

**ML**

**Dissecting Obesity**

Finding the causes with predictive modelling

**Group05**

**[DATE\_TO\_INSERT]**

**Christian Deluca 20241264**

**Dino Meng 20241265**

**Marek Rewolinski 20241452**

**Michal Marcin Wójcik 20241255**

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

Abstract

The abstract is a paragraph containing 250-350 words summarising your report and findings. As a rule of thumb, it should be arranged in five sections: subject matter and purposes of the study (Introduction), the methodology adopted (Methodology), your results (Results), the interpretation of your results (Discussion) and a conclusion of your work (Conclusions) based on your findings. Sentences which do not convey useful information should be avoided.

While the Abstract is the first visible section of your work, it will often be the last part you write. In conclusion, an abstract should be viewed as a miniversion of the report - it should provide a brief summary of each main section [1].

Keywords

Keyword 1; Keyword 2; Keyword 3; Keyword 4

*[ You must have at least one keyword. You can enter up to 6 keywords separated by a semi-colon. Capitalize the first letter of keywords. Use full phrases rather than acronyms or abbreviations (use Machine Learning rather than ML). ]*

# 

# Introduction

The Introduction section shows the current need for studying a specific problem [4]. The I**ntroduction** should be brief and clearly state the question you tried to answer in the study [3]. In this way, this section should focus on the Introduction of the problem. Every scientific report needs an introduction which presents background information a reader needs to understand the rest of the author’s work [4].

Script: 1) what is obesity 2) increase in obesity 3) goal: find lifestyle factors which might influence in someone’s obesity

Obesity is a health condition where someone’s quantity of excessive fat deposits potentially has negative impacts on their health. As of lately, the rate in obesity has been increasing in a worryingly manner: quoting the W.H.O. organization, “in 2022, 43% of adults aged 18 years and over were overweight and 16% were living with obesity” [1].

As obesity is a multifactorial phenomenon [1], in this study we will attempt to discern the major factors - in particular lifestyle and genetic ones - in impacting individuals’ obesity levels. To do this, we will use a predictive model to accurately classify obesity levels and interpret such model with the results.

# Data Exploration

This section should outline the main observations and findings you have achieved throughout your data exploration procedures. In this section, you should not focus on providing lenghty descriptions of mere statistical properties of the data. Instead, you should highlight the main univariate and multivariate observations that inform subsequent steps (the ones you explain later in Data Preprocessing and other subsequent steps.

## Level 2 title

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.



Figure 2.1 – Illustrative figure

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables

Table 2.1 – Illustrative table

|  |  |
| --- | --- |
| **Title** | **Title** |
| Text | Number |
| Text | Number |
| Text | Number |

Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables Sample text with the inclusion of figures and tables.

### Level 3 title

Example of an unnumbered list:

* Item 1
* Item 2
* Item 3

# Methodology

The Methodology section describes the procedures and the materials used to conduct the study [4]. The primary purposes of the methods section are to describe and sometimes defend the experimental design and to provide enough detail that a competent worker could repeat the study [3]. In here, your report should be clear and unambiguous on the following aspects:

1. What was your model assessment strategy (e.g. Hold-out, Cross-validation, etc…)
2. What did you do in terms of Data Preprocessing
3. What was your feature selection strategy
4. What predictive algorithms did you use and how did you select the best one
5. What model optimization efforts did you undergo

Note: The methods section is not meant to show or interpret any results. Instead, you should focus on outlining the steps you did and which criteria you adopted.

In this project we have used the following pipeline to construct and evaluate the model.

To preprocess the data, we have used the following pipeline:

1. Manually remove outliers detected in the exploratory data analysis phase
2. Standardize numerical values, impute numerical values with KNN and de-standardize numerical values to preserve interpretability
3. Combine age, weight and height into a BMI class
4. Fill missing values in physical activity with “No Activity”
5. Encode categorical variables with ordinal values following a precise order
6. Impute missing categorical values with decision classifiers
7. Combine the categorical variables, which represent a lifestyle factor, into a linear combination with weights equal to 1, denoted as “life score”

In terms of feature selection strategy, we used the main filter, wrapped and embedded methods; in particular:

* As for the filter methods, we used the Pearson correlation to determine if there is any correlation between numerical variables; then we used the Spearman correlation to determine if there is any correlation between all variables; in the end we used the Spearman rank correlation to test the categorical variables’ independence from the target variable
* As for the wrapped methods, we used the Recursive Features Elimination method with decision trees as a baseline model
* As for the embedded methods, we fit a random forest classifier and extracted the feature importances

For the model itself, we fit and evaluated the logistic regression, simple decision trees, random forest classifiers and gradient-boosted decision trees. To select the model, we evaluated them with repeated 10-folds and selected the model with best macro average f1-score in evaluation data.

