Heart Failure: predicting hospital re-admission after 6 months

Statistical Learning for Healthcare Data (056867) – A.Y. 2022/2023

Teo Bucci, Giulia Montani and Alice Traversa

POLITECNICO
MILANO 1863

June 6, 2023

Problem statement

Heart failure (HF) is a prevalent condition with high re-admission rates.
Number of HF cases worldwide:

- 33.5 million in 1990;
- 64.3 million in 2017.

Half of the patients diagnosed with HF will be re-admitted **once within a year** and 20% will be re-admitted twice or more.

Problem statement

Heart failure (HF) is a prevalent condition with high re-admission rates.
Number of HF cases worldwide:

- 33.5 million in 1990;
- 64.3 million in 2017.

Half of the patients diagnosed with HF will be re-admitted **once within a year** and 20% will be re-admitted twice or more.

Primary goal of the project

Develop a **prediction** model with focus on **interpretability**.

Parallel objective

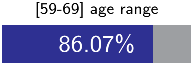
Assess the **importance of drugs** assumption.

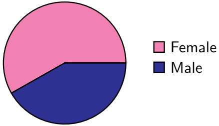
Data

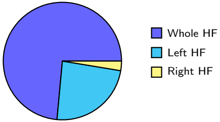
- 2008 patients admitted to a hospital, of which were discarded:
  - 57 dead patients
  - 5 patients with inconsistent information
- 168 variables provided, including:
  - Demographic data (height, sex, occupation, . . .).
  - Medical history (diabetes, comorbidities, . . .).
  - Clinical measurements (pressure, hemoglobyn, . . .).
  - Drugs taken.
  - Re-hospitalizations prior to 6 months (discarded).

Data

- 2008 patients admitted to a hospital, of which were discarded:
  - 57 dead patients
  - 5 patients with inconsistent information
- 168 variables provided, including:
  - Demographic data (height, sex, occupation, . . .).
  - Medical history (diabetes, comorbidities, . . .).
  - Clinical measurements (pressure, hemoglobyn, . . .).
  - Drugs taken.
  - Re-hospitalizations prior to 6 months (discarded).







Data

- 2008 patients admitted to a hospital, of which were discarded:
  - 57 dead patients
  - 5 patients with inconsistent information
- 168 variables provided, including:
  - Demographic data (height, sex, occupation, . . .).
  - Medical history (diabetes, comorbidities, . . .).
  - Clinical measurements (pressure, hemoglobyn, . . .).
  - Drugs taken.
  - Re-hospitalizations prior to 6 months (discarded).

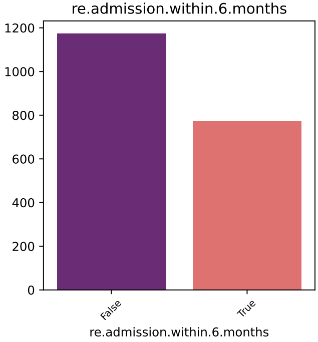
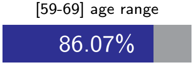
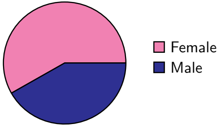
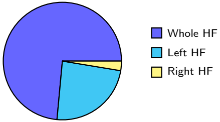


Figure: Target distribution.







Handling missing values

Categorical features

- `Occupation` 1.34%
- Imputation: most frequent

Handling missing values

Categorical features

- `Occupation` 1.34%
- Imputation: most frequent

Numerical features

- 14 features with over 60% missing are discarded
- 9 features between 50% and 60% are discarded after further analysis
- Imputation: KNN with 5 neighbors

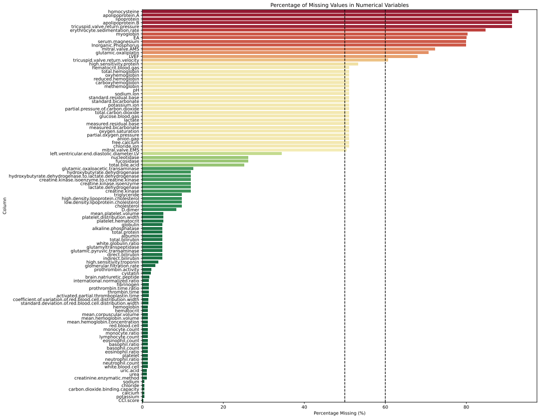


Figure: Percentage of missing values in numeric.

Further cleaning

- **Outlier analysis**
  1. Identification:
     - Sample Z-scores.
     - Physiological limits checked using literature.
  2. Replace with NaN for imputation.

