Myocardial infarction complications prediction

Statistical Learning for Healthcare Data

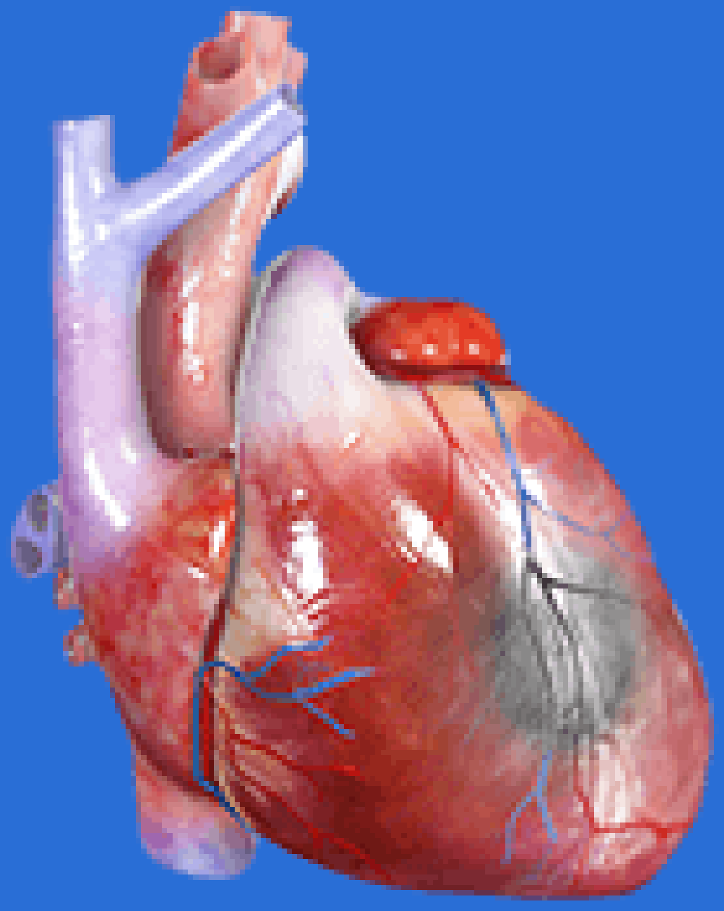
MSc. Biomedical Engineering
MSc. Mathematical Engineering

POLITECNICO
MILANO 1863

Presented by:
Sofia Di Filippo
Chiara Mocetti
Gabriele Moro
Filippo Pagella

Date Presented:
June 26, 2023

Problem description and project goals



The problem:

- **Myocardial Infarction** (MI) occurs when the blood stops flowing properly in a part of the heart causing injuries in its tissues because of lack of oxygen.
- MI is **widely spread in all countries** where people are more exposed to **chronic stress and unhealthy lifestyles**.
- It is **difficult to foresee** outcomes for MI patients.

Project goals:

1. To present a model able **to predict MI complications**.
2. To show **how the prediction** of our model **changes** if we consider only the information obtained at the time of admission and those **at the end of the** 3 days **hospitalization**.

