Implementación de un DD-Classifier en un contexto multivariado

Autores

Tanya Spee - tspee@eafit.edu.co
Daniel Ansia Dibuja - dansiad@eafit.edu.co

5 de septiembre de 2022

Índice

1. Introducción

En este proyecto se lleva a cabo la implementación de un procedimiento de clasificación no paramétrico basado en el concepto de profundidad y su extrapolación a las gráficas profundidad-profundidad (DD-plot). Para ello, se implementa la metodología establecida en el material de estudio [1].

Se conoce como profundidad a la medición de la centralidad de un punto con respecto a su nube de de datos. Cuanto más al centro del conjunto de datos se encuentre un punto, más "profundo"será, y viceversa.

El clasificador basado en DD-plots (DD-classifier) utiliza el concepto de profundidad para clasificar un conjunto de datos en base a las profundidades relativas con respecto a sus distribuciones específicas. Para medir la profundidad, se utilizarán en este proyecto diversos métodos, como las profundidades de Mahalanobis, Tukey y Likelihood Depth.

Es relevante mencionar que este análisis es no-paramétrico, por lo tanto no es necesario conocer la distribución del conjunto de datos para poder clasificarlos ni tampoco asumir dichas características. La curva de separación de dos distribuciones realizada por el DD-classifier es completamente determinada por la estructura de los datos.

En la sección 2 se expone la metodología seguida para la realización de este proyecto. La sección 3 incluye su implementación en distribuciones de datos generadas aleatoriamente y, finalmente, en la sección 4 se presenta su aplicación en un caso real con datos de clasificación relativos a la salud cardiaca de aproximadamente 400 pacientes en los que la mitad sufrieron un ataque cardíaco.

2. Metodología

Hay diversos métodos disponibles para medir la profundidad de una nube de puntos en un set de datos. Los métodos utilizados en este proyecto fueron las profundidades de Mahalanobis, de Tukey y Likelihood Depth.

La profundidad de Mahalanobis se basa en la medida de outliers definida de la siguiente manera:

$$MD(x, F) = \left[1 + (x - \mu)'\Sigma^{-1}(x - \mu)\right]^{-1}$$

La profundidad de Tukey en un punto es el número más pequeño de puntos en cualquier semiespacio cerrado que contenga un punto con respecto a un conjunto de datos. Un semiespacio es un subconjunto de un vector espacio caracterizado por una desigualdad $a'x \leq b$.

La profundidad de Tukey viene dada por:

$$HD_n(x, z) = \min_{\|a\|=1} \#\{i : a'x_i \leq a'x\}$$

Para el proyecto se han utilizado diferentes curvas polinómicas de clasificación (DD1, DD2 y DD3) y LDA.

2.1. Procedimiento

El procedimiento seguido para la realización del proyecto se detalla a continuación.

1. Análisis y entendimiento de las diferentes medidas de profundidad.
2. Análisis del paquete `fda.usc()` en `R`.
3. Implementación de los distintos algoritmos y medidas de profundidad a diferentes distribuciones generadas aleatoriamente.
4. Validación cruzada con diferentes algoritmos y medidas de profundidad para elegir los que mejor se adaptan a cada caso.

3. Implementación

En esta sección nos hemos centrado en implementar la metodología descrita para diferentes situaciones con respecto a la distribución de los datos. Hemos definido cuatro casos en los que la media y dispersión de las distribuciones generadas varía de diferentes maneras. Los resultados se muestran a continuación junto con las efectividades de las curvas de clasificación y las medidas de profundidad.

3.1. Igual media y dispersión

En un caso con dos distribuciones idénticas se puede observar como los datos se encuentran en la linea de 45^{o} ya que la profundidad con respecto a las dos distribuciones es la misma. El ratio de clasificación es obviamente 50 %.

