Detección temprana de cardiopatías con Machine Learning

Índice

Punto de partida

Según la OMS, a nivel mundial, **cada año mueren más personas por enfermedades cardiovasculares que por cualquier otra causa**, principalmente enfermedad cardíaca isquémica y accidente cerebrovascular.

>3/4

de estas muertes se producen **en países de bajos y medianos ingresos**, donde los casos siguen aumentando.

Ratios globales de mortalidad por cardiopatías

Journal of Healthcare Engineering

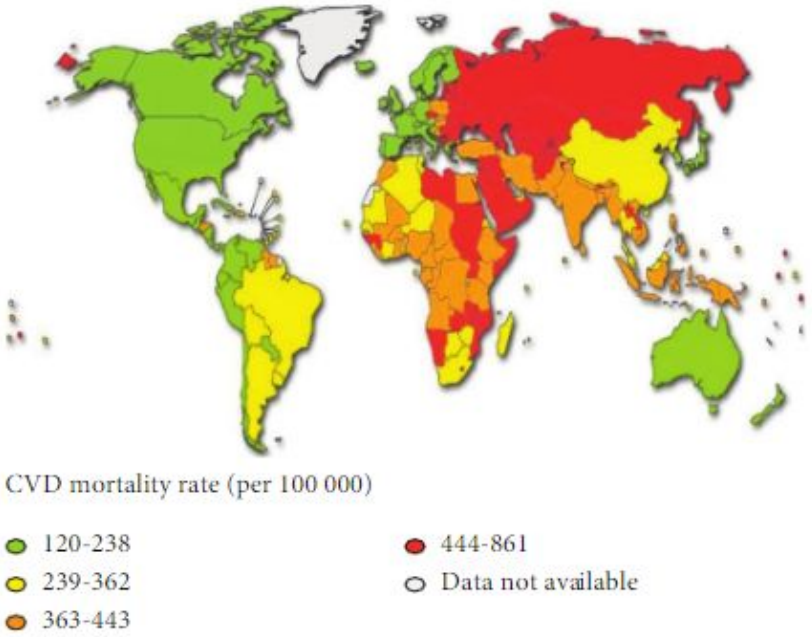


FIGURE 1: World map showing the global distribution of heart disease mortality rates [11].

Objetivo del proyecto

“Es fundamental un diagnóstico precoz de las cardiopatías si queremos disminuir el riesgo.”

¿Cómo?

- **Creando un algoritmo que permita la detección temprana de cardiopatías** en países de bajos y medianos ingresos sin acceso a atención médica de calidad.

¿Por qué así?

- **Es más eficiente y escalable un algoritmo que formar y desplazar médicos**, y en algunos casos es incluso más eficaz.


El Algoritmo

Exploratory Data Analysis

Dataset disponible en Kaggle:

12 VARIABLES

918 INSTANCIAS

2 CLASES -> **Cardiopatía / No Cardiopatía**

Variables

- **Age:** age of the patient [years]
- **Sex:** sex of the patient [M: Male, F: Female]
- **ChestPainType:** chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- **RestingBP:** resting blood pressure [mm Hg] -> presión arterial en reposo
- **Cholesterol:** serum cholesterol [mm/dl]
- **FastingBS:** fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- **RestingECG:** resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- **MaxHR:** maximum heart rate achieved [Numeric value between 60 and 202]
- **ExerciseAngina:** exercise-induced angina [Y: Yes, N: No]
- **Oldpeak:** oldpeak = ST [Numeric value measured in depression]

    oldpeak = Depresión del ST inducida por el ejercicio en relación con el reposo (la depresión del ST en el ECG al ingreso indica lesiones coronarias graves)

    El segmento ST abarca la región entre el final de la despolarización ventricular y el comienzo de la repolarización ventricular.
- **ST_Slope:** the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping] -> la pendiente del segmento ST de ejercicio máximo
- **TARGET -> HeartDisease:** output class [1: heart disease, 0: Normal]

Primer análisis

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

Los datos con valor 0 en la tensión arterial y el colesterol los descartamos por inválidos.

Los datos con valor 0 en el azúcar en sangre sí son válidos porque representan niveles < 120 mg.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
449	55	M	NAP	0	0	0	Normal	155	N	1.5	Flat	1

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
293	65	M	ASY	115	0	0	Normal	93	Y	0.0	Flat	1
294	32	M	TA	95	0	1	Normal	127	N	0.7	Up	1
295	61	M	ASY	105	0	1	Normal	110	Y	1.5	Up	1
296	50	M	ASY	145	0	1	Normal	139	Y	0.7	Flat	1
297	57	M	ASY	110	0	1	ST	131	Y	1.4	Up	1
...	...	...	...	...	...	...	...	...	...	...	...	...
514	43	M	ASY	122	0	0	Normal	120	N	0.5	Up	1
515	63	M	NAP	130	0	1	ST	160	N	3.0	Flat	0
518	48	M	NAP	102	0	1	ST	110	Y	1.0	Down	1
535	56	M	ASY	130	0	0	LVH	122	Y	1.0	Flat	1
536	62	M	NAP	133	0	1	ST	119	Y	1.2	Flat	1

Para evitar inputs inválidos en el dataset, he eliminado las entradas con valor 0 en el Colesterol.

