Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Inteligencia Artificial y Redes Neuronales

Actividad 5.- Árbol de decisión en Weka

Equipo N°7

Juan Pablo Ambriz Amador 1864353 N.L. 18 IB

Profesor. **Ing Ernesto Zambrano Serrano**

Programa Educativo. **IB**

Semestre Agosto – Diciembre 2021

Miércoles - V4

San Nicolás de los Garza, N.L. Miércoles 11 de Noviembre de 2021

Objetivo: Entrenar un árbol de decisión en Weka con el dataset "Heart Disease UCI".

Antecedentes: Esta base de datos contiene 76 atributos, pero todos los experimentos publicados se refieren al uso de un subconjunto de 14 de ellos. En particular, la base de datos de Cleveland es la única que han utilizado los investigadores de ML para
esta fecha. El campo "objetivo" se refiere a la presencia de enfermedad cardíaca en el paciente. Tiene un valor entero de 0 (sin presencia) a 4.

Información de atributos:

I. la edad
II. sexo
III. tipo de dolor en el pecho (4 valores)
IV. presión arterial en reposo
V. colestoral sérica en mg / dl
VI. azúcar en sangre en ayunas> 120 mg / dl
VII. resultados electrocardiográficos en reposo (valores 0,1,2)
VIII. frecuencia cardíaca máxima alcanzada
IX. angina inducida por ejercicio
X. oldpeak = depresión del ST inducida por el ejercicio en relación con el reposo
XI. la pendiente del segmento ST de ejercicio pico
XII. número de vasos principales (0-3) coloreados por la floración
XIII. thal: 3 = normal; 6 = defecto fijo; 7 = defecto reversible

Descripción del dataset: qué información nos da cada columna

1. la edad	Variable rango abierto. Años
2. sexo	Hombre (1) Mujer (0)
3. tipo de dolor en el pecho (4 valores)	Tipo de dolor CP (0,1,2,3)
4. presión arterial en reposo	Variable rango abierto. Presión Arterial
5. colesterol sérica en mg / dl	Variable rango abierto. Colesterol
6. azúcar en sangre en ayunas> 120 mg / dl	Variable Azucar en Sangre >120mg (1) <120mg (0)
7. resultados electrocardiográficos en reposo (valores 0,1,2)	Tipo de Valor Electrocardiografo (0,1,2,3)
8. frecuencia cardíaca máxima alcanzada	Variable rango (70 - 202) Frecuencia C.
9. angina inducida por ejercicio	Inducida (1) No inducida (0)

10. oldpeak = depresión del ST inducida por el ejercicio en relación con el reposo	Variable rango abierto. Depresión del ST
11. la pendiente del segmento ST de ejercicio pico	Slope, Pendiente (1,2,3)
12. número de vasos principales (0-3) coloreados por la floración	Ca (0,1,2,3)
13. thal: 3 = normal; 6 = defecto fijo; 7 = defecto reversible	Rango de Thal (1,2,3)
14. Target / Objetivo	Sobreviviente (1) No sobreviviente (0)

Metodología y Experimentación:

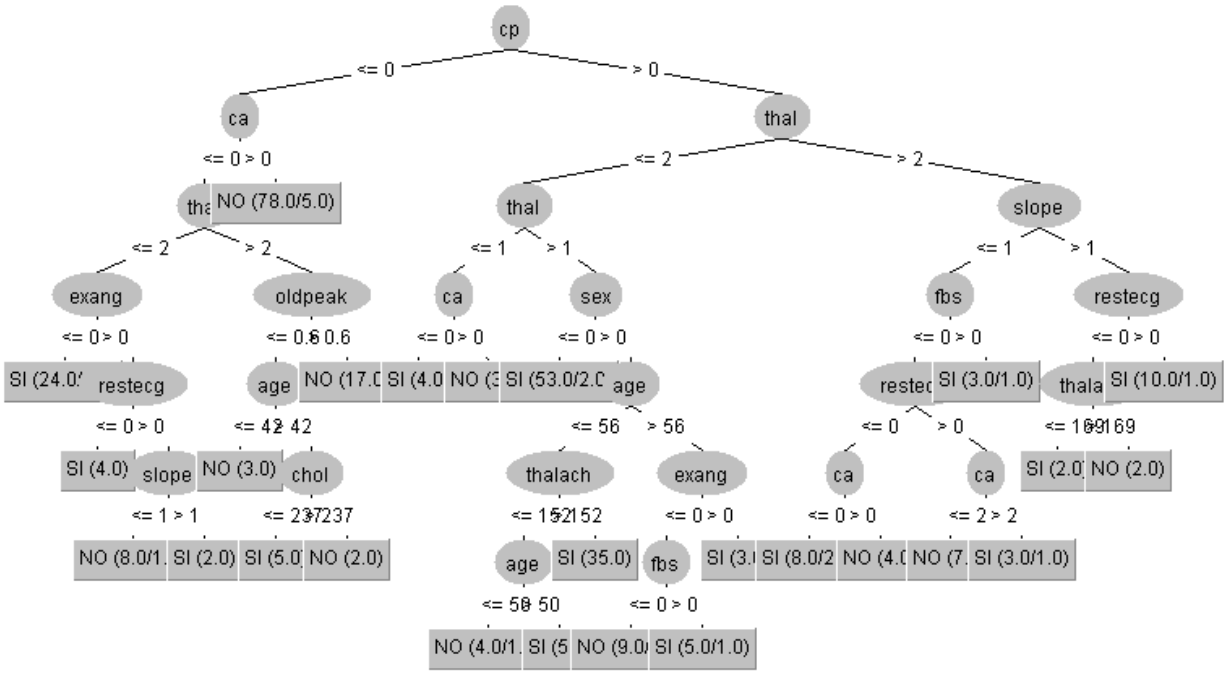
El Árbol #1 fue clasificado dentro de weka por Cross Validation aprovechando las 13 columnas de análisis NUM y 1 de resultado por NOM lo que nos mostró la salida de una exactitud baja del **78%.**

El Árbol #2 fue clasificado dentro de weka por Percentage Split de 70% aprovechando las 13 columnas NUM y 1 resultado NOM lo que aumentó la exactitud a **82%.**

El Árbol #3 fue clasificado dentro de weka por Training Split discriminando una columna de análisis, (trestbps) presión arterial en reposo, y mantuvimos de resultado por NOM lo que mejoró nuestra exactitud de gran manera presentando **93.7%.**

Resultados:

