
**UNIVERSIDAD DE GRANADA**

**MÁSTER DE CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES**

**CURSO ACADÉMICO 2019-2020**

**MINERÍA DE DATOS: APRENDIZAJE NO SUPERVISADO Y DETECCIÓN DE ANOMALÍAS**

# Minería de Datos con Reglas asociativas

*Análisis y extracción de conocimiento oculto de un dataset mediante Reglas Asociativas.*

**Nicolás Cubero Torres**

13 de Febrero de 2020

# Índice

# Índice de figuras

# Índice de tablas

# 1. Descripción del dataset: *Statlog (Heart)*

El *dataset* **Statlog (Heart)** [1] consituye una simplificación del *dataset* **heart dissease** [2]. Creada en 1988, recoge diferentes datos sanitarios sobre diferentes pacientes con la finalidad de predecir la presencia de enfermedades cardiacas.

El *dataset* original **heart dissease** recogía un total de 76 características de un total de 303 pacientes, la base de datos **Statlog (Heart)** considera sólo 13 características obtenidas a partir de las 76 características originales, midiendo la presencia o ausencia de enfermedades cardiacas con un único valor nominal binario. Este *dataset* considera sólo 270 instancias de las originales.

Ambos *datasets* fueron publicados en el repositorio de la *UCI Machine Learning*.

La descripción de los atributos se expone a continuación:

1. **age**: Valor numérico real que mide la edad del paciente.
2. **sex**: Valor categórico binario que representa el sexo del paciente: 1 para los hombres y 0 para las mujeres.
3. **chest pain type**: Valor categórico nominal que refleja el tipo de dolor de pecho que presenta el paciente: 1 para angina típica, 2 para angina atípica, 3 para no dolor de angina y 4 para asintomática
4. **resting blood pressure**: Valor real que representa la presión arterial en reposo.
5. **serum cholestoral in mg/dl**: Valor real que representa los niveles de colesterol sérico en mg/dl.
6. **fasting blood sugar > 120 mg/dl**: Valor nominal binario que determina si la glucemia en ayunas es superior a 120mg/dl (1) o no (0).
7. **resting electrocardiographic results**: Valor categórico nominal que refleja los resultados de las pruebas con electrocardiogramas en reposo:

[1] Enlace al repositorio de la UCI Machine learning sobre el dataset Statlog (Heart): https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart)

[2] Enlace al repositorio de la UCI Machine learning sobre el dataset Heart: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Adopta los valores 0 para un valor normal, 1 si se las pruebas revelan una anormalidad en las ondas ST-T y 2 si se muestra probable o segura una hipertrofía ventricular izquierda según el criterio de Estes.

8. **maximum heart rate achieved**: Valor real que representa la frecuencia cardiaca máxima alcanzable por el paciente.

9. **exercise induced angina**: Valor categórico binario que indica si el paciente realiza ejercicios que le pudieran inducir dolor de anginas o no.

10. **oldpeak**: Valor numérico real que refleja la depresión del segmento ST inducida por el ejercicio relativo al reposo.

11. **the slope of the peak exercise ST segment**: Valor categórico nominal que evalúa la pendiente del pico del segmento ST medido en ejercicio. Adopta los siguientes valores: 1 (ascendente), 2 (plana), 3 (descendente)

12. **number of major vessels (0-3) colored by flourosopy**: Valor numérico entero que representa el número de vasos principales coloreados por fluoroscopio. Comprende los valores entre 0 y 3 (ambos incluídos)

13. **thal**: Valor categórico nominal que representa un estado de salud normal (3), un defecto solucionado (6) o un defecto reversible (7).

Por último, este *dataset* asigna a cada paciente una **clase indicativa de la ausencia o presencia de enfermedades cardiacas** mediante un valor nominal que adopta los valores 1 o 2 respectivamente.

# 2. Transformación en transacciones de ítems

El *dataset* a tratar fue diseñado originalmente para un problema de clasificación, en el que se pretendía determinar la presencia o ausencia de enfermedades cardiacas en función de los otros datos sanitarios evaluados en cada ítem.

En este proyecto, por su parte, se persigue la aplicación de técnicas de minería mediante Reglas Asociativas extraídas sobre este *dataset*, para lo cual, el *dataset* requerirá su transformación en un conjunto de transacciones de ítems.

A continuación, y dadas las exigencias de este proyecto, se llevará a cabo un preprocesamiento de las variables de este *dataset* para su posterior transformación en ítems.
En especial, se requerirá la transformación de las variables reales en intervalos discretos y replicar algunos atributos nominales en binarios con la finalidad de obtener reglas que consideren estos atributos negados.

El proceso de transformación es el siguiente:

- *age*: Este atributo numérico entero se halla definido en el intervalo [29,77]. Para su tratamento con reglas asociativas conviene discretizarlo mediante su definición como un conjunto de intervalos.

  Para realizar esta división de forma significativa, se decide estudiar los estadísticos de posición más comunes con la finalidad de conocer cómo se distribuyen los datos:

| | age |
|---|---|
| Valor mínimo | 29 |
| Primer cuantil | 48 |
| Mediana | 55 |
| Media | 54.38 |
| Tercer cuantil | 61 |
| Valor máximo | 77 |

Tabla 1: Estadísticos de posición de *age*

  Se propone, teniendo en cuenta la semántica asociada a la edad de una persona, discretizar este atributo en los siguientes intervalos:

  - Adult: [29, 60).
  - Elderly: [60, 77].

```
summary(heart$age)

# Discretización del atributo age
heart[['age']] <- ordered(cut(heart[['age']],
                              c(29,60,+Inf),
```

Script 1: Conjunto de sentencias para discretizar el atributo age

- *resting blood pressure*: Este atributo numérico entero adopta valores en el intervalo [94,200], por lo que requiere ser discretizado en un conjunto de intervalos.
  Nuevamente, se estudian sus estadísticos de posición:

| | resting blood pressure |
|---|---|
| Valor mínimo | 94 |
| Primer cuantil | 120 |
| Mediana | 130 |
| Media | 131.3 |
| Tercer cuantil | 140 |
| Valor máximo | 200 |

Tabla 2: Estadísticos de posición de *resting blood pressure*

Consultando fuentes médicas específicas [3], se considera la discretización de la distribución en el siguiente conjunto de atributos:

- Normal: [94, 120)
- Elevated: [120, 130).
- Hypertension-stage1: [130, 140).
- Hypertension-stage2: [140, $+\infty$).

Si bien en el repositorio web de la UCI no se especifica el tipo de presión sanguínea evaluada (diastólica o sistólica), dado la distribución de valores, **se asuimirá que la presión medida es sistólica**

```
summary(heart[['resting blood pressure']])

# Discretizar el atributo resting blood presssure
heart[['resting blood pressure']] <- ordered(
                cut(heart[['resting blood pressure']],
                    c(94,120,130,140,+Inf),
                    labels=c('Normal', 'Elevated',
                             'Hypertension-stage1',
```

Script 2: Conjunto de sentencias para discretizar el atributo resting blood pressure

[3] Artículo de heart.org con información sobre la presión sanguínea y la división en intervalos, aceptada internacionalmente por la comunidad médica: https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

- *serum cholestoral*: Este atributo numérico entero adopta valores en el intervalo [126, 564], para su tratamiento mediante reglas asociativas, convendrá su transformación en intervalos.

  Nuevamente, se analizan los estadísticos de posición de este atributo:

| | serum cholestoral |
|---|---|
| Valor mínimo | 126 |
| Primer cuantil | 213 |
| Mediana | 245 |
| Media | 249.4 |
| Tercer cuantil | 277 |
| Valor máximo | 564 |

Tabla 3: Estadísticos de posición de *serum cholestoral*

  Teniendo en cuenta información médica específica sobre el nivel de colesterol total recomendados [4], se decide discretizar este atributo en función de los intervalos considerados por la comunidad médica:

  - Normal: [126,200).
  - High: [200, 240).
  - Dangerous: [240,+$\infty$).

```
summary(heart[['serum cholestoral']])

# Discretizar el atributo serum cholesterol
heart[['serum cholestoral']] <- ordered(
          cut(heart[['serum cholestoral']],
              c(126,200,240,+Inf),
              labels=c('Normal level',
                       'High level',
```

Script 3: Conjunto de sentencias para discretizar el atributo serum cholestoral

- *maximum heart rate achieved*: Este atributo numérico entero adopta valores en el intervalo [71,202], por lo que también debe de ser discretizado en un conjunto de atributos numéricos.
  Los estadísticos de posición de este atributo se recogen nuevamente en la tabla 4:

[4] Artículo en medlineplus.org con información médica sobre el colésterol sérico: https://medlineplus.gov/spanish/pruebas-de-laboratorio/niveles-de-colesterol/

| | maximum heart rate achieved |
|---|---|
| Valor mínimo | 71 |
| Primer cuantil | 113 |
| Mediana | 154 |
| Media | 149.8 |
| Tercer cuantil | 166 |
| Valor máximo | 202 |

Tabla 4: Estadísticos de posición de *maximum heart rate achieved*

Puesto que la semántica de este valor es asignado por la comunidad médica de acuerdo a diversos tests médicos y en función de otros parámetros como la edad, el nivel de actividad física de cada paciente además y otros factores, se decide discretizar esta variable en un conjunto de intervalos de forma conveniente según su distribución.

Se analiza más detalladamente la distribución de esta variable de forma gráfica haciendo uso de un histograma:

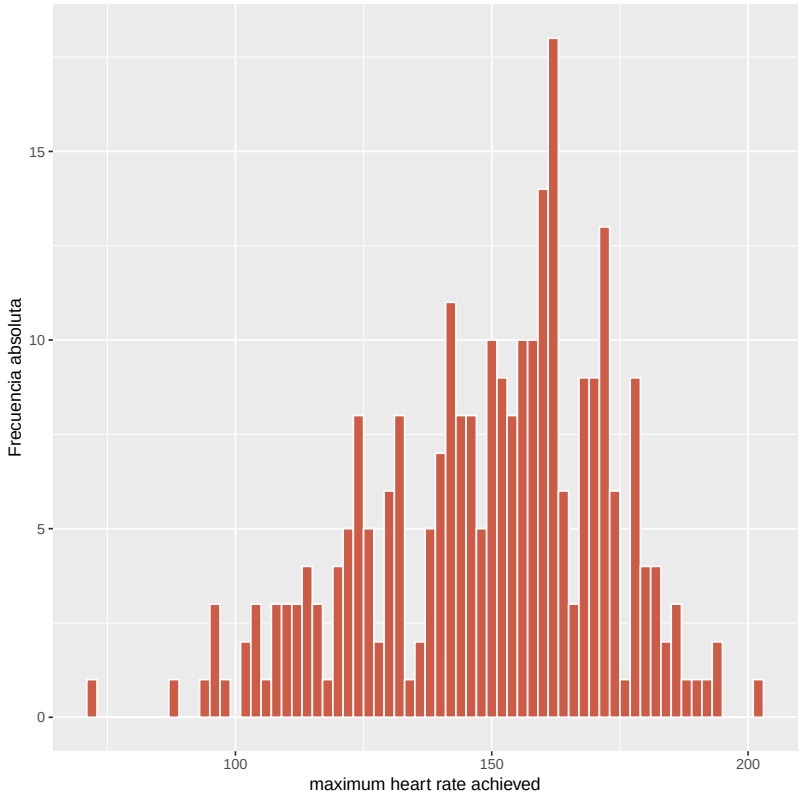


Figura 1: Histograma de la variable *maximum heart rate achieved* del dataset *Statlog (Heart)* considerando una amplitud de 2 unidades.

El anterior gráfico, a priori, no proporciona ninguna idea clara para la división del dominio. Con la finalidad de discretizar esta variable en intervalos que permitan obtener información relevante, se decide aplicar un método de **división en intervalos de igual frecuencia y se considera un número de intervalos igual a 4 intervalos**.

```
graf
dev.off()

# Discretizar en 4 intervalos
```

Script 4: Conjunto de sentencias para aplicar una discretización del atributo maximum heart rate achieved en 4 intervalos de igual frecuencia

El método nos permitió obtener los siguientes intervalos: [71,133), [133,154), [154,166) y [166,202]. En un principio, se aprecia que esta división podría resultar más o menos oportuna dada la semántica de este atributo.

- *oldpeak*: Este atributo numérico continuo toma valores en el intervalo [0, 6.2]. Para su tratamiento con reglas asociativas, nuevamente conviene discretizarlo en un conjunto de atributos.

  Estudiamos en primer lugar, sus estadísticos de posición, los cuales se resumen en la siguiente tabla:

| | oldpeak |
|---|---|
| Valor mínimo | 0 |
| Primer cuantil | 0 |
| Mediana | 0.8 |
| Media | 1.045 |
| Tercer cuantil | 1.6 |
| Valor máximo | 6.2 |

Tabla 5: Estadísticos de posición de *oldpeak*

  Por su parte, el descenso del segmento ST de forma aguda, se asocia con la presencia de daño miocárdico en cardiología [5], lo que implica que un breve incremento de este valor resulta significativo.

  Se propone analizar más detalladamente la distribución de valores de forma gráfica:

[5] Artículo de my-ekg.com sobre el segmento ST-T: https://www.my-ekg.com/como-leer-ekg/segmento-st.html

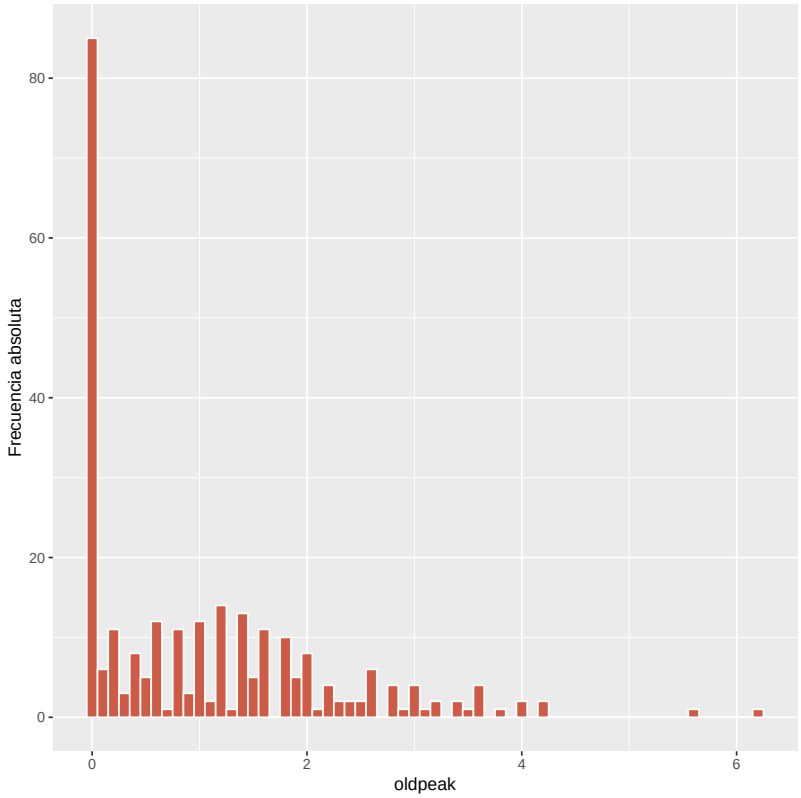


Figura 2: Histograma de la variable *oldpeak* del dataset *Statlog (Heart)*.

Se aprecia una concentración de la distribución muy pronunciado en torno al valor 0, de forma que, al incrementarse el valor de esta variable, la densidad de la distribución sufre una fuerte reducción, asemejándose (a grandes rasgos) a una reducción exponencial.
Se propone nuevamente, la discretización de esta variable en 3 intervalos de frecuencia similar:

```
  ggplot2::ylab('Frecuencia absoluta')
graf
dev.off()
```

Script 5: Conjunto de sentencias para aplicar una discretización del atributo oldpeak en 3 intervalos de igual frecuencia

Los intervalos considerados son los siguientes: [0,0.1), [0.1,1.4) y [1.4,6.2]. Observando el histograma de la distribución y teniendo en cuenta la semántica asociada a la variable, se considera que esta división resulta significativa.

- Por último, los atributos *fasting blood sugar*, *exercise induced angina* y *heart dissease* constituyen atributos categóricos binarios.
  Se decide considerar ítems positivos y negativos para los atributos *exercise induced angina* y *heart dissease*, mientras que el atributo *fasting blood sugar*, al presentar sólo 40 instancias con valor *True* frente a 229 instancias con valor *False*, se considera que resulta más significativo considerar unicamente ítems positivos para este atributo.

  Por consiguiente, para que los métodos de extracción de reglas a usar en este proyecto consideren ítems positivos y negativos para los atributos *exercise induced angina* y **heart dissease**, se requiere su transformación en el tipo de dato **factor**, mientras que el atributo **fasting blood sugar** se tratará como atributo binario.

Por último, para generar ítems descriptivos de los hechos que representan, se decide renombrar los valores numéricos nominales por cadenas que expresen de forma directa el hecho que representan:

- *thal*. Esta variable adoptaba los valores 3 para un estado de salud normal, 6 si se padece un defecto solucionado o 7 para un defecto de salud reversible. Estos valores serán renombrados por los siguientes de forma respectiva: *normal*, *Fixed defect* y *Reversible defect*.

- *resting electrocardiography result*: Esta variable indicaba con 0 unos resultados normales, 1 si las pruebas revelaban una anormalidad en las ondas ST-T y 2 si se muestra probable o segura una hipertrofía ventricular izquierda según el criterio de Estes. Estos valores van a ser renombrados respectivamente como sigue: *Normal*, *ST-T wave anormality* y *left ventricular hipertrophy*.

- *heart dissease*: Esta variable reflejaba con 1 la ausencia de enfermedades cardiacas y con 2 la presencia de enfermedades cardiacas. Estos valores serán renombrados a los valores booleanos FALSE y TRUE respectivamente.

- *chest pain type*: Asigna 1 para angina típica, 2 para angina atípica, 3 para no dolor de angina y 4 para asintomática. Estos valores se han renombrado respectivamente por lo siguientes: *non-anginal pain*, *typical angina*, *atypical angina* y *asymptomatic*.

- *slope of the peak exercise ST segment*: Esta variable adopta los valores 1 para pendiente ascendente, 2 para pendiente plana, 3 para pendiente

descendente. Estos valores pasan a llamarse respectivamente *Upsloping Flat Downsloping*.

- *fasting blood sugar > 120 mg/dl:*: Esta variable asignaba 1 si la condición que refleja es cierta y 0 si es falsa. Estos valores han sido renombrados por TRUE y FALSE respectivamente.

# 3. Búsqueda de reglas asociativas

El *dataset* resultante que, tras el preprocesamiento realizado en la anterior sección contiene 19 atributos, es convertido en un conjunto de transacciones de *itemsets* con el siguiente resumen:

```
transactions as itemMatrix in sparse format with
 270 rows (elements/itemsets/transactions) and
 40 columns (items) and a density of 0.3287037

most frequent items:
  age=Adult sex=male exercise induced angina=FALSE number of major vessels=0
thal=Normal (Other)
186 183 181 160 152 2688

element (itemset/transaction) length distribution:
sizes
 13  14
230  40

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.00   13.00   13.00   13.15   13.00   14.00

includes extended item information - examples:
       labels variables  levels
1   age=Adult       age   Adult
2 age=Elderly       age Elderly
3    sex=male       sex    male

includes extended transaction information - examples:
  transactionID
1             1
2             2
3             3
```

Del anterior conjunto de transacciones, destacamos los siguientes hechos:

- La longitud de las transacciones oscila entre valores de 13 y 14 ítems. El *dataset* original no presentaba valores perdidos, por lo que la causa de esta variación se explica por la presencia o ausencia del atributo *fasting blood sugar*, que aparece sólo si adoptaba el valor TRUE en alguna instancia del *dataset* original.

- Los ítems más frecuentes son los siguientes: *age=Adult*, *sex=male*, *exercise induced angina=FALSE*, *number of major vessels=0* y *thal=Normal*.

## 3.1. Análisis de la frecuencia de los ítems

Como paso previo a la búsqueda de itemsets frecuentes, se pretende analizar el soporte de todos los ítems del *dataset*, lo cual en este problema resulta viable, puesto que el *dataset* incluye únicamente 14 ítems.

Este análisis permitirá, de este modo, obtener una referencia de los soportes más adecuados a considerar como umbrales mínimos en la búsqueda de itemsets frecuentes, puesto que cualquier itemset frecuente es un superconjunto de los ítems del conjunto de transacciones.

En la tabla 8, se expone el análisis de los soportes de cada ítem:

| Ítem | Soporte | Ocurrencias |
|---|---|---|
| age=Adult | 0.688888889 | 186 |
| sex=male | 0.677777778 | 183 |
| exercise induced angina=FALSE | 0.670370370 | 181 |
| number of major vessels=0 | 0.592592593 | 160 |
| thal=Normal | 0.562962963 | 152 |
| heart dissease=FALSE | 0.555555556 | 150 |
| serum cholestoral=Dangerous level | 0.544444444 | 147 |
| resting electrocardiographic results=left ventricular hypertrophy | 0.507407407 | 137 |
| resting electrocardiographic results=Normal | 0.485185185 | 131 |
| slope of the peak exercise ST segment=Upsloping | 0.481481481 | 130 |
| chest pain type=asymptomatic | 0.477777778 | 129 |
| slope of the peak exercise ST segment=Flat | 0.451851852 | 122 |
| heart dissease=TRUE | 0.444444444 | 120 |
| thal=Reversable defect | 0.385185185 | 104 |
| oldpeak=[1.4,6.2] | 0.355555556 | 96 |
| exercise induced angina=TRUE | 0.329629630 | 89 |

| | | |
|---|---|---|
| oldpeak=[0.1,1.4) | 0.329629630 | 89 |
| resting blood pressure=Hypertension-stage2 | 0.325925926 | 88 |
| sex=female | 0.322222222 | 87 |
| oldpeak=[0,0.1) | 0.314814815 | 85 |
| age=Elderly | 0.311111111 | 84 |
| serum cholestoral=High level | 0.311111111 | 84 |
| chest pain type=non-anginal pain | 0.292592593 | 79 |
| maximum heart rate achieved=[166,202] | 0.255555556 | 69 |
| maximum heart rate achieved=[133,154) | 0.251851852 | 68 |
| maximum heart rate achieved=[71,133) | 0.248148148 | 67 |
| resting blood pressure=Elevated | 0.244444444 | 66 |
| maximum heart rate achieved=[154,166) | 0.244444444 | 66 |
| resting blood pressure=Hypertension-stage1 | 0.218518519 | 59 |
| number of major vessels=1 | 0.214814815 | 58 |
| resting blood pressure=Normal | 0.211111111 | 57 |
| chest pain type=atypical angina | 0.155555556 | 42 |
| fasting blood sugar | 0.148148148 | 40 |
| serum cholestoral=Normal level | 0.144444444 | 39 |
| number of major vessels=2 | 0.122222222 | 33 |
| chest pain type=typical angina | 0.074074074 | 20 |
| number of major vessels=3 | 0.070370370 | 19 |
| slope of the peak exercise ST segment=Downsloping | 0.066666667 | 18 |
| thal=Fixed defect | 0.051851852 | 14 |
| resting electrocardiographic results=ST-T wave anormality | 0.007407407 | 2 |

Tabla 6: Medidas de soporte y número de apariciones para cada uno de los ítems del *dataset*, los cuales han sido expuestos en orden descendiente de estas medidas

Se proporciona además, un diagrama de barras con los soportes de los ítems:

