PROYECTO DE MINERÍA DE DATOS

Alcantar Arenas Rogelio Garduño Vargas Juan Carlos

ALMACÉN Y MINERÍA DE DATOS

Universidad Nacional Autónoma de México

Facultad de Ciencias

Minería de Datos

Proyecto Final

Integrantes:

Alcantar Arenas Rogelio
Garduño Vargas Juan Carlos

Semestre: 2018-2

1.- Descripción del problema

El problema consiste en determinar, a partir de los atributos en el dataset proporcionado, los factores que contribuyen a que una persona padezca enfermedades cardiacas.

2.- Conocimiento de los Datos

En esta tarea se trabajará con el dataset acerca de enfermedades cardiacas, almacenado en la página del curso: heart.csv y heartDescription

2.1. Elabora una tabla que contenga la siguiente informaciónon para cada atributo:

- Tipo de atributo (nominal, ordinal, numerico, etc)
- Porcentaje de valores perdidos.
- Valor minimo, maximo, media, desviaciónon estandar.
- ¿Existen registros que tengan un valor para ese atributo que no aparezca en otros registros?
- ¿Tiene valores atipicos?

Atributos	Tipo	Porcentaje Valores Perdidos	Valor Mínimo	Valor máximo	Media	Desviación Estándar	Valor Atípico	Registros
Age	Numérico	0%	29	77	54.366	9.082	No	Si
Sex	Nominal	0%	-	-	-	-	No	No
Cp	Nominal	0%	-	-	-	-	No	No
Trestbps	Numérico	0%	94	200	131.62	17.538	No	Si
Chol	Numérico	0%	126	564	246.26	51.831	No	Si
Fbs	Nominal	0%	-	-	-	-	No	No
Restcg	Nominal	0%	-	-	-	-	No	No
Thalach	Numérico	0%	71	202	149.54	22.905	No	Si
Exang	Nominal	0%	-	-	-	-	No	No
Oldspeak	Numérico	0%	0	6.2	1.04	1.161	No	Si
Slope	Nominal	0%	-	-	-	-	No	No
Ca	Numérico	1.65%	0	3	0.674	0.938	Si	No

Thal	Nominal	0.66%	-	-	-	-	Si	No
Num	Nominal	0%	-	-	-	-	No	No

2.2 Haz una interpretación de los datos de acuerdo al estudio previo. Trata de determinar los atributos que aumentan el riesgo de enfermedades cardiacas.

Información de los datos:
1.- age: Edad en años
2.- sex: genero (1 = hombre; 0 = mujer)
3.- cp: Tipo de dolor de pecho
- asympt: asymptomatic(asintomático)
- non_anginal: non-anginal pain(dolor no anginal)
- atyp_angina: atypical angina(angina atípica)
- typ_angina: typical angina(angina típica)

4.- trestbps: Presion Arterial en reposo en mm Hg al ingreso en el hospital
5.- chol: colesterol sérico en mg / dl
6.- fbs: azúcar en sangre en ayunas > 120 mg/dl (1 = true; 0 = false)
7.- restecg: Resultados electrocardiograficos en reposo
- normal
- left_vent_hyper: showing probable or definite left ventricular hypertrophy by Estes' criteria(mostrando hipertrofia ventricular izquierda probable o definida según los criterios de Estes.).
- st_t_wave_abnormality: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) - tener anormalidad de la onda ST-T (inversiones de la onda T y / o elevación del ST o depresión> 0.05 mV).

8.- thalach: maxima frecuencia cardiaca lograda
9.- exang: angina inducida por el ejercicio (1 = yes; 0 = no)
10.- oldpeak: Depresión ST inducida por el ejercicio en relación con el reposo.
11.- slope: la pendiente del segmento ST de ejercicio pico
- -- Value 1: upsloping
- -- Value 2: flat
- -- Value 3: downsloping

12.- ca: número de vasos principales (0-3) coloreados por fluoroscopia
13.- thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14.- num: diagnóstico de enfermedad cardíaca, estado angiográfico de la enfermedad
- <50: < 50% diameter narrowing(50% de diámetro de estrechamiento)
- >50_1: > 50% diameter narrowing(50% de diámetro de estrechamiento).

Analisis de los atributos que aumentan el riesgo de enfermedades cardiacas

Age: La edad es importante ya que va muy ligado a al sexo, los hombres tienen más riesgo cardiovascular después de los 50 años de edad. En si su riesgo es alto. El factor protector de las mujeres son los estrógenos, empiezan con un riesgo cardiovascular después de la menopausia.

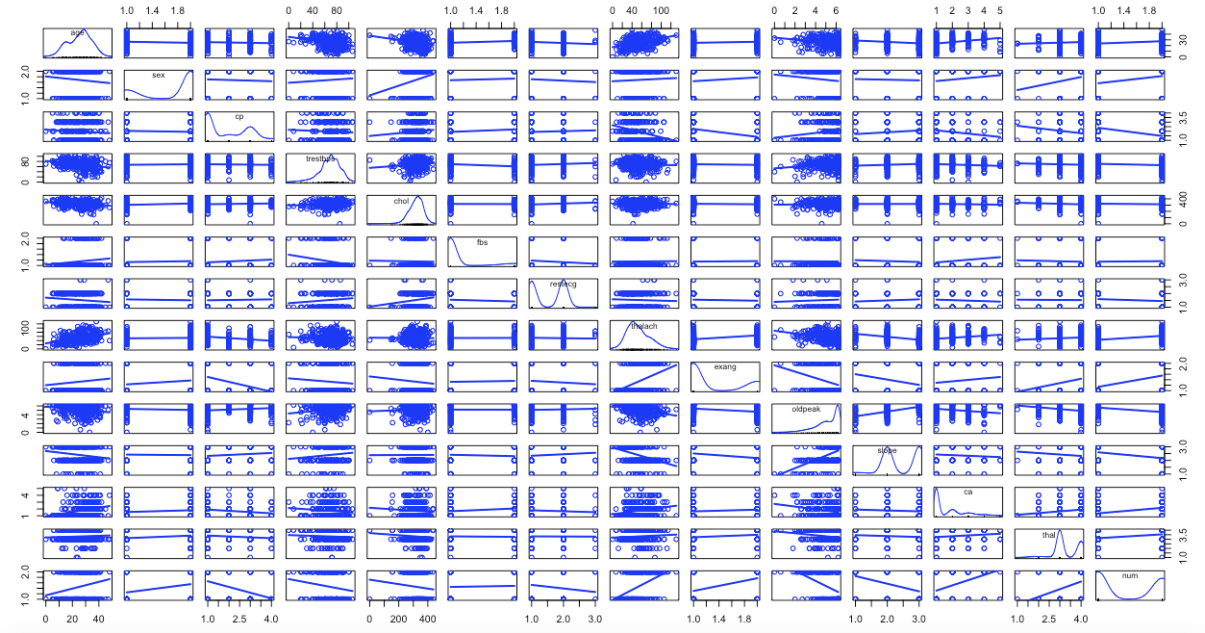
La presión de sangre: Es el principal factor de riesgo para el desarrollo de enfermedad cardiaca y evento vascular cerebral.