Data description

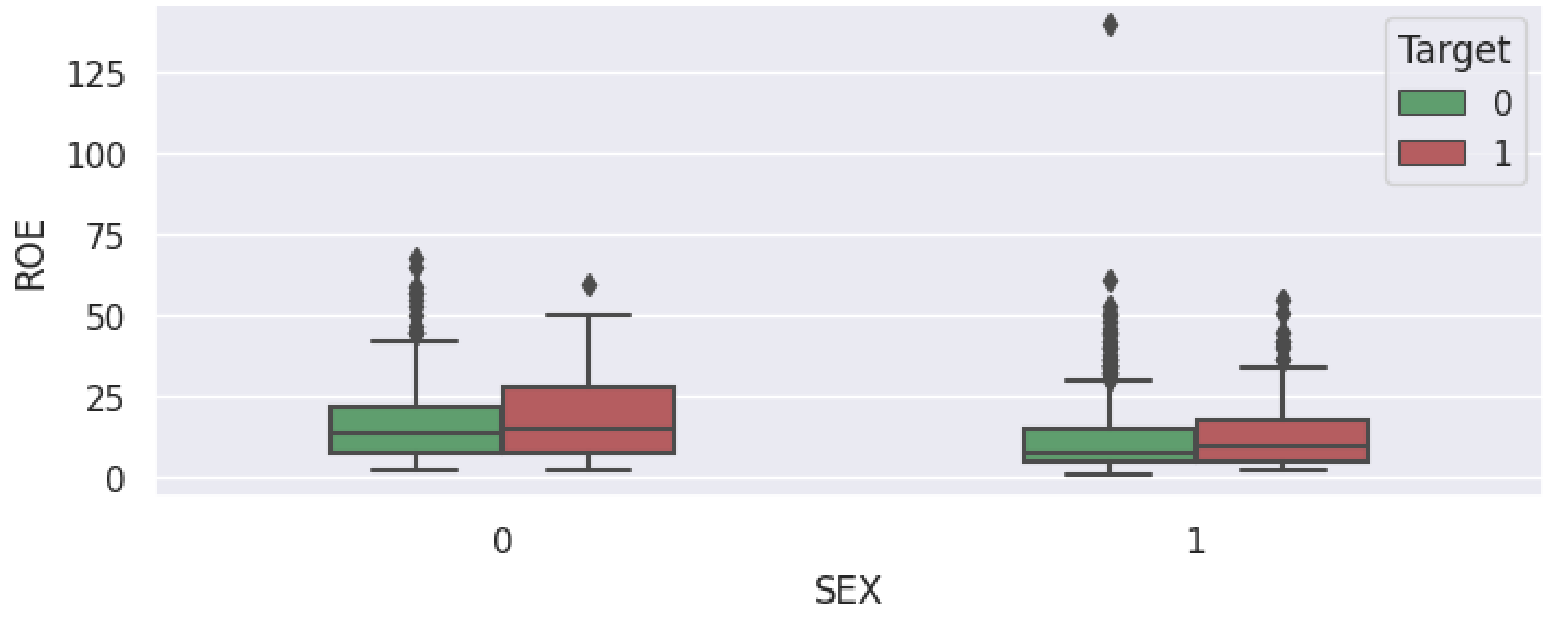
- Data from the Krasnoyarsk Interdistrict Clinical Hospital, Russia
- Collected between 1992 and 1995
- **1700** patients, **124** features
  - 1-112, clinical picture of the patients
  - 113-124, possible complications

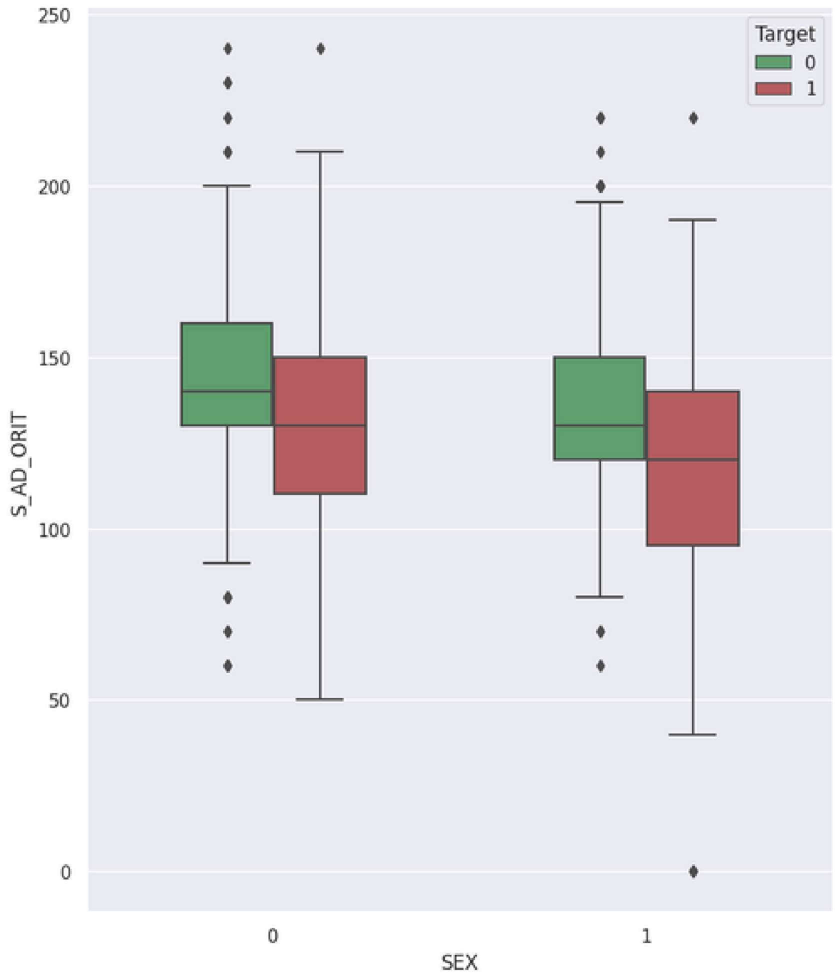
ID	AGE	SEX	INF_ANAM	STENOK_AN	FK_STENOK	IBS_POST	IBS_NASL	GB	SIM_GIPERT	...	RESSLER	ZSN	REC_IM	P_IM_STEN	LET_IS
1	77.0	1	2.0	1.0	1.0	2.0	NaN	3.0	0.0	...	0	0	0	0	0
2	55.0	1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0	0	0	0	0
3	52.0	1	0.0	0.0	0.0	2.0	NaN	2.0	0.0	...	0	0	0	0	0

