Preparant les dades dels qüestionaris per analitzar

Xavier Duran
GCAT Genomes for Life

15 de Juny del 2016

Qüestionari epidemiològic

- Onyx
- Opal
- 4000+ variables del qüestionari
- 3000+ variables filtrades
- Exportació diària a P:/gcat-cohort/output/export (R + cron)

Selecció i agrupació de variables

- Hàbits alimentaris
- Dieta mediterrània (PREDIMED)
- Activitat física
- Tabac
- Alcohol
- Cribatges
- ...

Magma Javascript API

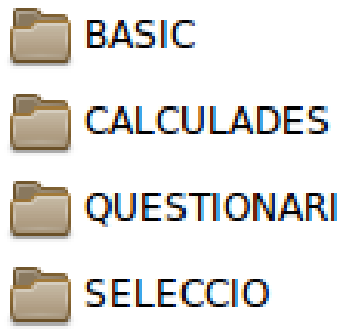
```
resultado = 0

cribado = $('CRIBADO_MUJERES_MAMOGRAFIA_RESULTADO')
resultado_anormal = $('CRIBADO_MUJERES_MAMOGRAFIA_RESULTADO_ANORMAL')

if (cribado == 1) {
    resultado = 1
}

if (cribado == 2) {
    resultado = resultado_anormal.map({
        0: 20,
        1: 21,
        2: 22,
        3: 23,
        4: 24,
        5: 25,
        6: 26
    }, 0, 0)
}

resultado
```



