Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Inteligencia Artificial

Actividad 5

Nombre: José Emanuel Martínez Rodríguez Matricula: 1851368

Carrera: Ingeniero Biomédico

Monterrey, Nuevo León a 10 de Noviembre del 2021

Antecedentes

Esta base de datos contiene 76 atributos, pero para 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 "objetivo" se refiere a la presencia de enfermedad cardíaca en el paciente. Tiene un valor entero de 0 (sin presencia) a 4. Los experimentos con la base de datos de Cleveland se han concentrado en simplemente intentar distinguir la presencia (valores 1, 2, 3, 4) de la ausencia (valor 0).

Los nombres y números de seguro social de los pacientes se eliminaron recientemente de la base de datos y se reemplazaron con valores ficticios.

Descripción de los atributos

De los 76 atributos que se mencionan en la parte anterior, para el entrenamiento del árbol de decisiones se utilizaron 14, y en esta parte se definirá a que se refiere cada uno de los atributos que se seleccionaron para la realización del árbol de decisiones:

- Age=Edad en años
- Sex= Sexo de la persona
- Cp= Tipo de dolor de pecho
- Trestbps= Presión arterial en reposo (mmHg)
- Chol= Colesterol sérico (mg/dl)
- Fbs= Glucemia en sangre > 120mg/dl)
- Restiecg= Resultados del electrocardiograma en reposo
- Thatlach= Pulso cardiaco máximo archivado
- Exang= Angina inducida por ejercicio
- Oldpeak= Depresión del ST por ejercicio en relación con el reposo
- Slope= Pendiente del segmento ST en el ejercicio pico
- Ca= Numero de vasos principales coloreados por fluoroscopia
- Thal= Existencia de algún defecto
- Target= Presenta el padecimiento

Metodología y Resultados

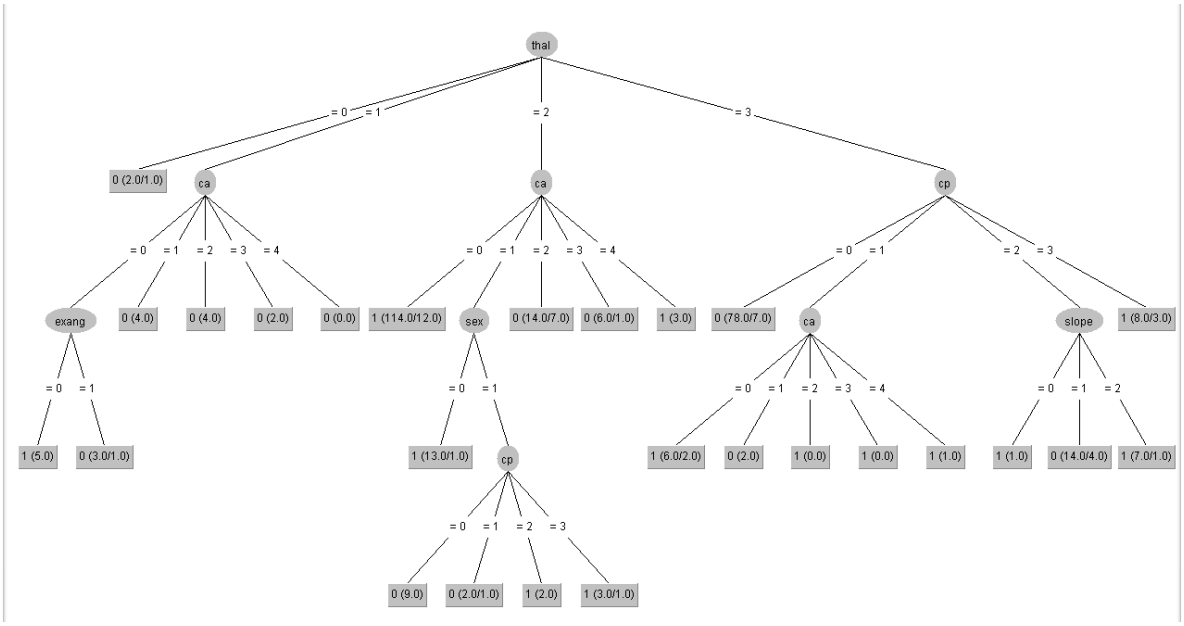
Modelo elegido para la creación del árbol de decisiones es el J48, el cual es una implementación del algoritmo C4.5, uno de los algoritmos de minería de datos más utilizado. Se trata de un refinamiento del modelo generado con OneR. Supone una mejora moderada en las prestaciones, y podría conseguir una probabilidad de acierto ligeramente superior al del anterior clasificador.

