

Exploración de los datos y Preprocesamiento

Roberto Saborit Roig

22/3/2021

Contents

Dataset	1
Exploración de los datos	2
Tipos de variables	2
Datos y valores ausentes	3
Distribución de la variable respuesta	4
Distribución de las variables predictoras	6
Random forest	16
Preprocesamiento	17
Tratamiento de los valores ausentes	18
Variables con varianza cercana 0	19
Normalización	20

Dataset

Lo primero es ver que dataset vamos a utilizar, disponemos de dos conjuntos de datos para realizar el estudio, uno es una cohorte de datos longitudinal, mientras que el otro que tenemos es uno seccional. Estos conjuntos pertenecen al proyecto OASIS, que es un proyecto que pretende poner a libre disposición datos de estudios realizados en MRI. Iremos realizando el preprocesamiento en ambos conjuntos de manera paralela. Y dispondremos de los dos conjuntos para generar el modelo, probablemente utilizaremos uno como conjunto de entrenamiento y otro para el test. Pero eso lo valoraremos más adelante en función de los resultados que obtengamos en la exploración de los datos y el preprocesamiento.

```
#Lo primero será cargar los datos y guardarlos en ls dataframes oasis_longitudinal y oasis_cross_seccional
oasis_cross_sectional <- read.csv("oasis_cross-sectional.csv")
oasis_longitudinal <- readxl::read_excel("oasis_longitudinal_demographics.xlsx")
```

Exploración de los datos

Antes de comenzar a generar el modelo, incluso antes del preprocesamiento, vamos a realizar una exploración de los datos con los siguientes objetivos:

- Ver si existen valores ausentes en el conjunto de datos y ver su distribución entre las distintas variables.
- Explorar los tipos de variable y ver si necesitamos cambiar la clase de alguna variable.
- Ver la distribución de las variables, tanto de la respuesta como de las variables descriptivas.

Tipos de variables

La primera comprobación que haremos será ver los tipos de variables que hay y si todas tienen el tipo de valor que le corresponde:

```
#Exporamos las variables, sus tipos y vemos si es necesario cambiar algún tipo para posteriores análisis.  
str(oasis_cross_sectional)
```

```
## 'data.frame':   436 obs. of  12 variables:  
## $ ID : chr "OAS1_0001_MR1" "OAS1_0002_MR1" "OAS1_0003_MR1" "OAS1_0004_MR1" ...  
## $ M.F : chr "F" "F" "F" "M" ...  
## $ Hand : chr "R" "R" "R" "R" ...  
## $ Age : int 74 55 73 28 18 24 21 20 74 52 ...  
## $ Educ : int 2 4 4 NA NA NA NA NA 5 3 ...  
## $ SES : int 3 1 3 NA NA NA NA NA 2 2 ...  
## $ MMSE : int 29 29 27 NA NA NA NA NA 30 30 ...  
## $ CDR : num 0 0 0.5 NA NA NA NA NA 0 0 ...  
## $ eTIV : int 1344 1147 1454 1588 1737 1131 1516 1505 1636 1321 ...  
## $ nWBV : num 0.743 0.81 0.708 0.803 0.848 0.862 0.83 0.843 0.689 0.827 ...  
## $ ASF : num 1.31 1.53 1.21 1.1 1.01 ...  
## $ Delay: chr "N/A" "N/A" "N/A" "N/A" ...
```

```
str(oasis_longitudinal)
```

```
## tibble [373 x 15] (S3: tbl_df/tbl/data.frame)  
## $ Subject ID: chr [1:373] "OAS2_0001" "OAS2_0001" "OAS2_0002" "OAS2_0002" ...  
## $ MRI ID : chr [1:373] "OAS2_0001_MR1" "OAS2_0001_MR2" "OAS2_0002_MR1" "OAS2_0002_MR2" ...  
## $ Group : chr [1:373] "Nondemented" "Nondemented" "Demented" "Demented" ...  
## $ Visit : chr [1:373] "1" "2" "1" "2" ...  
## $ MR Delay : num [1:373] 0 457 0 560 1895 ...  
## $ M/F : chr [1:373] "M" "M" "M" "M" ...  
## $ Hand : chr [1:373] "R" "R" "R" "R" ...  
## $ Age : num [1:373] 87 88 75 76 80 88 90 80 83 85 ...  
## $ EDUC : num [1:373] 14 14 12 12 12 18 18 12 12 12 ...  
## $ SES : num [1:373] 2 2 NA NA NA 3 3 4 4 4 ...  
## $ MMSE : num [1:373] 27 30 23 28 22 28 27 28 29 30 ...  
## $ CDR : num [1:373] 0 0 0.5 0.5 0.5 0 0 0 0.5 0 ...  
## $ eTIV : num [1:373] 1987 2004 1678 1738 1698 ...  
## $ nWBV : num [1:373] 0.696 0.681 0.736 0.713 0.701 ...  
## $ ASF : num [1:373] 0.883 0.876 1.046 1.01 1.034 ...
```

Lo primero que tenemos que tener en cuenta en ambos conjuntos es que el conjunto longitudinal esta etiquetado por grupos en la variable `Group`, estos son “demented”, “non demented” y “converted”.Hacen

referencia a sujetos con demencia, sin demencia y que desarrollaron demencia a lo largo del experimento, repectivamente. En cambio el estudio seccional no tiene estas etiquetas, pero realmente la variable CDR (Clasificación Clínica de Demencia), hace referencia a lo mismo, es decir, al nivel de demencia que tienen los pacientes según este test, que va desde 0 (no tiene demencia), 0.5 (demencia muy leve), 1 (demencia leve) o 2 (demencia severa). Podemos abordar el problema simplemente creando la variable `Group` y etiquetando como demented aquellos que tienen un nivel de demencia mayor de 0.

Además vemos que tenemos 2 variables más en el estudio longitudinal que son `visit` que hace referencia al número de visita de esa observación y `MR delay`, que es el tiempo que ha pasado entre visita y visita.

