Estudio de Caso: Enfermedad Cardíaca - Enzo Cozza

Introducción

Existen distintos tipos de enfermedades cardíacas, pero la más común es el estrechamiento o bloqueo de las arterias coronarias (enfermedad de las arterias coronarias). Esta enfermedad es la principal causa de infartos. Otros tipos de enfermedades cardíacas pueden afectar los músculos o las válvulas del corazón, o el mismo puede no latir bien a causa de una insuficiencia cardíaca. También hay personas que nacen con una enfermedad cardíaca (congénita).

Algunas de las principales causas que provocan estas enfermedades son: fumar, algunos tratamientos para el cáncer (quimioterapias por ejemplo), dietas altas en grasas o sal, presión alta, colesterol alto, obesidad, falta de actividad física, estrés y mala higiene. La edad y el sexo afectan también (pero estas no son controlables).

Muchas de estas enfermedades pueden ser prevenidas o tratadas llevando un estilo de vida saludable (dieta balanceada, buena higiene, manejo del estrés, hacer ejercicio).

En Estados Unidos, las muertes por enfermedades cardiovasculares representan aproximadamente 1 de cada 3 muertes (esto equivale a más de 801.000 muertes). Estas enfermedades provocan más muertes que el cáncer y enfermedades crónicas de vías respiratorias combinadas.

En el año 2013, las muertes por motivos cardiovasculares representaron el 31% de todas las muertes a nivel mundial.

Información del dataset

Atributos:

El dataset cuenta con 76 atributos en total. En primer lugar, las columnas que la mayoría de sus datos sean valores faltantes, o que contengan en su descripción del dataset “not used” o “dummy” serán eliminadas. Luego, se eliminarán variables correlacionadas, y por último se realizará un PCA para obtener los atributos más valiosos entre los restantes.

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	0.753	0.312	0.312
PC 2	0.465	0.119	0.431
PC 3	0.426	0.100	0.531
PC 4	0.380	0.079	0.610
PC 5	0.336	0.062	0.672
PC 6	0.312	0.054	0.726
PC 7	0.274	0.041	0.767
PC 8	0.265	0.039	0.806
PC 9	0.244	0.033	0.838
PC 10	0.216	0.026	0.864
PC 11	0.211	0.024	0.889
PC 12	0.184	0.019	0.907
PC 13	0.164	0.015	0.922
PC 14	0.152	0.013	0.934
PC 15	0.139	0.011	0.945
PC 16	0.127	0.009	0.954

Imagen 1 - Principales Componentes

Los primeros 16 PCs son los que logran una varianza acumulada mayor a 0.95. Tomando para cada uno de ellos los tres atributos que más aportan, se obtienen los siguientes (18) atributos: 3, 4, 9, 10, 11, 12, 16, 21, 23, 24, 25, 28, 30, 33, 34, 40, 42, 43, 56.

Age: edad en años.
Sex: sexo de la persona (1 hombre, 0 mujer).
CP: tipo del dolor de pecho (1 angina típica, 2 angina atípica, 3 dolor no anginal, 4 asintomático)
Trestbps: presión sanguínea en reposo.
Chol: colesterol en mg/dl.
Fbs: nivel de azúcar en ayunas (1 si es mayor a 120mg/dl, 0 si no lo es).
Ekgday: ejercicios ECG por día.
Dig: digitalis usados durante ejercicios ECG (1 sí, 0 no).
Prop: beta bloqueadores usados durante ejercicios ECG (1 sí, 0 no).
Nitr: nitratos usados durante ejercicios ECG (1 sí, 0 no).
Proto: protocolo de ejercicio
Thaltime: tiempo en el que se registró una depresión ST.
Thalrest: frecuencia cardíaca en reposo.
Tpeakbps: pico de presión sanguínea en ejercicio (primera parte)
Oldpeak: depresión ST inducida por ejercicio tras reposo.
Rldv5: altura en reposo.
Rldv5e: altura en ejercicio pico.
Cday: cateterismo cardíaco por día.

