NOVA IMS
Information Management School

MARCH, 2025

WHO-LIFE

A GLOBAL INITIATIVE TO UNDERSTAND LIFESTYLES WORLDWIDE

GROUP PROJECT
MACHINE LEARNING 2024/2025

01 I. INTRODUCTION

In an era of rapid globalization, urbanization, and shifting societal values, understanding how people structure their daily lives has become more critical than ever. Recognizing the importance of these insights, the World Health Organization (WHO) has launched WHO-LIFE, a global research initiative aimed at analyzing and categorizing **lifestyle patterns** across different countries and regions.

From the wellness-driven routines of Scandinavian countries to the fast-paced urban lifestyles of major metropolitan hubs, people's choices are influenced by health, economics, culture, and social structures. But what are the dominant lifestyle patterns worldwide? And how do factors like income, work-life balance and financial planning shape how people live?

II. PROJECT GOALS

The primary components and goals of the project are:

1. **Preprocessing and Exploratory Data Analysis (EDA):** Conduct a through data exploration to understand the dataset's main characteristics and prepare the data for effective modeling.
2. **Additional Insights:** True innovation often comes from exploring beyond predefined methods. Think critically and creatively by investigating aspects of the dataset that could lead to **meaningful discoveries with real impact.** You are given the flexibility to choose your own approach, as long as the analysis provides **valuable insights into global lifestyles**. This could include:
    a. *Comparing lifestyle trends across the world*—Are there lifestyles that are predominant in certain regions/continents? What trends can you find?
    b. *Investigating citizen behaviors*—Do eco-conscious individuals exhibit different spending or wellness habits? Is there a link between financial stability and a fitness-conscious lifestyle?
    c. *New data* — Integrate your data with data from other sources to generate new insights.
3. **Descriptive Modeling:** In this phase, you will **apply unsupervised learning to uncover natural lifestyle segments** within the population. Using clustering techniques, the objective is to identify and characterize distinct groups based on behaviors and characteristics, effectively defining different lifestyle choices. Your focus is on segmenting the population, analyzing the key features that define each group, and understanding what differentiates them. Based on these findings, you need to describe each lifestyle and determine the defining traits of the individuals within them.

02

4. Predictive Modeling: Based on the findings from the previous section, the WHO has created a **lifestyles** variable that categorizes each citizen's way of living (target). Your goal is to develop a **classification model to accurately predict which lifestyle each citizen belongs to**.
To achieve this, you will need to identify and analyze the most significant predictors of different lifestyles and apply machine learning algorithms to classify individuals based on the provided dataset.
5. Kaggle Competition: Teams can submit multiple predictions on Kaggle, with the scoreboard ranking submissions based on **F1 Score 'macro'**. Before the competition ends, you must choose one submission to compete.
6. Action Plan & Critical Insight: With the completion of your work, it is essential to **reflect on the key findings and assess their broader impact**. This section should summarize the most important insights uncovered throughout the project and critically evaluate their implications, highlighting the most actionable takeaways for the World Health Organization (WHO) into an action plan.

III. DATASET

You have access to two different datasets:

- *world_citizens.csv*: Containing the data you will use throughout the project to train you models. This corresponds to your **training set**, where you will find detailed information about each citizen. The goal is to apply the models you have created to make predictions on unseen data (i.e., the test set). **Important note: You should not consider the target variable as feature for any of the descriptive and predictive models.**
- *kaggle_test_citizens.csv*: Containing the data you will use for the Kaggle competition. This corresponds to your **test set**, where you have access to the same attributes, but the target variable, which you are trying to predict, is not be available.

The available data contains the following attributes:

ATTRIBUTE	DESCRIPTION
citizen_id	Unique identifier of the citizen.
name	First name of each citizen.
title	Title of each citizen.
date_of_birth	Date of birth of each citizen.
city	Name of citizen´s city.

03

ATTRIBUTE	DESCRIPTION
country	Name of citizen´s country.
last_year_charity_donations	The charitable donations made by each citizen in the last year, in in thousands of dollars ($).
environmental_awareness_rating	A rating [0, 10] of each individual's awareness of and engagement with environmental issues.
financial_wellness_index	An index indicating each citizen´s overall financial health.
investment_portfolio_value	The value, in thousands of dollars ($), of each citizen´s investment portfolio.
investments_risk_appetite	A measure of each individual's willingness to take risks in their investments.
investments_risk_tolerance	A measure of each individual's tolerance for risk in investment choices.
hapiness_level	A rank indicating how happy each citizen feels.
social_media_influence_ score	A score representing each citizen´s influence and activity on social media platforms.
marital_status	Marital Status of each citizen.
avg_monthly_entertainment_expenses	The monthly expenditure on entertainment for each citizen, in dollars ($)
avg_weekly_exercise_hours	The average number of hours each citizen spends on exercise weekly.
health_consciousness_rating	A rating [0, 10] of each citizen´s awareness and proactive behavior towards their health.
education_level	The highest level of education attained by each individual.
stress_management_score	A score indicating how effectively each citizen manages stress.
eco_consciousness_score	A metric evaluating each person's consciousness and actions towards ecological sustainability.
well_being_level	A score indicating each citizen overall well being.
lifestyle	A categorization of the predominant lifestyle choice for each citizen **(Target Variable)**.

04

IV. DELIVERABLES

Upon the project's deadline, you will be required to submit:

- A *Jupyter* **notebook (or zip file) containing all the code** used throughout the project **and any extra data you may have used**. Make sure to include detailed markdown cells and comments within the notebook to guide readers through the code. Clearly explain the purpose of each code segment, the insights gained, and the decisions made.
    - Name your file in the format ***ML_GroupXX_Notebook,*** where ***GroupXX*** should be your group number.