Further cleaning

- **Outlier analysis**
  1. Identification:
     - Sample Z-scores.
     - Physiological limits checked using literature.
  2. Replace with NaN for imputation.
- **Low variance variable:** Remove 16 variables with more than 95% dominance

Further cleaning

- **Outlier analysis**
  1. Identification:
     - Sample Z-scores.
     - Physiological limits checked using literature.
  2. Replace with NaN for imputation.
- **Low variance variable:** Remove 16 variables with more than 95% dominance
- **Correlation analysis:** Remove 12 variables with more than 85% correlation

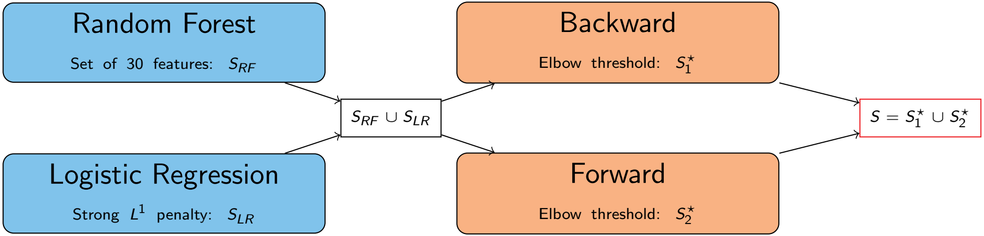
Further cleaning

- **Outlier analysis**
  1. Identification:
     - Sample Z-scores.
     - Physiological limits checked using literature.
  2. Replace with NaN for imputation.
- **Low variance variable:** Remove 16 variables with more than 95% dominance
- **Correlation analysis:** Remove 12 variables with more than 85% correlation

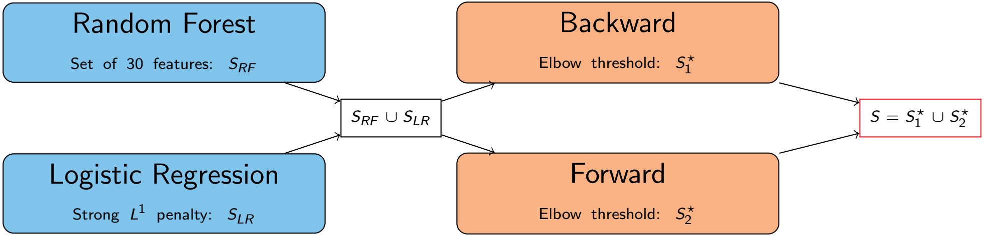
Three variables with possible outliers retained due to their importance:

- `eosinophil.count`
- `high.sensitivity.troponin`
- `glutamic.pyruvic.transaminase`

Feature selection



Feature selection



Strength of the method

- Faster than performing backward selection immediately.
- Takes away a lot of the *greedyness*.
- Takes advantages of both RF and LR (intersection is very small).

Final set of 13 selected variables.

Model selection

Train 6 different classifiers.

Model selection

Train 6 different classifiers.

Metric

- Compare performance between models using **AUC**.

Model selection

Train 6 different classifiers.

Metric

- Compare performance between models using **AUC**.

Training setting

- Preprocessing: one-hot encoding and scaling.
- **Tune hyperparameters** with GridSearchCV.
- Evaluate performance using **Stratified 5-fold** cross-validation (CV).
- 85:15 stratified train-test ratio.
- Always set seed for reproducibility.
- Class imbalance addressed by passing class weights based on sample proportions.

Results

Model	AUC
RandomForestClassifier	0.6769
LogisticRegression	0.6702
GaussianNB	0.6452
DecisionTreeClassifier	0.5943
KNeighborsClassifier	0.5681
MLPClassifier	0.5028

Table: Comparison of performance.

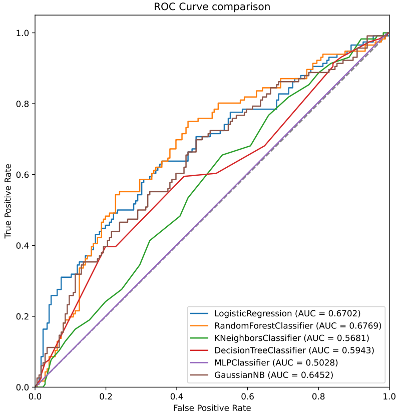


Figure: ROC curves comparison.

Results (cont.)

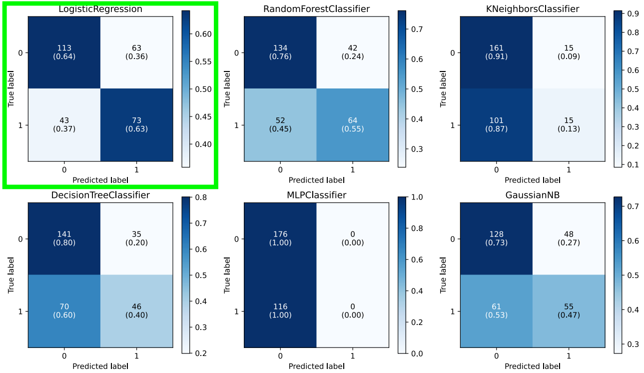


Figure: Confusion matrices comparison (threshold: 0.5).