Colesterol: Las placas de ateroma son principales causas de infarto.

Angina inducida por el ejercicio: tenemos los síndromes coronarios agudos que son: angina estable, angina inestable e infarto agudo amiocardio.

la pendiente del segmento ST de ejercicio: Es un dato electrocardiográfico que habla de infarto, la elevación del segmento ST nos va a hablar de infarto.

2.3 Mediante una matriz de gráficas de dispersión, determina:



a) ¿Cuáles atributos parecen estar más ligados a enfermedades cardiacas?

Age = edad
Trestbps = Presión de sangre
chol = Colesterol
thalach = frecuencia cardíaca máxima lograda
Oldpeak = depresión inducida por el ejercicio en relación con el descanso

b) ¿Cuáles atributos parecen estar menos ligados a enfermedades cardiacas?

Sex = sexo
cp = Dolor de pecho
fbs= glucemia en ayunas
restecg = resultado de electrocardiográfico en reposo

exang = angina inducida por el ejercicio -
slope= la pendiente del segmento ST de ejercicio -
ca = número de vasos principales coloreados por fluoroscopia -

thal = 3 = normal; 6 = fixed defect; 7 = reversable defect
num: diagnóstico de enfermedad cardiaca

c) Resume en una tabla tus hallazgos relativos a la predicción de los valores de cada atributo.

Atributo del eje x	Atributo del eje y	Hallazgos
age	age	Tiene una correlación debil
age	sex	Tiene una correlacion debil
age	cp	Tiene una correlacion debil
age	trestbps	Tiene una correlacion media
age	chol	Tiene una correlacion media
age	fbs	Tiene una correlacion debil
age	restecg	Tiene una correlacion debil
age	thalach	Tiene una correlacion media
age	exang	Tiene una correlacion debil
age	oldpeak	Tiene una correlacion media
age	slope	Tiene una correlacion debil
age	ca	Tiene una correlacion debil
age	thal	Tiene una correlacion debil
age	num	Tiene una correlacion debil

Atributo del eje x	Atributo del eje y	Hallazgos
sex	sex	Tiene una correlación fuerte
sex	cp	Tiene una correlación débil
sex	trestbps	Tiene una correlación débil
sex	chol	Tiene una correlación débil

sex	fbs	Tiene una correlación débil
sex	restecg	Tiene una correlación débil
sex	thalach	Tiene una correlación débil
sex	exang	Tiene una correlación débil
sex	oldpeak	Tiene una correlación débil
sex	slope	Tiene una correlación débil
sex	ca	Tiene una correlación débil
sex	thal	Tiene una correlación débil
sex	num	Tiene una correlación débil

Atributo del eje x	Atributo del eje y	Hallazgos
cp	cp	Tiene una correlación fuerte
cp	trestbps	Tiene una correlación débil
cp	chol	Tiene una correlación débil
cp	fbs	Tiene una correlación débil
cp	restecg	Tiene una correlación débil
cp	thalach	Tiene una correlación débil
cp	exang	Tiene una correlación débil
cp	oldpeak	Tiene una correlación débil
cp	slope	Tiene una correlación débil
cp	ca	Tiene una correlación débil
cp	thal	Tiene una correlación débil
cp	num	Tiene una correlación débil

Atributo del eje x	Atributo del eje y	Hallazgos
trestbps	trestbps	Tiene una correlación fuerte
trestbps	chol	Tiene una correlación media
trestbps	fbs	Tiene una correlación débil

trestbps	restecg	Tiene una correlación débil
trestbps	thalach	Tiene una correlación media
trestbps	exang	Tiene una correlación débil
trestbps	oldpeak	Tiene una correlación media
trestbps	slope	Tiene una correlación débil
trestbps	ca	Tiene una correlación débil
trestbps	thal	Tiene una correlación débil
trestbps	num	Tiene una correlación débil

Atributo del eje x	Atributo del eje y	Hallazgos
chol	chol	Tiene una correlación fuerte
chol	fbs	Tiene una correlación débil
chol	restecg	Tiene una correlación débil
chol	thalach	Tiene una correlación media
chol	exang	Tiene una correlación débil
chol	oldpeak	Tiene una correlación media
chol	slope	Tiene una correlación débil
chol	ca	Tiene una correlación débil
chol	thal	Tiene una correlación débil
chol	num	Tiene una correlación débil

Atributo del eje x	Atributo del eje y	Hallazgos
thalach	thalach	Tiene una correlación fuerte
thalach	exang	Tiene una correlación débil
thalach	oldpeak	Tiene una correlación media
thalach	slope	Tiene una correlación débil
thalach	ca	Tiene una correlación débil
thalach	thal	Tiene una correlación débil
thalach	num	Tiene una correlación débil

Atributo del eje x	Atributo del eje y	Hallazgos
oldpeak	oldpeak	Tiene una correlación fuerte
oldpeak	slope	Tiene una correlación débil
oldpeak	ca	Tiene una correlación débil
oldpeak	thal	Tiene una correlación débil
oldpeak	num	Tiene una correlación débil

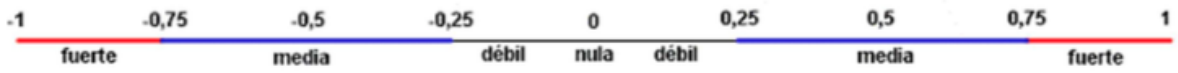
d) ¿Existen atributos correlacionados?