- A **report** that describes the analytical processes and the conclusions obtained, with, at most **15 pages** (excluding cover but including annexes). The body of text should only include figures and tables that are essential to understand the work. Supporting figures and Tables can be added to annexes but must be referenced in the text.
    - Name your file in the format ***ML_GroupXX_Report.pdf***, like the notebook.
    - Use the template provided in the file *Report_Template.docx* with the following setting:
        - Heading 1: Calibri, Size 14pt, in bold.
        - Heading 2 (if needed): Calibri, Size 14pt, in bold.
        - Text: Calibri, Size 11pt, line spacing of 1.15pt and paragraph spacing of 6pt.

- A **presentation** (max. 10 minutes) highlighting your general work, main findings and recommendations. The file name should follow ***ML_GroupXX_Presentation.pdf***. **Your group will be required to present it in your defense (time and date tbd).**

- In all previous cases, your group number should replace the expression ***XX*** in Group ***XX***.

Deadline: June 6th at 18h00

05 VI. EVALUATION

Your work will be evaluated according to the following criteria:

CRITERIA	PERCENTAGE (%)	MAXIMUM GRADE (OUT OF 20)
Report Quality & Storytelling	10	2
Methodology	5	1
Preprocessing & EDA	5	2
Additional Insight	5	1
Descriptive Modeling	20	4
Predictive Modeling & Optimization	25	5
Action Plan & Conclusions	5	1
Kaggle Performance	5	1
Presentation & Discussion	15	3

Your grade will reflect our assessment of the quality of your work in terms of quality of writing, clarity, conciseness, correctness and efficiency. Please find below more details about what is taken for each topic:

- **Report Quality & Storytelling:** Each report should follow the report template provided. A good report should give the reader a clear picture of the problem you are tasked with , the steps you took, the rationale behind those steps , your main results and your insights. Adequate identification and reference to figures, sources are also considered here.
- **Methodology:** The methodology should clearly highlight the general approach taken and describe the various stages of the project.
- **Preprocessing and EDA:** Describe the dataset and extract meaningful insights that you consider relevant to the problem. Avoid adding unnecessary visualizations or elements. This section also addresses the initial preprocessing of the data, providing a clear explanation of each step taken and the rationale for your choice.

06

- **Additional Insights:** Describe your strategy for the additional insight's objective. This section is separated into different components:
  - Formulation and adequacy
  - Difficulty
  - Correctness/efficiency of implementation
  - Explanation and discussion of results
- **Descriptive Modeling:** Describe your strategy for the clustering objective. This section is separated into different components:
  - Additional preprocessing
  - Modelling approach – implementation and reasoning behind any clustering model used. You should explain the feature selection for the models.
  - Description of resulting clusters- Each cluster should be statistically and visually explored and described, emphasizing the characteristics that differentiate them.
- **Predictive Modeling & Optimization:** Describe your strategy for the classification objective. This section is separated into different components:
  - Additional preprocessing
  - Feature selection– strategy employed to select features included and excluded from the models.
  - Modelling approach –the predictive algorithms explored.
  - Performance assessment–rationale for choice of evaluation metric(s), interpretation of results and algorithm comparison.
  - Model optimization–fine-tuning strategies used to optimize models..
- **Action Plan & Conclusions:** You should provide a succinct but well-oriented action plan that includes recommendations on how the WHO can leverage the insights and results obtained throughout the project.
- **Model performance on Kaggle:** Grading based on groups ranking.
- **Presentation & Discussion:** In this section, we will not only evaluate your presentation as a standalone document, but also how you, as a group, deliver the message you want to convey while presenting. How comfortable in answering questions about your data and methods is also under evaluation.

07 VII. PARTING NOTES

1. Deliveries after the deadline will be penalized at 1 point per day.
2. Deliveries made before the deadline will receive a bonus of 0.15 points per day of delivery in advance (up to a maximum of 1 point).
3. For modelling purposes, any **algorithm implementation outside the vanilla scikit-learn is explicitly off-limits** and will result in a 1-point penalty.
4. The report will be the primary method of evaluating your work. When preparing it, remember that a **reader should be able to understand your work without needing to check your notebook**. We won't be able to consider any steps or results not mentioned in your report.
5. Ensure your report is concise, focused and based on reliable sources. You should look to source information provided from peer-reviewed journals (thus, avoid citing Medium, TowardsDataScience and similar sources). Avoid irrelevant, unimportant, or redundant information. Don't provide theoretical explanations of topics covered in class.
6. Before submitting, run your notebook from the start one last time (if you used a GridSearch, you can comment this cell, but you should run the final model with the GS parameters in a different cell).
7. **Your submitted notebook should include all the unneeded code you used to obtain your final solution, but it should also be commented**.
8. We will run your Jupyter Notebooks if we have any doubts. So, please **make sure we can run the notebook from start to finish in one go. Notebooks that do not fulfil this condition will be penalized.**
9. The report and code will pass through a process of plagiarism and AI generation checking.
10. You must submit your predictions on the Kaggle competition to get points for that component (more details on this can be found on Moodle).
11. When determining the grade for your work, there will be a comparative component between it and the work presented by your peers.

Friendly Reminders:

1. Attendance at the **defense is mandatory for approval** in the project. The defense has a group component and an individual component.
2. Questions during the discussion will be individualized, and every group member should be able to explain, to some extent, what was done in the project at every step of the way.
3. **Do not include techniques/algorithms/steps you cannot explain** in your report: we will ask about them in the defense.
4. Finished is better than Perfect.