Entre los atributos seleccionados se pueden destacar:
Age: como fue mencionado previamente, la edad es un factor muy importante. Aproximadamente, el 80% de las personas que sufren de enfermedad coronaria son personas de 65 años (o más).
Sex: los hombres tienen un riesgo mayor de contraer alguna enfermedad del corazón. Luego de la menopausia, el riesgo para una mujer aumenta casi al nivel de un hombre.
Dolor de pecho: una angina es producida cuando los músculos del corazón no reciben suficiente sangre oxigenada.

Trestbps: una presión sanguínea alta puede dañar las arterias.
Chol: un colesterol alto muy probablemente estrechen las arterias.
Fbs: no producir suficiente insulina causa que tu nivel de azúcar en sangre suba, aumentando el riesgo de tener una enfermedad del corazón.

Distribución por clase:

Dataset	0	1	2	3	4	Total
Cleveland	164	55	36	35	13	303
Hungarian	188	37	26	28	15	294
Switzerland	8	48	32	30	5	123
Long Beach VA	51	56	41	42	10	200

Tabla 1 - Distribución por clases según cada dataset

Debido a varias incongruencias en el orden de los atributos obtenidos a partir de los distintos datasets (observando los distintos rangos uno puede darse cuenta que las bases de datos no son 100% coherentes entre ellas mismas), se decide modelar con los conjuntos de datos ya procesados previamente por UCI, los cuales han sido utilizados de manera eficiente en muchos estudios ya realizados por otras personas.

Los 14 atributos seleccionados para el modelado de información son:

Age: edad en años.
Sex: sexo de la persona (1 hombre, 0 mujer).
CP: tipo del dolor de pecho (1 angina típica, 2 angina atípica, 3 dolor no anginal, 4 asintomático)
Trestbps: presión sanguínea en reposo.
Chol: colesterol en mg/dl.
Fbs: nivel de azúcar en ayunas (1 si es mayor a 120mg/dl, 0 si no lo es).
Restecg: resultados electrocardiográficos en reposo (0 normal, 1 onda ST-T, 2 hipertrofia probable o definitiva en el ventrículo izquierdo).
Thalach: máxima frecuencia cardíaca obtenida.
Exang: angina producida por ejercicio (1 sí, 0 no).
Oldpeak: depresión ST inducida por ejercicio.
Slope: pendiente del pico del segmento ST (1 positiva, 2 cero, 3 negativa).
Ca: número de venas mayores.
Thal: 3 normal, 6 defecto fijo, 7 defecto reversible.
Num: diagnóstico de enfermedad del corazón. Valores del 0 al 4 (0 es absencia).

Aquí se puede observar que varios de los atributos obtenidos a partir de la totalidad de los datos coinciden con los propuestos por UCI.

Rangos y Distribuciones

Se analizarán rangos y distribuciones de los atributos no binominales o no polinominales (para estos otros casos, sus rangos ya fueron definidos previamente).

Age: de 28 a 77 años, con una distribución normal. El promedio se sitúa en los 52.9 años, con una desviación de 9.5.