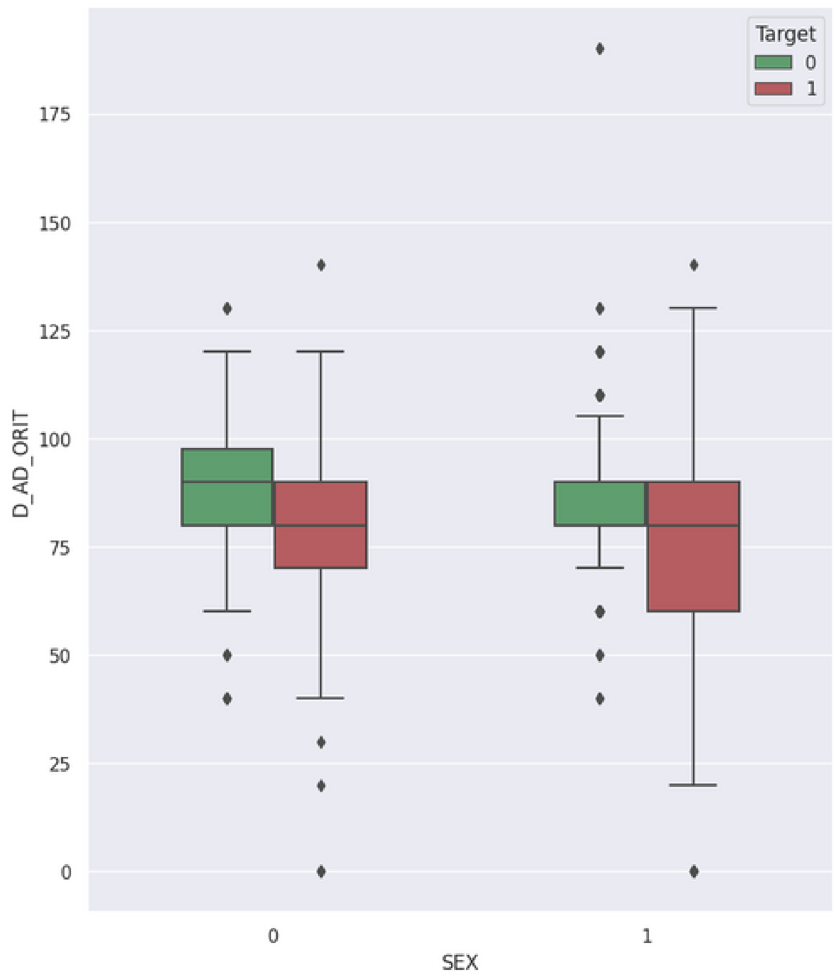
1. 113-123 were dropped
2. '**LET_IS**' becomes '**Target**'
3. 9 features referring to different times of the hospitalization, reorganized in just 3

Data cleaning and outliers detection

1. Deletion of features with > **60%** of missing values
2. Deletion of patients with > **20%** of missing values
3. Outliers detection on the continuous features







Missing values imputation

"Manual" imputation:

1. Categorical features with < 20% of missing values imputed by **most frequent value**
2. Continuous features with < 20% of missing values imputed by **mean**
3. Features with > 20% of missing values
   a. "Not important" features were dropped
   b. **Deterministic regression imputation**

Multiple Imputation by Chained Equations (MICE):

1. Multiple imputation on the continuous features, initialization by **mean**
2. Multiple imputation on the categorical features, initialization by **most frequent value**

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.90	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.47	1.14	1.28
0.89	1.23	1.45

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	1.24	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45

