Aprendizagem Computacional
Machine Learning

2025/2026

Project Assignment

Detecting breast cancer from blood samples

1 Background

Detecting cancer at the initial stages is a major aspect concerning treatment success and survival rate. A minimally invasive way, without submitting the patient to specialized procedures that usually include imaging by ionizing radiation (X-Ray), is through the analysis of blood samples.

2 Objective

Your task is to develop machine learning pipelines to identify if a blood sample is or is not collected from a person with breast cancer.

3 Practical Assignment

3.1 Dataset Description

Consider the dataset available at https://archive-beta.ics.uci.edu/dataset/451/breast+cancer+coimbra[1] that contains clinical features for 64 patients with breast cancer and 52 healthy controls.

Ten features were considered and are related to anthropometric data and parameters gathered in routine blood analysis.

3.2 Data partitioning

Implement a training, validation, and testing data partitioning approach to perform a trustworthy performance assessment of your pipelines.

3.3 Feature Selection and Reduction

Some of the supplied features may be useless, redundant, or highly correlated with others. Consider using feature selection (Kruskall-Wallis, ROC-AUC and correlation analysis) and dimensionality reduction (PCA and LDA) techniques to evaluate their impact on the performance. Analyze the distribution of your feature values and calculate the correlation between them. Ensure you understand your features! Remember to include your findings in the final report.

3.4 Classification

Implement just the classifiers studied during the course. Other classifiers are not important. Thus, you should consider the classifiers: Minimum Distance, Fisher LDA, Bayes classifier (not Naive Bayes), kNN, SVM (Linear+Non-Linear), Adaboost, Random Forests and Decision Trees.

3.5 Performance Assessment & Causality Analysis

Consider the performance metric learned during the course to report the performance of your pipelines, such as: Sensitivity, Specificity, Precision, F1score and ROC-AUC.

Run the experiments multiple times and present average/median results and standard deviations/IQR (for the metrics used).

Remember that manually inspecting your algorithm's predictions can provide valuable insights into where it is failing and why, as well as how to improve it (e.g., what causes the algorithm to fail in specific cases? What unique characteristic makes it particularly challenging? How can I help the algorithm handle those cases more effectively?). Reassess the Pre-processing, Feature Reduction, and Feature Selection phases until you are satisfied with the results. It's wise to monitor the evolution of your algorithm's performance throughout this process. Aim to display these trends in your final report to support all related issues (parameter selection, model fit, etc.).

You can write your own code or utilize the available functions and methods. The methods employed in your work should be described and understood by you. Be aware of what is the impact of selected parameters.

Present and discuss the results obtained from your project assignment. This dataset has already been studied by the authors of the dataset in [2]. Compare your findings with the results they achieved.

3.6 Submissions

In both intermediate and final submissions you have to submit a report (in Portuguese or English) in PDF format and the developed code.

3.6.1 Intermediate

The intermediate submission should include the initial steps, such as:

- Data loading;
- Data partitioning;
- Feature inspection and dimensionality reduction (feature selection+feature. transformation);
- Minimum Distance classifier + Fisher LDA.

⇒ **Deadline: October 31, 2025!**

3.6.2 Final

The final submission submission should consider all the pipeline steps, including all the previosly referred classifiers. You are free to reformulate the steps implemented for the intermediate submission.

⇒ **Deadline: December 12th, 2025!**

3.7 Requirements

The practical assignment should be developed in pairs. However, students who prefer to work individually are also welcome. Larger groups are not permitted!

References

[1] Patrcio, Miguel, et al. "Breast Cancer Coimbra." UCI Machine Learning Repository, 2018, https://doi.org/10.24432/C52P59.

[2] Patrício, Miguel et al. “Using Resistin, glucose, age and BMI to predict the presence of breast cancer.” BMC Cancer 18 (2018).