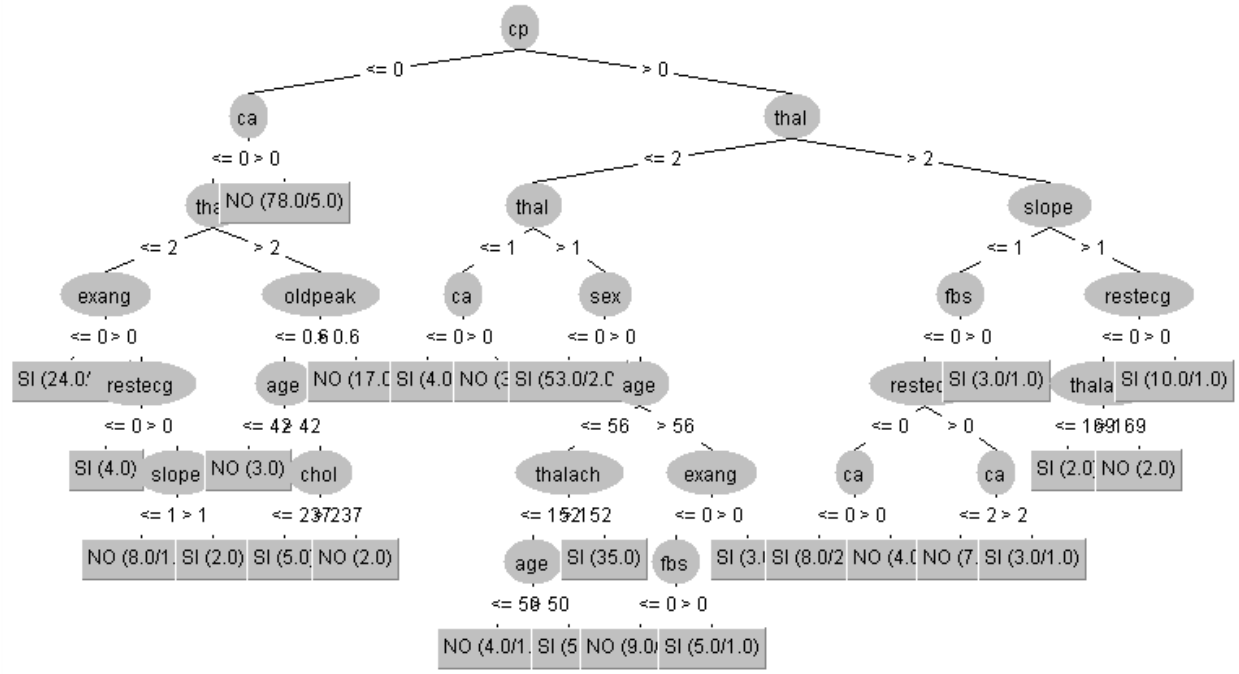
> ***Árbol 1 Cross Validation***



```
cp <= 0
|   ca <= 0
|   |   thal <= 2
|   |   |   exang <= 0: SI (24.0/2.0)
|   |   |   exang > 0
|   |   |   |   restecg <= 0: SI (4.0)
|   |   |   |   restecg > 0
|   |   |   |   |   slope <= 1: NO (8.0/1.0)
|   |   |   |   |   slope > 1: SI (2.0)
|   |   thal > 2
|   |   |   oldpeak <= 0.6
|   |   |   |   age <= 42: NO (3.0)
|   |   |   |   age > 42
|   |   |   |   |   chol <= 237: SI (5.0)
|   |   |   |   |   chol > 237: NO (2.0)
|   |   |   oldpeak > 0.6: NO (17.0)
|   ca > 0: NO (78.0/5.0)
cp > 0
|   thal <= 2
|   |   thal <= 1
|   |   |   ca <= 0: SI (4.0)
|   |   |   ca > 0: NO (3.0)
|   |   thal > 1
|   |   |   sex <= 0: SI (53.0/2.0)
|   |   |   sex > 0
|   |   |   |   age <= 56
|   |   |   |   |   thalach <= 152
|   |   |   |   |   |   age <= 50: NO (4.0/1.0)
|   |   |   |   |   |   age > 50: SI (5.0)
|   |   |   |   |   thalach > 152: SI (35.0)
|   |   |   |   age > 56
|   |   |   |   |   exang <= 0
|   |   |   |   |   |   fbs <= 0: NO (9.0/2.0)
|   |   |   |   |   |   fbs > 0: SI (5.0/1.0)
|   |   |   |   |   exang > 0: SI (3.0)
|   thal > 2
|   |   slope <= 1
|   |   |   fbs <= 0
|   |   |   |   restecg <= 0
|   |   |   |   |   ca <= 0: SI (8.0/2.0)
|   |   |   |   |   ca > 0: NO (4.0)
|   |   |   |   restecg > 0
|   |   |   |   |   ca <= 2: NO (7.0)
|   |   |   |   |   ca > 2: SI (3.0/1.0)
|   |   |   fbs > 0: SI (3.0/1.0)
|   |   slope > 1
|   |   |   restecg <= 0
|   |   |   |   thalach <= 169: SI (2.0)
|   |   |   |   thalach > 169: NO (2.0)
|   |   |   restecg > 0: SI (10.0/1.0)

Number of Leaves  :     26

Size of the tree :      51


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         238               78.5479 %
Incorrectly Classified Instances        65               21.4521 %
```