Eines de preprocés de dades

Figure 1: Python

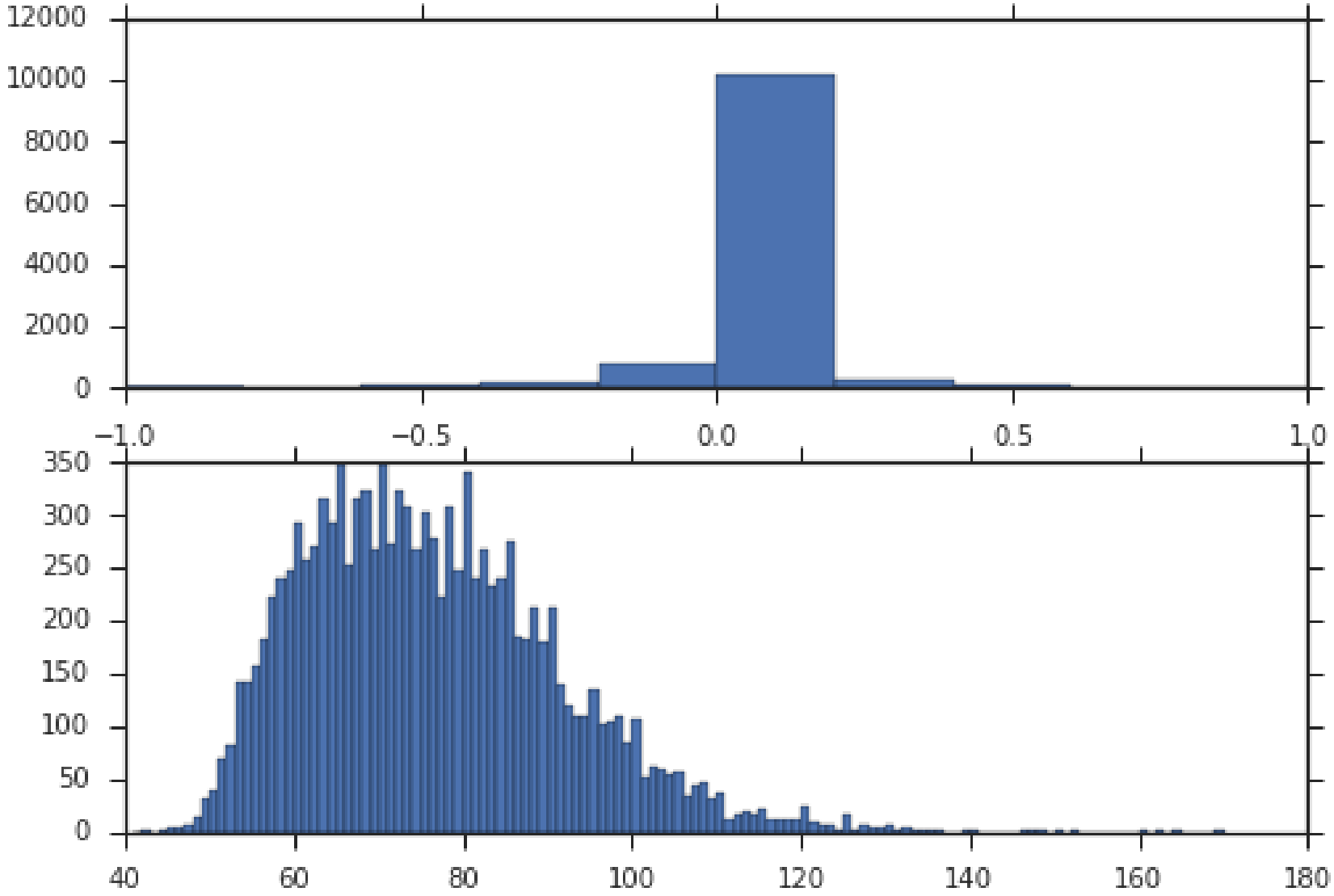
Preprocessament de dades

Mesures antropomètriques

- Alçada
- Pes
- Cintura
- Maluc
- Tensió arterial
- Pols
- BMI
- WHR

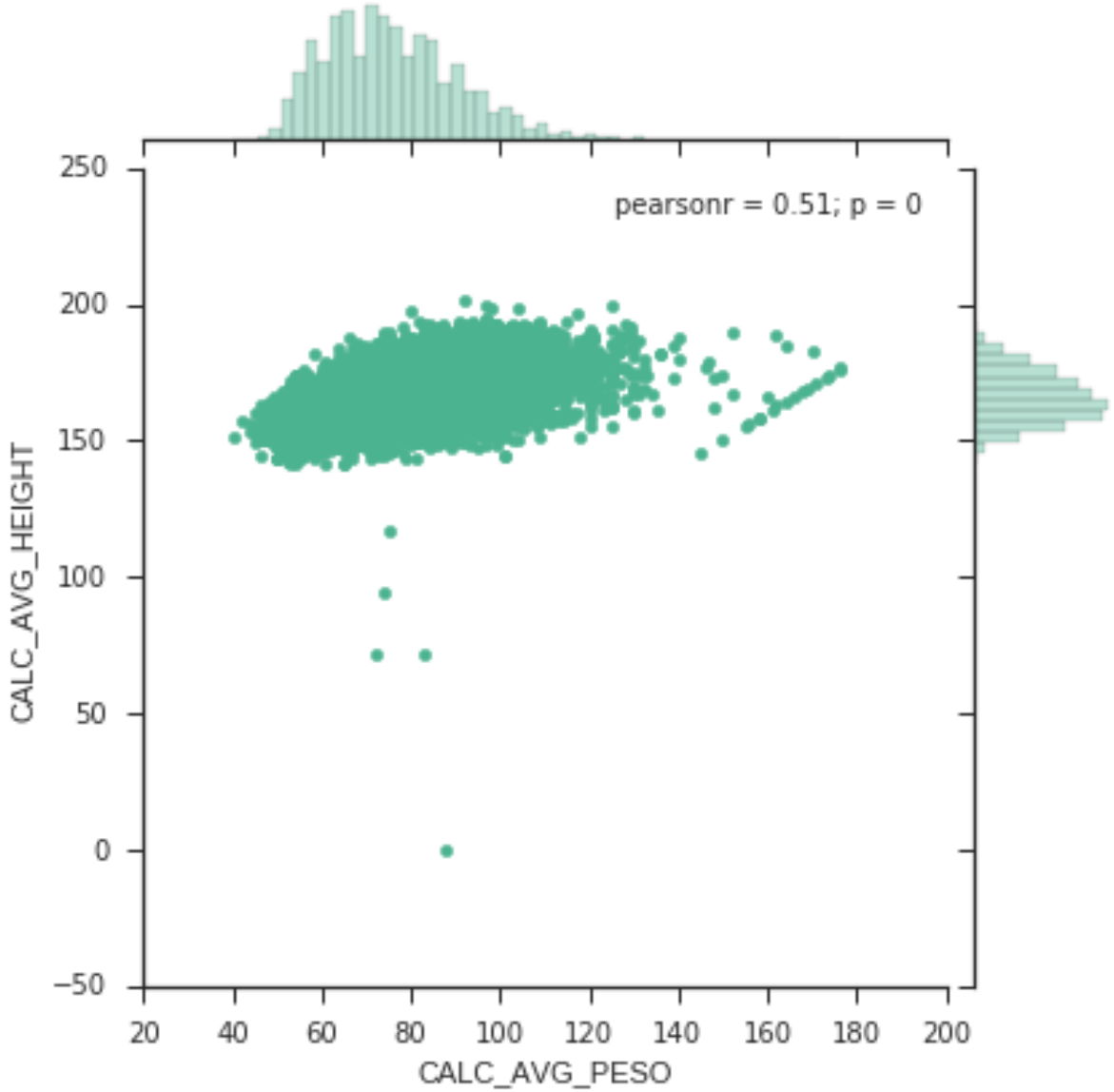
Preprocessament de dades

Pes



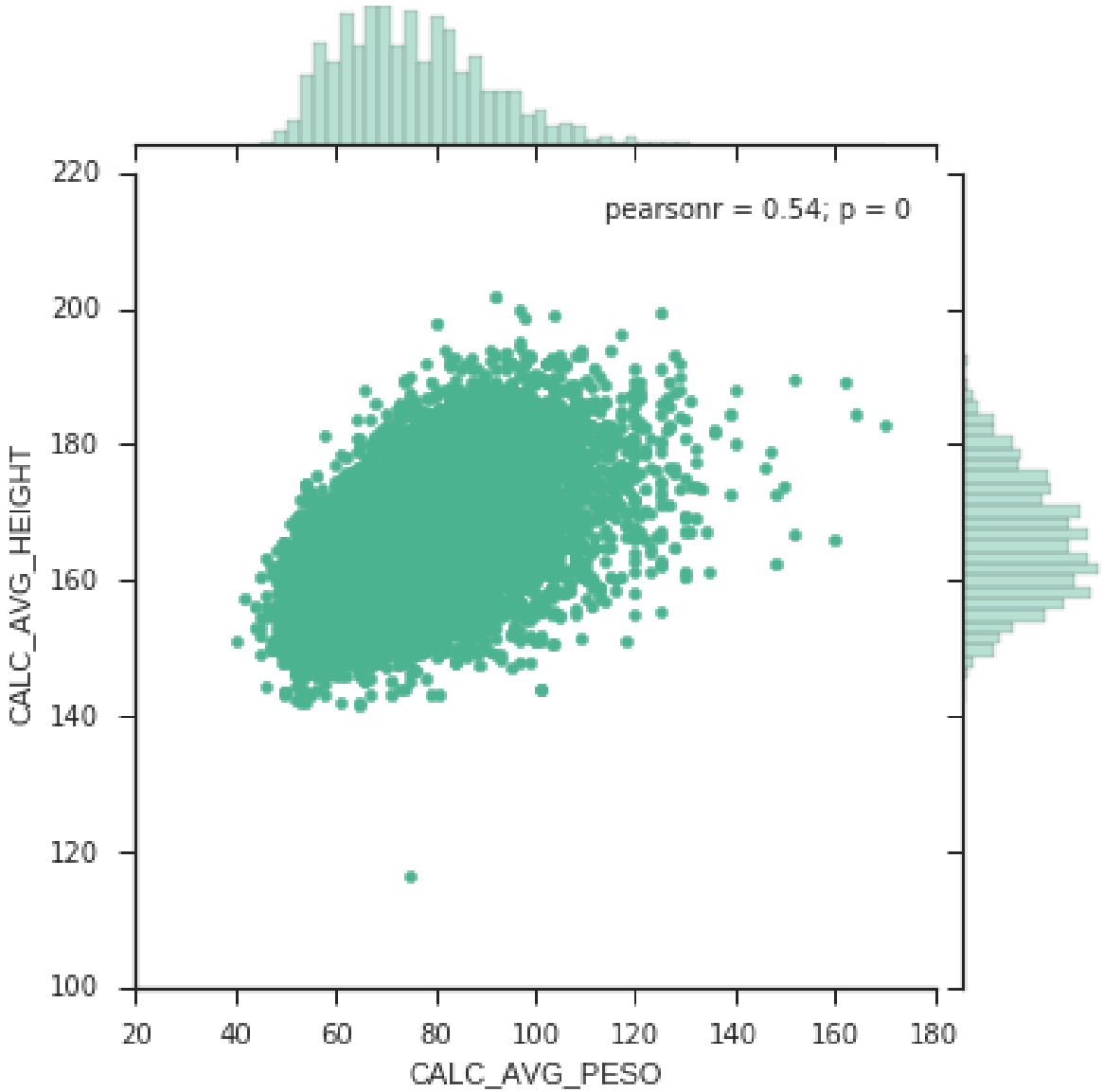
Preprocessament de dades

Errors sistemàtics



