Examen 2

Efrén Jiménez

7 de diciembre de 2016

Pregunta 1

Árbol de Decisión:Es un modelo utilizado en el ámbito de la inteligencia artificial, por medio de construcciones de diagramas lógicos que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva.

Red Neuronal:Es un modelo utilizado en el ámbito de la inteligencia artificial, inspirado en la forma en que funciona el sistema nervioso biológico en un sistema de interconexión de neuromas que colaboran entre sí para obtener un elemento de salida.

Bosque Aleatorio:Es un modelo que utiliza una combinación de árboles de decisión, dado que cada uno de estos árboles depende de un vector aleatorio y una misma distribución para cada uno.

Serie de Tiempo: Es una secuencia de observaciones obtenidos en un momento determinado y ordenados cronológicamente.

Aprendizaje en Conjunto:Es una técnica para deducir una valor a partir de datos de entrenamiento, el objetivo del aprendizajes es ser capaz de predecir el valor correspondientes a cualquier entrada de valor valida.

Agrupamiento Jerárquico:Es un método de análisis de grupos el cual busca construir una jerarquía en distintos grupos.

Agrupamiento por Particionamiento:Es un método de análisis de grupos el cual busca construir una distancia o similitud entre el conjunto de datos.

Matriz Documento – Términos:Es la frecuencia de ocurrencia de un término en la colección de documentos, es una medida numérica que indica el nivel de relevancia de una palabra en un documento.

Segmentación o Tokenización ("Tokenization"):Es la función de separar las palabras o frases por una longitud n.

Autoregresión:Es un proceso estadístico para estimar las relaciones entre variables,en donde se incluye varias técnicas para el modelado y análisis de diversas variables y la relación entre una variable dependiente y una o más variables independientes.

Pregunta 2

Análisis del Problema

La piel es el órgano principal de localización de las infecciones en el hombre, siendo estas infecciones clasificadas en superficiales y profundas. La incidencia del eritematoso escamoso aumenta desde hace años en todo el mundo y muchas veces la adquisición de conocimientos imprescindibles sobre el tema se dificultan por la existencia de diversas tendencias para su estudio. Para realizar un estudio en este conjunto de datos, se ha propuesto un análisis para obtener la clasificación del historial familiar en cada una de las nuevas observaciones.

Entendimiento de los Datos

Este conjunto de datos contiene 365 observaciones y 35 variables.

- **Eritema** : Valor numérico entre 0 y 3.
- **Escalamiento** : Valor numérico entre 0 y 3.
- **Fronteras definidas** : Valor numérico entre 0 y 3.
- **Picazón** : Valor numérico entre 0 y 3.
- **Fenómeno koebner** : Valor numérico entre 0 y 3.
- **Pápulas poligonales** : Valor numérico entre 0 y 3.
- **Pápulas foliculares** : Valor numérico entre 0 y 3.
- **Compromiso de la mucosa oral** : Valor numérico entre 0 y 3.
- **Afectación de la rodilla y del codo** : Valor numérico entre 0 y 3.
- **Afectación del cuero cabelludo** : Valor numérico entre 0 y 3.
- **Historia familiar** : Valor numérico entre 0 y 1.
- **Incontinencia de melanina** : Valor numérico entre 0 y 3.
- **Eosinófilos en el infiltrado** : Valor numérico entre 0 y 2.
- **Infiltrado de PNL** : Valor numérico entre 0 y 3.
- **Fibrosis de la dermis papilar** : Valor numérico entre 0 y 3.
- **Exocitosis** : Valor numérico entre 0 y 3.
- **Acantosis** : Valor numérico entre 0 y 3.
- **Hiperqueratosis** : Valor numérico entre 0 y 3.
- **Paraqueratosis** : Valor numérico entre 0 y 3.
- **Clubbing de las crestas rete** : Valor numérico entre 0 y 3.
- **Elongación de las crestas rete** : Valor numérico entre 0 y 3.
- **Adelgazamiento de la epidermis suprapapilar** : Valor numérico entre 0 y 3.
- **Pústula espongiforme** : Valor numérico entre 0 y 3.
- **Microabcès munro** : Valor numérico entre 0 y 3.
- **Hipergranulosis focal** : Valor numérico entre 0 y 3.
- **Desaparición de la capa granular** : Valor numérico entre 0 y 3.
- **Vacuolización y daño de la capa basal** : Valor numérico entre 0 y 3.
- **Espongiosis** : Valor numérico entre 0 y 3.
- **Aspecto de retes de los dientes de sierra** : Valor numérico entre 0 y 3.
- **Enchufe de cuerno folicular** : V Valor numérico entre 0 y 3.alor numérico entre 0 y 3.
- **Paraqueratosis perifolicular** : Valor numérico entre 0 y 3.
- **Inflamación monoluclear inflamatoria** : Valor numérico entre 0 y 3.
- **Infiltrado en banda** : Valor numérico entre 0 y 3.
- **Edad** : Valor numérico entre 0 y 75.

Exploración de los Datos

```
#librerías utilizadas
library(caTools)
library(rpart)
library(rpart.plot)
library(rattle)
```

```
## Rattle: A free graphical interface for data mining with R.
## Versión 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
## Escriba 'rattle()' para agitar, sacudir y  rotar sus datos.
```

```
library(lattice)
library(neuralnet)
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

##
## Attaching package: 'ROCR'

## The following object is masked from 'package:neuralnet':
##
##     prediction
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(randomForest)
```

```
## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.
```

```
#Setear el lugar de trabajo
setwd('D:\\Drive\\Universidad\\Cenfotec\\MBD\\2016 Cuatrimestre 3\\MBD-305 Minería de datos 1\\Semana 1

#Cargar los datos
datos <- read.csv('dermatology.csv',na.strings = "?")

#Cargar el nombre d las columnas
colnames(datos) <- c("erythema","scaling","definite","itching","koebner","polygonal","follicular","oral

#Crear los factores
datos$erythema=factor(datos$erythema)
datos$scaling=factor(datos$scaling)
datos$definite=factor(datos$definite)
datos$itching=factor(datos$itching)
datos$koebner=factor(datos$koebner)
datos$polygonal=factor(datos$polygonal)
datos$follicular=factor(datos$follicular)
datos$oral=factor(datos$oral)
datos$knee=factor(datos$knee)
```

```r
datos$scalp=factor(datos$scalp)
datos$family=factor(datos$family)
datos$melanin=factor(datos$melanin)
datos$eosinophils=factor(datos$eosinophils)
datos$PNL=factor(datos$PNL)
datos$fibrosis=factor(datos$fibrosis)
datos$exocytosis=factor(datos$exocytosis)
datos$acanthosis=factor(datos$acanthosis)
datos$hyperkeratosis=factor(datos$hyperkeratosis)
datos$parakeratosis=factor(datos$parakeratosis)
datos$clubbing=factor(datos$clubbing)
datos$elongation=factor(datos$elongation)
datos$thinning=factor(datos$thinning)
datos$spongiform=factor(datos$spongiform)
datos$munro=factor(datos$munro)
datos$focal=factor(datos$focal)
datos$disappearance=factor(datos$disappearance)
datos$vacuolisation=factor(datos$vacuolisation)
datos$spongiosis=factor(datos$spongiosis)
datos$saw=factor(datos$saw)
datos$follicular=factor(datos$follicular)
datos$perifollicular=factor(datos$perifollicular)
datos$inflammatory=factor(datos$inflammatory)
datos$band=factor(datos$band)

#Visualizar los datos
str(datos)
```

```
## 'data.frame':    365 obs. of  35 variables:
##  $ erythema      : Factor w/ 4 levels "0","1","2","3": 4 3 3 3 3 3 3 3 3 4 ...
##  $ scaling       : Factor w/ 4 levels "0","1","2","3": 4 2 3 4 4 2 3 3 3 4 ...
##  $ definite      : Factor w/ 4 levels "0","1","2","3": 4 3 3 3 3 1 4 2 2 3 ...
##  $ itching       : Factor w/ 4 levels "0","1","2","3": 3 4 1 3 1 3 4 1 1 2 ...
##  $ koebner       : Factor w/ 4 levels "0","1","2","3": 2 2 1 3 1 1 4 3 2 2 ...
##  $ polygonal     : Factor w/ 4 levels "0","1","2","3": 1 4 1 3 1 1 4 1 1 1 ...
##  $ follicular    : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ oral          : Factor w/ 4 levels "0","1","2","3": 1 4 1 3 1 1 3 1 1 1 ...
##  $ knee          : Factor w/ 4 levels "0","1","2","3": 2 1 4 1 1 1 1 1 1 3 ...
##  $ scalp         : Factor w/ 4 levels "0","1","2","3": 2 1 3 1 1 1 1 1 1 3 ...
##  $ family        : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 2 ...
##  $ melanin       : Factor w/ 4 levels "0","1","2","3": 1 2 1 2 1 1 3 1 1 1 ...
##  $ eosinophils   : Factor w/ 3 levels "0","1","2": 1 1 1 1 3 1 1 1 1 1 ...
##  $ PNL           : Factor w/ 4 levels "0","1","2","3": 2 1 4 1 2 1 1 1 1 1 ...
##  $ fibrosis      : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 4 1 1 1 1 ...
##  $ exocytosis    : Factor w/ 4 levels "0","1","2","3": 2 2 1 2 3 2 3 3 4 1 ...
##  $ acanthosis    : Factor w/ 4 levels "0","1","2","3": 3 3 3 3 3 4 4 2 3 4 ...
##  $ hyperkeratosis: Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 3 ...
##  $ parakeratosis : Factor w/ 4 levels "0","1","2","3": 3 3 4 1 3 1 1 2 3 4 ...
##  $ clubbing      : Factor w/ 4 levels "0","1","2","3": 3 1 3 1 1 1 1 1 1 3 ...
##  $ elongation    : Factor w/ 4 levels "0","1","2","3": 3 1 3 1 1 3 1 1 1 3 ...
##  $ thinning      : Factor w/ 4 levels "0","1","2","3": 3 1 3 1 1 1 1 1 1 3 ...
##  $ spongiform    : Factor w/ 4 levels "0","1","2","3": 3 1 3 1 2 1 1 1 1 2 ...
##  $ munro         : Factor w/ 4 levels "0","1","2","3": 2 1 1 1 1 1 1 1 1 2 ...
```

```
##  $ focal         : Factor w/ 4 levels "0","1","2","3": 1 3 1 3 1 1 1 1 1 1 ...
##  $ disappearance : Factor w/ 4 levels "0","1","2","3": 1 1 4 3 1 1 3 1 1 1 ...
##  $ vacuolisation : Factor w/ 4 levels "0","1","2","3": 1 3 1 4 1 1 3 1 1 1 ...
##  $ spongiosis    : Factor w/ 4 levels "0","1","2","3": 1 4 1 3 3 1 4 3 3 1 ...
##  $ saw           : Factor w/ 4 levels "0","1","2","3": 1 3 1 4 1 1 3 1 1 1 ...
##  $ follicular    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ perifollicular: Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ inflammatory  : Factor w/ 4 levels "0","1","2","3": 2 3 4 3 2 3 4 3 3 2 ...
##  $ band          : Factor w/ 4 levels "0","1","2","3": 1 4 1 4 1 1 4 1 1 1 ...
##  $ Age           : int  8 26 40 45 41 18 57 22 30 20 ...
##  $ unknown       : int  1 3 1 3 2 5 3 4 4 1 ...
```

```
#Datos resumidos
summary(datos)
```

```
##  erythema scaling definite itching koebner polygonal follicular oral
##  0:  4    0:  8   0: 58    0:118   0:223   0:296     0:332      0:298
##  1: 57    1:111   1: 93    1: 72   1: 70   1:  1     1: 11      1:  9
##  2:214    2:194   2:168    2:100   2: 54   2: 41     2: 16      2: 45
##  3: 90    3: 52   3: 46    3: 75   3: 18   3: 27     3:  6      3: 13
##
##
##
##  knee    scalp   family  melanin eosinophils PNL     fibrosis exocytosis
##  0:251   0:263   0:319   0:295   0:323       0:234   0:311    0:118
##  1: 27   1: 30   1: 46   1:  8   1: 33       1: 69   1:  8    1: 57
##  2: 64   2: 56           2: 46   2:  9       2: 55   2: 23    2:129
##  3: 23   3: 16           3: 16               3:  7   3: 23    3: 61
##
##
##
##  acanthosis hyperkeratosis parakeratosis clubbing elongation thinning
##  0: 10      0:226          0: 85         0:251    0:197      0:255
##  1: 71      1: 90          1:118         1: 19    1: 23      1: 19
##  2:209      2: 44          2:132         2: 61    2: 95      2: 60
##  3: 75      3:  5          3: 30         3: 34    3: 50      3: 31
##
##
##
##  spongiform munro   focal   disappearance vacuolisation spongiosis saw
##  0:295      0:285   0:294   0:272         0:293         0:199      0:293
##  1: 38      1: 37   1: 13   1: 30         1:  3         1: 28      1:  5
##  2: 26      2: 33   2: 43   2: 49         2: 43         2: 96      2: 40
##  3:  6      3: 10   3: 15   3: 14         3: 26         3: 42      3: 27
##
##
##
##    follicular     perifollicular inflammatory band         Age
##  Min.   :0.0000   0:344          0: 13        0:288   Min.   : 0.00
##  1st Qu.:0.0000   1:  4          1: 84        1:  3   1st Qu.:25.00
##  Median :0.0000   2: 13          2:206        2: 22   Median :35.00
##  Mean   :0.1041   3:  4          3: 62        3: 52   Mean   :36.24
##  3rd Qu.:0.0000                                      3rd Qu.:49.00
##  Max.   :3.0000                                      Max.   :75.00
```

```
##                                                 NA's   :8
##     unknown
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :3.000
##  Mean   :2.805
##  3rd Qu.:4.000
##  Max.   :6.000
##
```

```
#dividir el conjunto de datos en entrenamiento y prueba
set.seed(1234)
splt <- sample.split(datos$family, SplitRatio = 0.7)
entrenamiento <- datos[splt, ]
prueba <- datos[!splt, ]

#Datos resumidos de entrenamiento
summary(entrenamiento)
```

```
##  erythema scaling definite itching koebner polygonal follicular oral
##  0:  1    0:  5   0: 42    0:87    0:152   0:204     0:231      0:206
##  1: 44    1: 80   1: 66    1:47    1: 54   1:  1     1:  7      1:  7
##  2:144    2:138   2:113    2:71    2: 37   2: 29     2: 12      2: 34
##  3: 66    3: 32   3: 34    3:50    3: 12   3: 21     3:  5      3:  8
##
##
##
##  knee    scalp   family  melanin eosinophils PNL     fibrosis exocytosis
##  0:179   0:188   0:223   0:204   0:226       0:166   0:217    0:76
##  1: 20   1: 21   1: 32   1:  5   1: 22       1: 43   1:  6    1:41
##  2: 43   2: 36           2: 35   2:  7       2: 42   2: 17    2:94
##  3: 13   3: 10           3: 11               3:  4   3: 15    3:44
##
##
##
##  acanthosis hyperkeratosis parakeratosis clubbing elongation thinning
##  0:  6      0:158          0:64          0:181    0:141      0:181
##  1: 48      1: 63          1:84          1: 11    1: 15      1: 14
##  2:148      2: 30          2:90          2: 42    2: 66      2: 41
##  3: 53      3:  4          3:17          3: 21    3: 33      3: 19
##
##
##
##  spongiform munro   focal   disappearance vacuolisation spongiosis saw
##  0:208      0:206   0:203   0:192         0:202         0:137      0:203
##  1: 25      1: 24   1:  9   1: 23         1:  3         1: 18      1:  4
##  2: 17      2: 21   2: 32   2: 29         2: 32         2: 71      2: 31
##  3:  5      3:  4   3: 11   3: 11         3: 18         3: 29      3: 17
##
##
##
##    follicular     perifollicular inflammatory band         Age
##  Min.   :0.0000   0:240          0:  8        0:201   Min.   : 0.00
##  1st Qu.:0.0000   1:  1          1: 59        1:  0   1st Qu.:26.25
```

```
##  Median :0.0000   2: 10          2:142        2: 16   Median :36.00
##  Mean   :0.1176   3:  4          3: 46        3: 38   Mean   :37.22
##  3rd Qu.:0.0000                                      3rd Qu.:50.00
##  Max.   :3.0000                                      Max.   :75.00
##                                                      NA's   :5
##     unknown
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :3.000
##  Mean   :2.851
##  3rd Qu.:4.000
##  Max.   :6.000
##
```

