Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Inteligencia Artificial y Redes Neuronales

Jueves N4

Actividad 5. – Árbol de decisión en Weka

Tania Elizabeth Frías López

Matrícula: 1843664

Carrera: Ingeniero Biomédico

Docente: Daniel Isaías López Páez

Semestre Agosto– Diciembre 2021

San Nicolás de los Garza, N.L. 14/11/2021

OBJETIVO

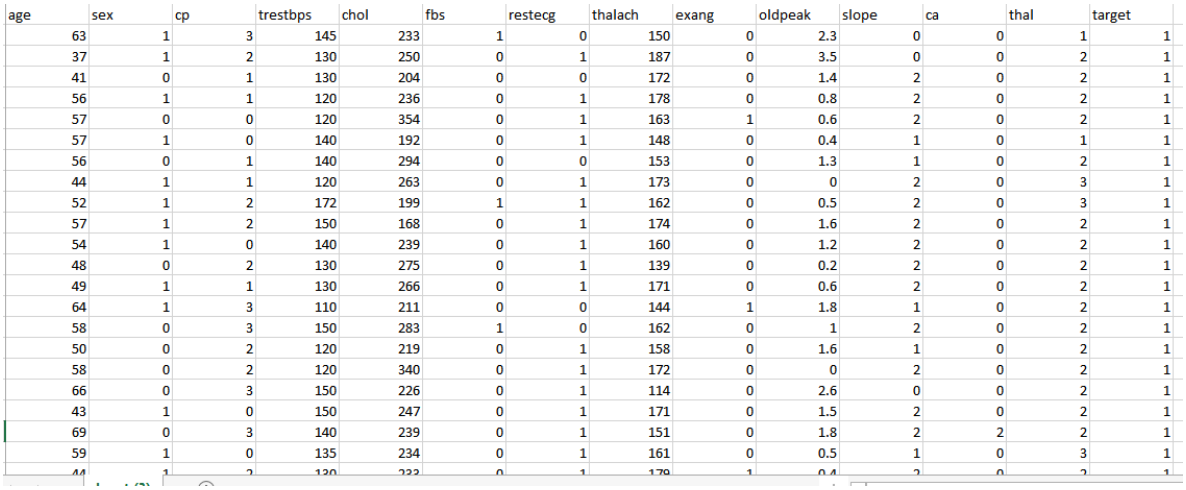
Entrenar un árbol de decisión en Weka con el dataset "Heart Disease UCI", el cual deberá de poder clasificar si existe o no existe alguna enfermedad del corazón.

ANTECEDENTES

La información para realizar el entrenamiento del árbol Weka se basó en la siguiente dataset:

https://www.kaggle.com/ronitf/heart-disease-uci

En la siguiente imagen podemos visualizar la tabla que nos muestra el documento con formato csv de manera general:



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1

En esta dataset encontramos 303 registros, es decir, los datos que se muestran en cada columna son provenientes de 303 personas.

En esa tabla se utilizaron 14 atributos:

1. age
2. sex
3. cp
4. trestbps
5. chol
6. fbs
7. restecg
8. thalach
9. exang
10. oldpeak

11. slope
12. ca
13. thal
14. (num) (the predicted attribute)

DESCRIPCIÓN

Podemos encontrar 14 columnas, las cuales, cada una de ellas nos brinda un dato específico, siendo estos datos los siguientes:

1) Age: Edad de la persona
2) Sex: Sexo de la persona
3) Cp: Tipo de dolor en el pecho
   Esta se clasifica en 4 valores, siendo cada uno lo siguiente:
   - Valor 0: Angina típica
   - Valor 1: Angina atípica
   - Valor 2: Tipo de dolor no anginoso
   - Valor 3: Asintomático - Sin dolor
4) Trestbps: Presión arterial en estado de reposo. Esta se expresa en mmHg y se refiere al nivel de la presión arterial al momento de ingresar al hospital.
5) Chol: Nivel de colesterol sérico en mg/dl
6) Fbs: Nivel de azúcar en ayunas
7) Restecg: Resultados del electrocardiograma cuando la persona está en reposo.
   Esta característica se clasifica en 3 valores:
   - Valor 0: Resultados normales
   - Valor 1: La persona muestra un segmento ST-T anormal (inversiones de la onda T y / o elevación o depresión del ST> 0,05 mV)
   - Valor 2: El electrocardiograma muestra una posible o definitiva hipertrofia ventricular izquierda
8) Thalach: Valor máximo de frecuencia cardíaca alcanzado
9) Exang: Dolor de angina originada por hacer ejercicio (1 = SI, 0 = NO)
10) Oldpeak: Depresión del segmento ST inducido por el ejercicio en relación con a la posición en reposo
11) Slope: Pendiente del segmento ST durante ejercicio
    Esta característica se clasifica en 3 valores:
    - Valor 1: Ascendente
    - Valor 2: Plana
    - Valor 3: Descendente
12) Ca: Número de vasos sanguíneos principales (0-3) coloreados por la floración
13) Thal: Esta característica se representa de la siguiente manera:
    - 1 = Normal
    - 2 = Defecto fijo