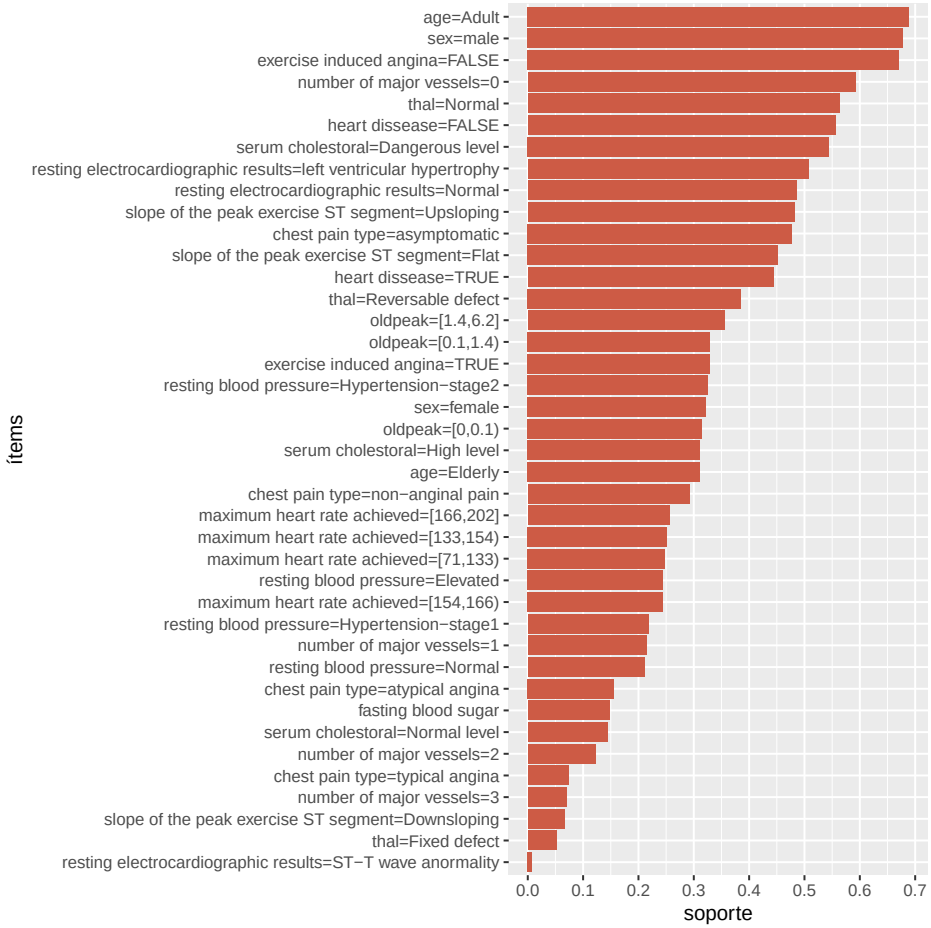


Figura 3: Diagrama de barras con la representación gráfica de los soportes de todos los posibles ítems.

De este modo, se deduce que cualquier itemset que se forme a partir de estos ítems, tendrá un soporte igual o inferior a 0.688888889.

Por su parte, teniendo en cuenta la información mostrada en el anterior diagrama, se propone utilizar el siguiente conjunto de umbrales mínimos de soporte en función de la cantidad de itemsets frecuentes que se deseen explorar: {0.6, 0.5, 0.4, 0.3, 0.2, 0.1 y 0.05}.

## 3.2. Búsqueda de itemsets frecuentes

Tomando como referencia los soportes de los ítems individuales, realizamos diferentes análisis exploratorios de itemsets frecuentes considerando los umbrales propuestos anteiormente:

- **Considerando un soporte mínimo de 0.5**: Se realizaron las siguientes acciones:

```
dev.off()

# Extracción de itemsets frecuentes  minSupport de 0.5
```

Script 6: Conjunto de sentencias para aplicar una búsqueda de itemsets frecuentes mediante el método A priori considerando un soporte mínimo de 0.5

La anterior búsqueda permitió obtener únicamente los siguientes 9 itemsets frecuentes:

| Itemset | Soporte | Ocurrencias |
|---|---|---|
| age=Adult | 0.6888889 | 186 |
| sex=male | 0.6777778 | 183 |
| exercise induced angina=FALSE | 0.6703704 | 181 |
| number of major vessels=0 | 0.5925926 | 160 |
| thal=Normal | 0.5629630 | 152 |
| heart dissease=FALSE | 0.5555556 | 150 |
| serum cholestoral=Dangerous level | 0.5444444 | 147 |
| resting electrocardiographic results=left ventricular hypertrophy | 0.5074074 | 137 |
| age=Adult, sex=male | 0.5037037 | 136 |

Tabla 7: Medidas de soporte y número de apariciones para cada uno de los ítems frecuentes que aparecen con un soporte mínimo de 0.5

De estos itemsets, sólo el último presenta más de 1 ítem, por lo que se considera, que este umbral de mínimo soporte es demasiado elevado para generar reglas a partir de él.

- **Considerando un soporte mínimo de 0.3**: Esta configuración, permitió obtener un número mayor de itemsets, haciendo un total de 79 itemsets frecuentes.
  Se analizan las medidas resúmen de este conjunto de itemsets:

```
set of 79 itemsets

most frequent items:
exercise induced angina=FALSE age=Adult heart dissease=FALSE number of major vess
thal=Normal (Other)
22 21 18 17 16 62
```

```
element (itemset/transaction) length distribution:sizes
 1  2  3  4
22 39 16  2

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   1.975   2.000   4.000

summary of quality measures:
    support           count
 Min.   :0.3000   Min.   : 81.0
 1st Qu.:0.3222   1st Qu.: 87.0
 Median :0.3481   Median : 94.0
 Mean   :0.3866   Mean   :104.4
 3rd Qu.:0.4426   3rd Qu.:119.5
 Max.   :0.6889   Max.   :186.0

includes transaction ID lists: FALSE

mining info:
  data ntransactions support confidence
 heart           270     0.3          1
```

Se observa que, en este caso se parte de un buen conjunto de itemsets frecuentes para analizar.

- **Considerando un soporte mínimo de 0.2**: Se obtiene un total de 280 itemsets con las siguientes medidas resumen:

```
set of 280 itemsets

most frequent items:
exercise induced angina=FALSE age=Adult  heart dissease=FALSE
                           85        76                    68

number of major vessels=0  sex=male (Other)
                      62        59     351

element (itemset/transaction) length distribution:sizes
  1   2   3   4   5
 31 113 104  28   4

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
  1.000   2.000   2.000   2.504   3.000   5.000

summary of quality measures:
    support           count
 Min.   :0.2000   Min.   : 54.00
 1st Qu.:0.2148   1st Qu.: 58.00
 Median :0.2444   Median : 66.00
 Mean   :0.2758   Mean   : 74.48
 3rd Qu.:0.3111   3rd Qu.: 84.00
 Max.   :0.6889   Max.   :186.00

includes transaction ID lists: FALSE

mining info:
  data ntransactions support confidence
 heart           270     0.2          1
```

Para este umbral mínimo de soporte, se ha obtenido un número considerable de itemsets que podrían dar lugar a un conjunto interesante de reglas, dado que este umbral es de por sí, relativamente bajo, se **se decide usar este conjunto de itemsets frecuentes como punto de partida en la búsqueda de reglas de interés**.

En el anterior conjunto de itemsets, las longitudes de los itemsets varían entre 1 y 5 ítems. Para conocer y comparar más detalladamente la cantidad de itemsets que poseen un número determinado cada longitud de ítems, en la siguiente figura 4 se representa un diagrama de barras de la proporción de ítemsets de cada longitud de ítem determinada:

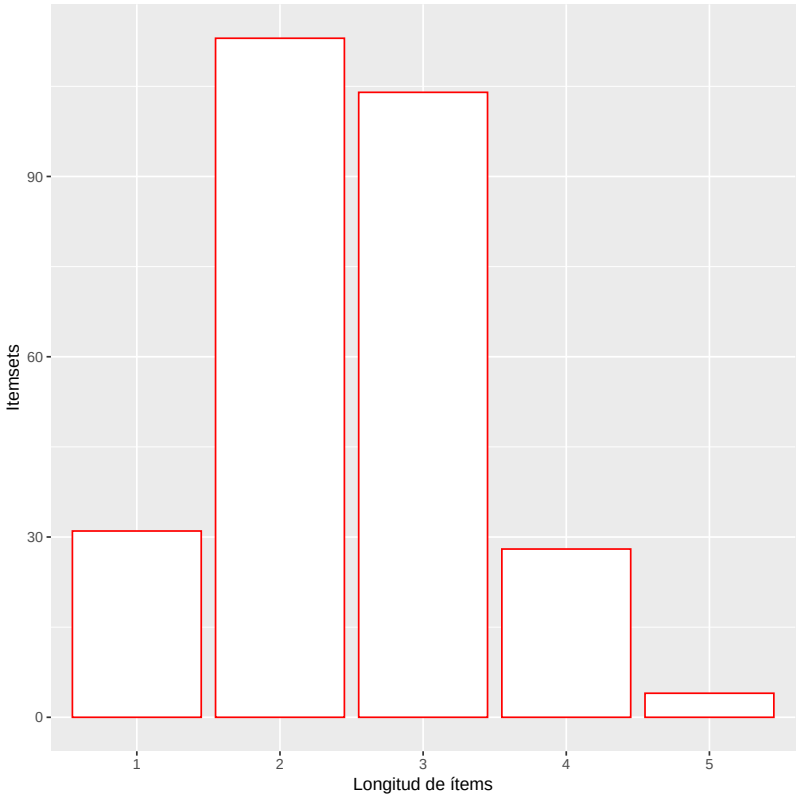


Figura 4: Diagrama de barras que muestra las cantidad de ítems que posee una longitud de ítem determinadas

## 3.3. Búsqueda de reglas de alta confianza

La búsqueda de reglas asociativas se enfocará en primer lugar, en el análisis de reglas con alta confianza a partir del conjunto de itemsets frecuentes seleccionado que se irá reduciendo según convenga:

- **Considerando una confianza mínima de 0.9**: Se obtuvieron un total de 24 reglas asociativas con los siguientes resultados resumen:

```
set of 24 rules

rule length distribution (lhs + rhs):sizes
 3  4  5
 9 11  4

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
  3.000   3.000   4.000   3.792   4.000   5.000

summary of quality measures:
    support         confidence          lift           count
 Min.   :0.2000   Min.   :0.9000   Min.   :1.332   Min.   :54.00
 1st Qu.:0.2102   1st Qu.:0.9024   1st Qu.:1.384   1st Qu.:56.75
 Median :0.2241   Median :0.9098   Median :1.654   Median :60.50
 Mean   :0.2281   Mean   :0.9189   Mean   :1.639   Mean   :61.58
 3rd Qu.:0.2380   3rd Qu.:0.9262   3rd Qu.:1.762   3rd Qu.:64.25
 Max.   :0.3037   Max.   :0.9701   Max.   :2.025   Max.   :82.00

mining info:
  data ntransactions support confidence
 heart           270     0.2        0.9
```

El **soporte máximo alcanzado por las reglas obtenidas es de 0.3037, mientras que la confianza máxima alcanzada es de 0.9701**.