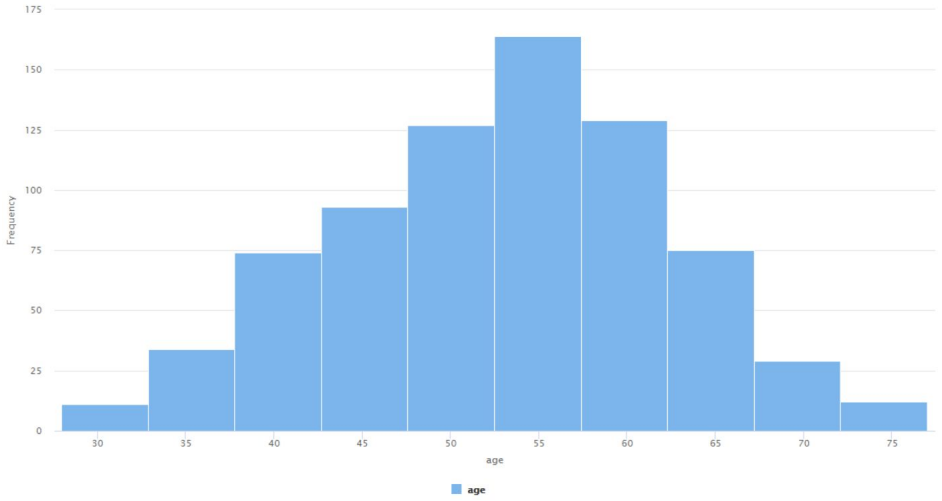

Imagen 2 - Distribución de age

Trestbps: de 92 a 200 mmHg, con una distribución normal. El promedio se sitúa en los 132.7 mmHg, con una desviación de 17.8.

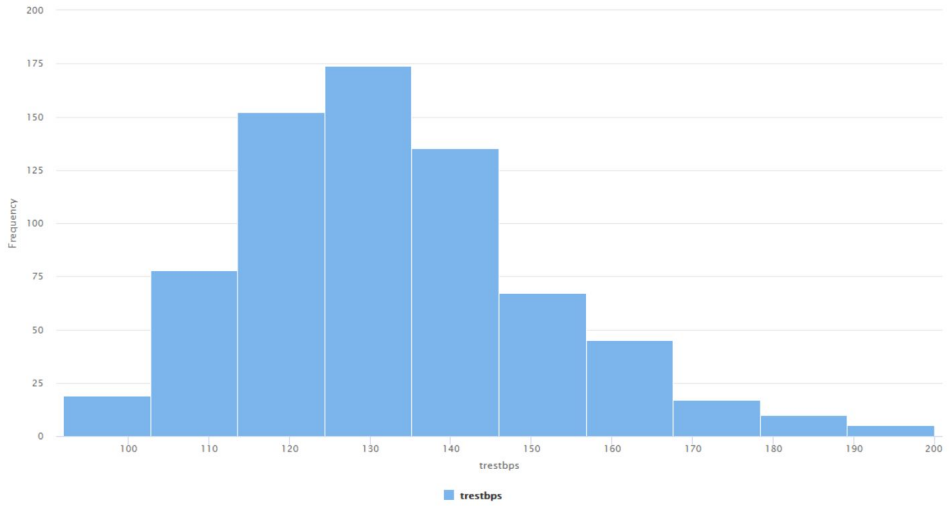

Imagen 3 - Distribución de trestbps

Chol: de 85 a 603 mg/dl, con una distribución normal. El promedio se sitúa en los 246.8 mg/dl, con una desviación de 58.5.

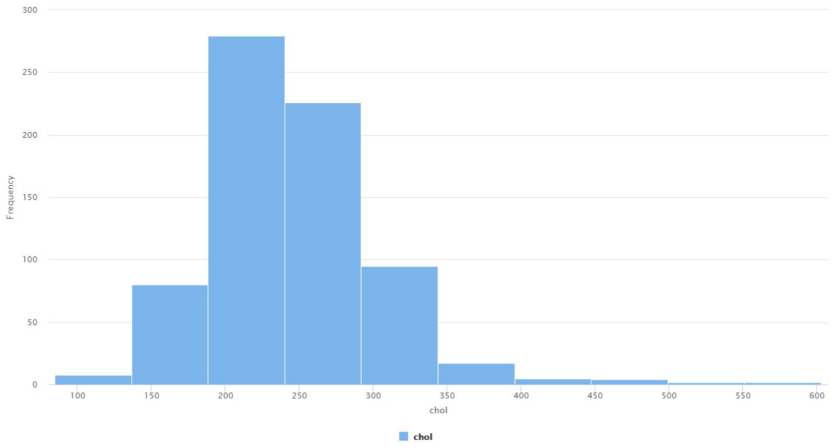

Imagen 4 - Distribución de chol

Thalach: de 69 a 202 latidos/minuto, con una distribución similar a la normal. El promedio se sitúa en los 141.1 latidos/minuto, con una desviación de 24.9.

