heart-disease

El dataset escogido se llama **Heart Disease Data Set** y se trata de un dataset del repositorio de datos UCI que tiene un conjunto de predictores y una variable de salida **num**. Se trata de un dataset que contiene datos de pacientes que han sido tratados por enfermedades cardíacas

Esta base de datos contiene 76 atributos, pero todos los experimentos publicados se refieren al uso de un subconjunto de 14 de ellos. En particular, la base de datos de Cleveland es la única que ha sido utilizada por investigadores de ML para esta fecha. El campo **num** se refiere a la presencia de una enfermedad cardíaca en el paciente. Es un valor entero de 0 (sin presencia) a 4. Los experimentos con la base de datos de Cleveland se han concentrado en simplemente intentar distinguir la presencia (valores 1,2,3,4) de la ausencia (valor 0).

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age                : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex                : num  1 1 1 1 0 1 0 0 1 1 ...
##  $ cp                 : num  1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps           : num  145 160 120 130 130 120 140 120 130 140 ...
##  $ cholestoral        : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ fasting_blood_sugar: num  1 0 0 0 0 0 0 0 0 1 ...
##  $ restecg            : num  2 2 2 0 2 0 2 0 2 2 ...
##  $ thalach            : num  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang              : num  0 1 1 0 0 0 0 1 0 1 ...
##  $ oldpeak            : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope              : num  3 2 2 3 1 1 3 1 2 3 ...
##  $ ca                 : num  0 3 2 0 0 0 2 0 1 0 ...
##  $ thal               : num  6 3 7 3 3 3 3 3 7 7 ...
##  $ num                : int  0 2 1 0 0 0 3 0 2 1 ...
```

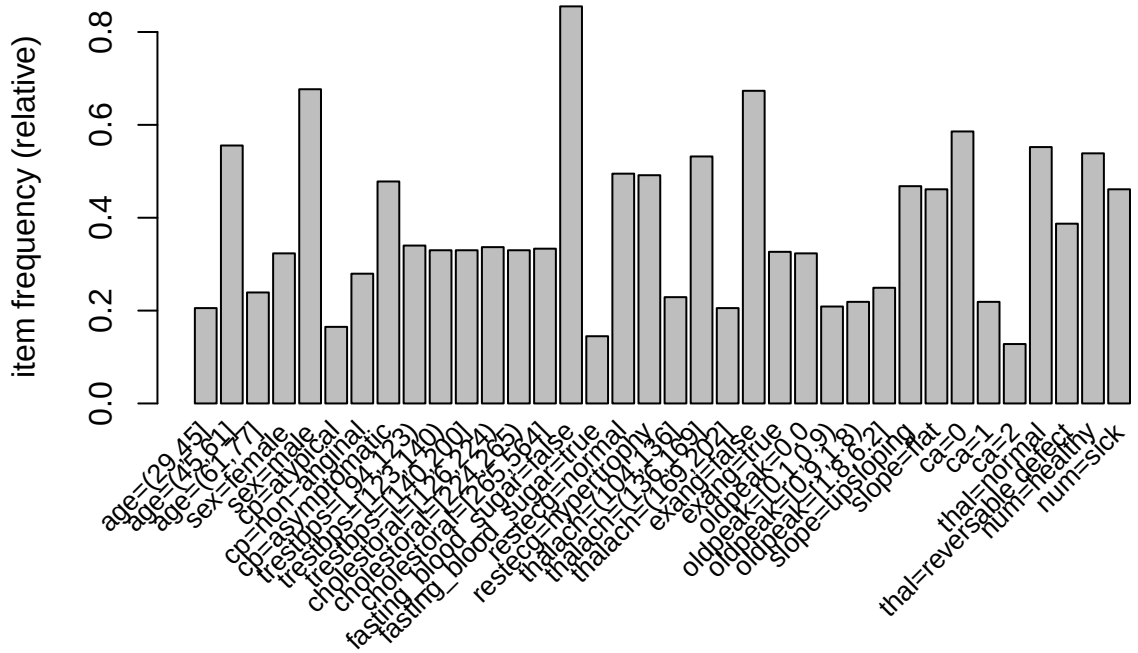
El primer paso es el proceso de limpieza de datos para eliminar posibles valores faltante en el conjunto de datos, luego discretizar posibles variables continuas. Comprobamos la presencia de N.A. Solo 6 casos presentan símptomas de tener missing values, por ello seran eliminados del dataser

El segundo paso es la discretización de los datos. La discretización cambia el tipo de datos de atributos de tipo numérico a tipo discreto Algunos atributos con tipo numérico son la edad, trestbps, chol, thalach, oldpeak and ca, y se transforman a tipo discreto.

```
## 'data.frame':    297 obs. of  14 variables:
##  $ age                : Factor w/ 3 levels "(29,45]","(45,61]",..: 3 3 3 1 1 2 3 2 3 2 ...
##  $ sex                : Factor w/ 2 levels "female","male": 2 2 2 2 1 2 1 1 2 2 ...
##  $ cp                 : Factor w/ 4 levels "typical","atypical",..: 1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps           : Factor w/ 3 levels "[ 94,123)","[123,140)",..: 3 3 1 2 2 1 3 1 2 3 ...