El conjunto de reglas encontrado se expone a continuación, junto con las siguientes medidas de calidad: soporte, confianza, lift, Confianza confirmada y Convicción:

| Regla | Soporte | Confianza | lift | Conteo | Convicción | Conf. Confirmada |
|---|---|---|---|---|---|---|
| exercise induced angina=FALSE, number of major vessels=0, thal=Normal ⇒ heart dissease=FALSE | 0.3037 | 0.9011 | 1.622 | 82 | 4.4938 | 0.8022 |
| age=Adult, number of major vessels=0, thal=Normal ⇒ heart dissease=FALSE | 0.2889 | 0.9286 | 1.6714 | 78 | 6.2222 | 0.8571 |
| age=Adult, thal=Reversable defect ⇒ sex=male | 0.2481 | 0.9437 | 1.3923 | 67 | 5.7194 | 0.8873 |
| sex=female, heart dissease=FALSE ⇒ thal=Normal | 0.2407 | 0.9701 | 1.7233 | 65 | 14.6407 | 0.9403 |
| age=Adult, heart dissease=TRUE ⇒ sex=male | 0.2407 | 0.9028 | 1.332 | 65 | 3.3143 | 0.8056 |
| age=Adult, slope of the peak exercise ST segment=Upsloping, thal=Normal ⇒ heart dissease=FALSE | 0.237 | 0.9014 | 1.6225 | 64 | 4.5079 | 0.8028 |
| age=Adult, slope of the peak exercise ST segment=Upsloping, thal=Normal ⇒ exercise induced angina=FALSE | 0.237 | 0.9014 | 1.3446 | 64 | 3.3434 | 0.8028 |
| chest pain type=asymptomatic, thal=Reversable defect ⇒ heart dissease=TRUE | 0.2333 | 0.9 | 2.025 | 63 | 5.5556 | 0.8 |
| exercise induced angina=FALSE, oldpeak=[0,0.1) ⇒ slope of the peak exercise ST segment=Upsloping | 0.2296 | 0.9254 | 1.9219 | 62 | 6.9481 | 0.8507 |
| slope of the peak exercise ST segment=Upsloping, number of major vessels=0,thal=Normal ⇒ heart dissease=FALSE | 0.2296 | 0.9254 | 1.6657 | 62 | 5.9556 | 0.8507 |
| slope of the peak exercise ST segment=Upsloping, number of major vessels=0,thal=Normal ⇒ exercise induced angina=FALSE | 0.2259 | 0.9104 | 1.3581 | 61 | 3.6809 | 0.8209 |

| | | | | | | |
|---|---|---|---|---|---|---|
| oldpeak=[0,0.1),thal=Normal ⇒ slope of the peak exercise ST segment=Upsloping | 0.2222 | 0.9091 | 1.8881 | 60 | 5.7037 | 0.8182 |
| exercise induced angina=TRUE, heart dissease=TRUE ⇒ chest pain type=asymptomatic | 0.2222 | 0.9091 | 1.9027 | 60 | 5.7444 | 0.8182 |
| resting electrocardiographic results=Normal, thal=Normal, heart dissease=FALSE ⇒ exercise induced angina=FALSE | 0.2148 | 0.9062 | 1.3519 | 58 | 3.516 | 0.8125 |
| oldpeak=[0,0.1),heart dissease=FALSE ⇒ slope of the peak exercise ST segment=Upsloping | 0.2111 | 0.9048 | 1.8791 | 57 | 5.4444 | 0.8095 |
| sex=female, number of major vessels=0 ⇒ thal=Normal | 0.2 | 0.9153 | 1.6258 | 54 | 5.157 | 0.8305 |
| age=Adult, exercise induced angina=FALSE, slope of the peak exercise ST segment=Upsloping, number of major vessels=0 ⇒ heart dissease=FALSE | 0.2 | 0.9 | 1.62 | 54 | 4.4444 | 0.8 |

Tabla 8: Conjuntos de reglas soporte superior a 0.2 y confianza superior a 0.9. Para cada regla se muestra su valor de soporte, confianza, lift, conteo, convicción y confianza confirmada

Este conjunto de reglas, presentan una confianza elevada máxima de 0.9701, se plantea ahora, la búsqueda de reglas con confianza cercana a 1 relajando el umbral mínimo de soporte.

- **Considerando una confianza mínima de 0.97 y un soporte máximo de 0.2**: Se obtuvieron un total de 68 reglas con las siguientes características resumen:

```
set of 68 rules

rule length distribution (lhs + rhs):sizes
 3  4  5  6  7
 6 27 26  8  1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   5.000   4.574   5.000   7.000

summary of quality measures:
    support         confidence          lift           count
 Min.   :0.1000   Min.   :0.9706   Min.   :1.432   Min.   :27.00
 1st Qu.:0.1111   1st Qu.:0.9737   1st Qu.:1.475   1st Qu.:30.00
 Median :0.1222   Median :1.0000   Median :1.754   Median :33.00
 Mean   :0.1244   Mean   :0.9894   Mean   :1.743   Mean   :33.59
 3rd Qu.:0.1370   3rd Qu.:1.0000   3rd Qu.:1.800   3rd Qu.:37.00
 Max.   :0.1630   Max.   :1.0000   Max.   :2.250   Max.   :44.00

mining info:
  data ntransactions support confidence
 heart           270     0.1       0.97
```

Aplicando eliminación de reglas redundantes, el conjunto queda reducido a 29 reglas que se muestran a continuación:

| Reglas | Soporte | Confianza | Lift | Conteo | Convicción | Conf. Confirmada |
|---|---|---|---|---|---|---|
| number of major vessels=1,heart dissease=TRUE ⇒ sex=male | 0.137 | 0.9737 | 1.4366 | 37 | 12.2444 | 0.9474 |
| resting blood pressure=Elevated, thal=Reversable defect ⇒ sex=male | 0.1111 | 1 | 1.4754 | 30 | - | 1 |
| resting blood pressure=Elevated, heart dissease=TRUE ⇒ sex=male | 0.1148 | 1 | 1.4754 | 31 | - | 1 |
| sex=female, oldpeak=[0,0.1) ⇒ thal=Normal | 0.1148 | 1 | 1.7763 | 31 | - | 1 |
| sex=female, slope of the peak exercise ST segment=Upsloping ⇒ thal=Normal | 0.163 | 0.9778 | 1.7368 | 44 | 19.6667 | 0.9556 |
| slope of the peak exercise ST segment=Upsloping, thal=Reversable defect ⇒ sex=male | 0.1222 | 0.9706 | 1.432 | 33 | 10.9556 | 0.9412 |
| maximum heart rate achieved=[71,133),exercise induced angina=TRUE, heart dissease=TRUE ⇒ sex=male | 0.1296 | 0.9722 | 1.4344 | 35 | 11.6 | 0.9444 |
| chest pain type=asymptomatic, maximum heart rate achieved=[71,133),exercise induced angina=TRUE ⇒ sex=male | 0.1296 | 0.9722 | 1.4344 | 35 | 11.6 | 0.9444 |
| sex=male, maximum heart rate achieved=[166,202],slope of the peak exercise ST segment=Upsloping ⇒ age=Adult | 0.1296 | 1 | 1.4516 | 35 | - | 1 |
| maximum heart rate achieved=[166,202],number of major vessels=0,thal=Normal ⇒ heart dissease=FALSE | 0.1556 | 0.9767 | 1.7581 | 42 | 19.1111 | 0.9535 |
| age=Adult, maximum heart rate achieved=[166,202],thal=Normal ⇒ heart dissease=FALSE | 0.163 | 0.9778 | 1.76 | 44 | 20 | 0.9556 |

| | | | | | | |
|---|---|---|---|---|---|---|
| sex=male, maximum heart rate achieved=[166,202],heart dissease=FALSE ⇒ age=Adult | 0.1333 | 1 | 1.4516 | 36 | - | 1 |
| sex=male, maximum heart rate achieved=[166,202],number of major vessels=0 ⇒ age=Adult | 0.137 | 1 | 1.4516 | 37 | - | 1 |
| sex=female, chest pain type=non-anginal pain, thal=Normal ⇒ heart dissease=FALSE | 0.1074 | 1 | 1.8 | 29 | - | 1 |
| serum cholestoral=Dangerous level, exercise induced angina=FALSE, oldpeak=[0,0.1) ⇒ slope of the peak exercise ST segment=Upsloping | 0.1222 | 0.9706 | 2.0158 | 33 | 17.6296 | 0.9412 |
| sex=male, oldpeak=[0,0.1),thal=Normal ⇒ slope of the peak exercise ST segment=Upsloping | 0.1259 | 0.9714 | 2.0176 | 34 | 18.1481 | 0.9429 |
| sex=female, resting electrocardiographic results=Normal, heart dissease=FALSE ⇒ thal=Normal | 0.1333 | 0.973 | 1.7283 | 36 | 16.1704 | 0.9459 |
| age=Adult, sex=female, heart dissease=FALSE ⇒ thal=Normal | 0.1593 | 1 | 1.7763 | 43 | - | 1 |
| age=Adult, sex=female, exercise induced angina=FALSE ⇒ heart dissease=FALSE | 0.137 | 0.9737 | 1.7526 | 37 | 16.8889 | 0.9474 |
| age=Adult, sex=female, exercise induced angina=FALSE ⇒ thal=Normal | 0.1407 | 1 | 1.7763 | 38 | - | 1 |
| exercise induced angina=TRUE, oldpeak=[1.4,6.2],thal=Reversable defect ⇒ heart dissease=TRUE | 0.1222 | 0.9706 | 2.1838 | 33 | 18.8889 | 0.9412 |
| serum cholestoral=Dangerous level, exercise induced angina=TRUE, oldpeak=[1.4,6.2] ⇒ heart dissease=TRUE | 0.1 | 1 | 2.25 | 27 | - | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| oldpeak=[0.1,1.4),number of major vessels=0,thal=Normal ⇒ heart dissease=FALSE | 0.1259 | 0.9714 | 1.7486 | 34 | 15.5556 | 0.9429 |
| chest pain type=asymptomatic, oldpeak=[1.4,6.2],thal=Reversable defect ⇒ heart dissease=TRUE | 0.137 | 1 | 2.25 | 37 | - | 1 |
| chest pain type=asymptomatic, resting electrocardiographic results=left ventricular hypertrophy, thal=Reversable defect ⇒ heart dissease=TRUE | 0.1407 | 0.9744 | 2.1923 | 38 | 21.6667 | 0.9487 |
| chest pain type=asymptomatic, serum cholestoral=Dangerous level, thal=Reversable defect ⇒ heart dissease=TRUE | 0.1407 | 0.9744 | 2.1923 | 38 | 21.6667 | 0.9487 |
| age=Adult, resting electrocardiographic results=Normal, thal=Reversable defect ⇒ sex=male | 0.1296 | 0.9722 | 1.4344 | 35 | 11.6 | 0.9444 |
| age=Adult, exercise induced angina=FALSE, thal=Reversable defect ⇒ sex=male | 0.1296 | 1 | 1.4754 | 35 | - | 1 |
| age=Adult, resting electrocardiographic results=Normal, slope of the peak exercise ST segment=Upsloping, number of major vessels=0, thal=Normal ⇒ heart dissease=FALSE | 0.1222 | 0.9706 | 1.7471 | 33 | 15.1111 | 0.9412 |

Tabla 9: Conjuntos de reglas con confianza superior a 0.9706 y soporte comprendido entre 0.1 y un máximo de 0.2. Para cada regla se muestra su valor de soporte, confianza, lift, conteo, convicción y confianza confirmada. Nótese que para las reglas con Confianza 1, no es posible calcular su valor de Convicción

En las dos anteriores búsquedas de reglas ejecutadas, pese a haber exigido y obtenido un conjunto de reglas con una confianza elevada, el soporte más alto logrado por el conjunto de reglas es de 0.3037 en la primera búsqueda y 0.1607 en la última búsqueda. Por tanto, resulta de interés, la posibilidad de encontrar otro conjunto de reglas con un soporte mayor considerando un umbral mínimo de confianza menor.

