NOVA IMS Information Management School

september, 2024

Obesity in Focus: data for healthier lives

Group Project
Machine Learning I
2024/2025

(Image source: Britannica, 2024)

01

I. Introduction

Obesity is one of the most pressing public health issues of our time, affecting millions of individuals worldwide (Haththotuwa, Wijeyaratne, & Senarath, 2020). Studies have shown that the prevalence of obesity is driven by factors such as sedentary lifestyles and unhealthy eating habits (Sultana et al., 2021). This trend has highlighted the urgent need to develop effective tools for assessing the risk of obesity based on behavioral patterns, physical conditions, and other information about individuals.

In recognition of this need, a global public health organization has partnered with machine learning researchers (you!) to create a predictive model aimed at identifying obesity levels based on individuals' dietary habits, physical conditions and background.
The project's goal is to support prevention campaigns by providing a better understanding of the relationship between lifestyle factors and the risk of obesity, thereby enabling more effective and personalized interventions.

II. Project Goals

Your team's challenge is to develop a predictive model that accurately estimates obesity level for any given individual, using the data provided.

Objectives:

- Data Exploration and Preprocessing: Conduct thorough data exploration to understand the main characteristics of the dataset and prepare it for effective modeling.
- Feature Analysis: Identify and analyze the most significant predictors of obesity levels based on lifestyle habits, physical condition and background.
- Predictive Modeling: Apply machine learning algorithms to predict an individual's obesity level using the variables in the dataset.
- Critical Insights: Beyond predicting obesity levels, derive insights that can inform public health strategies, preventive interventions, and policies aimed at reducing obesity risk.

02

III. Dataset

The data was gathered through interviews done to individuals between ages 16 and 56. The dataset is divided between a training set (close to 1600 participants) and a testing set (close to 500 participants).

In the training set, you will find the features and the ground truth associated with each instance, i.e. the participant's obesity level. Use it to build your machine learning models. The goal will be to use the model you created and make predictions on unseen data (i.e. your test set).
In the test set, you will see the same features presented in the training set. However, you will not have access to the ground truth of the test set. Your goal will be to predict the ground truth value by using the model you created using the training set, therefore you should train a predictive model with the training set and use the model to make predictions on the test set.
The available data contains the following attributes:

Attribute	Description
IDcode	Unique identifier for the participant
age	Participant's age in years.
alcohol_freq	How often the participant consumes alcohol.
caloric_freq	Whether the participant frequently eats chocolate/sweets or not.
devices_perday	The amount of time (in hours) the participant spends using electronic devices daily.
eat_between_meals	Whether the participant consumes food between meals.
gender	The gender of the participant.
height	Participant's height in meters.
marital_status	The participant's marital status.
meals_perday	How many main meals the participant consumes daily.
monitor_calories	Whether the participant monitors their daily caloric intake.

03

Attribute	Description
parent_overweight	Whether the participant's parents (father or mother) suffer from overweight.
physical_activity_per week	How many days per week the participant engages in physical activity.
region	The participant's region of origin.
siblings	Number of siblings the participant has.
smoke	Whether the participant smokes or not.
transportation	The primary mode of transportation used by the participant.
veggies_freq	How frequently does the participant consume vegetables in their meals.
water_daily	The amount of water (in liters) the participant consumes daily.
weight	Participant's weight in kilograms.
obese_level	The participant's obesity level (target variable).

IV. Deliverables

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

- A report that describes the analytical processes and the conclusions obtained with, at most, 15 pages (excluding cover, but including annexes). The file naming format should follow ML_GroupXX_Report.pdf, where GroupXX should be your group number. You should follow the template provided in the file Report_Template.docx and costumize it but you must keep the following settings or you will be penalized:
  - Heading 1: Calibri, Size 14 pt, in bold
  - Heading 2 (if needed): Calibri, Size 13 pt, in bold
  - Text: Calibri, Size 11 pt, line spacing of 1.15 pt and paragraph spacing of 6 pt
- A notebook with your code implementation. The file naming format should be ML_GroupXX_Notebook.ipynb, where "GroupXX" should be your group number.

- The deadline for the delivery is 23:59 December 20th. All evaluation documents (report and notebook) must be submitted with a maximum delay of 3 days, incurring a penalty of 1 point per day. After these three days, no further submissions will be accepted.

V. Evaluation

Your work will be evaluated according to the following criteria:

Criteria	Percentage (%)	Maximum Grade (out of 20)
Kaggle Performance	10	2
Report Quality and Storytelling	20	4
Data Exploration	15	3
Feature Selection	15	3
Modelling	15	3
Performance Assessment	15	3
Creativity and other self-studies	10	2

05

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

- Kaggle Performance: The performance obtained on the test set.
- Report Quality and Storytelling: Each report should follow the report template provided. A good report should, by itself, 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.
- Data Exploration: Describe the data and extract meaningful insights that may be helpful to address the problem at hand.
- Feature Selection: Describe the strategy you employed to select which features you included in your models and which features you excluded. Do not forget to mention which features were selected for the next phases. Blindly and strictly following the strategies implemented in class will be heavily penalized.
- Modelling: Implement different predictive algorithms and compare their performances. Moreover, you are also expected to employ fine-tuning strategies to your models. The application of models not covered in class is optional and considered in the Creativity and other self-studies section.
- Performance Assessment: Your strategy for comparing different models and their performance. Pay special attention to the metrics you choose and how you interpret them.
- Creativity and other self-studies: If other techniques not given during practical classes are applied, a theoretical explanation of the algorithm/technique should be provided in the background section (do not use the background for topics covered in class). This topic includes not only the application of different techniques but also aspects of creativity, such as the quality of visualizations, plots and others.

06

VI. Parting Notes

1. For modelling purposes, any algorithm implementation outside the vanilla scikit-learn is explicitly off-limits. Moreover, using Lazy Predict or similar AutoML packages is also not allowed.
2. 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.
3. Please don't provide long theoretical explanations of topics covered in class in your report.
4. Everything featured in your report must have a clear purpose. Avoid including irrelevant/unimportant/redundant information, as the space is limited and you will need it.
5. Trustworthiness of the information you provide is key. You should look to source information you provide from peer-reviewed journals (thus, avoid citing Medium, TowardsDataScience and similar sources).
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 parameters found by the GS in a different cell).
7. All the unneeded code you used to obtain your final solution should be part of your submitted notebook, but it should 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 fulfill 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 to the Kaggle competition to get points for that component (more details on this can be found on Moodle).

Friendly Reminder:

1. Attendance at the defense is mandatory for approval in the project. Grades can change significantly during the defense (improve or decrease) without any limitations to the extent of the change, depending on the answers provided.
2. As questions are individualized, every group member should be able to understand what was done at every step of the way.
3. If something is good enough to be mentioned in the report, it is also good enough to know. DO NOT include techniques/algorithms/steps you cannot explain in your report: we may (and probably will) ask about them in the defense.
4. Finished is better than perfect.

07

VII. References

Britannica, T. Editors of Encyclopaedia (2024, September 21). obesity. Encyclopedia Britannica. https://www.britannica.com/science/obesity

Sultana, S., Rahman, M. M., Sigel, B., & Hashizume, M. (2021). Associations of lifestyle risk factors with overweight or obesity among adolescents: a multicountry analysis. The American journal of clinical nutrition, 113(3), 742-750.

Haththotuwa, R. N., Wijeyaratne, C. N., & Senarath, U. (2020). Worldwide epidemic of obesity. In Obesity and obstetrics (pp. 3-8). Elsevier.