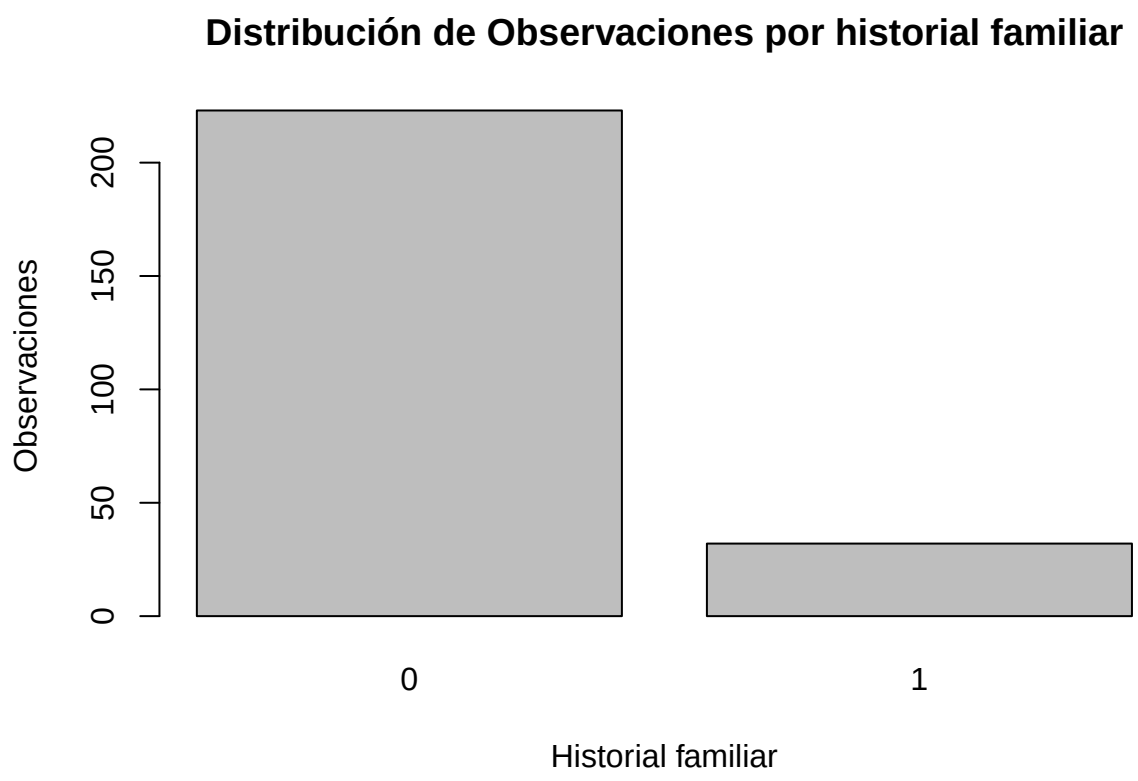
```
#Datos resumidos de prueba
summary(prueba)
```

```
##  erythema scaling definite itching koebner polygonal follicular oral
##  0: 3     0: 3    0:16     0:31    0:71    0:92      0:101      0:92
##  1:13     1:31    1:27     1:25    1:16    1: 0      1:  4      1: 2
##  2:70     2:56    2:55     2:29    2:17    2:12      2:  4      2:11
##  3:24     3:20    3:12     3:25    3: 6    3: 6      3:  1      3: 5
##
##
##
##  knee   scalp  family melanin eosinophils PNL    fibrosis exocytosis
##  0:72   0:75   0:96   0:91    0:97        0:68   0:94     0:42
##  1: 7   1: 9   1:14   1: 3    1:11        1:26   1: 2     1:16
##  2:21   2:20          2:11    2: 2        2:13   2: 6     2:35
##  3:10   3: 6          3: 5                3: 3   3: 8     3:17
##
##
##
##  acanthosis hyperkeratosis parakeratosis clubbing elongation thinning
##  0: 4       0:68           0:21          0:70     0:56       0:74
##  1:23       1:27           1:34          1: 8     1: 8       1: 5
##  2:61       2:14           2:42          2:19     2:29       2:19
##  3:22       3: 1           3:13          3:13     3:17       3:12
##
##
##
##  spongiform munro  focal  disappearance vacuolisation spongiosis saw
##  0:87       0:79   0:91   0:80          0:91          0:62       0:90
##  1:13       1:13   1: 4   1: 7          1: 0          1:10       1: 1
##  2: 9       2:12   2:11   2:20          2:11          2:25       2: 9
##  3: 1       3: 6   3: 4   3: 3          3: 8          3:13       3:10
##
##
##
##    follicular      perifollicular inflammatory band        Age
##  Min.   :0.00000   0:104          0: 5         0:87   Min.   : 7.00
##  1st Qu.:0.00000   1:  3          1:25         1: 3   1st Qu.:22.00
##  Median :0.00000   2:  3          2:64         2: 6   Median :33.00
##  Mean   :0.07273   3:  0          3:16         3:14   Mean   :33.96
```

```
##  3rd Qu.:0.00000                                          3rd Qu.:45.50
##  Max.   :2.00000                                          Max.   :62.00
##                                                           NA's   :3
##     unknown
##  Min.   :1.0
##  1st Qu.:1.0
##  Median :2.0
##  Mean   :2.7
##  3rd Qu.:4.0
##  Max.   :6.0
##
```

En este grafico podemos observar que los registros son mayor mente sin historial.

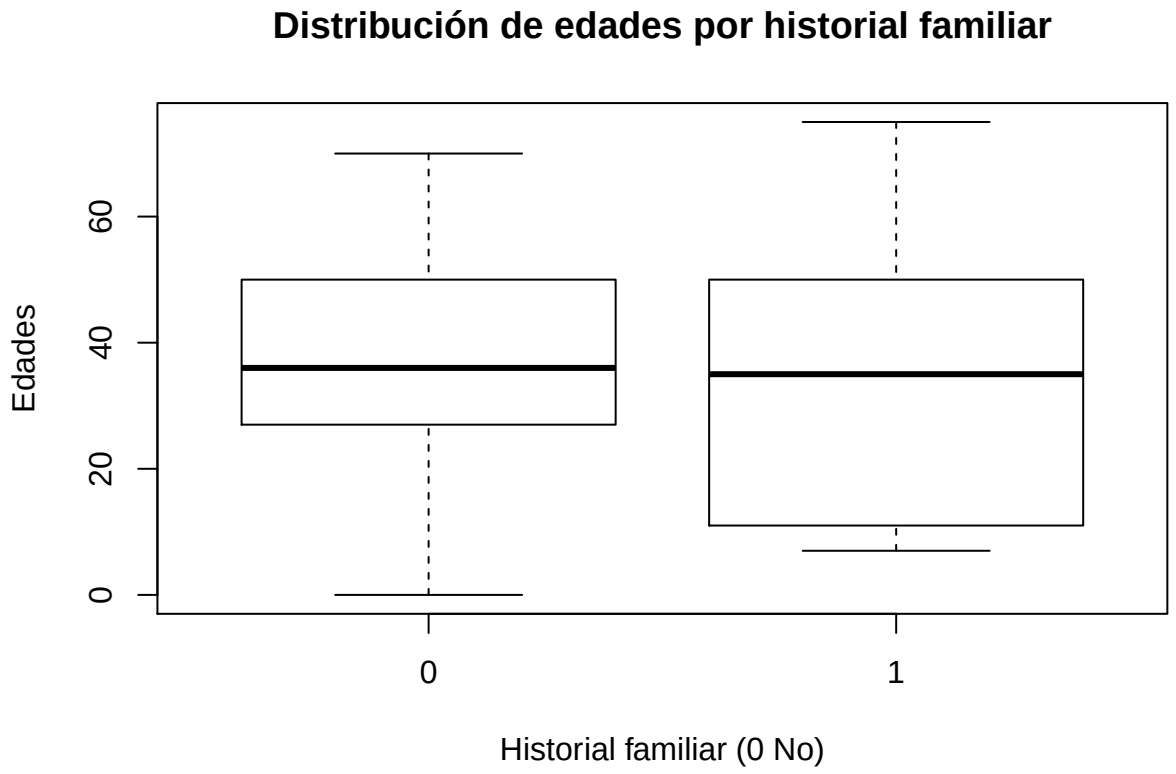
```
barplot(table(entrenamiento$family),
main = 'Distribución de Observaciones por historial familiar',
ylab = 'Observaciones',
xlab = 'Historial familiar')
```



En este grafico podemos observar que los registros que no tienen historial familiar rondan los 25 a 50 años y los que si entre los 15 y 50 años.

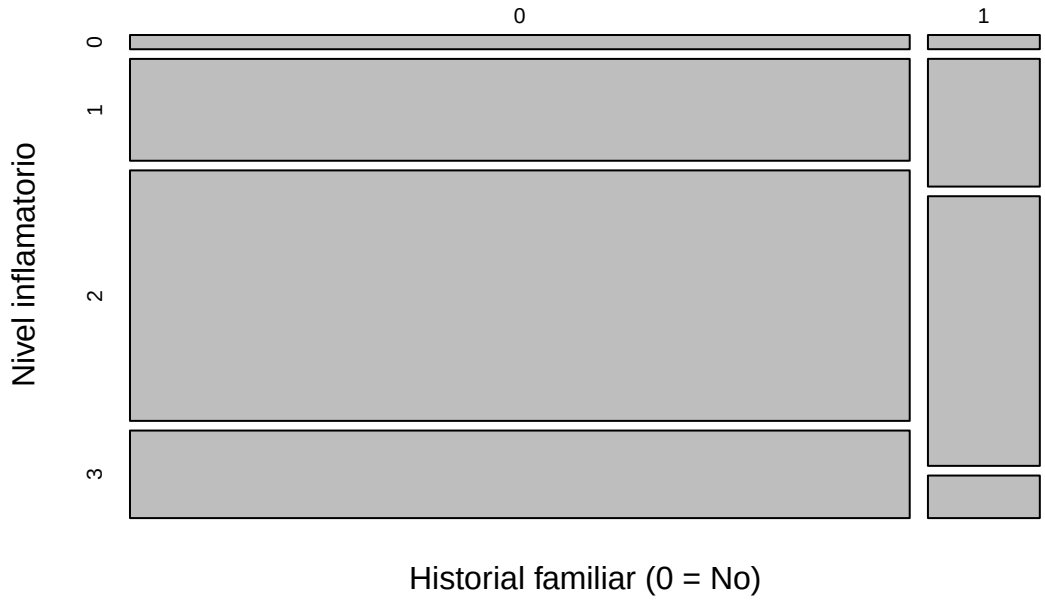
```
boxplot(Age ~ family,
data = entrenamiento,
main = 'Distribución de edades por historial familiar',
```

```
xlab = 'Historial familiar (0 No)',
ylab = 'Edades')
```



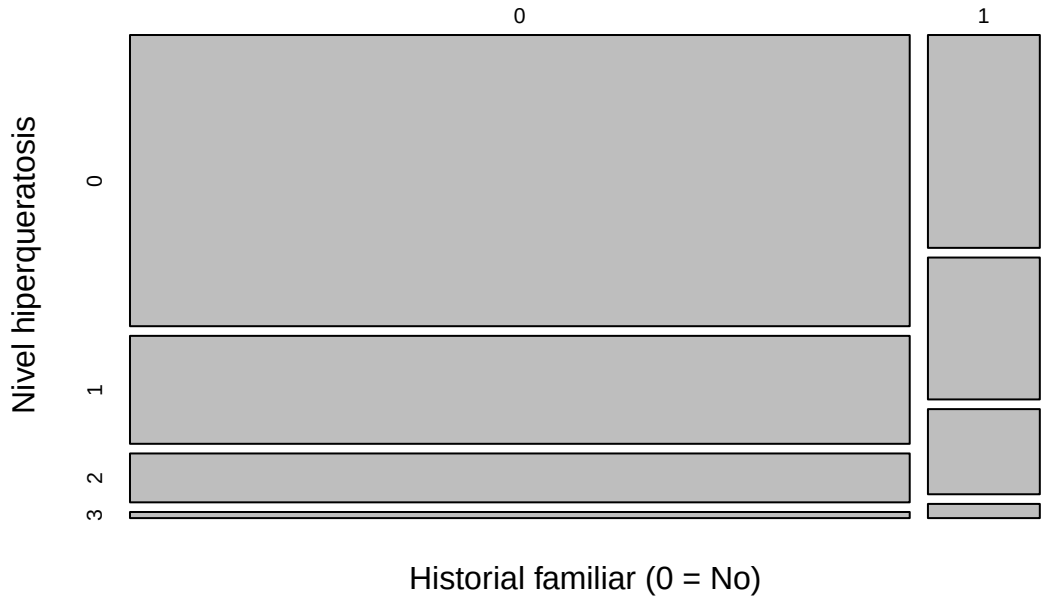
En este grafico podemos observar que los registros que no y si tienen historial tienen mayormente en el valor 2 en inflamantorio.

```
mosaicplot(~entrenamiento$family + entrenamiento$inflammatory,
main = 'Proporción de historial familiar por el nivel inflamatorio',
ylab = 'Nivel inflamatorio',
xlab = 'Historial familiar (0 = No)')
```



En este grafico podemos observar que los registros que no y si tienen historial tienen mayormente en el valor 0 en hiperqueratosis.

```
mosaicplot(~entrenamiento$family + entrenamiento$hyperkeratosis,
main = 'Proporción de historial familiar por el nivel hiperqueratosis',
ylab = 'Nivel hiperqueratosis',
xlab = 'Historial familiar (0 = No)')
```



Creación del Modelo

En este caso, se van a utilizar una red neuronal, un árbol de decisión y un bosque aleatorio. Se comienza por crear la red neuronal y por aplicarla a los datos de prueba:

Modelo de Minería de Datos Redes neuronales.

```
#crear matrices numéricas para ser consumidas por las redes neuronales.
entrenamiento.red <- model.matrix(~ family+erythema + scaling + definite + itching + koebner + polygonal
data = entrenamiento)

prueba.red <- model.matrix(~ family+erythema + scaling + definite + itching + koebner + polygonal + foll
data = prueba)

#Ajustar los nombres de las columnas
colnames(entrenamiento.red) <- make.names(colnames(entrenamiento.red))
colnames(prueba.red) <- make.names(colnames(prueba.red))

set.seed(12345)

#crear red neuronal con 7 unidades en la capa oculta
modelo.red <- neuralnet(family1 ~ erythema1+erythema2+erythema3+scaling1+scaling2+scaling3+definite1+de
itching1+itching2+itching3+koebner1+koebner2+koebner3+polygonal1+polygonal2+polygonal3+follicular1+foll
data = entrenamiento.red,hidden = 7)
```

```
#realizar predicciones
predicciones.red <- compute(modelo.red, prueba.red[ , c(3:ncol(prueba.red))])

detach("package:neuralnet", unload=TRUE) #descargar la librería neural net para poder usar la función p
```

Modelo de Minería de Datos (árbol de decisión)

También podemos crear un modelo utilizando árboles de decisión

```
#crear modelo
set.seed(12345)
modelo.arbol <- rpart(family ~ erythema + scaling + definite + itching + koebner + polygonal + follicula

#realizar predicciones
predicciones.arbol <- predict(modelo.arbol, newdata = prueba, type = 'prob')
```

Modelo de Minería de Datos (bosque aleatorio)

Finalmente, el bosque aleatorio:

```
#crear modelo
set.seed(12345)
modelo.bosque <- randomForest(family ~ erythema + scaling + definite + itching + koebner + polygonal + i

#realizar predicciones
predicciones.bosque <- predict(modelo.bosque, newdata = prueba, type = 'prob')
```

Evaluación

Evaluación (modelo ingenuo)

El primer punto de comparación es contra un modelo ingenuo: (pronostica siempre 'no')

```
modelo.ingenuo <- rep(0, nrow(prueba))
table(prueba$family, modelo.ingenuo)
```

```
##    modelo.ingenuo
##       0
##   0 96
##   1 14
```

```
prediccionROC.ingenuo <- prediction(modelo.ingenuo, prueba$family)
# ROC
as.numeric(performance(prediccionROC.ingenuo, "auc")@y.values)
```

```
## [1] 0.5
```

Métricas del modelo ingenuo:

- Exactitud: 87.27%
- Sensibilidad: 0%
- Especificidad: 100%
- Área bajo la curva: 50%

Evaluación (red neuronal)

La evaluación de la red neuronal.