##  $ cholestoral        : Factor w/ 3 levels "[126,224)","[224,265)",..: 2 3 2 2 1 2 3 3 2 1 ...
##  $ fasting_blood_sugar: Factor w/ 2 levels "false","true": 2 1 1 1 1 1 1 1 1 2 ...
##  $ restecg            : Factor w/ 3 levels "normal","stt",..: 3 3 3 1 3 1 3 1 3 3 ...
##  $ thalach            : Factor w/ 4 levels "(70.9,104]","(104,136]",..: 3 2 2 4 4 4 3 3 3 3 ...
##  $ exang              : Factor w/ 2 levels "false","true": 1 2 2 1 1 1 1 2 1 2 ...
##  $ oldpeak            : Factor w/ 4 levels "0.0","[0.1,0.9)",..: 4 3 4 4 3 2 4 2 3 4 ...
##  $ slope              : Factor w/ 3 levels "upsloping","flat",..: 3 2 2 3 1 1 3 1 2 3 ...
##  $ ca                 : Factor w/ 4 levels "0","1","2","3": 1 4 3 1 1 1 3 1 2 1 ...
##  $ thal               : Factor w/ 3 levels "normal","fixed defect",..: 2 1 3 1 1 1 1 1 3 3 ...
##  $ num                : Factor w/ 2 levels "healthy","sick": 1 2 2 1 1 1 2 1 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int  88 167 193 267 288 303
##    ..- attr(*, "names")= chr  "88" "167" "193" "267" ...
```

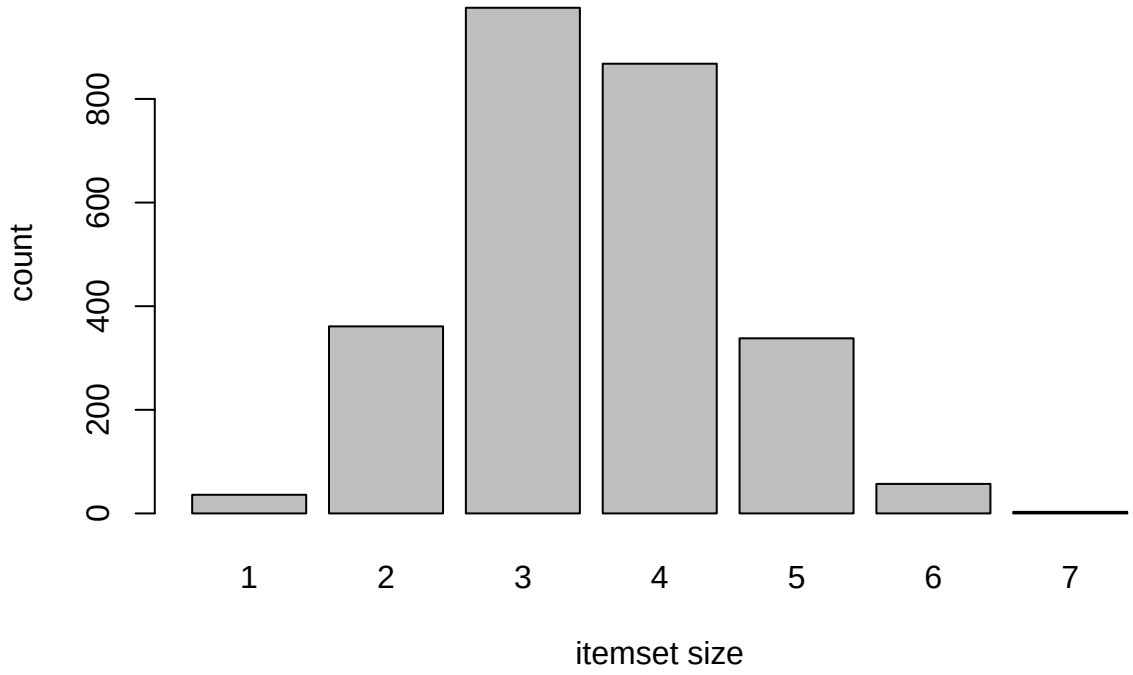
En el siguiente gráfico se muestran los items con una frecuencia de más de 0.1 (soporte > 0.1). En este

gráfico podemos decidir qué items estudiar más en profundidad. Para obtener reglas con valor, podemos analizar reglas en las que aparezcan items no muy frecuentes. Como se muestra, hay algunos items bastante frecuentes, como son por ejemplo lapresión arterial en reposo **trestbps=Desirable** y el azúcar en la sangre no en ayunas **fasting_blood_sugar=false** son itemsets muy frecuentes. Sin embargo, estudiar los items opuestos puede resultar bastante más interesante. Además cabe destacar la presencia superior del sexo masculino respecto al femenino. Será interesante poder ver si existe una presencia de enfermedad cardíaca en los hombres superior respecto a las mujeres.