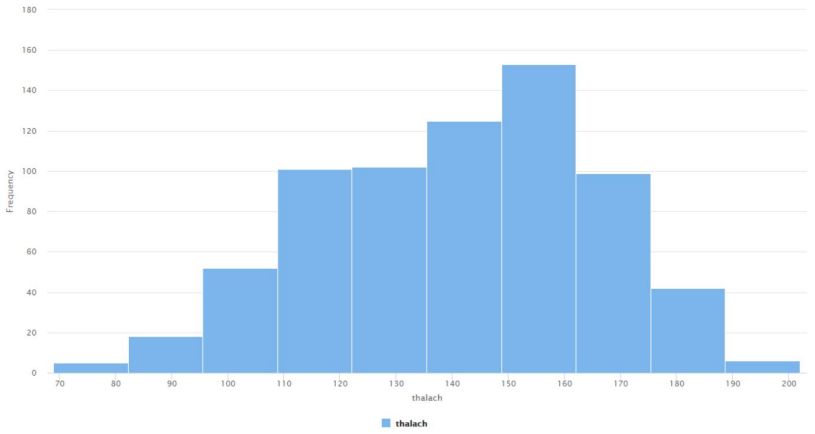

Imagen 5 - Distribución de thalach

Oldpeak: de 0 a 6.2, con una distribución sesgada a la derecha. El promedio se sitúa en 0.89, con una desviación de 1.09.