Árbol de decisiones 1

Modo de testeo: 10-folds validación-cruzada

En que consiste la validación cruzada, pues consiste en una evaluación. Donde se dividirán las instancias en tantas carpetas como indica el parámetro “Folds”, y en cada evaluación se toman las instancias de cada carpeta como datos de test, y el resto como datos de entrenamiento para construir el modelo. Los errores calculados serán el promedio de todas las ejecuciones

Valor de exactitud del entrenamiento 75.9076%



Matriz de confusión:

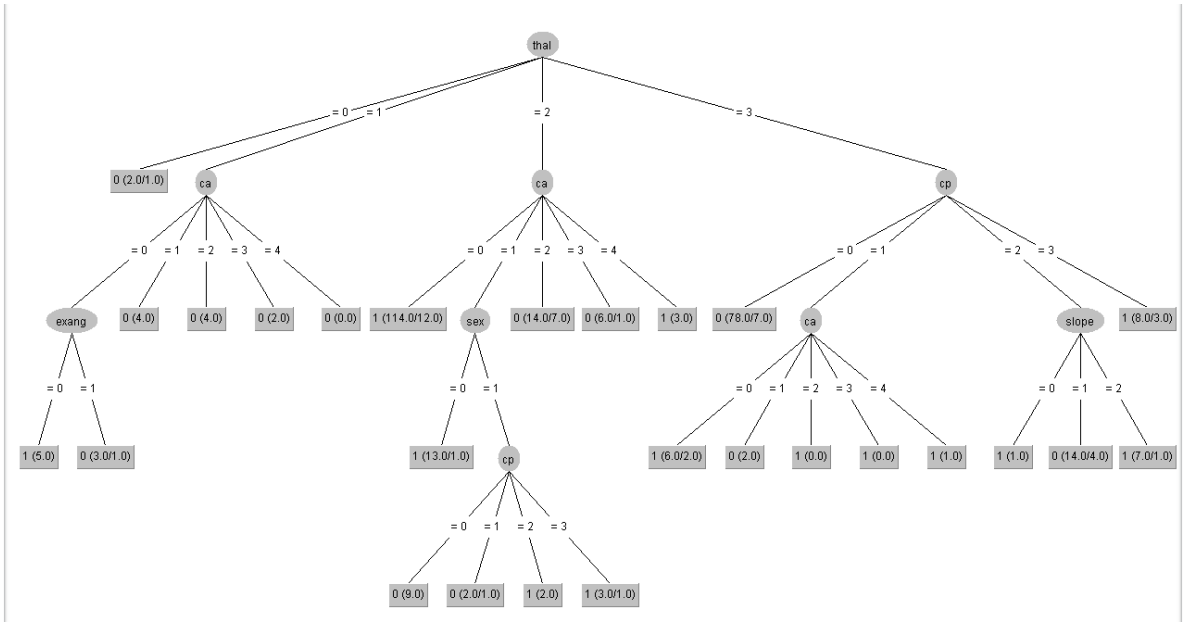
A	B		
104	34	|	A=0
39	126	|	B=1

Árbol de decisiones 2

Modo de testeo: evaluate on training data

Este modo de testeo se trata de una evaluación del clasificador sobre el mismo conjunto sobre el que se construye el modelo predictivo para determinar el error, que en este caso se denomina “error de resustitucion”.