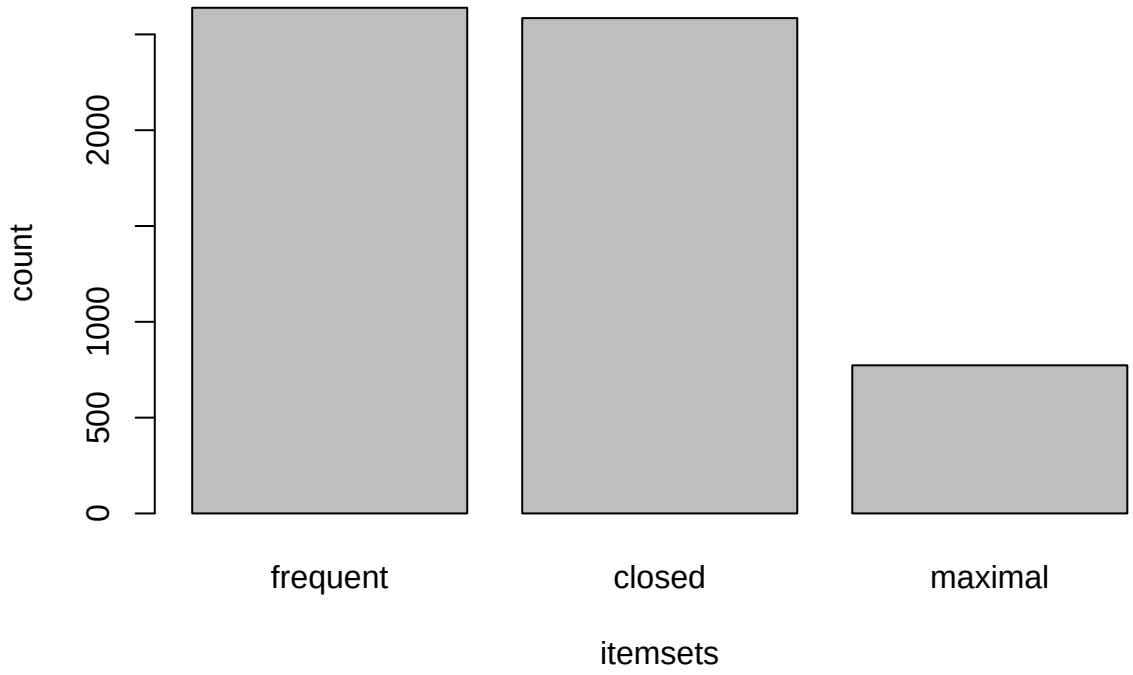
En el siguiente gráfico vemos la distribución de los tamaños de los itemsets frecuentes. Vemos cómo el tamaño más común es 5 items por itemset. También se muestran los primeros 10 itemsets frecuentes.

Usamos apriori para obtener los itemsets frecuentees. Primera información importante que **sex=male** o **fasting_blood_sugar=false** es muy frecuente. En el barplot se muestran los tamaños de itemsets frecuentes. Se ve cómo el tamaño de itemsets más frecuente es 4.




```
##      items                                      support   count
## [1]  {fasting_blood_sugar=false}                0.8552189 254
## [2]  {sex=male}                                 0.6767677 201
## [3]  {exang=false}                              0.6734007 200
## [4]  {ca=0}                                     0.5858586 174
## [5]  {fasting_blood_sugar=false,exang=false}    0.5757576 171
## [6]  {sex=male,fasting_blood_sugar=false}       0.5723906 170
## [7]  {age=(45,61]}                              0.5555556 165
## [8]  {thal=normal}                              0.5521886 164
## [9]  {num=healthy}                              0.5387205 160
## [10] {thalach=(136,169]}                        0.5319865 158
```

Vamos a ver cuál es la cantidad de itemsets frecuentes, de itemsets cerrados y de itemsets maximales. Como se muestra en el siguiente gráfico, hay una gran diferencia entre la cantidad de itemsets frecuentes y los cerrados y maximales.



Reglas general

A continuación empezamos ya a aplicar apriori para la obtención de reglas. En esta ejecución indicaré que el mínimo soporte sea de 0.1 y la mínima confianza de 0.79. También indicaré el mínimo del tamaño de las reglas, que será 2.

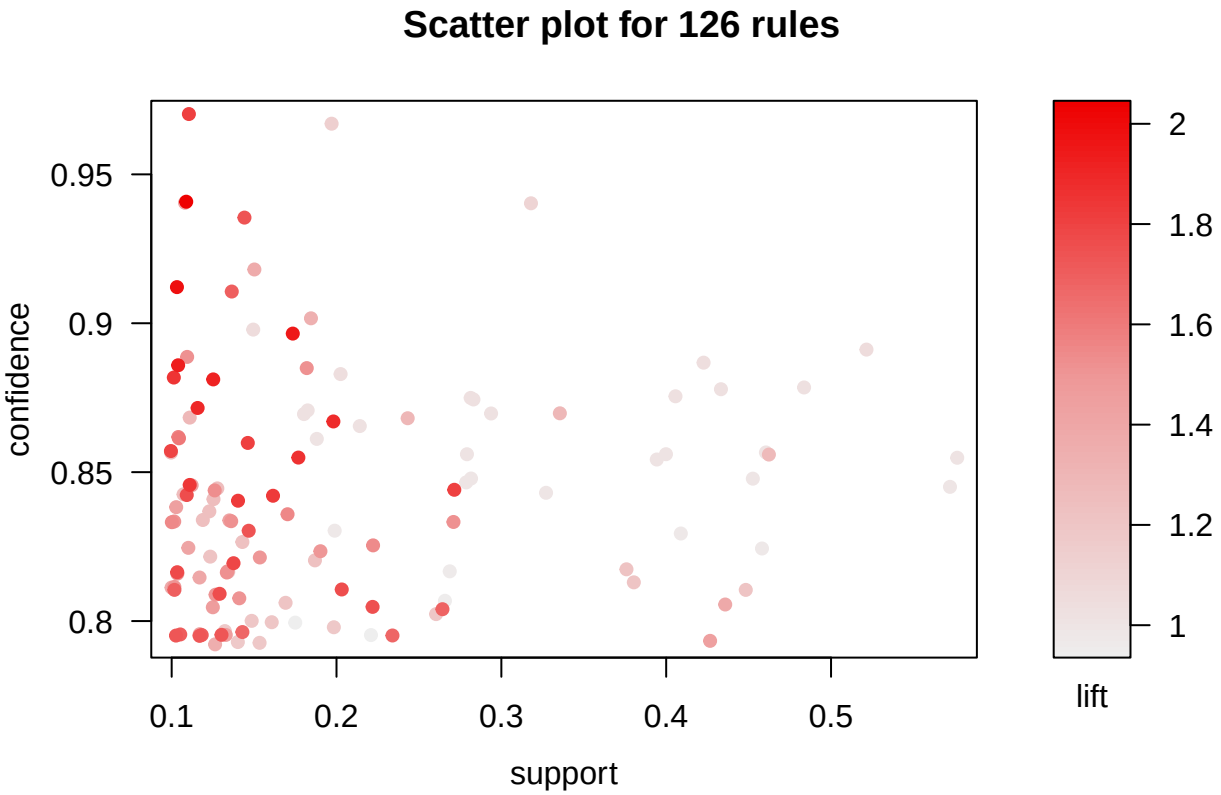
Posteriormente tras cada filtrado según los items que quiera explorar, aplicaré el filtro de reglas redundantes. También muestro un gráfico en el que se muestran como puntos las distintas reglas generadas por nuestro método.

