Practica 2

Los archivos utilizados en la practica se encuentran en el repositorio practica2 de Github accesible en el siguiente link: https://github.com/carolinekonig2/practica2

1) Descripción del dataset: El dataset 'arrhytmia' son datos de ensayos clinicos que comprenden mediciciones de electrocardiograma de diferente pacientes. El dataset fue utilizado en estudios anteriores para predecir la presencia de arritmia y de clasificarlas en 16 subcategorias de arritmia.
   En este estudio el objetivo es analizar los atributos y determinar cual son los mas relevantes para la predicción de la variable (el diagnostico) . Aparte de la existencia de un programa de predicción automatico, que permite a los medicos a realizar un diagnostico mas rapido y detallado, los expertos quieren conocer en muchas ocasiones las variables que mas influencian la predicción para entender mejor el razonamiento del modelo .

   El conjunto de datos 'arrhytmia' esta publicado en el 'UCI machine learning repositorio' en el siguiente link http://archive.ics.uci.edu/ml/datasets/Arrhythmia
   El dataset original esta contenido en la carpeta dataset/original del proyecto y corresponde a los ficheros 'arrhytmia.txt' y 'arrhytmia.names'.

2) Limpieza de datos:
   El conjunto de datos original comprende 452 entradas de 279 atributos cada una. En el presente estudio llevamos una analisis mas simplificado a cabo y por tanto seleccionamos un subconjunto de atributos. Se incluyen los atributos sobre las caracteristicas personales, como edad, genero, altura, peso (Atributos 1-4). Los siguientes atributos corresponden a los datos medidos en el ECG (Atributos 5-15). Las arithmias se predicen por la duración, amplitud y morfologia del complejo QRS en uno de 16 diferentes categorias (V280) en el dataset original. Simplificamos la clasificación en este estudio y distinguimos el diagnostico en sano (clase=1), diagnostico con arrithmia (clase 2-15) y la categoria 'no conocido' (clase 16). Las dos clases mas importantes (sano y arrithmia) tienen 245 y 185 instancias respectivamente, por tanto la distribución entre las dos clases es casi equitativa y podemos trabajar con un conjunto de datos equilibrado.

   Mostramos la descripción de los atributos originales a continuación:

- 1 Age: Age in years , linear
- 2 Sex: Sex (0 = male; 1 = female) , nominal
- 3 Height: Height in centimeters , linear
- 4 Weight: Weight in kilograms , linear
- 5 QRS duration: Average of QRS duration in msec., linear
- 6 P-R interval: Average duration between onset of P and Q waves in msec., linear
- 7 Q-T interval: Average duration between onset of Q and offset of T waves in msec., linear
- 8 T interval: Average duration of T wave in msec., linear
- 9 P interval: Average duration of P wave in msec., linear
- Vector angles in degrees on front plane of:, linear
- 10 QRS
- 11 T

- 12 P
- 13 QRST
- 14 J
- 15 Heart rate: Number of heart beats per minute ,linear
- 280 Arrhytmia diagnostic (16 values):

Class code :	Class :	Number of instances:
01	Normal	245
02	Ischemic changes (Coronary Artery Disease)	44
03	Old Anterior Myocardial Infarction	15
04	Old Inferior Myocardial Infarction	15
05	Sinus tachycardy	13
06	Sinus bradycardy	25
07	Ventricular Premature Contraction (PVC)	3
08	Supraventricular Premature Contraction	2
09	Left bundle branch block	9
10	Right bundle branch block	50
11	1. degree AtrioVentricular block	0
12	2. degree AV block	0
13	3. degree AV block	0
14	Left ventricule hypertrophy	4
15	Atrial Fibrillation or Flutter	5
16	Others	22

2b) Tras cargar los datos en el R Studio realizamos un analisis estadistico de los atributos personales (V1-V4)

Analisis de los datos personales

```
 V1 (Edad)       V2 (Genero)      V3 (Altura)     V4 (Peso)
Min.   : 0.00   Min.   :0.0000   Min.   :105.0   Min.   :  6.00
1st Qu.:36.00   1st Qu.:0.0000   1st Qu.:160.0   1st Qu.: 59.00
Median :47.00   Median :1.0000   Median :164.0   Median : 68.00
Mean   :46.47   Mean   :0.5509   Mean   :166.2   Mean   : 68.17
3rd Qu.:58.00   3rd Qu.:1.0000   3rd Qu.:170.0   3rd Qu.: 79.00
Max.   :83.00   Max.   :1.0000   Max.   :780.0   Max.   :176.00
```

Los datos estadisticos indican la existencia de valores extremos en los datos. Por ejemplo el valor de 780 cm como maximo de la altura indica la existencia de valores extremos.
Al igual el valor minimo de 6 como peso de una persona podria indicar un dato erroneo o la edad de 0 como valor minimo. Una comprobación de los registros de entrada correspondiente a la edad de 0 y la de peso 6 kg muestra que la relación entre edad y peso seria coherente, pero que los valores de la altura son erroneos (608 y 780 cm no son posibles). En consecuencia se consideran los registros 142 y 317 erroneos :

```
      Edad        Genero      altura      peso
[142] 0           0           608         10

[317] 1           1           780         6
```

Decidimos transformar las alturas indicadas, dado que es factible que se han reportado en 'mm' en lugar de 'cm'. Transformamos 608 a 61 y 780 a 78, que serian alturas racionales para niños de la edad de 0-1 año.
A continuación llevamos a cabo un analisis de los registros acerca de valores vacios. Un test de valores vacios mediante el comando ' which(is.na(dataset[c]))' para cada variable V1 -V4 muestra que el conjunto de datos no contiene valores vacios.