Por otro lado la variable `Hand` no aporta ninguna información ya que todos los sujetos son diestros, por tanto, la eliminaremos del conjunto:

```
#La función levels muestra los niveles de la variable $Hand  
levels(as.factor(oasis_cross_sectional$Hand))
```

```
## [1] "R"
```

```
levels(as.factor(oasis_longitudinal$Hand))
```

```
## [1] "R"
```

Como vemos solo existe un nivel en el factor que es “R” (Right/Diestro), por lo que no nos aportará ningún valor al modelo, pero si hay que tener en cuenta que el modelo lo habremos realizado solo en personas diestras y si esta variable tuviera importancia en la aparición de demencia los resultados podrían no ser tan precisos en personas zurdas.

Datos y valores ausentes

Vamos a comprobar ahora el número de datos que tenemos y la cantidad de valores ausentes que hay:

```
#El número total de filas nos indica la cantidad de observaciones de cada uno de los datasets  
nrow(oasis_cross_sectional); ncol(oasis_cross_sectional)
```

```
## [1] 436
```

```
## [1] 12
```

```
nrow(oasis_longitudinal); ncol(oasis_longitudinal)
```

```
## [1] 373
```

```
## [1] 15
```

```
any(is.na(oasis_cross_sectional)); any(is.na(oasis_longitudinal))
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
#Tenemos datos ausentes en ambos conjuntos
#Podemos comprobar que variables tienen mayor porcentaje de NA y la cantidad total
apply(is.na(oasis_cross_sectional), 2, sum)
```

```
##      ID      M.F      Hand      Age      Educ      SES      MMSE      CDR      eTIV      nWBV      ASF      Delay
##      0        0        0        0      201      220      201      201        0        0        0        0
```

```
apply(is.na(oasis_longitudinal), 2, sum)
```

```
## Subject ID      MRI ID      Group      Visit      MR Delay      M/F      Hand
##          0          0          0          0          0          0          0
##      Age      EDUC      SES      MMSE      CDR      eTIV      nWBV
##          0          0          19          2          0          0          0
##      ASF
##          0
```

En el caso del estudio seccional tenemos una gran cantidad de datos ausentes en las variables Educ, SES, MMSE, CDR, que suponen casi un 50% de los datos, en esas variables, lo que puede suponer un problema si durante el preprocesamiento optamos por eliminar las observaciones con datos ausentes, ya que prederíamos una gran cantidad de información.

En cambio en el conjunto de datos longitudinal no tenemos apenas datos ausentes, solo unos pocos en la variable SES, que es el estatus socioeconómico, y solo 2 en MMSE (test de deterioro cognitivo).

Distribución de la variable respuesta

Otra información importante que debemos conocer antes de comenzar con el modelo es la distribución de la variable respuesta, es decir la cantidad de observaciones que hay según los grupos que tenemos en la variable respuesta, en nuestro caso nos interesa conocer como está distribuida la variable **Group** en el estudio longitudinal y la variable **CDR** en el estudio longitudinal. Para ello vamos a generar una tabla con los datos y vamos a verlo también gráficamente:

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.4
```

```
par(mfrow=c(1,2))
```

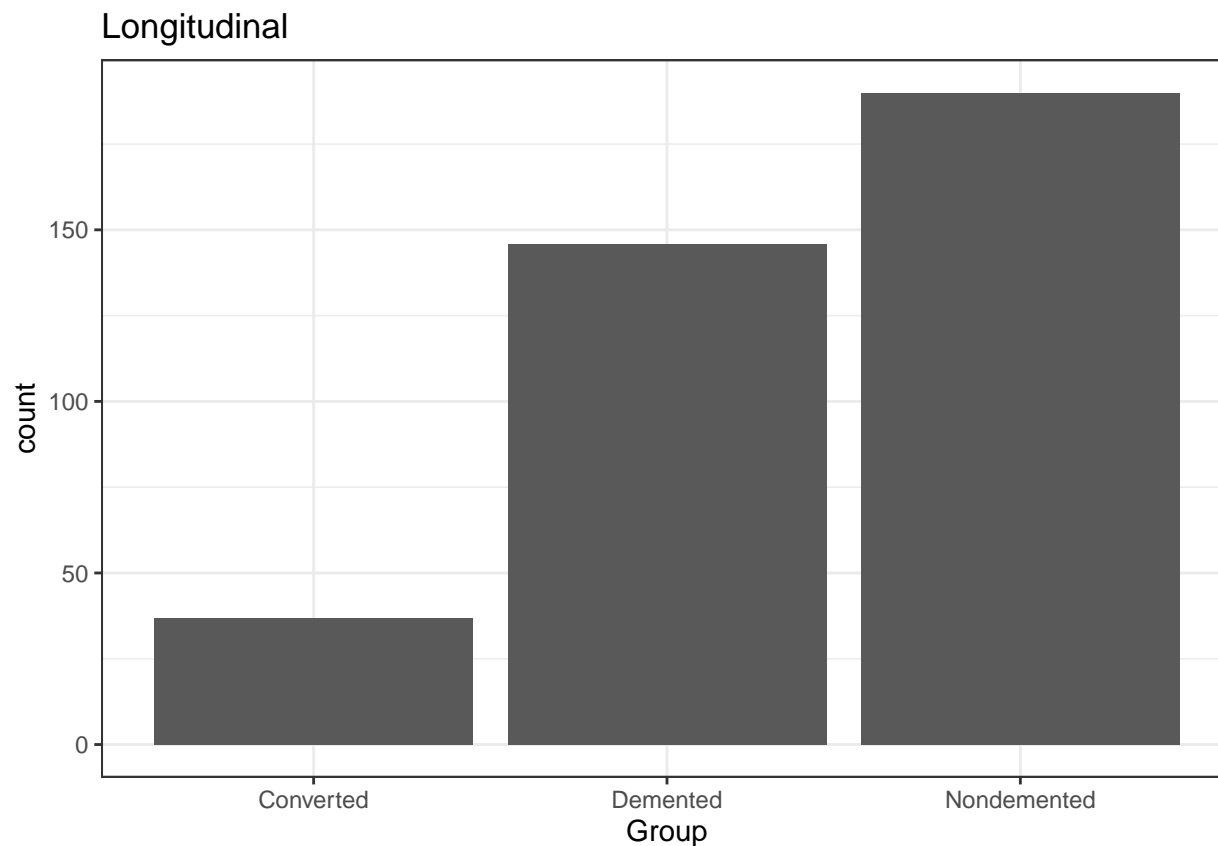
```
#Generamos las tablas, con datos absolutos y relativos
table(oasis_longitudinal$Group)
```

```
##
##      Converted      Demented      Nondemented
##           37          146          190
```

```
round(prop.table(table(oasis_longitudinal$Group)), 2)
```

```
##
##      Converted      Demented      Nondemented
##          0.10          0.39          0.51
```

```
#Y generamos el gráfico
ggplot(data = oasis_longitudinal, aes(x = Group, y = ..count.., )) +
  geom_bar() +
  labs(title = "Longitudinal") +
  theme_bw() +
  theme(legend.position = "bottom")
```



```
#Hacemos lo mismo con el estudio seccional
table(oasis_cross_sectional$CDR)
```

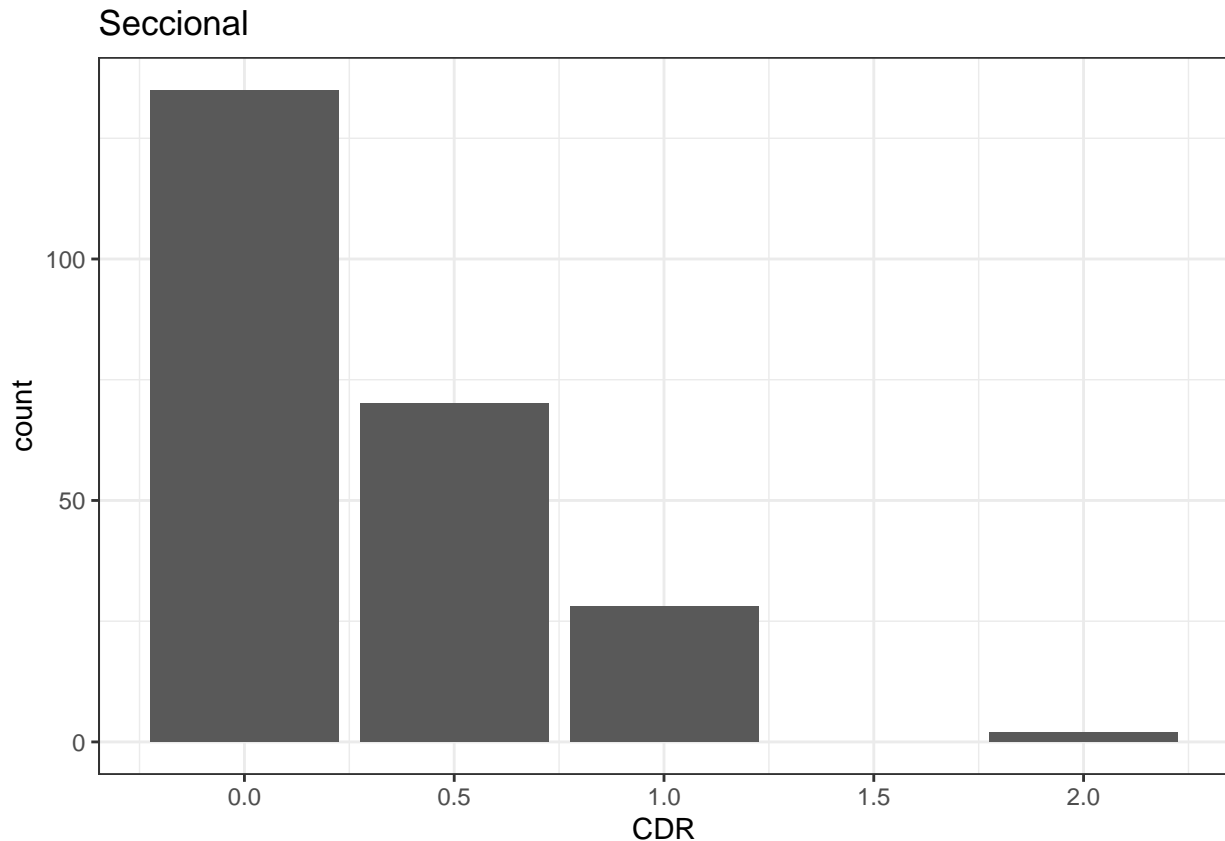
```
##
##  0 0.5  1  2
## 135 70 28  2
```

```
round(prop.table(table(oasis_cross_sectional$CDR)), 2)
```

```
##
##  0 0.5  1  2
## 0.57 0.30 0.12 0.01
```

```
ggplot(data = oasis_cross_sectional, aes(x = CDR, y = ..count.., )) +
  geom_bar() +
  labs(title = "Seccional") +
  theme_bw() +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 201 rows containing non-finite values (stat_count).
```



Como vemos el porcentaje de converted que son aquellos que al principio del experimento no tenían demencia y en las sucesivas medidas la desarrollaron, es del 10% de los datos, estos es importante conocerlo, si desamos crear un modelo efcitivo es importante que acierte más del 10% de converted, que podría acertarse si simplemente calsificamos todos los sujetos como converted.

En el caso del estudio seccional lo hemos dividido en grupos según la variable CDR que muestra si se no se tiene demencia, y también vemos una distribución desigual, donde los pacientes con demencia son menos que los que no tienen demencia, sobre todo aquellos que tienen demencia leve y pacientes con demencia moderada apenas hay.

Distribución de las variables predictoras

Tras ver la distribución de la variable respuesta, nos interesa conocer la de las variables predictoras

Variables continuas

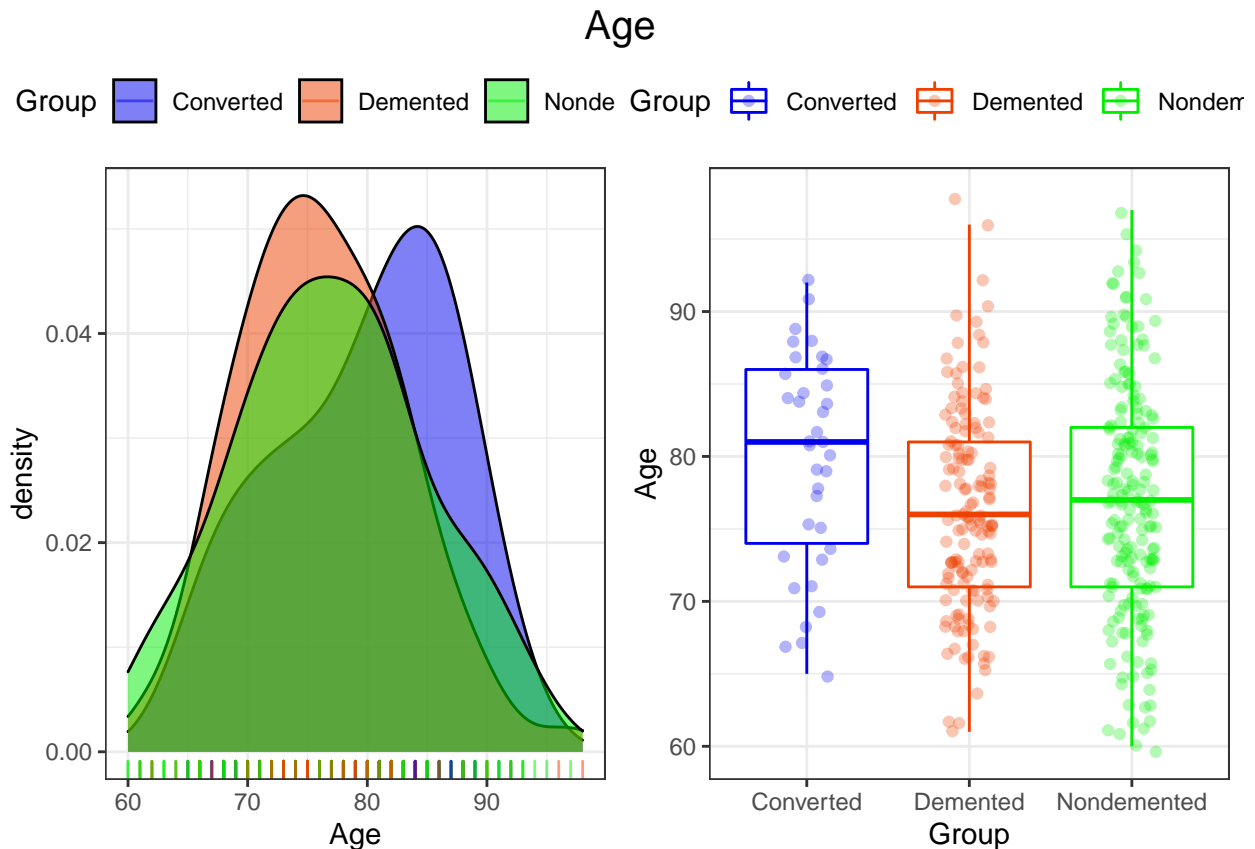
```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.0.3
```

```

#Creamos un gráfico para ver la distribución de la variable Age
p1 <- ggplot(data = oasis_longitudinal, aes(x = Age, fill = Group)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("blue2", "orangered2", "green2")) +
  geom_rug(aes(color = Group), alpha = 0.5) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
p2 <- ggplot(data = oasis_longitudinal, aes(x = Group, y = Age, color = Group)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
final_plot <- ggarrange(p1, p2, legend = "top")
final_plot <- annotate_figure(final_plot, top = text_grob("Age", size = 15))
final_plot

```



```

#Calculamos la media y la mediana
tapply(oasis_longitudinal$Age, oasis_longitudinal$Group, mean)

```

