

 FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO   FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO	First Degree in Artificial Intelligence and Data Science Elements of Artificial Intelligence and Data Science	2024/2025 1 st Year 2 nd Semester
TEACHERS: Luís Paulo Reis, Miriam Santos, Rita Ribeiro, Moisés Santos		

Assignment No. 2

Machine Learning Project: Building a Student Intervention System

Theme

The second practical assignment of this course consists in the development of a full machine learning pipeline, from exploratory data analysis and data preprocessing to the application of supervised learning techniques for classification and their respective performance evaluation. Optionally, the project may also consider the exploration of additional techniques such as clustering, more advanced data preprocessing methodologies, or other approaches to foster Responsible AI practices.

Building a Student Intervention System

In this project, the goal is to address a real data science case study using the **UCI Student Performance** dataset. This data was collected from two secondary education schools in Portugal, and includes information about students' grades and demographics, as well as social and school-related features. The main goal of this project is to develop a machine learning pipeline capable of predicting whether a given student will pass their final exam, thus allowing to build an intervention system that may flag individual students requiring extra attention and support. To address this project, you should focus on each step of a standard machine learning process and explore suitable solutions to develop an efficient solution:

- **Data Exploration:** An initial exploratory data analysis should be carried out, including examining feature types, number of features/records, class distribution, values per attribute, etc., and highlighting feature inconsistencies such as missing values, outliers, underrepresented concepts, irrelevant features, etc. The analysis can and should be supported with visualization techniques.
- **Data Cleaning and Preprocessing:** This refers to feature cleaning and preprocessing (e.g., imputation of missing values, outlier removal, resampling, data transformation, data scaling, etc.) and feature engineering (e.g., building new features or removing redundant features) and other tasks considered relevant.
- **Data Modeling (Supervised Learning):** Supervised learning includes the identification of the target concept, definition of the training and test sets, selection and parameterization of the learning algorithms to apply, and evaluation of the learning process. You can explore several classifiers (e.g., Decision Trees, KNN, Logistic Regression, Neural Networks, among others) available on scikit-learn to build and evaluate suitable solutions.
- **Performance Evaluation:** Classification results should be compared across different evaluation metrics (learning curves, confusion matrix, ROC/AUC, precision, recall, accuracy) using different cross-validation strategies for train/test splits. Results should be compared using tables and plots (e.g., using seaborn or matplotlib libraries).
- **Interpretation of Results:** This involves extracting meaningful insights from the obtained results: explaining the behavior of the models, drawing conclusions about the effectiveness of the chosen algorithms and preprocessing techniques, providing recommendations for future analysis, investigating discrepancies of unexpected findings, etc.

Extra Elements:

The incorporation of elements beyond the core requirements of the project are given a bonus of 10%. These elements can either focus on the technical implementation of data science software to support user interaction and data analysis, or refer to subsidiary tasks along the experimental setup. Some suggestions are as follows:

- Implementing advanced techniques for missing data imputation and assessing their effect in classification performance (e.g., sensitivity/specificity results) and imputation quality (e.g., RMSE).
- In case of class imbalance, assessing the impact of data balancing techniques, either through data down-sampling of the most frequent class or through more advanced imbalanced data techniques (e.g., SMOTE, ADASYN) and compare their impact on the final performance results.
- Experimenting with additional algorithms and hyperparameters to optimize model performance.
- Exploring the dataset from a Responsible AI perspective, focusing on bias and fairness or explainability, among others.
- Deploying the solution to an external source, such as creating a Streamlit app to create an interactive web application for the project.

Programming Language/Libraries

The programs should be developed using Python language due to the availability of very strong machine learning libraries for this language. It is highly advisable that the main libraries used are the ones lectured on the course such as pandas, numpy/scipy, scikit-learn and matplotlib/seaborn. **The final result should be a jupyter notebook containing all of the source code, images, tables, and discussion of results.**

Groups

Groups must be composed of 3 students. Groups should be composed of students from the same practical class. All students should be present in the final presentation/demonstration of the work. The establishment of groups composed of students from different classes is not advised, given the logistic difficulties that this can cause and is only accepted in exceptional conditions.

Final Delivery

Each group must submit in Moodle the following deliverable:

- **A Jupyter Notebook (.ipynb)** with all the details of the final work: data preprocessing, the developed models and their evaluation and comparison using appropriate graphical elements (tables, plots, etc.). The notebook should also contain a proper discussion of results about the steps of the project development, written in markdown (similar to an interactive report).

You may submit your final work until May 30, but note that the presentations of the project will take place on the week of May 26 to May 30, during the practical classes or in another period to be designated by the teachers of the course.

Evaluation

The project will be evaluated regarding the following expected outcomes:

- **Student Involvement (10%):** Student involvement and contribution to the project;
- **Presentation Quality (10%):** Quality of presentation delivered by the students, domain over the project's goals, methodology, results, and conclusions;
- **Jupyter Notebook Quality (10%):** Quality of the delivered implementation and discussion of insights;
- **Data Characterization (10%):** Understanding feature types and overall dataset and feature characteristics;

- **Data Quality Assessment and Data Preprocessing (20%):** Exploring feature selection, transformation, and engineering methods, identifying data quality issues in the data and applying suitable techniques to handle them effectively;
- **Data Visualization (15%):** Visually exploring the data and producing meaningful insights from the chosen visualizations (e.g., barplots, boxplots, histograms, heatmaps, correlation matrices).
- **Supervised Learning (20%):** Exploring suitable classification algorithms and interpreting performance metrics;
- **Critical Thinking and Communication of Results (10%):** Examining the information yielded by the data analysis and sharing insights regarding the problem domain.
- **Extra Elements (10%):** Any creative methods that go beyond the scope of the project, either theoretical (e.g., exploring clustering solutions, other classifiers, distance metrics, missing data, imbalance data, fairness and bias issues) or practical (e.g., developing a Streamlit application to showcase the project results).

Bibliography

1. UC Irvine Machine Learning Repository. Student Performance, <https://archive.ics.uci.edu/dataset/320/student+performance>
2. Cortez, P., & Silva, A. M. G. (2008). "Using data mining to predict secondary school student performance". <https://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf>