# Trabajo Final de Aprendizaje Automático y Minería de Datos

Rubén García Macho Inés Ibañez Valle

Mayo 2022

## 1. Introduction

Nuestro objetivo es detectar enfermedades cardiacas probando distintos modelos de predicción con la herramienta de Knime. Para ello hemos utilizado algoritmos de clasificación.
El link a la carpeta de drive es:
https://unicancloud-my.sharepoint.com/:f:/g/personal/rgm177_alumnos_unican_es/EhNGQ6a_gABPijhnGtu3Y18BtbXPHptedRW7UqXd-Hzd5Q?e=hXd3YJ

## 2. Procesado del dato

Para poder hallar unas buenas predicciones a o largo de este trabajo hemos tenido que hacer un preprocesado de los datos obtenidos mediante el dataset. A continuación describiremos los métodos usados para el procesado del dato, su descripción y el porqué los usamos.

### 2.1. CSV Reader

Es el nodo que usamos para leer los datos del dataset, como nuestro dataset tenía el formato CSV utilizamos este nodo antes que un file reader parq tener una mayor optimalidad.

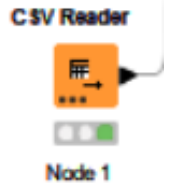


Figura 1: CSV reader

### 2.2. Normalizer

Para poder pasar los datos y utilizarlos de una forma correcta teníamos que normalizarlos, para esto utilizamos el nodo *Normalizer* en el se aplica una normalización de tipo *Min-Max* con un mínimo de 0.0 y máximo de 1.0 .En este nodo normalizaremos todos los parámetros no discretos

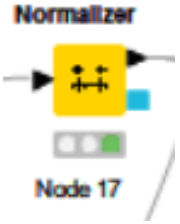


Figura 2: Normalizer

## 2.3. PCA

El nodo que halla el PCA(*PrincipalComponent Analysis*) es un nodo que se utiliza para reducir la dimensionalidad de un dataset de $R^n$ a $R^k$ oteniendo el mayor nivel de información.
El segundo nodo del PCA(*PCA Apply*) Es el nodo que aplica la computación del primer nodo.

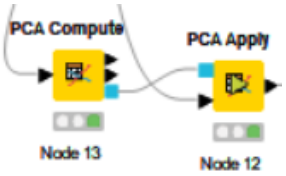


Figura 3: nodos PCA

## 2.4. Column Filter

El *Column Filter* es el nodo que utilizaremos para reducir la dimensionalidad del dataset, hemos obtenido las columnas nuevas del PCA para reducir la dimensionalidad, ahora tendremos que quitar las columnas "viejas"para tener una mayor precisión en la predicción.

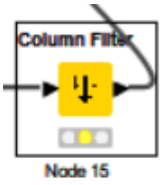


Figura 4: nodo del Column Filter

## 2.5. Particionado del Dataset

Hemos particionado el dataset en 3 distintas particiones: *Training, Cross-validation* y test, para ello hemos utilizado 2 nodos de particionado, uno donde se dividen los datos en un 70 % y un 30 % y otro que particiona el previo 30 % en un 60 % y 40 %, de esta forma podremos dividir los datos de una forma que se puedan utilizar bien sin que haya fallos de ajuste en las predicciones.

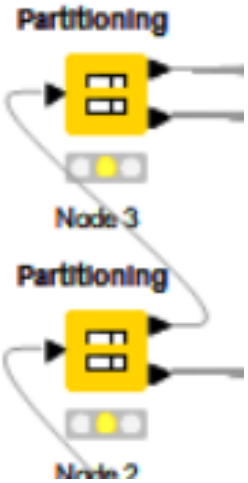


Figura 5: nodos de particionado

# 3. Modelos de predicción utilizados

Hemos usado diversos métodos de predicción de la clasificación.

## 3.1. Regresión logística

En este método hemos clasificado los resultados a partir del algoritmo de regresión logística.