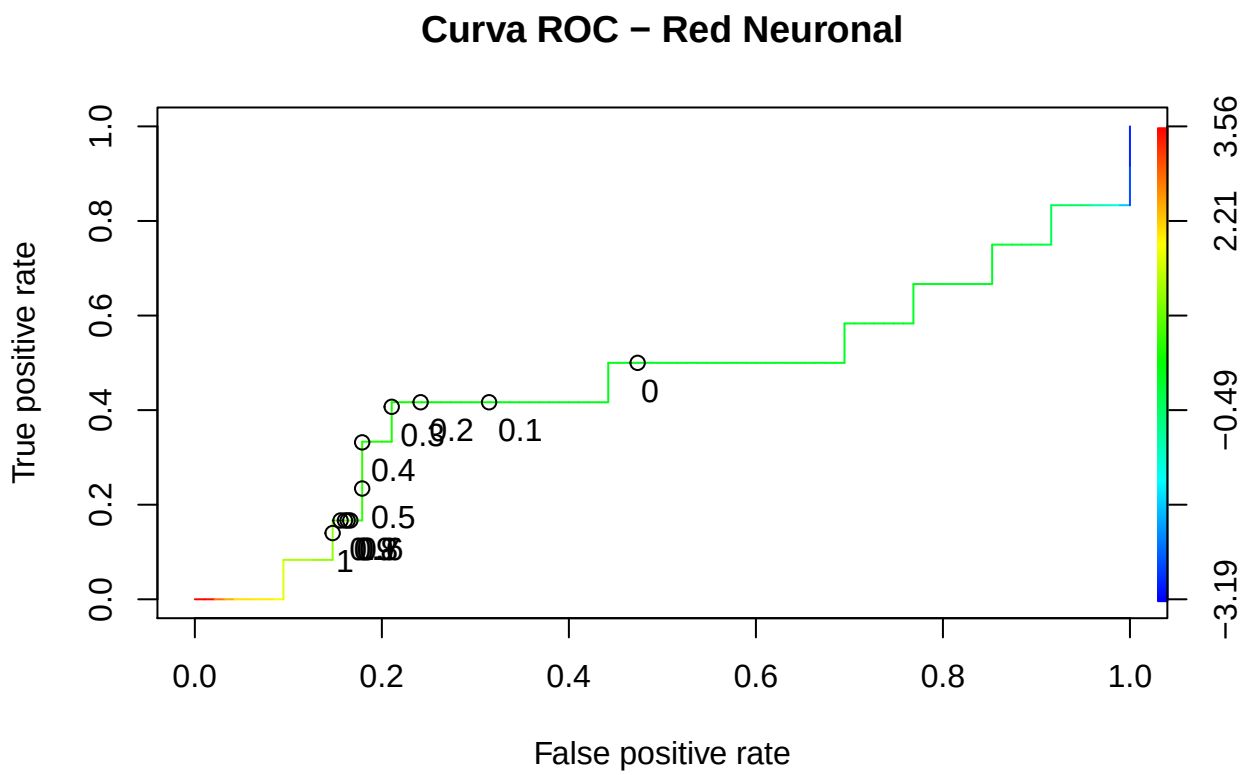
```
# Ver resultados de las redes
resultado.red <- table(prueba.red[, "family1"], predicciones.red$net.result >=
0.5)
resultado.red
```

```
##
##     FALSE TRUE
##   0    78   17
##   1    10    2
```

```
#predict de ROCR
prediccionROC.red <- prediction(predicciones.red$net.result, prueba.red[, "family1"])

# ROC
as.numeric(performance(prediccionROC.red, "auc")@y.values)
```

```
## [1] 0.4596491228
```

```
ROCR.red <- performance(prediccionROC.red, "tpr", "fpr")
plot(ROCR.red, main = "Curva ROC - Red Neuronal", colorize = TRUE, print.cutoffs.at = seq(0,
1, by = 0.1), text.adj = c(-0.2, 1.7))
```



Métricas del Modelo red neuronal:

- Exactitud: 74.76%
- Sensibilidad: 16.66%
- Especificidad: 82.10%
- Área bajo la curva: 69.21%

Evaluación (árbol de decisión)

La evaluación del árbol de decisión.

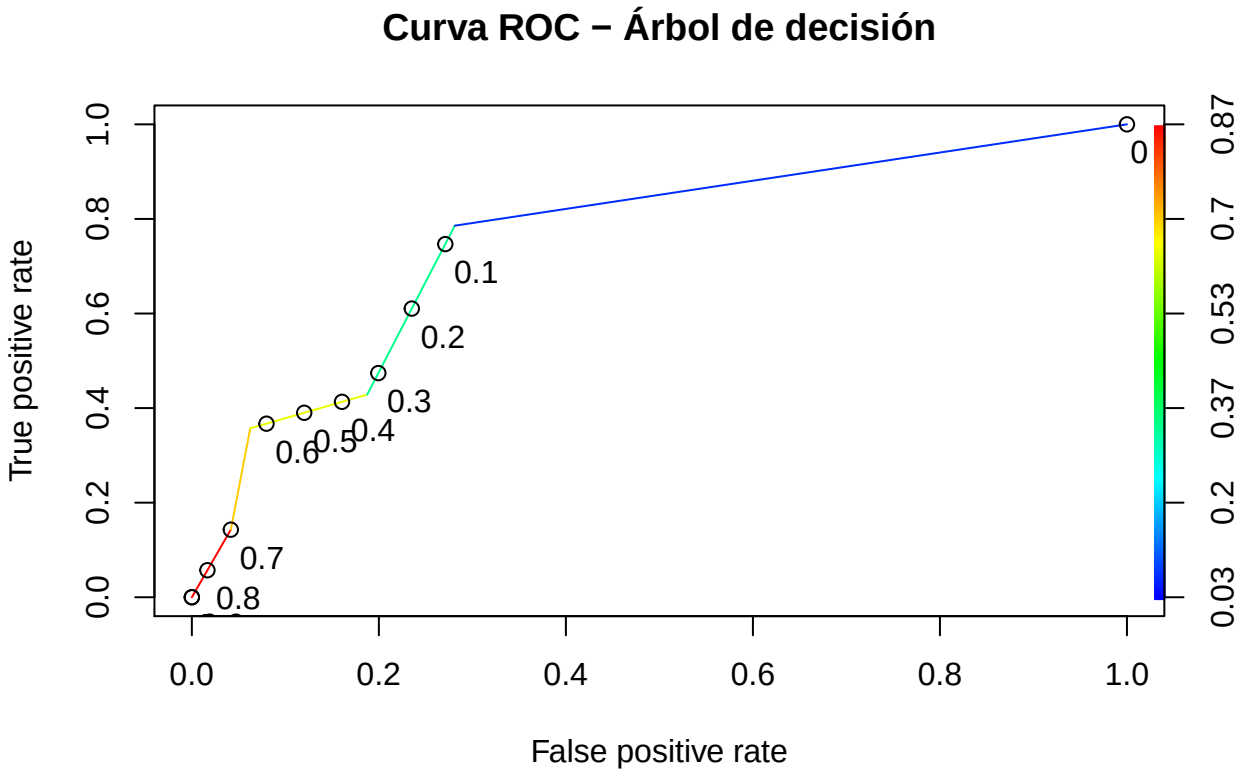
```
# Evaluar el modelo
resultados.arbol <- table(prueba$family, predicciones.arbol[,2]>= 0.5)
resultados.arbol
```

```
##
##     FALSE TRUE
##   0    90    6
##   1     9    5
```

```
#predict de ROCR
prediccionROC.arbol <- prediction(predicciones.arbol[,2],prueba$family)

# ROC
as.numeric(performance(prediccionROC.arbol, "auc")@y.values)
```

```
## [1] 0.755952381
```

```
ROCR.arbol <- performance(prediccionROC.arbol, "tpr", "fpr")
plot(ROCR.arbol, main = "Curva ROC - Árbol de decisión", colorize = TRUE, print.cutoffs.at = seq(0,
1, by = 0.1), text.adj = c(-0.2, 1.7))
```



Métricas del Modelo árbol de decisión:

- Exactitud: 86.36%
- Sensibilidad: 35.71%
- Especificidad: 93.75%
- Área bajo la curva: 75.59%

Evaluación (bosque aleatorio)

La evaluación del bosque aleatorio.

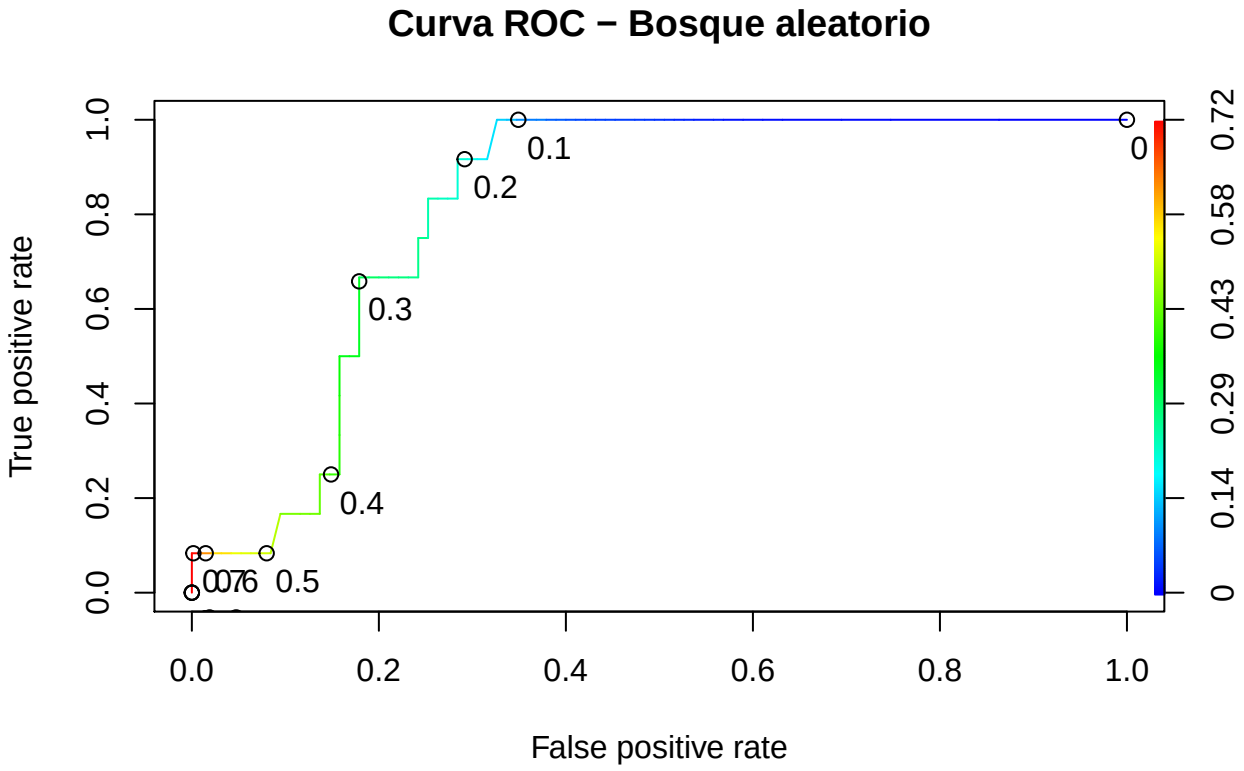
```
# Evaluar el modelo
resultados.bosque <- table(prueba$family, predicciones.bosque[,2] >= 0.5)
resultados.bosque
```

```
##
##     FALSE TRUE
##   0    88    7
##   1    11    1
```

```
#predict de ROCR
prediccionROC.bosque <- prediction(predicciones.bosque[,2],prueba$family)

# ROC
as.numeric(performance(prediccionROC.bosque, "auc")@y.values)
```

```
## [1] 0.8201754386
```

```
ROCR.bosque <- performance(prediccionROC.bosque, "tpr", "fpr")
plot(ROCR.bosque, main = "Curva ROC - Bosque aleatorio", colorize = TRUE, print.cutoffs.at = seq(0,
1, by = 0.1), text.adj = c(-0.2, 1.7))
```



Métricas del Modelo bosque aleatorio:

- Exactitud: 83.17%
- Sensibilidad: 8.33%
- Especificidad: 92.63%
- Área bajo la curva: 82.01%

Resultados

En conclusion podemos obtener el siguiente resumen:

- El modelo ingenuo es muy bueno obteniendo verdaderos negativos, pero malo obteniendo verdadero positivo.
- El modelo de árbol de decisión tiene el mayor valor de exactitud de los 4 modelos.
- El modelo de árbol de decisión tiene el mayor valor de especificidad entre los modelos red neuronal y bosque aleatorio.
- El modelo de árbol de decisión tiene el mayor valor de sensibilidad de los 4 modelos.
- El modelo que tiene mayor cobertura bajo la curva es bosque aleatorio.
- El mejor modelo para utilizar es el de árbol de decisión.

- Ningún modelo es adecuado ni recomendable para obtener verdaderos positivos.

Métricas del modelo ingenuo:

- Exactitud: 87.27%
- Sensibilidad: 0%
- Especificidad: 100%
- Área bajo la curva: 50%

Métricas del Modelo red neuronal:

- Exactitud: 74.76%
- Sensibilidad: 16.66%
- Especificidad: 82.10%
- Área bajo la curva: 69.21%

Métricas del Modelo árbol de decisión:

- Exactitud: 86.36%
- Sensibilidad: 35.71%
- Especificidad: 93.75%
- Área bajo la curva: 75.59%

Métricas del Modelo bosque aleatorio:

- Exactitud: 83.17%
- Sensibilidad: 8.33%
- Especificidad: 92.63%
- Área bajo la curva: 82.01%

Pregunta 3

Análisis del Problema

El objetivo es distinguir entre la presencia y ausencia de arritmia cardíaca y clasificarlo en grupos. Para el tiempo, existe un programa de computadora que hace una clasificación. Sin embargo, existen diferencias entre cardiología y la clasificación de los programas. Tomando los cardiólogos como un patrón de oro para pretender minimizar esa diferencia mediante herramientas de aprendizaje automático.

Se quiere determinar cuántos grupos se deben crear y cuáles características tienen estos grupos.

Entendimiento de los Datos

Este conjunto de datos contiene 9 atributos y 451 observaciones.

- **Edad** : Valor numérico entre 0 y 83.
- **Sexo** : Valor numérico entre 0 y 1 (0 Hombre - 1 mujer).
- **Altura** : Valor numérico entre 105 y 780.
- **Peso** : Valor numérico entre 6 y 176.
- **Duración del QRS** : Valor numérico entre 55 y 188.
- **Intervalo P-R** : Valor numérico entre 0 y 524.
- **Intervalo Q-T** : Valor numérico entre 232 y 509.
- **Intervalo T** : Valor numérico entre 108 y 381.
- **Intervalo P** : Valor numérico entre 0 y 205.

Exploración de los Datos

```
#limpiar variables
rm(list=ls(all=TRUE))

#librerías utilizadas
library(cluster)

setwd('D:\\Drive\\Universidad\\Cenfotec\\MBD\\2016 Cuatrimestre 3\\MBD-305 Minería de datos 1\\Semana 13
datos <- read.csv('arrhythmia.csv',na.strings = "?")

colnames(datos) <- c("Age","Sex","Height","Weight","QRS","PR","QT","T","P")

datos=datos[,1:9]

#Visualizar los datos
str(datos)
```

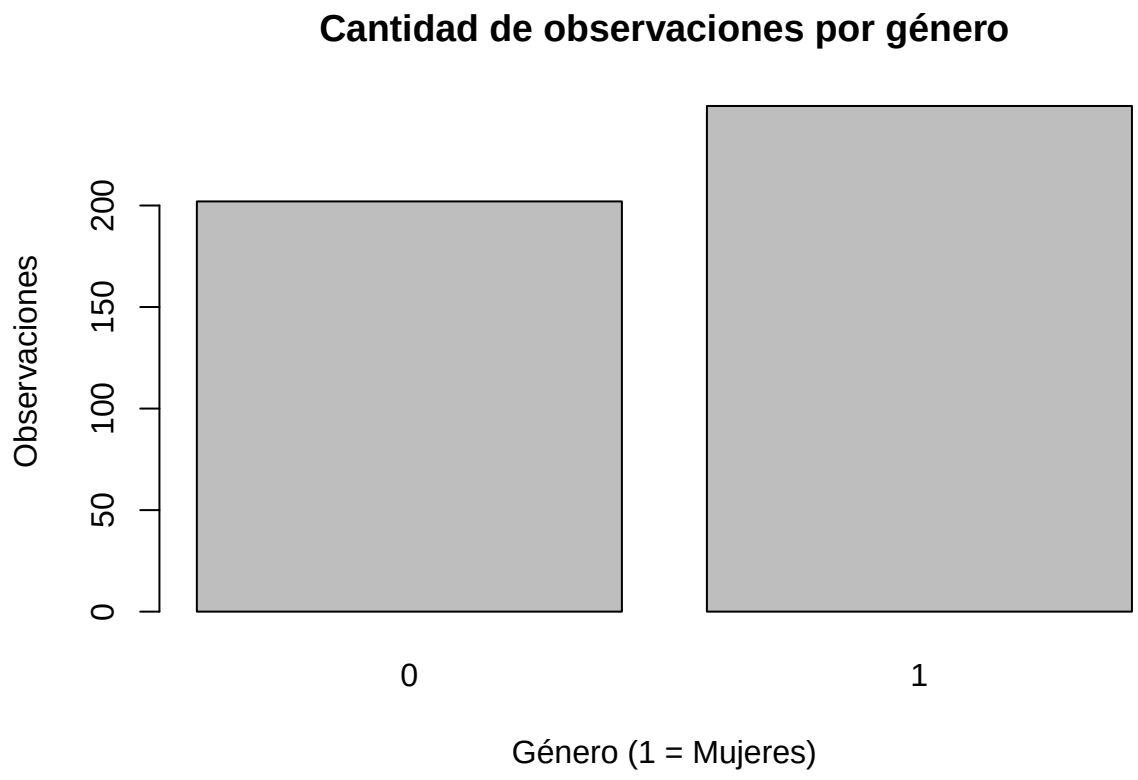
```
## 'data.frame':    451 obs. of  9 variables:
##  $ Age   : int  56 54 55 75 13 40 49 44 50 62 ...
##  $ Sex   : int  1 0 0 0 0 1 1 0 1 0 ...
##  $ Height: int  165 172 175 190 169 160 162 168 167 170 ...
##  $ Weight: int  64 95 94 80 51 52 54 56 67 72 ...
##  $ QRS   : int  81 138 100 88 100 77 78 84 89 102 ...
##  $ PR    : int  174 163 202 181 167 129 0 118 130 135 ...
##  $ QT    : int  401 386 380 360 321 377 376 354 383 401 ...
##  $ T     : int  149 185 179 177 174 133 157 160 156 156 ...
##  $ P     : int  39 102 143 103 91 77 70 63 73 83 ...
```

```
#Datos resumidos
summary(datos)
```

```
##       Age               Sex               Height
##  Min.   : 0.00000   Min.   :0.0000000   Min.   :105.0000
##  1st Qu.:36.00000   1st Qu.:0.0000000   1st Qu.:160.0000
##  Median :47.00000   Median :1.0000000   Median :164.0000
##  Mean   :46.40798   Mean   :0.5521064   Mean   :166.1353
##  3rd Qu.:58.00000   3rd Qu.:1.0000000   3rd Qu.:170.0000
##  Max.   :83.00000   Max.   :1.0000000   Max.   :780.0000
##      Weight              QRS                 PR
##  Min.   :  6.00000   Min.   : 55.00000   Min.   :  0.0000
##  1st Qu.: 59.00000   1st Qu.: 80.00000   1st Qu.:142.0000
##  Median : 68.00000   Median : 86.00000   Median :157.0000
##  Mean   : 68.14412   Mean   : 88.91574   Mean   :155.0687
##  3rd Qu.: 78.50000   3rd Qu.: 94.00000   3rd Qu.:174.5000
##  Max.   :176.00000   Max.   :188.00000   Max.   :524.0000
##        QT                  T                   P
##  Min.   :232.0000   Min.   :108.0000   Min.   :  0.0000
##  1st Qu.:350.0000   1st Qu.:148.0000   1st Qu.: 79.0000
##  Median :367.0000   Median :162.0000   Median : 91.0000
##  Mean   :367.1996   Mean   :169.9401   Mean   : 89.9357
##  3rd Qu.:384.0000   3rd Qu.:179.0000   3rd Qu.:102.0000
##  Max.   :509.0000   Max.   :381.0000   Max.   :205.0000
```