Obtenemos la matriz de correlación, por medio de la herramienta R

	age	sex	cp	trestbps	chol	fbs	restecg
age	1.00000000	0.09844660	-0.10295202	0.27935091	0.213677957	0.121307648	0.17119947
sex	0.09844660	1.00000000	0.01401255	0.05676882	0.197912174	-0.045031789	0.01202636
cp	-0.10295202	0.01401255	1.00000000	0.04081803	-0.069791939	0.038584641	-0.08498156
trestbps	0.27935091	0.05676882	0.04081803	1.00000000	0.123174207	0.177530542	0.15216530
chol	0.21367796	0.19791217	-0.06979194	0.12317421	1.000000000	0.013293602	0.17341105
fbs	0.12130765	-0.04503179	0.03858464	0.17753054	0.013293602	1.000000000	0.05262144
restecg	0.17119947	0.01202636	-0.08498156	0.15216530	0.173411050	0.052621445	1.00000000
thalach	-0.39852194	0.04401991	0.33639203	-0.04669773	-0.009939839	-0.008567107	-0.12281166
exang	0.09680083	-0.14166381	-0.38693543	0.06761612	0.067022783	0.025665147	0.09848464
oldpeak	0.21001257	-0.09609288	-0.20561817	0.19321647	0.053951920	0.005747223	0.16825560
slope	0.16881424	-0.03071057	-0.15560404	0.12147458	0.004037770	0.059894178	0.17246310
ca	0.35953460	-0.08974222	-0.23492389	0.10274275	0.121192155	0.143098453	0.13606449
thal	0.13481816	-0.37531740	-0.26548160	0.13825267	0.025713421	0.064624763	0.02910558
num	0.22543872	-0.28093658	-0.41252222	0.14493113	0.085239105	0.028045760	0.18163263

	thalach	exang	oldpeak	slope	ca	thal	num
age	-0.398521938	0.09680083	0.210012567	0.16881424	0.35953460	0.13481816	0.22543872
sex	0.044019908	-0.14166381	-0.096092877	-0.03071057	-0.08974222	-0.37531740	-0.28093658
cp	0.336392029	-0.38693543	-0.205618171	-0.15560404	-0.23492389	-0.26548160	-0.41252222
trestbps	-0.046697728	0.06761612	0.193216472	0.12147458	0.10274275	0.13825267	0.14493113
chol	-0.009939839	0.06702278	0.053951920	0.00403777	0.12119215	0.02571342	0.08523911
fbs	-0.008567107	0.02566515	0.005747223	0.05989418	0.14309845	0.06462476	0.02804576
restecg	-0.122811656	0.09848464	0.168255597	0.17246310	0.13606449	0.02910558	0.18163263
thalach	1.000000000	-0.37881209	-0.344186948	-0.38678441	-0.26208824	-0.27532286	-0.42174093
exang	-0.378812094	1.00000000	0.288222808	0.25774837	0.14317517	0.32524030	0.43675708
oldpeak	-0.344186948	0.28822281	1.000000000	0.57753682	0.29264621	0.34240455	0.43069600
slope	-0.386784410	0.25774837	0.577536817	1.00000000	0.10737100	0.28679183	0.34587708
ca	-0.262088239	0.14317517	0.292646215	0.10737100	1.00000000	0.25577715	0.46085196
thal	-0.275322864	0.32524030	0.342404551	0.28679183	0.25577715	1.00000000	0.52825411
num	-0.421740934	0.43675708	0.430696002	0.34587708	0.46085196	0.52825411	1.00000000

Por medio de esta tabla podemos analizar la correlación entre los atributos



Entonces con esto, podemos observar que la diagonal de la matriz tiene una correlación con un valor de 1 lo cual es una correlación fuerte pues son atributos correlacionados a sí mismos.

Para observar una correlación nula podemos ver en la coordenada de “slope” y “chol” con una correlación de 0.004

Para una correlación media podemos observar la correlación de los atributos “ca” con “num” con una correlación de 0.46

2.4 Investiga posibles asociaciones de atributos con el atributo de clase. Es decir, estudia las gráficas de dispersión elaboradas en el punto anterior y trata de identificar posibles áreas “densas” de enfermedades cardíacas.

-Si hubiera áreas “densas” en alguna(s) gráfica(s) de dispersión, cuantifican las enfermedades cardíacas en ellas con respecto al dataset completo.

Basándonos en la matriz de correlación descrita en los ejercicios anteriores, podemos observar una densidad de valores en los atributos “age” y “trestbps”, “age” y “chol”, “oldpeak” y “thalach”, etc.

3- Preprocesamiento de datos (Preparación de datos).

En este paso se preparan los datos de acuerdo con las tareas de minería que se van a realizar.

3.1 Selección de atributos

Selecciona los atributos que consideres apropiados para una tarea predictiva. Justifica tu respuesta. Guarda los atributos seleccionados en un archivo llamado heart-c1.csv.

Los atributos que nosotros consideramos son:

- age
- sex
- trestbps
- chol
- restecg
- thalach

- exang
- oldpeak
- slope
- ca
- num

Quitamos dolor de pecho, ya que este no es relevante para que se tenga una enfermedad cardiaca, el dolor de pecho puede ser producido por varios factores.

3.2 Manejo de valores perdidos.

Considera los siguientes métodos para tratar con valores perdidos:

a) *Reemplaza los valores perdidos por la media o la moda del atributo, de acuerdo con el tipo de dato del atributo. Guarda el dataset resultante en un archivo con el nombre de heart-c2.csv*

b) Utiliza regresión lineal para estimar los valores perdidos de cada atributo. Guarda el dataset resultante en el archivo heart-c3.csv.

3.3 Eliminado de atípicos. Elimina los registros atípicos y guarda el resultado en el archivo heart-c4.csv

4.- Minado de datos (Construcción del modelo).

Tareas de clasificación

Estas tareas con los diferentes data sets, se pueden verificar en los documentos: heart.doc, heart-c1.doc, heart-c2.doc, heart-c3.doc, heart-c4.doc