- 3 = Defecto reversible

14) Target: Atributo predicho

METODOLOGÍA

Entrenamiento Árbol 1

Para este primer entrenamiento se eliminaron los siguientes atributos:

- Age: Edad de la persona
- Trestbps: Presión arterial en estado de reposo. Esta se expresa en mmHg y se refiere al nivel de la presión arterial al momento de ingresar al hospital
- Chol: Nivel de colesterol sérico en mg/dl
- Thalach: Valor máximo de frecuencia cardíaca alcanzado
- Oldpeak: Depresión del segmento ST inducido por el ejercicio en relación con a la posición en reposo

El criterio que se utilizó para realizar el entrenamiento de este primer árbol fue en eliminar las columnas en las que se almacenaran mayor cantidad de datos distintos.

Por ejemplo, en la edad podemos encontrar valores desde 29 a 77; en la columna para ***trestbps*** hay valores entre 94 y 200; en el caso del atributo ***chol***, aparecen valores desde 126 a 564, siendo este el atributo en donde hay mayor cantidad de valores distintos; para la columna ***thalach*** los valores van desde 71 a 202; y por ultimo para el atributo ***oldpeak***, encontramos valores que van desde 0 a 6.2.

Además para este entrenamiento se decidió utilizar 20 iteraciones, ya que nos mostraba un mayor valor para el porcentaje de exactitud.

Para este entrenamiento se requirió cambiar los valores de la columna target, para que en lugar de brindarnos información numérica nos mostrara información de tipo carácter.

Los resultados que se obtuvieron fueron los siguientes:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     heart (2)
Instances:    303
Attributes:   9
              ï»¿sex
              cp
              fbs
              restecg
              exang
              slope
              ca
              thal
              target
Test mode:    20-fold cross-validation


=== Classifier model (full training set) ===

J48 pruned tree
------------------

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: NO (27.0/5.0)
|   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: SI (114.0/13.0)
|   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
|   |   |   |   ca <= 0: SI (2.0)
|   |   |   |   ca > 0: NO (2.0)
|   |   |   restecg > 0: SI (10.0/1.0)

Number of Leaves  :     17

Size of the tree :      33


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

Correctly Classified Instances         253               83.4983 %
Incorrectly Classified Instances        50               16.5017 %
Kappa statistic                          0.6657
Mean absolute error                      0.2335
Root mean squared error                  0.3756
Relative absolute error                 47.0695 %
Root relative squared error             75.4047 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.873     0.210     0.832       0.873   0.852      0.667    0.833     0.787     SI
               0.790     0.127     0.838       0.790   0.813      0.667    0.833     0.804     NO
Weighted Avg.  0.835     0.172     0.835       0.835   0.834      0.667    0.833     0.795