Logistic Regression performance:

- AUC: **0.6702**
- Accuracy: **0.6370**
- Precision: 0.5368
- Recall: 0.6293
- F1-score: 0.5794

Conclusions

feature	beta	exp_beta
occupation_farmer	-0.6973	0.4979
glutamic.pyruvic.transaminase_log	-0.1305	0.8776
D.dimer	-0.0911	0.9129
partial.pressure.of.carbon.dioxide	-0.0261	0.9743
sodium	-0.0049	0.9951
basophil.ratio	0.0055	1.0055
creatinine.enzymatic.method	0.0066	1.0066
dischargeDay	0.0294	1.0298
eosinophil.ratio	0.0486	1.0498
NYHA.cardiac.function.classification_IV	0.3599	1.4332
diabetes_True	0.4049	1.4991
international.normalized.ratio	0.4074	1.5029
type.of.heart.failure_Both	0.5502	1.7336

Table: LR coefficients.

We found meaningful **interpretations** with clinical facts:

- **Farmers** are less likely to be re-admitted (probably external confounder).
- **D-dimer** seems associated with tissue repair.
- Higher **discharge day** (i.e. longer stay in hospital) is associated with higher risk.
- **Level 4 NYHA**, presence of **diabetes** and having suffered a **Whole HF** highly associated with re-admission.

Conclusions (cont.)

Regarding **drugs**, they were divided into 4 categories:

- Diuretics
- Vasodilatory
- Inhibitor
- Increase force of heart contraction (IFHC)

Conclusions (cont.)

Regarding **drugs**, they were divided into 4 categories:

- Diuretics
- Vasodilatory
- Inhibitor
- Increase force of heart contraction (IFHC)

Takeaways:

- None made it to the final set of features, so their importance is **not fundamental**.

Conclusions (cont.)

Regarding **drugs**, they were divided into 4 categories:

- Diuretics
- Vasodilatory
- Inhibitor
- Increase force of heart contraction (IFHC)

Takeaways:

- None made it to the final set of features, so their importance is **not fundamental**.
- **Most patients** are treated with both **diuretics and vasodilators**, therefore they don't help separation.

Conclusions (cont.)

Regarding **drugs**, they were divided into 4 categories:

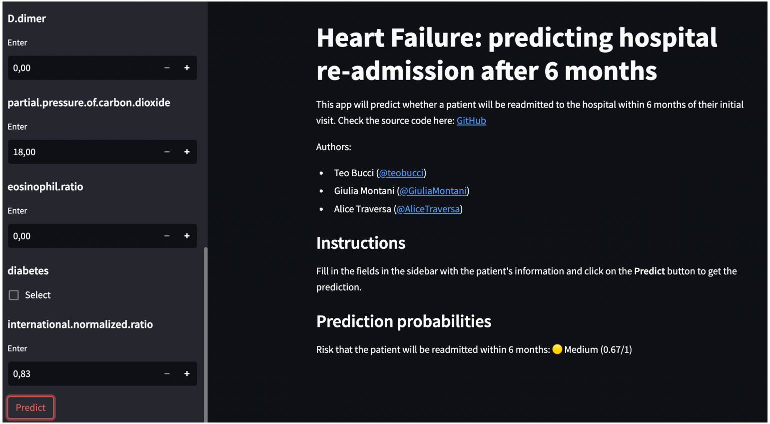
- Diuretics
- Vasodilatory
- Inhibitor
- Increase force of heart contraction (IFHC)

Takeaways:

- None made it to the final set of features, so their importance is **not fundamental**.
- **Most patients** are treated with both **diuretics and vasodilators**, therefore they don't help separation.
- **IFHC** made it to the second step of feature selection, so it's the most informative category.

Deployment

Web app for easy usage by clinicians: `https://teobucci-slhd-app-3iahgf.streamlit.app/`



Limitations and Recommendations

Limitations

- reliance on **single dataset**
- difficulty in **comparing results**
- the sample is **not representative**
- missing data imputation
- lacking data coming from electrocardiography.

Limitations and Recommendations

Limitations

- reliance on **single dataset**
- difficulty in **comparing results**
- the sample is **not representative**
- missing data imputation
- lacking data coming from electrocardiography.

Recommendations

- better **management of missing values** in the data
- further validation with **external datasets**
- for the sake of **performance** only, keep more variables and explore more models, at the cost of simplicity

Thank You!

https://github.com/teobucci/slhd

References

[1] N. L. Bragazzi, W. Zhong, J. Shu, *et al.*, “Burden of heart failure and underlying causes in 195 countries and territories from 1990 to 2017,” *European Journal of Preventive Cardiology*, vol. 28, no. 15, pp. 1682–1690, Feb. 2021.

[2] A. Groenewegen, F. H. Rutten, A. Mosterd, and A. W. Hoes, “Epidemiology of heart failure,” *European Journal of Heart Failure*, vol. 22, no. 8, pp. 1342–1356, Jun. 2020.

[3] Z. Zhang, L. Cao, R. Chen, *et al.*, “Electronic healthcare records and external outcome data for hospitalized patients with heart failure,” *Scientific Data*, vol. 8, no. 1, Feb. 2021.

[4] E. Lonn, “Regular review: Drug treatment in heart failure,” *BMJ*, vol. 320, no. 7243, pp. 1188–1192, Apr. 2000.