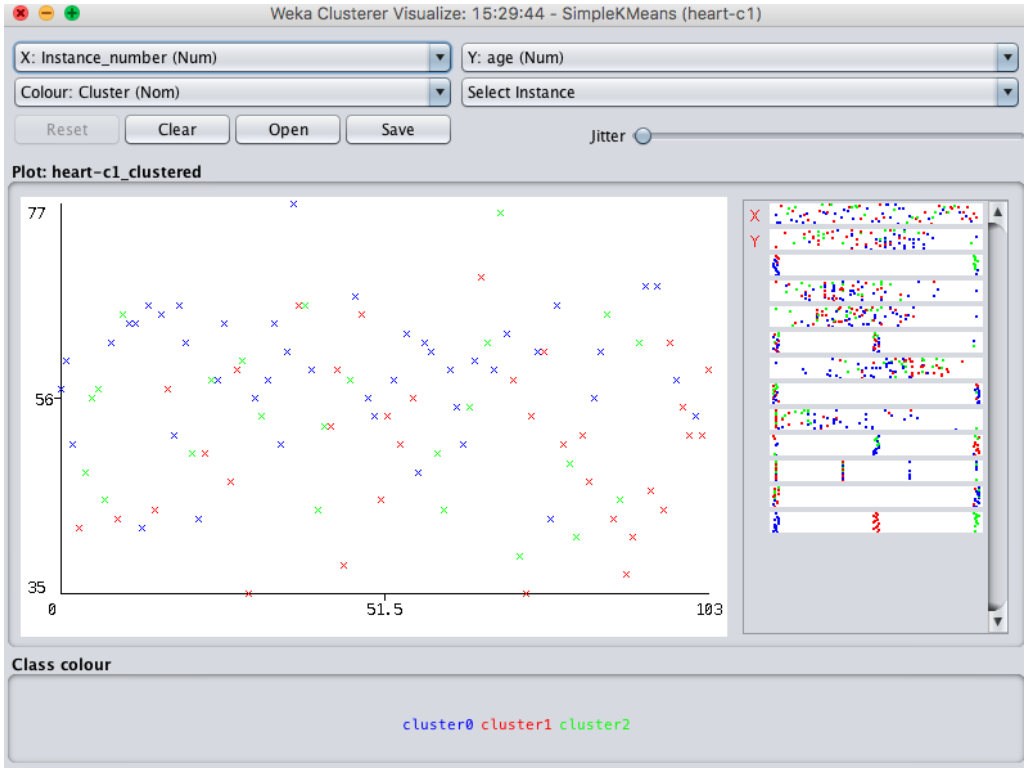
4.2 Tareas de agrupamiento

Investiga si hay una tendencia de clustering en el dataset. Empieza agrupando los datos con el algoritmo k-medias, para algunas k, $2 \leq k \leq 10$.

No uses el atributo de clase num.

b) Encuentra un valor adecuado para k. Justifica tu respuesta.

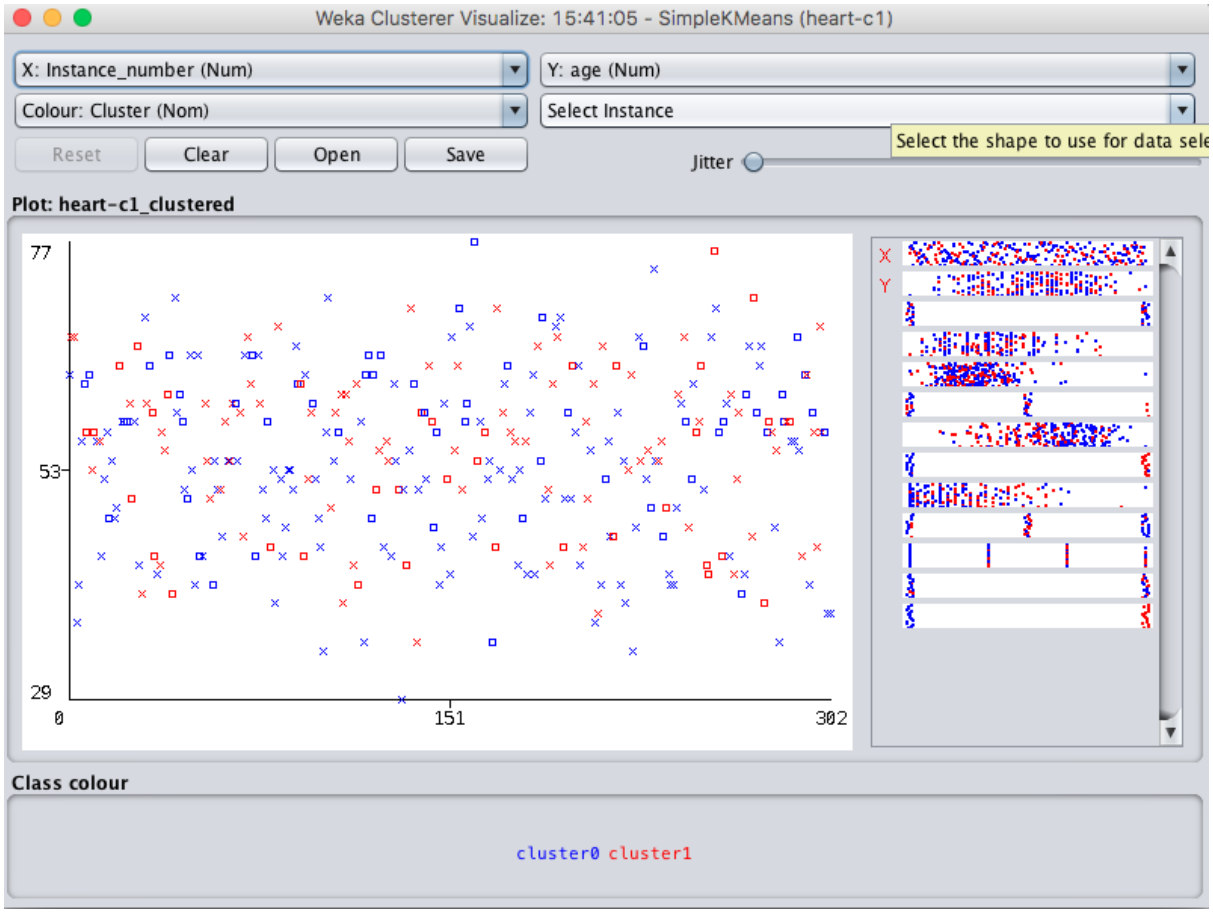
```
=== Model and evaluation on test split ===

kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 268.86042633233706

Initial starting points (random):

Cluster 0: 50,male,140,233,normal,163,no,0.6,flat,1,>50_1
Cluster 1: 41,female,130,204,left_vent_hyper,172,no,1.4,up,0,<50
Cluster 2: 65,female,140,417,left_vent_hyper,157,no,0.8,up,1,<50

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                  Full Data           0              1              2
                             (199.0)        (68.0)         (75.0)         (56.0)
=============================================================================================
age                           53.9497       55.8235        51.0133        55.6071
sex                              male          male           male         female
trestbps                     130.7538      132.2353       130.6133       129.1429
chol                         247.4121      246.8676       231.3867       269.5357
restecg             left_vent_hyper        normal left_vent_hyper left_vent_hyper
thalach                      152.3518      136.3529       164.9867       154.8571
exang                              no           yes             no             no
oldpeak                        0.9734        1.6941          0.516         0.7107
slope                              up          flat             up             up
ca                             0.6583        1.1618         0.3333         0.4821
num                               <50          >50_1            <50            <50



Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0       45 ( 43%)
1       35 ( 34%)
2       24 ( 23%)
```



Una vez hecho el inciso a) y b) decidimos que un valor óptimo para k es de 3, puesto que vimos una mejor distribución de los datos.