Decidimos tambien crear un nuveo atributo, el mass body index (MBI) como dato derivado de los atributos altura y peso. Vemos la distribución de sus valores a continuación:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.203  22.459  25.036  25.269  27.893  61.622
```

Analisis de los datos de ECG

```
V5              V6              V7              V8              V9
Min.   : 55.00   Min.   :  0.0   Min.   :232.0   Min.   :108.0   Min.   :  0
1st Qu.: 80.00   1st Qu.:142.0   1st Qu.:350.0   1st Qu.:148.0   1st Qu.: 79
Median : 86.00   Median :157.0   Median :367.0   Median :162.0   Median : 91
Mean   : 88.92   Mean   :155.2   Mean   :367.2   Mean   :169.9   Mean   : 90
3rd Qu.: 94.00   3rd Qu.:175.0   3rd Qu.:384.0   3rd Qu.:179.0   3rd Qu.:102
Max.   :188.00   Max.   :524.0   Max.   :509.0   Max.   :381.0   Max.   :205


V9           V10               V11 (T)        V12 (P)       V13 (QRST)
Min.   :  0   Min.   :-172.00   52     : 13   60     : 23   49     :  9
1st Qu.: 79   1st Qu.:   3.75   36     : 10   ?      : 22   55     :  9
Median : 91   Median :  40.00   42     :  9   61     : 16   59     :  9
Mean   : 90   Mean   :  33.68   ?      :  8   56     : 14   62     :  9
3rd Qu.:102   3rd Qu.:  66.00   10     :  8   58     : 13   26     :  8
Max.   :205   Max.   : 169.00   33     :  8   68     : 12   33     :  8
                                (Other):396   (Other):352   (Other):400

V14 (J)       V15 (Heart rate)
?      :376   63     : 21
84     :  3   72     : 21
-157   :  2   70     : 20
-164   :  2   73     : 19
-93    :  2   81     : 18
103    :  2   68     : 17
(Other): 65   (Other):336
```

Un analisis acerca de los valores vacios indica que los atributos V11-V15 contienen valores vacios indicados con el signo '?' en el conjunto de datos. Por otro lado los valores de las variables v11-V15 estan guardados como strings y no como valores enteros. Se lleva a cabo la substitución de las entradas con valor '? ' por el valor '0' y se transforma el tipo de datos a enteros. Decidimos descartar la variable V14, dado que 376 entradas tenian un valor desconocido.

Tras las correciones y la transformación las variables V11-V13 y V15 tienen las siguientes estadisticas:

```
      V11                V12                V13
 Min.   :-170.00   Min.   :-170.00   Min.   :-135.00
 1st Qu.:  36.00   1st Qu.:  36.00   1st Qu.:  12.00
 Median :  54.50   Median :  54.50   Median :  40.00
 Mean   :  46.53   Mean   :  46.53   Mean   :  36.63
 3rd Qu.:  64.00   3rd Qu.:  64.00   3rd Qu.:  62.00
 Max.   : 176.00   Max.   : 176.00   Max.   : 166.00
 V15
 Min.   :  0.0
 1st Qu.: 65.0
 Median : 72.0
 Mean   : 74.3
 3rd Qu.: 81.0
 Max.   :163.0
```

El nuevo dataset esta contenido en la carpeta dataset/new del proyecto y corresponde a los ficheros 'matriz.txt' y 'matriz_standardized.txt' (contiene los datos estandardizados).

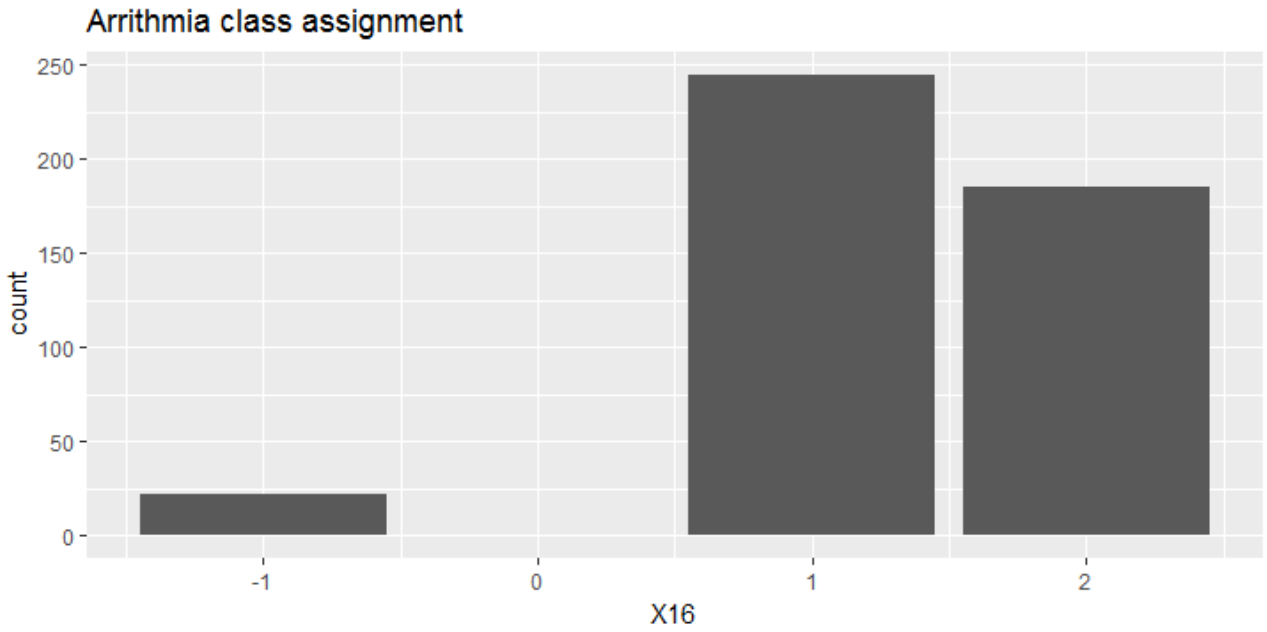
3) Analisis de los datos

El conjunto de datos a analizar se compone de los atributos V1- V4 (datos personales), el nuevo atributo MBI, los atributos V5-V13 y V15 (datos de ECG) y la variable de diagnostico V280. Distinguimos el grupo de los datos personales (V1-V4 y MBI), el grupo de los datos de ECG (V5-V13/V15). Los datos de ECG estan ademas diferenciados en los atributos V5-V9 (tiempos de ciclo QSR) y atributos V10-V13 (angulos) y la frecuencia cardiaca (V15).

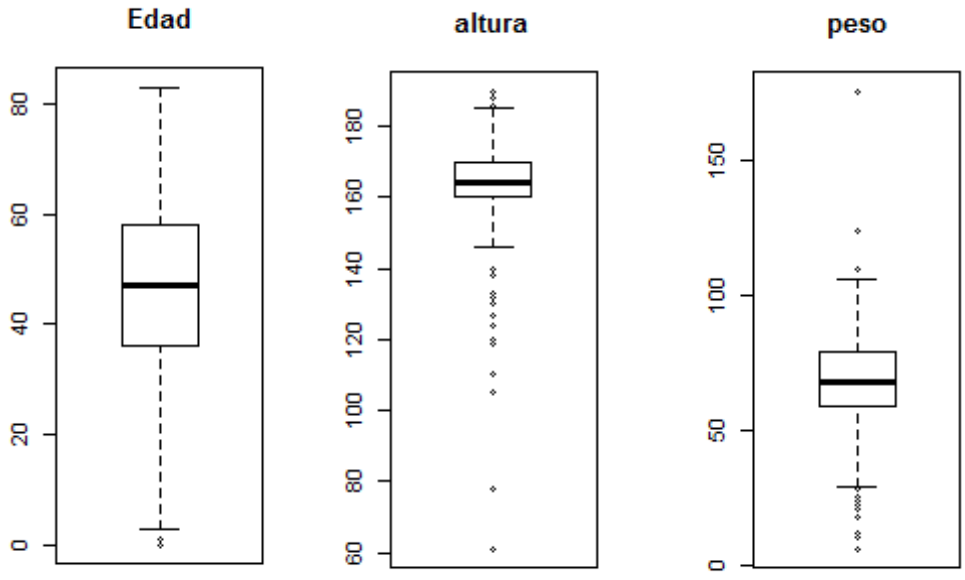
La siguiente tabla muestra la relación entre el nuevo conjunto de datos y los atributos del antiguo conjunto de datos:

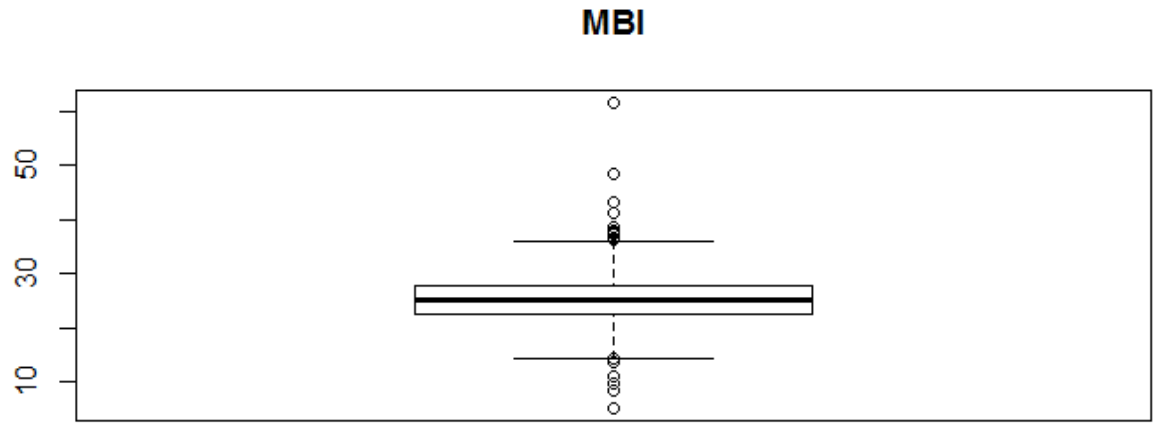
Atributo nuevo	**Descripción**	**Atributo nuevo**	**Descripción**
X1	Edad (V1)	X9	T intervalo (V8)
X2	Genero (V2)	X10	P intervalo (V9)
X3	Altura (V3)	X11	Angulo QRS (V10)
X4	Peso (V4)	X12	Angulo T (V11)
X5	MBI	X13	Angulo P (V12)
X6	QRS duracion (V5)	X14	Angulo QRST (V13)
X7	P-R intervalo (V6)	X15	Frequ. Cardiaca (V15)
X8	Q-T intervalo (V7)	X16	Diagnostico (V280)

La variables de diagnostico X16 puede tomar 3 valores distintos y diferencia los tres diagnosticos posibles: diagnostico como sano (clase 01), arrithmia (clase 02) y otros que no tengan posible diagnostivo (clase -1). El grafico de barras a continuación explica la distribución de las instancias en estos tres grupos: 245 con diagnostico sano, 185 con diagnostico de arrithmia y 22 casos sin posible diagnostico.

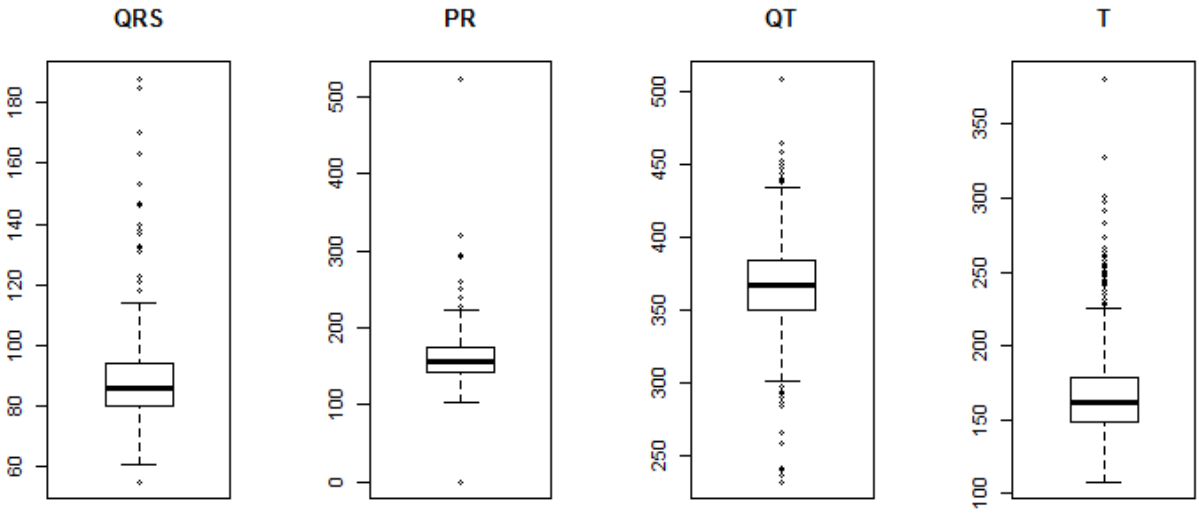


A continuación se muestra mediante boxplots la distribución de los valores de los atributos. Los boxplots informan de la distribución de los valores y indican la existencia de outliers, que son puntos fuera del rango inter-cuartil. A continuación se muestran los boxplots de los diferentes atributos. Se puede apreciar la existencia de outliers en los diferentes distribuciones:

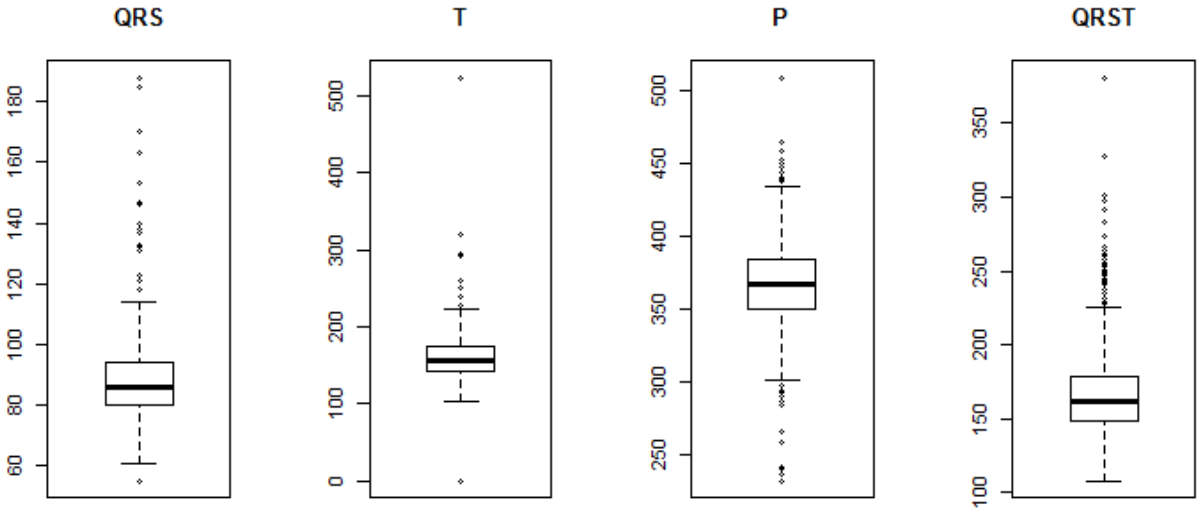




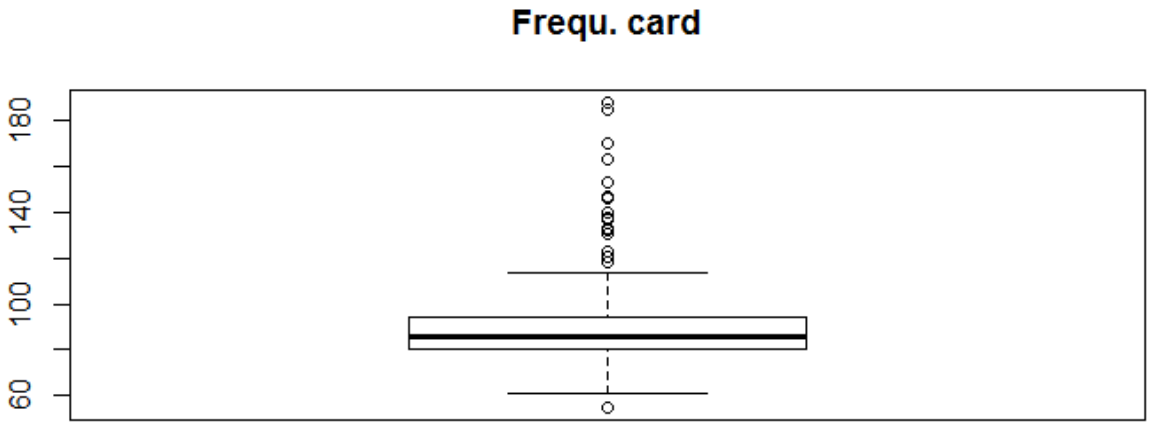
A continuación se muestra un boxplot de la distribución de los valores de las medidas de QSR:



A continuación se muestra un boxplot de la distribución de los valores de los angulos medidos en el ECG:



A continuación se muestra un boxplot de la distribución de la frecuencia cardiaca:



3b) De acuerdo con los datos estadisticos de los atributos mostrados en el apartado anterior, no hay ningun atributo con valores estandarizados, es decir con media 0 y desviación estandar igual a 1. Como primer paso aplicamos una estandarización a los atributos numericos (excepto a la variable X2 y X16, que es el genero y el diagnostico) para que los valores tengan un valor medio de 0 y desviación estandar 1.

3c) Aplicación de pruebas estadisticas para la comprobación de los grupos de datos.

En este apartado reportamos los tests estadisticos realizados para determinar si los atributos diferencian el grupo diagnosticado como sano de las con diagnostico de aritmia (No consideramos el grupo de instancias sin posible diagnostico, al tratarse de un grupo minoritario de solo 22 instancias, que no tienen interes para el estudio).

Shapiro Test para la comprobación de normalidad en la distribución

El Shapirto Test comprueba si una distribución de valores proviene de una distribución normal. La hipotesis nula es que la distribución es normal. A continuación se muestra el resultado del test de Shapiro para las distribuciones de cada variable y grupo.

Variable	Class 01 (Sano)		Class 02 (arritmia)	
	p-value	Normal	p-value	normal
X1(edad)	0.1734	si	1,86E-002	no
X3(altura)	3.66e-06	no	2.2e-16	no
X4 (peso	0.00731	si	5.555e-10	no
X5 (MBI)	6.695e-09	no	7.113e-09	no
X6 (V5)	0.04373	no	2.047e-14	no
X7 (V6)	2.2e-16	no	2.2e-16	no
X8 (V7)	9.146e-06	no	0.004372	no
X9 (V8)	9.88e-11	no	8.694e-11	no
X10 (V9)	1.004e-11	no	2.349e-12	no
X11 (V10)	7.594e-05	no	1.836e-0	no
X12 (V11)	2.2e-16	no	9.997e-11	no
X13 (V12)	2.2e-16	no	9.997e-11	no
X14 (V13)	0.0113	no	0.00251	no
X15 (V15)	8.543e-05	no	6.224e-07	no

El resultado del test de Shapiro muestra que la gran mayoria de las distribuciones de los atributos según grupo no siguen una distribución normal.

A continuación analizamos mediante test de hypothesis si la distribución de valores para cada variable según su grupo (sano o aritmia) provienen de distribuciones de datos independientes. Para ello analizamos si existe evidencia que la media de las dos distribuciones de datos sea igual, es decir que ambas muestran provienen de la misma población. La hipothesis nula seria que ambos particiones tienen la misma media.