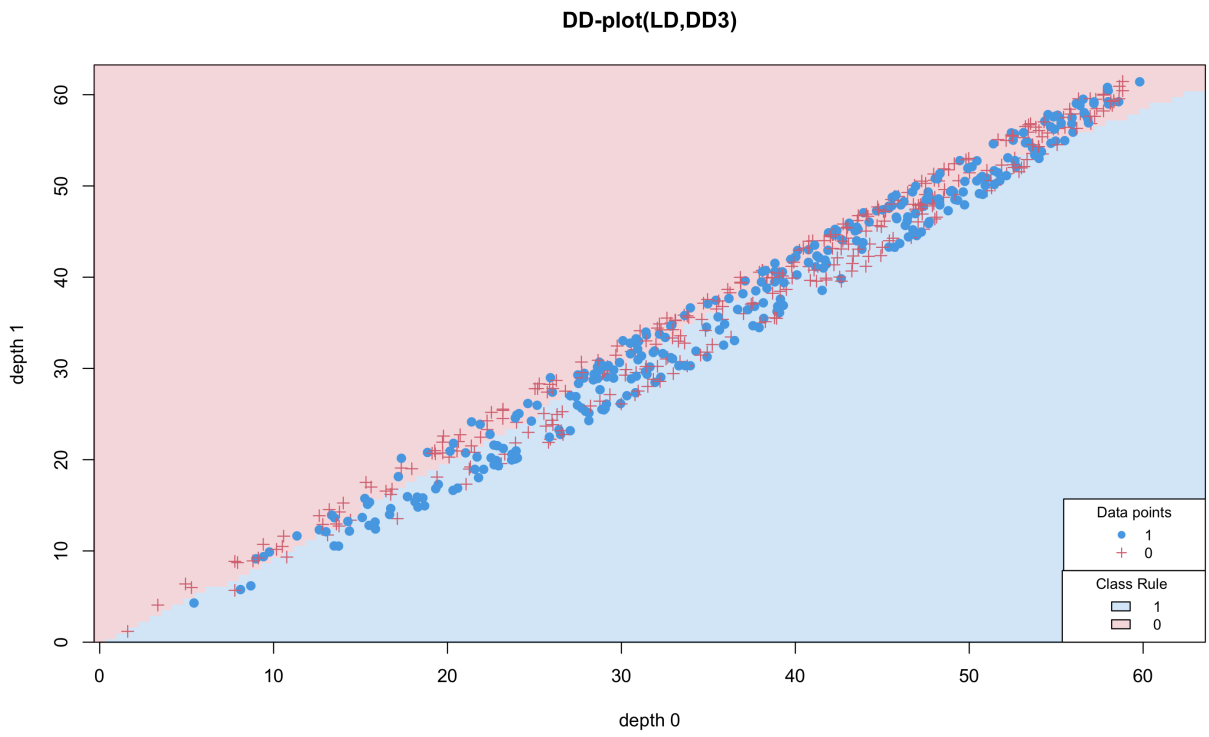

Figura 1: Resultados del clasificador (DD3) para dos distribuciones idénticas.

3.2. Igual media y diferente dispersión

En el segundo caso hemos creado dos distribuciones de datos con la misma media y diferentes matrices de dispersión (Σ). Podemos observar que la nube de datos en el DD-plot se arquea, lo que quiere explica que una de las densidades tiene una dispersión menor que la otra y por lo tanto sus profundidades con respecto a una distribución es mayor que con respecto a la otra.

Hemos probado diferentes clasificadores para comparar sus resultados. La siguiente gráfica muestra la curva de clasificación y el mejor clasificador (DD2).

