iscte
INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Introduction to Machine Learning — 2024/2025

Final Project

> This project should be solved using Python notebooks due to the ability to generate a report integrated with the code. It is assumed you are proficient with programming. All answers must be justified and the results discussed and compared to the appropriate baselines. In addition to the technical report integrated with the code, a report documenting the application of the CRISP-DM methodology should also be submitted.
> Max score of the project is 8 points. The work should be done in groups (two students) or individually. In the case of groups with two members, the report should indicate an estimate of each member's contribution to the work. For example: manuel: 60%, pedro: 40%, together with a short justification. *It is mandatory to make an oral presentation and discussion of the project.*
> **Deadline:** January 9th, 2025 (1st season), or January, 20th, 2025 (2nd season)

The objective of this project is to apply the CRISP-DM methodology to solve a cardiovascular diseases risk prediction problem, using Machine Learning methods.

To that end, a dataset based on information from United States Behavioral Risk Factor Surveillance System[1] should be used [Lupague et al., 2023]. This surveillance system conducts health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.

The project and the report should follow the phases of the CRISP-DM methodology.

To experiment with the different machine learning models, the scikit-learn[2] toolkit [Pedregosa et al., 2011] should be used.

[1] `https://www.cdc.gov/brfss/annual_data/annual_2021.html`

[2] `https://scikit-learn.org/stable/index.html`

Dataset

This comprehensive dataset [Lupague et al., 2023], obtained from the 2021 annual Behavioral Risk Factor Surveillance System dataset by the Center for Disease Control and Prevention, originally comprised 438,693 records with 304 attributes. However, for the purpose of constructing a predictive machine learning model for cardiovascular disease, a focused subset of 19 relevant attributes was selected. This selection process reduced the dataset to 308,854 records, which are to be used for analysis and model development.

Each sample is characterized by 19 features as follows:

General Health Would you say that in general your health is …

Checkup About how long has it been since you last visited a doctor for a routine checkup?

Exercise During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?

Heart Disease Respondents that reported having coronary heart disease or mycardial-infarction.

Skin Cancer Respondents that reported having skin cancer.

Other Cancer Respondents that reported having any other types of cancer.

Depression Respondents that reported having a depressive disorder (including depression, major depression, dysthymia, or minor depression).

Diabetes Respondents that reported having a diabetes. If yes, what type of diabetes it is/was.

Arthritis Respondents that reported having an Arthritis.

Sex Respondent' sex.

Age Category Respondent' age category.

Height Respondent' height in cm.

Weight Respondent' weight in kg.

BMI Respondent' body mass index.

Smoking History Does the respondent has a smoking history.

Alcohol Consumption Self-reported data regarding the frequency and quantity of alcohol intake by the respondent.

Fruit Consumption Self-reported data regarding the frequency and quantity of fruit intake by the respondent.

Green Vegetables Consumption Self-reported data regarding the frequency and quantity of green vegetables intake by the respondent.

Fried Potato Consumption Self-reported data regarding the frequency and quantity of fried potatoes intake by the respondent.

Note that the in Data Understanding step of the CRISP-DM methodology these aspects must be clearly addressed, providing an adequate understanding of this data.

The dataset is available at `moodle.iscte-iul.pt`.

Experiments

You should perform the following experiments:

- Use supervised and unsupervised methods (see following sections);
- Randomly remove 10% and 20% of the values of the features the dataset and explore two different strategies to handle missing values;
- Experiment with data normalization, data discretization, and data reduction. Apply these steps to the original, unchanged, dataset.

Do not forget to visually explore your data, namely presenting correlations between pairs of features.

The technical evaluation should include different metrics to better understand the errors of the supervised machine learning approaches. The assessment of the unsupervised machine learning approaches should be based on the different strategies and metrics addressed in the respective lecture.

Supervised Learning Algorithms

The target for the supervised experiments should be the attribute *Heart Disease*. Experiment with the following supervised learning algorithms and comment the results, based on your knowledge of how they work:

- Decision Trees;
- Multi-layer perceptron;
- k-NN.

Unsupervised learning algorithms

The unsupervised experiments should focus on understanding groups of people and how they are related to having or not heart disease. Experiment with the following unsupervised learning algorithms and comment the results, based on your knowledge of how they work:

- k-Means;
- DBScan.

References

R. M. J. M. Lupague, R. C. Mabborang, A. G. Bansil, and M. M. Lupague. Integrated machine learning model for comprehensive heart disease risk assessment based on multi-dimensional health factors. *European Journal of Computer Science and Information Technology*, 11(3):44–58, 2023.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.