PROYECTO - MINERÍA DE DATOS

->Flores González Luis Brandon

->Santaella Marín Héctor

METODOLOGÍA CRISP-DM

- **Comprensión del negocio**
- **Comprensión de los datos**
- **Preparación de los datos**
- **Modelado**
- **Evaluación**
- **Despliegue**

Comprensión del negocio

Este trabajo se basa en una contratación con nuestros conocimientos de Minería de datos.

Los interesados quieren crear un programa dentro del hospital que trate a las personas que padecen enfermedades cardíacas a partir de diversos factores .

Objetivos del negocio

“Objetivos y requerimientos desde una perspectiva no técnica”

Determinar qué factores contribuyen a que una persona padezca problemas cardíacos.

Se considerará un éxito si se tienen claros los factores que generan problemas cardiacos.

Evaluación de la situación

Contamos con un conjunto de datos proporcionada por el cliente en donde se contiene valores importantes para cumplir el objetivo del negocio.

El conjunto de datos describe los contenidos del directorio de enfermedades cardíacas.

Comprensión de datos

Objetivo

- Descripción de los datos.
- Identificar problemas con la calidad de los datos.
- Formar hipótesis de los datos.

Diccionario de datos

Atributo	Significado	Valores
age	edad en años.	0, …., n
sex	sexo	● male(hombre): es representado con 1. ● female(mujer): es representado con 0.
cp	tipo de dolor en el pecho	● asympt: asymptomatic(asintomático) ● non_anginal: non-anginal pain(dolor no anginal) ● atyp_angina: atypical angina(angina atípica) ● typ_angina: typical angina(angina típica)
trestbps	presión arterial en reposo	en mm Hg(presión manométrica) al ingreso en el hospital
chol	colesterol sérico	en mg / dl (miligramos por decilitro)
fbs	azúcar en sangre en ayunas) > 120 mg/dl	1 = true; 0 = false
restecg	resultados electrocardiográficos en reposo	● normal ● left_vent_hyper: mostrando hipertrofia ventricular izquierda probable ● st_t_wave_abnormality: tener anormalidad de la onda ST

Diccionario de datos

Atributo	Significado	Valores
thalach	Frecuencia cardíaca máxima lograda	-
exang	Angina inducida por el ejercicio	1 = yes; 0 = no
oldpeak	Depresión ST inducida por el ejercicio en relación con el reposo	-
slope	La pendiente del segmento ST de ejercicio pico	● up: upsloping ● flat: flat(plano) ● down: downsloping
ca	número de vasos principales (0-3) coloreados por fluoroscopia	-
thal	-	3 = normal 6 = fixed defect 7 = reversable defect
num	diagnóstico de enfermedad cardíaca, estado angiográfico de la enfermedad	● <50: < 50% de diámetro de estrechamiento) ● >50_1: > 50% de diámetro de estrechamiento).

Diccionario de datos

Atributo	Tipo	%valores perdidos	Mínimo	Máximo	Media	Desviación estándar	Valores únicos	valores atípicos
age	numérico	0	29	77	54.36	9.082	Si	No
sex	nominal	0	-	-	-	-	No	No
cp	nominal	0	-	-	-	-	No	No
trestbps	numérico	0	94	200	131.624	17.538	Si	No
chol	numérico	0	126	564	246.264	51.831	Si	No
fbs	nominal	0	-	-	-	-	No	No
restecg	nominal	0	-	-	-	-	No	No

Diccionario de datos

Atributo	Tipo	%valores perdidos	Mínimo	Máximo	Media	Desviación estándar	Valores únicos	valores atípicos
thalach	numérico	0	71	202	149.647	22.905	Si	No
exang	nominal	0	-	-	-	-	No	No
oldpeak	numérico	0	0	6.2	1.04	1.161	Si	No
slope	nominal	0	-	-	-	-	No	No
ca	numérico	1.65	0	3	0.674	0.938	No	Si
thal	nominal	0.66	-	-	-	-	No	Si
num	nominal	0	-	-	-	-	No	No

Matriz de correlación

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
age	1.00	0.10	-0.10	0.28	0.21	0.12	0.17	-0.40	0.10	0.21	0.17	0.36	0.13	0.23
sex	0.10	1.00	0.01	0.06	0.20	-0.05	0.01	0.04	-0.14	-0.10	-0.03	-0.09	-0.38	-0.28
cp	-0.10	0.01	1.00	0.04	-0.07	0.04	-0.08	0.34	-0.39	-0.21	-0.16	-0.23	-0.27	-0.41
trestbps	0.28	0.06	0.04	1.00	0.12	0.18	0.15	-0.05	0.07	0.19	0.12	0.10	0.14	0.14
chol	0.21	0.20	-0.07	0.12	1.00	0.01	0.17	-0.01	0.07	0.05	0.00	0.12	0.03	0.09
fbs	0.12	-0.05	0.04	0.18	0.01	1.00	0.05	-0.01	0.03	0.01	0.06	0.14	0.06	0.03
restecg	0.17	0.01	-0.08	0.15	0.17	0.05	1.00	-0.12	0.10	0.17	0.17	0.14	0.03	0.18