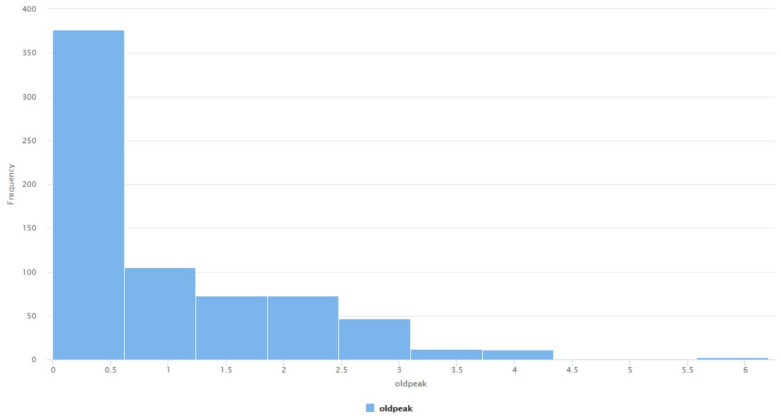

Imagen 6 - DIstribución de oldpeak

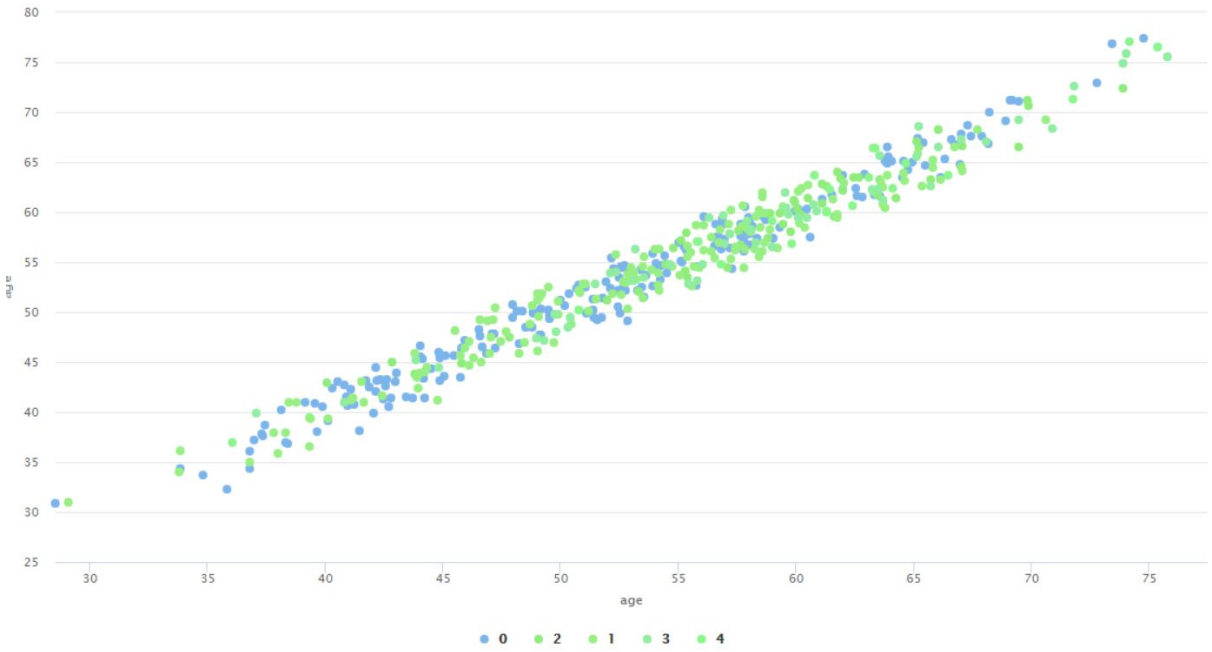

Imagen 7 - Ausencia/Presencia de una enfermedad según edad

En la imagen 7 se puede observar que a partir de cierta edad (53 años aproximadamente), la presencia de enfermedades del corazón se hace cada vez más frecuente que en las personas menores a esa edad (lo que no significa que no se pueda dar).

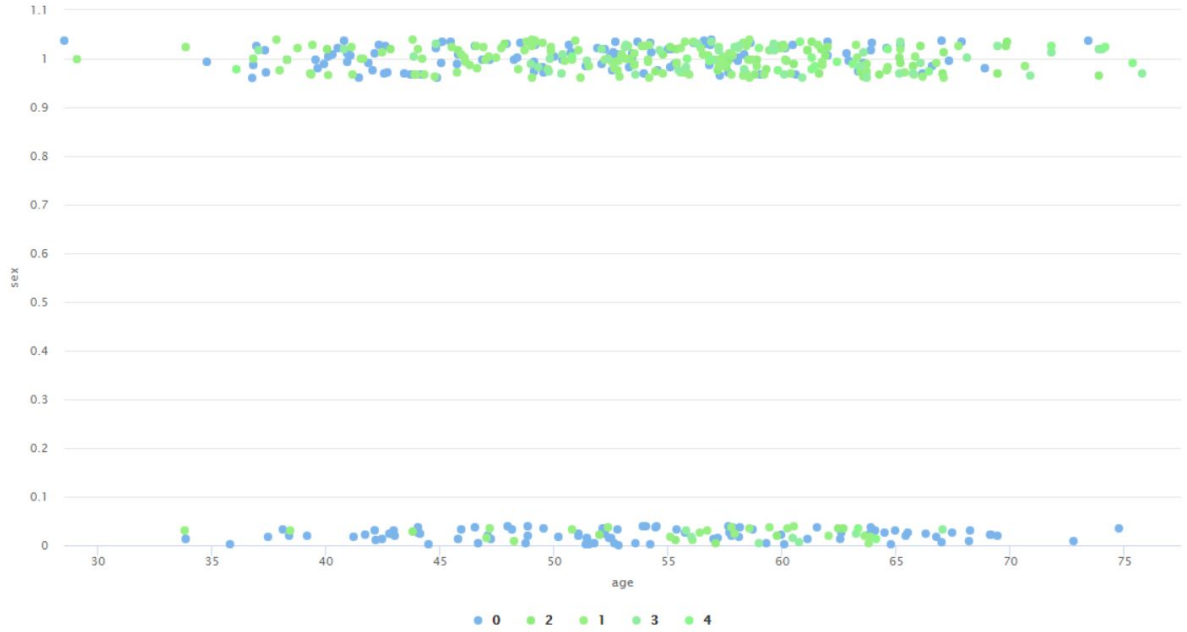

Imagen 8 - Ausencia/Presencia de una enfermedad según sexo y edad

En la imagen 8 se realiza un filtro tanto por edad como por sexo, donde se puede ver que las enfermedades del corazón se dan mucho más seguido en hombres (sexo ‘1’) que en mujeres.