Se muestra cómo las reglas con mejor Lift son por lo general con menor soporte que 0.3 o 0.4 y hay algunas reglas con un "confidence"" muy alto donde el valor es cercano a 1, sin embargo, estas reglas las consideraré como triviales siempre y cuando su soporte sea muy elevado, superior al 50%. Los lifts más "realistas" o con más equilibrio los encontramos en un nivel de confianza entre 0.5 y 0.8 aproximadamente. A continuación se muestran algunas reglas con más lift.

```
##      lhs                          rhs                support confidence     lift count
## [1]  {cp=asymptomatic,
##       ca=1}                    => {num=sick}        0.1077441  0.9411765 2.040361    32
## [2]  {thalach=(104,136],
##       thal=reversable defect} => {num=sick}        0.1043771  0.9117647 1.976599    31
## [3]  {exang=true,
##       thal=reversable defect} => {num=sick}        0.1750842  0.8965517 1.943619    52
## [4]  {slope=flat,
##       ca=1}                    => {num=sick}        0.1043771  0.8857143 1.920125    31
## [5]  {thalach=(104,136],
##       exang=true}              => {num=sick}        0.1245791  0.8809524 1.909802    37
## [6]  {cholestoral=[265,564],
##       exang=true}              => {num=sick}        0.1144781  0.8717949 1.889949    34
## [7]  {slope=flat,
##       thal=reversable defect} => {num=sick}        0.1986532  0.8676471 1.880957    59
## [8]  {exang=true,
##       slope=flat}              => {num=sick}        0.1784512  0.8548387 1.853190    53
```

```
## [9]  {thalach=(104,136],
##       thal=reversable defect} => {cp=asymptomatic} 0.1010101  0.8823529 1.845485    30
## [10] {cholestoral=[265,564],
##       thal=reversable defect} => {num=sick}        0.1111111  0.8461538 1.834363    33
```

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```



En las anteriores reglas se ve cómo la principal consecuencia que nos permite tener un lift bueno es **num=sick** es decir, si el usuario esta enfermo. Reglas interesantes a partir de estas, por ejemplo, cómo la reversión **thal=reversable defect** es un itemset que aparecec en varias reglas en el antececendente y en todas ellas el consecuente es un estado no saludable del paciente. Lo más destacable de este antecendente es el hecho de que independiente del resto de antecedentes, su aparición tienen un peso destacable en el consecuente enfermo. El colesterol alto **cholestoral=[265,564]** tiene una presencia en distintos itemsets. Será interesante su estudio y si tiene alguna relación respecto a la enfermedad coronaria. Otro aspecto a destacar y que contradice mis creencias al respecto, es que la edad no aparece en ninguna regla importante y por tanto no es un factor relevante.

Reglas específicas mediante filtrado

Como puede haber muchas de estas reglas, solo las reglas que contenían la clase 'enfermo' o 'saludable' en el lado derecho (RHS) seran considerados.

```
##      lhs                          rhs             support confidence     lift count
## [1]  {exang=false,
##       thal=normal}           => {num=healthy} 0.3737374  0.8345865 1.549201   111
## [2]  {exang=false,
##       ca=0}                  => {num=healthy} 0.3737374  0.8473282 1.572853   111
## [3]  {ca=0,
##       thal=normal}           => {num=healthy} 0.3434343  0.8869565 1.646413   102
```

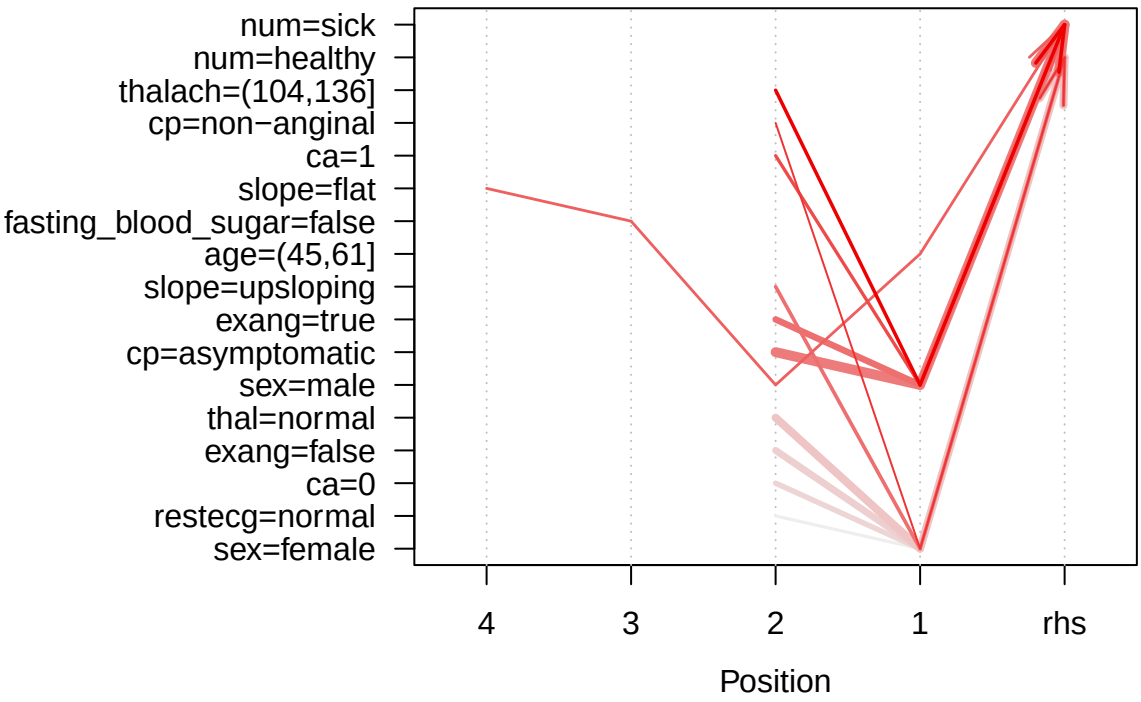
```
