Machine Learning 2025 Course – Homework

1. **Background**

The SUPPORT2 (Study to Understand Prognoses Preferences Outcomes and

Risks of Treatment) dataset comprises 9105 individual critically ill patients

across 5 United States medical centers, accessioned throughout 1989-1991 and

1992-1994. Each row concerns hospitalized patient records who met the inclu-

sion and exclusion criteria for nine disease categories: acute respiratory failure,

chronic obstructive pulmonary disease, congestive heart failure, liver disease,

coma, colon cancer, lung cancer, multiple organ system failure with malignancy,

and multiple organ system failure with sepsis. The goal is to determine these

patients’ short-term and mid-term survival rates based on several physiologic,

demographic, and disease severity factors. It is an important problem because

it addresses the growing national concern over patients’ loss of control near the

end of life. It enables earlier decisions and planning to reduce the frequency of

a mechanical, painful, and prolonged dying process.

1. **Dataset**

An accurate description of the dataset features can be found at https://

archive.ics.uci.edu/dataset/880/support2.

Remark 1 The dataset contains information about previously developed

models. Variable aps, sps, surv2m, surv6m, prg2m, prg6m, dnr, and dnrday

contain information that could heavily influence your model predictions and

should thus be excluded.

Remark 2. Due to the high percentage of missing values, there are a

couple of recommended imputation values. According to the HBiostat Reposi-

tory (https://hbiostat.org/data/repo/supportdesc, Professor Frank Har-

rell) the following default values have been found to be useful in imputing miss-

ing baseline physiologic data in Table 1.

Moreover, there are 159 patients surviving 2 months for whom there were

no patient or surrogate interviews. These patients have missing sfdm2.

1. **Requests**

The dataset is provided as is. It is part of the project to develop a fully struc-

tured and correct analysis based on Machine Learning techniques and method-

ologies provided during lessons and laboratories.

**3.1 Tasks**

The project will consist of the following tasks:

* Perform a preliminary analysis of the data. For instance, but not limited to, visualize samples, identify if features are correlated, determine which are most correlated with the target class (for the regression task), and inspect the distribution of samples among classes;
* Develop a regression model to estimate the target variable totcst. Compare different regression algorithms for this task, and eventually perform feature selection procedures.