

DATA MINING

Definition and project assignment

1 Name of group components

- Ibáñez Pérez, Raúl
- Lao Tebar, Diego
- Lázaro Costa, Carlos
- López Alcácer, Albert
- Ribes Marzá, Albert
- Roldán Montaner, Carlos

2 Data Source

We got our dataset from the Machine Learning Repository of the Center for Machine Learning And Intelligent Systems (Bren School of Information and Computer Science)

Repository url: <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>

Dataset url: <http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>

We are going to use the Hungarian, Long Beach and Switzerland non processed .data sources.

3 Process to get data

Data was downloaded from <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>.

The process to get data was done by the doctors and analysts from Hungary, Long Beach and Switzerland hospitals.

Each data record is a copy of the patient medical results. All the centers used the same unit representation for the medical analysis to keep the data consistency.

4 What data is about

Our data is a merge of four different datasets concerning heart disease diagnosis. Each instance collects different conditions or variables (numerical, binary and qualitative) of a patient, which can be used to predict the presence of a heart disease in that patient.

The data was collected from:

- Hungarian Institute of Cardiology, Budapest by Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland by William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland by Matthias Pfisterer, M.D.

- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation (The data is corrupted - Discarded) by Robert Detrano, M.D., Ph.D.

The Cleveland data is corrupted, so we discarded it.

5 Structure of data matrix

- **Number of records:**
 - ~~Cleveland: 303~~ (Discarded)
 - Hungarian: 294
 - Switzerland: 123
 - Long Beach VA: 200
 - **Total: 617**
- **Number of variables:** 76 (including the predicted one)
- **Number of numerical variables:** 55
- **Number of binary variables:** 15
- **Number of qualitative variables:** 6
- **Number and % of missing data per each variable:**

<i>ITEM</i>	<i>#Entities</i>	<i>MissPercentage</i>
ID	0	0%
CCF	0	0%
AGE	0	0%
SEX	0	0%
PAINLOC	282	31,37%
PAINEXER	282	31,37 %
RESTREL	286	31,81%
PNCADEN	899	100%
CP	0	0%
TRESTBPS	59	06,56%
HTN	34	03,78%
CHOL	30	03,34%
SMOKE	669	74,42%
CIGS	420	46,72%
YEARS	432	48,05%
FBS	90	10,01%
DM	804	89,43%
FAMHIST	422	46,94%
RESTECG	2	0,22%
EKGMO	53	05,90%
EKGDAY	54	06,01%
EKGYR	53	05,90%
DIG	68	07,56%
PROP	66	07,34%
NITR	65	07,23%
PRO	63	07,01%

<i>ITEM</i>	<i>#Entities</i>	<i>MissPercentage</i>
DIURETIC	82	09,12%
PROTO	112	12,46%
THALDUR	56	6,23%
THALTIME	453	50,39%
MET	105	11,68%
THALACH	55	6,12%
THALREST	56	6,23%
TPEAKBPS	63	7,01%
TPEAKBPD	63	7,01%
DUMMY	59	6,56%
TRESTBPD	59	6,56%
EXANG	55	6,12%
XHYPO	58	6,45%
OLDPEAK	62	6,90%
SLOPE	308	34,26%
RLDV5	425	47,27%
RLDV5E	142	15,80%
CA	608	67,63%
RESTCKM	899	100%
EXERCKM	898	99,89%
RESTEF	871	96,89%
RESTWM	869	96,66%
EXEREF	897	99,78%
EXERWM	894	99,44%
THAL	477	53,06%
THALSEV	769	85,54%

<i>ITEM</i>	<i>#Entities</i>	<i>MissPercentage</i>
THALPUL	855	95,11%
EARLOBE	898	99,89%
CMO	11	01,22%
CDAY	9	01,00%
CYR	9	01,00%
NUM	0	0%
LMT	275	30,59%
LADPROX	236	26,25%
LADDIST	246	27,36%
DIAG	558	62,07%
CXMAIN	235	26,14%
RAMUS	567	63,07%

<i>ITEM</i>	<i>#Entities</i>	<i>MissPercentage</i>
OM1	271	30,14%
OM2	572	63,63%
RCAPROX	245	27,25%
RCADIST	270	30,03%
LVX1	19	02,11%
LVX2	19	02,11%
LVX3	19	02,11%
LVX4	19	02,11%
LVF	16	01,78%
CATHEF	588	65,41%
JUNK	780	86,76%
NAME	0	0%

- % of missing data in the whole data matrix: 31,09%