AA1 – 24/25

Machine Learning Competition

1. Introduction

The purpose of this work is to solve a classification problem proposed as a competition in the Kaggle InClass platform, where each team will try to get the maximum score. You can apply any of the concepts and techniques studied in class for exploratory data analysis, feature selection and classification.

2. Problem description

Context:

Patients with liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. This dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors.

Content:

The original data set contains 416 liver patient records and 167 non-liver patient records collected from North East of Andhra Pradesh, India. Any patient whose age exceeded 89 is listed as being of age "90".

The dataset has been preprocessed to remove samples with missing values and convert categorical features to numerical data. After this stage, the dataset contains 414 liver patient records and 165 records for healthy subjects.

Attribute information:

1. Age: Age of the patient
2. Female: Gender of the patient (1 if Female, 0 if Male)
3. TB: Total Bilirubin
4. DB: Direct Bilirubin
5. Alkphos: Alkaline Phosphotase
6. Sgpt: Alamine Aminotransferase
7. Sgot: Aspartate Aminotransferase
8. TP: Total Protiens
9. ALB: Albumin
10. A/R: Albumin and Globulin Ratio

Acknowledgements

This dataset was downloaded from the UCI ML Repository:

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Goal:

The goal is to create a classifier that predicts whether a subject is healthy (non-liver patient) or ill (liver patient) based on some clinical and demographic features.

3. Dataset

The dataset has been split into training (80%) and test (20%) subsets provided as two separate files.

- The training dataset: the file `train_features_ILDS.csv` contains a 463x10 matrix, where rows correspond to subjects and columns to features, and file `train_labels_ILDS.csv` contains the 463x1 vector of labels.

- The test dataset: `test_data_ILDS.csv` contains a 116x10 matrix with data for the 116 test subjects. No labels are provided for the test set.

You will have to upload the predictions for the test set and Kaggle will compare your predictions with the ground-truth labels in order to compute a score.

4. The competition in Kaggle InClass platform

The competition is available at https://www.kaggle.com/t/1ccb38a0e7054e6dbe2c953c10cf0124

Each group needs to create a Kaggle account to participate in the competition. Using this account, you will be able to download the dataset, participate in the competition forum, make submissions and see your scores on the live leaderboard.

One account per team: each team will use just one Kaggle account. The username must contain the last name of the group members, separated by underscore (e.g. name1_name2).

Code sharing: it is not allowed to share your code with other teams, but you can use the discussion forum to post questions or comments.

Team alliances: alliances between groups are not allowed.

Submission limits: there is a limit of 5 submissions per day per team.

Competition timeline:
Start date: May 1th, 2025
End date: Jun 20th, 2025

5. Evaluations

Metric:

The metric used by Kaggle to compute each submission score is the F1-score (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score). However, in addition to the F1 score you can compute other performance metrics and include them in your final report. Have in mind the unbalance between classes and test at least one solution to fix it.

Submission file format:

You should submit a csv file with exactly 116 entries plus a header row. Your submission will show an error if you have extra columns or rows.

The file should have exactly 2 columns:

- ID index corresponding to each sample from file `test_data_ILDS.csv`
- Label binary prediction (0 for liver-patient, 1 for healthy subject)

Example:

```
ID,Label
1,0
2,0
3,1
4,0
```

Homework evaluation:

The evaluation of this assignment will be based on the quality of your report, appropriateness of conditioning and analysis of features, selection of ML methods adopted, management of hyperparameters, justification of the performance indicators used, code efficiency, and your final rank in the leadership board.

6. Code and report handling

Handling instructions:

Submit your work though Atenea. Save your code and report in a folder (name `AA1_name1_name2`) and upload the compressed folder (zip, rar).

Code:

The code should contain all the python scripts used for visualizing data, training and testing of models and creation of Kaggle submission files.

Report:

The report must be written using the IEEE template provided in https://www.ieee.org/conferences/publishing/templates.html, either in MSWord or latex. The Overleaft format is also provided (it is recommended for collaborative work). The report has to include the following sections:

- The Abstract should contain a maximum of 200 words.
- The "Introduction" section should describe the problem
- The titles of the following sections should be:
    - "Feature analysis" (optional, if you use any method for feature selection, feature engineering or exploratory data analysis)
    - "Classification methods" (where you explain which of the models studied in class were used to solve the problem, including tricks to deal with class imbalance, use of ensemble classifiers: boosting, random forest, etc.)
    - "Evaluation metrics" (you can include other performance metrics in addition to the F1 score used in Kaggle, e.g. evaluation of P and R may explain the observed performance of F1)
    - "Experiments" (describe all the experiments, validation of hyperparemeters, etc).
    - "Results"
    - "Conclusions"
    - "References"
- You can use any number of subsections
- The maximum length of the report is 5 pages.
- All figures and tables should be one-column wide. Each figure/table should contain a numbered caption (below the figure/table). Figures and tables should be cited and explained in the text (ex. Fig.1, Fig2,…,Tab.1,…).
- All equations should be numbered and cited in the text.
- The document should contain useful numeric and graphical results and explanations, that is, those that give insight into the problem and proposed solutions. Explanations should be based on theoretical concepts studied in class.