c) Usa el atributo de clase para evaluar el cluster y asegurate que las desviaciones estándar se calculan sobre los atributos numéricos.

```
Final cluster centroids:
                                          Cluster#
Attribute                     Full Data           0              1
                                (303.0)     (184.0)        (119.0)
=============================================================================
age                             54.3663     53.8207        55.2101
sex                                male        male           male
trestbps                       131.6238    131.2609       132.1849
chol                            246.264    247.9293       243.6891
restecg                          normal left_vent_hyper       normal
thalach                        149.6469    157.5761       137.3866
exang                                no          no            yes
oldpeak                          1.0396      0.7283          1.521
slope                                up          up           flat
ca                               0.6634       0.587         0.7815



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      184 ( 61%)
1      119 ( 39%)


Class attribute: num
Classes to Clusters:

   0   1  <-- assigned to cluster
 128  37 | <50
  56  82 | >50_1

Cluster 0 <-- <50
Cluster 1 <-- >50_1

Incorrectly clustered instances :       93.0     30.6931 %
```

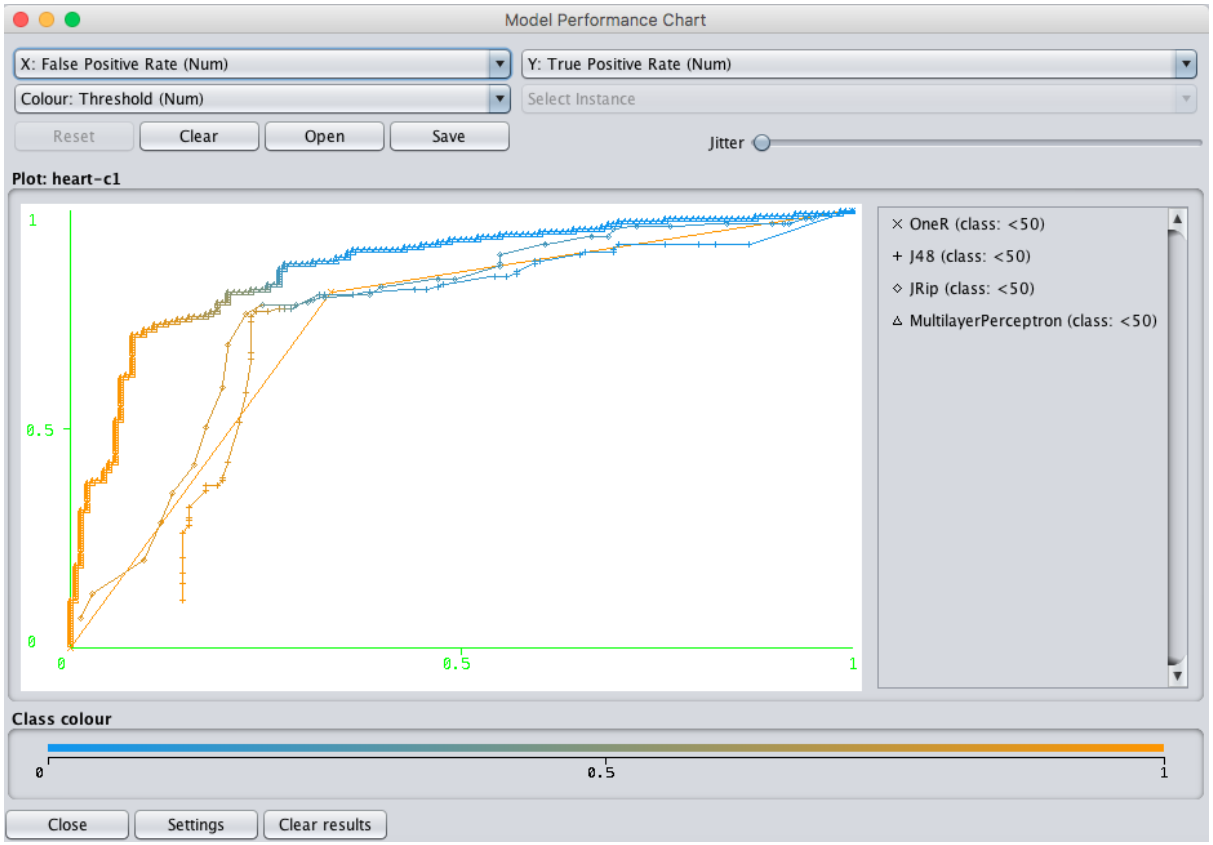


Una vez realizado el algoritmo, podemos observar que el margen de error al clasificar instancias de clase es de un 30.69% con una k de valor 2 lo cual proporciona una aproximación muy buena, con un rendimiento eficiente.

d) Saca conclusiones de las medidas numéricas desplegadas para cada cluster

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 408.1532428251373

Initial starting points (random):

Cluster 0: 65,female,140,417,left_vent_hyper,157,no,0.8,up,1
Cluster 1: 58,male,114,318,st_t_wave_abnormality,140,no,4.4,down,3
```

e) Investiga la posibilidad de usar la información del cluster para construir un clasificador para la variable num. Compara los resultados con los obtenidos con las técnicas anteriores.¿Cuál es mejor clasificador?



A la vista de los resultados, hay que concluir (en principio, sorprendentemente) que el perceptrón multicapa es el algoritmo que mejor rendimiento ofrece, en cualquier caso. Si bien en la fase de entrenamiento era el que peores resultados obtenía en comparación con los otros modelos, finalmente obtuvo una tasa de éxito del 50%, que es bastante si nos fijamos en los pobres resultados de las otras alternativas.

Por otro lado, esta experiencia muestra que hay que valorar con mucha cautela los resultados de la evaluación del entrenamiento, ya que pueden conducir a confianzas erróneas en determinados algoritmos.

5.- Conclusión (Prueba y evaluación).

Vamos a comparar los modelos creados en el paso 4.1, de aquí obtendremos el mejor clasificador para cada dataset.

Heart

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:43 PM


Dataset                   (1) rules.On | (2) rules (3) trees (4) funct
----------------------------------------------------------------------
heart                     (10)   73.78 |   76.60     77.66     79.90 v
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (1/0/0)


Key:
(1) rules.OneR
(2) rules.JRip
(3) trees.J48
(4) functions.MultilayerPerceptron
```