```

##   Converted   Demented Nondemented
##   79.75676   76.26027   77.05789

```

```

tapply(oasis_longitudinal$Age, oasis_longitudinal$Group, median)

```

```

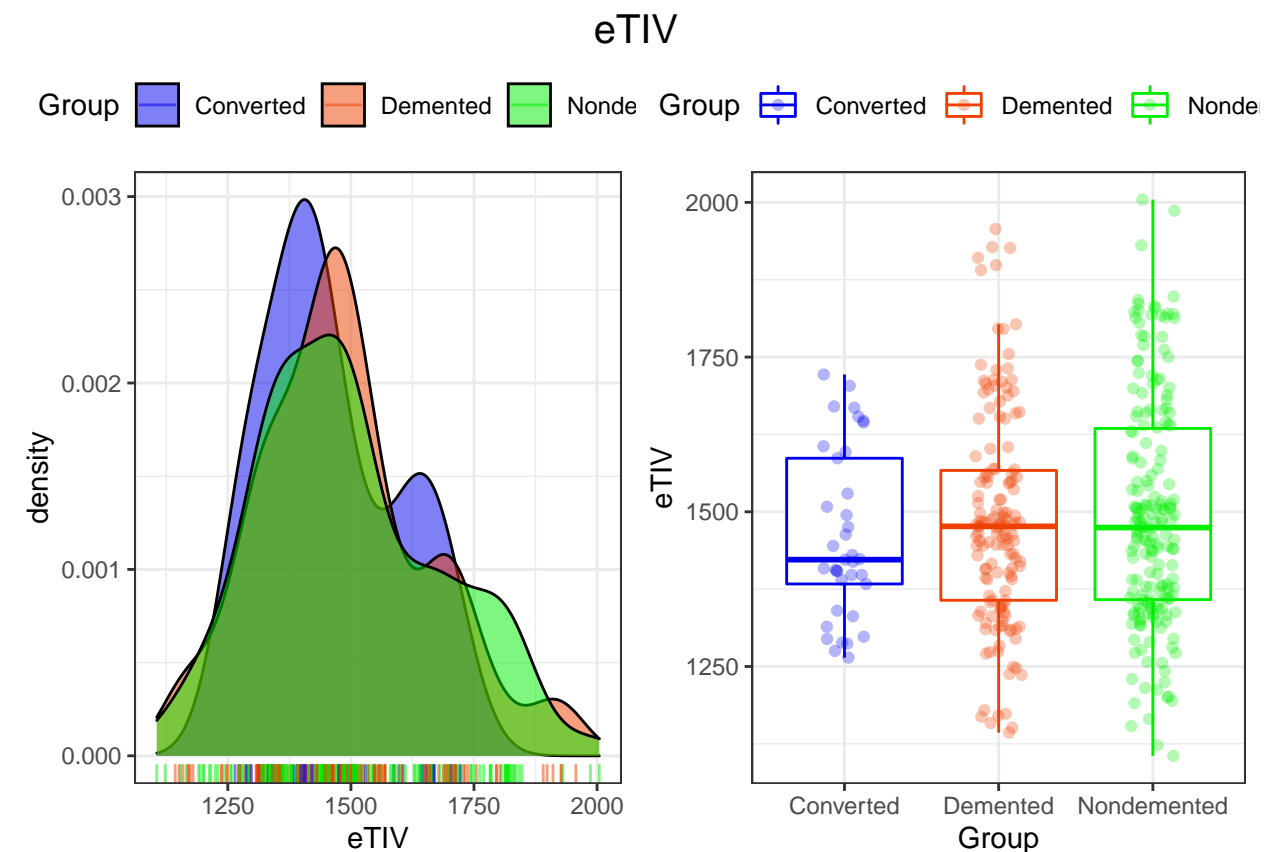
##   Converted   Demented Nondemented

```

```
##           81           76           77
```

Como se ve el grupo converted tiene una edad media significativamente más baja que las otras dos variables, al igual que pasa con la mediana.

```
#Volumen intracraneal
p1 <- ggplot(data = oasis_longitudinal, aes(x = eTIV, fill = Group)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("blue2", "orangered2", "green2")) +
  geom_rug(aes(color = Group), alpha = 0.5) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
p2 <- ggplot(data = oasis_longitudinal, aes(x = Group, y = eTIV, color = Group)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
final_plot <- ggarrange(p1, p2, legend = "top")
final_plot <- annotate_figure(final_plot, top = text_grob("eTIV", size = 15))
final_plot
```




```
tapply(oasis_longitudinal$eTIV, oasis_longitudinal$Group, median)
```





```
##      Converted      Demented      Nondemented
##      1422.623      1476.460      1474.505
```

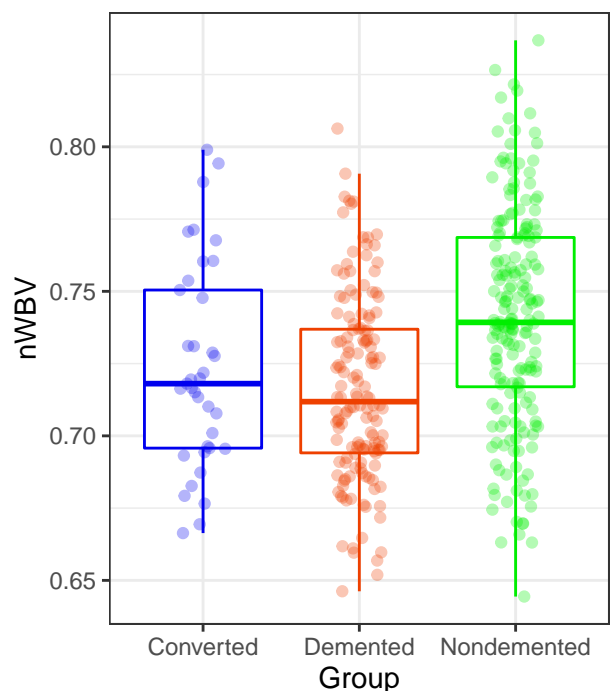
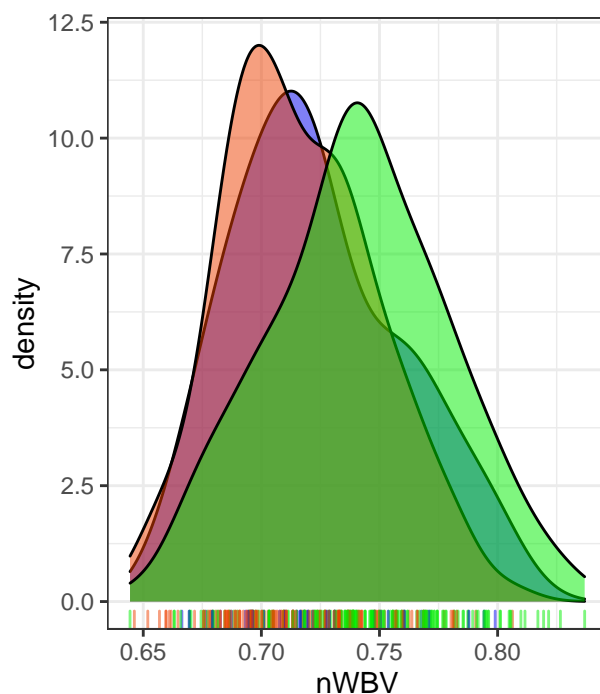
En esta variable el grupo converted tiene una mediana significativamente más baja que las otras dos, en la media en cambio no hay tanta diferencia. La variable Nondemented tiene una media por encima de Demented, en cambio una mediana por debajo.

```
#Volumen total normalizado
```

```
p1 <- ggplot(data = oasis_longitudinal, aes(x = nWBV, fill = Group)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("blue2", "orangered2", "green2")) +
  geom_rug(aes(color = Group), alpha = 0.5) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
p2 <- ggplot(data = oasis_longitudinal, aes(x = Group, y = nWBV, color = Group)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
final_plot <- ggarrange(p1, p2, legend = "top")
final_plot <- annotate_figure(final_plot, top = text_grob("nWBV", size = 15))
final_plot
```

nWBV

Group  Converted  Demented  Nonder Group  Converted  Demented  Nonder



```
tapply(oasis_longitudinal$nWBV, oasis_longitudinal$Group, mean)
```

```
##   Converted   Demented Nondemented  
##   0.7237336   0.7163034   0.7408726
```

```
tapply(oasis_longitudinal$nWBV, oasis_longitudinal$Group, median)
```

```
##   Converted   Demented Nondemented  
##   0.7180650   0.7118355   0.7392630
```

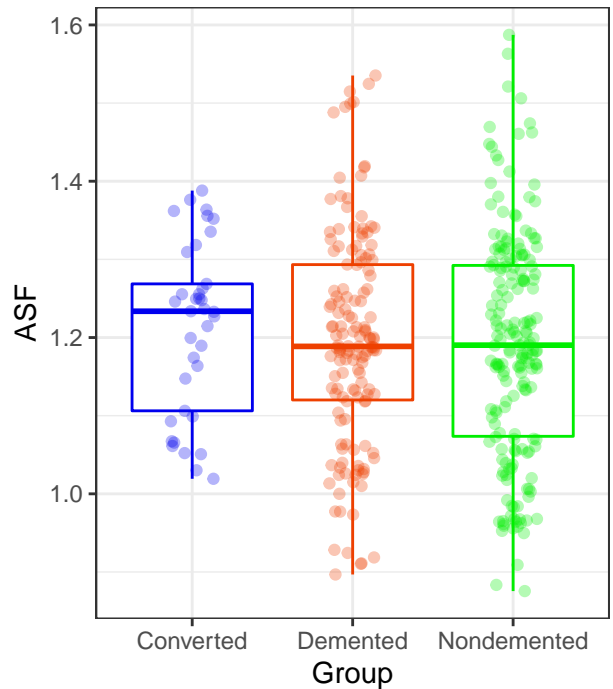
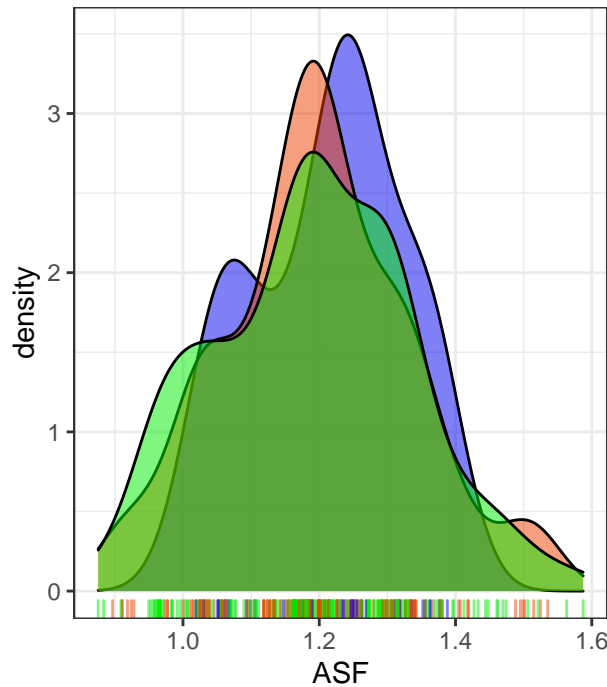
En esta variable sí se ven diferencias entre los grupos, los no dementes tienen valores en promedio más altos, que los dementes y los converted.

```
#Atlas Scaling Factor
```

```
p1 <- ggplot(data = oasis_longitudinal, aes(x = ASF, fill = Group)) +  
  geom_density(alpha = 0.5) +  
  scale_fill_manual(values = c("blue2", "orangered2", "green2")) +  
  geom_rug(aes(color = Group), alpha = 0.5) +  
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +  
  theme_bw()  
p2 <- ggplot(data = oasis_longitudinal, aes(x = Group, y = ASF, color = Group)) +  
  geom_boxplot(outlier.shape = NA) +  
  geom_jitter(alpha = 0.3, width = 0.15) +  
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +  
  theme_bw()  
final_plot <- ggarrange(p1, p2, legend = "top")  
final_plot <- annotate_figure(final_plot, top = text_grob("ASF", size = 15))  
final_plot
```

ASF

Group ■ Converted ■ Demented ■ Nondem Group ▢ Converted ▢ Demented ▢ Nonden



```
tapply(oasis_longitudinal$ASF, oasis_longitudinal$Group, mean)
```

```
##   Converted   Demented Nondemented
##   1.212422   1.196880   1.191066
```

```
tapply(oasis_longitudinal$ASF, oasis_longitudinal$Group, median)
```

```
##   Converted   Demented Nondemented
##   1.233637   1.188655   1.190225
```

En este caso no parece haber diferencias en la distribución.