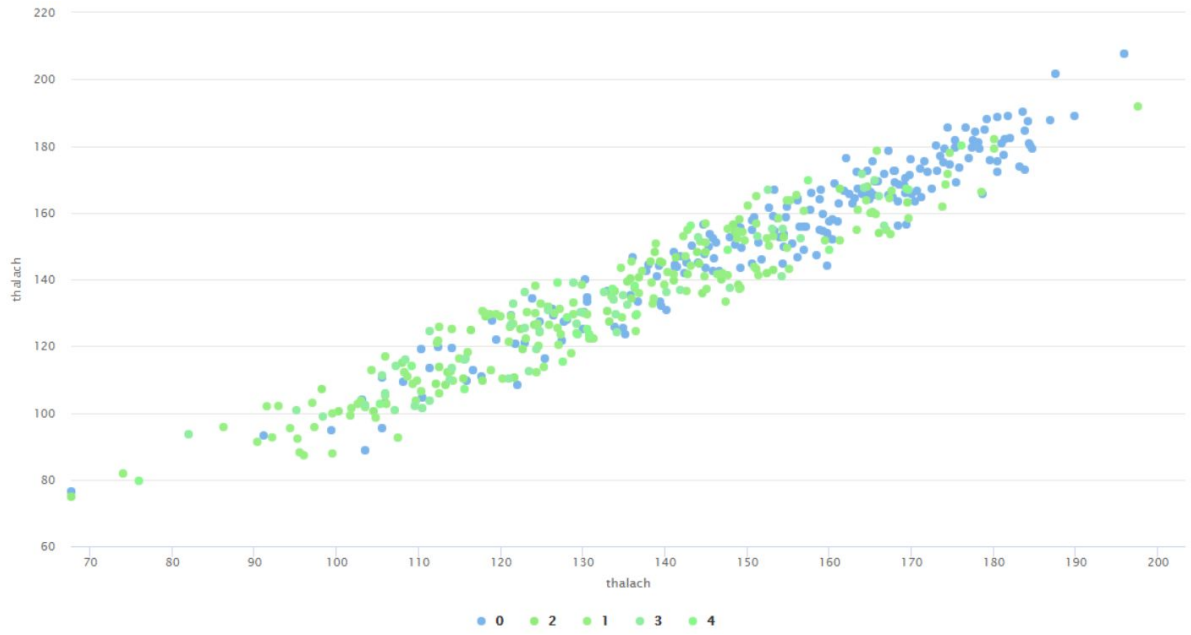

Imagen 9 - Ausencia/Presencia de una enfermedad según el ritmo cardíaco máximo registrado

Por último, en la imagen 9 se analiza cuánto afecta el ritmo cardíaco máximo registrado para las personas del dataset, y se puede observar que cuanto más alto sea este, más chances de que la persona no tenga una enfermedad del corazón se presentan. Esto podría darse ya que cuanto menor sea el pico de ritmo cardíaco, probablemente haya alguna insuficiencia que marque una posible enfermedad.

Tratamiento previo de los datos:

En el dataset se encuentran casos incoherentes que podrían afectar el correcto modelado de los datos. Por ejemplo: hay casi 200 casos de personas con colesterol 0, lo cual es imposible que sea un caso real, por lo que esas personas fueron excluidas del dataset.

Las columnas ca y thal cuentan con más del 66% de datos faltantes, por lo que se decide removerlas. Para slope, que cuenta con 200 datos faltantes, se decide eliminar las filas con estos datos faltantes ya que si se incluyera un valor (sea moda, máximo o mínimo), este terminaría afectando en gran manera la distribución original de la columna. Para el resto de las columnas, donde los datos faltantes son relativamente pocos, se decide cambiarlos por la media o por la moda (dependiendo si son numéricos o polinominales).

Modelado:

A continuación, para el modelado de los datos lo primero que se hizo fue una división de la información de 70-30%, para ser utilizados como datasets de entrenamiento y evaluación respectivamente.