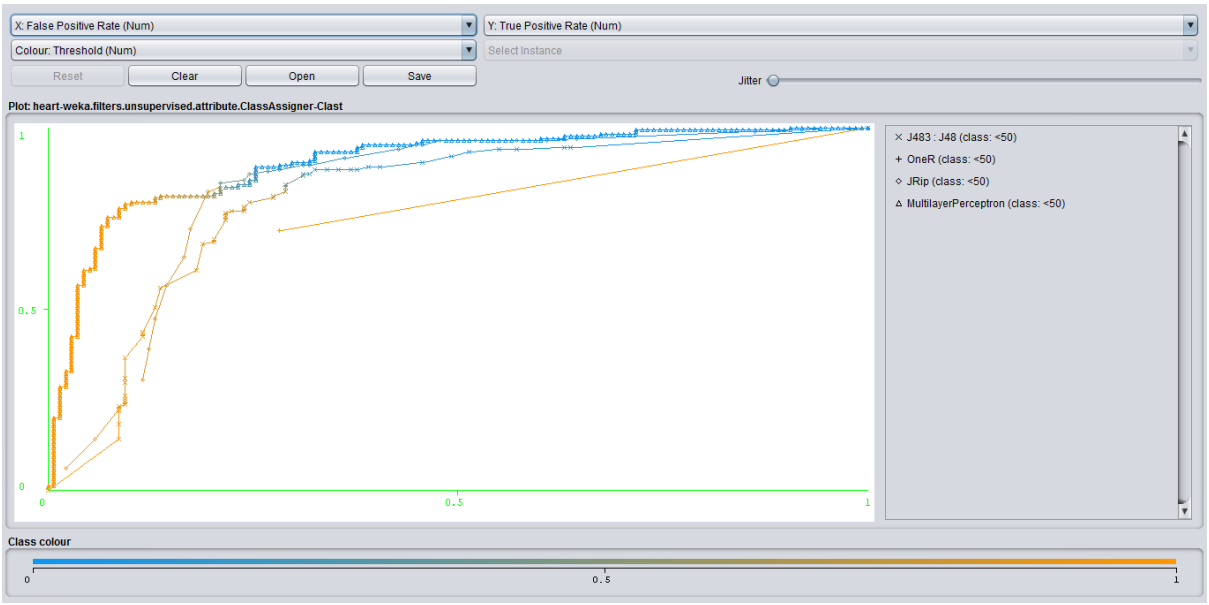
```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:49 AM


Dataset                   (2) rules.JR | (1) rules (3) trees (4) funct
----------------------------------------------------------------------
heart                     (10)   76.60 |   73.78     77.66     79.90
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.OneR
(2) rules.JRip
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:50 AM


Dataset                   (4) function | (1) rules (2) rules (3) trees
----------------------------------------------------------------------
heart                     (10)   79.90 |   73.78 *   76.60     77.66
----------------------------------------------------------------------
                               (v/ /*) |   (0/0/1)   (0/1/0)   (0/1/0)


Key:
(1) rules.OneR
(2) rules.JRip
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:49 AM


Dataset                   (3) trees.J4 | (1) rules (2) rules (4) funct
------------------------------------------------------------------------
heart                     (10)   77.66 |   73.78     76.60     79.90
------------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.OneR
(2) rules.JRip
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:45 PM


     a     b     c     d  (No. of datasets where [col] >> [row])
     - 1 (0) 1 (0) 1 (1) | a = (1) rules.OneR
 0 (0)     - 1 (0) 1 (0) | b = (2) rules.JRip
 0 (0) 0 (0)     - 1 (0) | c = (3) trees.J48
 0 (0) 0 (0) 0 (0)     - | d = (4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:45 PM


>-<   >   < Resultset
  1   1   0 functions.MultilayerPerceptron
  0   0   0 trees.J48
  0   0   0 rules.JRip
 -1   0   1 rules.OneR
```



Una vez analizando los datos y la tabla ROC podemos ver que MultilayerPerceptron tiene mejor predicción a comparación de los otros 3 clasificadores, esto lo aseguramos ya que en la tabla ROC la grafica de MultilayerPerceptron tiene mayor inclinación hacia la parte superior izquierda.

Heart-C1

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:46 PM


Dataset                   (1) rules.On | (2) rules (3) trees (4) funct
----------------------------------------------------------------------
heart-c1                  (10)   71.45 |   76.20     76.59     78.45 v
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (1/0/0)


Key:
(1) rules.OneR
(2) rules.JRip
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:51 AM


Dataset                   (2) rules.JR | (1) rules (3) trees (4) funct
----------------------------------------------------------------------
heart-c1                  (10)   76.20 |   71.45     76.59     78.45
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.OneR
(2) rules.JRip
(3) trees.J48
(4) functions.MultilayerPerceptron
```

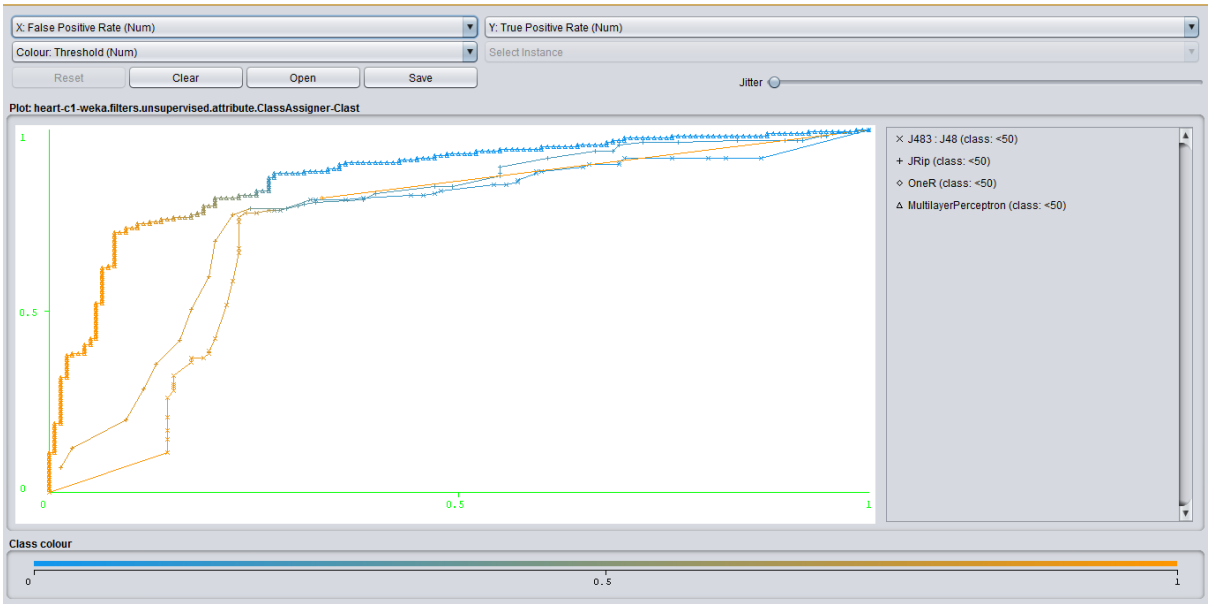
```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:51 AM