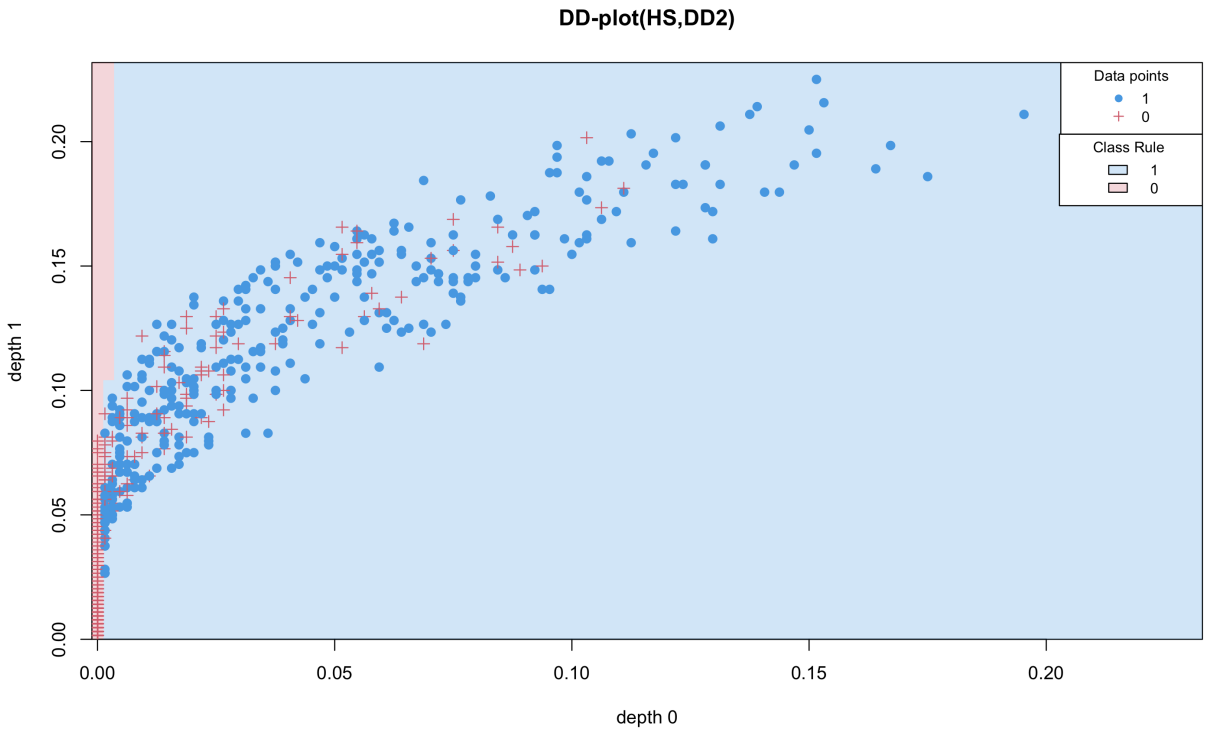

Figura 2: Resultados del clasificador (DD2 y HS) para dos distribuciones con igual media y diferente dispersión.

Se puede observar que, con respecto al DD-plot, la curva de clasificación es una linea recta con un ángulo específico. Transponiendo este clasificador a un plano tridimensional – ya que los conjuntos de datos tienen 3 dimensiones – se correspondería con un polígono esférico irregular.

Finalmente, la siguiente tabla muestra la efectividad de cada uno de los clasificadores utilizados.

Cuadro 1: Efectividad de cada clasificador y medida de profundidad para dos distribuciones con igual media y diferente dispersión.

Clasificador	**Profundidad**	**Probabilidad**
DD2	HS	0.864
DD3	HS	0.864
DD2	MhD	0.830
DD2	LD	0.828
DD3	MhD	0.826
LDA	MhD	0.797
DD1	MhD	0.490

Donde HS, MhD and LD se corresponden con las medidas de profundidad Half-Space,

distancia de Mahalanobis y Likelihood Depth. El clasificador y medidas de profundidad más efectivos son tanto DD2 como DD3 y Half-Space.

3.3. Diferente media e igual dispersión

El tercer caso se compone de dos distribuciones con distinta media e igual ratio de dispersión. En este caso se puede observar como los diferentes conjuntos de datos en el DD-plot comienzan a alejarse entre ellos y acercarse a cada uno de los ejes del gráfico, respectivamente.

En este caso el clasificador y medida de profundidad que mejor resultado dan son, otra vez, DD2 y Half-Space.