## [4]  {exang=false,
##       slope=upsloping}        => {num=healthy} 0.3030303  0.7964602 1.478429    90
## [5]  {slope=upsloping,
##       thal=normal}            => {num=healthy} 0.2895623  0.8600000 1.596375    86
## [6]  {sex=male,
##       cp=asymptomatic}        => {num=sick}    0.2727273  0.7941176 1.721554    81
## [7]  {slope=upsloping,
##       ca=0}                   => {num=healthy} 0.2659933  0.8681319 1.611470    79
## [8]  {restecg=normal,
##       thal=normal}            => {num=healthy} 0.2390572  0.8554217 1.587877    71
## [9]  {cp=asymptomatic,
##       thal=reversable defect} => {num=sick}    0.2356902  0.9090909 1.970803    70
## [10] {sex=female,
##       thal=normal}            => {num=healthy} 0.2323232  0.8625000 1.601016    69
```

En las anteriores reglas se puede observar como la angina inducida por el ejercicio es falsa **exang=false** y estado del corazón normal **thal=normal** demuestran que son buenos indicadores de la salud. Este dos items aparecen en varias ocasiones en los primeros 10 itemsets con un valor de confianza y de lift bastante buenos. La angina estable es un síndrome clínico caracterizado por malestar en el pecho, que aparece con el ejercicio o estrés emocional y remite con el descanso o con la administración de nitroglicerina. El número de vasos coloreados que son cero **ca=0** también indica condiciones saludables.

- {exang=false,thal=normal} => {num=healthy}
- {exang=false,ca=0} => {num=healthy}
- {ca=0,thal=normal} => {num=healthy}

Si filtramos todas reglas donde en el antecedente aparezca el género del paciente, nos fijamos que todas reglas para la clase "saludable" se atribuyeron al género femenino, lo que indica que, en base a este conjunto de datos en particular, las mujeres tienen más posibilidades de estar libres de enfermedad coronaria.

```
##      lhs                             rhs             support confidence     lift count
## [1]  {sex=male,
##       cp=asymptomatic}           => {num=sick}    0.2727273  0.7941176 1.721554    81
## [2]  {sex=female,
##       thal=normal}               => {num=healthy} 0.2323232  0.8625000 1.601016    69
## [3]  {sex=female,
##       exang=false}               => {num=healthy} 0.2121212  0.8513514 1.580321    63
## [4]  {sex=male,
##       exang=true}                => {num=sick}    0.2020202  0.8000000 1.734307    60
## [5]  {sex=female,
##       ca=0}                      => {num=healthy} 0.1818182  0.8437500 1.566211    54
## [6]  {sex=female,
##       slope=upsloping}           => {num=healthy} 0.1447811  0.9347826 1.735190    43
## [7]  {sex=male,
##       thalach=(104,136]}         => {num=sick}    0.1414141  0.8571429 1.858186    42
## [8]  {sex=male,
##       ca=1}                      => {num=sick}    0.1380471  0.8200000 1.777664    41
## [9]  {sex=female,
##       restecg=normal}            => {num=healthy} 0.1346801  0.8163265 1.515306    40
## [10] {age=(45,61],
##       sex=male,
##       fasting_blood_sugar=false,
##       slope=flat}                => {num=sick}    0.1279461  0.8085106 1.752757    38
```



- {sex=male,cp=asymptomatic} => {num=sick}
- {sex=female,thal=normal} => {num=healthy}
- {sex=female,exang=false} => {num=healthy}
- {sex=male,exang=true} => {num=sick}

Además, si los resultados mostraron que cuando la angina inducida por el ejercicio (dolor en el pecho) era falsa **exang=false**, era un buen indicador de que una persona estaba sana, independientemente del género (la angina inducida por el ejercicio = falsa ha aparecido en el LHS de todas las reglas de alta confianza).