Dataset                   (3) trees.J4 | (1) rules (2) rules (4) funct
----------------------------------------------------------------------
heart-c1                  (10)   76.59 |   71.45     76.20     78.45
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.OneR
(2) rules.JRip
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:52 AM


Dataset                   (4) function | (1) rules (2) rules (3) trees
----------------------------------------------------------------------
heart-c1                  (10)   78.45 |   71.45 *   76.20     76.59
----------------------------------------------------------------------
                               (v/ /*) |   (0/0/1)   (0/1/0)   (0/1/0)


Key:
(1) rules.OneR
(2) rules.JRip
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:49 PM


     a     b     c     d  (No. of datasets where [col] >> [row])
     - 1 (0) 1 (0) 1 (1) | a = (1) rules.OneR
 0 (0)     - 1 (0) 1 (0) | b = (2) rules.JRip
 0 (0) 0 (0)     - 1 (0) | c = (3) trees.J48
 0 (0) 0 (0) 0 (0)     - | d = (4) functions.MultilayerPerceptron
```


```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:49 PM
```


```
>-<   >   < Resultset
  1   1   0 functions.MultilayerPerceptron
  0   0   0 trees.J48
  0   0   0 rules.JRip
 -1   0   1 rules.OneR
```



Una vez analizando los datos y la tabla ROC podemos ver que MultilayerPerceptron tiene mejor predicción a comparación de los otros 3 clasificadores, esto lo aseguramos ya que en la tabla ROC la gráfica de MultilayerPerceptron tiene mayor inclinación hacia la parte superior izquierda, para este dataset se ve a simple vista ya que se distingue en apariencia.

Heart-C2

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:53 PM


Dataset                   (1) rules.JR | (2) rules (3) trees (4) funct
----------------------------------------------------------------------
heart-c2                  (10)   76.60 |   73.78     73.39     79.23
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.JRip
(2) rules.OneR
(3) trees.J48
(4) functions.MultilayerPerceptron
```

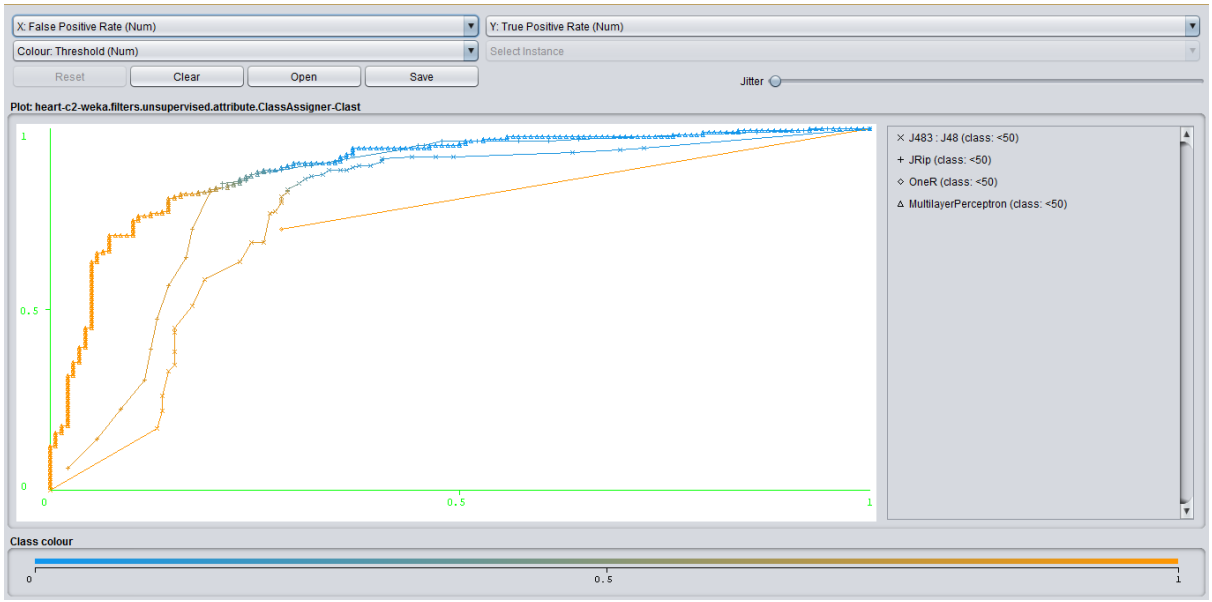
```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:54 AM


Dataset                   (2) rules.On | (1) rules (3) trees (4) funct
----------------------------------------------------------------------
heart-c2                  (10)   73.78 |   76.60     73.39     79.23
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.JRip
(2) rules.OneR
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:54 AM


Dataset                   (3) trees.J4 | (1) rules (2) rules (4) funct
----------------------------------------------------------------------
heart-c2                  (10)   73.39 |   76.60     73.78     79.23
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.JRip
(2) rules.OneR
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:55 AM


Dataset                  (4) function | (1) rules (2) rules (3) trees
----------------------------------------------------------------------
heart-c2                 (10)   79.23 |   76.60     73.78     73.39
----------------------------------------------------------------------
                              (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.JRip
(2) rules.OneR
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:55 PM


     a     b     c     d  (No. of datasets where [col] >> [row])
     - 0 (0) 0 (0) 1 (0) | a = (1) rules.JRip
 1 (0)     - 0 (0) 1 (0) | b = (2) rules.OneR
 1 (0) 1 (0)     - 1 (0) | c = (3) trees.J48
 0 (0) 0 (0) 0 (0)     - | d = (4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:55 PM


>-<   >   < Resultset
  0   0   0 functions.MultilayerPerceptron
  0   0   0 trees.J48
  0   0   0 rules.OneR
  0   0   0 rules.JRip
```



Una vez analizando los datos y la tabla ROC podemos ver que MultilayerPerceptron tiene mejor predicción a comparación de los otros 3 clasificadores, esto lo aseguramos ya que en la tabla ROC la gráfica de MultilayerPerceptron tiene mayor inclinación hacia la parte superior izquierda, aquí podemos notar que desde que empieza a clasificar se distingue de las demás rectas.