To optimize the final model, we used Grid Search to find the best hyperparameters for our model.

# Results

The **Results** section organizes the findings [4]. This section is the meat of a report, the most important part of a study. All other sections serve subordinate roles, either preparing the reader for the **Results** or providing supplemental information to augment the findings. **Results** are general statements that present the research's key results (data) without interpreting their meaning. The author should not include the raw data but present them as text, illustrations, and tables. All these three forms may be used, but the same data should not be repeated in more than one form. The results of statistical analyses should also be stated in this section [4]. Consider the following guidelines when writing the **Results** section [4]:

1. It is not necessary to include all the collected data during the research. This isn’t a diary. Instead, select and emphasize only important and relevant data that will answer the question or solve the problem raised in the Introduction section.
2. Do not include information belonging to other sections of the paper, such as Materials and Methods, or Discussions (if Results and Discussions are separated).
3. Prevent repeating the legends for figures or the titles of tables in the text.
4. Explain in the text only those illustrations and tables whose significance is not obvious to the reader. Important features that are readily apparent from the illustrations and tables should be highlighted in the text. Do not repeat the data presented in the illustrations and tables.
5. Be sure that the text, illustrations, and tables are consistent with one another. In addition, make sure that all numerical values in every table agree with the figures or data presented.
6. Analyze your data by statistical methods, if appropriate.

Be honest. Do not omit data that does not support your hypothesis and conclusion or do not answer the research question.

To conclude, the results section of a report has two key features: an overall description of the study's major findings; and the data should be presented clearly and concisely [3].

DRAFT: idk, take from EDA and feature selection?!?!?

# discussion

The Discussion section attaches the findings to other existing scientific papers to form new ideas [4]. The Discussion section of a scientific article reiterates the main findings but in the context of furthering knowledge or impacting on teaching practice, or future research. In other words, the **Discussion** takes and interprets the findings reported in the Results section, evaluates their significance, and examines the implications. This is probably the most challenging to write among all sections in a research article and will demonstrate how well the author understands the results. Nevertheless, it does not mean that the Discussion should be long, especially if there is little to discuss. [4]

DRAFT: idk we need to research for papers

# Conclusion

The conclusion presents the outcome of the work. In it, you should interpret the findings at a higher level of abstraction than the one you did in the ***Discussion***. Moreover, it would help if you related these findings to the motivation stated in the ***Introduction*** [2]. Finally, this section is where the author restates the contribution of the research, with a particular emphasis on what it allows others to do; and proposes new research directions to prevent duplication of effort or encourage collaboration[4].

DRAFT: idk to be honest, maybe look at the values indicated in the feature selection part (ex: features importance, et cetera…)

# References

Every research project usually relies in part upon the work of other scientific works. Therefore, any time an author cites external materials, he/she must identify his/her sources in the form of systematic references. The importance of the References section in a report is not only for giving credit to the ideas and work of other scientists but also to provide the readers with access to these sources [4]. [You should use a reference management system such as Endnote, Mendeley, Citavi or Zotero.]

[1] Robert A Day. How to write and publish scientific reports, 1998.

[2] Jean-Luc Doumont, Laura Grossenbacher, Christina Matta, and Jorge Cham. English communication for scientists. 2014.

[3] George M Hall. How to write a report. John Wiley & Sons, 2012.

[4] Parlindungan Pardede. Scientific articles structure. In Scientific Writing Workshop, volume 16, 2012.

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

# Appendix (optional)

*[An* ***annex*** *can stand alone. For example, if you are attaching additional documents to the end of your report that would make complete sense and provide important information even outside the context of your report, you can categorize them as annexes.*

*An* ***appendix*** *tends to be more closely connected than an annex to the main body of the report. An appendix would not be as informative or valuable outside the context of your report. While an appendix enhances or expands upon your research report by adding details like illustrations or case studies, it is never presented to readers by itself.*

*Retrieved from* [*https://www.enago.com/academy/annex-vs-appendix-do-you-know-the-difference/*](https://www.enago.com/academy/annex-vs-appendix-do-you-know-the-difference/)*]*

# Annexes (optional)