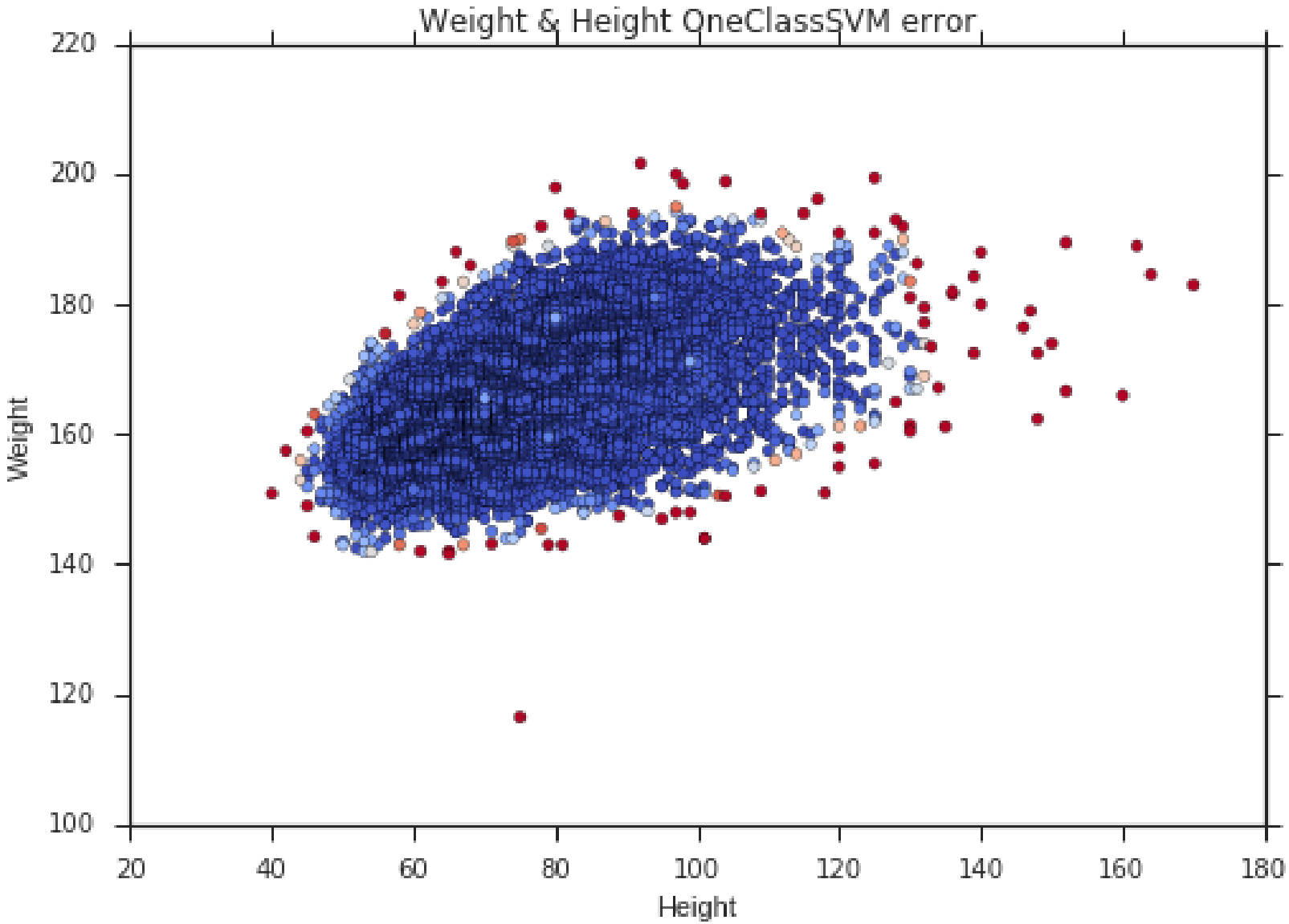
Preprocessament de dades

Errors sistemàtics



Preprocessament de dades

Outliers



Preprocessament de dades

Coherència en les dades

- Gènere
- Edat i data de naixement
- Localitzacions

Preprocessament de dades

Medicació

	entity_id	variable	value	CODI_ATC	NOMBRE
212502	=E00251439519321	HTA	C09AA;LISINOPRIL	C09AA	LISINOPRIL
212507	=E00251513629021	HTA	C09CA;LOSARTAN	C09CA	LOSARTAN