Matriz de correlación

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
thalach	-0.40	0.04	0.34	-0.05	-0.01	-0.01	-0.12	1.00	-0.38	-0.34	-0.39	-0.26	-0.28	-0.42
exang	0.10	-0.14	-0.39	0.07	0.07	0.03	0.10	-0.38	1.00	0.29	0.26	0.14	0.33	0.44
oldpeak	0.21	-0.10	-0.21	0.19	0.05	0.01	0.17	-0.34	0.29	1.00	0.58	0.29	0.34	0.43
slope	0.17	-0.03	-0.16	0.12	0.00	0.06	0.17	-0.39	0.26	0.58	1.00	0.11	0.29	0.35
ca	0.36	-0.09	-0.23	0.10	0.12	0.14	0.14	-0.26	0.14	0.29	0.11	1.00	0.26	0.46
thal	0.13	-0.38	-0.27	0.14	0.03	0.06	0.03	-0.28	0.33	0.34	0.29	0.26	1.00	0.53
num	0.23	-0.28	-0.41	0.14	0.09	0.03	0.18	-0.42	0.44	0.43	0.35	0.46	0.53	1.00

Preparación de datos

- Selección de los datos.
-
- Limpieza de datos.
  - Reemplaza los valores perdidos.
  - Regresión lineal para estimar los valores perdidos.
  - Eliminado de atípicos.
-
- Construcción de datos
-
- Integración de datos.

Construcción del modelo

Para tratar con los datos obtenidos, se eligieron:

- R
- RapidMiner
- Weka
- DataCleaner

Modelado

Aplicar las técnicas de minería de datos a los dataset

- Selección de la técnica de modelado
- Diseño de la evaluación
- Construcción del modelo
- Evaluación del modelo

Modelado

Las Tareas de clasificación se realizaron con los siguientes clasificadores.

OneR
RIPPER
Árbol de D. C4.5
Red neuronal

OneR

Para la mayoria de los dataset obtuvimos la siguiente regla del modelo de clasificación:

```
thal:
        fixed_defect     -> >50_1
        normal   -> <50
        reversable_defect        -> >50_1
        ?        -> <50
```

A excepción del heart-c2 y heart-c4 ya que no se contenían datos anómalos y quedo de esta forma:

```
thal:
        fixed_defect     -> >50_1
        normal   -> <50
        reversable_defect        -> >50_1
        ?        -> <50
```

Ripper

Para heart obtuvimos la siguiente regla del modelo de clasificación:

```
(cp = asympt) and (thal = reversable_defect) => num=>50_1 (78.0/7.0)
(ca >= 1) and (slope = flat) => num=>50_1 (34.0/5.0)
(ca >= 1) and (cp = asympt) => num=>50_1 (11.0/2.0)
 => num=<50 (180.0/29.0)
```

Para heart-c1 obtuvimos la siguiente regla del modelo de clasificación:

```
(ca >= 1) and (oldpeak >= 0.6) => num=>50_1 (82.0/10.0)
(thal = reversable_defect) => num=>50_1 (67.0/26.0)
 => num=<50 (154.0/25.0)
```

Ripper

Para heart-c2 obtuvimos la siguiente regla del modelo de clasificación:

```
(ca >= 1) and (oldpeak >= 0.6) => num=>50_1 (82.0/10.0)
(thal = reversable_defect) => num=>50_1 (67.0/26.0)
 => num=<50 (154.0/25.0)
```

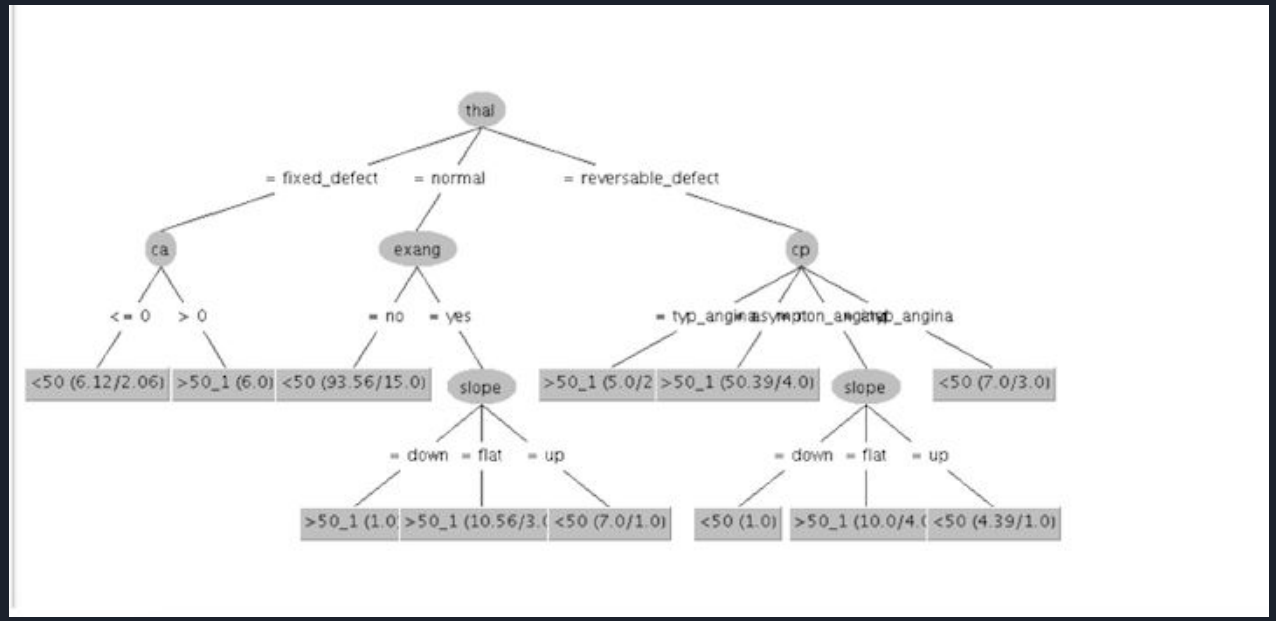
Para heart-c4 obtuvimos la siguiente regla del modelo de clasificación:

```
(thal = reversable_defect) and (thalach <= 150) => num=>50_1 (71.0/8.0)
(ca >= 1) and (slope = flat) => num=>50_1 (32.0/4.0)
(ca >= 1) and (sex = male) => num=>50_1 (30.0/10.0)
 => num=<50 (163.0/25.0)
```