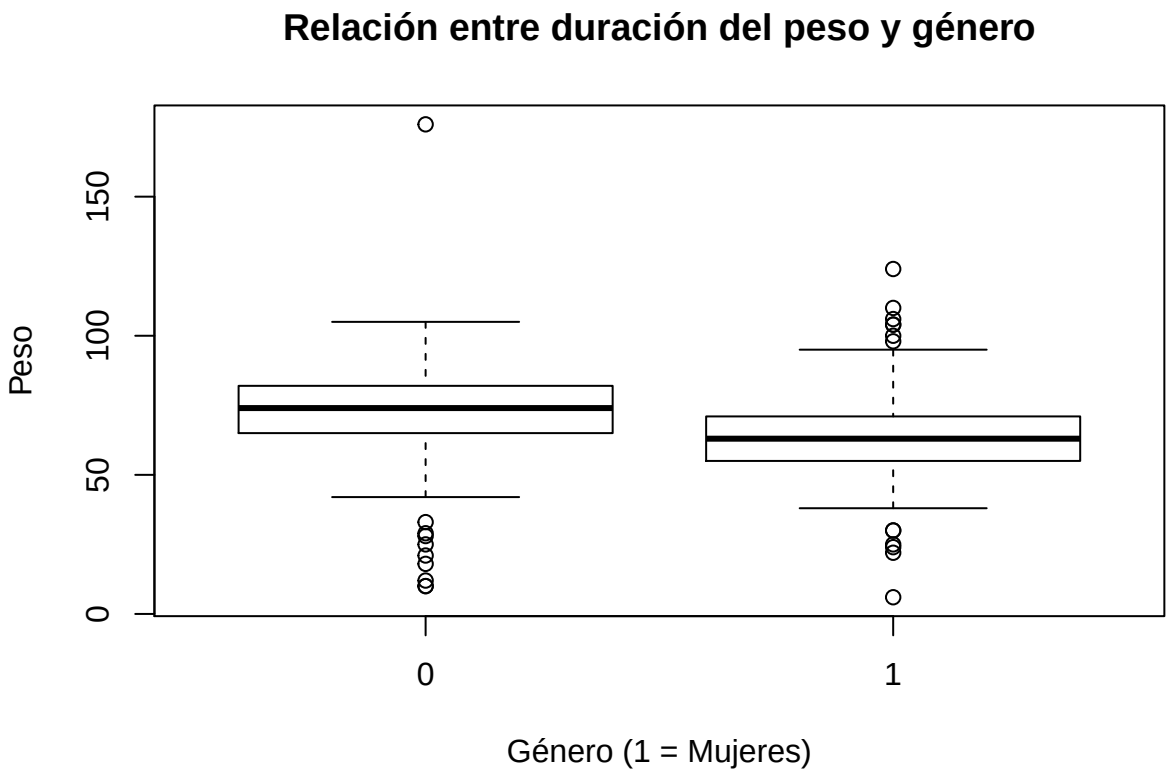
En este grafico podemos observar que los registros son mayormente mujeres.

```
barplot(table(datos$Sex),
main = 'Cantidad de observaciones por género',
xlab = 'Género (1 = Mujeres)',
ylab = 'Observaciones')
```



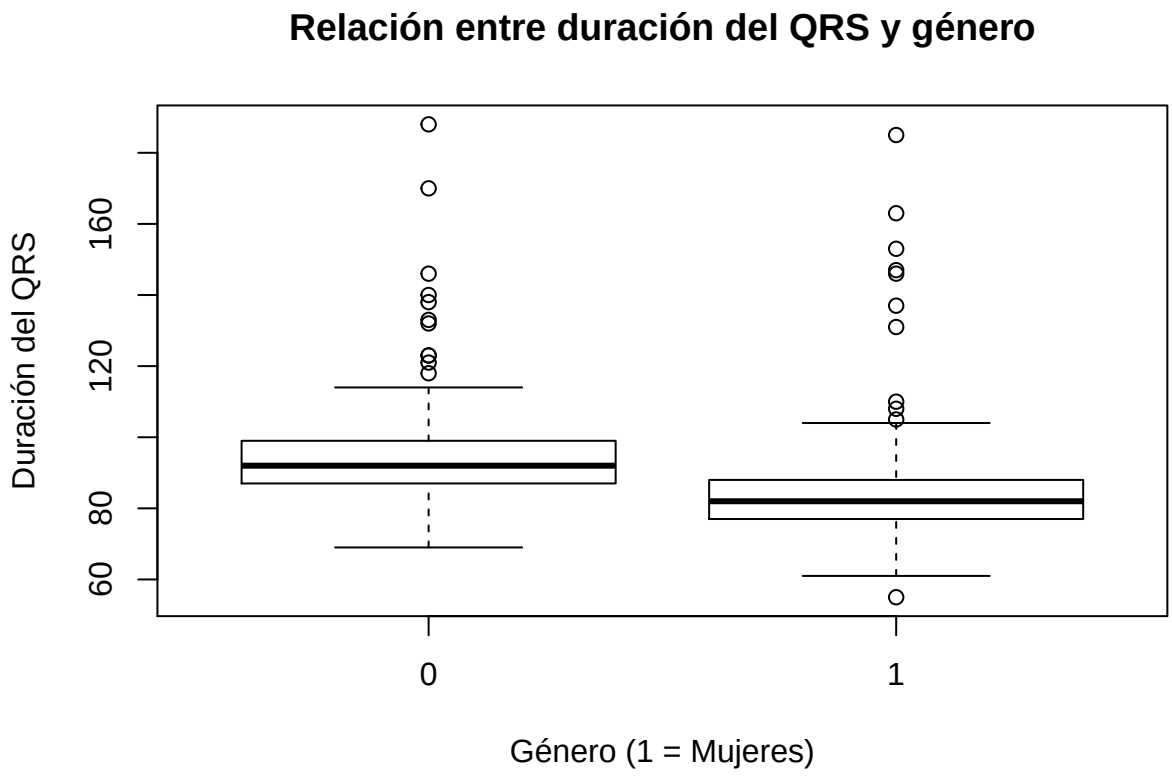
En este grafico podemos observar que el peso de un hombre está alrededor de 60 y 80 y de una mujer de 60 a 70.

```
boxplot(datos$Weight ~ factor(datos$Sex),
main = 'Relación entre duración del peso y género',
xlab = 'Género (1 = Mujeres)',
ylab = 'Peso')
```



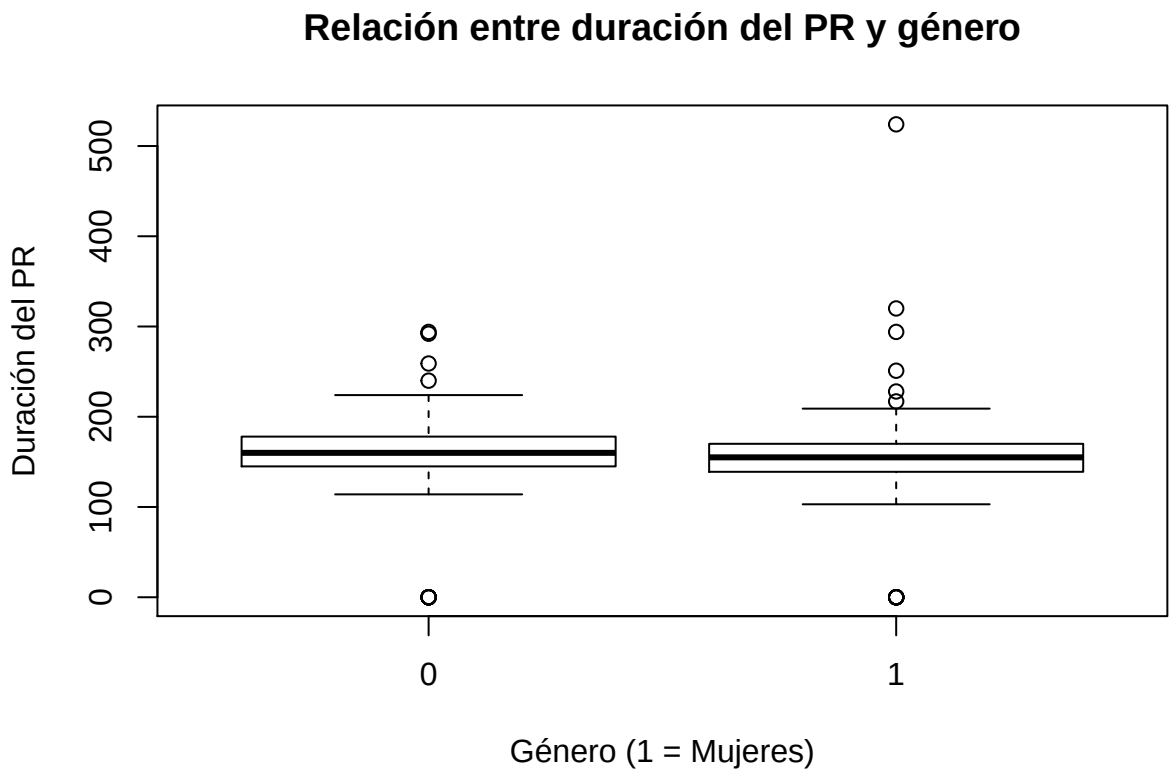
En este grafico podemos observar que la duración del QRS de un hombre está alrededor de 80 y 100 y de una mujer de 70 a 90.

```
boxplot(datos$QRS ~ factor(datos$Sex),
main = 'Relación entre duración del QRS y género',
xlab = 'Género (1 = Mujeres)',
ylab = 'Duración del QRS')
```



En este grafico podemos observar que la duración del PR de un hombre y de una mujer son muy similares.

```
boxplot(datos$PR ~ factor(datos$Sex),
main = 'Relación entre duración del PR y género',
xlab = 'Género (1 = Mujeres)',
ylab = 'Duración del PR')
```



Creación del Modelo

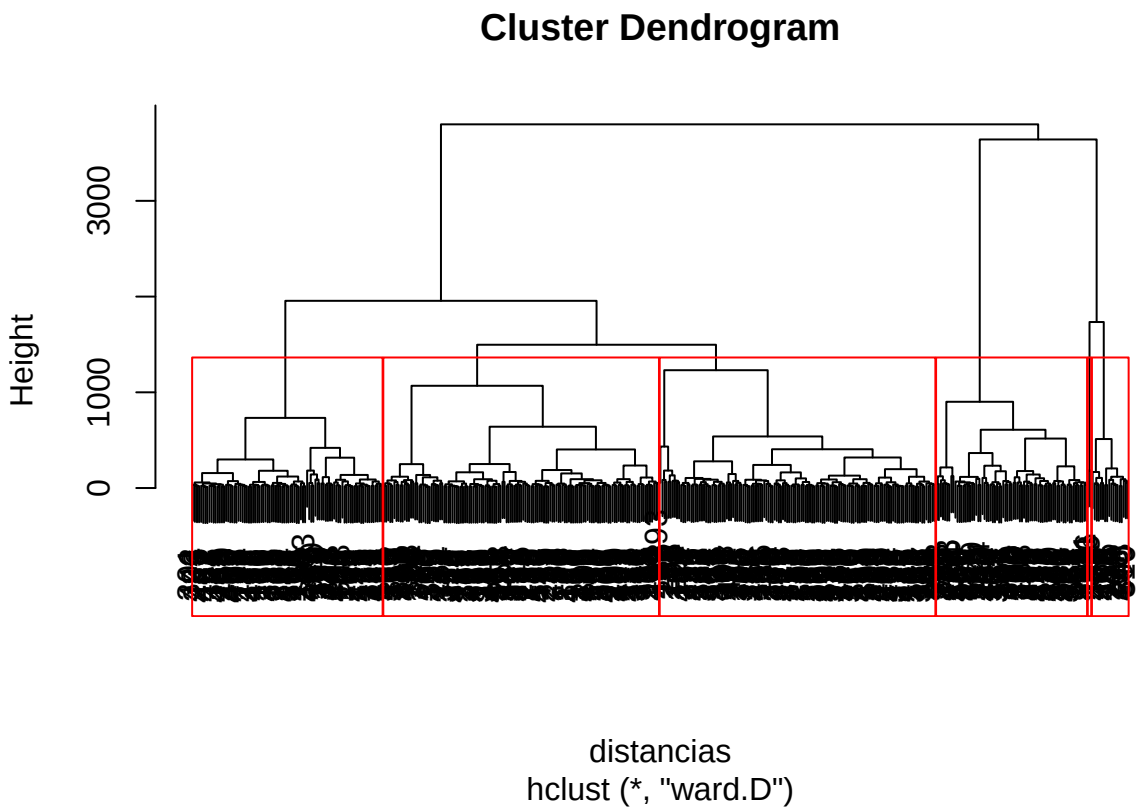
Creación del Modelo jerárquico:

Para determinar la cantidad de clústeres que se pueden crear, se procede a hacer un agrupamiento jerárquico:

```
distancias <- dist(datos, method="euclidean")
datos.jerarquico <- hclust(distancias, method="ward.D")
```

Para este análisis, vamos a utilizar 6 grupos.

```
plot(datos.jerarquico)
rect.hclust(datos.jerarquico, k = 6, border = "red")
```



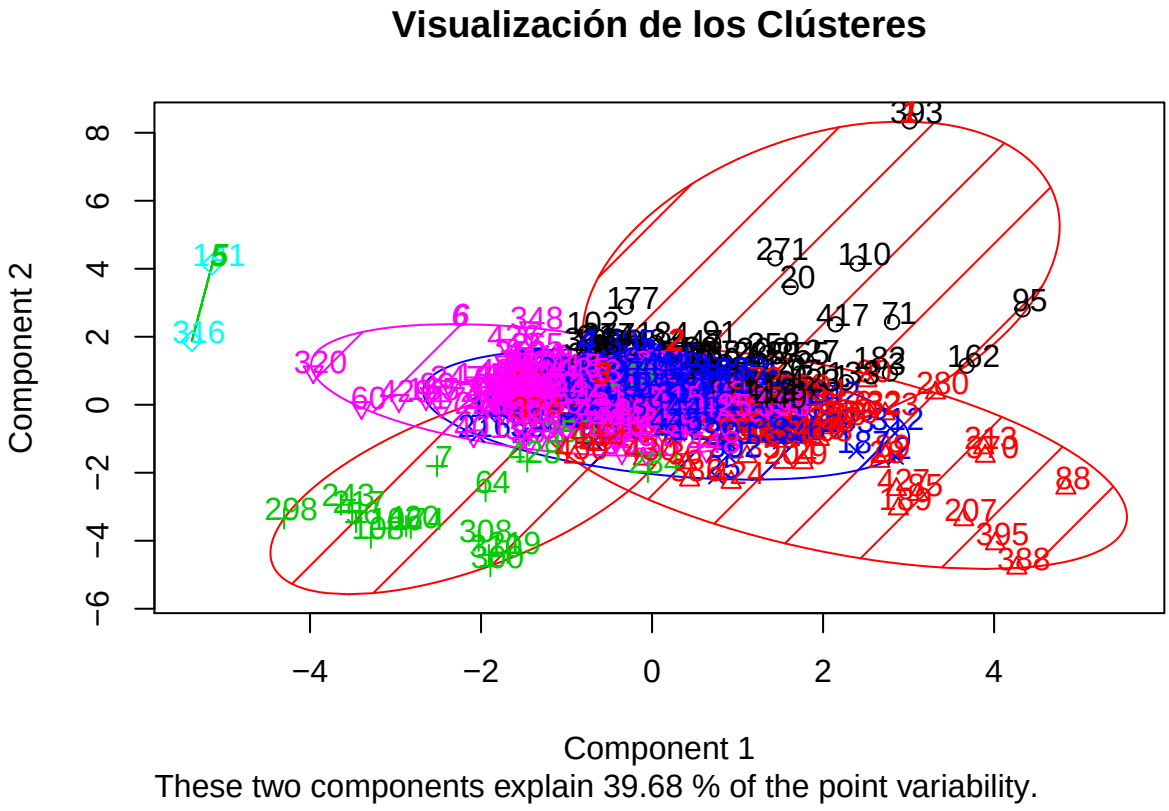

```
cluster.jerarquico <- factor(cutree(datos.jerarquico, k=6))
```