Valor de exactitud del entrenamiento 86.1386%



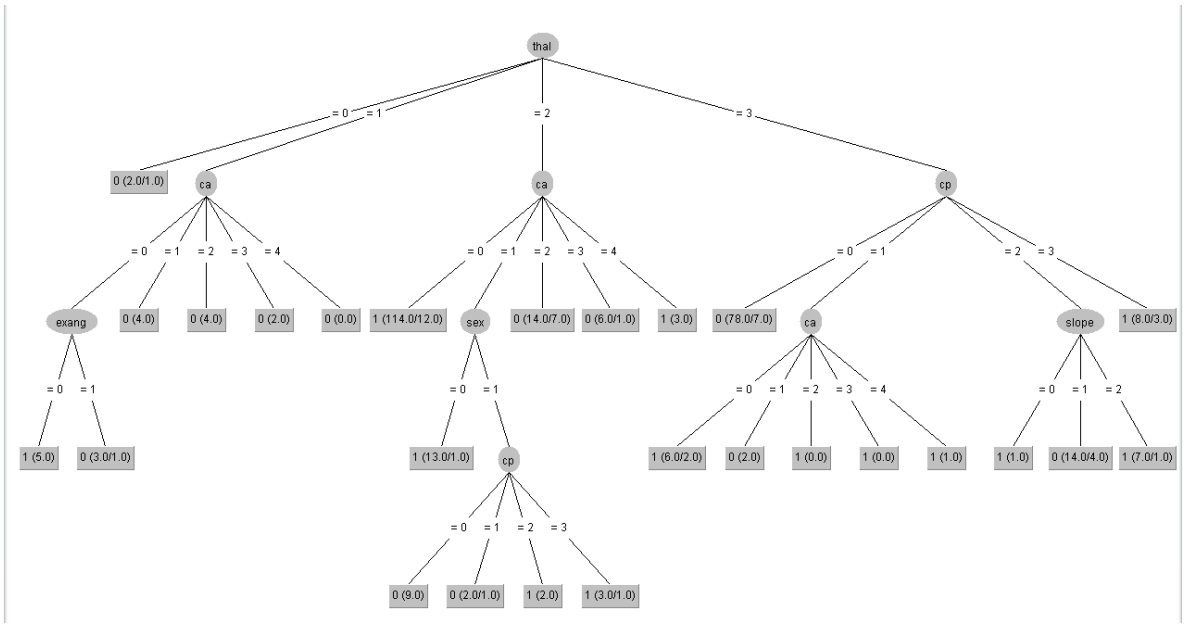
Matriz de confusión:

A	B		
118	20	|	A=0
22	143	|	B=1

Árbol de decisiones 3

Modo de testeo: 25-folds validación-cruzada

Valor de exactitud del entrenamiento 73.5974%



Matriz de confusión:

A	B		
101	37	|	A=0
43	122	|	B=1

Conclusión:

Con lo visto por los porcentajes del valor de entrenamiento la que tuvo mayor porcentaje fue la de “evaluate on training data” con un 86% aproximadamente, mientras que la validación cruzada con 10 folds mostro un porcentaje de 755 aprox, por ultimo la de 25 folds resulto con un porcentaje aprox de 73%, además se experimento con mas valores de “folds” y se obtuvo que el mayor porcentaje de valor de entrenamiento ronda los 77% mientras que el menor fue de 73% aprox.

También se observo que los arboles de decisiones son prácticamente iguales, lo cual me pareció algo interesante, posiblemente si se hubiera descartado alguna de las columnas de los atributos, el árbol de decisiones hubiera cambiado y el porcentaje de entrenamiento hubiera sido mayor a lo hemos visto.

Bibliografía:

- García, M., & Alvarez, A. (s. f.). Análisis de Datos en WEKA – Pruebas de Selectividad. Unknown. Recuperado 10 de noviembre de 2021, de http://www.it.uc3m.es/~jvillena/irc/practicas/06-07/28.pdf
- Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (2018, 25 junio). Heart Disease UCI. Kaggle. Recuperado 10 de noviembre de 2021, de https://www.kaggle.com/ronitf/heart-disease-uci