=== Confusion Matrix ===

   a   b   <-- classified as
 144  21 |   a = SI
  29 109 |   b = NO
```

Entrenamiento Árbol 2

Para este entrenamiento se eliminaron los siguientes atributos:

- Trestbps: Presión arterial en estado de reposo. Esta se expresa en mmHg y se refiere al nivel de la presión arterial al momento de ingresar al hospital
- Thalach: Valor máximo de frecuencia cardíaca alcanzado
- Oldpeak: Depresión del segmento ST inducido por el ejercicio en relación con a la posición en reposo

El criterio que se utilizó para este entrenamiento consistió en eliminar las columnas que tuvieran cierta relación con otra, por ejemplo, se eliminó la columna de ***trestbps***, la cual como se mencionó anteriormente nos muestra los niveles de la presión arterial, por lo tanto, se decidió dejar la columna del atributo ***chol***, ya que esa nos muestra los niveles de colesterol. También se eliminaron las columnas t***halach*** y ***oldpeak***, ya que el atributo ***restecg*** podría darnos una idea de esos valores, ya que ese atributo nos muestra los resultados del electrocardiograma, además de que también el atributo ***slope*** podría ayudarnos con más datos importantes.

Además para este entrenamiento se decidió utilizar 20 iteraciones, ya que nos mostraba un mayor valor para el porcentaje de exactitud.

De igual manera en este entrenamiento se requirió cambiar los valores de la columna target, para que en lugar de brindarnos información numérica nos mostrara información de tipo carácter.

Los resultados que se obtuvieron fueron los siguientes:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     ARBOL3
Instances:    303
Attributes:   11
              ï»¿age
              sex
              cp
              chol
              fbs
              restecg
              exang
              slope
              ca
              thal
              target
Test mode:    20-fold cross-validation

=== Classifier model (full training set) ===
```

```
J48 pruned tree
------------------

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
|   |   |   slope <= 1: NO (17.0/1.0)
|   |   |   slope > 1
|   |   |   |   chol <= 237
|   |   |   |   |   ï»¿age <= 42: NO (2.0)
|   |   |   |   |   ï»¿age > 42: SI (4.0)
|   |   |   |   chol > 237: NO (4.0)
|   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
|   |   |   |   ï»¿age <= 56: SI (44.0/3.0)
|   |   |   |   ï»¿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
|   |   |   |   ca <= 0: SI (2.0)
|   |   |   |   ca > 0: NO (2.0)
|   |   |   restecg > 0: SI (10.0/1.0)

Number of Leaves  :     24

Size of the tree :      47


Time taken to build model: 0 seconds


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

Correctly Classified Instances         247               81.5182 %
Incorrectly Classified Instances        56               18.4818 %
Kappa statistic                          0.6256
Mean absolute error                      0.2375
Root mean squared error                  0.396
Relative absolute error                 47.8754 %
Root relative squared error             79.5089 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.855    0.232    0.815      0.855    0.834      0.627    0.821     0.779     SI
                 0.768    0.145    0.815      0.768    0.791      0.627    0.821     0.770     NO
Weighted Avg.    0.815    0.193    0.815      0.815    0.815      0.627    0.821     0.775

=== Confusion Matrix ===

   a   b   <-- classified as
 141  24 |   a = SI
  32 106 |   b = NO
```

Entrenamiento Árbol 3

Para este entrenamiento se eliminaron los siguientes atributos:

- Fbs: Nivel de azúcar en ayunas
- Oldpeak: Depresión del segmento ST inducido por el ejercicio en relación con a la posición en reposo
- Slope: Pendiente del segmento ST durante ejercicio

El criterio que se utilizó para este entrenamiento fue eliminar el atributo **fbs**, ya que sólo 45 personas que fueron registradas, presentaban un nivel más alto de lo normal de azúcar en ayunas. De igual manera se decidió eliminar los atributos ***slope*** y ***oldpeak***, ya que el atributo ***restecg*** podría darnos una idea de esos valores, porque es el atributo que nos muestra si los resultados del electrocardiograma son normales o no.

De igual manera para este caso se intentó eliminar algunos registros, pero al momento de realizarlo se notó que el porcentaje de exactitud en lugar de incrementarse se disminuía, por lo tanto, se decidió realizar el entrenamiento con los 303 registros.

Además para este entrenamiento se decidió utilizar 20 iteraciones, ya que nos mostraba un mayor valor para el porcentaje de exactitud.

De igual manera en este entrenamiento se requirió cambiar los valores de la columna target, para que en lugar de brindarnos información numérica nos mostrara información de tipo carácter.

Los resultados que se obtuvieron fueron los siguientes:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     heart (4)
Instances:    303
Attributes:   11
              ï»¿age
              sex
              cp
              trestbps
              chol
              restecg
              thalach
              exang
              ca
              thal
              target
Test mode:    20-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------
```

