HW#2 – Machine Learning in Healthcare 336546

This assignment relates to the detection death in the case of heart failure patient. Your goal is to predict if a patient will death from his heart failure.

Introduction to the database :

The database contains 299 patients of heart failure comprising of 105 women and 194 men. All the patients were more than 40 years old, having left ventricular systolic dysfunction and falling in NYHA class III and IV. Follow up time was 4–285 days with an average of 130 days. Disease was diagnosed by cardiac echo report or notes written by physician. Age, serum sodium, serum creatinine, gender, smoking, Blood Pressure (BP), Ejection Fraction (EF), anemia, platelets, Creatinine Phosphokinase (CPK) and diabetes were considered as potential variables explaining mortality caused by CHD.

References paper that uses this database (just in case you want to know a little bit more about it):

Note that they are not doing classification task with this database.

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001

Assignment

This assignment has only a prebuilt notebook that download the database and open it with pandas. You are required to build and present an appropriate notebook to show your experiments and results. Please provide all answers within the notebook (in a markdown cell), labeled carefully based on the question number. In this assignment, you will do the following:

- Explore the data provided.
- Implement linear and non-linear classifiers.
- Model optimization with k-fold cross validation
- Evaluate your model performances with appropriate metrics.
- Present a 2d visualization of multi-featured data.
- Use feature selection tools.

Use the provided HW2 environment and any additional packages you need for this assignment.

Theory Questions (28%)

1) To evaluate how well our model performs at death classification, we need to have evaluation metrics that measures of its performances/accuracy. Which evaluation metric is more important to us: model accuracy or model performance? Give a simple example that illustrates your claim.

2) You are asked to design a ML algorithm to predict which patients are going to death from a heart attack. Relevant patient features for the algorithm may include Age, serum sodium, serum creatinine, gender, smoking, Blood Pressure (BP), Ejection Fraction (EF), anemia, platelets, Creatinine Phosphokinase (CPK) and diabetes. You should choose between two classifiers: the first uses only BP and EF features and the other one uses all of the features available to you. Explain the pros and cons of each choice.

3) Let's consider that we have the choice between linear SVM and logistic regression. Give 2 notable differences between what you will obtain from these models. (You are not expected to give the mathematical definition but more practical differences).

4) What are the differences between LR and linear SVM and what is the difference in the effect/concept of their hyper-parameters tuning?

Coding Assignment (72%)

1) Load the data. Explain any preprocessing. (5%)

2) Perform a test-train split of 20% test. (5%)

3) Provide detailed visualization and exploration of the data. (10%)

   You should at least include:

   a. An analysis to show that the distribution of the features is similar between test and train. See table 1 below.
      i. What issues could an imbalance of features between train and test cause?
      ii. How could you solve the issue?
   b. Plots to show the relationship between feature and label. See Figure 1 below.
   c. Additional plots that make sense given the mostly binary nature of this dataset.
   d. State any insights you have
      i. Was there anything unexpected?
      ii. Are there any features that you feel will be particularly important to your model? Explain why.

4) **Encode** all your categorical data as one hot vector. (5%)

5) Choose, build and optimize Machine Learning Models: (20%)
   a. Use 5k cross fold validation and **tune** the models to achieve the highest test AUC:
      i. Train one or more linear models on your training set
      ii. Train one or more non-linear models on your training set
   b. Report the appropriate evaluation metrics of the train and test sets (AUC, F1, LOSS, ACC).
   c. What performs best on this dataset? Linear or non-linear models?

6) Feature Selection (10%)
   a. As seen previously, a Random Forest Network can be used to explore feature importance. Train a Random Forest on your data.
      i. What are the 2 most important features according to the random forest.
      ii. Does this match up exactly with the feature exploration you did?

Note: Question 7 should only be completed after your lecture on dimensionality reduction

7) Data Separability Visualization: (20%)

a. Perform dimensionality reduction on the dataset so that you can **plot your data in a 2d plot** (show samples with positive and negative labels in different colors).
b. How separable is your data when reduced to just two features?
c. Train the same models above on the dimensionality-reduced training set.
d. Train the same models on the best two features from section 6.
e. What performs better? 2 features of the reduced dimensionality.