- **Considerando un soporte mínimo de 0.3037 y una confianza mínima de 0.6**: Se relaja el umbral de confianza mínima exigiendo un soporte mayor, dando lugar al siguiente conjunto de 104 reglas con las siguientes características resumen:

```
set of 104 rules

rule length distribution (lhs + rhs):sizes
 2  3  4
51 45  8

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.587   3.000   4.000

summary of quality measures:
    support         confidence          lift           count
 Min.   :0.3037   Min.   :0.6120   Min.   :0.913   Min.   : 82.00
 1st Qu.:0.3185   1st Qu.:0.6802   1st Qu.:1.081   1st Qu.: 86.00
 Median :0.3481   Median :0.7549   Median :1.230   Median : 94.00
 Mean   :0.3602   Mean   :0.7502   Mean   :1.228   Mean   : 97.25
 3rd Qu.:0.3815   3rd Qu.:0.8099   3rd Qu.:1.357   3rd Qu.:103.00
 Max.   :0.5037   Max.   :0.9011   Max.   :1.622   Max.   :136.00

mining info:
  data ntransactions support confidence
 heart           270  0.3037        0.6
```

Al anterior conjunto de reglas se le aplica un proceso de eliminación de reglas redundantes:

El conjunto de reglas queda reducido a 23 reglas:

| Regla | Soporte | Confianza | lift | Conteo | Convicción | Conf. Confirmada |
|---|---|---|---|---|---|---|
| thal=Reversable defect ⇒ sex=male | 0.337 | 0.875 | 1.291 | 91 | 2.5778 | 0.75 |
| heart dissease=TRUE ⇒ sex=male | 0.3704 | 0.8333 | 1.2295 | 100 | 1.9333 | 0.6667 |
| slope of the peak exercise ST segment=Flat ⇒ sex=male | 0.3148 | 0.6967 | 1.0279 | 85 | 1.0625 | 0.3934 |
| chest pain type=asymptomatic ⇒ sex=male | 0.3481 | 0.7287 | 1.0751 | 94 | 1.1876 | 0.4574 |
| chest pain type=asymptomatic ⇒ age=Adult | 0.3037 | 0.6357 | 0.9227 | 82 | 0.8539 | 0.2713 |
| slope of the peak exercise ST segment=Upsloping ⇒ number of major vessels=0 | 0.3185 | 0.6615 | 1.1163 | 86 | 1.2037 | 0.3231 |
| slope of the peak exercise ST segment=Upsloping ⇒ exercise induced angina=FALSE | 0.3889 | 0.8077 | 1.2048 | 105 | 1.7141 | 0.6154 |
| slope of the peak exercise ST segment=Upsloping ⇒ sex=male | 0.3148 | 0.6538 | 0.9647 | 85 | 0.9309 | 0.3077 |
| slope of the peak exercise ST segment=Upsloping ⇒ age=Adult | 0.363 | 0.7538 | 1.0943 | 98 | 1.2639 | 0.5077 |
| resting electrocardiographic results=Normal ⇒ heart dissease=FALSE | 0.3148 | 0.6489 | 1.1679 | 85 | 1.2657 | 0.2977 |
| resting electrocardiographic results=Normal ⇒ number of major vessels=0 | 0.3111 | 0.6412 | 1.0821 | 84 | 1.1355 | 0.2824 |
| resting electrocardiographic results=Normal ⇒ exercise induced angina=FALSE | 0.3481 | 0.7176 | 1.0704 | 94 | 1.1671 | 0.4351 |
| resting electrocardiographic results=Normal ⇒ sex=male | 0.3222 | 0.6641 | 0.9799 | 87 | 0.9593 | 0.3282 |
| resting electrocardiographic results=Normal ⇒ age=Adult | 0.3481 | 0.7176 | 1.0416 | 94 | 1.1015 | 0.4351 |

| | | | | | | |
|---|---|---|---|---|---|---|
| resting electrocardiographic results=left ventricular hypertrophy ⇒ serum cholestoral=Dangerous level | 0.3259 | 0.6423 | 1.1798 | 88 | 1.2737 | 0.2847 |
| resting electrocardiographic results=left ventricular hypertrophy ⇒ exercise induced angina=FALSE | 0.3185 | 0.6277 | 0.9364 | 86 | 0.8855 | 0.2555 |
| resting electrocardiographic results=left ventricular hypertrophy ⇒ sex=male | 0.3556 | 0.7007 | 1.0339 | 96 | 1.0767 | 0.4015 |
| resting electrocardiographic results=left ventricular hypertrophy ⇒ age=Adult | 0.337 | 0.6642 | 0.9642 | 91 | 0.9266 | 0.3285 |
| serum cholestoral=Dangerous level ⇒ exercise induced angina=FALSE | 0.3407 | 0.6259 | 0.9336 | 92 | 0.881 | 0.2517 |
| serum cholestoral=Dangerous level ⇒ sex=male | 0.337 | 0.619 | 0.9133 | 91 | 0.8458 | 0.2381 |
| serum cholestoral=Dangerous level ⇒ age=Adult | 0.3444 | 0.6327 | 0.9184 | 93 | 0.8469 | 0.2653 |
| thal=Normal ⇒ age=Adult | 0.3926 | 0.6974 | 1.0123 | 106 | 1.028 | 0.3947 |
| number of major vessels=0 ⇒ sex=male | 0.3741 | 0.6312 | 0.9314 | 101 | 0.8738 | 0.2625 |

Tabla 10: Conjuntos de reglas obtenidos considerando un soporte mínimo de 0.5 y una confianza mínima de 0.6, para las cuales, se indica su valor de soporte, confianza y lift

Para el anterior conjunto de reglas se representa en la figura 5, un diagrama de puntos (*scatterplot*) donde se relacionan los valores de **confianza** y **soporte** de las reglas:

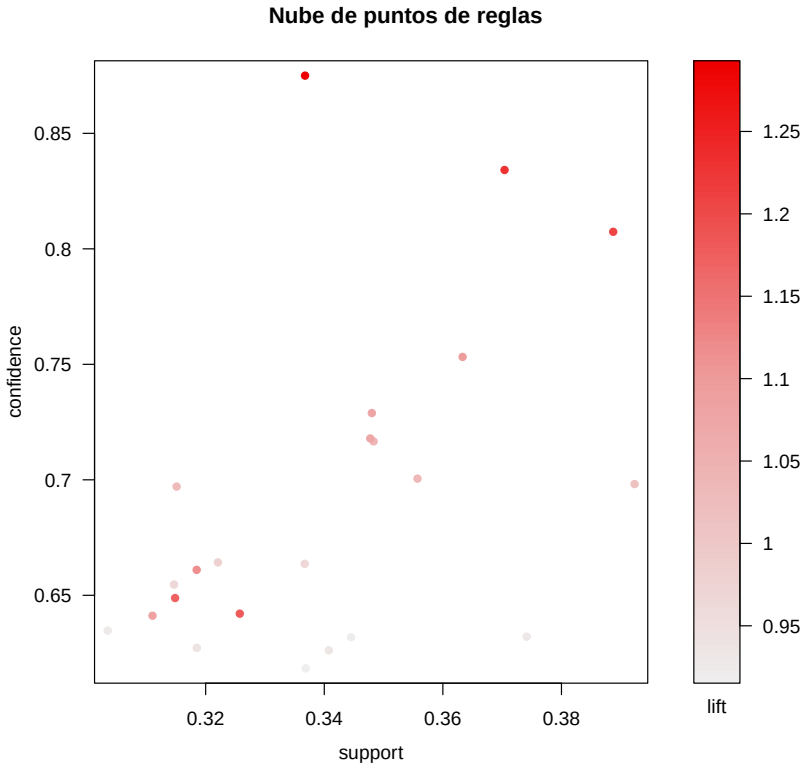


Figura 5: Diagrama de puntos donde para cada punto (regla de asociación) se indica su valor de confianza (eje de ordenadas) frente a su valor de soporte (eje de abscisas), la coloración cada punto indica el grado de lift de la regla

Con ayuda del anterior diagrama se identifican como reglas con mayor relación confianza-soporte las siguientes:

| Regla | Soporte | Confianza | lift | Conteo | Convicción | Conf. Confirmada |
|---|---|---|---|---|---|---|
| thal=Reversable defect ⇒ sex=male | 0.337 | 0.875 | 1.291 | 91 | 2.5778 | 0.75 |
| heart dissease=TRUE ⇒ sex=male | 0.3704 | 0.8333 | 1.2295 | 100 | 1.9333 | 0.6667 |
| slope of the peak exercise ST segment=Upsloping ⇒ exercise induced angina=FALSE | 0.3889 | 0.8077 | 1.2048 | 105 | 1.7141 | 0.6154 |
| slope of the peak exercise ST segment=Upsloping ⇒ age=Adult | 0.363 | 0.7538 | 1.0943 | 98 | 1.2639 | 0.5077 |

Tabla 11: Selección de reglas con mayor relación confianza y soporte del anterior conjunto de reglas

# 4. Análisis de reglas por grupos

Una vez identificados conjuntos de reglas de interés se procede a analizar con mayor detalle el conjunto de reglas realizando un análisis por grupos de las reglas, lo cual va a permitir ampliar la información semántica reflejada por el conjunto de reglas, así como obtener más información sobre cada regla individual:

## 4.1. Conjunto de reglas que determinan la ausencia de enfermedades cardiacas

En el siguiente conjunto de reglas se indica una tendencia a la ausencia enfermedades cardiacas (*heart dissease=FALSE*):

| Regla | Soporte | Confianza | Lift | Confianza Confirmada |
|---|---|---|---|---|
| exercise induced angina=FALSE, number of major vessels=0,thal=Normal ⇒ heart dissease=FALSE | 0.3037 | 0.9011 | 1.622 | 0.8022 |
| age=Adult, number of major vessels=0,thal=Normal ⇒ heart dissease=FALSE | 0.2889 | 0.9286 | 1.6714 | 0.8571 |
| age=Adult, slope of the peak exercise ST segment=Upsloping, thal=Normal ⇒ heart dissease=FALSE | 0.237 | 0.9014 | 1.6225 | 0.8028 |
| slope of the peak exercise ST segment=Upsloping, number of major vessels=0,thal=Normal ⇒ heart dissease=FALSE | 0.2296 | 0.9254 | 1.6657 | 0.8507 |
| age=Adult, exercise induced angina=FALSE, slope of the peak exercise ST segment=Upsloping, number of major vessels=0 ⇒ heart dissease=FALSE | 0.2 | 0.9 | 1.62 | 0.8 |

| | | | | |
|---|---|---|---|---|
| maximum heart rate achieved=[166,202],number of major vessels=0,thal=Normal ⇒ heart dissease=FALSE | 0.1556 | 0.9767 | 1.7581 | 0.9535 |
| age=Adult, maximum heart rate achieved=[166,202],thal=Normal ⇒ heart dissease=FALSE | 0.163 | 0.9778 | 1.76 | 0.9556 |
| sex=female, chest pain type=non-anginal pain, thal=Normal ⇒ heart dissease=FALSE | 0.1074 | 1.8 | 1 | |
| age=Adult, sex=female, exercise induced angina=FALSE ⇒ heart dissease=FALSE | 0.137 | 0.9737 | 1.7526 | 0.9474 |
| oldpeak=[0.1,1.4),number of major vessels=0,thal=Normal ⇒ heart dissease=FALSE | 0.1259 | 0.9714 | 1.7486 | 0.9429 |
| age=Adult, resting electrocardiographic results=Normal, slope of the peak exercise ST segment=Upsloping, number of major vessels=0,thal=Normal ⇒ heart dissease=FALSE | 0.1222 | 0.9706 | 1.7471 | 0.9412 |

Tabla 12: Conjunto de reglas que reflejan tendencia a no padecer enfermedades cardiacas

Este conjunto de reglas se caracteriza por presentar altos grados de confianza y confianza confirmada, lo cual nos da una referencia de que los factores reflejados en los antecedentes de la regla tienden a reflejar de forma casi absulta la tendencia a la ausencia de enfermedades cardiacas frente a la presencia de enfermedades cardiacas. El anterior conjunto de reglas, se caracteriza además por no presentar ítems diferentes de un mismo atributo del *dataset*, es decir, el conjunto de reglas no presenta factores contradictorios.

El anterior conjunto de reglas nos permitiría, de este modo, enumerar una serie de factores que tienden a estar relacionados con la ausencia de enfermedades cardiacas:

- **exercise induced angina=FALSE**: Los pacientes no realizan ejercicios que les pudieran inducir la aparicion de anginas.

- **number of major vessels=0**: Un valor nulo de números de vasos mayores coloreados por fluoroscopio estaría relacionado con la ausencia de enfermedades cardiacas, esto en sí se relacionaría, a priori, con la ausencia de episodios cardiacos previos u análisis clínicos anteriores que llevaran a la necesidad de analizar el estado de los vasos sanguíneos mayores del paciente mediante esta técnica de coloración.

- **thal=Normal**: La ausencia de enfermedades y/o la ausencia de alteraciones del estado de salud del paciente también estaría relacionada con la ausencia de enfermedades cardiacas.

- **age=Adult**: Los adultos presentarían menos enfermedades cardiacas que las personas de edad avanzada.

- **slope of the peak exercise ST segment=Upsloping**: La presencia de una pendiente ascendente en el pico del segmento ST reflejaría la ausencia de enfermedades cardiacas.

- **maximum heart rate achieved=[166,202]**: Una frecuencia cardiaca máxima alcanzable comprendida en el intervalo [166, 202] (niveles normales), estaría relacionado con la ausencia de enfermedades cardiacas. Esto se explicaría realmente por la ausencia de enfermedades cardiacas que llevarían a una alteración del rítmo cardiaco normal del paciente.