Árbol de decisión C4.5

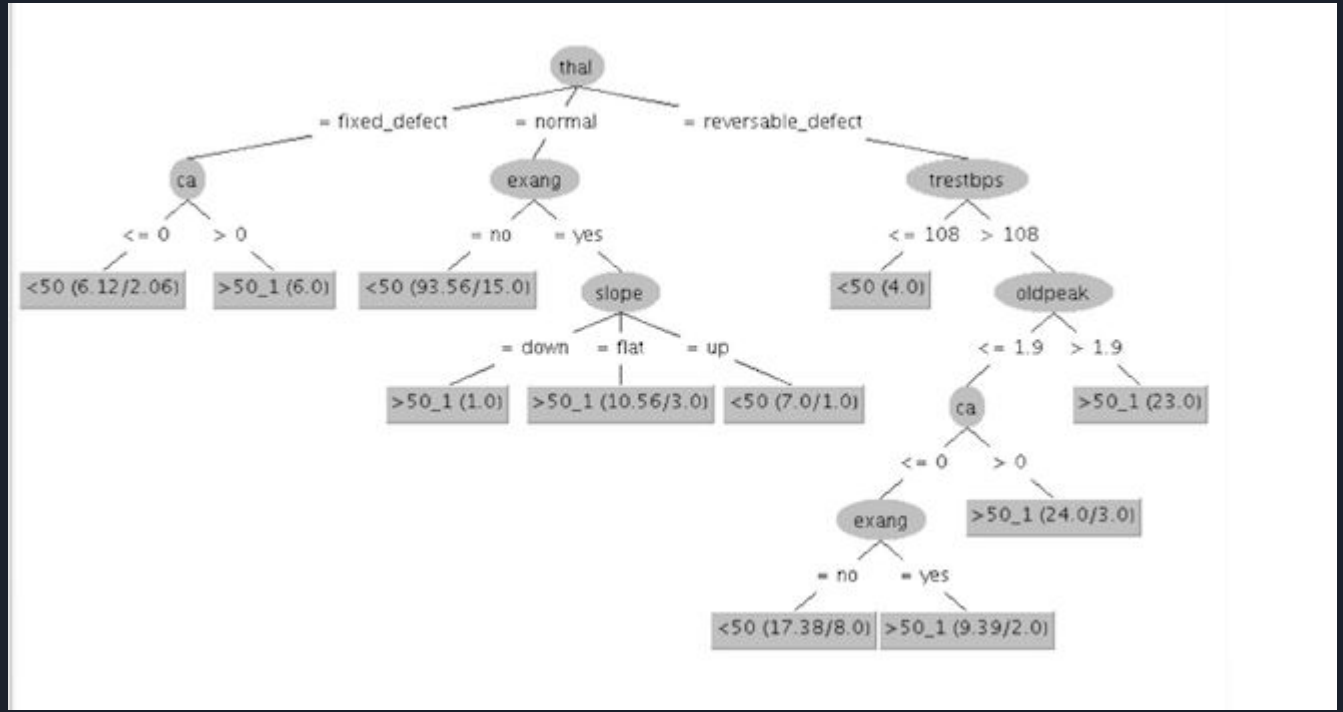
A continuación se mostrará la representación del árbol para cada daset.

Para el original:



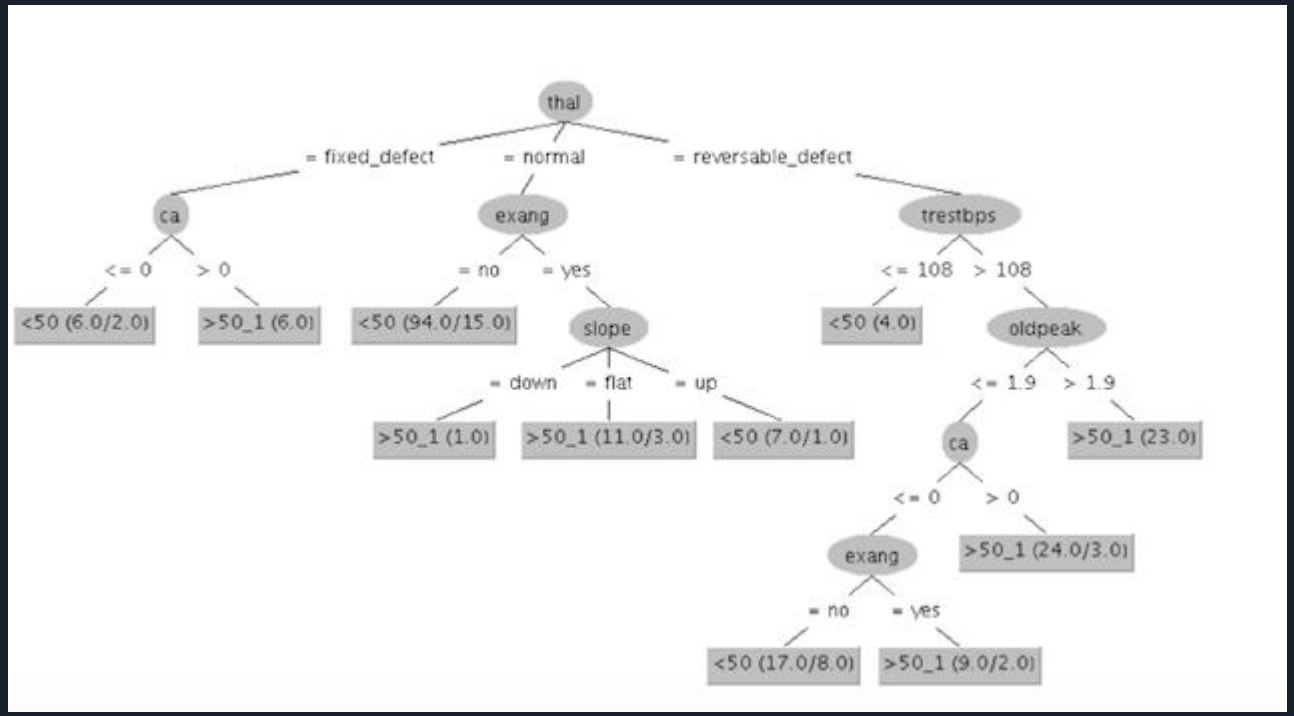
Árbol de decisión C4.5

Para heart-c1



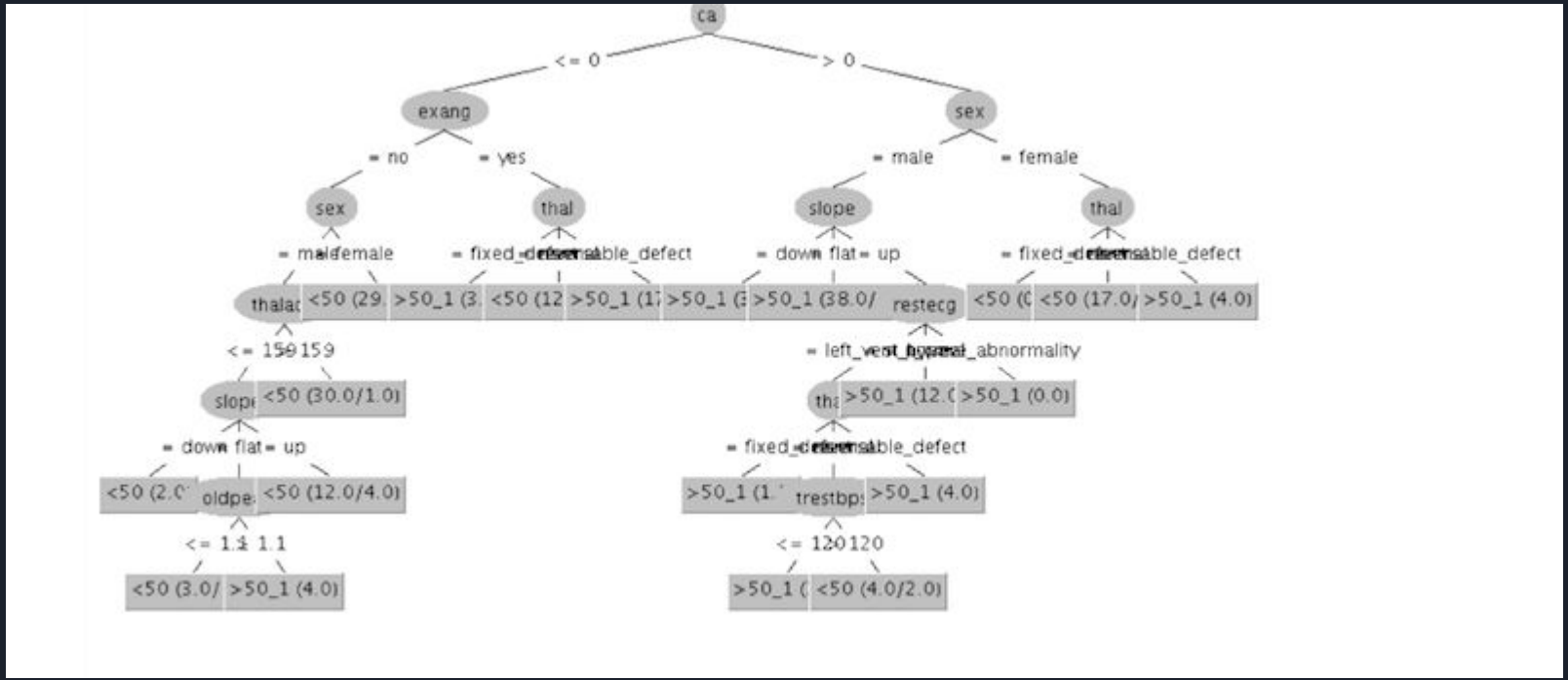
Árbol de decisión C4.5

Para el heart-c2:

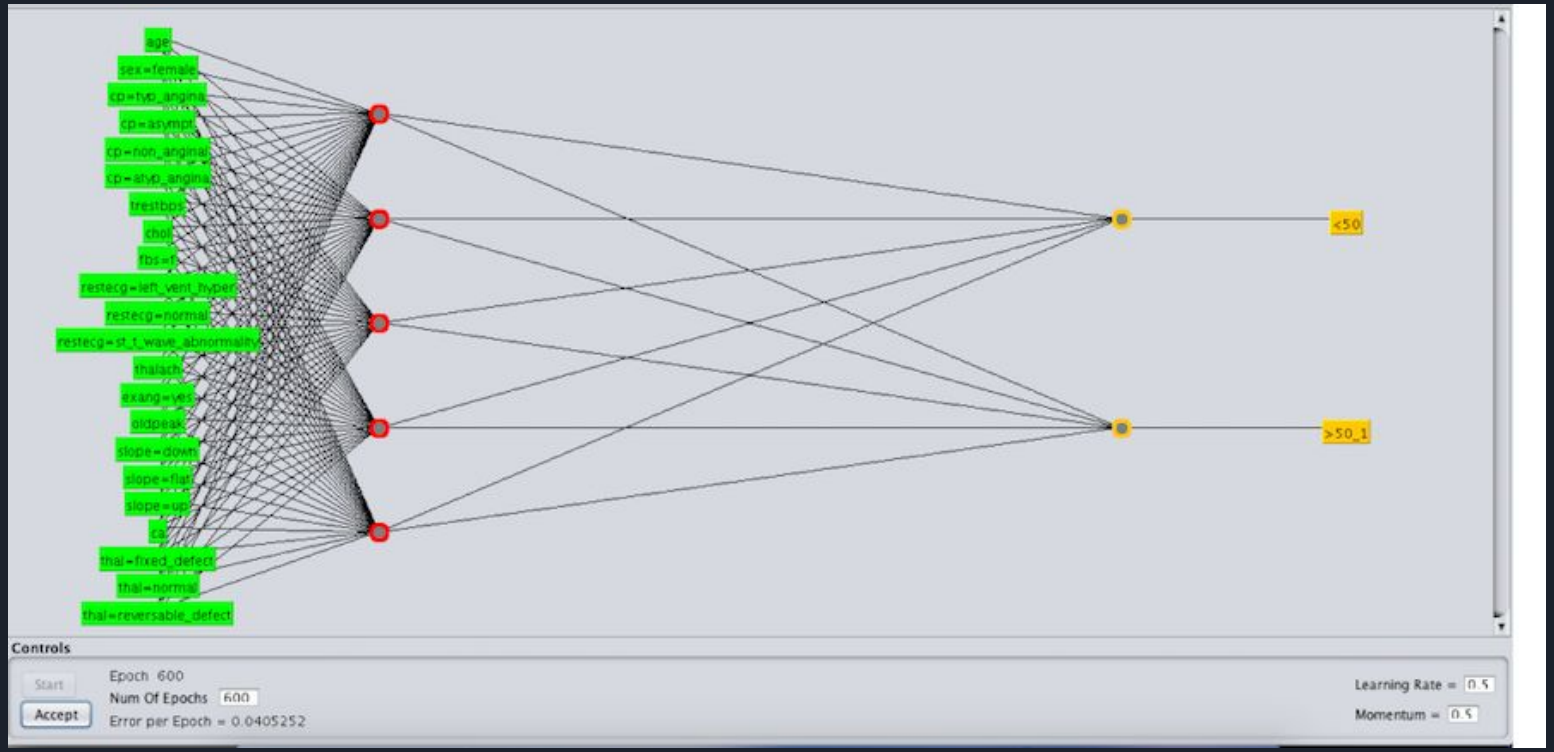


Árbol de decisión C4.5

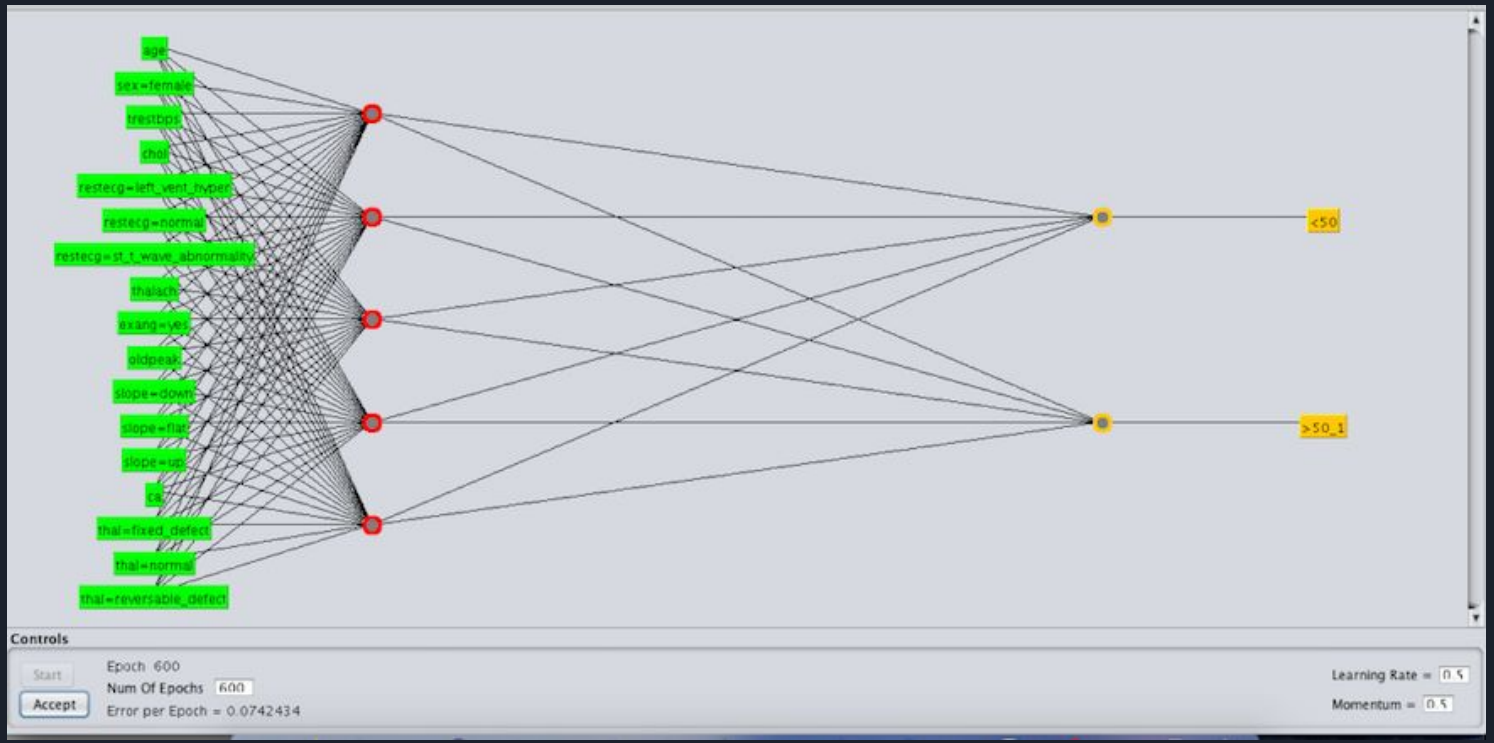
Para heart-c4:



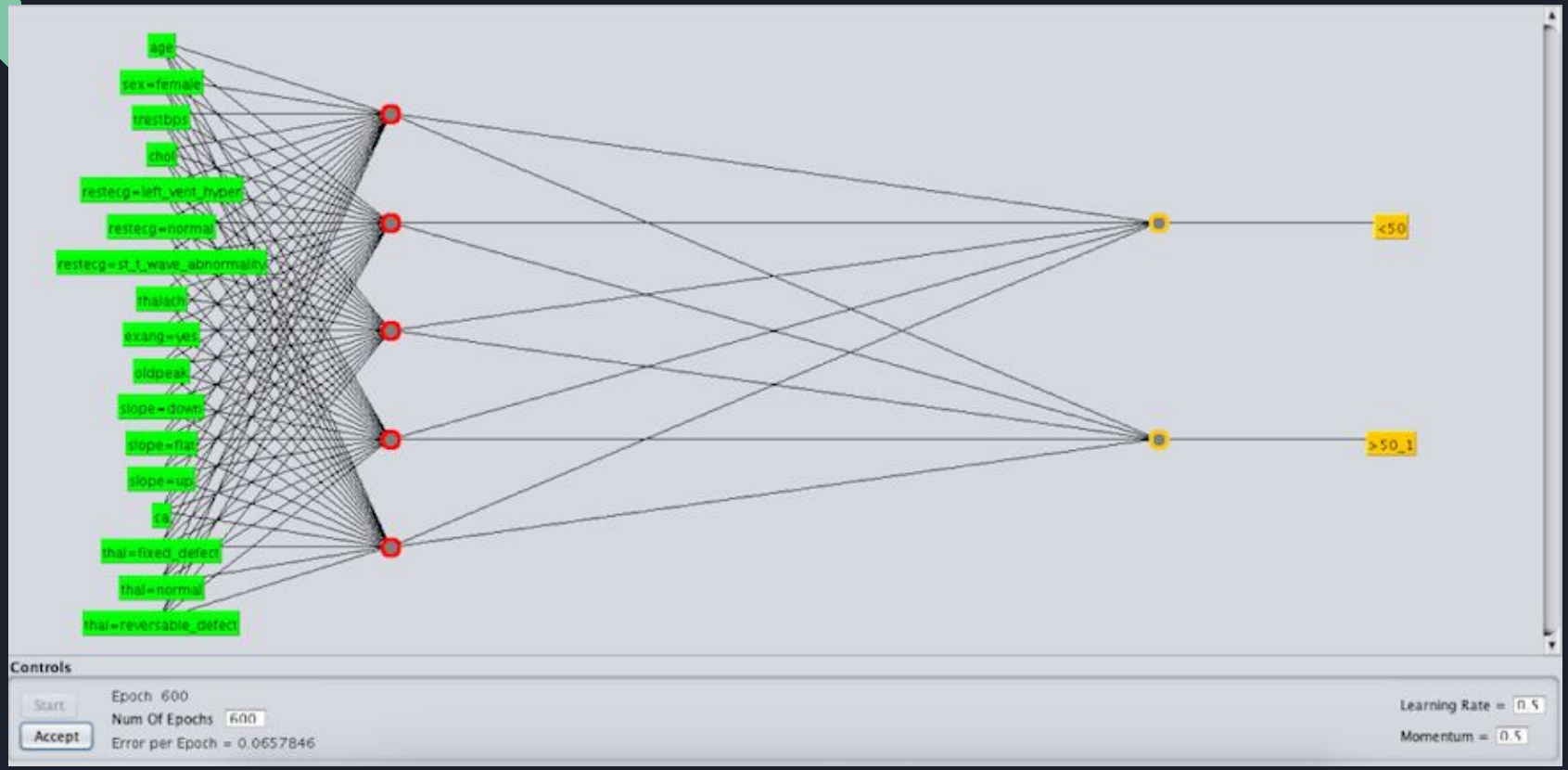
Red neuronal



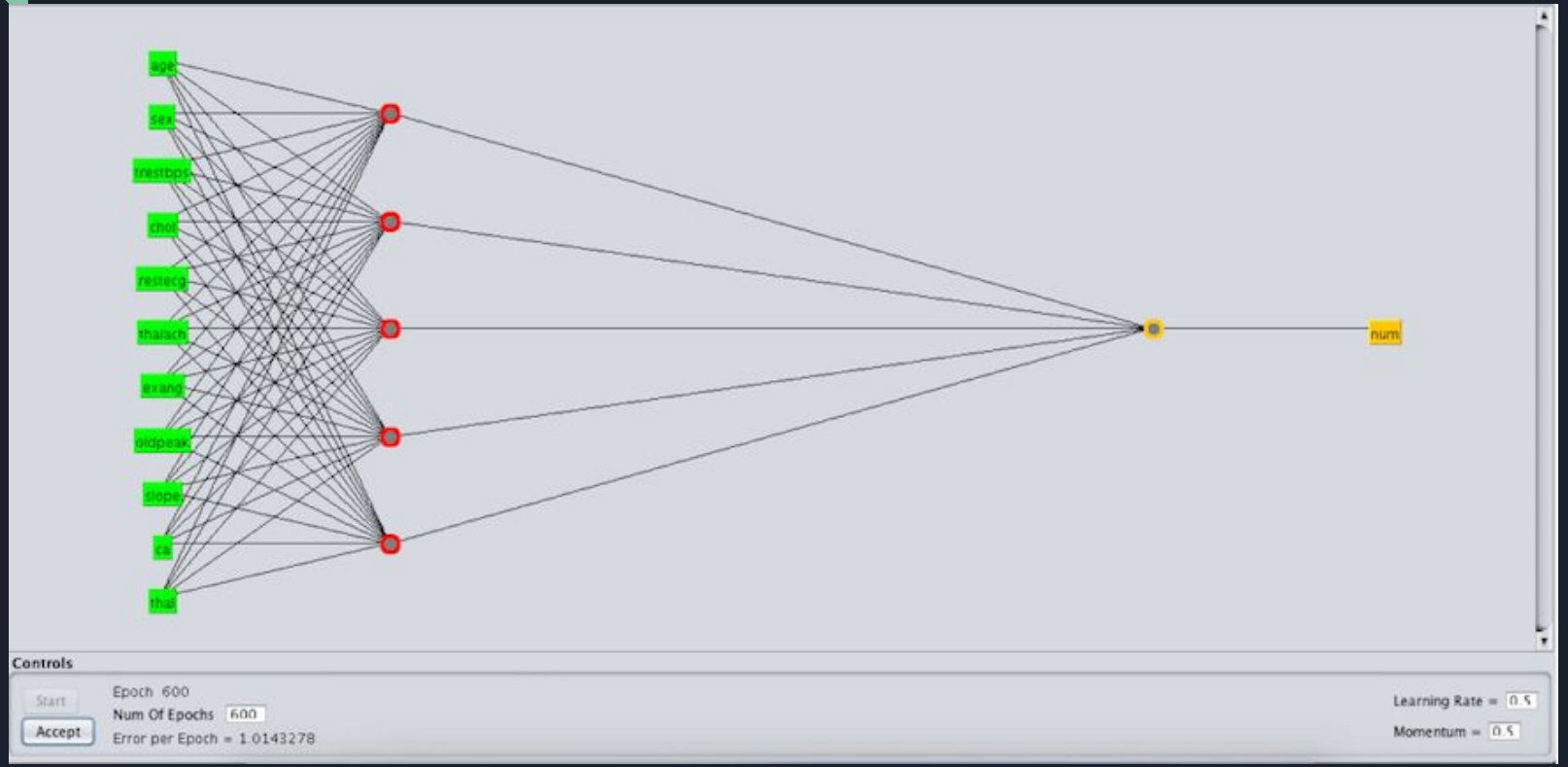
Red neuronal



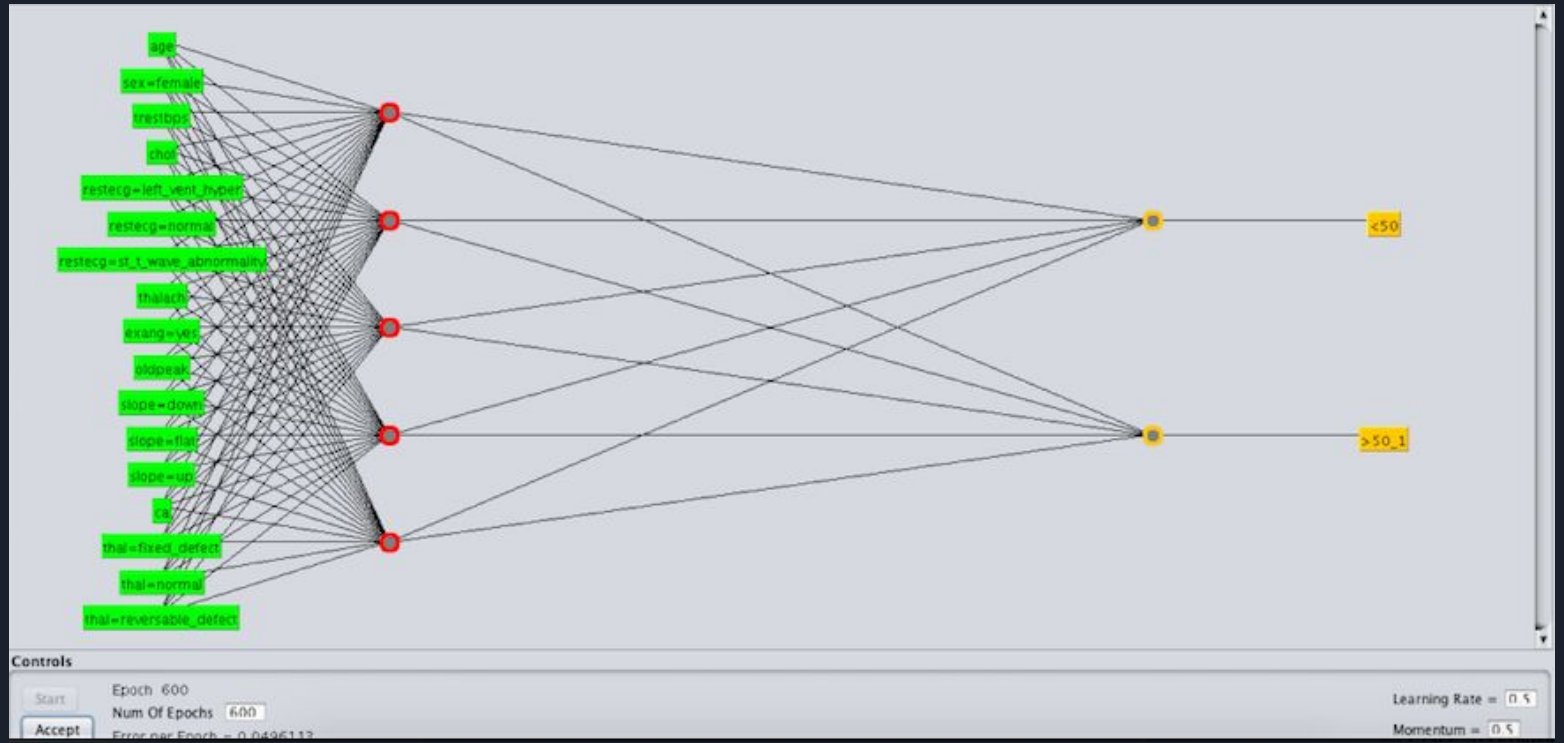
Red neuronal



Red neuronal



Red neuronal



Prueba y evaluación

Función de puntuación basada en la matriz de confusión, donde cada celda indica el porcentaje de **instancias clasificadas correctamente**.

	OneR	10 ‘fold-cross’	RIPPER	Árbol de D. C4.5	Red neuronal
Dataset original	77.6699%	71.6172	82.5243	78.6408	49.5146