- {sex=male,exang=true} => {num=sick}
- {sex=female,exang=false} => {num=healthy}

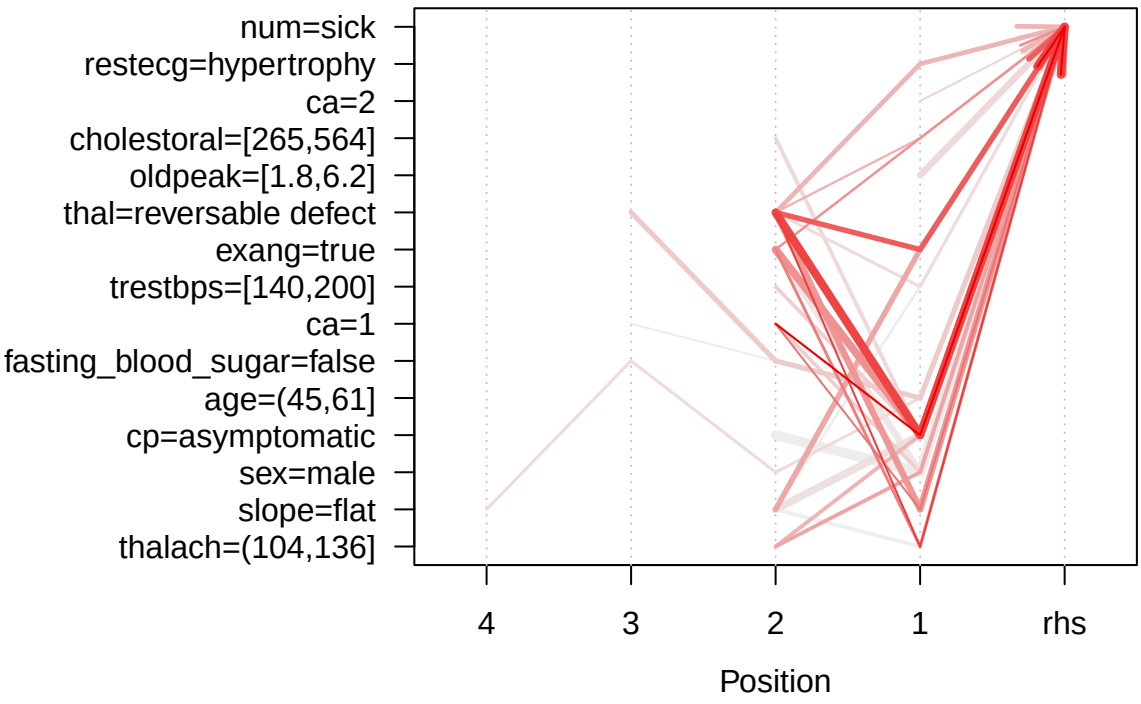
El número de vasos coloreados que son cero **ca=0** y que el estado del corazón (normal) **thal=normal** también se mostró como buenos indicadores de salud.

```
##      lhs                                    rhs           support
## [1]  {exang=false,thal=normal}           => {num=healthy} 0.3737374
## [2]  {exang=false,ca=0}                  => {num=healthy} 0.3737374
## [3]  {exang=false,slope=upsloping}       => {num=healthy} 0.3030303
## [4]  {cp=asymptomatic,exang=true}        => {num=sick}    0.2289562
## [5]  {sex=female,exang=false}            => {num=healthy} 0.2121212
## [6]  {exang=false,oldpeak=0.0}           => {num=healthy} 0.2053872
## [7]  {sex=male,exang=true}               => {num=sick}    0.2020202
## [8]  {cp=non-anginal,exang=false}        => {num=healthy} 0.1952862
## [9]  {exang=true,slope=flat}             => {num=sick}    0.1784512
## [10] {exang=true,thal=reversable defect} => {num=sick}    0.1750842
##      confidence lift     count
## [1]  0.8345865  1.549201 111
## [2]  0.8473282  1.572853 111
## [3]  0.7964602  1.478429  90
## [4]  0.8717949  1.889949  68
```

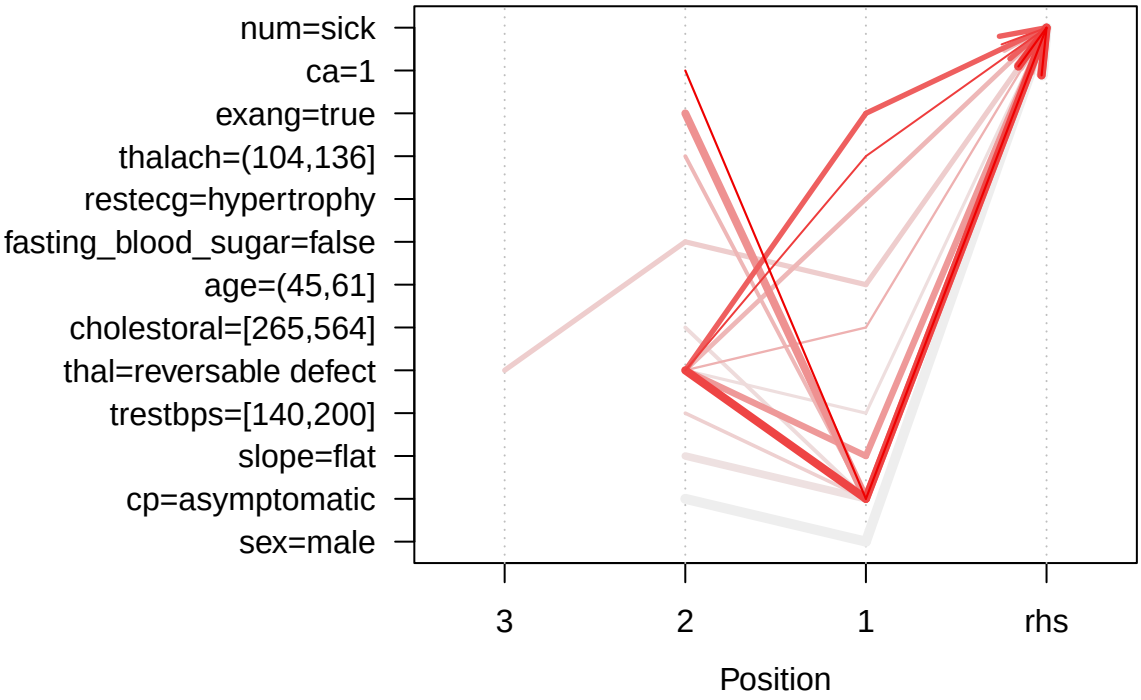
```
## [5]  0.8513514  1.580321  63
## [6]  0.7922078  1.470536  61
## [7]  0.8000000  1.734307  60
## [8]  0.8055556  1.495313  58
## [9]  0.8548387  1.853190  53
## [10] 0.8965517  1.943619  52
```

Las reglas filtradas para la clase "enfermo", por otro lado, mostraron que el tipo de dolor torácico es asintomático **cp=asymptomatic** y que la reversión **thal=reversable defect** es un indicador probable de que una persona está enferma (ambas reglas de alta confianza tienen estos dos factores en LHS).