Data pre-processing

01

Shuffling and train/test splitting

02

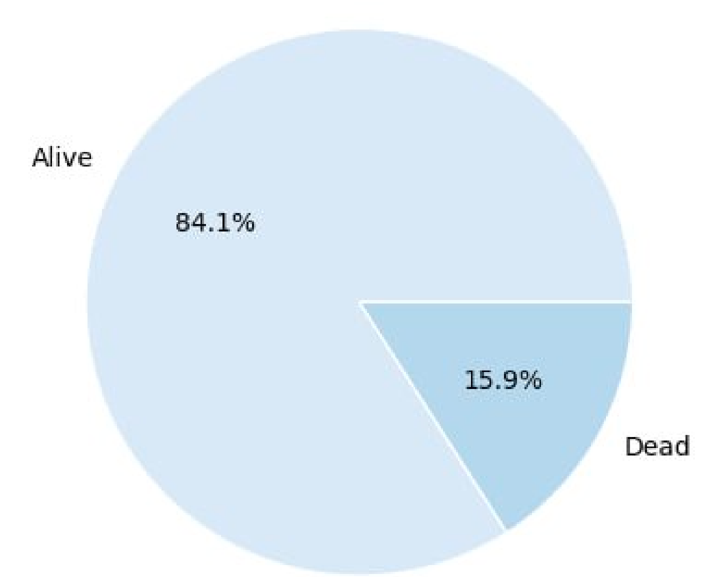
Normalization

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

03

Unbalanced classes

→ **SMOTE**



Measure of Performance (MoP)

Due to the **healthcare domain**, we need a prudential model that classifies dead accurately, without losing reliability:

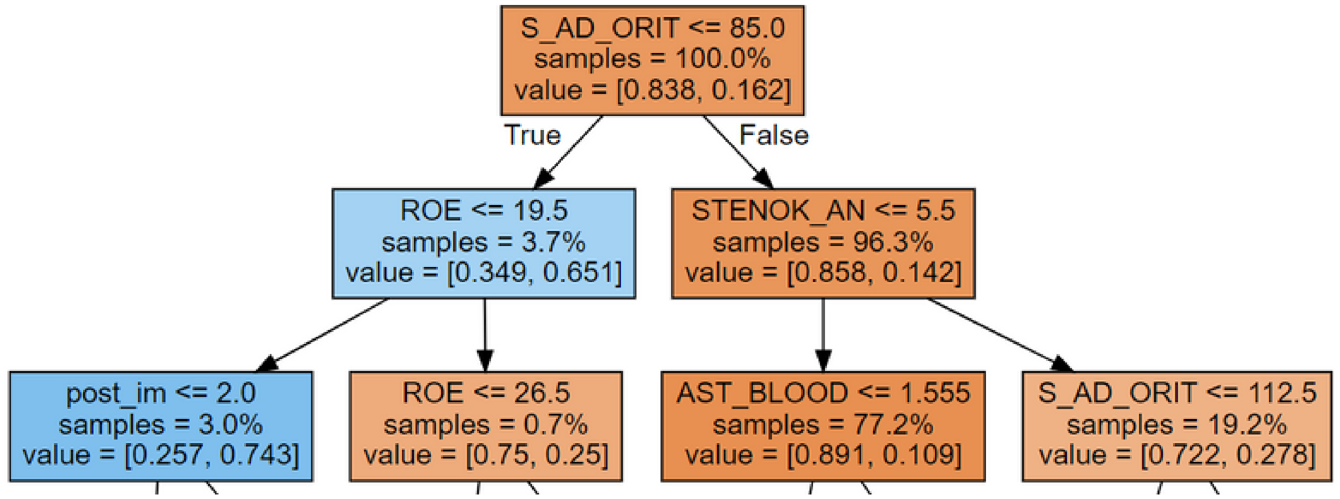
$$\mathrm{MoP} = 0.5 * \mathrm{Sensitivity} + 0.3 * \mathrm{Precision} + 0.2 * \mathrm{Specificity}$$

Later, models hyperparameter tuning will be done maximizing this new measure.