212509	=E00251514739921	HTA	C09DA;LOSARTAN/HIDROCLOROTIAZIDA	C09DA	LOSARTAN/HIDROCLOROTIAZIDA
212546	=E00251518701421	HTA	C09AA;ENALAPRIL	C09AA	ENALAPRIL
212559	=E00251511135221	HTA	C03AA;HIDROSALURETIL	C03AA	HIDROSALURETIL

	entity_id	variable	value
212600	=E00251513756521	HTA	ENAPRIL
212746	=E00251432344121	HTA	Amlodipino 5
212861	=E00251510885721	HTA	simvastatina
212931	=E00251511139621	HTA	Amlodipino
212957	=E00251514740321	HTA	Esídres

Preprocessament de dades

Medicació

	variable	value	CODI_ATC	NOMBRE
213182	HTA	COVALS	C09DA	CO VALS
213275	HTA	Lisonopril	C09AA	LISINOPRIL
213524	HTA	CO-DIOVAN	C09DA	CO DIOVAN
213558	HTA	CO-DIOVAN	C09DA	CO DIOVAN
213616	HTA	ENNALAPRIL	C09AA	ENALAPRIL

Falta comprovar coherència entre la condició i la medicació reportada

Codificació de variables

Creació de dos datasets per utilitzar segons l'algorisme de machine learning

- Categoritzat
- Binari: Dummy Variables (one-hot encoding)

	ETNIA_PARTICIPANTE
0	3
1	1
2	1
3	NaN
4	1

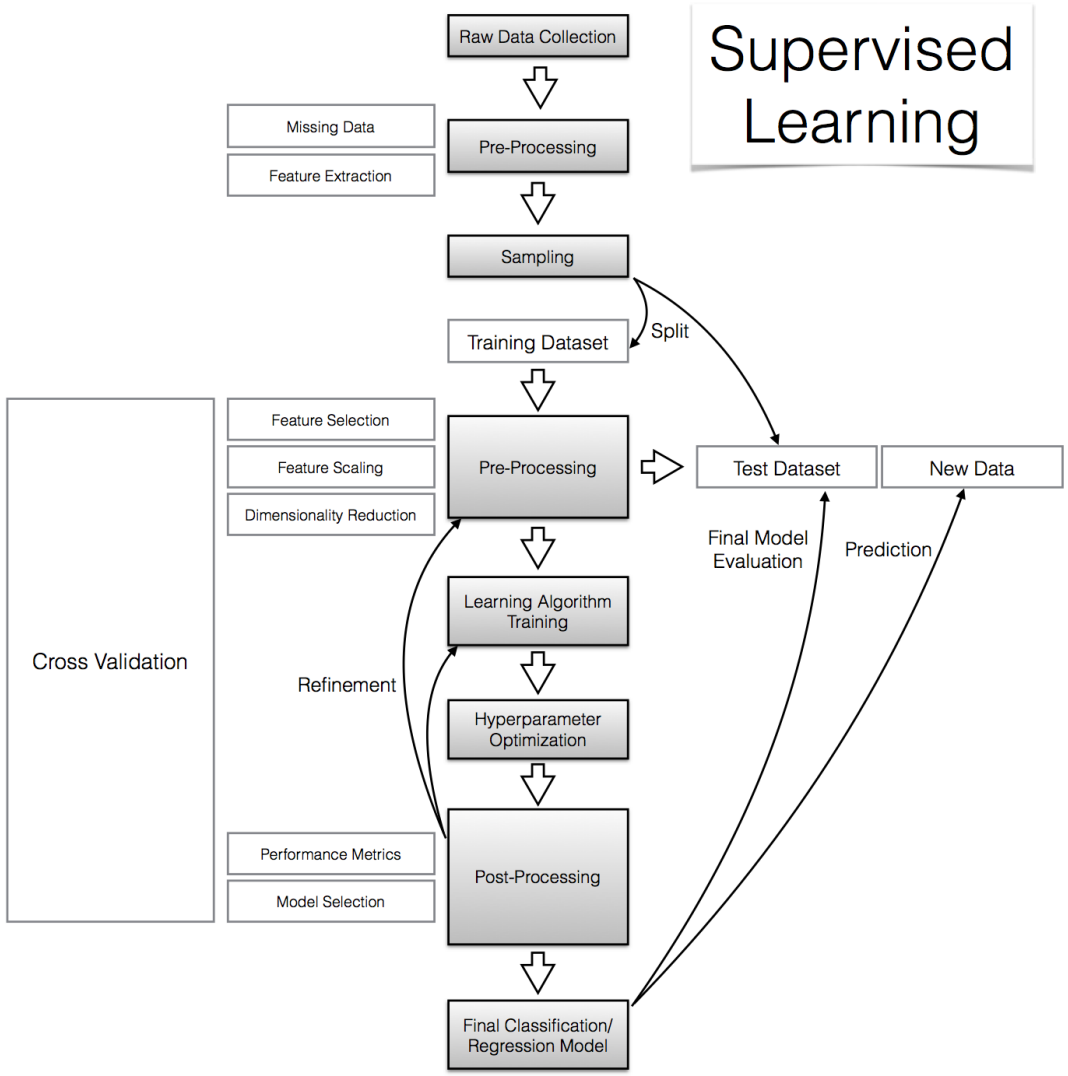
	ETNIA_PARTICIPANTE_1	ETNIA_PARTICIPANTE_2	ETNIA_PARTICIPANTE_3	ETNIA_PARTICIPANTE_4
0	0	0	1	0
1	1	0	0	0
2	1	0	0	0
3	0	0	0	0
4	1	0	0	0