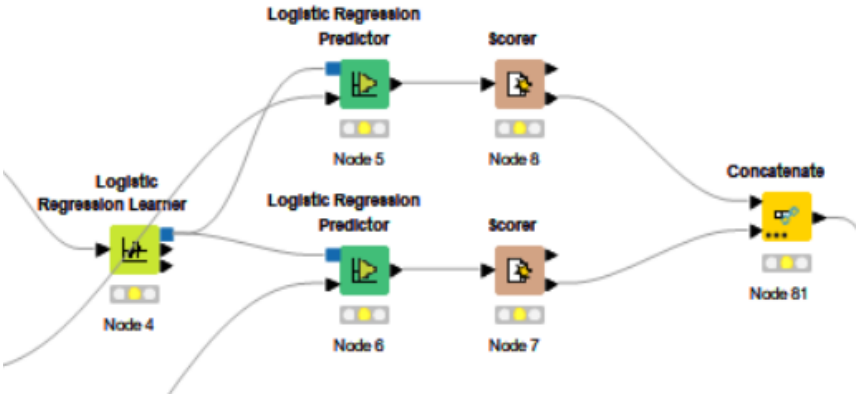


Figura 6: Regresión logística

Aquí vemos como hemos realizado la regresión logística utilizando el *learner* y los *predictor*. Hemos configurado el learner con 10000 *epochs* para que el algoritmo converja, de esta forma, hemos llegado a una precisión de un 82.352 por ciento de acierto con la dimensionalidad reducida a 5, 6 y 7 dimensinoes por el PCA.

## 3.2. Red Neuronal Probabilista(PNN)

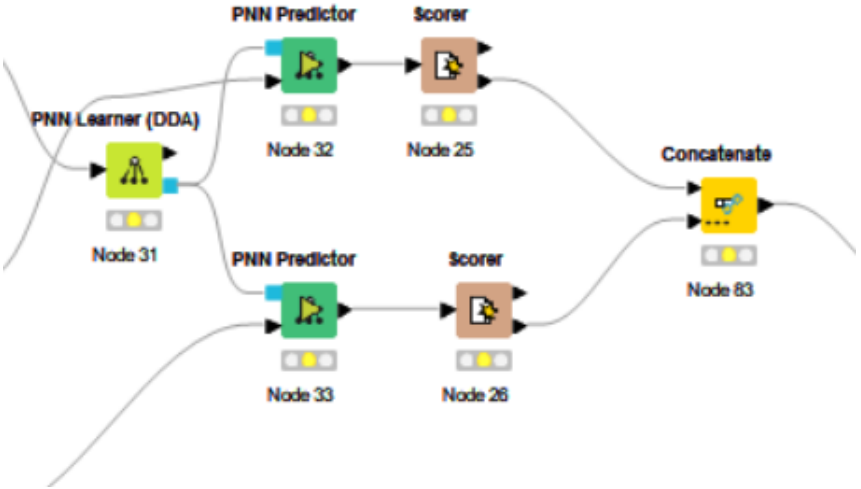


Figura 7: Red Neuronal Probabilista

Aquí vemos la distribución de los nodos en este método, hemos aplicado el PNN utilizando, como en el método anterior y como en la mayoría de los que vamos a dar durante el discurso de este apartado de Modelos de predicción, un *learner* y 2 *predictors*.

El nodo de PNN es un nodo que implementa una red neuronal probabilista basada en algoritmos de propagación hacia delante.

Para poder saber que tipo de red era la mas precisa, hemos implementado los 7 tipos de red(*Best Guess, Incorp, Mean, Min, Max, One, Zero*) y hemos comparado sus resultados, de la siguiente forma, hemos llegado a los siguientes resultados:

Analizando los resultados hemos obtenido los mismos dando igual que algoritmo usásemos, siempre se ha obtenido como valor máximo 75.294 por ciento de acierto con una dimensionalidad de 2 y 7.