Luego de hacer el análisis jerárquico, se procede a hacer el análisis utilizando el algoritmo KMeans con 6 centros:

```
set.seed(12345)
km <- kmeans(datos, centers = 6)
cluster.kmeans <- factor(km$cluster)
```

El resultado se puede visualizar así: (Los colores de los puntos representan el grupo al cual pertenecen)

```
#Visualizar los Clústeres
clusplot(datos,
km$cluster,
col.p = km$cluster,
color=TRUE,
shade=TRUE,
labels=2,
lines=0,
main = 'Visualización de los Clústeres')
```



Evaluación

Con el fin de comparar ambos agrupamientos, podemos generar tablas resumen para comparar los valores promedios de cada variable en cada grupo:

```
resultado.jerarquico <- rbind(tapply(datos$Age, cluster.jerarquico, mean),
                              tapply(datos$Sex, cluster.jerarquico, mean),
                              tapply(datos$Height, cluster.jerarquico, mean),
                              tapply(datos$Weight, cluster.jerarquico, mean),
                              tapply(datos$QRS, cluster.jerarquico, mean),
                              tapply(datos$PR, cluster.jerarquico, mean),
                              tapply(datos$QT, cluster.jerarquico, mean),
                              tapply(datos$T, cluster.jerarquico, mean),
                              tapply(datos$P, cluster.jerarquico, mean)
                              )

rownames(resultado.jerarquico) <- c('Age', 'Sex', 'Height','Weight', 'QRS', 'PR', 'QT', 'T', 'P')
resultado.jerarquico[2,]<-round(resultado.jerarquico[2,],0)
resultado.jerarquico
```

```
##                   1           2           3           4           5
## Age      46.65413534  46.64383562  53.29323308  36.16304348  50.22222222
## Sex       0.00000000   0.00000000   1.00000000   1.00000000   0.00000000
## Height  166.87969925 163.36986301 164.41353383 158.52173913 164.83333333
## Weight   74.53383459  69.82191781  69.69172932  56.75000000  67.61111111
```

```
## QRS     89.99248120 102.52054795  84.24812030  83.55434783  88.22222222
## PR     150.74436090 166.43835616 183.47368421 141.78260870   0.00000000
## QT     382.03759398 371.05479452 369.13533835 344.97826087 355.94444444
## T      167.74436090 233.27397260 151.80451128 152.63043478 155.22222222
## P       87.02255639  95.39726027 104.49624060  79.34782609  36.61111111
##            6
## Age      0.5
## Sex      0.0
## Height 694.0
## Weight   8.0
## QRS     84.0
## PR     145.5
## QT     234.5
## T      139.0
## P       83.0
```

Los grupos creados se pueden resumir así:

- **Grupo 1**: Hombres con edad promedio de 52 años,altura 167,peso 72,duración del QRS 87,intervalo P-R 208,intervalo Q-T 363,intervalo T 158,intervalo P 121.
- **Grupo 2**: Hombres con edad promedio de 46 años,altura 162,peso 70,duración del QRS 104,intervalo P-R 162,intervalo Q-T 376,intervalo T 243,intervalo P 94.
- **Grupo 3**: Mujeres con edad promedio de 50 años,altura 164,peso 67,duración del QRS 88,intervalo P-R 0,intervalo Q-T 356,intervalo T 155,intervalo P 36.
- **Grupo 4**: Mujeres con edad promedio de 50 años,altura 164,peso 69,duración del QRS 87,intervalo P-R 161,intervalo Q-T 392,intervalo T 160,intervalo P 90.
- **Grupo 5**: Hombres con edad promedio de 0.5 años,altura 694,peso 8,duración del QRS 84,intervalo P-R 145,intervalo Q-T 234,intervalo T 139,intervalo P 83.
- **Grupo 6**: Hombres con edad promedio de 42 años,altura 163,peso 65,duración del QRS 85,intervalo P-R 146,intervalo Q-T 347,intervalo T 158,intervalo P 85.

De manera similar, se puede generar un resumen para los grupos creados con el algoritmo KMeans:

```r
resultado.kmeans <- rbind(tapply(datos$Age, cluster.kmeans, mean),
                          tapply(datos$Sex, cluster.kmeans, mean),
                          tapply(datos$Height, cluster.kmeans, mean),
                          tapply(datos$Weight, cluster.kmeans, mean),
                          tapply(datos$QRS, cluster.kmeans, mean),
                          tapply(datos$PR, cluster.kmeans, mean),
                          tapply(datos$QT, cluster.kmeans, mean),
                          tapply(datos$T, cluster.kmeans, mean),
                          tapply(datos$P, cluster.kmeans, mean))
rownames(resultado.kmeans) <- c('Age', 'Sex', 'Height','Weight', 'QRS', 'PR', 'QT', 'T', 'P')
resultado.kmeans[2,]<-round(resultado.kmeans[2,],0)
resultado.kmeans
```

```
##                  1           2           3           4     5
## Age     51.01754386  46.16393443  50.22222222  50.46979866   0.5
## Sex      0.00000000   0.00000000   0.00000000   1.00000000   0.0
## Height 166.77192982 162.85245902 164.83333333 164.02684564 694.0
## Weight  71.92982456  70.90163934  67.61111111  69.40939597   8.0
## QRS     86.68421053 103.54098361  88.22222222  87.62416107  84.0
## PR     207.91228070 161.75409836   0.00000000 160.19463087 145.5
```

```
## QT     362.22807018 375.55737705 355.94444444 391.91275168 234.5
## T      157.17543860 242.45901639 155.22222222 160.61744966 139.0
## P      120.49122807  93.67213115  36.61111111  89.15436242  83.0
##                   6
## Age     41.34756098
## Sex      1.00000000
## Height 162.75609756
## Weight  65.44512195
## QRS     85.56097561
## PR     146.69512195
## QT     346.21951220
## T      157.86585366
## P       84.57317073
```

Los grupos creados se pueden resumir así:

- **Grupo 1**: Hombres con edad promedio de 47 años,altura 167,peso 75,duración del QRS 90,intervalo P-R 151,intervalo Q-T 382,intervalo T 168,intervalo P 87.
- **Grupo 2**: Hombres con edad promedio de 47 años,altura 163,peso 70,duración del QRS 103,intervalo P-R 166,intervalo Q-T 371,intervalo T 233,intervalo P 95.
- **Grupo 3**: Mujeres con edad promedio de 54 años,altura 164,peso 70,duración del QRS 84,intervalo P-R 183,intervalo Q-T 369,intervalo T 151,intervalo P 104.
- **Grupo 4**: Mujeres con edad promedio de 36 años,altura 159,peso 57,duración del QRS 84,intervalo P-R 142,intervalo Q-T 345,intervalo T 153,intervalo P 79.
- **Grupo 5**: Hombres con edad promedio de 50 años,altura 165,peso 68,duración del QRS 88,intervalo P-R 0,intervalo Q-T 356,intervalo T 155,intervalo P 37.
- **Grupo 6**: Hombres con edad promedio de 0.5 años,altura 694,peso 8,duración del QRS 84,intervalo P-R 145,intervalo Q-T 234,intervalo T 139,intervalo P 83.

Resultados

De los anteriores datos podemos concluir que los datos Edad,Sexo,Altura,Duración del QRS,Intervalo P-R,Intervalo Q-T,Intervalo T,Intervalo P, idependiendemente del algoritmo utilizado(jerárquico o KMeans), los grupos en general son básicamente los mismo. Existen varios grupos que varían pequeñas cosas, pero en general no representan grandes diferencias. En resumen ambos modelos son muy válidos para cualquiera de la expectativas esperadas en el análisis.

Pregunta 4

Análisis del Problema

Australia ha estado habitada desde hace por lo menos cuarenta y seis mil años por los aborígenes australianos. Su descubrimiento se habría producido tras las esporádicas visitas de españoles y portugueses que exploraron la costa septentrional y occidental de Australia. La mayor parte de los aproximadamente 25 millones de australianos viven concentrados en las principales ciudades. La población de Australia se ha cuadriplicado desde el final de la Primera Guerra Mundial,39 incentivada por un ambicioso programa de inmigración.

Entendimiento de los Datos

Números (en miles) de residentes australianos medidos trimestralmente de marzo de 1971 a marzo de 1994. Contiene 89 observaciones.

- **Serie de tiempo** : Valor numérico entre 1971 y 1993

Exploración de los Datos

```
#limpiar variables
rm(list=ls(all=TRUE))

#librerías utilizadas
library(forecast)
```

```
## Loading required package: zoo
```

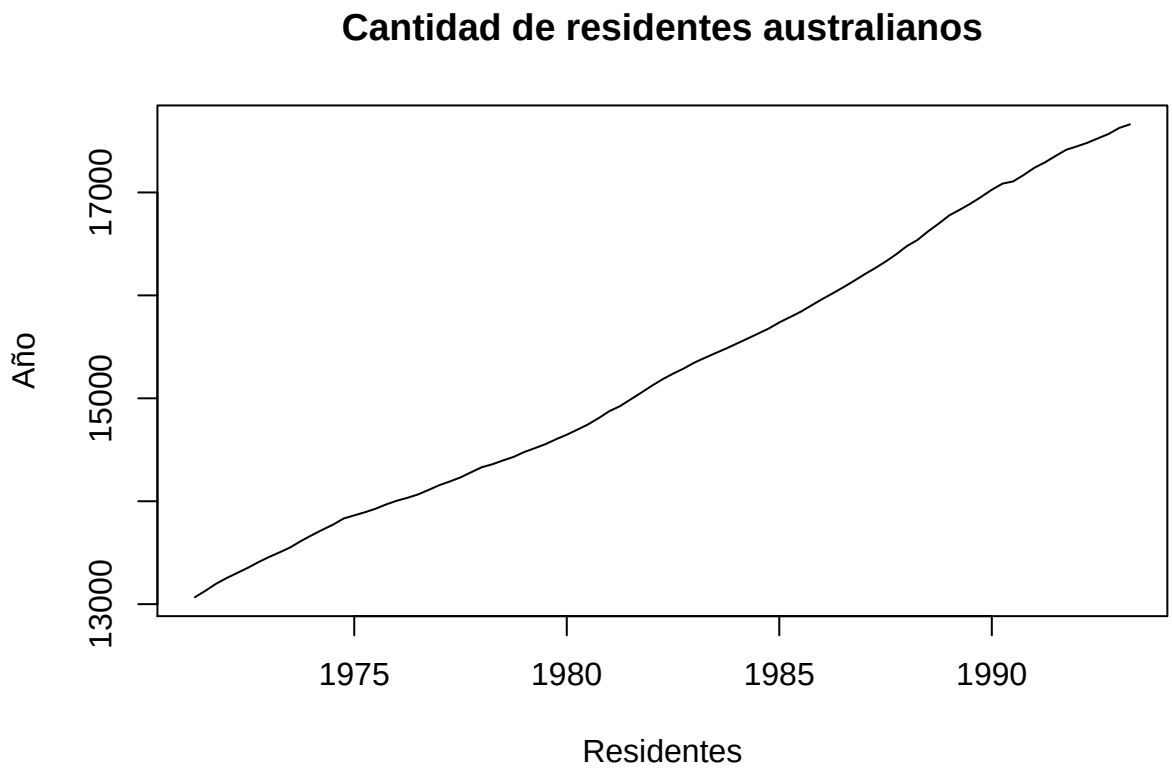
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

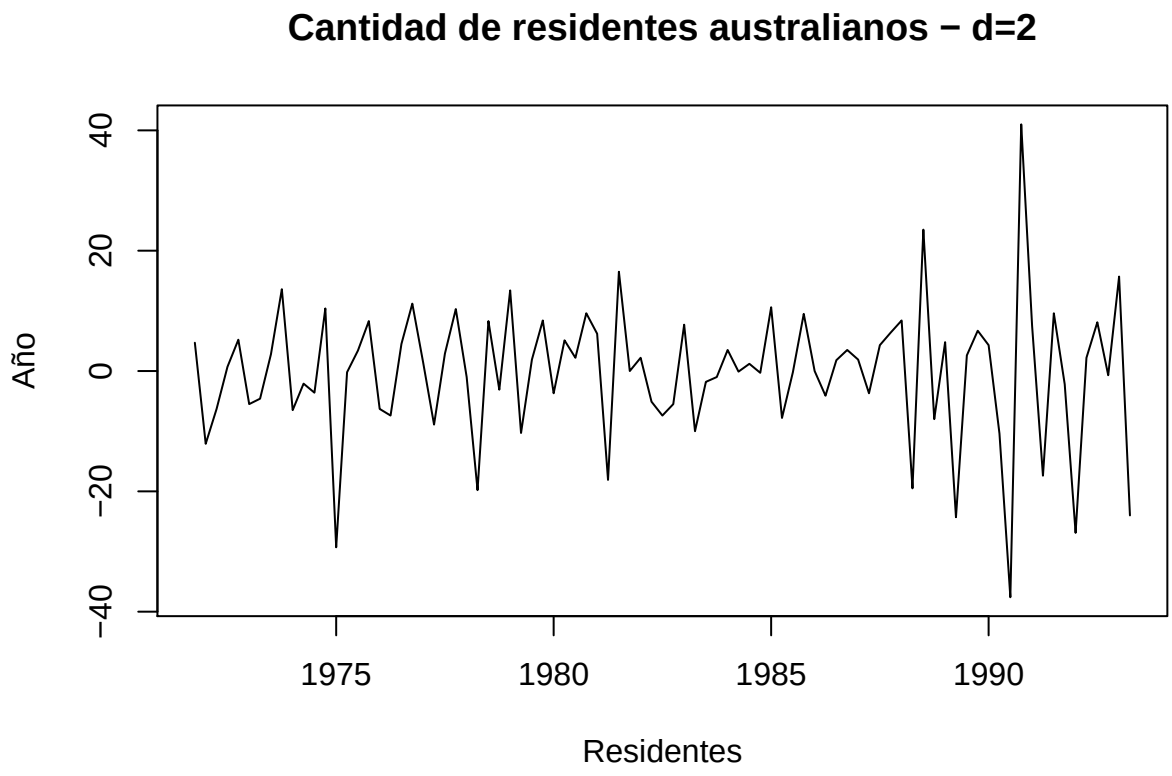
```
## Loading required package: timeDate
```

```
## This is forecast 7.3
```

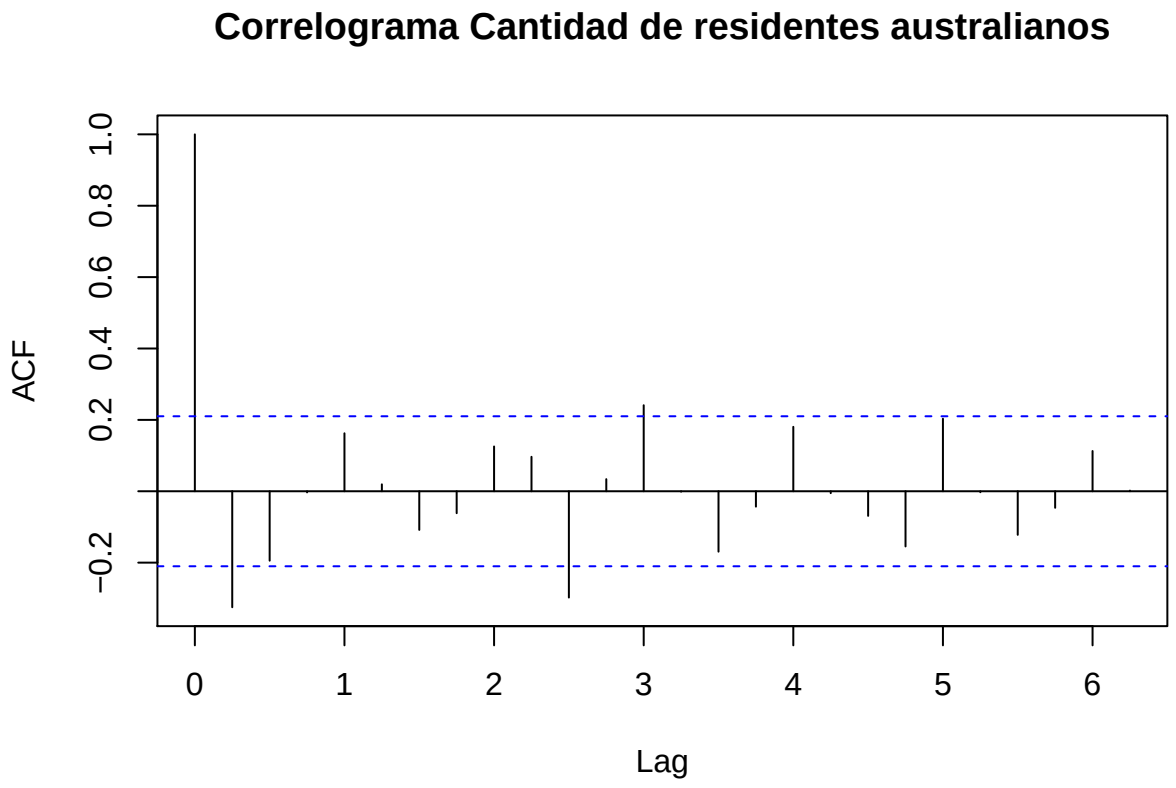
```
data(austres)

#Visualizacíon de los datos
plot(austres,
main = 'Cantidad de residentes australianos',
xlab = 'Residentes',
ylab = 'Año')
```




```r
#Diferenciación,2 nivel

austresDiff2 <- diff(austres,differences = 2)

#Visualizacíon de los datos diferenciados
plot(austresDiff2,
main = 'Cantidad de residentes australianos - d=2',
xlab = 'Residentes',
ylab = 'Año')
```