```
##      lhs                              rhs          support confidence     lift count
## [1]  {sex=male,
##       cp=asymptomatic}           => {num=sick} 0.2727273  0.7941176 1.721554    81
## [2]  {cp=asymptomatic,
##       thal=reversable defect}    => {num=sick} 0.2356902  0.9090909 1.970803    70
## [3]  {cp=asymptomatic,
##       exang=true}                => {num=sick} 0.2289562  0.8717949 1.889949    68
## [4]  {cp=asymptomatic,
##       slope=flat}                => {num=sick} 0.2222222  0.8048780 1.744882    66
## [5]  {oldpeak=[1.8,6.2]}         => {num=sick} 0.2020202  0.8108108 1.757743    60
## [6]  {sex=male,
##       exang=true}                => {num=sick} 0.2020202  0.8000000 1.734307    60
## [7]  {slope=flat,
##       thal=reversable defect}    => {num=sick} 0.1986532  0.8676471 1.880957    59
## [8]  {exang=true,
##       slope=flat}                => {num=sick} 0.1784512  0.8548387 1.853190    53
## [9]  {exang=true,
##       thal=reversable defect}    => {num=sick} 0.1750842  0.8965517 1.943619    52
## [10] {age=(45,61],
##       fasting_blood_sugar=false,
##       thal=reversable defect}    => {num=sick} 0.1717172  0.8225806 1.783259    51
```



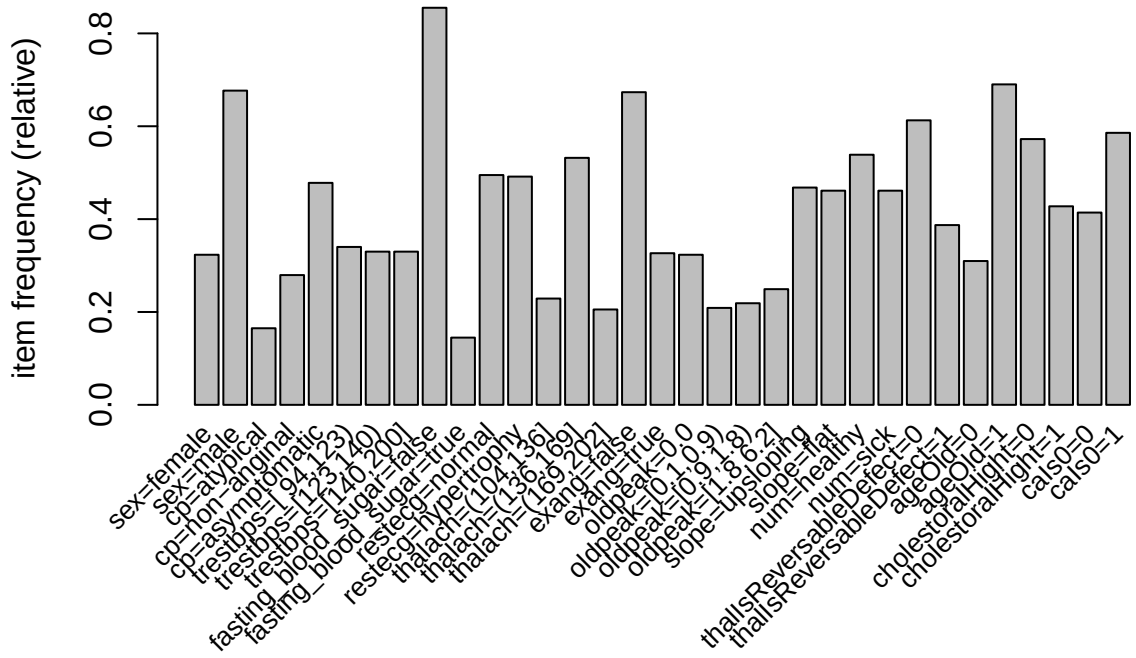

```
##      lhs                                rhs          support confidence     lift count
## [1]  {sex=male,
##       cp=asymptomatic}                => {num=sick} 0.2727273  0.7941176 1.721554    81
## [2]  {cp=asymptomatic,
##       thal=reversable defect}         => {num=sick} 0.2356902  0.9090909 1.970803    70
## [3]  {cp=asymptomatic,
##       exang=true}                     => {num=sick} 0.2289562  0.8717949 1.889949    68
## [4]  {cp=asymptomatic,
##       slope=flat}                     => {num=sick} 0.2222222  0.8048780 1.744882    66
## [5]  {slope=flat,
##       thal=reversable defect}         => {num=sick} 0.1986532  0.8676471 1.880957    59
## [6]  {exang=true,
##       thal=reversable defect}         => {num=sick} 0.1750842  0.8965517 1.943619    52
## [7]  {age=(45,61],
##       fasting_blood_sugar=false,
##       thal=reversable defect}         => {num=sick} 0.1717172  0.8225806 1.783259    51
## [8]  {restecg=hypertrophy,
##       thal=reversable defect}         => {num=sick} 0.1616162  0.8421053 1.825586    48
## [9]  {cp=asymptomatic,
##       thalach=(104,136]}              => {num=sick} 0.1447811  0.8431373 1.827823    43
## [10] {cp=asymptomatic,
##       cholestoral=[265,564]}          => {num=sick} 0.1447811  0.8113208 1.758849    43
```



A continuación voy a probar a ver reglas con items negados.