## 3.3. Máquinas de soporte vectorial(SVM)

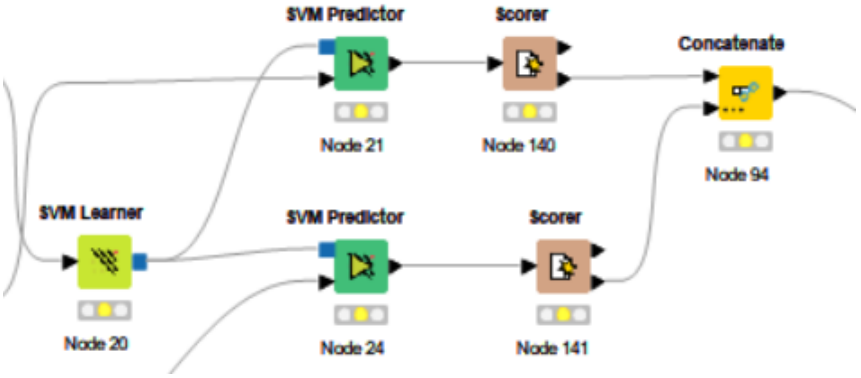


Figura 8: Máquina de soporte vectorial

Hemos utilizado las Máquinas de soporte vectorial para la clasificación de los datos también, de forma análoga a la anterior, hemos implemetado una SVM para cada variación del método que haya, de esta forma hemos obtenido 3 *Support Vector Machines*: *HyperTangent, Polynomial y RBF*.

Hemos obtenido los siguientes resultados por cada Método:

- SVM Polynomial: 74.117 por ciento reduciendo la dimensionalidad a 1
- SVM: HyperTangent: 81.176 por ciento reduciendo la dimensionalidad a 2
- SVM RBF: 80 por ciento bajando la dimensionalidad a 8 y 10

### 3.4. MLP: *Multilayer Perceptron*

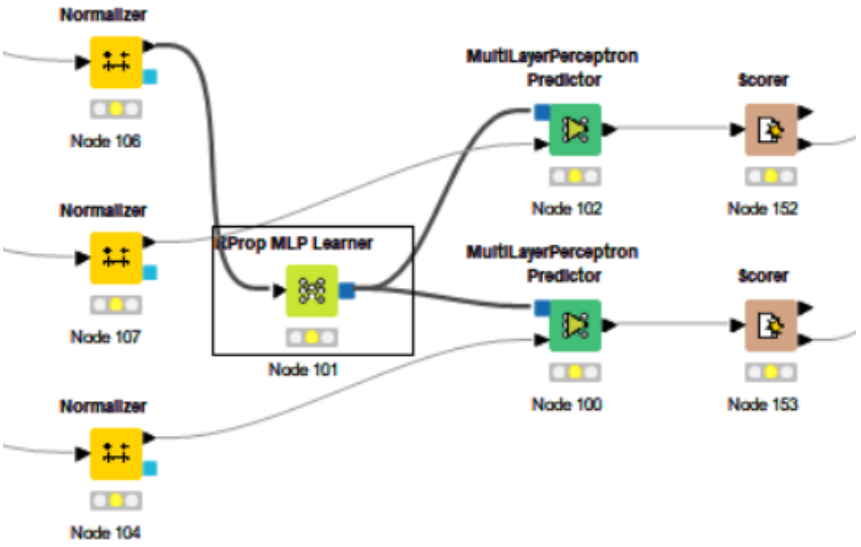


Figura 9: *Multilayer Perceptron*

En este nodo se ve una particularidad, hemos tenido que volver a normalizar los resultados, esto es porque la MLP necesita que se vuelvan a normalizar.

En este Método hemos utilizado iteradores para hallar en cada caso el número de capas y nuronas por capa más óptimo. El mayor *accuracy* obtenido ha sido con 9 capas y 51 neuronas, con un 81.176 por ciento de precisión, como se ve en la teoría, hemos hallado la mayor precisión en un numero de neuronas intermedio, ni muy alto, ni muy bajo, esto sería porque podria haber fallos de ajuste, como el *underfitting* en caso de que hubiese pocas neuronas o el *overfitting* en caso de que haya demasiadas neuronas. Esto lo hicimos para encontrar las neuronas más óptimas, después de esto las dejamos constantes y iteramos en función de las dimensiones del PCA para hallar la dimensinalidad más óptima, Con esto hallamos una precision del 78.823 por ciento con una dimensionalidad bajada a 3 dimensiones.