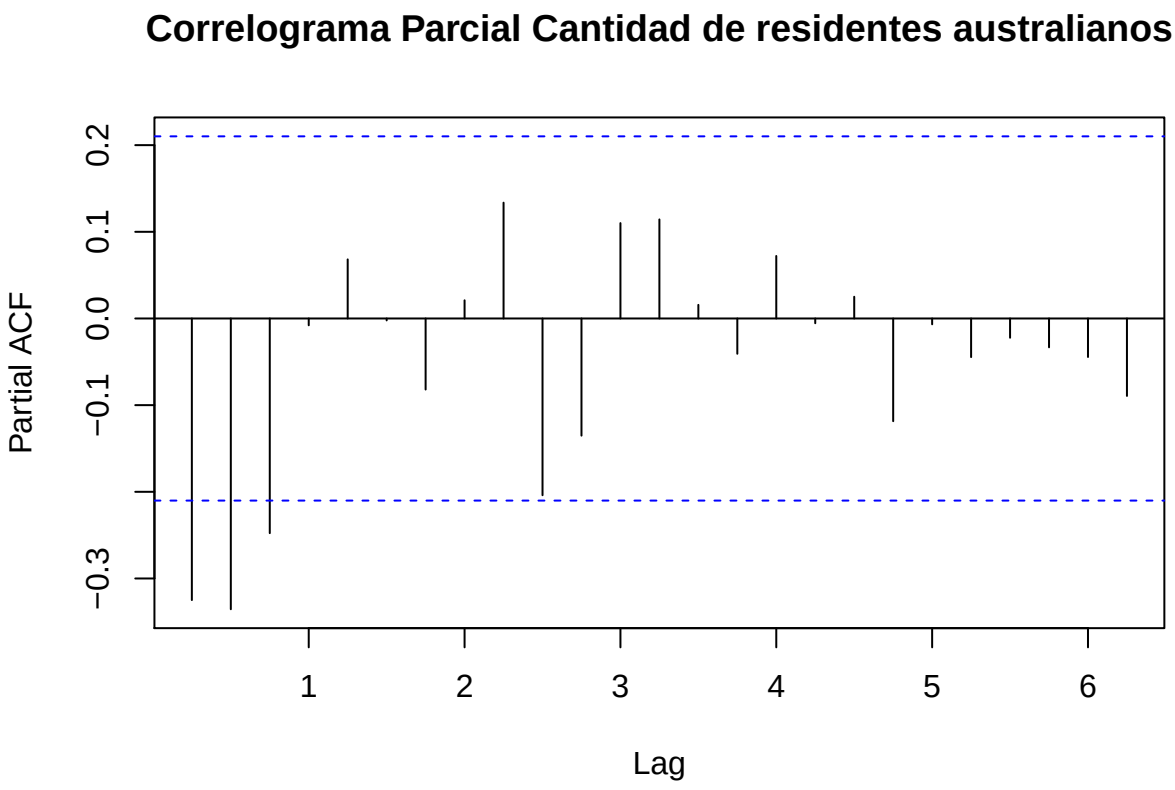

```r
#Creación del correlograma
acf(austresDiff2,lag.max = 25,#visualizar 25 retrasos o 'lags'
    main="Correlograma Cantidad de residentes australianos")
```




```
#Ver solo los valores
acf(austresDiff2,lag.max = 25,#visualizar 25 retrasos o 'lags'
    plot = FALSE)
```

```
##
## Autocorrelations of series 'austresDiff2', by lag
##
##   0.00   0.25   0.50   0.75   1.00   1.25   1.50   1.75   2.00   2.25
##  1.000 -0.325 -0.194 -0.003  0.162  0.019 -0.108 -0.061  0.126  0.097
##   2.50   2.75   3.00   3.25   3.50   3.75   4.00   4.25   4.50   4.75
## -0.298  0.034  0.241 -0.001 -0.169 -0.043  0.181 -0.005 -0.069 -0.155
##   5.00   5.25   5.50   5.75   6.00   6.25
##  0.203 -0.003 -0.122 -0.046  0.113  0.002
```

```
#Creación del correlograma parcial
pacf(austresDiff2,lag.max = 25,#visualizar 25 retrasos o 'lags'
    main="Correlograma Parcial Cantidad de residentes australianos")
```




```
#Ver solo los valores
pacf(austresDiff2,lag.max = 25,#visualizar 25 retrasos o 'lags'
    plot = FALSE)
```

```
##
## Partial autocorrelations of series 'austresDiff2', by lag
##
##   0.25   0.50   0.75   1.00   1.25   1.50   1.75   2.00   2.25   2.50
## -0.325 -0.336 -0.248 -0.008  0.068 -0.002 -0.082  0.021  0.134 -0.204
##   2.75   3.00   3.25   3.50   3.75   4.00   4.25   4.50   4.75   5.00
## -0.135  0.110  0.114  0.016 -0.041  0.072 -0.006  0.025 -0.119 -0.007
##   5.25   5.50   5.75   6.00   6.25
## -0.045 -0.022 -0.033 -0.044 -0.089
```

Creación del Modelo

Modelo IMA(0,2,0.25)

```
#Crear modelo
modelo1<-arima(austres,order = c(0,2,0.25)) #p=0,d=2,q=0.25
#Ver modelo
modelo1
```

```
##
```

```
## Call:
## arima(x = austres, order = c(0, 2, 0.25))
##
##
## sigma^2 estimated as 130.1297:  log likelihood = -335.23,  aic = 672.46
```

```
#Hacer predicciones
predicciones.modelo1<-forecast.Arima(modelo1,h=10)#Pronosticar los próximos 10 periodos
```

Modelo IMA(0,2,0.50)

```
#Crear modelo
modelo2<-arima(austres,order = c(0,2,0.50)) #p=0,d=2,q=0.50
#Ver modelo
modelo2
```

```
##
## Call:
## arima(x = austres, order = c(0, 2, 0.5))
##
##
## sigma^2 estimated as 130.1297:  log likelihood = -335.23,  aic = 672.46
```

```
#Hacer predicciones
predicciones.modelo2<-forecast.Arima(modelo2,h=10)#Pronosticar los próximos 10 periodos
```

Modelo IMA(0,2,0.75)

```
#Crear modelo
modelo3<-arima(austres,order = c(0,2,0.75)) #p=0,d=2,q=0.75
#Ver modelo
modelo3
```

```
##
## Call:
## arima(x = austres, order = c(0, 2, 0.75))
##
##
## sigma^2 estimated as 130.1297:  log likelihood = -335.23,  aic = 672.46
```

```
#Hacer predicciones
predicciones.modelo3<-forecast.Arima(modelo3,h=10)#Pronosticar los próximos 10 periodos
```

Modelo ARIMA(2.50,2,0.25)

```
#Crear modelo
modelo4<-arima(austres,order = c(2.50,2,0.25)) #p=2.50,d=2,q=0.25
#Ver modelo
modelo4
```

```
##
## Call:
## arima(x = austres, order = c(2.5, 2, 0.25))
##
## Coefficients:
##              ar1         ar2
##       -0.4440368  -0.3448637
## s.e.   0.1022489   0.1029951
##
## sigma^2 estimated as 102.2594:  log likelihood = -324.93,  aic = 655.86
```

```
#Hacer predicciones
predicciones.modelo4<-forecast.Arima(modelo4,h=10)#Pronosticar los próximos 10 periodos
```

Modelo ARIMA(3,2,0.50)

```
#Crear modelo
modelo5<-arima(austres,order = c(3,2,0.50)) #p=3,d=2,q=0.50
#Ver modelo
modelo5
```

```
##
## Call:
## arima(x = austres, order = c(3, 2, 0.5))
##
## Coefficients:
##              ar1         ar2         ar3
##       -0.5315557  -0.4555178  -0.2576215
## s.e.   0.1051911   0.1093818   0.1059603
##
## sigma^2 estimated as 95.56576:  log likelihood = -322.08,  aic = 652.17
```

```
#Hacer predicciones
predicciones.modelo5<-forecast.Arima(modelo5,h=10)#Pronosticar los próximos 10 periodos
```

Modelo ARIMA(5,2,0.75)

```
#Crear modelo
modelo6<-arima(austres,order = c(5,2,0.75)) #p=5,d=2,q=0.75
#Ver modelo
modelo6
```

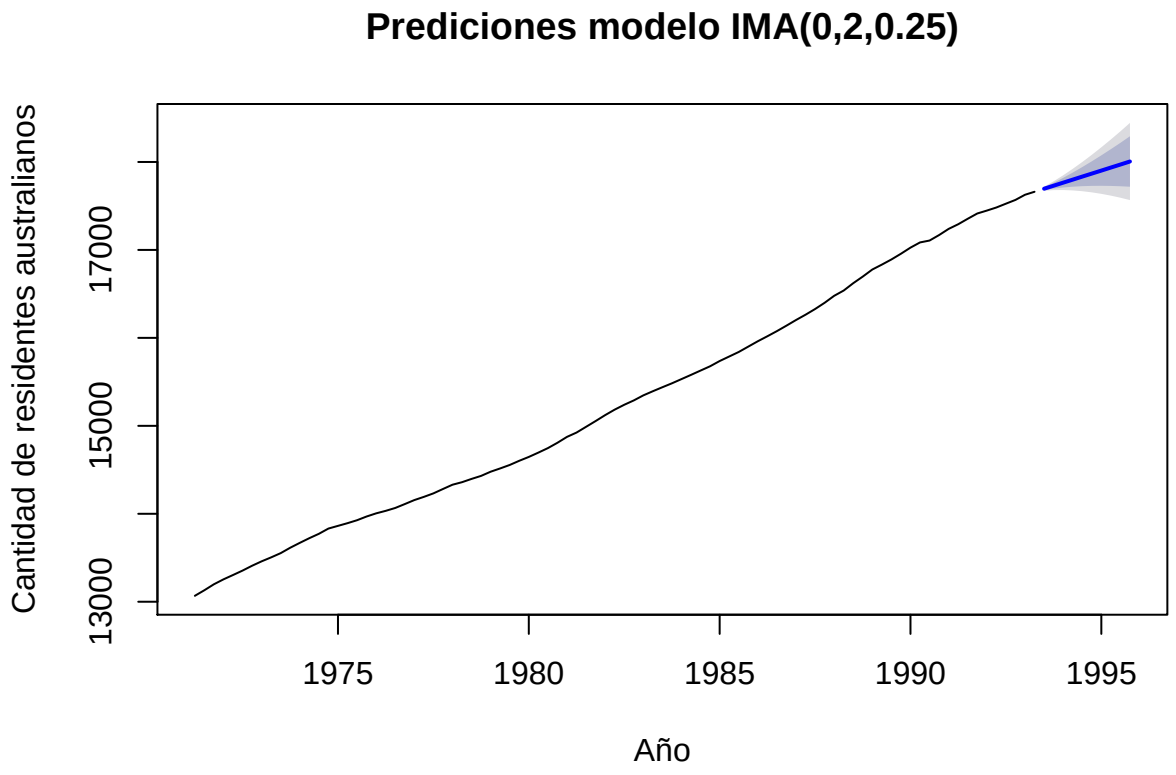
```
##
## Call:
## arima(x = austres, order = c(5, 2, 0.75))
##
## Coefficients:
##              ar1         ar2         ar3        ar4        ar5
##       -0.5335880  -0.4345317  -0.2170806  0.0442851  0.0889461
## s.e.   0.1079563   0.1235816   0.1308744  0.1243981  0.1096536
##
## sigma^2 estimated as 94.8056:  log likelihood = -321.76,  aic = 655.51
```

```
#Hacer predicciones
predicciones.modelo6<-forecast.Arima(modelo6,h=10)#Pronosticar los próximos 10 periodos
```

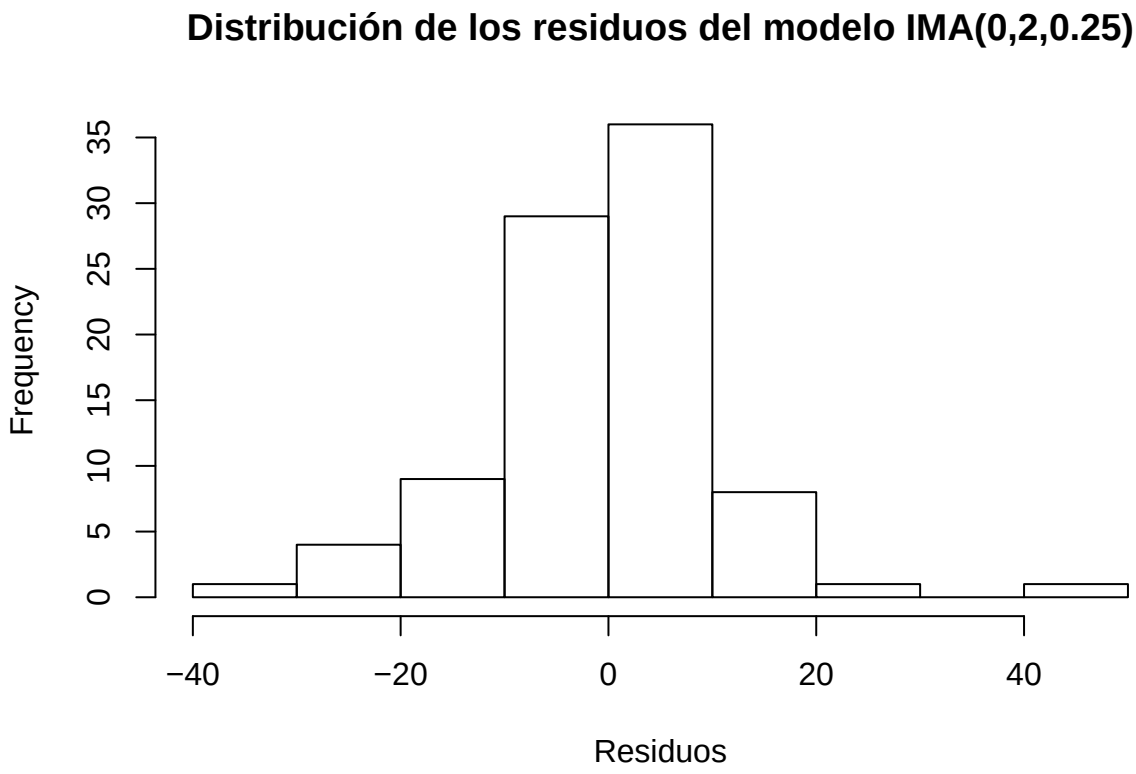
Evaluación

Modelo IMA(0,2,0.25)

```
#Ver serie de tiempo con las prediciones
plot.forecast(predicciones.modelo1,
              main = 'Prediciones modelo IMA(0,2,0.25)',
              xlab='Año',
              ylab ='Cantidad de residentes australianos')
```

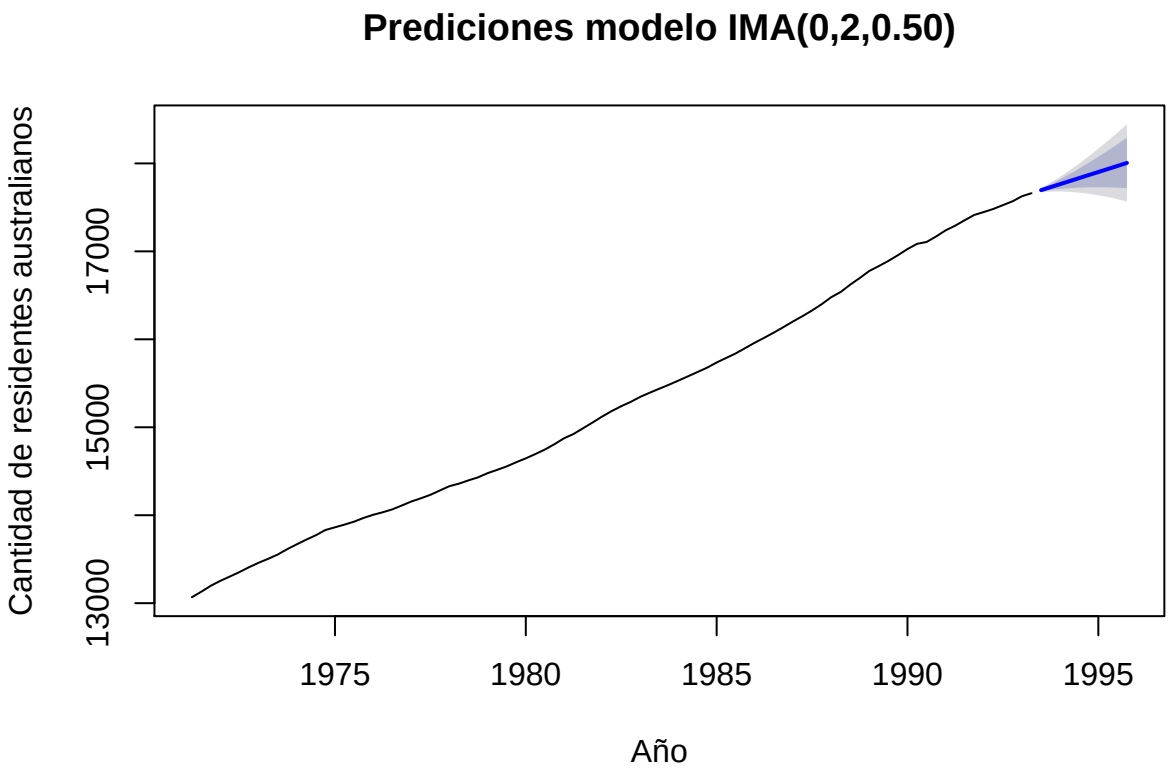



```
#Distribución de los residuos
hist(predicciones.modelo1$residuals,
     main = 'Distribución de los residuos del modelo IMA(0,2,0.25)',
     xlab='Residuos')
```