Voy a negar la variable cp y ca y el colesterol.

```
## 'data.frame':    297 obs. of  14 variables:
##  $ sex                   : Factor w/ 2 levels "female","male": 2 2 2 2 1 2 1 1 2 2 ...
##  $ cp                    : Factor w/ 4 levels "typical","atypical",..: 1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps              : Factor w/ 3 levels "[ 94,123)","[123,140)",..: 3 3 1 2 2 1 3 1 2 3 ...
##  $ fasting_blood_sugar   : Factor w/ 2 levels "false","true": 2 1 1 1 1 1 1 1 1 2 ...
##  $ restecg               : Factor w/ 3 levels "normal","stt",..: 3 3 3 1 3 1 3 1 3 3 ...
##  $ thalach               : Factor w/ 4 levels "(70.9,104]","(104,136]",..: 3 2 2 4 4 4 3 3 3 3 ...
##  $ exang                 : Factor w/ 2 levels "false","true": 1 2 2 1 1 1 1 2 1 2 ...
##  $ oldpeak               : Factor w/ 4 levels "0.0","[0.1,0.9)",..: 4 3 4 4 3 2 4 2 3 4 ...
##  $ slope                 : Factor w/ 3 levels "upsloping","flat",..: 3 2 2 3 1 1 3 1 2 3 ...
##  $ num                   : Factor w/ 2 levels "healthy","sick": 1 2 2 1 1 1 2 1 2 2 ...
##  $ thalIsReversableDefect: Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 2 2 ...
##  $ ageOld                : Factor w/ 2 levels "0","1": 2 2 2 1 1 2 2 2 2 2 ...
##  $ cholestoralHight      : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 2 2 2 1 ...
##  $ caIs0                 : Factor w/ 2 levels "0","1": 2 1 1 2 2 2 1 2 1 2 ...
##  - attr(*, "na.action")= 'omit' Named int  88 167 193 267 288 303
##    ..- attr(*, "names")= chr  "88" "167" "193" "267" ...
```




```
##      lhs                            rhs                          support confidence      lift count
## [1]  {exang=false,
##       thalIsReversableDefect=0} => {fasting_blood_sugar=false} 0.4141414  0.8601399 1.0057541   123
## [2]  {exang=false,
##       caIs0=1}                  => {fasting_blood_sugar=false} 0.3939394  0.8931298 1.0443289   117
## [3]  {num=healthy,
##       thalIsReversableDefect=0} => {exang=false}               0.3905724  0.8721805 1.2951880   116
## [4]  {num=healthy,
##       thalIsReversableDefect=0} => {fasting_blood_sugar=false} 0.3905724  0.8721805 1.0198330   116
## [5]  {exang=false,
##       num=healthy}              => {fasting_blood_sugar=false} 0.3905724  0.8467153 0.9900569   116
## [6]  {num=healthy,
##       caIs0=1}                  => {fasting_blood_sugar=false} 0.3838384  0.8837209 1.0333272   114
## [7]  {thalIsReversableDefect=0,
##       caIs0=1}                  => {fasting_blood_sugar=false} 0.3771044  0.9105691 1.0647206   112
## [8]  {num=healthy,
##       caIs0=1}                  => {exang=false}               0.3737374  0.8604651 1.2777907   111
## [9]  {sex=male,
##       ageOld=1}                 => {fasting_blood_sugar=false} 0.3670034  0.8014706 0.9371526   109
## [10] {num=healthy,
##       caIs0=1}                  => {thalIsReversableDefect=0}  0.3636364  0.8372093 1.3662152   108
```

Se muestran las 10 primeras reglas ordenadas por Lift. Observándolas, vemos como no obtenemos ninguna regla demasiado informativa. Quizás la más destacable sea que si el el número de vasos principales coloreados por fluoroscopia **ca** es distinto de 0 es muy probables que tengan una enfermedad coronaria. Respecto a la variable ageOld no nos aporta ninguna regla como al igual sucede con el colesterol alto. Aparentemente no existe relacion ninguna entre el colesterol y la enfermedad, por tanto descartamos las hipotesi hecha al principio del estudio hacerca del colesterol.

CONCLUSIONES

Esta investigación ha presentado un experimento de extracción de reglas en datos de enfermedades del corazón utilizando algoritmos de extracción de reglas (Apriori). Se realizó un análisis adicional basado en la minería

de reglas al clasificar los datos según el género y se encontraron factores de riesgo significativos para las enfermedades cardíacas tanto para hombres como para mujeres. Curiosamente, se encuentra en el conjunto de reglas saludables, ser “femenino” es uno de los factores para una condición cardíaca saludable. En otras palabras, los resultados indicaron que las mujeres tienen más probabilidades de estar libres de enfermedad coronaria que los hombres.