Preliminary methods

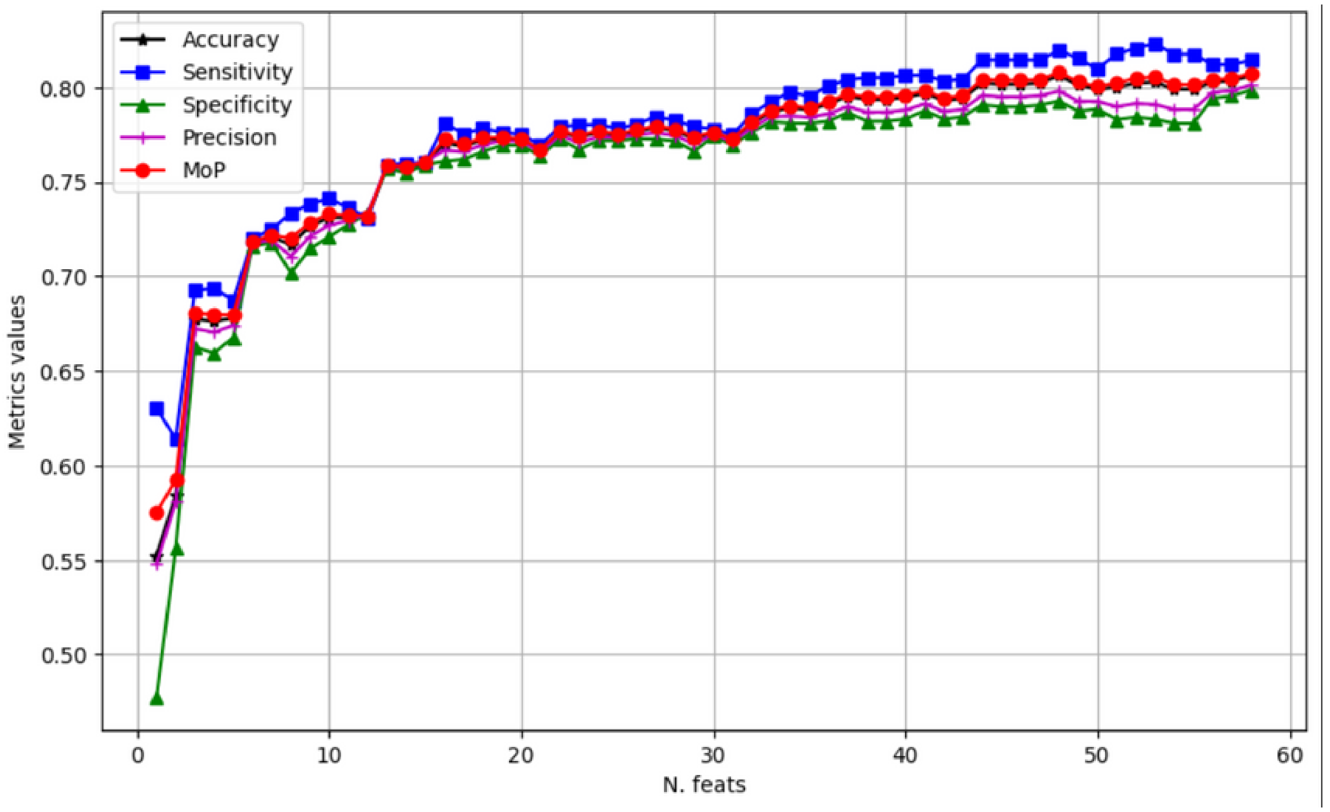
First method: Random Forest Classifier

- Grid search cross-validation
- Pros: categorical data
- Cons: interpretability



Second method: Stepwise Logistic Regression

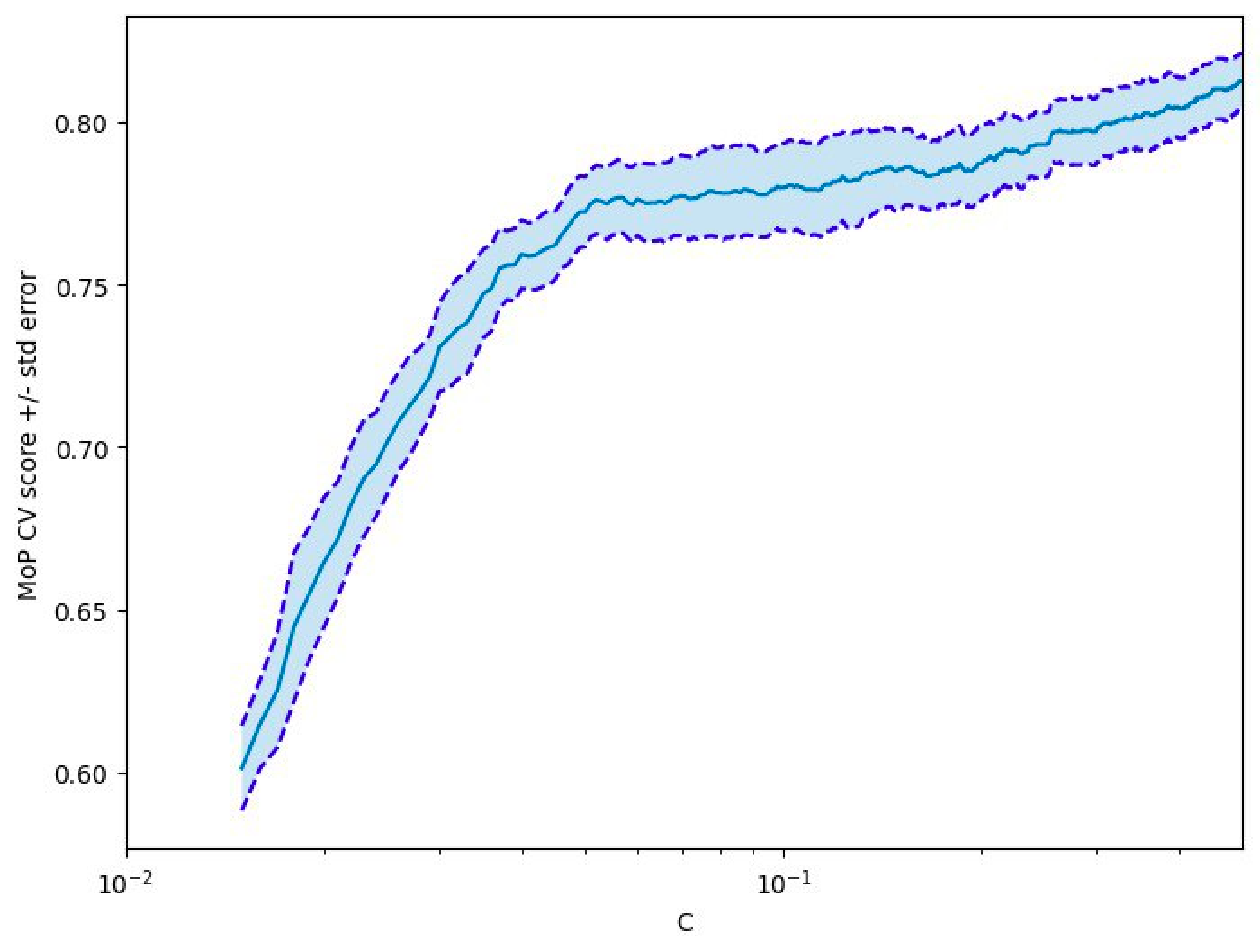
- Features importance ranking imported from RF
- Evaluate the "**elbow**" of MoP to perform feature selection
- Interpretability but lost of non linear information