Llevamos a cabo el analisis mediante t-test, Wilcoxon rank test y el test de Kolmogorov-Smirov. El Wilcoxon rank test no necesita que la distribución de las muestras sea una distribución normal. Hemos visto en el apartado anterior que la mayoria de atributos no tienen distribución normal por tanto el Wilcxon rank test parece espcialmente adecuado para nuestros datos. Sin embargo reportamos tambien los valores del hipothesis test para t-test a modo de comparación, dado que aunque el t-test asume normalidad en las muestras este test esta considerado robusto a la falta de normalidad en las muestras y deberia proporcionar tambien resultados validos. El test de Kolmogorov-Smirov no establece requerimientos sobre la distribución de las muestras por tanto lo hemos elegido como otro test de interes.

Variable	T-test		Wilcoxon rank test		Kolmogorov-Smirov	
	p-value	Sign. difference	p-value	Sign difference	p-value	Sign difference
X1(edad)	0.6608	no	0.1493	no	0.009021	no
X3(altura)	0.1695	no	0.3158	no	0.1617	no
X4 (peso	0.4899	no	0.8219	no	0.6047	no
X5 (MBI)	0.5673	no	0.6085	no	0.7413	no
X6 (V5)	2.215e-10	si	7.125e-11	si	2.034e-09	si
X7 (V6)	0.1951	no	0.512	no	0.4906	no
X8 (V7)	0.9383	no	0.7447	no	0.146	no
X9 (V8)	2.059e-05	si	0.01451	si	0.0001103	si
X10 (V9)	0.04405	no	0.307	no	0.5393	no
X11 (V10)	0.06856	no	0.2838	no	0.04733	si
X12 (V11)	0.3367	no	0.3906	no	0.5466	no
X13 (V12)	0.3367	no	0.3906	no	0.5466	no
X14 (V13)	0.3404	no	0.9759	no	0.01201	si
X15 (V15)	0.08699	no	0.3708	no	0.009861	si

Los resultados de los tres tests de hipothesis diferentes señalan que los valores de los atributos X6(V5) y X9(V8) provienen de dos poblaciones distintas (es decir, se podia rechazar con certeza que provengan de la misma población en los tres tests diferentes). En consecuencia sus valores son distintas y estas variables podrian ser buenos candidadatos para diferenciar entre ambos grupos del diagnostico. El test de Kolmogorov-Smirov ademas ha encontrado los atributos X11 (V10), X14 (V13) y X15(V15) tambien como variables que provienen de dos poblaciones independientes, por tanto consideramos estos variables tambien posibles candidatos para distinguir bien entre los dos grupos de diagnostico.

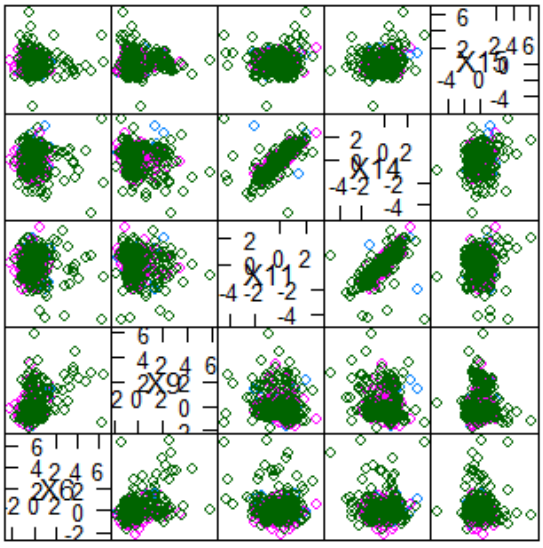
4) Representación de resultados