- **sex=female**: Se muestra una tendencia entre el hecho de ser mujer y la ausencia de enfermedades cardiacas. Esto llevaría a considerar y a analizar si los hombres presentan mayor tendencia que las mujeres a padecer enfermedades cardiacas o no por el hecho de ser hombres y/o determinar las causas de esta tendencia.

- **chest pain type=non-anginal pain**: La ausencia de dolor de angina en el pecho estaría relacionada con la ausencia de enfermedades cardiacas.

- **oldpeak**: Una depresión del segmento ST inducida por el ejecicio relativo al reposo comprendida entre 0.1 y 1.4 (valores relativamente bajos en relación a la distribución de valores de este atributo) estaría relacionado con la ausencia de enfermedades cardiacas.

- **resting electrocardiographic results=Normal**: Valores normales de resultados de pruebas de eletrocardiografía medidas en reposo estaría relacionado con la ausencia de enfermedades cardiacas.

Aunque estos factores estén relacionados con la ausencia de enfermedades cardiacas, para evaluar de forma más precisa la tendencia entre la aparición de cada uno de estos factores y la ausencia de enfermedades cardiacas sería necesario un análisis más profundo mediante reglas más específicas.

## 4.2. Conjunto de reglas que determinan la presencia de enfermedades cardiacas

Se analizan ahora, el siguiente conjunto de reglas que determina la tendencia a padecer enfermedades cardiacas debido a diversos factores:

| Regla | Soporte | Confianza | Lift | Confianza Confirmada |
|---|---|---|---|---|
| chest pain type=asymptomatic, thal=Reversable defect ⇒ heart dissease=TRUE | 0.2333 | 0.9 | 2.025 | 0.8 |
| exercise induced angina=TRUE, oldpeak=[1.4,6.2],thal=Reversable defect ⇒ heart dissease=TRUE | 0.1222 | 0.9706 | 2.1838 | 0.9412 |
| serum cholestoral=Dangerous level, exercise induced angina=TRUE, oldpeak=[1.4,6.2] ⇒ heart dissease=TRUE | 0.1 | 1 | 2.25 | 1 |
| chest pain type=asymptomatic, oldpeak=[1.4,6.2],thal=Reversable defect ⇒ heart dissease=TRUE | 0.137 | 1 | 2.25 | 1 |
| chest pain type=asymptomatic, resting electrocardiographic results=left ventricular hypertrophy, thal=Reversable defect ⇒ heart dissease=TRUE | 0.1407 | 0.9744 | 2.1923 | 0.9487 |
| chest pain type=asymptomatic, serum cholestoral=Dangerous level, thal=Reversable defect ⇒ heart dissease=TRUE | 0.1407 | 0.9744 | 2.1923 | 0.9487 |

Tabla 13: Conjunto de reglas que reflejan tendencia a padecer enfermedades cardiacas

Al igual que en el caso anterior, el conjunto de reglas presenta altos valores de confianza y confianza confirmada, lo que nos da una idea del alto

cumplimiento de esta reglas y de la mayor tendencia de los factores enumerados en los antecedentes para relacionar su aparición con la presencia de enfermedades cardiacas que la ausencia de enfermedades cardiacas.

De este modo, enumeramos a continuación, los factores que muestran una fuerte relación con la presencia de enfermedades cardiacas:

- **chest pain type = asymptomatic**: Un dolor de pecho asimtomático estaría relacionado con la presencia de enfermedades cardiacas.
- **thal=Reversable defect**: La presencia de una enfermedad o alteración del estado normal de salud reversible estaría relacionada con la presencia de enfermedades cardiacas.
- **exercise induced angina=TRUE**: La realización de ejercicios inducidos a angina estaría relacionado con la presencia de enfermedades cardiacas.
- **oldpeak=[1.4,6.2]**: Una depresión del segmento ST inducida por el ejecicio relativo al reposo comprendida entre 1.4 y 6.2 (se corresponde con los valores más altos en la distribución de este atributo), estaría relacionado con la presencia de enfermedades cardiacas.
- **serum cholestoral=Dangerous level**: Niveles elevados y considerados peligrosos de colésterol sérico estarían relacionados con la presencia de enfermedades cardiacas.
- **resting electrocardiographic results=left ventricular hypertrophy**: Valores de hipertrofía en el ventrículo izquierdo en resultados de pruebas de electrocardiografía estarían relacionados con la presencia de enfermedades cardiacas.

Por último, se puede apreciar que estos factores que están relacionados con la presencia de enfermedades cardiacas son mutuamente excluyentes con los factores relacionados con la ausencia de enfermedades cardiacas. Esta característica unida a la alta confiabilidad de las reglas permitiría **usar estos 2 grupos de reglas para determinar de forma clara la presencia o ausencia de enfermedades cardiacas** conociendo la existencia de los factores incluídos en el antecedente de las reglas.

## 4.3. Relación entre el sexo y la presencia de enfermedades cardiacas

A continuación, se analiza el siguiente conjunto de reglas que muestran relaciones entre la presencia de enfermedades cardiacas y el sexo del paciente:

| Regla | Soporte | Confianza | Lift | Confianza Confirmada |
|---|---|---|---|---|
| age=Adult, heart dissease=TRUE ⇒ sex=male | 0.2407 | 0.9028 | 1.332 | 0.8056 |
| number of major vessels=1,heart disease=TRUE ⇒ sex=male | 0.137 | 0.9737 | 1.4366 | 0.9474 |
| resting blood pressure=Elevated, heart dissease=TRUE ⇒ sex=male | 0.1148 | 1 | 1.4754 | 1 |
| maximum heart rate achieved=[71,133),exercise induced angina=TRUE, heart dissease=TRUE ⇒ sex=male | 0.1296 | 0.9722 | 1.4344 | 0.9444 |
| sex=female, chest pain type=non-anginal pain, thal=Normal ⇒ heart dissease=FALSE | 0.1074 | 1.0000 | 1.8000 | 1.0000 |
| age=Adult, sex=female, exercise induced angina=FALSE ⇒ heart dissease=FALSE | 0.1370 | 0.9737 | 1.7526 | 0.9474 |
| heart dissease=TRUE ⇒ sex=male | 0.3704 | 0.8333 | 1.2295 | 0.6667 |

Tabla 14: Conjunto de reglas que reflejan tendencia entre el padecimiento de enfermedades cardiacas y el hecho de ser hombre

La última regla refleja una **clara tendencia entre el padecimiento de enfermedades cardiacas y el hecho de ser hombre**, otorgando una confianza de 0.8333 a esta relación.

Por su parte, las 4 primeras reglas reflejan factores en su antecedente que, unidos al hecho de presentar enfermedades cardiacas, **otorgan una mayor confianza de que el sexo del enfermo será hombre**.

Estos factores se enumeran a continuación:

- age=Adult
- number of major vessels=1

- resting blood pressure=Elevated
- maximum heart rate achieved=[71,133)
- exercise induced angina=TRUE

Por último, la 5º y la 6º regla, son las únicas reglas que involucran el hecho de ser mujer con la ausencia de enfermedades cardiacas.

Si bien, para analizar con mayor detalle estos fenómenos sería necesario realizar un análisis más profundo con ayuda de otras reglas no exploradas, de este conjunto de reglas podemos concluir un observación muy interesante: **Los pacientes hombres presentan una mayor tendencia a padecer enfermedades cardiacas que las mujeres, y esta tendencia se afianza con la presencia del conjunto de factores enumerados anteriormente**.

# 5. Selección de reglas más interesantes

Finalmente, una vez finalizado el análisis y extracción de reglas, se lleva a cabo una selección de 8 reglas que se consideradan de mayor interés y que identifican asociaciones relevantes:

1. **age=Adult, heart dissease=TRUE ⇒ sex=male**: Esta regla tal y como se ha podido comprobar en el análisis de reglas por grupos, resulta de interés por reflejar una mayor tendencia de los hombres a presentar enfermedades cardiacas.

2. **chest pain type=asymptomatic, thal=Reversable defect ⇒ heart dissease=TRUE**: Se propone el conjunto de reglas que determinan la presencia de enfermedades cardiacas como un conjunto de reglás de interés, por relacionar, con una alta confiabilidad (igual o superior a 0.9 en todos los casos), la presencia de enfermedades cardiacas con un conjunto bien definido de factores.

   En esta regla, se sugiere que un dolor de pecho asimtomático y el padecimiento de un defecto de salud reversible están relacionados con la aparición de enfermedades cardiacas.

3. **exercise induced angina=TRUE, oldpeak=[1.4,6.2],thal=Reversable defect ⇒ heart dissease=TRUE**: Esta regla relaciona la realización,

por parte del paciente de ejercicios que pudieran inducir a dolor de angina, una depresión en el pico del segmento ST-T alta y nuevamente el padecimiento de un defecto de salud reversible con la presencia de enfermedades cardiacas.

4. **serum cholestoral=Dangerous level, exercise induced angina=TRUE, oldpeak=[1.4,6.2] ⇒ heart dissease=TRUE**: Esta regla relaciona la presencia de niveles de colesterol sérico considerados peligrosos, la realización por parte del paciente de ejercicios que pudieran inducir a dolor de angina, y una depresión en el pico del segmento ST-T alta con la aparición de enfermedades cardiacas.

5. **chest pain type=asymptomatic, oldpeak=[1.4,6.2],thal=Reversable defect ⇒ heart dissease=TRUE**: Esta regla relaciona ahora el padecimiento de dolores asimtomáticos en el pecho, además de la presencia de una depresión en el pico del segmento ST-T alta junto con el padecimiento de un defecto de salud reversible con la presencia de enfermedades cardiacas.

6. **chest pain type=asymptomatic, resting electrocardiographic results=left ventricular hypertrophy, thal=Reversable defect ⇒ heart dissease=TRUE**: Esta regla relaciona el padecimiento de dolores asimtomáticos en el pecho, el padecimiento de un defecto de salud reversible y, ahora, una hipertrofía ventricular izquierda reflejado en los resultados de las pruebas con electrocardigramas en reposo con la presencia de enfermedades cardiacas.

7. **chest pain type=asymptomatic, serum cholestoral=Dangerous level, thal=Reversable defect ⇒ heart dissease=TRUE**: Esta regla relaciona el padecimiento de dolores asimtomáticos en el pecho, la presencia de niveles de colesterol sérico considerados peligrosos el padecimiento de un defecto de salud reversible con la presencia de enfermedades cardiacas.

8. **resting blood pressure=Elevated, thal=Reversable defect ⇒ sex=male**: Esta regla identifica una fuerte relación entre el hecho de padecer una elevada presión sanguínea unido al padecidimiento de un defecto de salud reversible con el hecho de ser hombre. Si bien considerar esta afirmación es muy fuerte, esta regla refleja que los pacientes con estas dolencias son en su mayoría hombres.