Target

2 CLASES -> Cardiopatía / No Cardiopatía

```
csv.HeartDisease.value_counts()
✓ 0.5s
1    508
0    410
Name: HeartDisease, dtype: int64
```

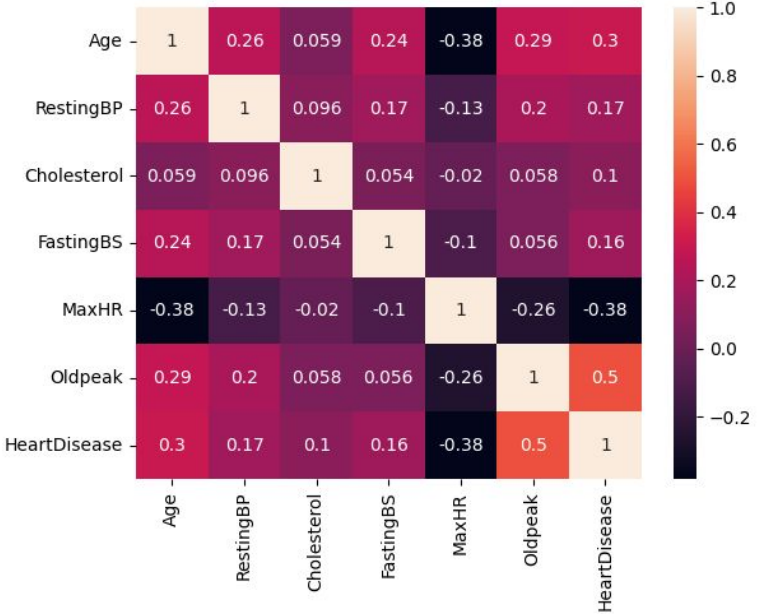
Gráficas de correlación

```
Logistic Regression Test
 [[66  5]
 [11 68]]
Gaussian Test
 [[64  7]
 [10 69]]
```

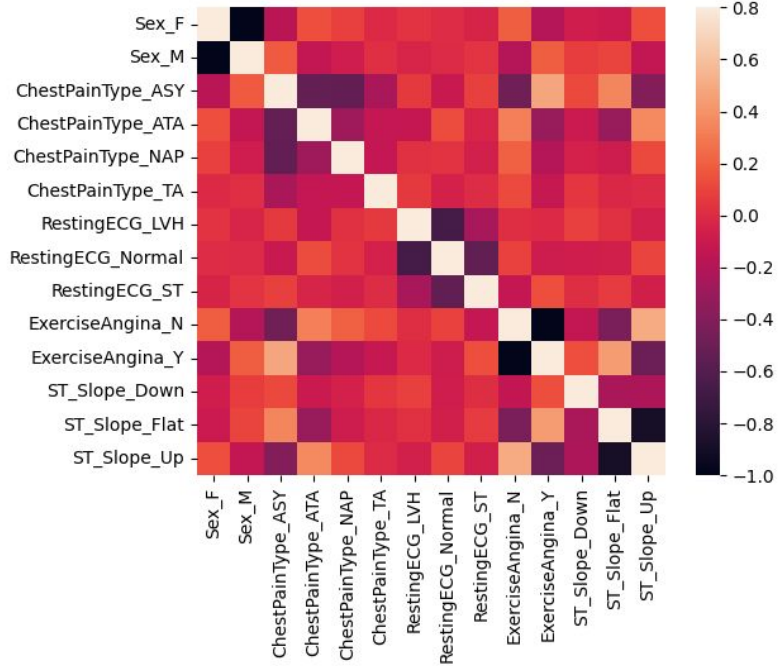
Matrices de correlación

21 variables

Numéricas



Categóricas



Datos finales

Tras limpiar el dataset y organizar las variables:

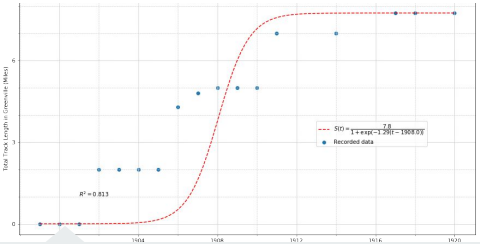
18 VARIABLES

746 INSTANCIAS

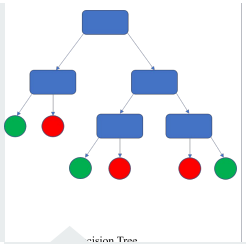
2 CLASES -> **Cardiopatía / No Cardiopatía**

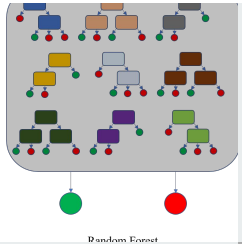
Modelado

- Pipeline



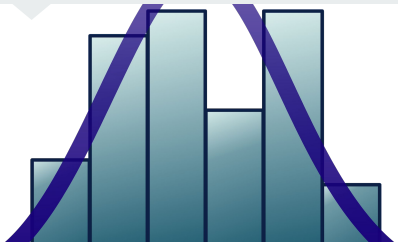
Naive-Bayes Gaussian



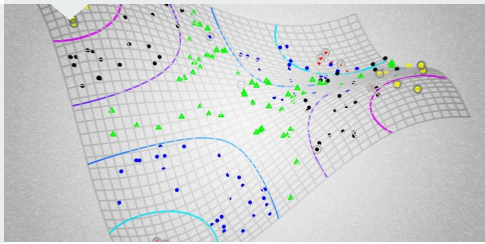


Support Vector Machine

Logistic Regression



Random Forest



Métrica

RECALL

Objetivo: Minimizar los falsos negativos.

¿Por qué? Es preferible asegurar el diagnóstico de todas las posibles cardiopatías.

- Tiene un impacto menos negativo asignar falsos positivos que falsos negativos.

The Best!

- SVM

```
Best estimator: Pipeline(steps=[('scaler', StandardScaler()), ('selectkbest', SelectKBest(k=1)),
                ('svc', SVC(C=0.1, kernel='linear'))])
Best params: {'selectkbest__k': 1, 'svc__C': 0.1, 'svc__kernel': 'linear'}
Best score: 0.87
```

```
SVM Test
 [[59 12]
 [ 9 70]]
```

Gracias.