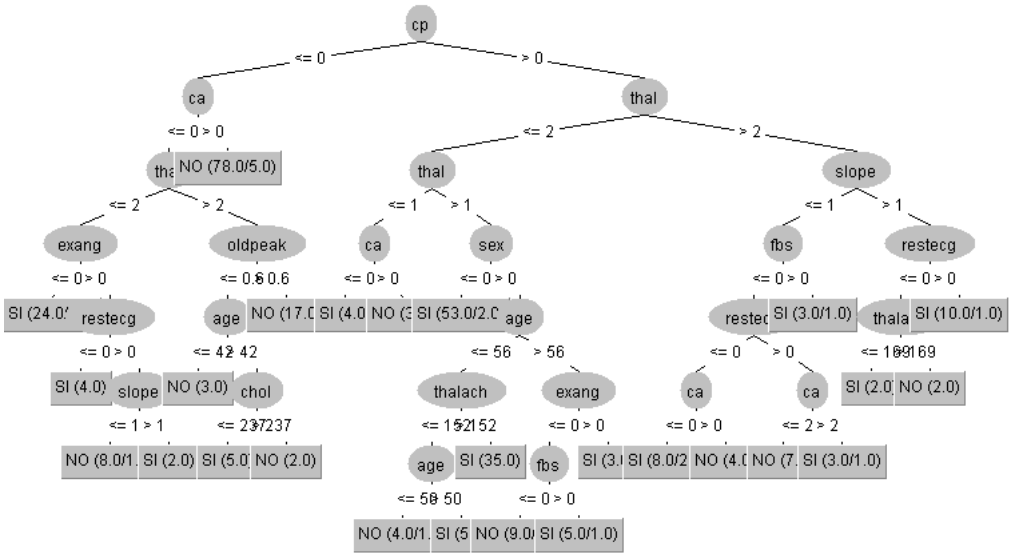
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         238               78.5479 %
Incorrectly Classified Instances        65               21.4521 %
Kappa statistic                          0.5657
Mean absolute error                      0.2494
Root mean squared error                  0.4305
Relative absolute error                 50.2792 %
Root relative squared error             86.4445 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.824    0.261    0.791     0.824    0.807      0.566    0.788     0.771     SI
                 0.739    0.176    0.779     0.739    0.758      0.566    0.788     0.720     NO
Weighted Avg.    0.785    0.222    0.785     0.785    0.785      0.566    0.788     0.748

=== Confusion Matrix ===

   a   b   <-- classified as
 136  29 |   a = SI
  36 102 |   b = NO
```

> *Árbol 2 Percentage Split 70%*



```
cp <= 0
|   ca <= 0
|   |   thal <= 2
|   |   |   exang <= 0: SI (24.0/2.0)
|   |   |   exang > 0
|   |   |   |   restecg <= 0: SI (4.0)
|   |   |   |   restecg > 0
|   |   |   |   |   slope <= 1: NO (8.0/1.0)
|   |   |   |   |   slope > 1: SI (2.0)
|   |   thal > 2
|   |   |   oldpeak <= 0.6
|   |   |   |   age <= 42: NO (3.0)
|   |   |   |   age > 42
|   |   |   |   |   chol <= 237: SI (5.0)
|   |   |   |   |   chol > 237: NO (2.0)
|   |   |   oldpeak > 0.6: NO (17.0)
|   ca > 0: NO (78.0/5.0)
cp > 0
|   thal <= 2
|   |   thal <= 1
|   |   |   ca <= 0: SI (4.0)
|   |   |   ca > 0: NO (3.0)
|   |   thal > 1
|   |   |   sex <= 0: SI (53.0/2.0)
|   |   |   sex > 0
|   |   |   |   age <= 56
|   |   |   |   |   thalach <= 152
|   |   |   |   |   |   age <= 50: NO (4.0/1.0)
|   |   |   |   |   |   age > 50: SI (5.0)
|   |   |   |   |   thalach > 152: SI (35.0)
|   |   |   |   age > 56
|   |   |   |   |   exang <= 0
|   |   |   |   |   |   fbs <= 0: NO (9.0/2.0)
|   |   |   |   |   |   fbs > 0: SI (5.0/1.0)
|   |   |   |   |   exang > 0: SI (3.0)
|   thal > 2
|   |   slope <= 1
|   |   |   fbs <= 0
|   |   |   |   restecg <= 0
|   |   |   |   |   ca <= 0: SI (8.0/2.0)
|   |   |   |   |   ca > 0: NO (4.0)
|   |   |   |   restecg > 0
|   |   |   |   |   ca <= 2: NO (7.0)
|   |   |   |   |   ca > 2: SI (3.0/1.0)
|   |   |   fbs > 0: SI (3.0/1.0)
|   |   slope > 1
|   |   |   restecg <= 0
|   |   |   |   thalach <= 169: SI (2.0)
|   |   |   |   thalach > 169: NO (2.0)
|   |   |   restecg > 0: SI (10.0/1.0)