Heart-C4

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:59 PM


Dataset                   (1) rules.JR | (2) rules (3) trees (4) funct
----------------------------------------------------------------------
heart-c4                  (10)   76.87 |   72.13     76.77     80.43
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.JRip
(2) rules.OneR
(3) trees.J48
(4) functions.MultilayerPerceptron
```

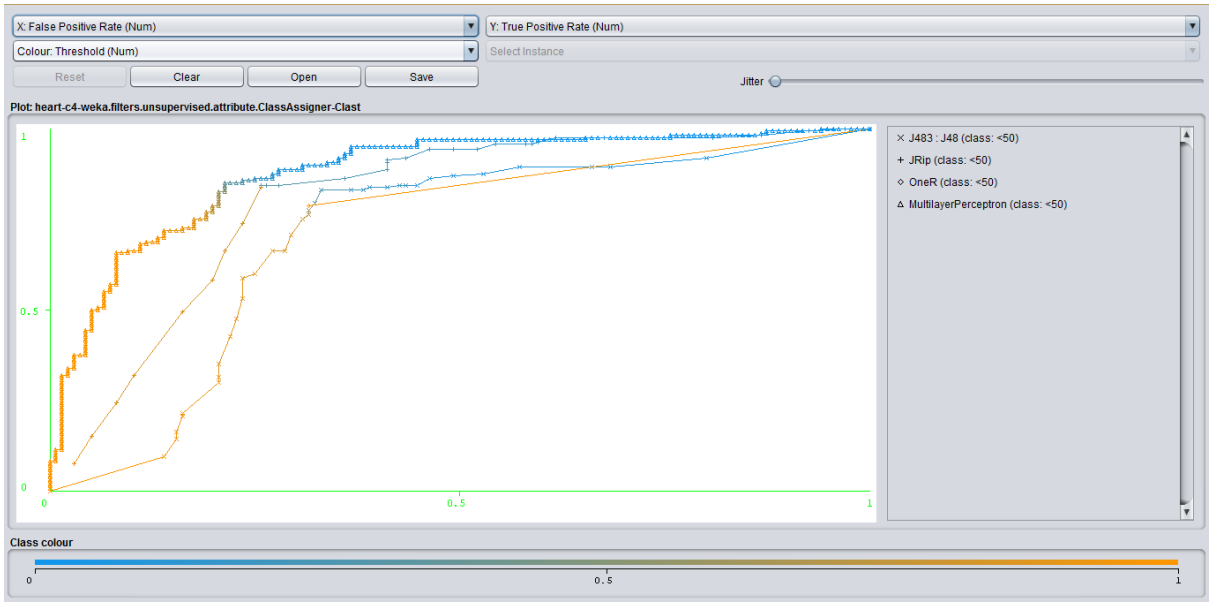
```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:55 AM


Dataset                   (1) rules.JR | (2) rules (3) trees (4) funct
----------------------------------------------------------------------
heart-c4                  (10)   76.87 |   72.13     76.77     80.43
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.JRip
(2) rules.OneR
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:56 AM


Dataset                   (3) trees.J4 | (1) rules (2) rules (4) funct
----------------------------------------------------------------------
heart-c4                  (10)   76.77 |   76.87     72.13     80.43
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)


Key:
(1) rules.JRip
(2) rules.OneR
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       9/06/18 12:56 AM


Dataset                   (4) function | (1) rules (2) rules (3) trees
----------------------------------------------------------------------
heart-c4                  (10)   80.43 |   76.87     72.13 *   76.77
----------------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/0/1)   (0/1/0)


Key:
(1) rules.JRip
(2) rules.OneR
(3) trees.J48
(4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 09:59 PM


     a     b     c     d  (No. of datasets where [col] >> [row])
     - 0 (0) 0 (0) 1 (0) | a = (1) rules.JRip
 1 (0)     - 1 (0) 1 (1) | b = (2) rules.OneR
 1 (0) 0 (0)     - 1 (0) | c = (3) trees.J48
 0 (0) 0 (0) 0 (0)     - | d = (4) functions.MultilayerPerceptron
```

```
Tester:     weka.experiment.PairedCorrectedTTester -G 3 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-p
Analysing:  Percent_correct
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       7/06/18 10:00 PM


>-<   >   < Resultset
  1   1   0 functions.MultilayerPerceptron
  0   0   0 trees.J48
  0   0   0 rules.JRip
 -1   0   1 rules.OneR
```



Una vez analizando los datos y la tabla ROC podemos ver que MultilayerPerceptron tiene mejor predicción a comparación de los otros 3 clasificadores, esto lo aseguramos ya que en la tabla ROC la gráfica de MultilayerPerceptron tiene mayor inclinación hacia la parte superior izquierda, en la tabla ROC se distingue a simple vista MultilayerPerceptron.

Conclusión:

En el paso 4 se puede observar que al generar MultilayerPerceptron no es el mejor clasificador, de echo tiene muy pocos aciertos, pero a la hora de comparar sus verdaderos positivos y sus falsos positivos, en la tabla ROC la cual es sacada por verdaderos positivos y sus falsos positivos nos indica que tanta certeza tiene al clasificar, y este algoritmo en particular resalto en todos los datasets. En general en estos datasets se tiene poco conocimiento de las relaciones entre los atributos y clases. También uno de los factores es que MultilayerPerceptron tiene alta tolerancia con valores perdidos. Por estas razones es que MultilayerPerceptron es el mejor algoritmo de predicción para este conjunto de datos.

Conclusiones para clientes:

Este conocimiento es bastante útil, pues con estos resultados se ha logrado crear una clasificación de los problemas cardiacos bastante homogénea con respecto a la cardinalidad de cada bloque de atributos. A pesar de que la red neuronal fue la más eficiente de todos los otros métodos de clasificación, esta tiene una desventaja de una complejidad de interpretación. Un éxito del 50% o menos quiere decir que estos sistemas de predicción diseñados son absolutamente inútiles, ya que no arrojan ninguna luz sobre el posible futuro de los pacientes sometidos al análisis, siendo esto así peor que tirar una moneda al aire. Su pobre rendimiento puede ser debido a factores diversos, como un mal preprocesador o una mala elección de los parámetros de los algoritmos.

Nota.

(a) El modelo generado en el Explorer, una vez que han ejecutado el algoritmo (se genera un archivo.model).

Estos archivos se encuentran en la carpeta Tareas-Heart, Tareas-Heart-C1, Tareas-Heart-C2, Tareas-Heart-C3, Tareas-Heart-C4.

(b) El archivo CSV generado, en el Experimentador, con los resultados.

Estos CSV se encuentran a simple vista y están enumerado conforme fueron creados.

(c) El flujo de conocimiento (Knowlege Flow) que genera en un archivo.kf

Estos archivos se encuentran en las carpetas; Mejor Algo 0, Mejor Algo 1, Mejor Algo 2, Mejor Algo 4. En estas se encontrará los Knowlege Flow y los archivos correspondientes a el Experiment.