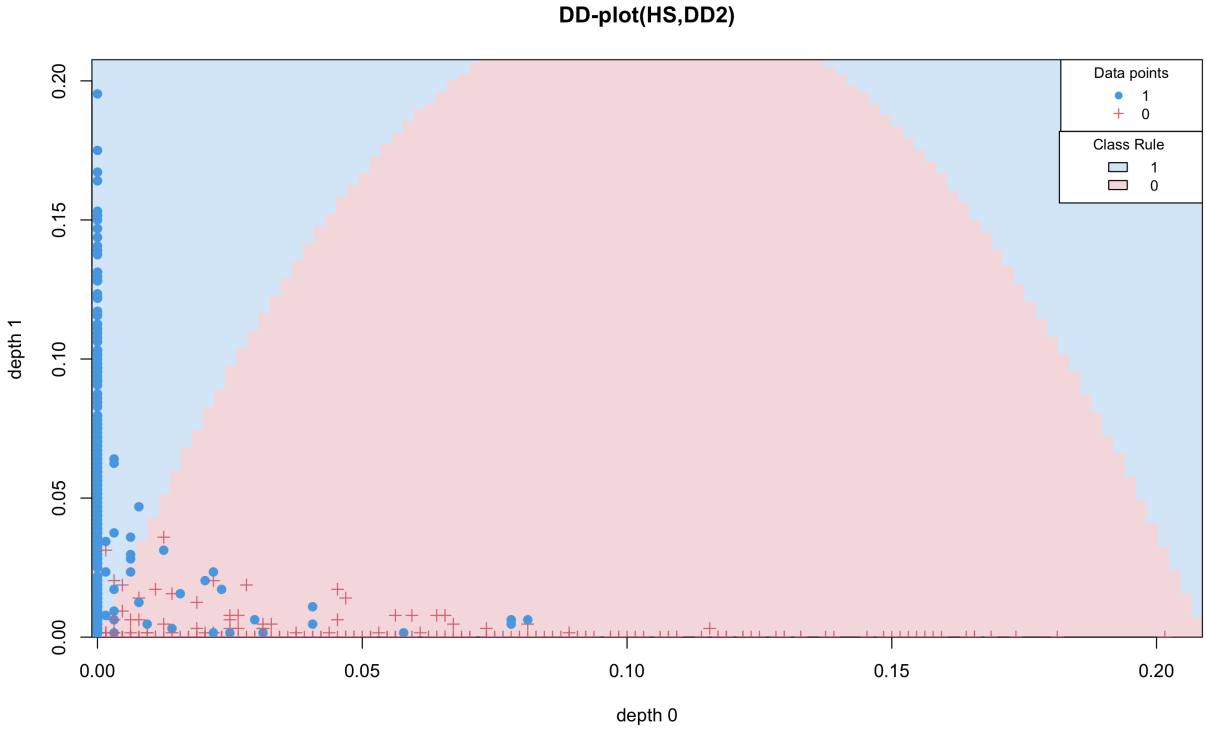

Figura 3: Resultados del clasificador (DD2 y HS) para dos distribuciones con diferente media e igual dispersión.

Cuadro 2: Efectividad de cada clasificador y medida de profundidad para dos distribuciones con diferente media e igual dispersión.

Clasificador	Profundidad	Probabilidad
DD2	HS	0.961
DD2	MhD	0.955
DD2	LD	0.955
DD3	MhD	0.955
DD1	MhD	0.955
DD3	HS	0.953
LDA	MhD	0.950

Todos los conjuntos de clasificadores y medidas de profundidad se comportan de manera muy positiva, con clasificaciones correctas el 95 % de las veces.

3.4. Diferente media y dispersión

En el último caso, con diferente media y dispersión, se puede observar como una de las distribuciones se aproxima a un eje mientras que la otra se encuentra dispersa por todo la parte inferior del gráfico.