Final model: Logistic Regression

We added an **L1 penalty**, in order to solve collinearity and perform feature selection.

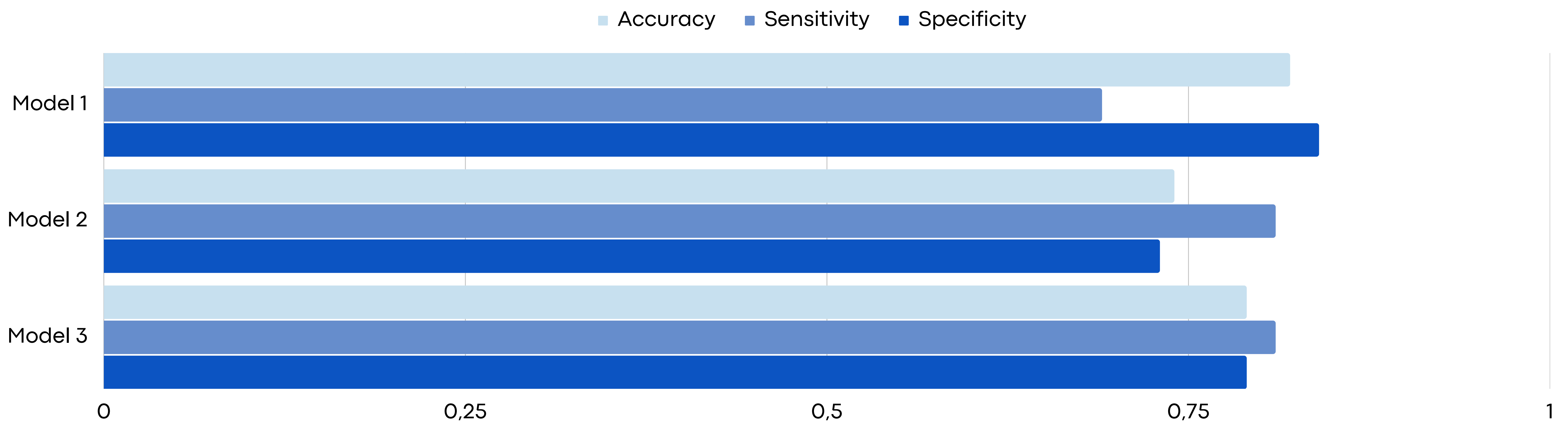
- The choice of the optimal **parameter C** has been made through the maximization of MoP.
- Observing the plot, we noticed an **elbow** in correspondence of the value C=0.1.
- We obtained a model with satisfying performances, making use of **28 out of the original 93 variables**.