El objetivo del estudio era analizar los atributos y determinar cuales son los atributos que mejor distinguen entre ambos grupos o cuales son las que mas importancia tienen para distringuir entre los dos diagnosticos posibles (sanos o aritmia). A continuación reportamos los resultados del analisis sobre la importancia de los atributos para la predicción del diagnostico mediante una exploración visual de los datos y mediante metodos de selección de atributos.

4a) Exploración visual de los atributos:

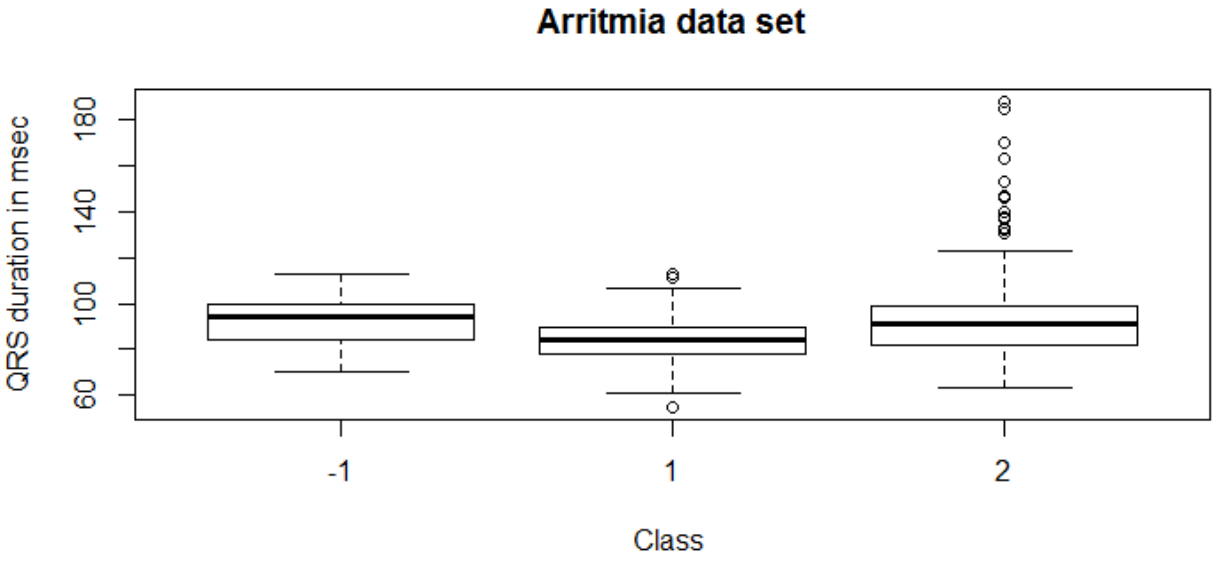
En este apartado realizamos la exploración visual de la distribución de los atributos señalados como posibles candidatos para distinguir entre los dos grupos de acuerdo con los resultados de los tests de hipothesis.

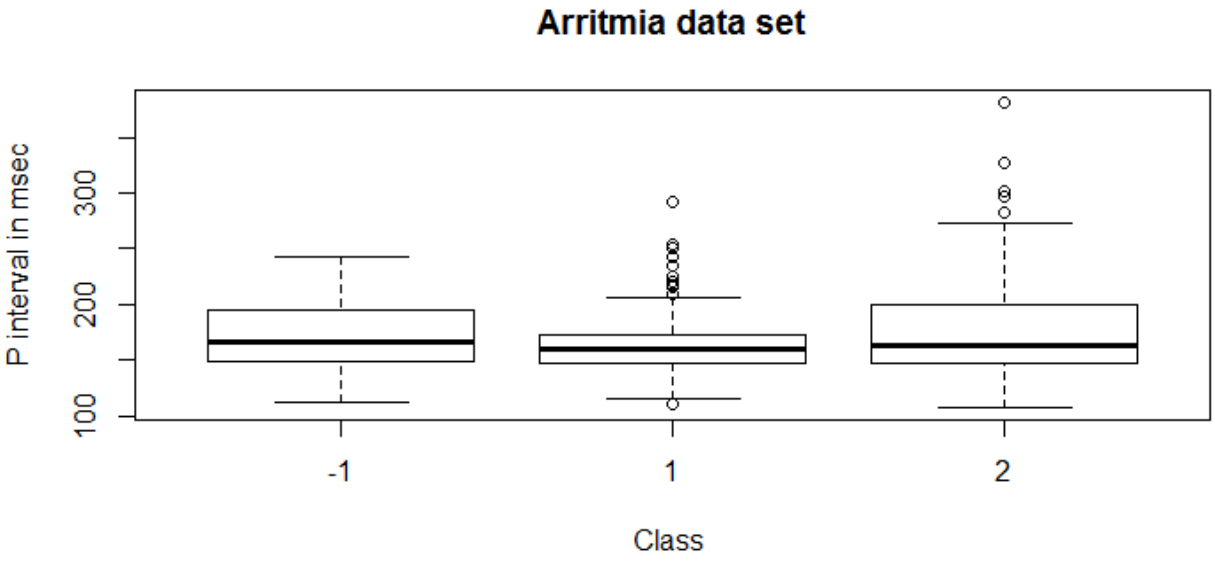
Un analisis mediante scatter plot de los atributos X6(V5), X9(V8), X11(V10),X14(V13) y X15 (V15) muestra que la combinación de dos atributos no es suficiciente para separar los dos grupos. En el scatter plot no se aprecia una separación clara entre el grupo de los puntos rosas y verdes.



Scatter Plot Matrix

Tambien la representación mediante boxplot de los atributos X6(V5) y X9(V8) no señala una clara diferencia entre el valor medio de ambos grupos ni en su distribución. En el caso de la variable X6 (V5-QRS duration) se aprecia que la clase 02 tiene mas valores altos, que aparecen como outliers fuera del cuartil. Lo mismo sucede para la variable X9 (V8-intervalo de P): aunque el valor medio de la clase 01 y 02 son parecidos, se aprecia una distribución mas alargada para la clase 02 y la existencia de mas valores elevados que aparecen como outliers.

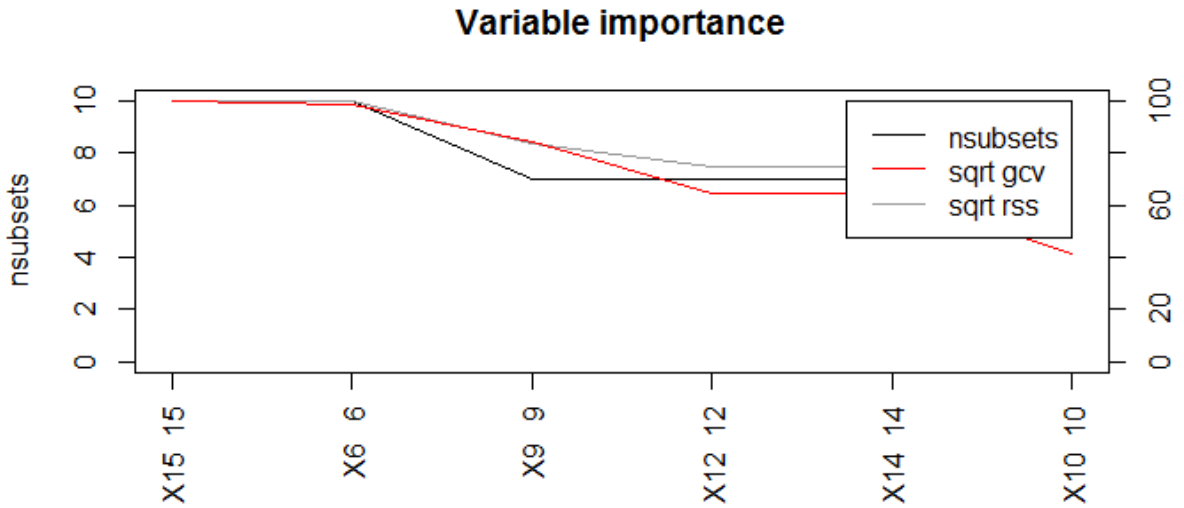




3b) Analisis mediante mediante metodos de selección de atributos:

En este apartado reportamos los resultados obtenidos mediante diferentes metodos de selección de atributos. Consideramos el metodo MARS (Multi-angle Regression and Shrinkage del package earth de R), selección de atributos mediante Random Forest y el metodo wrapper 'Boruta'.
El metodo MARS establece un ranking de importancia sobre los atributos. De los resultados del algoritmo vemos que los atributos X15 (V15) y X6 (V5) han obtenido un alto peso al aparecer en los 10 conjuntos de pruebas , seguido de variable X9(V8) que aparece en 7 conjuntos de prueba. El grafico 'Variable importance' representa graficamente el ranking de atributos y representa el numero de conjuntos donde la variable ha sido incluida (nsubsets), la raiz caudrado del general cross validation valor (gcv) y de la suma residual (rss) .

	nsubsets	gcv	rss
X15	10	100.0	100.0
X6	10	98.8	99.7
X9	7	84.6	83.9
X12	7	64.3	74.8
X14	7	64.3	74.8
X10	5	41.5	58.5