heart-c1.csv	77.6699%	76.2376	74.7573	77.6699	49.5146
heart-c2.csv	77.6699%	76.5677	74.7573	77.6699	49.5146
heart-c3.csv	-	-	-	-	-
heart-c4.csv	79.2079%	73.9865	79.2079	75.2475	49.5146

Despliegue

- Clasificador **RIPPER**

1. **(cp = aympt) and (thal = reversable_defect) => num => 50_1 (78.0/7.0)**

   Si el tipo de dolor en el pecho es asintomático y es efecto reversible entonces el diagnóstico de enfermedad es mayor.

2. **(ca >= 1) and (slope = flat) => num=>50_1 (34.0/5.0)**

   Si el número de vasos principales (0-3) coloreados por fluoroscopia es mayor o igual que 1 y la pendiente del segmento ST de ejercicio pico es plano entonces el diagnóstico de enfermedad es mayor.

3. (ca >= 1) and (cp = asympt) => num=>50_1 (11.0/2.0)

Si el número de vasos principales (0-3) coloreados por fluoroscopia es mayor o igual que 1 y la pendiente del segmento ST de ejercicio pico es plano entonces el diagnóstico de enfermedad es mayor.

4. => num <= 50 (180.0/29.0)

En otro caso, el diagnóstico de enfermedad es menor.