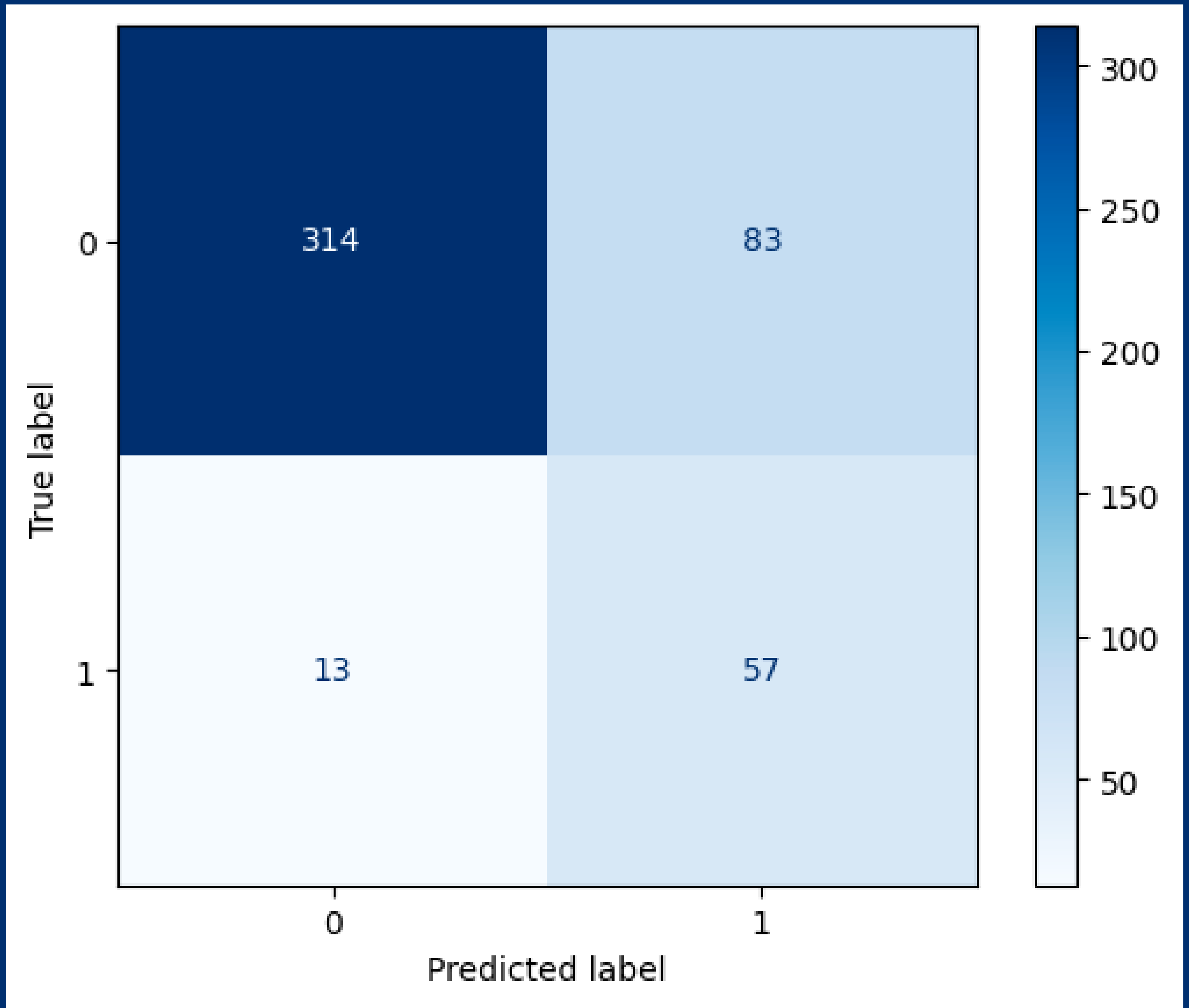
Performance comparison

From the results on the test set Model 3 is the more complete

- We achieve **81% of sensitivity**, still obtaining good values for both accuracy and specificity