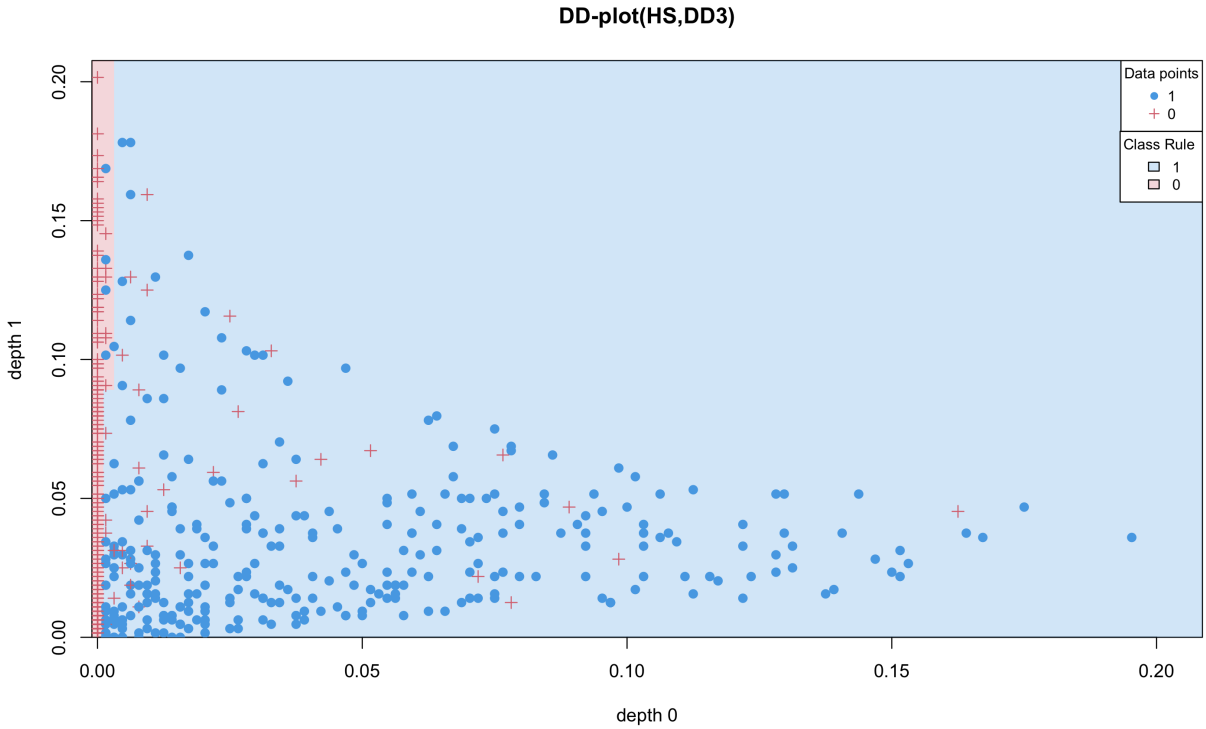

Figura 4: Resultados del clasificador (DD3 y HS) para dos distribuciones con diferente media y dispersión.

Cuadro 3: Efectividad de cada clasificador y medida de profundidad para dos distribuciones con diferente media y dispersión.

Clasificador	**Profundidad**	**Probabilidad**
DD3	HS	0.937
DD2	HS	0.934
DD2	MhD	0.925
DD2	LD	0.925
DD3	MhD	0.917
DD1	MhD	0.910
LDA	MhD	0.844

4. Aplicación

En esta sección hemos implementado la metodología en un conjunto de datos de predicción de insuficiencia cardíaca [2]. Se han utilizado los siguientes datos numéricos para predecir la insuficiencia cardíaca de una persona, cada característica consta de unos 400 valores:

- Age: Edad de los pacientes
- Oldpeak: Prueba de aterosclerosis coronaria obstructiva
- Hemoglobin: Niveles de hemoglobina en sangre
- Trestbps: Presión sanguínea en reposo
- Chol: Nivel de colesterol
- Thalach: Ritmo cardíaco máximo conseguido
- BGR: Niveles del regulador de glucosa en sangre
- RBC: Densidad de glóbulos rojos en sangre
- Mortalidad: muertes causadas por ataques cardíacos

Para la validación cruzada de las diferentes medidas de profundidad y métodos de clasificación, hemos utilizado un 80 % aleatorio del conjunto de datos. Esto es para fines de entrenamiento y predicción más adelante.

Los resultados se muestran a continuación junto con la eficacia de las curvas de valoración y las medidas de profundidad. En la cuadro 4 sólo se muestran las mejores combinaciones de medidas de profundidad y curvas de clasificación. La mejor clasificación, la profundidad Mahalanobis con la curva DD3 se muestra en el gráfico. También se puede ver visualmente que el clasificador tiene un buen rendimiento.