# 4. Resultados

Los resultados obtenidos han sido comentados en el transcurso de esta memoria, pero, en resumen, a parte de habiendo confirmado lo visto en teoría, hemos observado que la regresión logística es el mejor método de predicción. A continuación vamos a adjuntar una imagen de la mayoría de los métodos de predicción.

| | | Accuracy | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | | 0.741176 | 0.729412 | 0.741176 | 0.741176 | 0.823529 | 0.823529 | 0.823529 | 0.811765 | 0.811765 | 0.776471 | 0.823529 |
| LR CrossValidation | | 0.714286 | 0.736264 | 0.758242 | 0.758242 | 0.758242 | 0.813187 | 0.802198 | 0.802198 | 0.802198 | 0.78022 | |
| PNN Best Guess | | 0.705882 | 0.752941 | 0.752941 | 0.752941 | 0.752941 | 0.741176 | 0.752941 | 0.729412 | 0.741176 | 0.741176 | 0.752941 |
| CrossValidation | | 0.714286 | 0.714286 | 0.725275 | 0.769231 | 0.725275 | 0.725275 | 0.703297 | 0.692308 | 0.747253 | 0.736264 | |
| PNN Incorp | | 0.705882 | 0.752941 | 0.752941 | 0.752941 | 0.752941 | 0.741176 | 0.752941 | 0.729412 | 0.741176 | 0.741176 | 0.752941 |
| CrossValidation | | 0.714286 | 0.714286 | 0.725275 | 0.769231 | 0.725275 | 0.725275 | 0.703297 | 0.692308 | 0.747253 | 0.736264 | |
| PNN mean | | 0.705882 | 0.752941 | 0.752941 | 0.752941 | 0.752941 | 0.741176 | 0.752941 | 0.729412 | 0.741176 | 0.741176 | 0.752941 |
| CrossValidation | | 0.714286 | 0.714286 | 0.725275 | 0.769231 | 0.725275 | 0.725275 | 0.703297 | 0.692308 | 0.747253 | 0.736264 | |
| PNN min | | 0.705882 | 0.752941 | 0.752941 | 0.752941 | 0.752941 | 0.741176 | 0.752941 | 0.729412 | 0.741176 | 0.741176 | 0.752941 |
| CrossValidation | | 0.714286 | 0.714286 | 0.725275 | 0.769231 | 0.725275 | 0.725275 | 0.703297 | 0.692308 | 0.747253 | 0.736264 | |
| PNN max | | 0.705882 | 0.752941 | 0.752941 | 0.752941 | 0.752941 | 0.741176 | 0.752941 | 0.729412 | 0.741176 | 0.741176 | 0.752941 |
| CrossValidation | | 0.714286 | 0.714286 | 0.725275 | 0.769231 | 0.725275 | 0.725275 | 0.703297 | 0.692308 | 0.747253 | 0.736264 | |
| PNN one | | 0.705882 | 0.752941 | 0.752941 | 0.752941 | 0.752941 | 0.741176 | 0.752941 | 0.729412 | 0.741176 | 0.741176 | 0.752941 |
| CrossValidation | | 0.714286 | 0.714286 | 0.725275 | 0.769231 | 0.725275 | 0.725275 | 0.703297 | 0.692308 | 0.747253 | 0.736264 | |
| PNN Zero | | 0.705882 | 0.752941 | 0.752941 | 0.752941 | 0.752941 | 0.741176 | 0.752941 | 0.729412 | 0.741176 | 0.741176 | 0.752941 |
| CrossValidation | | 0.714286 | 0.714286 | 0.725275 | 0.769231 | 0.725275 | 0.725275 | 0.703297 | 0.692308 | 0.747253 | 0.736264 | |
| SVM HyperTangent | | 0.741176 | 0.682353 | 0.705882 | 0.705882 | 0.705882 | 0.717647 | 0.717647 | 0.717647 | 0.717647 | 0.717647 | 0.741176 |
| CrossValidation | | 0.714286 | 0.736264 | 0.747253 | 0.747253 | 0.736264 | 0.736264 | 0.736264 | 0.736264 | 0.736264 | 0.736264 | |
| SVM RBF | | 0.741176 | 0.811765 | 0.788235 | 0.776471 | 0.741176 | 0.694118 | 0.658824 | 0.658824 | 0.588235 | 0.529412 | 0.811765 |
| CrossValidation | | 0.659341 | 0.725275 | 0.758242 | 0.725275 | 0.714286 | 0.648352 | 0.626374 | 0.571429 | 0.571429 | 0.571429 | |
| SVM polynomial | | 0.741176 | 0.741176 | 0.705882 | 0.705882 | 0.705882 | 0.752941 | 0.776471 | 0.8 | 0.788235 | 0.8 | 0.8 |
| CrossValidation | | 0.714286 | 0.736264 | 0.747253 | 0.747253 | 0.747253 | 0.758242 | 0.747253 | 0.78022 | 0.78022 | 0.747253 | |
| MLP | | 0.682353 | 0.764706 | 0.788235 | 0.764706 | 0.788235 | 0.741176 | 0.752941 | 0.717647 | 0.729412 | 0.705882 | 0.788235 |
| CrossValidation | | 0.681319 | 0.736264 | 0.758242 | 0.758242 | 0.802198 | 0.813187 | 0.736264 | 0.747253 | 0.747253 | 0.703297 | |
| | Dimension | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| | | | | | | | | | | | | |
| | max | 0.741176 | 0.811765 | 0.788235 | 0.776471 | 0.823529 | 0.823529 | 0.823529 | 0.811765 | 0.811765 | 0.8 | |