Results



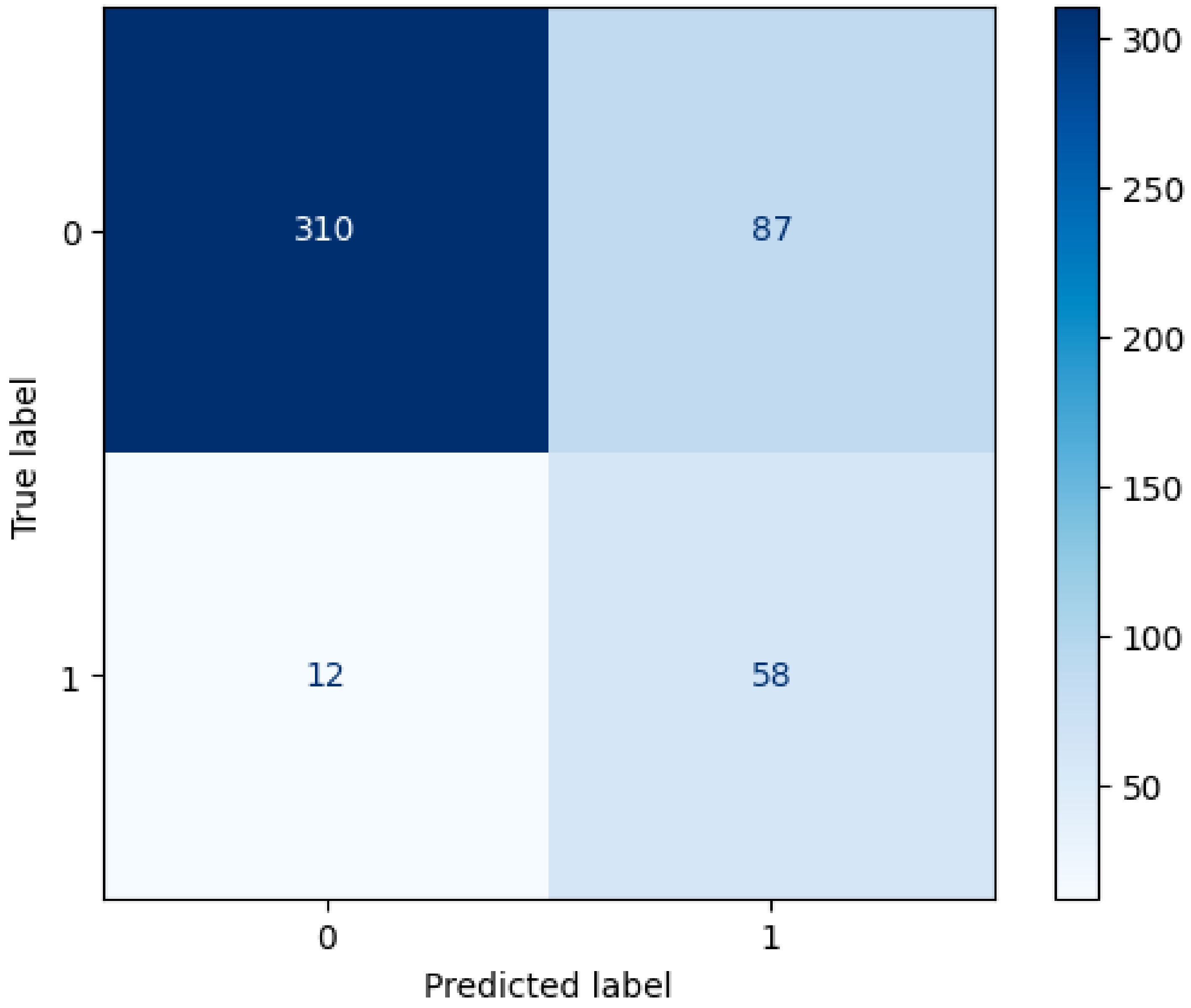
The model controls false negatives

As remarked before, the main goal of our study has been, in this medical framework, to build a non trivial classifier, able to **predict correctly whether a patient will suffer from a lethal outcome after a MI**.

The model is interpretable

The context in which we conduct our analysis needs interpretability of results.
This is the reason why we decided to award a model that not only had a performance power, but also could give **understandable outcomes**, exploitable by the medical community and applicable in clinical scenarios.

Adding information



In the dataset also information relative to the hospitalization period were available.

We added those variables to our final model and we evaluated if and how the results change.

We did not notice any substancial modification, with respect to the previous confusion matrix.

Conclusions

The final model selects 28 features which consider:

- **Sociodemographic characteristics** (Age, Sex);
- **Risk factors** (Diabetes, Obesity, Hypertension);
- **Presentation characteristics** (Arrhythmias, Cardiogenic shock);
- **Initial diagnostic studies** (ECG);
- **Pharmacological treatments** (Beta Blockers, Nitrates, Calcium Channel Blockers, Aspirin).

Analysis of coeffiecients

Atrial Fibrillation	Cardiogenic Shock	Diabetes
0,5603	0,1495	0,1835

Beta Blockers	Calcium Channel Blockers	Acetylsalicylic acid
-0,3055	-0,5294	-0,6735

To sum up

- There is **coherence** between the current models used to estimate patient's risk of mortality after MI and our model.
- The main important factors emerging from the final model are: white blood cell counts, systolic blood pressure, age, use of liquid nitrates, complete RBBB on ECG. These variables are relevant in literature.
- Adding information of the hospitalization period does not improve the predictive performances of the model.

We are satisfied by the final results. This model could be of help in real clinical scenarios.

Thank you for your attention