# A. preprocesamiento_variables.R

Este script fue utilizado para realizar un análisis breve de las variables que constituyen el *dataset* original de los datos y para la ejecución de las tranformaciones consideradas antes de convertir el *dataset* en un conjunto de transacciones.

```
# Librerías importados
library('arules')
library('ggplot2')

# Función para lectura del dataset
read.statlog.heart.dataset <- function(filename) {
  dat <- read.csv(filename, sep=' ')

  colnames(dat) <- c('age', 'sex', 'chest pain type',
                     'resting blood pressure', 'serum cholestoral',
                     'fasting blood sugar',
                     'resting electrocardiographic results',
                     'maximum heart rate achieved',
                     'exercise induced angina', 'oldpeak',
                     'slope of the peak exercise ST segment',
                     'number of major vessels', 'thal', 'heart dissease
   ')

  # Preprocesar el dataset para asignar los tipos de datos correctos
  # Convertir age en integer
  dat$age <- as.integer(dat$age)

  # Convertir sex en factor
  dat$sex <- factor(x=dat$sex, levels=c(1,0), labels=c('male','female')
   )

  # Convertir chest pain type en factor
  dat[,3] <- factor(dat[,3], levels=1:4, labels=c('typical angina',
                                                  'atypical angina',
                                                  'non-anginal pain',
                                                  'asymptomatic'))

  # Convertir fasting blood sugar en factor binario
  dat[,6] <- as.logical(dat[,6])

  # Convertir resting electrocardiographic results en factores
  dat[,7] <- factor(dat[,7], levels=0:2, labels=c('Normal',
                                                  'ST-T wave anormality',
                                                  'left ventricular
   hypertrophy'))

  # Convertir exercise induced angina en factores binarios
  dat[,9] <- as.logical(dat[,9])

  # Convertir the slope of the peak exercise ST segment en factores
  dat[,11] <- factor(dat[,11], levels=1:3, labels=c('Upsloping',
                                                    'Flat',
                                                    'Downsloping'))

  # Convertir number of major vessels
```

```
  dat[,12] <- as.integer(dat[,12])

  # Convertir thal en factores
  dat[,13] <- factor(dat[,13], levels=c(3,6,7), labels=c('Normal',
                                                         'Fixed defect',
                                                         'Reversable
   defect'))
  # Convertir la clase en factor binaria
  dat[,14] <- dat[,14] == 2.0

  return(dat)
}

# Cargar dataset
heart <- read.statlog.heart.dataset('./heart.dat')

# Información breve del dataset
head(heart)
summary(heart)

# Estudiar los estadísticos de posición de age
summary(heart$age)

# Discretización del atributo age
heart[['age']] <- ordered(cut(heart[['age']],
                              c(29,60,+Inf),
                              labels=c('Adult', 'Elderly'),
                              right = F))

# Estudiar los estadísticos de posición resting blood pressure
summary(heart[['resting blood pressure']])

# Discretizar el atributo resting blood presssure
heart[['resting blood pressure']] <- ordered(
                cut(heart[['resting blood pressure']],
                    c(94,120,130,140,+Inf),
                    labels=c('Normal', 'Elevated',
                             'Hypertension-stage1',
                             'Hypertension-stage2'),
                    right = F))

# Estudiar los estadísticos de posición de serum cholesterol
summary(heart[['serum cholestoral']])

# Discretizar el atributo serum cholesterol
heart[['serum cholestoral']] <- ordered(
          cut(heart[['serum cholestoral']],
              c(126,200,240,+Inf),
              labels=c('Normal level',
                       'High level',
                       'Dangerous level'),
              right = F))

# Estudiar los estadísticos de posición de maximum heart rate achieved
summary(heart[['maximum heart rate achieved']])

# Estudiar gráficamente el atributo "maximum heart rate achieved"
pdf('maximum_heart_rate_achieved.pdf')
```

```
graf <- ggplot2::ggplot(data=heart, aes(x=`maximum heart rate achieved
    `)) +
  ggplot2::geom_histogram(binwidth = 2, colour='white',
                          fill='coral3') +
  ggplot2::ylab('Frecuencia absoluta')
graf
dev.off()

# Discretizar en 4 intervalos
heart[['maximum heart rate achieved']] <- discretize(
  heart[['maximum heart rate achieved']],
  method = 'frequency', breaks=4)

# Estudiar los estadísticos de posición de oldpeek
summary(heart[['oldpeak']])

# Estudiar gráficamente el atributo "oldpeak"
pdf('oldpeak.pdf')
graf <- ggplot2::ggplot(data=heart, aes(x=oldpeak)) +
  ggplot2::geom_histogram(binwidth = 0.1, colour='white',
                          fill='coral3') +
  ggplot2::ylab('Frecuencia absoluta')
graf
dev.off()

# Discretizar el atributo oldpeak en 3 intervalos de igual frecuencia
heart[['oldpeak']] <- discretize(
  heart[['oldpeak']], method = 'frequency',
  breaks=3)

# Convertir exercise induced angina y heart dissease en factores
heart[['exercise induced angina']] <- factor(
  heart[['exercise induced angina']], ordered = TRUE)
heart[['heart dissease']] <- factor(
  heart[['heart dissease']], ordered = TRUE)
```

# B. analisis_reglas.R

Este *script* fue escrito para llevar a cabo una exploración y análisis de itemsets frecuentes, así como de reglas de interés generadas a partir de los mismos.

```
# Librerías importados
library('arules')

# Función para lectura del dataset
read.statlog.heart.transactions <- function(filename) {
  dat <- read.table(filename, sep=' ')

  colnames(dat) <- c('age', 'sex', 'chest pain type',
                     'resting blood pressure', 'serum cholestoral',
                     'fasting blood sugar',
                     'resting electrocardiographic results',
                     'maximum heart rate achieved',
                     'exercise induced angina', 'oldpeak',
```

```
                    'slope of the peak exercise ST segment',
                    'number of major vessels', 'thal', 'heart dissease
  ')

# Preprocesar el dataset para asignar los tipos de datos correctos
# Convertir age en integer
dat$age <- as.integer(dat$age)

# Convertir sex en factor
dat$sex <- factor(x=dat$sex, levels=c(1,0), labels=c('male','female')
 )

# Convertir chest pain type en factor
dat[,3] <- factor(dat[,3], levels=1:4, labels=c('typical angina',
                                                'atypical angina',
                                                'non-anginal pain',
                                                'asymptomatic'))

# Convertir fasting blood sugar en factor binario
dat[,6] <- as.logical(dat[,6])

# Convertir resting electrocardiographic results en factores
dat[,7] <- factor(dat[,7], levels=0:2, labels=c('Normal',
                                    'ST-T wave anormality',
                                    'left ventricular hypertrophy'))

# Convertir exercise induced angina en factores binarios
dat[,9] <- as.logical(dat[,9])

# Convertir the slope of the peak exercise ST segment en factores
dat[,11] <- factor(dat[,11], levels=1:3, labels=c('Upsloping',
                                                  'Flat',
                                                  'Downsloping'))

# Convertir number of major vessels
dat[,12] <- as.factor(as.integer(dat[,12]))

# Convertir thal en factores
dat[,13] <- factor(dat[,13], levels=c(3,6,7), labels=c(
                                            'Normal',
                                            'Fixed defect',
                                            'Reversable defect'))
# Convertir la clase en factor binaria
dat[,14] <- dat[,14] == 2.0

# Discretización del atributo age
dat[['age']] <- ordered(cut(dat[['age']],
                            c(29,60,+Inf),
                            labels=c('Adult', 'Elderly'),
                            right = F))

# Discretizar el atributo resting blood presssure
dat[['resting blood pressure']] <- ordered(
  cut(dat[['resting blood pressure']],
      c(94,120,130,140,+Inf),
      labels=c('Normal', 'Elevated',
               'Hypertension-stage1',
               'Hypertension-stage2'),
      right = F))
```

```
71
72    # Discretizar el atributo serum cholesterol
73    dat[['serum cholestoral']] <- ordered(
74      cut(dat[['serum cholestoral']],
75          c(126,200,240,+Inf),
76          labels=c('Normal level',
77                   'High level',
78                   'Dangerous level'),
79          right = F))
80
81    # Discretizar en 4 intervalos
82    dat[['maximum heart rate achieved']] <- discretize(
83      dat[['maximum heart rate achieved']],
84      method = 'frequency', breaks=4)
85
86    # Discretizar el atributo oldpeak en 3 intervalos de igual frecuencia
87    dat[['oldpeak']] <- discretize(
88      dat[['oldpeak']], method = 'frequency',
89      breaks=3)
90
91    # Convertir exercise induced angina y heart dissease en factores
92    dat[['exercise induced angina']] <- factor(
93      dat[['exercise induced angina']], ordered = TRUE)
94    dat[['heart dissease']] <- factor(
95      dat[['heart dissease']], ordered = TRUE)
96
97    # Convertir el dataset en un conjunto de transacciones
98    dat <- as(dat, 'transactions')
99
100   return(dat)
101 }
102
103 # Definir función para eliminar reglas redundantes
104 remove.redundat.rules <- function(rule.set) {
105
106   # Identificar las reglas que son superconjuntos de otras
107   subset.rules.mask <- is.subset(rule.set)
108   redundant.rules.mask <- colSums(subset.rules.mask,
109                                   na.rm = TRUE) >= 2
110
111   # Eliminar reglas redundates
112   rule.set[!redundant.rules.mask]
113 }
114
115 # Cargar dataset
116 heart <- read.statlog.heart.transactions('./heart.dat')
117
118 # Extraer reglas con un minConf de 0.9
119 rules.conf1 <- arules::apriori(heart, parameter = list(support=0.2,
120                                                        confidence=0.9,
121                                                        minlen=2))
122 rules.conf1 <- sort(rules.conf1, by='support')
123
124 #  Se eliminan las reglas redundantes
125 rules.conf1.pruned <- remove.redundat.rules(rules.conf1)
126
127 # Extraer reglas con un minConf de 0.97 y un soporte máximo de 0.2
128 rules.conf2 <- arules::apriori(heart, parameter = list(confidence=0.97,
129                                                        minlen=2,
```

```
                                                    smax=0.2))

#  Se eliminan las reglas redundantes
rules.conf2.pruned <- remove.redundat.rules(rules.conf2)

# Explorar reglas de soporte mínimo 0.3037 y confianza mínima 0.6
rules.conf3 <- arules::apriori(heart, parameter = list(support=0.3037,
                                                       confidence=0.6,
                                                       minlen=2))

#   Aplicar poda al conjunto de reglas rules.conf3
rules.conf3.pruned <- remove.redundat.rules(rules.conf3)

#   Seleccionar sólo las reglas que se eligieron
rules.conf3.pruned <- subset(rules.conf3.pruned,
                             subset=support>0.336 & confidence>0.7537)

# Combinar todas las reglas en un sólo conjunto
rule.set <- union(union(rules.conf1.pruned, rules.conf2.pruned),
                  rules.conf3.pruned)

# Calcular métricas extra
extra.measures.rules.set <- arules::interestMeasure(rule.set,
                                        measure=c('
    confirmedConfidence',
                                                  'lift'),
                                        transactions = heart)
quality(rule.set) <-round(
  cbind(quality(rule.set), extra.measures.rules.set),
  4)

# Analizar las reglas con heart dissease=FALSE
rules.heart_dissease.false <- subset(rule.set,
        subset=rhs %in% 'heart dissease=FALSE')

# Analizar las reglas con heart dissease=TRUE
rules.heart_dissease.true <- subset(rule.set,
        subset=rhs %in% 'heart dissease=TRUE')

# Analizar las reglas con sex=male y heart dissease=TRUE en el
    antecedente o consecuente
rules.male_heart_dissease <- subset(rule.set,
            subset=lhs %in% c('heart dissease=TRUE',
                              'heart dissease=FALSE',
                              'sex=male', 'sex=female') &
              rhs %in% c('heart dissease=TRUE',
                         'heart dissease=FALSE',
                         'sex=male', 'sex=female'))

# Búsqueda de otros grupos de reglas que no llevó al descubrimiento de
    ningún hecho relevante
inspect(subset(rule.set,
     subset=lhs %in% c('exercise induced angina=FALSE',
                                'exercise induced angina=TRUE') |
       rhs %in% c('exercise induced angina=FALSE',
                  'exercise induced angina=TRUE')))

inspect(subset(rule.set,
     subset=lhs %in% c('number of major vessels=0',
```

```
                                    'number of major vessels=1',
                                    'number of major vessels=2',
                                    'number of major vessels=3') |
       rhs %in% c('number of major vessels=0',
                  'number of major vessels=1',
                  'number of major vessels=2',
                  'number of major vessels=3')))

inspect(subset(rule.set,
       subset=lhs %in% c('thal=Normal',
                         'thal=Reversable defect',
                         'thal=Fixed defect',
                         'heart dissease=TRUE',
                         'heart dissease=FALSE') &
         rhs %in% c('thal=Normal', 'thal=Reversable defect',
                    'thal=Fixed defect', 'heart dissease=TRUE',
                    'heart dissease=FALSE')))

inspect(subset(rule.set,
       subset=lhs %in% c('serum cholestoral=Dangerous level',
                         'serum cholestoral=High level',
                         'serum cholestoral=Normal level') |
         rhs %in% c('serum cholestoral=Dangerous level',
                    'serum cholestoral=High level',
                    'serum cholestoral=Normal level')))

inspect(subset(rule.set, subset=lhs %in% c(
  'resting electrocardiographic results=left ventricular hypertrophy',
  'resting electrocardiographic results=Normal',
  'resting electrocardiographic results=ST-T wave anormality') |
                 rhs %in% c(
  'resting electrocardiographic results=left ventricular hypertrophy',
  'resting electrocardiographic results=Normal',
  'resting electrocardiographic results=ST-T wave anormality')))

inspect(subset(rule.set, subset=lhs %in% c(
  'slope of the peak exercise ST segment=Upsloping',
  'slope of the peak exercise ST segment=Downsloping',
  'slope of the peak exercise ST segment=Flat') |
                 rhs %in% c(
  'slope of the peak exercise ST segment=Upsloping',
  'slope of the peak exercise ST segment=Downsloping',
  'slope of the peak exercise ST segment=Flat')))

inspect(subset(rule.set, subset=lhs %in% c(
  'chest pain type=asymptomatic',
  'chest pain type=non-anginal pain',
  'chest pain type=atypical angina',
  'chest pain type=typical angina') |
                 rhs %in% c(
  'chest pain type=asymptomatic',
  'chest pain type=non-anginal pain',
  'chest pain type=atypical angina',
  'chest pain type=typical angina')))

inspect(subset(rule.set, subset=lhs %in% c(
  'oldpeak=[0,0.1)',
  'oldpeak=[0.1,1.4)',
  'oldpeak=[1.4,6.2]') |
```

```
                    rhs %in% c(
  'oldpeak=[0,0.1)',
  'oldpeak=[0.1,1.4)', 'oldpeak=[1.4,6.2]')))

inspect(subset(rule.set,
   subset=lhs %in% c('maximum heart rate achieved=[71,133)',
                     'maximum heart rate achieved=[133,154)',
                     'maximum heart rate achieved=[166,202]') |
     rhs %in% c('maximum heart rate achieved=[71,133)',
                'maximum heart rate achieved=[133,154)',
                'maximum heart rate achieved=[166,202]')))
```