Figura 10: Resultados de la mayoría de los elementos de predicción en diversas dimensiones

## 4.1. Reglas de asociación

Presentamos a continuación las reglas de asociación que tienen como consecuencia la enfermedad cardiaca o la ausencia de esta. Las hemos obtenido con el nodo correspondiente fijando el soporte mínimo a 0.21 y la confianza mínima 0.8.

| **EXPLICACIÓN DE ATRIBUTOS** |
|---|
| *adult*: personas de 36 años o mayores y menores de 60. |
| *angina*: el ejercicio induce a dolor de angina. |
| *asymptomatic*: ninguno de los criterios del dolor de angina son relacionables con los dolores padecidos. |
| $female$: mujer. |
| $flat$: valor referido a la pendiente del segmento ST de ejercicio máximo. |
| $fluoroscopy_0$: 0 venas importantes coloreadas por el fluoroscopio. |
| $highBP$: presión sanguinea en reposo ¿120 mmHg. |
| $highChol$: grado de colesterol en sangre ¿200 mg/dL. |
| $noAngina$: el ejercicio no induce a dolor de angina. |
| $noDiabetic$: nivel de azúcar en sangre ¡120 mg/dL. |
| $noDiference$: la diferencia entre el resultado de una prueba en reposo y durante el ejercicio es 0. |
| $normalECD$: resultado normal de un electrocardiograma en reposo. |
| $normalHR$: latidos por minuto que el corazón puede alcanzar durante el ejercicio (160-190) |
| $normalThal$: el tejido del corazón absorve talio de forma normal. |
| $reversableThal$: el tejido del corazón no absorve talio en la parte de ejercicio de la prueba. |
| $upsloping$: valor referido a la pendiente del segmento ST de ejercicio máximo. |

| Soporte | Confianza | REGLA DE ASOCIACIÓN |
|---|---|---|
| 0.22442244 | 0.81927711 | $reversableThal, highBP \rightarrow YesAHD$ |
| 0.22442244 | 0.80952381 | $asymptomatic, flat \rightarrow YesAHD$ |
| 0.23102310 | 0.875 | $asymptomatic, angina \rightarrow YesAHD$ |
| 0.2343234 | 0.91025641 | $reversableThal, asymptomatic \rightarrow YesAHD$ |

