4^{rd} assignment – Clustering MIMIC data

Introduction to Clinical Data Science, 2022b

1 Preface

- We will compare different imputation approaches for predicting deliveries' outcome.

2 The data

The data is of deliveries' data. The outcome you are asked to predict is apgar5 which is the new born's APGAR score measured 5 minutes after labor.

The data is compsed of two files:

1. Train - train.csv
2. Test - test.csv

3 Prediction model

We will use XGBoost model to predict the outcome. To evaluate the performance of the model we will use the RMSE measure. You will need to train the model based on data imputed using different approaches as described below.

4 Imputation

Use the next type of imputation for data preparation.

1. No imputation, leaving missing data as is. XGBoost can handle missing data.
2. Drop rows with missing data.
3. Mean for continous, mode (most frequent) for categorical
4. Median for continous, mode (most frequent) for categorical
5. kNN imputation
6. Iterative imputation
7. You can add your others imputations if you'd like to.

Compare the perofmance of the model when using different types of imputation. Print a table where rows are imputation methods, and a column is the RMSE on the train data and on the test data.

Don't: impute training and testing indepedently (seperately). Can you think why?

Don't: Merge the train and test data and then impute, and then split again. Can you think why?

Do: Learn how to impute using the training data, and apply the imputation with the same learn parameter on the test data. For example, for mean imputation, compute the mean of a feature in the training data, and the use this mean (and not the mean of the test data) to impute the test data.

5 Notes

- You can use methods from sklearn.impute or from impyute packages.

- Make sure to run all the notebook before submitting.
- Submission should be the notebook itself (if you used Colab, download it fist), not a URL.
- Use Moodle's forum for discussion, questions and answers. In this way everyone can benefit from the questions, answers and discussions.