# C. analisis_grupos.R

Este *script* fue utilizado para aislar subconjuntos del conjunto de reglas extraído como resultado de la ejecución del anterior *script* y faciitar su análisis grupal.

```
# Librerías importados
library('arules')

# Función para lectura del dataset
read.statlog.heart.transactions <- function(filename) {
  dat <- read.table(filename, sep=' ')

  colnames(dat) <- c('age', 'sex', 'chest pain type',
                     'resting blood pressure', 'serum cholestoral',
                     'fasting blood sugar',
                     'resting electrocardiographic results',
                     'maximum heart rate achieved',
                     'exercise induced angina', 'oldpeak',
                     'slope of the peak exercise ST segment',
                     'number of major vessels', 'thal', 'heart dissease
    ')

  # Preprocesar el dataset para asignar los tipos de datos correctos
  # Convertir age en integer
  dat$age <- as.integer(dat$age)

  # Convertir sex en factor
  dat$sex <- factor(x=dat$sex, levels=c(1,0), labels=c('male','female')
    )

  # Convertir chest pain type en factor
  dat[,3] <- factor(dat[,3], levels=1:4, labels=c('typical angina',
                                                  'atypical angina',
                                                  'non-anginal pain',
                                                  'asymptomatic'))

  # Convertir fasting blood sugar en factor binario
  dat[,6] <- as.logical(dat[,6])

  # Convertir resting electrocardiographic results en factores
  dat[,7] <- factor(dat[,7], levels=0:2, labels=c('Normal',
                                        'ST-T wave anormality',
```

```
                                        'left ventricular hypertrophy'))

# Convertir exercise induced angina en factores binarios
dat[,9] <- as.logical(dat[,9])

# Convertir the slope of the peak exercise ST segment en factores
dat[,11] <- factor(dat[,11], levels=1:3, labels=c('Upsloping',
                                                  'Flat',
                                                  'Downsloping'))

# Convertir number of major vessels
dat[,12] <- as.factor(as.integer(dat[,12]))

# Convertir thal en factores
dat[,13] <- factor(dat[,13], levels=c(3,6,7), labels=c(
                                                'Normal',
                                                'Fixed defect',
                                                'Reversable defect'))
# Convertir la clase en factor binaria
dat[,14] <- dat[,14] == 2.0

# Discretización del atributo age
dat[['age']] <- ordered(cut(dat[['age']],
                            c(29,60,+Inf),
                            labels=c('Adult', 'Elderly'),
                            right = F))

# Discretizar el atributo resting blood presssure
dat[['resting blood pressure']] <- ordered(
  cut(dat[['resting blood pressure']],
      c(94,120,130,140,+Inf),
      labels=c('Normal', 'Elevated',
               'Hypertension-stage1',
               'Hypertension-stage2'),
      right = F))

# Discretizar el atributo serum cholesterol
dat[['serum cholestoral']] <- ordered(
  cut(dat[['serum cholestoral']],
      c(126,200,240,+Inf),
      labels=c('Normal level',
               'High level',
               'Dangerous level'),
      right = F))

# Discretizar en 4 intervalos
dat[['maximum heart rate achieved']] <- discretize(
  dat[['maximum heart rate achieved']],
  method = 'frequency', breaks=4)

# Discretizar el atributo oldpeak en 3 intervalos de igual frecuencia
dat[['oldpeak']] <- discretize(
  dat[['oldpeak']], method = 'frequency',
  breaks=3)

# Convertir exercise induced angina y heart dissease en factores
dat[['exercise induced angina']] <- factor(
  dat[['exercise induced angina']], ordered = TRUE)
dat[['heart dissease']] <- factor(
```

```
    dat[['heart dissease']], ordered = TRUE)

  # Convertir el dataset en un conjunto de transacciones
  dat <- as(dat, 'transactions')

  return(dat)
}

# Definir función para eliminar reglas redundantes
remove.redundat.rules <- function(rule.set) {

  # Identificar las reglas que son superconjuntos de otras
  subset.rules.mask <- is.subset(rule.set)
  redundant.rules.mask <- colSums(subset.rules.mask,
                                  na.rm = TRUE) >= 2

  # Eliminar reglas redundates
  rule.set[!redundant.rules.mask]
}

# Cargar dataset
heart <- read.statlog.heart.transactions('./heart.dat')

# Extraer reglas con un minConf de 0.9
rules.conf1 <- arules::apriori(heart, parameter = list(support=0.2,
                                                       confidence=0.9,
                                                       minlen=2))
rules.conf1 <- sort(rules.conf1, by='support')

#  Se eliminan las reglas redundantes
rules.conf1.pruned <- remove.redundat.rules(rules.conf1)

# Extraer reglas con un minConf de 0.97 y un soporte máximo de 0.2
rules.conf2 <- arules::apriori(heart, parameter = list(confidence=0.97,
                                                       minlen=2,
                                                       smax=0.2))

#  Se eliminan las reglas redundantes
rules.conf2.pruned <- remove.redundat.rules(rules.conf2)

# Explorar reglas de soporte mínimo 0.3037 y confianza mínima 0.6
rules.conf3 <- arules::apriori(heart, parameter = list(support=0.3037,
                                                       confidence=0.6,
                                                       minlen=2))

#   Aplicar poda al conjunto de reglas rules.conf3
rules.conf3.pruned <- remove.redundat.rules(rules.conf3)

#   Seleccionar sólo las reglas que se eligieron
rules.conf3.pruned <- subset(rules.conf3.pruned,
                             subset=support>0.336 & confidence>0.7537)

# Combinar todas las reglas en un sólo conjunto
rule.set <- union(union(rules.conf1.pruned, rules.conf2.pruned),
                  rules.conf3.pruned)

# Calcular métricas extra
extra.measures.rules.set <- arules::interestMeasure(rule.set,
```

```
                                                measure=c('
    confirmedConfidence',
                                                          'lift'),
                                                transactions = heart)
quality(rule.set) <-round(
  cbind(quality(rule.set), extra.measures.rules.set),
  4)

# Analizar las reglas con heart dissease=FALSE
rules.heart_dissease.false <- subset(rule.set,
        subset=rhs %in% 'heart dissease=FALSE')

# Analizar las reglas con heart dissease=TRUE
rules.heart_dissease.true <- subset(rule.set,
        subset=rhs %in% 'heart dissease=TRUE')

# Analizar las reglas con sex=male y heart dissease=TRUE en el
    antecedente o consecuente
rules.male_heart_dissease <- subset(rule.set,
            subset=lhs %in% c('heart dissease=TRUE',
                              'heart dissease=FALSE',
                              'sex=male', 'sex=female') &
              rhs %in% c('heart dissease=TRUE',
                         'heart dissease=FALSE',
                         'sex=male', 'sex=female'))

# Búsqueda de otros grupos de reglas que no llevó al descubrimiento de
    ningún hecho relevante
inspect(subset(rule.set,
     subset=lhs %in% c('exercise induced angina=FALSE',
                                'exercise induced angina=TRUE') |
       rhs %in% c('exercise induced angina=FALSE',
                  'exercise induced angina=TRUE')))

inspect(subset(rule.set,
     subset=lhs %in% c('number of major vessels=0',
                                'number of major vessels=1',
                                'number of major vessels=2',
                                'number of major vessels=3') |
       rhs %in% c('number of major vessels=0',
                  'number of major vessels=1',
                  'number of major vessels=2',
                  'number of major vessels=3')))

inspect(subset(rule.set,
       subset=lhs %in% c('thal=Normal',
                         'thal=Reversable defect',
                         'thal=Fixed defect',
                         'heart dissease=TRUE',
                         'heart dissease=FALSE') &
         rhs %in% c('thal=Normal', 'thal=Reversable defect',
                    'thal=Fixed defect', 'heart dissease=TRUE',
                    'heart dissease=FALSE')))

inspect(subset(rule.set,
       subset=lhs %in% c('serum cholestoral=Dangerous level',
                         'serum cholestoral=High level',
                         'serum cholestoral=Normal level') |
         rhs %in% c('serum cholestoral=Dangerous level',
```

```
                    'serum cholestoral=High level',
                    'serum cholestoral=Normal level')))

inspect(subset(rule.set, subset=lhs %in% c(
  'resting electrocardiographic results=left ventricular hypertrophy',
  'resting electrocardiographic results=Normal',
  'resting electrocardiographic results=ST-T wave anormality') |
                  rhs %in% c(
  'resting electrocardiographic results=left ventricular hypertrophy',
  'resting electrocardiographic results=Normal',
  'resting electrocardiographic results=ST-T wave anormality')))

inspect(subset(rule.set, subset=lhs %in% c(
  'slope of the peak exercise ST segment=Upsloping',
  'slope of the peak exercise ST segment=Downsloping',
  'slope of the peak exercise ST segment=Flat') |
                  rhs %in% c(
  'slope of the peak exercise ST segment=Upsloping',
  'slope of the peak exercise ST segment=Downsloping',
  'slope of the peak exercise ST segment=Flat')))

inspect(subset(rule.set, subset=lhs %in% c(
  'chest pain type=asymptomatic',
  'chest pain type=non-anginal pain',
  'chest pain type=atypical angina',
  'chest pain type=typical angina') |
                  rhs %in% c(
  'chest pain type=asymptomatic',
  'chest pain type=non-anginal pain',
  'chest pain type=atypical angina',
  'chest pain type=typical angina')))

inspect(subset(rule.set, subset=lhs %in% c(
  'oldpeak=[0,0.1)',
  'oldpeak=[0.1,1.4)',
  'oldpeak=[1.4,6.2]') |
                  rhs %in% c(
  'oldpeak=[0,0.1)',
  'oldpeak=[0.1,1.4)', 'oldpeak=[1.4,6.2]')))

inspect(subset(rule.set,
   subset=lhs %in% c('maximum heart rate achieved=[71,133)',
                     'maximum heart rate achieved=[133,154)',
                     'maximum heart rate achieved=[166,202]') |
     rhs %in% c('maximum heart rate achieved=[71,133)',
                'maximum heart rate achieved=[133,154)',
                'maximum heart rate achieved=[166,202]')))
```