| Soporte | Confianza | REGLA DE ASOCIACIÓN |
|---|---|---|
| 0.21122112 | 0.853334 | $female, noAngina \rightarrow NoAHD$ |
| 0.21122112 | 0.8 | $noDiference, noAngina \rightarrow NoAHD$ |
| 0.21122112 | 0.82051282 | $normalHR, upsloping \rightarrow NoAHD$ |
| 0.21122112 | 0.87671233 | $normalECD, fluoroscopy_0, noAngina \rightarrow NoAHD$ |
| 0.21122112 | 0.90140845 | $normalThal, normalHR, noAngina \rightarrow NoAHD$ |
| 0.21122112 | 0.8 | $normalECD, adult, noAngina \rightarrow NoAHD$ |
| 0.21122112 | 0.8421053 | $normalHR, adult, noAngina \rightarrow NoAHD$ |
| 0.21122112 | 0.85333334 | $highChol, normalThal, noDiabetic, upsloping \rightarrow NoAHD$ |
| 0.21452145 | 0.87837837 | $normalThal, adult, upsloping \rightarrow NoAHD$ |
| 0.21452145 | 0.89041096 | $highChol, normalThal, fluoroscopy_0, noDiabetic, noAngina \rightarrow NoAHD$ |
| 0.21782178 | 0.85714286 | $highChol, fluoroscopy_0, upsloping \rightarrow NoAHD$ |
| 0.21782178 | 0.9295775 | $normalThal, fluoroscopy_0, upsloping \rightarrow NoAHD$ |
| 0.21782178 | 0.868421053 | $highChol, normalThal, noAngina, upsloping \rightarrow NoAHD$ |
| 0.2211221 | 0.8933334 | $normalECD, normalThal, noAngina \rightarrow NoAHD$ |
| 0.2211221 | 0.85897436 | $normalECD, normalThal, noDiabetic \rightarrow NoAHD$ |
| 0.22442244 | 0.90666667 | $normalHR, fluoroscopy_0 \rightarrow NoAHD$ |
| 0.22442244 | 0.82926829 | $normalECD, upsloping \rightarrow NoAHD$ |
| 0.22442244 | 0.85000001 | $fluoroscopy_0, highBP, noAngina \rightarrow NoAHD$ |
| 0.22442244 | 0.91891891 | $highChol, normalThal, fluoroscopy_0, adult \rightarrow NoAHD$ |
| 0.22442244 | 0.86075949 | $normalThal, noDiabetic, noAngina, upsloping \rightarrow NoAHD$ |
| 0.227722773 | 0.884615385 | $normalThal, normalHR \rightarrow NoAHD$ |
| 0.227722773 | 0.8625000002 | $normalThal, female \rightarrow NoAHD$ |
| 0.2310231 | 0.875 | $highChol, normalThal, adult, noAngina \rightarrow NoAHD$ |
| 0.2310231 | 0.94594594 | $normalThal, fluoroscopy_0, adult, noAngina \rightarrow NoAHD$ |
| 0.234323432 | 0.816091954 | $normalHR, adult \rightarrow NoAHD$ |
| 0.234323432 | 0.8875 | $fluoroscopy_0, noAngina, upsloping \rightarrow NoAHD$ |
| 0.234323432 | 0.8554217 | $fluoroscopy_0, noDiabetic, upsloping \rightarrow NoAHD$ |
| 0.234323432 | 0.8452381 | $highChol, fluoroscopy_0, adult, noAngina \rightarrow NoAHD$ |
| 0.23762376 | 0.9 | $highChol, normalThal, fluoroscopy_0, noAngina \rightarrow NoAHD$ |
| 0.23762376 | 0.8372093 | $highChol, normalThal, adult, noDiabetic \rightarrow NoAHD$ |
| 0.2409240924 | 0.80219780 | $normalHR, noAngina \rightarrow NoAHD$ |
| 0.2409240924 | 0.85882353 | $normalECD, normalThal \rightarrow NoAHD$ |
| 0.2409240924 | 0.9240506 | $normalThal, fluoroscopy_0, adult, noDiabetic \rightarrow NoAHD$ |
| 0.244224422 | 0.831460674 | $normalThal, highBP, noAngina \rightarrow NoAHD$ |
| 0.244224422 | 0.8705882353 | $normalThal, adult, noDiabetic, noAngina \rightarrow NoAHD$ |
| 0.24752475 | 0.862068965 | $highChol, normalThal, upsloping \rightarrow NoAHD$ |
| 0.2508250825 | 0.85393258 | $normalThal, noDiabetic, upsloping \rightarrow NoAHD$ |
| 0.2508250825 | 0.8735632184 | $highChol, normalThal, fluoroscopy_0, noDiabetic \rightarrow NoAHD$ |
| 0.2508250825 | 0.8444444 | $fluoroscopy_0, adult, noDiabetic, noAngina \rightarrow NoAHD$ |
| 0.2607260726 | 0.86813187 | $normalThal, noAngina, upsloping \rightarrow NoAHD$ |
| 0.2607260726 | 0.8144329897 | $highChol, fluoroscopy_0, noDiabetic, noAngina \rightarrow NoAHD$ |
| 0.2640264 | 0.8695652174 | $fluoroscopy_0, upsloping \rightarrow NoAHD$ |