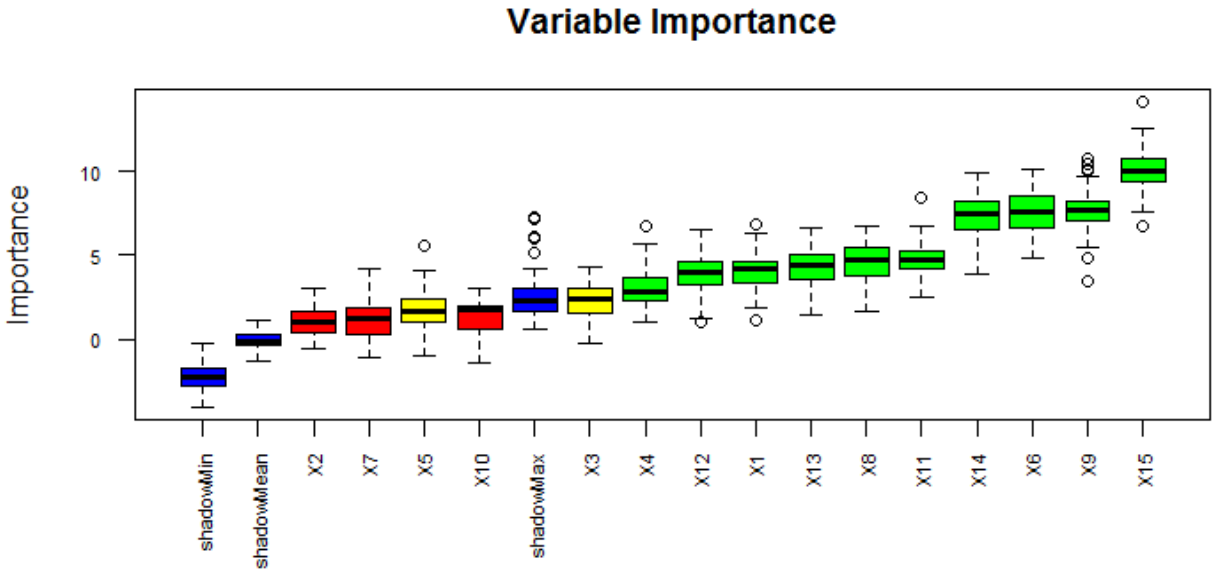


La selección de atributos mediante el algoritmo Random Forest permite establecer un ranking sobre la importancia de cada atributo midiendo la perdida media correspondiente a cada variable mediante la metrica 'area under the curve' (AUC):

X1	X2	X3	X4	X5
0.0086999085	-0.0017123478	0.0021996127	0.0018643965	0.0030628899
X6	X7	X8	X9	X10
0.0118809819	0.0029366861	0.0012013152	0.0127669766	0.0026834601
X11	X12	X13	X14	X15
0.0037602020	-0.0007096862	0.0023342846	0.0126079658	0.0120411754

Los resultados indican que la variable X6(V5), X9(V8), X14(V13) y X15(V15) han resultado los mas importantes al provocar mayor varianca en la variable a predecir.

Un analisis mediante el metodo wrapper 'Boruta' devuelve un ranking de importancia sobre los atributos. A continuación se muestra el resultado del algoritmo, donde los atributos X15(V15), X9(V8), X6(V5), X14(V13) y X11(V10) aparecen como los atributos de mas importancia. A continuación se muestra la representación grafica de la valoración de importancia de cada variable. La representación de los resultados es mediante un boxplot, dado que el algoritmo es iterativo. Se representa por tanto la distribución de los valores de importancia de cada atributos obtenido durante los diferentes ejecuciones del algoritmo.



5) Resolución del problema:
En este estudio se ha llevado a cabo un analisis simplificado del dataset 'Arrithmia' para determinar cuales son los atributos con mas relevancia para la prediccion del diagnostico.
Tras adecuar, homogeneizar e estandardizar los datos de entrada se han aplicado diferentes test de hypothesis que han permitido conocer la naturaleza de la distribución de los atributos según el grupo de diagnostico.
En primer lugar los resultados del test de Shapiro han indicado que la mayoria de atributos no tienen una distribución normal (diferenciando los valores según su grupo de diagnostico).
En segundo lugar la aplicación de diferentes tests de hypothesis (t-test, Wilcoxon rank test y test de Kolmogorov-Smirov) ha indicado cual de los atributos tiene una distribucion diferentes para los dos grupos de diagnostico. Estos metodos han señalado que los atributos 'QRS duration' (X6) y ' T-interval' (X9) tienen distribuciones distintas para ambos grupos de diagnostico y que podrian ser candidatos para distinguir entre ambos grupos. El test de Kolmogorov-Smirov ademas ha encontrado los atributos 'T' (X11), 'QRST' (X14) y la 'frecuencia cardiaca' (X15) como variables que provienen de dos poblaciones independientes y que podrian ser de interes para distinguir los dos grupos.
