DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

STATISTICAL METHODS IN DATA MINING

MEGIE, MECD, MMAC, OTHERS, 1^{st} SEMESTER, 2022/2023

Project 1

Handed out on October 10, 2022.

To be **handed back** by **November 3**, 2022.

Myocardial Infarction (MI) is one of the most challenging problems of modern medicine. Acute myocardial infarction is associated with high mortality in the first year after it. The incidence of MI remains high in all countries. This is especially true for the urban population of highly developed countries, which is exposed to chronic stress factors, irregular and not always balanced nutrition.In this regard, predicting complications of myocardial infarction in order to timely carry out the necessary preventive measures is an important task.

The researcher has rudimentary knowledge about Data Mining and wants your help to work on his/her problem.

The proposed dataset can be used to solve an important problem: predicting complications of Myocardial Infarction based on information about the patient at different time moments.

You can grab the dataset: here.

Problems to solve

In general columns 2-112 can be used as input data (explanatory variables). Possible complications (outputs - response variables) are listed in columns 113-124.

There are four possible time moments for complication prediction: on base of the information known at:

Question 1 - the time of admission to hospital: all input columns (2-112) except 93, 94, 95, 100, 101, 102, 103, 104, 105 can be used for prediction;

Question 2 - the end of the first day (24 hours after admission to the hospital): all input columns (2-112) except 94, 95, 101, 102, 104, 105 can be used for prediction;

Question 3 - the end of the second day (48 hours after admission to the hospital) all input columns (2-112) except 95, 102, 105 can be used for prediction;

Question 4 - the end of the third day (72 hours after admission to the hospital) all input columns (2-112) can be used for prediction.

The four problems associated with the above four questions and a corresponding response variable are distributed by each group as follows:

Table 1: Assignment of each problem to each group of students.

Number	Name	Group	Question/Response variable
102404	Gonçalo de Matos Silva Sousa Félix	1	Question 4
104078	Karen Rodrigues Villarroel	1	Atrial fibrillation (FIBR_PREDS)
87302	Ana Rita Teodoro Apolinário	1	Column 113
89671	Inês Ferro Passinhas de Medeiros Rodrigues	1	
93139	Miguel André Grácio	1	
105325	Klara Penko	1	
54563	Ângelo Eduardo de Oliveira Xavier Sanches da Silva	2	Question 3
86821	António Pedro Baptista Nunes Simões de Carvalho	2	Supraventricular tachycardia (PREDS_TAH)
87087	Miguel Pereira da Graça Mira de Oliveira	2	Column 114
90486	Maria Margarida De Araújo Matos Barbosa	2	
93530	Helena Barbier Tello	2	
87224	José Antunes Pedroso de Sousa Franco	3	Question 1
92637	João Vasco Certal Afonso	3	Ventricular tachycardia (JELUD_TAH)
92642	Miguel Filipe De Brito Morujo da Graça	3	Column 115
93359	Vasco Pimentel de Carvalho Marques Carneiro	3	
98461	Maria Araújo Fontaínha	3	
92630	Gonçalo Francisco de Meneses Marques	3	
95758	Nuno Pereira da Costa Figueiredo Marques	4	Question 3
96193	Diogo Miguel Pimpão Vilela	4	Ventricular fibrillation (FIBR_JELUD)
96260	José Pedro Martins Campião Antunes	4	Column 116
96301	Pedro Maria Dias Ferreira Fernandes Rodrigues	4	
96321	Sebastião Miguel Gomes Pereira Lemos Caldas	4	
88195	Leonardo Lourenço Monteiro	4	
92635	João Carlos Pereira Fernandes	5	Question 1
95744	Guilherme Marques Antunes	5	Dressler syndrome (DRESSLER)
95751	José Maria Nogueira de Pinho Marques	5	Column 120
95763	Rúben Bastos da Silva	5	
95764	Wanghao Zhu	5	
93434	Bernardo Pavoeiro Santos	5	
102190	Filipa Ferreira de Silva Real Marques	6	Question 3
83933	Manuel Oliveira Santos Leite Garcia	6	Post-infarction angina (P_IM_STEN)
90586	Alexandre Stepanov Bukovac	6	Column 123
92829	Maria Margarida Sá Costa Valente	6	
101207	Jessi Bashakan Vieira Arsénio	6	
96557	Nuno Miguel Matos Torgo Gonçalves	6	

Table 2: Assignment of each problem to each group of students.

Number	Name	Group	Question/Response variable
101256	Mina Golshenas Rad	7	Question 3
104798	Maarten Harmen van Veen	7	Relapse of the myocardial infarction (REC_IM)
104838	Wessel Bernardus Cornelis Geerlings	7	Column 122
86809	Thomas Fernando Dinis Gaehtgens	7	
98888	Davood Fanaei Sheikholeslami	7	
93119	Manuel António Gaspar Mendes	7	
104860	Erlend Nonaas Lokna	8	Question 2
104861	Thomas Fardal Rødland	8	Third-degree AV block (A_V_BLOK)
105070	Robert Siridol Kjellberg	8	Column 117
105180	Roméo Paul-Loup Axel Legoupil	8	
83963	Ricardo António Lourenço Pimentel	8	
95578	Francisco Manuel Leal Mitha Ribeiro	8	
102328	João André Resende Dias da Silva	9	Question 4
102363	João José Cardoso do Amaral	9	Pulmonary edema (OTEK_LANC)
81341	Diogo Miguel Branco Marques Calhamar Soares	9	Column 118
89844	Afonso Queiroz Minderico	9	
104785	Hadrien Marie Raoul Nobels	9	
102211	Miguel Bernardo Guerreiro Pereira	10	Question 3
105319	Guilherme Reis Prudêncio de Neto Lopes	10	Myocardial rupture (RAZRIV)
105700	Aníbal Tiago Marques Morais da Câmara Pires	10	Column 119
105774	José Pedro Santos Guedes	10	
	Luis Miguel Carmona	10	
63404	Carlos Correia Martins	10	

Steps

1. Perform a exploratory data analysis and discuss what you have learned from this analysis.

2. Solve your classification problem. Consider several classification methods and discuss how can they contribute for the solution of your problem. Include in your discussion topics such as the options that you have made in building each classifier; interpretation of results; validation of the methods used and possible assumptions; advantages and disadvantages of each alternative; etc. Have in mind that some of the explanatory variables may be irrelevant to the classification problem and that you may need to do some preprocessing methodologies of your data set e.g. dimensionality reduction techniques.

3. Imagine you are going to meet the researcher who contacted you. Report to him/her what you have learned about the problem. Discuss limitations of the analysis you have done and provide suggestions for future work.

About the report:

- The report should not exceed 25 pages.
- Do not forget topics such as objectives; estimation methods; decisions; conclusions of the study; and references.
- The R code (or other code), the data cleaned or transformed (where applicable) as well as the report, must be send by email to:
  `conceicao.amado@tecnico.ulisboa.pt`