```
cp <= 0
|   ca <= 0
|   |   thal <= 2
|   |   |   exang <= 0: SI (24.0/2.0)
|   |   |   exang > 0
|   |   |   |   restecg <= 0: SI (4.0)
|   |   |   |   restecg > 0
|   |   |   |   |   ï»¿age <= 56: NO (4.0)
|   |   |   |   |   ï»¿age > 56
|   |   |   |   |   |   ï»¿age <= 58: SI (2.0)
|   |   |   |   |   |   ï»¿age > 58: NO (4.0/1.0)
|   |   thal > 2
|   |   |   restecg <= 0: NO (15.0/1.0)
|   |   |   restecg > 0
|   |   |   |   exang <= 0
|   |   |   |   |   ï»¿age <= 41: NO (2.0)
|   |   |   |   |   ï»¿age > 41: SI (3.0)
|   |   |   |   exang > 0: NO (7.0/1.0)
|   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
|   |   |   |   ï»¿age <= 56
|   |   |   |   |   thalach <= 152
|   |   |   |   |   |   ï»¿age <= 50: NO (4.0/1.0)
|   |   |   |   |   |   ï»¿age > 50: SI (5.0)
|   |   |   |   |   thalach > 152: SI (35.0)
|   |   |   |   ï»¿age > 56
|   |   |   |   |   ca <= 1
|   |   |   |   |   |   chol <= 259: SI (11.0/2.0)
|   |   |   |   |   |   chol > 259: NO (4.0)
|   |   |   |   |   ca > 1: NO (2.0)
```

```
|   thal > 2
|   |   sex <= 0: SI (3.0/1.0)
|   |   sex > 0
|   |   |   restecg <= 0
|   |   |   |   ca <= 0
|   |   |   |   |   ï»¿age <= 65: SI (8.0/1.0)
|   |   |   |   |   ï»¿age > 65: NO (2.0)
|   |   |   |   ca > 0: NO (6.0)
|   |   |   restecg > 0
|   |   |   |   thalach <= 142: NO (5.0)
|   |   |   |   thalach > 142
|   |   |   |   |   ï»¿age <= 50
|   |   |   |   |   |   ï»¿age <= 47: SI (5.0/1.0)
|   |   |   |   |   |   ï»¿age > 47: NO (2.0)
|   |   |   |   |   ï»¿age > 50: SI (8.0)

Number of Leaves  :     27

Size of the tree :      53


Time taken to build model: 0 seconds

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

Correctly Classified Instances         249               82.1782 %
Incorrectly Classified Instances        54               17.8218 %
Kappa statistic                          0.6386
Mean absolute error                      0.2306
Root mean squared error                  0.4019
Relative absolute error                 46.4851 %
Root relative squared error             80.7    %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.867     0.232     0.817      0.867     0.841      0.640     0.802     0.779     SI
               0.768     0.133     0.828      0.768     0.797      0.640     0.802     0.722     NO
Weighted Avg.  0.822     0.187     0.822      0.822     0.821      0.640     0.802     0.753

=== Confusion Matrix ===

   a   b   <-- classified as
 143  22 |   a = SI
  32 106 |   b = NO
```

CONCLUSIÓN

Con esta actividad se pudo comprender en una mejor manera lo que se visualizó en la clase de IA, ya que como requisito se debían de entrenar tres árboles distintos con características diferentes cada uno.

Me pareció un tema muy interesante y que podría ayudar en gran medida a la medicina, para poder analizar estudios de una manera estadística y de esa manera poder sacar una conclusión. Por ejemplo, en la rama de la bioestadística se realizan análisis de distintos tipos con el propósito de conocer el comportamiento de una población o muestra, y ya con los resultados que se obtengan poder implementar una estrategia que ayude cubrir una posible necesidad.

Al realizar los distintos entrenamientos se hicieron varios intentos para aumentar el porcentaje de exactitud. En cada intento se eliminaban atributos de acuerdo a un criterio, en donde se pudo ver que al eliminar ciertos atributos el porcentaje de exactitud se aumentaba o disminuía. En un intento se había eliminado datos, basándose en la cantidad de registros que había de una cierta edad, en donde si una edad contaba con registros menores o igual a 3, esos datos se eliminaban, pero al obtener el porcentaje de exactitud se visualizó que entre más registros y atributos se eliminaban, menor sería el porcentaje de exactitud que se obtendría, por lo tanto, era necesario tener un criterio claro en cada entrenamiento y lograr obtener un porcentaje aceptable.

BIBLIOGRAFÍA

Janosi,, A., Steinbrunn, W., Pfisterer,, M., & Detrano, R. (s.f.). *UCI Machine Learning Repository*. Obtenido de Heart Disease Data Set: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

kaggle. (s.f.). Obtenido de Heart Disease UCI: https://www.kaggle.com/ronitf/heart-disease-uci/version/1