A continuación se han aplicado diferentes metodos de exploración visual de los datos sobre las variables de interes señaladas en los tests de hipothesis, es decir los variables X6, X9, X11, X14 y X15.
La representación grafica mediante scatterplot de estos 5 variables de interes no ha revelado que exista la posibilidad de distinguir ambos grupos del diagnostico utilizando una de las posibles combinaciones entre los atributos (utilizando combinaciones de dos atributos). En

el scatterplot aparecen los puntos del grupo sano mezclados y superpuesto con el grupo 'arrithmia'.
La representacion grafica de los atributos mediante boxplot según el grupo de diagnostico ha sido util para conocer la distribución de los atributos de interes. Los boxplots han mostrado que para la clase de diagnostico 'arrithmia' existen mas instancias con valor elevado (posibles outliers) para los atributos 'QRS duration' (X6) y ' P interval'(X9) que para el grupo con diagnostico 'sano'.
A continuación se han aplicado diferentes metodos de selección de atributos que proporcionan una valoración de la importancia de cada atributo. Estos metodos se han aplicado sobre el conjunto de todos los atributos (X1-X15). Curiosamente los tres metodos utilizados han señalado los mismos atributos como mas relevantes para la prediccíón de la variable de diagnostico:
El metodo MARS ha establecido el ranking de atributos según su importancia en la predicción del diagnostico de X15, X6, X9 ,X12, X14 y X10 (en orden decreciente).
El metodo Random Forest tambien ha señalado los X6, X9, X14 y X15 como los atributos de mas importancia. El metodo 'Boruta' ha indicado el siguiente ranking de atributos según su importancia: X15, X9, X6, X14 y X11.
Podemos concluir que los resultados encontrados por los tests de hipothesis y la de la selección de atributos son coherentes. Los tests de hipothesis han analizado que atributos tienen distribuciones diferenciados para cada grupo y han señalado los atributos X6, X9, X11, X14 y X15 como distribuciones de poblaciones distintas según su grupo de diagnostico. Los metodos de selección de atributo han confirmado la importancia de estos atributos al haber sido seleccionado los atributos X6, X9, X15 y X14 por los tres algoritmos de selección de atributos. El atributo X11 ha sido hallado por dos de los tres algoritmos de selección de atributos, por tanto tambien puede ser considerado como atributo de importancia.

6) Codigo en R: El código en R esta descrito en el fichero R_code.txt del proyecto.