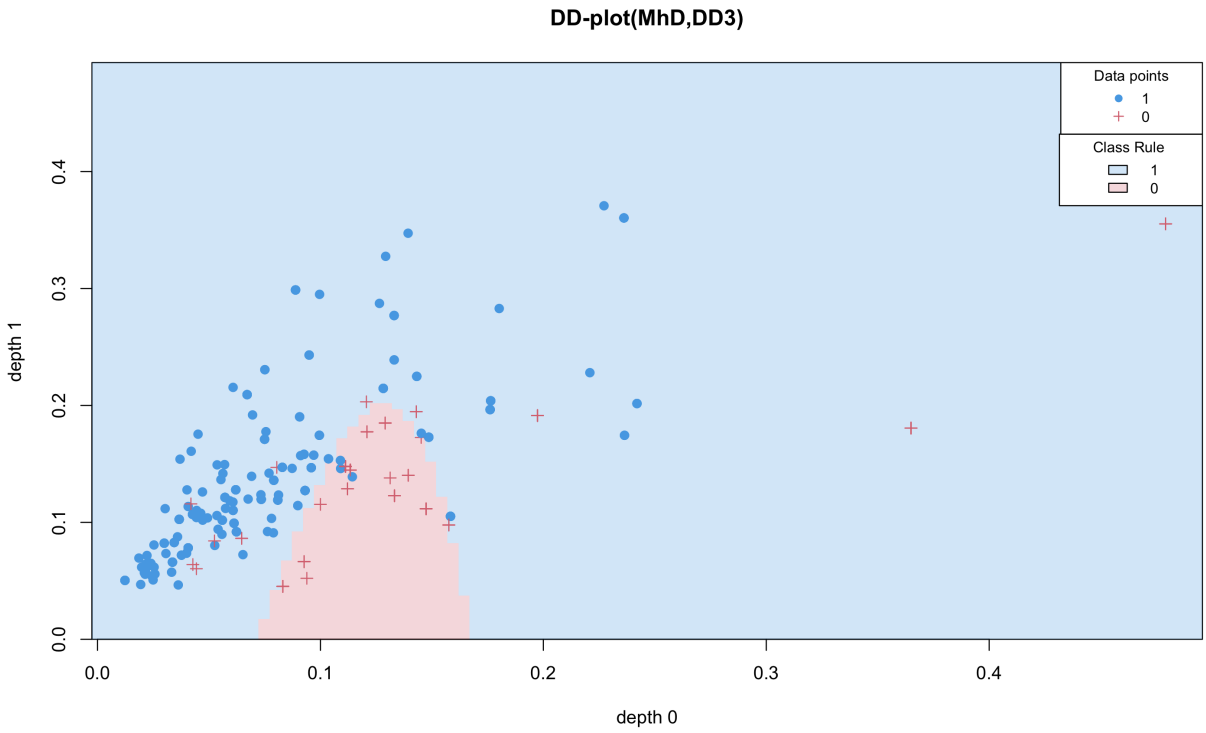

Figura 5: Resultados del clasificador para el caso de aplicación (DD3 y MhD).

Cuadro 4: Efectividad de cada clasificador y medida de profundidad para el caso de aplicación.

Clasificador	**Profundidad**	**Probabilidad**
DD3	HS	0.915
DD2	HS	0.867
DD2	MhD	0.857
DD2	LD	0.837
DD3	MhD	0.928
DD1	MhD	0.827
LDA	MhD	0.810

Una vez elegido el mejor método de profundidad y la mejor curva de clasificación, vamos a probar nuestro método. Hemos entrenado un modelo de clasificación utilizando la profundidad de Mahalanobis y la curva DD3 en el 80 % de los datos que hemos utilizado anteriormente. Hemos utilizado el 20 % restante de los datos para predecir el resultado, es decir, si es probable que alguien tenga un fallo cardíaco o no. Los resultados se muestran a continuación en una matriz de confusión.

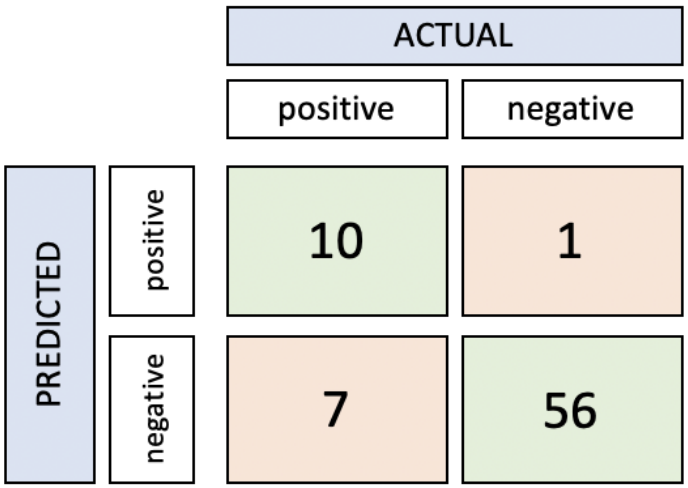

Figura 6: Matriz de confusión para el resultado previsto de que alguien tenga una enfermedad cardíaca

De la matriz de confusión se desprende que la predicción es correcta para el 89 %. Por lo tanto, podemos concluir que nuestro método funciona con éxito en esta aplicación.

Referencias

[1] J. Li, J. A. Cuesta-Albertos, and R. Y. Liu, "Dd-classifier: Nonparametric classification procedure based on dd-plot," *Journal of the American Statistical Association*, vol. 107, no. 498, pp. 737–753, 2012.

[2] "The heart failure prediction dataset." `https://www.kaggle.com/datasets/asgharalikhan/mortality-rate-heart-patient-pakistan-hospital`. 2022-09-05.