```
#MMSE (Test de deterioro cognitivo)
```

```
p1 <- ggplot(data = oasis_longitudinal, aes(x = MMSE, fill = Group)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("blue2", "orangered2", "green2")) +
  geom_rug(aes(color = Group), alpha = 0.5) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
p2 <- ggplot(data = oasis_longitudinal, aes(x = Group, y = MMSE, color = Group)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
```

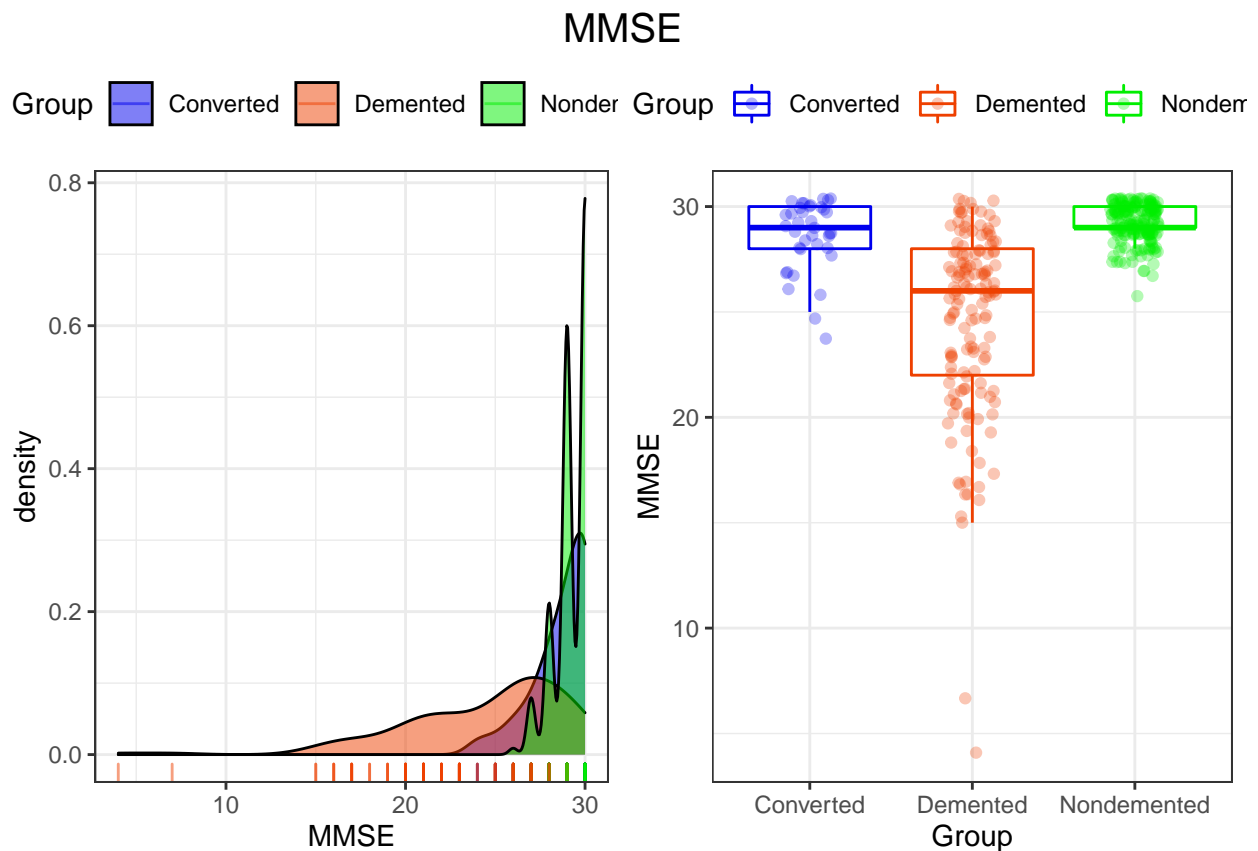
```
theme_bw()
final_plot <- ggarrange(p1, p2, legend = "top")
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
final_plot <- annotate_figure(final_plot, top = text_grob("MMSE", size = 15))
final_plot
```



```
tapply(oasis_longitudinal$MMSE, oasis_longitudinal$Group, na.rm= TRUE, mean)
```

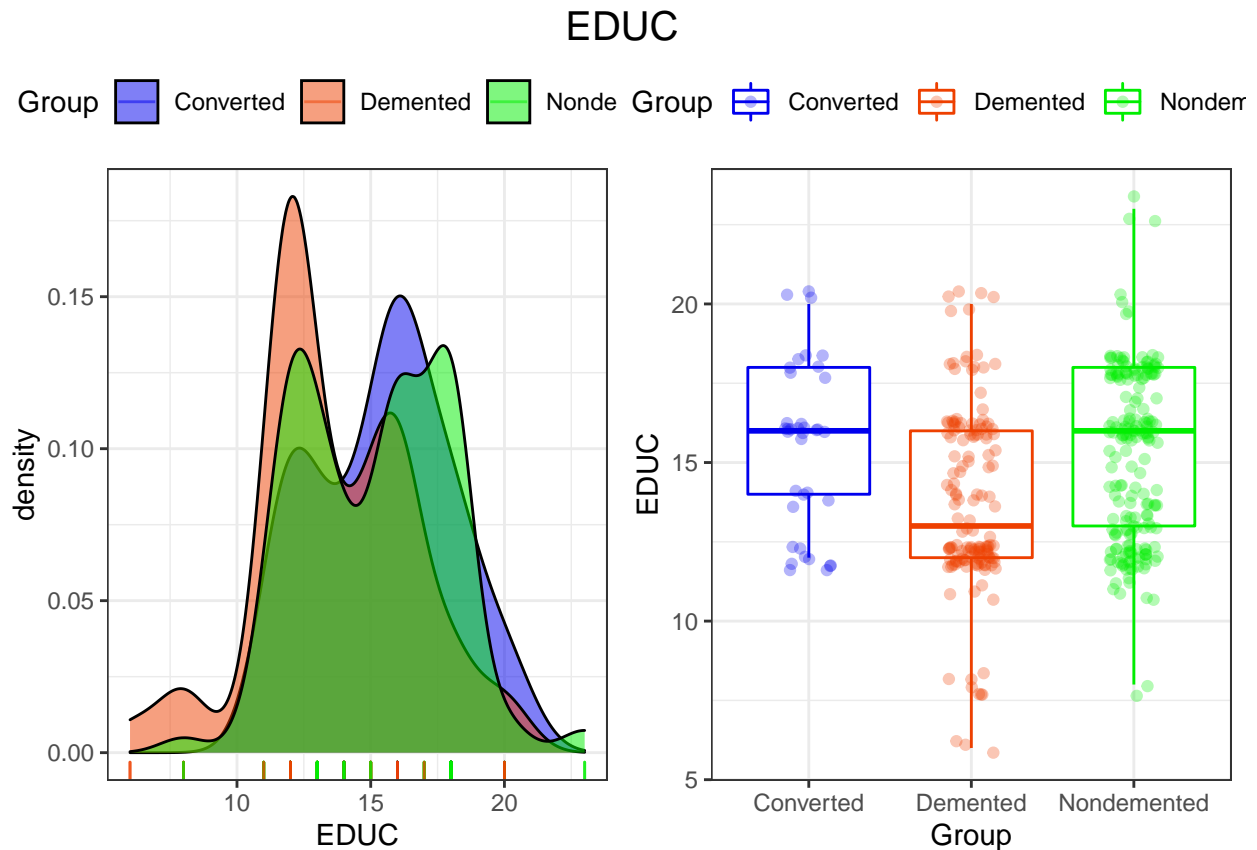
```
##   Converted   Demented Nondemented
##   28.67568   24.51389   29.22632
```