| Soporte | Confianza | REGLA DE ASOCIACIÓN |
|---|---|---|
| 0.2640264 | 0.930232558 | $normalThal, fluoroscopy_0, adult \rightarrow NoAHD$ |
| 0.2640264 | 0.8988764045 | $normalThal, fluoroscopy_0, noDiabetic, noAngina \rightarrow NoAHD$ |
| 0.26732673267 | 0.8350515464 | $highChol, normalThal, adult \rightarrow NoAHD$ |
| 0.26732673267 | 0.8350515464 | $highChol, normalThal, noDiabetic, noAngina \rightarrow NoAHD$ |
| 0.2739273927 | 0.87368421 | $normalThal, adult, noAngina \rightarrow NoAHD$ |
| 0.27722772 | 0.8842105263 | $highChol, normalThal, fluoroscopy_0 \rightarrow NoAHD$ |
| 0.2805280528 | 0.8415841584 | $normalThal, adult, noDiabetic \rightarrow NoAHD$ |
| 0.287128712 | 0.861386138 | $fluoroscopy_0, adult, noAngina \rightarrow NoAHD$ |
| 0.2904290429 | 0.862745098 | $normalThal, upsloping \rightarrow NoAHD$ |
| 0.29372937294 | 0.8317757 | $highChol, fluoroscopy_0, noAngina \rightarrow NoAHD$ |
| 0.29372937294 | 0.9081632653 | $normalThal, fluoroscopy_0, noAngina \rightarrow NoAHD$ |
| 0.3036303630363 | 0.8761904762 | $normalThal, fluoroscopy_0, noDiabetic \rightarrow NoAHD$ |
| 0.306930693 | 0.801724138 | $noAngina, upsloping \rightarrow NoAHD$ |
| 0.306930693 | 0.830357143 | $highChol, normalThal, noAngina \rightarrow NoAHD$ |
| 0.3135313531 | 0.840707964 | $normalThal, adult \rightarrow NoAHD$ |
| 0.3234323432 | 0.8305084746 | $fluoroscopy_0, noDiabetic, noAngina \rightarrow NoAHD$ |
| 0.32673267 | 0.83898305 | $normalThal, noDiabetic, noAngina \rightarrow NoAHD$ |
| 0.33663366 | 0.88695652174 | $normalThal, fluoroscopy_0 \rightarrow NoAHD$ |
| 0.36963696 | 0.84848484 | $fluoroscopy_0, noAngina \rightarrow NoAHD$ |
| 0.372937294 | 0.837037037 | $normalThal, noAngina \rightarrow NoAHD$ |

## 4.2. Metodologías alternativas

Hemos probado otras metodologías de clasificación como son Clusterin con el algoritmo Kmeans que resultó muy mal predictor y por eso sus resultados no se registrarán en este informe. También construimos un árbol de decisión del que no hemos sacado conclusiones precisas.