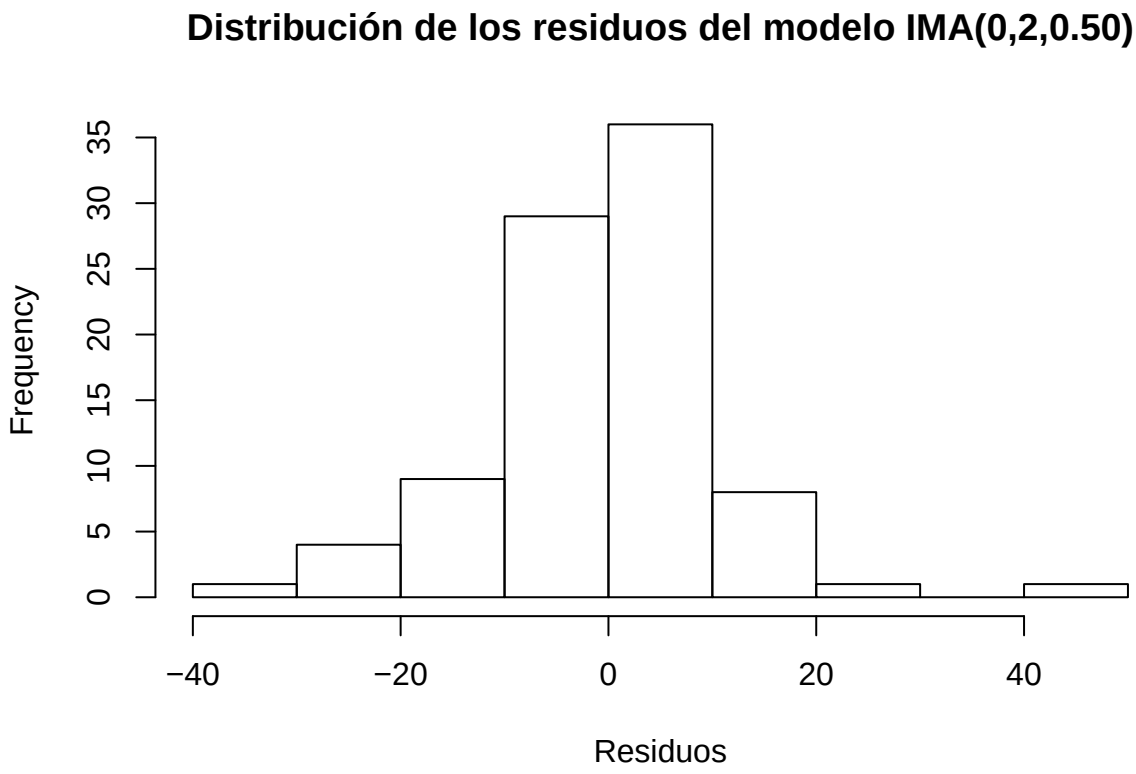


Modelo IMA(0,2,0.50)

```
#Ver serie de tiempo con las prediciones
plot.forecast(predicciones.modelo2,
              main = 'Prediciones modelo IMA(0,2,0.50)',
              xlab='Año',
              ylab ='Cantidad de residentes australianos')
```

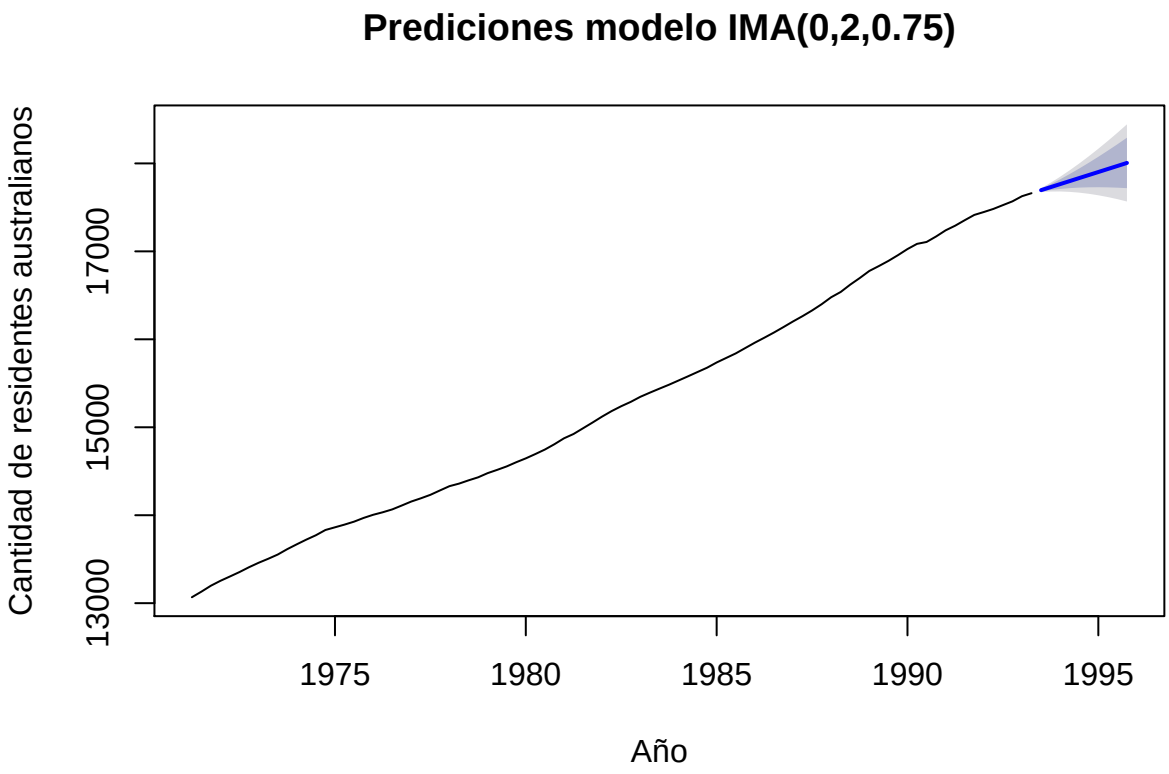



```
#Distribución de los residuos
hist(predicciones.modelo2$residuals,
     main = 'Distribución de los residuos del modelo IMA(0,2,0.50)',
     xlab='Residuos')
```

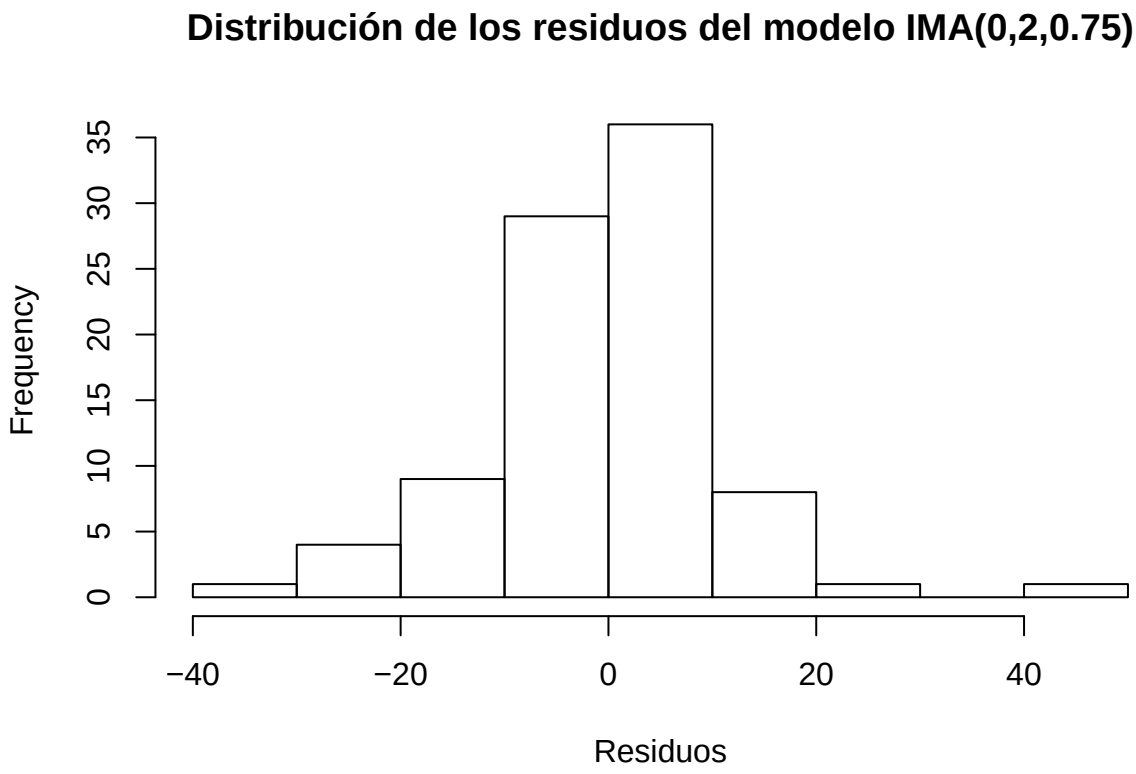


Modelo IMA(0,2,0.75)

```
#Ver serie de tiempo con las prediciones
plot.forecast(predicciones.modelo3,
              main = 'Prediciones modelo IMA(0,2,0.75)',
              xlab='Año',
              ylab ='Cantidad de residentes australianos')
```

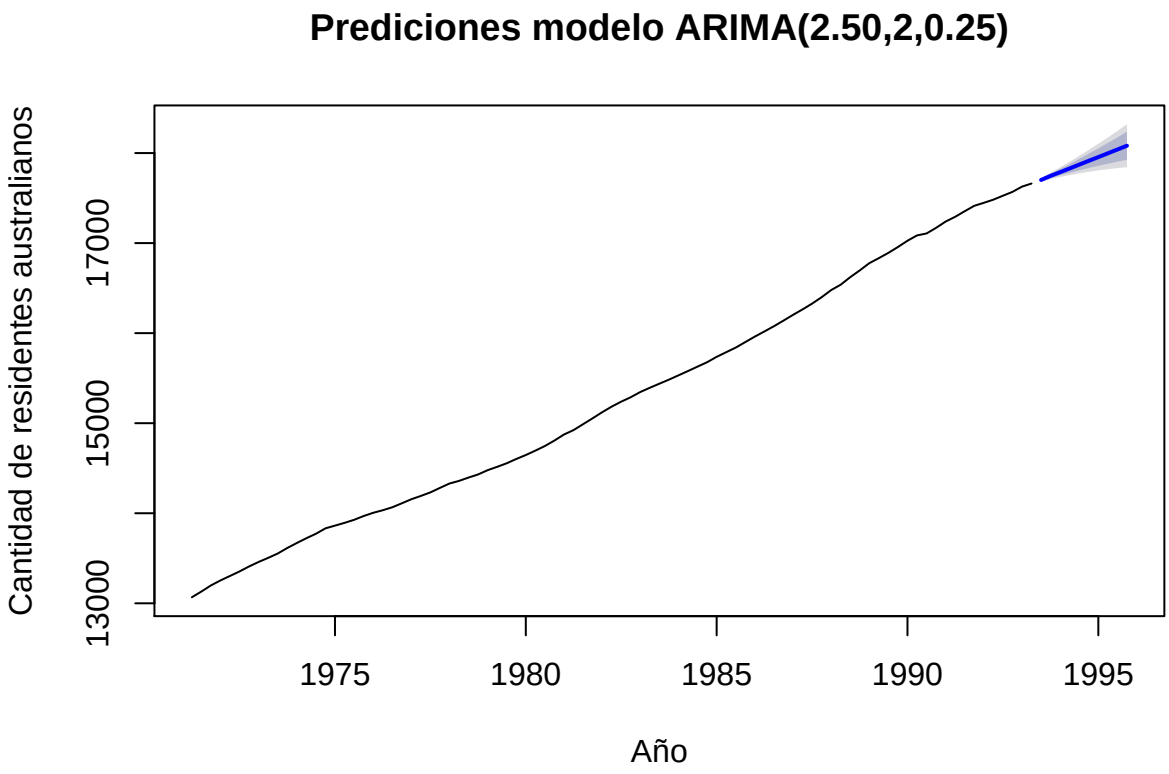



```r
#Distribución de los residuos
hist(predicciones.modelo3$residuals,
     main = 'Distribución de los residuos del modelo IMA(0,2,0.75)',
     xlab='Residuos')
```

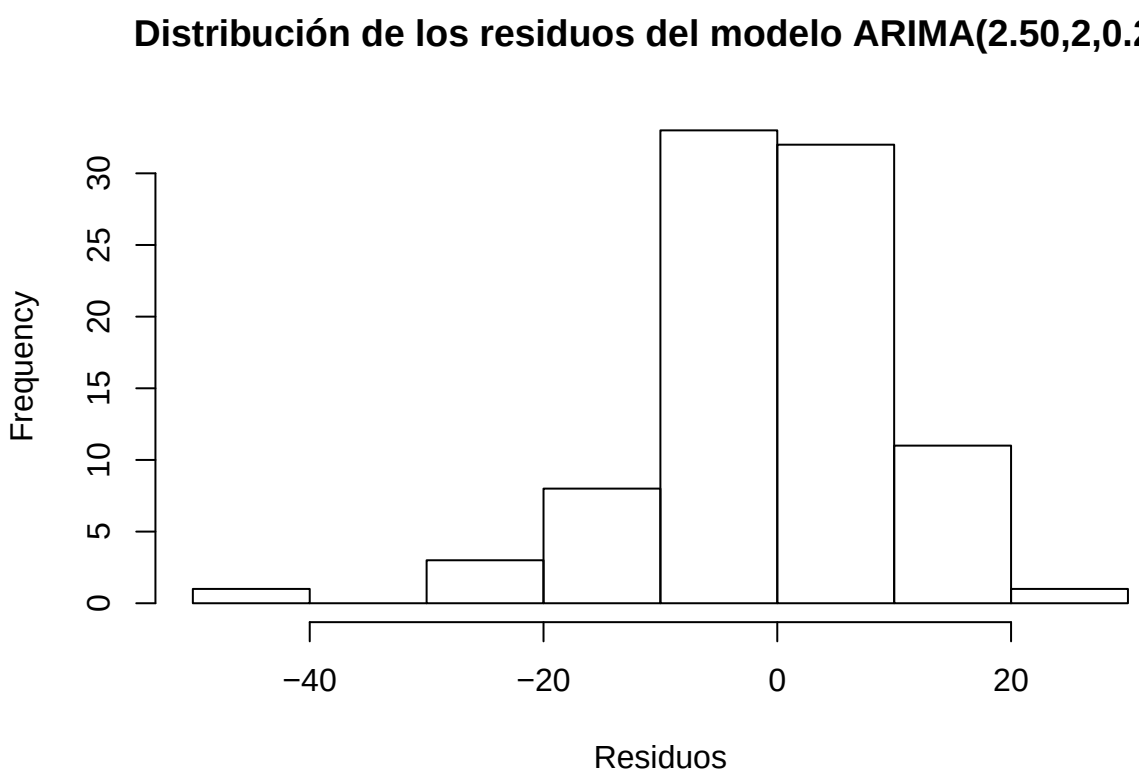


Modelo ARIMA(2.50,2,0.25)

```
#Ver serie de tiempo con las prediciones
plot.forecast(predicciones.modelo4,
              main = 'Prediciones modelo ARIMA(2.50,2,0.25)',
              xlab='Año',
              ylab ='Cantidad de residentes australianos')
```