Number of Leaves  :     26

Size of the tree :      51


Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          75               82.4176 %
Incorrectly Classified Instances        16               17.5824 %

 === Confusion Matrix ===

   a  b   <-- classified as
  37  7 |  a = SI
   9 38 |  b = NO
```

> *Árbol 3 Training Split 12 Columnas*




```
cp <= 0
|  ca <= 0
|  |  thal <= 2
|  |  |  exang <= 0: SI (24.0/2.0)
|  |  |  exang > 0
|  |  |  |  restecg <= 0: SI (4.0)
|  |  |  |  restecg > 0
|  |  |  |  |  slope <= 1: NO (8.0/1.0)
|  |  |  |  |  slope > 1: SI (2.0)
|  |  thal > 2
|  |  |  oldpeak <= 0.6
|  |  |  |  age <= 42: NO (3.0)
|  |  |  |  age > 42
|  |  |  |  |  chol <= 237: SI (5.0)
|  |  |  |  |  chol > 237: NO (2.0)
|  |  |  oldpeak > 0.6: NO (17.0)
|  ca > 0: NO (78.0/5.0)
cp > 0
|  thal <= 2
|  |  thal <= 1
|  |  |  ca <= 0: SI (4.0)
|  |  |  ca > 0: NO (3.0)
|  |  thal > 1
|  |  |  sex <= 0: SI (53.0/2.0)
|  |  |  sex > 0
|  |  |  |  age <= 56
|  |  |  |  |  thalach <= 152
|  |  |  |  |  |  age <= 50: NO (4.0/1.0)
|  |  |  |  |  |  age > 50: SI (5.0)
|  |  |  |  |  thalach > 152: SI (35.0)
|  |  |  |  age > 56
|  |  |  |  |  exang <= 0
|  |  |  |  |  |  fbs <= 0: NO (9.0/2.0)
|  |  |  |  |  |  fbs > 0: SI (5.0/1.0)
|  |  |  |  |  exang > 0: SI (3.0)
|  thal > 2
|  |  slope <= 1
|  |  |  fbs <= 0
|  |  |  |  restecg <= 0
|  |  |  |  |  ca <= 0: SI (8.0/2.0)
|  |  |  |  |  ca > 0: NO (4.0)
|  |  |  |  restecg > 0
|  |  |  |  |  ca <= 2: NO (7.0)
|  |  |  |  |  ca > 2: SI (3.0/1.0)
|  |  |  fbs > 0: SI (3.0/1.0)
|  |  slope > 1
|  |  |  restecg <= 0
|  |  |  |  thalach <= 169: SI (2.0)
|  |  |  |  thalach > 169: NO (2.0)
|  |  |  restecg > 0: SI (10.0/1.0)
```

```
Number of Leaves :      26

Size of the tree :      51


Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         284               93.7294 %
Incorrectly Classified Instances        19                6.2706 %


=== Confusion Matrix ===

   a   b   <-- classified as
 156   9 |   a = SI
  10 128 |   b = NO
```

Conclusiones

La herramienta Weka nos permite analizar rápidamente y de manera eficiente y versátil la información obtenida por un análisis de población considerable registrada en nuestra base de datos, presentándonos un resultado detallado de su validación y representación gráfica a través de los árboles de decisiones, en esta actividad implementamos el uso de esta plataforma imprimiendo 3 árboles de decisiones regidos por diferentes métodos de análisis definidos y controlados por nosotros el usuario, obteniendo diferente nivel de exactitud y precisión para el resultado de nuestra validación. Es una gran herramienta, no tuve dificultades con el uso del software.

Referencias

I. Heart Disease Data Set.
   1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D
   2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
   3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
   4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

II. Zambrano, E. (2021) Presentación, Inteligencia Artificial y Redes Neuronales, IB.

III. Ambriz, JP. (2021) Actividad 5, Inteligencia Artificial y Redes Neuronales, IB. GitHUB. https://github.com/AmbrizPapo/Inteligencia-Artificial-y-Redes-Neuronales-1864353/tree/main/LasActividadesVanAqui