lifestyle categories in our dataset.

	Train_Acc	Val_acc	Train_Precision	\
Logistic Regression	0.680717	0.678516	0.689916	
KNN Classifier	0.837290	0.706259	0.835863	
Random Forest	0.822065	0.767503	0.829482	
Gradient Boosting	0.782661	0.760658	0.786158	
Linear SVC	0.702054	0.700735	0.708987	
HistGradientBoostingClassifier	0.855557	0.770987	0.857702	

	Val_Precision	Train_Recall	Val_Recall	\
Logistic Regression	0.688216	0.681317	0.679134	
KNN Classifier	0.701583	0.837771	0.707034	
Random Forest	0.773957	0.822662	0.768251	
Gradient Boosting	0.764157	0.783324	0.761392	
Linear SVC	0.707680	0.702786	0.701470	
HistGradientBoostingClassifier	0.773081	0.856049	0.771687	

	Train_F1	Val_F1	Time
Logistic Regression	0.679392	0.677213	0.095820
KNN Classifier	0.836023	0.703061	0.076067
Random Forest	0.821121	0.765442	2.233084
Gradient Boosting	0.781394	0.758879	7.415023
Linear SVC	0.698726	0.697198	0.075212
HistGradientBoostingClassifier	0.855479	0.770279	1.063963

Figure 14 – Model Comparison Metrics for Lifestyle Prediction

ACTION PLAN

The findings from this project provide the World Health Organization (WHO) with a practical framework to understand and act on diverse lifestyle behaviors across countries and regions. Based on our clustering and predictive modeling results, we recommend the following actions:

Policy Design Based on Profiles:
The hybrid lifestyle profile that combines health and financial behaviors offers WHO a more comprehensive view of modern lifestyles. These insights can help shape policies that promote a balance between physical health, financial stability, and digital engagement.

Targeted Health Campaigns
Using the lifestyle clusters to design personalized awareness campaigns. For example, individuals in low exercise/high digital engagement clusters could be targeted with digital wellness programs or gamified physical activity apps.

Country-Specific Interventions:
By integrating GDP and population data, we found trends that suggest wealthier nations prioritize leisure differently. WHO can use this insight to tailor country-specific programs, such as promoting mental health and balanced living in high-income countries with low physical activity.

Enable Predictive Monitoring
Apply the predictive model to population data to detect shifts in lifestyle behavior over time. This enables early identification of risk-prone groups and supports real-time decision-making in fast-changing urban environments.

Promote Data-Driven Colaboration
Share the feature-based clustering and modeling framework with local health agencies and NGOs to support coordination. WHO can also invest in similar data science initiatives to support evidence-based decision-making in global health planning.

CONCLUSION

Looking back at the question we set out to explore "Can machine learning techniques help us identify and predict different lifestyle categories based on individual-level features?" We, as a team, can confidently say yes.

We approached this challenge in a structured way beginning with exploratory analysis, followed by detailed preprocessing, feature engineering, and finally applying both clustering and predictive modeling. Our goal was simple and straight, we were trying to identify lifestyle patterns s were meaningful and useful from different perspectives.

Throughout this project, we experimented with different clustering strategies, designed multiple profiles, train, and validated our predictive models using techniques like cross-validation. In the end, we decided to choose the results which we believed were most reliable.

This project allowed us to combine our technical knowledge critical thinking, making us understand of how data science can support public health research. It also made us more aware of the importance of clean data, proper feature selection, and thoughtful interpretation when dealing with complex datasets.

Overall, this project helped us improve our machine learning skills and showed us how data can be used to better understand people’s lifestyles and support various important decisions.

REFERENCES

- World Bank. (2023). *Total Population per Country*. Retrieved from https://data.worldbank.org/indicator/SP.POP.TOTL
- World Bank. (2023). *GDP per capita (current US$)*. Retrieved from https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. Retrieved from https://scikit-learn.org/stable/

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa