Clara Aibar Álvarez

PRÁCTICA 2 - MINERÍA DE DATOS BIOMÉDICOS

Nota: para compilar los cinco scripts, uno para cada algoritmo escogido, hay que ejecutar el script "Descripción y preparación de los datos" porque en él se generan las variables de entrenamiento y testeo de los datos ya normalizados y tipificados.

1. DESCRIPCIÓN DE LA BASE DE DATOS

La base de datos escogida para esta práctica está orientada a la predicción de enfermedades cardiacas (https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction), las cuales son la primera causa de muerte en el mundo. En ella se recogen los datos de trece variables de interés médico para la predicción de dichas patologías, entre las que se encuentran, aparte de los datos de sexo y edad, variables fisiológicas importantes para determinar la salud del sistema cardiovascular del paciente. La mayoría estos datos médicos son relativamente fáciles de obtener, ya que se suelen medir en pruebas rutinarias. Aun así, resulta muy complicado para los médicos realizar su análisis, ya que se trata de un conjunto de datos muy amplio. Es por ello por lo que surge la necesidad de desarrollar algoritmos que faciliten esta evaluación de los datos y aseguren un diagnóstico más preciso y rápido.

La base de datos escogida cuenta con 270 registros de pacientes, de los cuales 120 presentan algún tipo de enfermedad cardiaca y los 150 restantes se encuentran sanos. La variable que predecir de las 13 que se incluyen es "Heart.Disease", la cual es binaria y toma el valor "Presence" cuando el paciente tiene alguna patología cardiaca y "Absence" cuando no. Por lo tanto, las 12 variables restantes son las descriptivas de la variable de enfermedad a predecir.

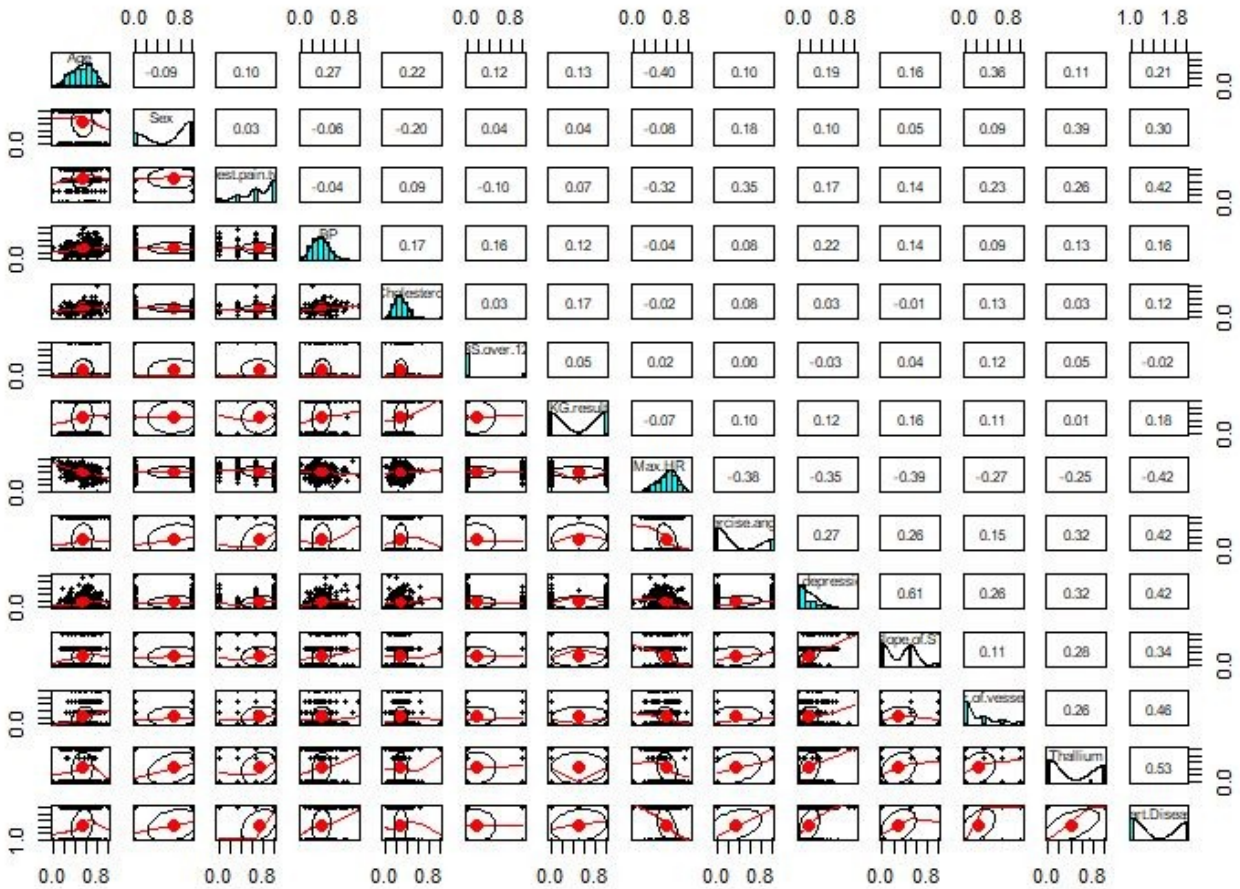
La siguiente tabla muestra los estadísticos descriptivos de las variables de interés:

	Media	Desviación típica	Media truncada	Mediana	Desviación media absoluta	Min	Max	Rango	Kurtosis	Asimetría (skew)
Age	54.43	9.11	54.56	55	10.38	29	77	48	-0.57	-0.16
Chest.pain. type	3.17	0.95	3.31	3	1.48	1	4	3	-0.33	-0.87
BP	131.34	17.86	130.12	130	14.83	94	200	106	0.86	0.71
Cholesterol	249.66	51.69	247.07	245	48.18	126	564	438	4.73	1.17
EKG.results	1.02	1	1.03	2	0	0	2	2	-2	-0.04
Max.HR	149.68	23.17	151	153.5	22.98	71	202	131	-0.14	-0.52
ST.depression	1.05	1.15	0.87	0.8	1.19	0	6.2	6.2	1.67	1.25
Slope.of.ST	1.59	0.61	1.52	2	1.48	1	3	2	-0.64	0.54
Number.of. vessels.fluro	0.67	0.94	0.5	0	0	0	3	3	0.25	1.2
Thallium	4.7	1.94	4.62	3	0	3	7	4	-1.9	0.28

En esta tabla no se incluyen las variables de “Sex”, “FBS.over.120” y “Exercise.angina”, ya que son variables booleanas. A continuación, se muestra su tabla de frecuencia:

	0	**1**
Variable “Sex”	0.32222	0.6777
Variable “FBS.over.120”	0.851851	0.1481
Variable “Exercise.angina”,	0.67	0.32

Tras normalizar y tipificar los datos, se puede observar la correlación de cada variable descriptiva respecto a la que se quiere predecir empleando la función pairs.panels() de la librería “psych”. A simple vista y antes de aplicar ningún algoritmo, se puede deducir que la variable que más peso influye en la presencia de una enfermedad cardiaca es el nivel de talio, seguido de la curva ST del electrocardiograma del paciente y del tipo de dolor de pecho.



2. ALGORITMOS IMPLEMENTADOS

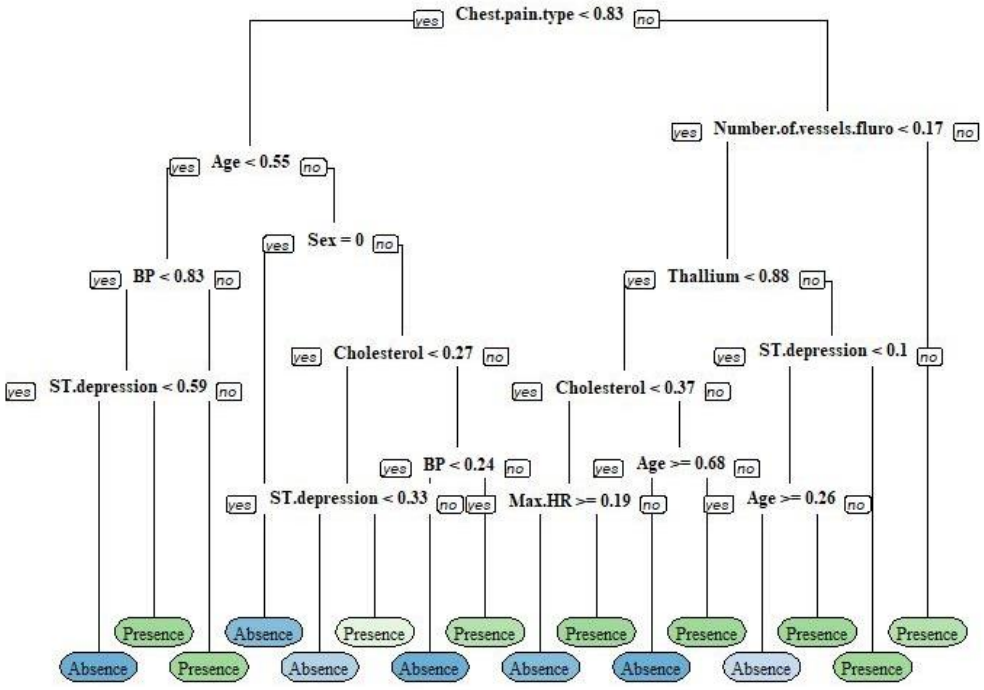
2.1 Supervisados

Árbol de decisión

El algoritmo del árbol de decisión resulta el más adecuado para aplicar con datos clínicos debido a que permite obtener la explicabilidad de los resultados que produce.

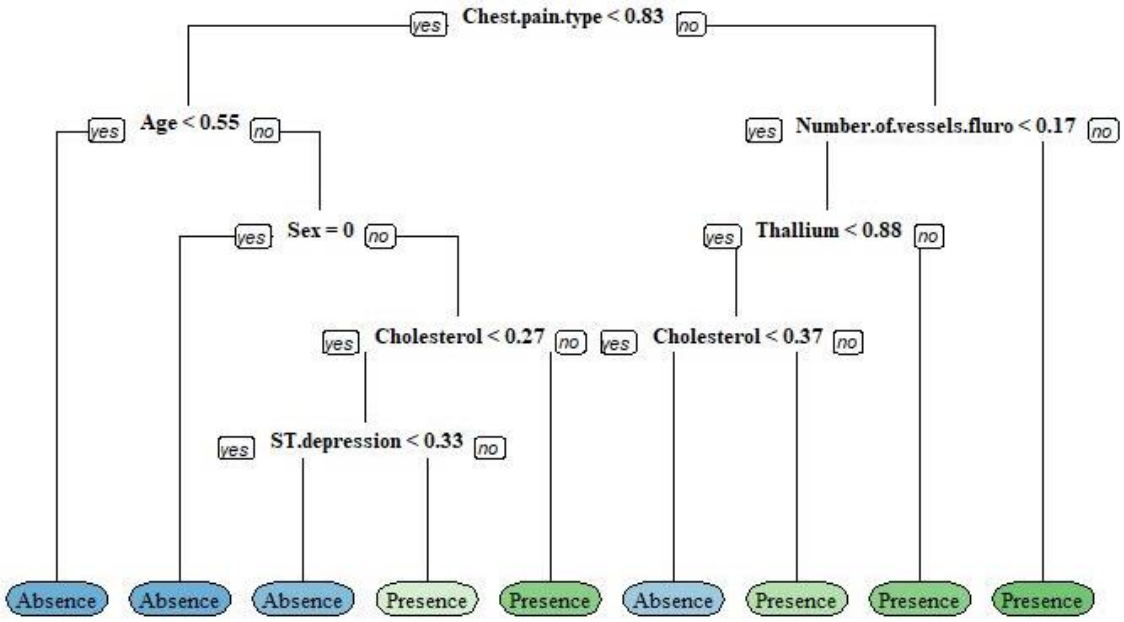
Para su desarrollo se emplearon dos librerías distintas: la del algoritmo C5.0 y la rpart. La primera permite generar árboles con distintos boosts para obtener modelos más complejos, pero no presenta resultados tan altos como la segunda (el máximo obtenido es el un 86,4% de precisión).

El siguiente modelo fue realizado con la librería rpart y fue el que más precisión mostró, concretamente de un 89%.



	Precisión	Sensibilidad	Especificidad	Medida F	Coef. Kappa
Resultado	89%	84%	92%	0.86	0.76

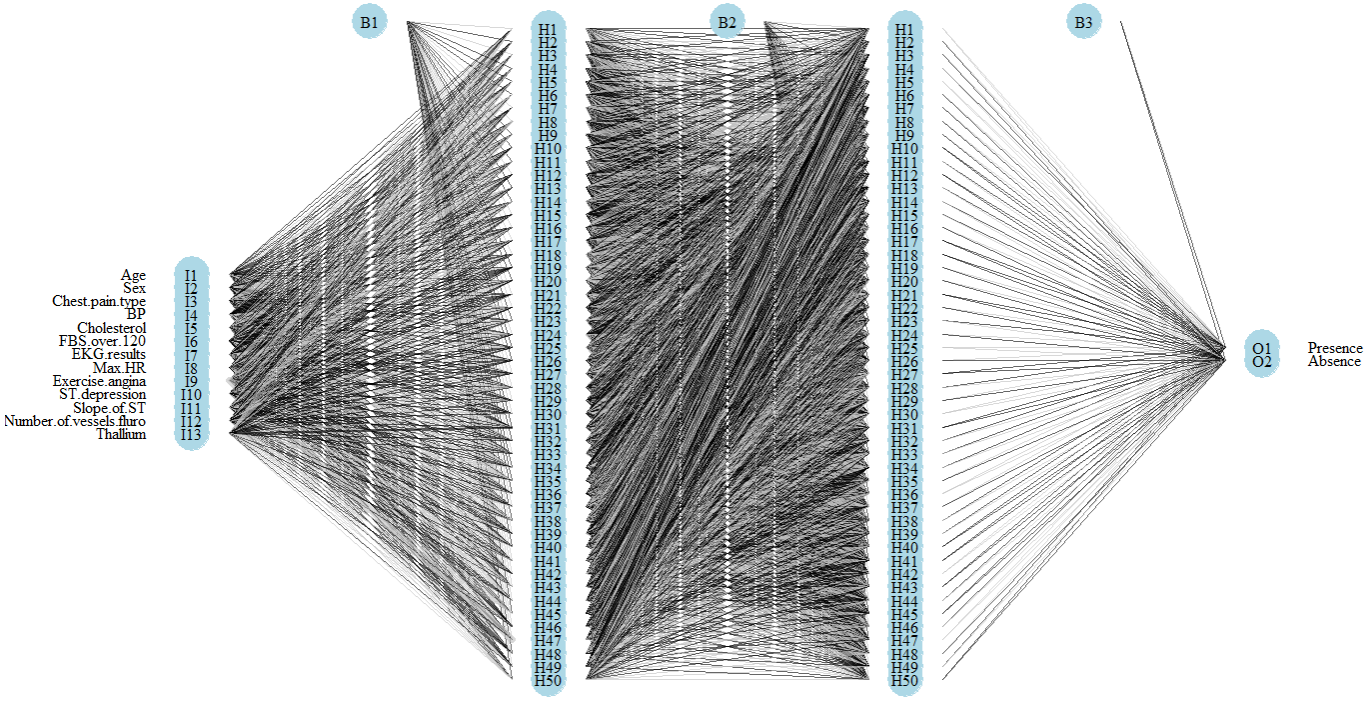
Además de este primer modelo de árbol de decisión también se generó uno enfocado a una mayor optimización del gasto computacional. Aun siendo un modelo mucho menos complejo que el anterior, sigue teniendo una precisión relativamente alta, por lo que podría ser una opción bastante práctica de usar.



Redes neuronales

La red neuronal que más precisión mostró fue la compuesta por dos capas de 50 neuronas cada una. Para generar el modelo se hizo uso de la librería neuralnet.

	Precisión	Sensibilidad	Especificidad	Medida F	Coef. Kappa
Resultado	82.72%	82%	83,87%	0.8235	0.6429



Máquina de soporte vectorial

Para generar un modelo basado en el algoritmo de soporte vectorial se empleó la librería kernlab. Esta ofrece distintos tipos de kernels, por lo que se creó un modelo con cada uno de ellos, obteniendo el mejor resultado con el kernel "laplacedot".

	Precisión	Sensibilidad	Especificidad	Medida F	Coef. Kappa
Resultado	90%	87%	92%	0.88	0.88

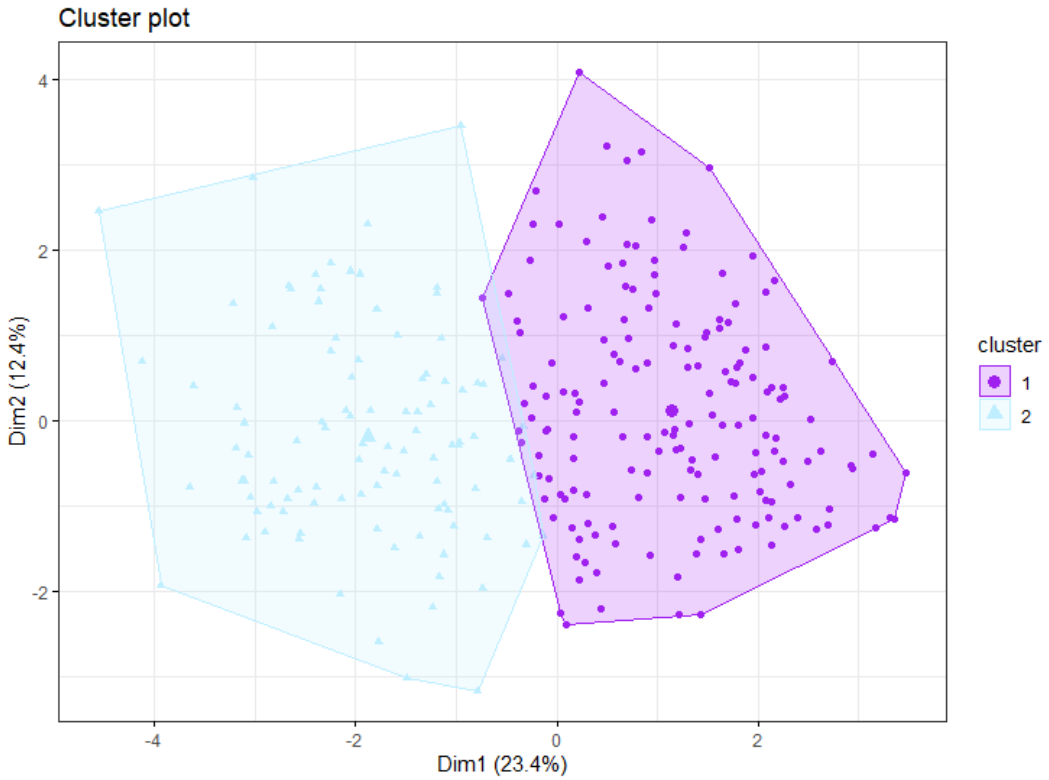
Naive Bayes

Al algoritmo Naive Bayes creado no se le aplicó la técnica de smoothing de Laplace, ya que no afectaba a los resultados de rendimiento del modelo, por lo que su rendimiento no pudo ser mejorado añadiendo una mayor complejidad.

	Precisión	Sensibilidad	Especificidad	Medida F	Coef. Kappa
Resultado	87,65%	86,21%	88,46%	0.8692	0.7355

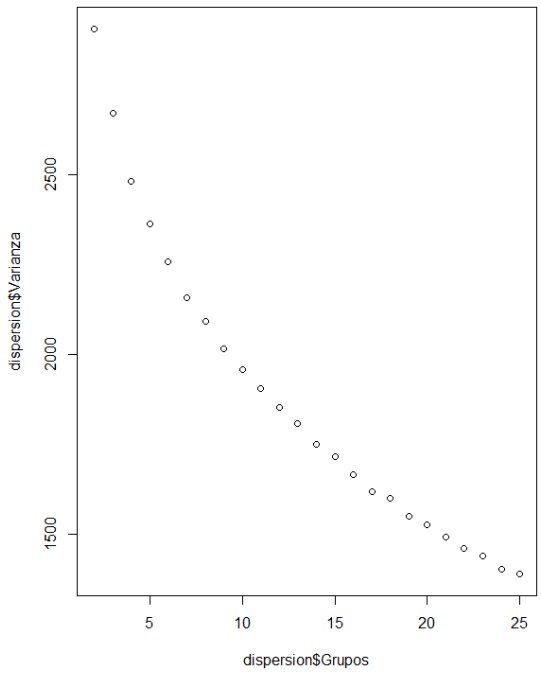
2.2 No supervisado

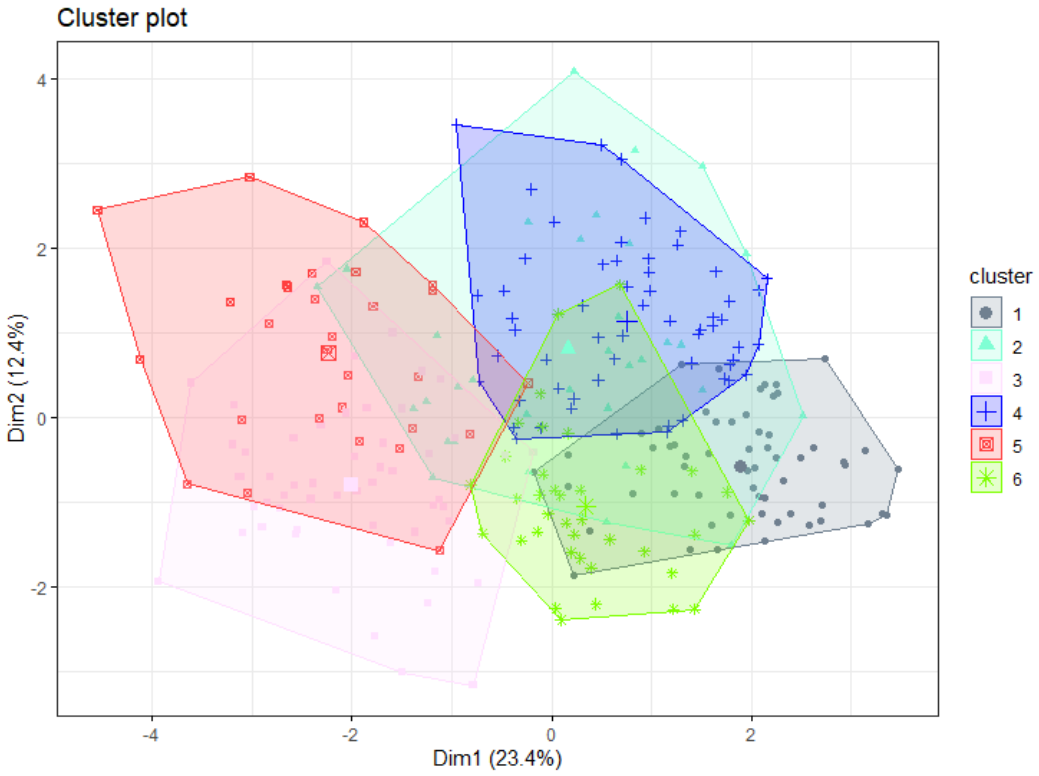
Para aplicar un algoritmo no supervisado al conjunto de datos, se escogió el de k-means.
Primero, se probó a desarrollar un algoritmo que dividiera el conjunto de datos en únicamente dos grupos. De este modo, se podría determinar su eficacia a la hora de clasificar los datos de la misma manera que un algoritmo supervisado: en función de la presencia o ausencia de la enfermedad. El resultado resultó ser bastante bueno obteniendo una precisión del 83,7%, aunque fue menor que los obtenidos con los algoritmos supervisados.



	Precisión	Sensibilidad	Especificidad	Medida F	Coef. Kappa
Resultado	83,7%	74,17%	91,33%	0.7864	0.665

En segundo lugar, se hizo uso del método del codo para determinar el modelo óptimo entre los que se generados con distintas varianzas interclusters. El punto en el que los valores se estabilizaban era aproximadamente con una k igual a 6, por lo que el modelo óptimo divide el conjunto de datos en 6 clusters que se muestran a continuación:





3. CONCLUSIONES

	Precisión	Sensibilidad	Especificidad	Medida F	Coef. Kappa
Árbol de decisión	89%	84%	92%	0.86	0.76
Redes neuronales	82.72%	82%	83,87%	0.8235	0.6429
Máquina de soporte vect	**90%**	**87%**	**92%**	**0.88**	**0.88**
Naive Bayes	87,65%	86,21%	88,46%	0.8692	0.7355
(K-medias)	83,7%	74,17%	91,33%	0.7864	0.665

Se puede concluir que el algoritmo con el que se consiguen mejores valores de rendimiento es el de soporte vectorial, aunque el árbol de decisión tiene una precisión muy cercana. Teniendo en cuenta que la base de datos empleada contiene datos clínicos, el algoritmo más recomendado de aplicar es el de árbol de decisión debido al fator de la explicabilidad. Además, su variante optimizada también podría ser un buen recurso al mantener una precisión alta con gasto computacional mínimo.