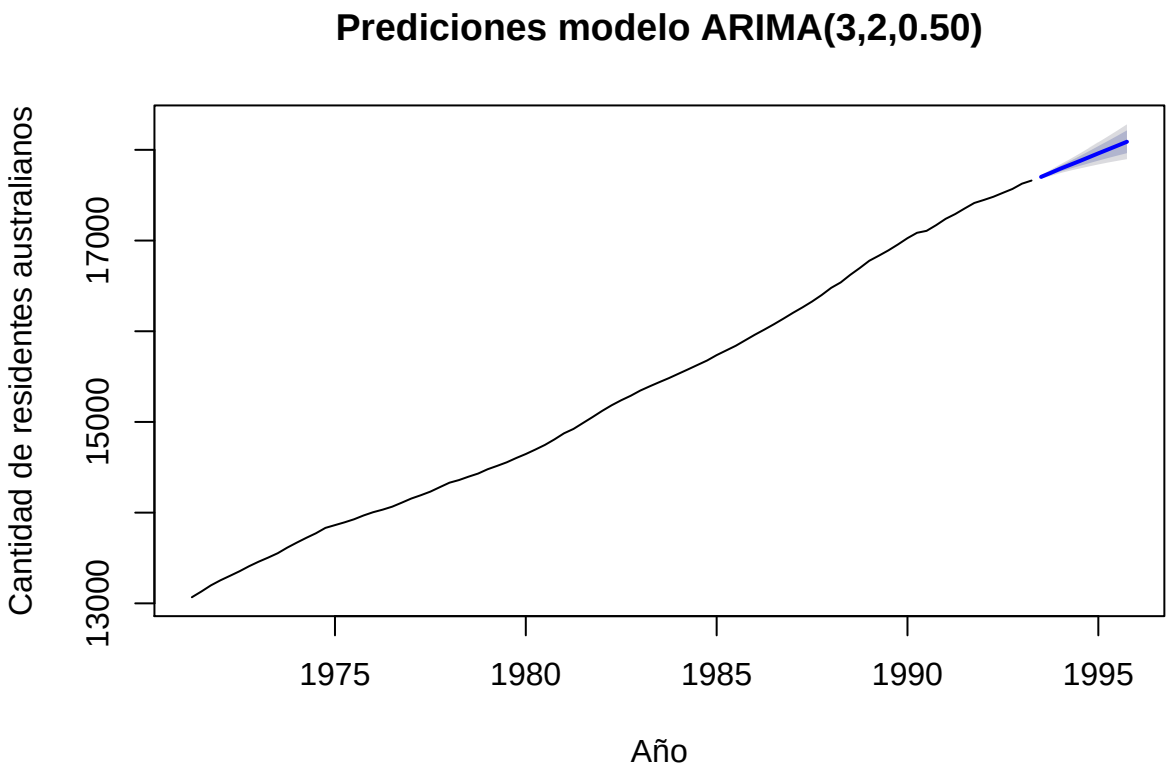



```
#Distribución de los residuos
hist(predicciones.modelo4$residuals,
     main = 'Distribución de los residuos del modelo ARIMA(2.50,2,0.25)',
     xlab='Residuos')
```

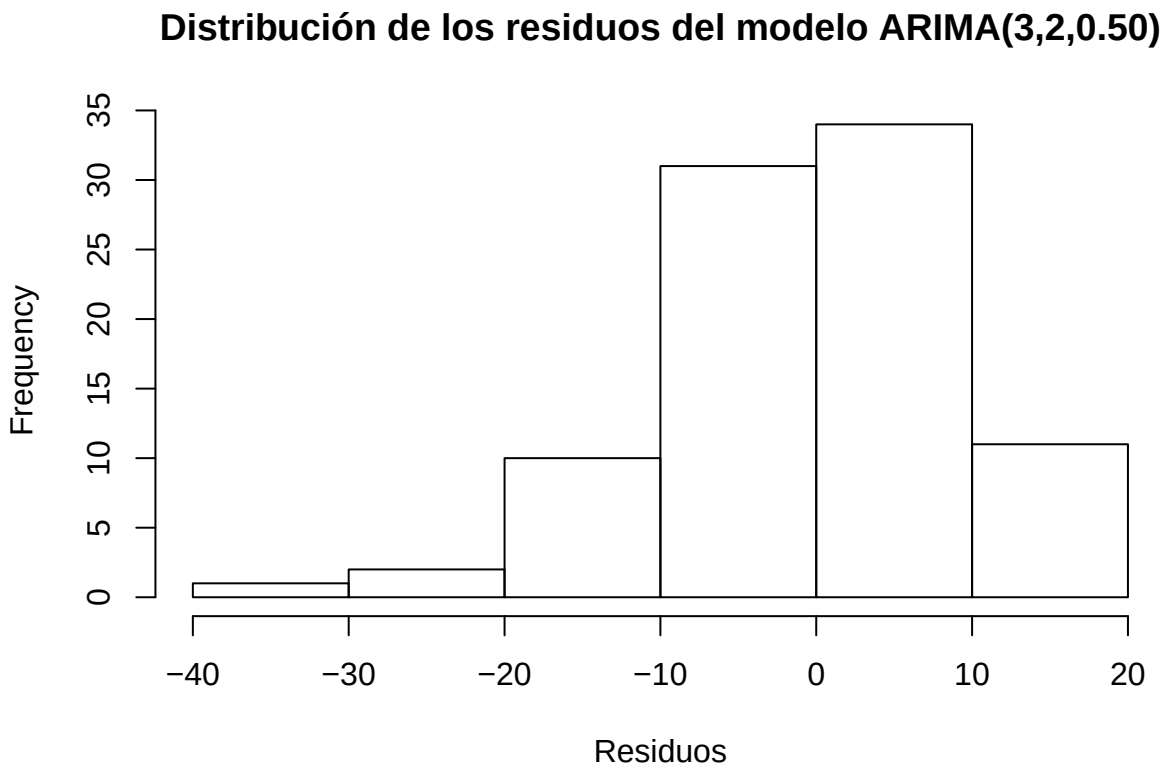


Modelo ARIMA(3,2,0.50)

```
#Ver serie de tiempo con las prediciones
plot.forecast(predicciones.modelo5,
              main = 'Prediciones modelo ARIMA(3,2,0.50)',
              xlab='Año',
              ylab ='Cantidad de residentes australianos')
```

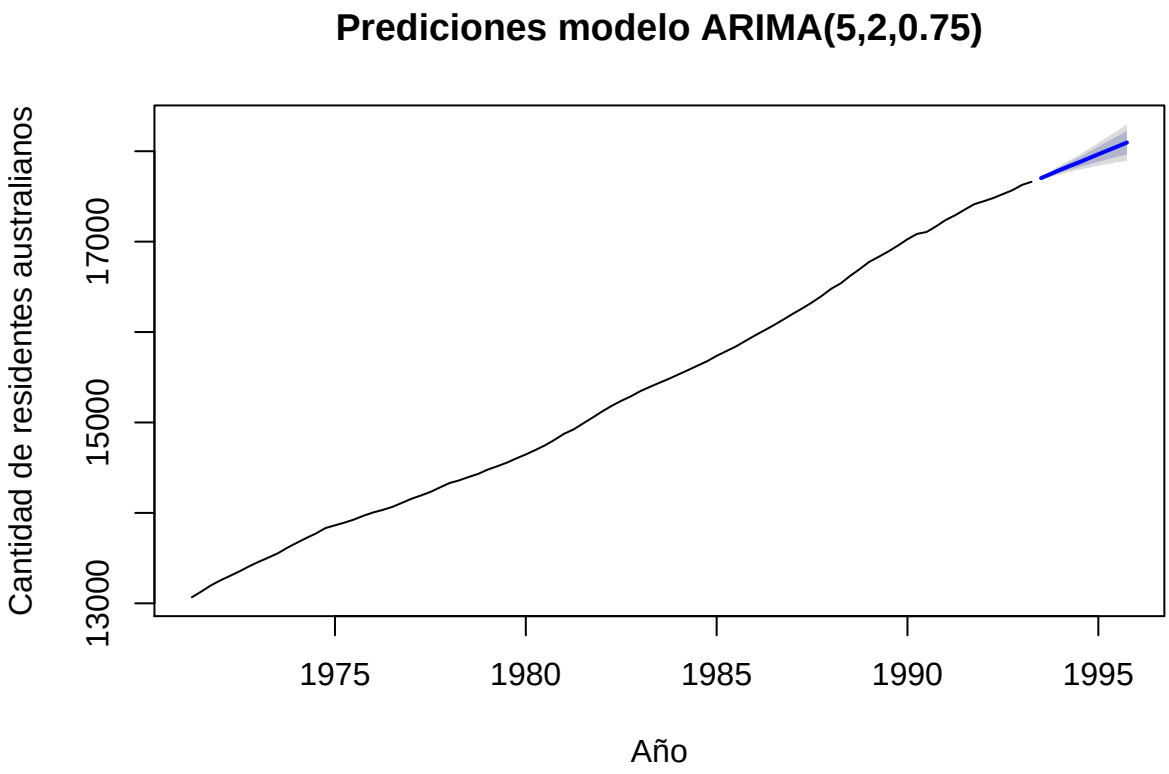



```r
#Distribución de los residuos
hist(predicciones.modelo5$residuals,
     main = 'Distribución de los residuos del modelo ARIMA(3,2,0.50)',
     xlab='Residuos')
```

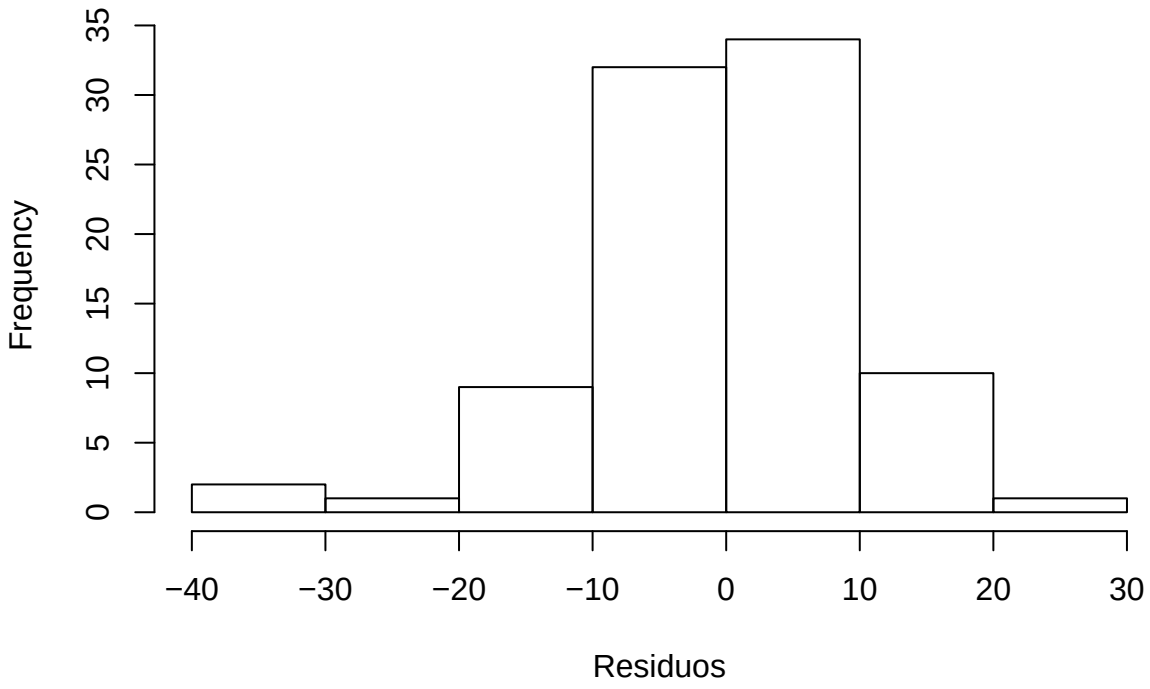


Modelo ARIMA(5,2,0.75)

```
#Ver serie de tiempo con las prediciones
plot.forecast(predicciones.modelo6,
              main = 'Prediciones modelo ARIMA(5,2,0.75)',
              xlab='Año',
              ylab ='Cantidad de residentes australianos')
```




```
#Distribución de los residuos
hist(predicciones.modelo6$residuals,
     main = 'Distribución de los residuos del modelo ARIMA(5,2,0.75)',
     xlab='Residuos')
```



Resultados

- Al crear los 6 posibles modelos ARIMA, se tiene el siguiente resumen:

Modelo IMA(0,2,0.25) log likelihood = -335.23, aic = 672.46

Modelo IMA(0,2,0.50) log likelihood = -335.23, aic = 672.46

Modelo IMA(0,2,0.75) log likelihood = -335.23, aic = 672.46

Modelo ARIMA(2.50,2,0.25) log likelihood = -324.93, aic = 655.86

Modelo ARIMA(3,2,0.50) log likelihood = -322.08, aic = 652.17

Modelo ARIMA(5,2,0.75) log likelihood = -321.76, aic = 655.51

- De acuerdo con el Akaike Information Criteria (AIC), el modelo donde se pierde menos información es el **Modelo ARIMA(3,2,0.50)**

Las predicciones son las siguientes:

```
predicciones.modelo5
```

```
##         Point Forecast       Lo 80       Hi 80       Lo 95       Hi 95
## 1993 Q3    17701.68604 17689.15788 17714.21420 17682.52588 17720.84621
## 1993 Q4    17745.68425 17723.42665 17767.94185 17711.64419 17779.72431
```

```
## 1994 Q1    17791.20335 17759.83851 17822.56819 17743.23497 17839.17174
## 1994 Q2    17832.68689 17791.45875 17873.91504 17769.63389 17895.73990
## 1994 Q3    17874.64067 17821.44771 17927.83364 17793.28906 17955.99229
## 1994 Q4    17917.79095 17851.58313 17983.99876 17816.53483 18019.04706
## 1995 Q1    17961.13066 17881.33610 18040.92523 17839.09541 18083.16592
## 1995 Q2    18003.70351 17909.57254 18097.83449 17859.74261 18147.66442
## 1995 Q3    18046.28946 17936.87943 18155.69949 17878.96125 18213.61766
## 1995 Q4    18089.16896 17963.69880 18214.63912 17897.27892 18281.05900
```

```
predicciones.modelo5$lower
```

```
##               80%         95%
##  [1,] 17689.15788 17682.52588
##  [2,] 17723.42665 17711.64419
##  [3,] 17759.83851 17743.23497
##  [4,] 17791.45875 17769.63389
##  [5,] 17821.44771 17793.28906
##  [6,] 17851.58313 17816.53483
##  [7,] 17881.33610 17839.09541
##  [8,] 17909.57254 17859.74261
##  [9,] 17936.87943 17878.96125
## [10,] 17963.69880 17897.27892
```

```
predicciones.modelo5$upper
```

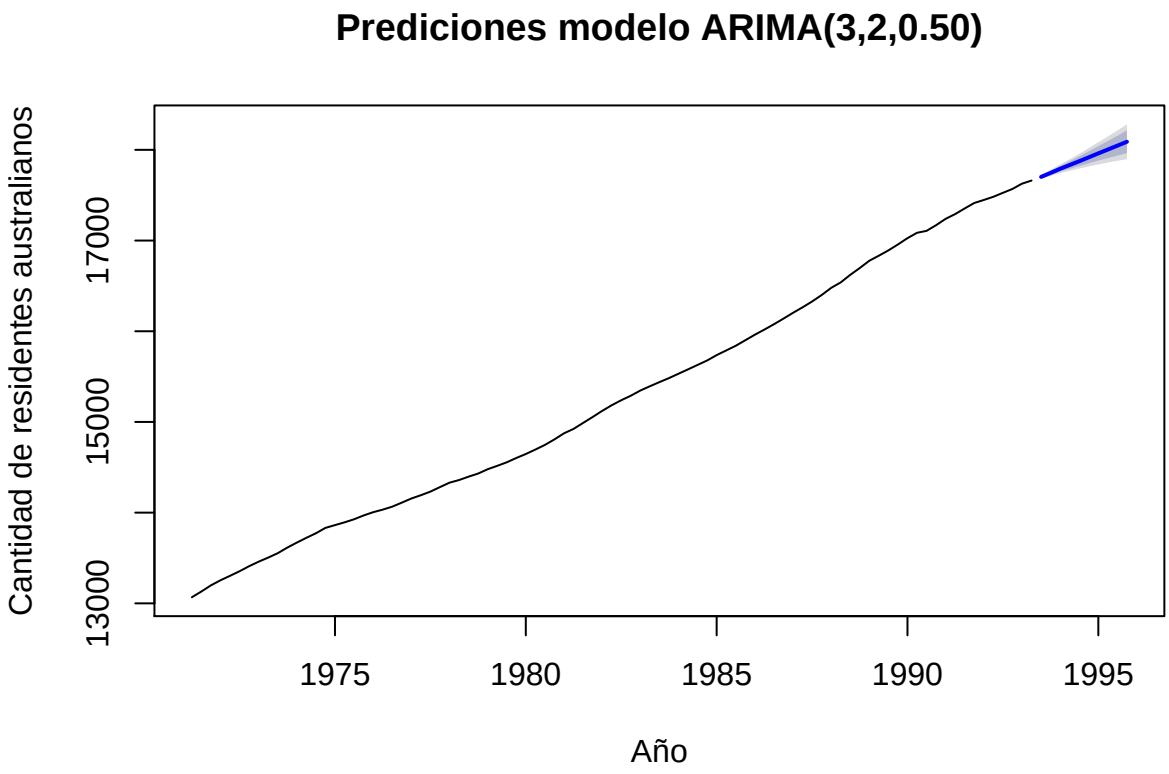
```
##               80%         95%
##  [1,] 17714.21420 17720.84621
##  [2,] 17767.94185 17779.72431
##  [3,] 17822.56819 17839.17174
##  [4,] 17873.91504 17895.73990
##  [5,] 17927.83364 17955.99229
##  [6,] 17983.99876 18019.04706
##  [7,] 18040.92523 18083.16592
##  [8,] 18097.83449 18147.66442
##  [9,] 18155.69949 18213.61766
## [10,] 18214.63912 18281.05900
```

```
as.numeric(predicciones.modelo5$mean)
```

```
##  [1] 17701.68604 17745.68425 17791.20335 17832.68689 17874.64067
##  [6] 17917.79095 17961.13066 18003.70351 18046.28946 18089.16896
```

En este grafico podemos observar que el crecimiento de residente va a incrementarse en los próximos años.

```
plot.forecast(predicciones.modelo5,
main = 'Prediciones modelo ARIMA(3,2,0.50)',
xlab='Año',
ylab ='Cantidad de residentes australianos')
```


Predicciones modelo ARIMA(3,2,0.50)
Cantidad de residentes australianos
13000
15000
17000
1975
1980
1985
1990
1995
Año