Missings

- Deixar-los com estan
- Eliminar de l'anàlisi participants amb missings en la variable d'interès
- Imputar variables
- Mitjana, mediana de la variable
- Inferir variable segons participants més 'semblants'

Problema exemple

Predictor hipercolesterolèmia



Selecció del model

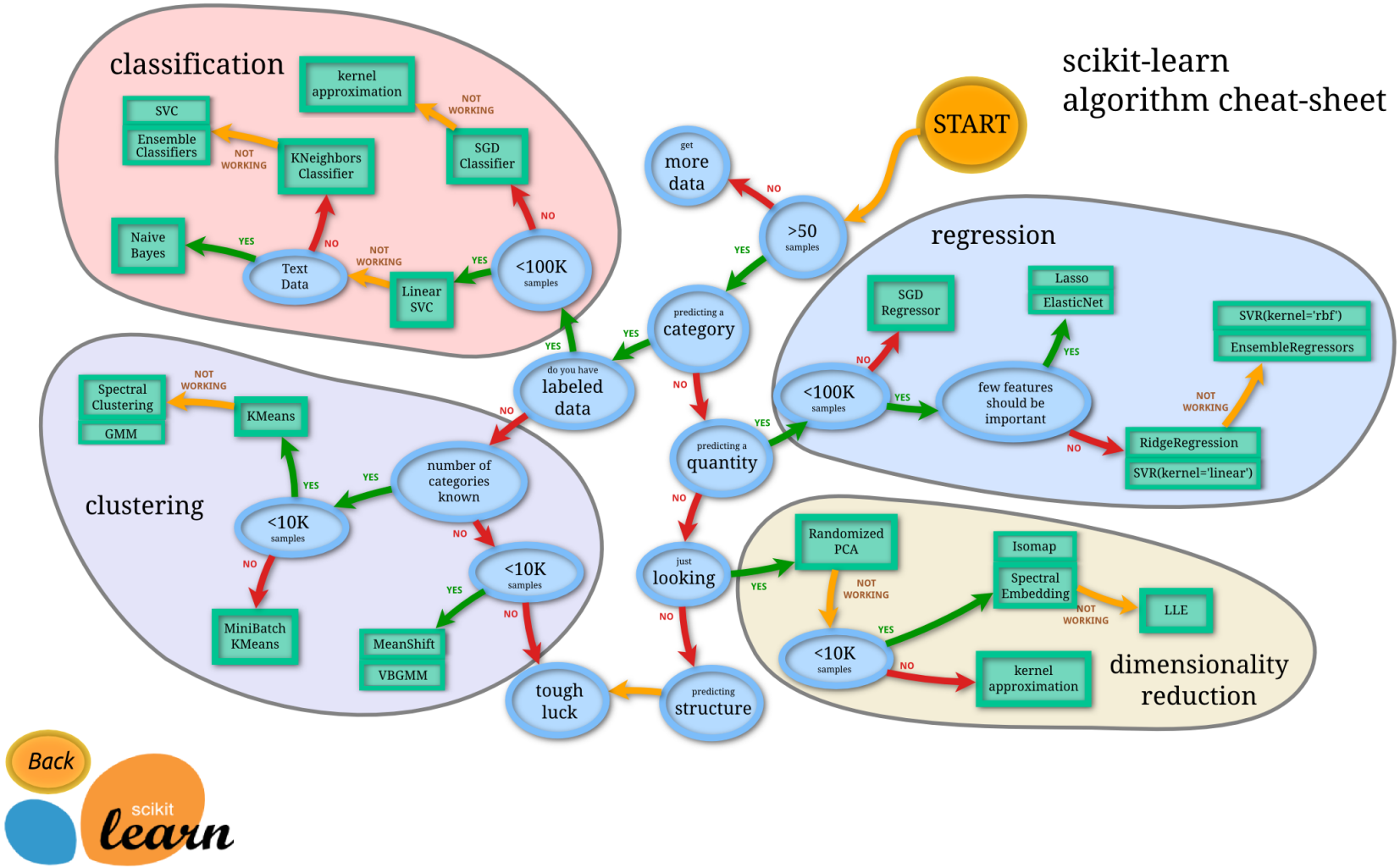


Figure 3:

Arbre de decisió resultat

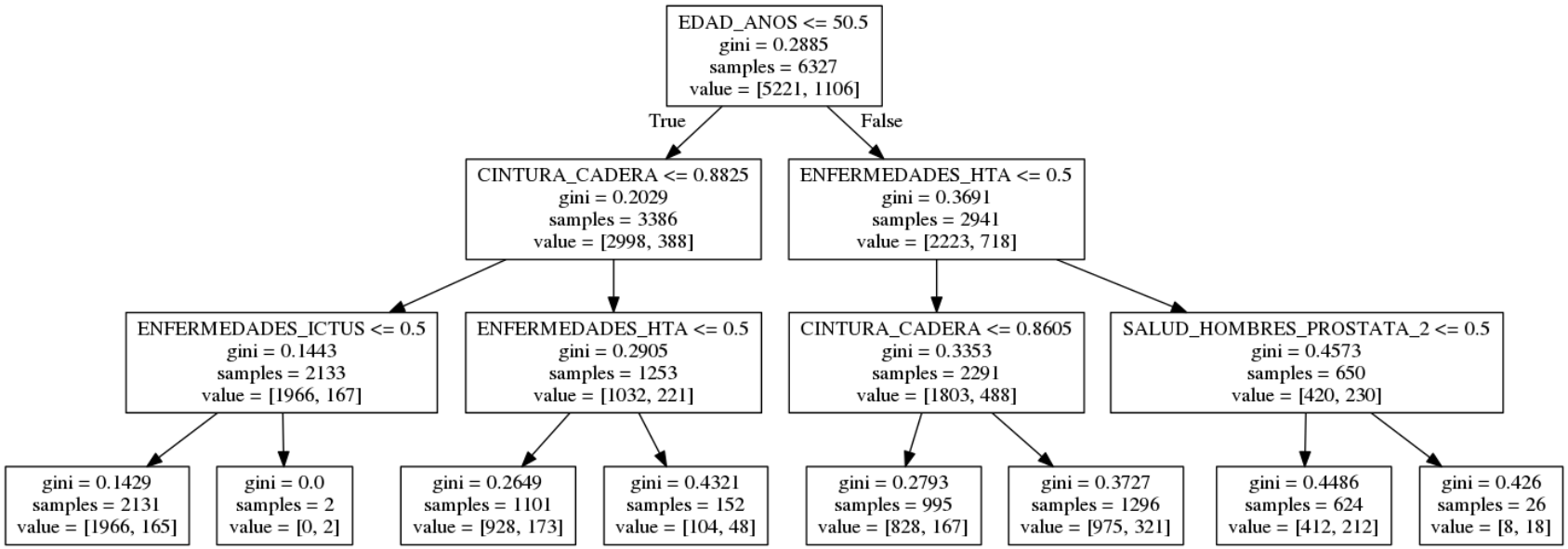


Figure 4:

Avaluació del model