Se realizarán modelos para: SVM, Naive Bayes, Regresión Logística, Árboles de decisión, Random Forest y k-NN. Además, se utilizará un modelo de ensamble y uno de AdaBoost.

Se evaluarán los distintos modelos según: precisión ((TN+TP)/(total de ejemplos)) y sensibilidad (TP/(TP+FN)) obtenidos. Se busca la sensibilidad, ya que se quiere saber que

tan bien funciona el modelo para encontrar las personas que de verdad tienen la enfermedad.

Para realizar este modelado, primero se transformó la clase de salida de polinominal a binominal. Las clases 1, 2, 3 y 4 todas representan la presencia de alguna enfermedad del corazón, mientras que 0 implica ausencia, por lo tanto se transformaron las clases 2, 3 y 4 a clase 1, bajo el concepto “la persona tiene una enfermedad” o “la persona no tiene una enfermedad”.

SVM

Matriz de confusión para SVM:

	true 0	**true 1**	**class precision**
predicted 0	54	20	72.97%
predicted 1	7	61	89.71%
class recall	88.52%	75.31%	

Precisión: 80.99%

Sensibilidad: 75.31%

Naive Bayes

Matriz de confusión para Naive Bayes:

	true 0	**true 1**	**class precision**
predicted 0	49	12	80.33%
predicted 1	12	69	85.19%
class recall	80.33%	85.19%	

Precisión: 83.10%

Sensibilidad: 85.19%

Regresión Logística

Matriz de confusión para Regresión Logística:

	true 0	**true 1**	**class precision**
predicted 0	43	11	79.63%

predicted 1	18	70	79.55%
class recall	70.49%	86.42%	

Precisión: 79.58%

Sensibilidad: 86.42%

Árbol de decisión

Matriz de confusión para Árbol de decisión:

	true 0	**true 1**	**class precision**
predicted 0	45	21	68.18%
predicted 1	16	60	78.95%
class recall	73.77%	74.07%	

Precisión: 73.94%

Sensibilidad: 74.07%

k-NN

Matriz de confusión para k-NN:

	true 0	**true 1**	**class precision**
predicted 0	42	11	79.25%
predicted 1	19	70	78.65%
class recall	78.85%	86.42%	

Precisión: 78.87%

Sensibilidad: 86.42%

Random Forest

Matriz de confusión para Random Forest:

	true 0	**true 1**	**class precision**
predicted 0	44	11	80.00%

predicted 1	17	70	80.46%
class recall	72.13%	86.42%	

Precisión: 80.28%

Sensibilidad: 86.42%

Ensamble

Para el ensamble se utilizaron k-NN, árbol de decisión y Naive Bayes.

Matriz de confusión para Ensamble:

	true 0	**true 1**	**class precision**
predicted 0	49	12	80.33%
predicted 1	12	69	85.19%
class recall	80.33%	85.19%	

Precisión: 83.10%

Sensibilidad: 85.19%

AdaBoost

Para el modelado se utilizó un árbol de decisión.

Matriz de confusión para AdaBoost:

	true 0	**true 1**	**class precision**
predicted 0	44	12	78.57%
predicted 1	17	69	80.23%
class recall	72.13%	85.19%	

Precisión: 79.58%

Sensibilidad: 85.19%

Conclusión

Los algoritmos que obtuvieron las mejores performances fueron: Naive Bayes y el Ensamble. Estos dos algoritmos son los que obtuvieron la precisión más elevada entre todos los algoritmos propuestos.

Además de la precisión, se consideró clave para el tema del que se está tratando la sensibilidad (es decir, que la cantidad de casos que efectivamente presentan una enfermedad sean marcados como con la enfermedad, y no que sean marcados como que no la tienen). En este aspecto, hay varios algoritmos que ofrecen un % superior al 85% de eficacia. De los evaluados, todos los algoritmos excepto SVM y árbol de decisión son los que obtuvieron ese %.

Bibliografía

- https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118
- https://medlineplus.gov/spanish/heartdiseases.html
- https://professional.heart.org/idc/groups/ahamah-public/@wcm/@sop/@smd/documents/downloadable/ucm_491392.pdf