```
tapply(oasis_longitudinal$MMSE, oasis_longitudinal$Group, na.rm= TRUE, median)
```

```
##   Converted   Demented Nondemented
##         29         26         29
```

En este caso se ven diferencias entre los que tienen demencia y los otros dos grupos, pero no entre converted y nondemented.

```
#Nivel de estudios
p1 <- ggplot(data = oasis_longitudinal, aes(x = EDUC, fill = Group)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("blue2", "orangered2", "green2")) +
  geom_rug(aes(color = Group), alpha = 0.5) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
p2 <- ggplot(data = oasis_longitudinal, aes(x = Group, y = EDUC, color = Group)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15) +
  scale_color_manual(values = c("blue2", "orangered2", "green2")) +
  theme_bw()
final_plot <- ggarrange(p1, p2, legend = "top")
final_plot <- annotate_figure(final_plot, top = text_grob("EDUC", size = 15))
final_plot
```



```
tapply(oasis_longitudinal$EDUC, oasis_longitudinal$Group, na.rm= TRUE, mean)
```

```
##   Converted   Demented Nondemented
##   15.45946   13.67123   15.14211
```

```
tapply(oasis_longitudinal$EDUC, oasis_longitudinal$Group, na.rm= TRUE, median)
```

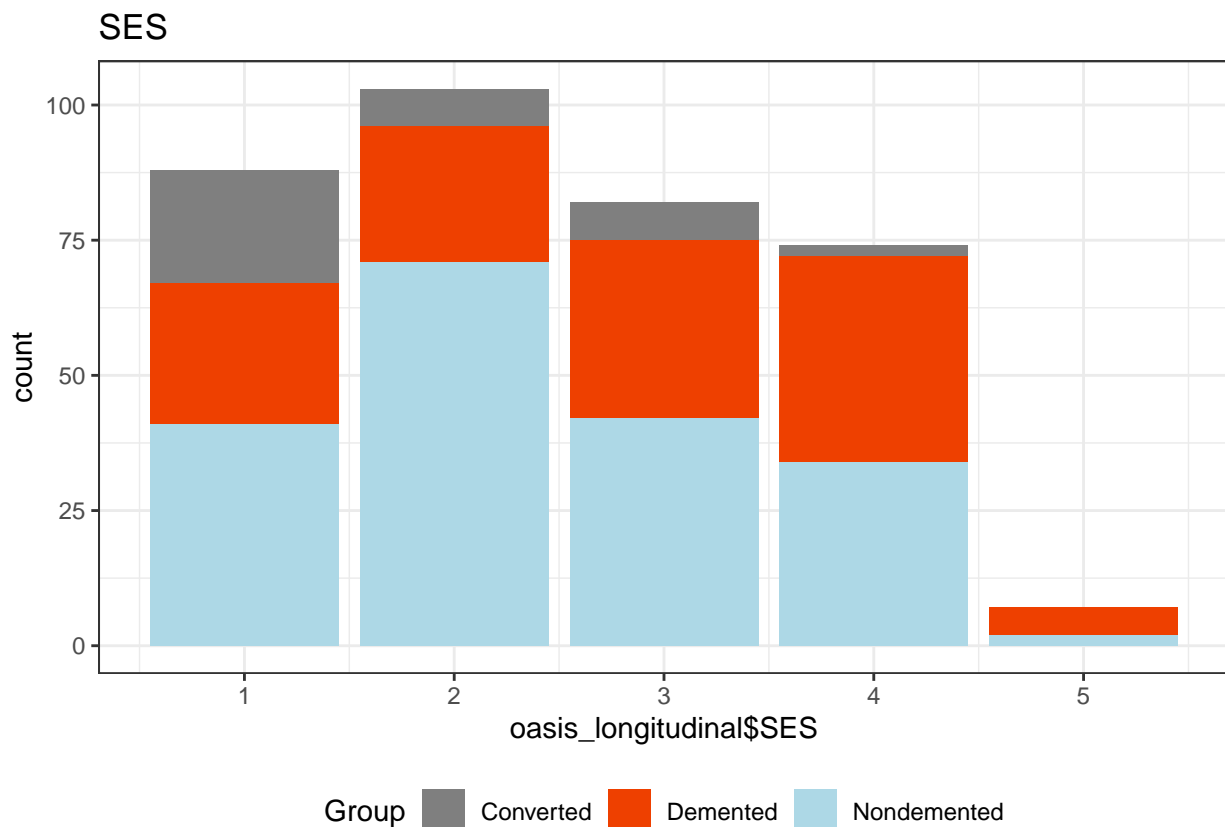
```
##      Converted      Demented Nondemented
##           16           13           16
```

En esta variable tambien se ven diferencias entre los sujetos con demencia y los otros dos grupos.

```
#Status socioeconómico
ggplot(data = oasis_longitudinal, aes(x = oasis_longitudinal$SES, y = ..count.., fill = Group)) +
  geom_bar() +
  labs(title = "SES") +
  scale_fill_manual(values = c("gray50", "orangered2", "lightblue")) +
  theme_bw() +
  theme(legend.position = "bottom")
```

```
## Warning: Use of 'oasis_longitudinal$SES' is discouraged. Use 'SES' instead.
```

```
## Warning: Removed 19 rows containing non-finite values (stat_count).
```



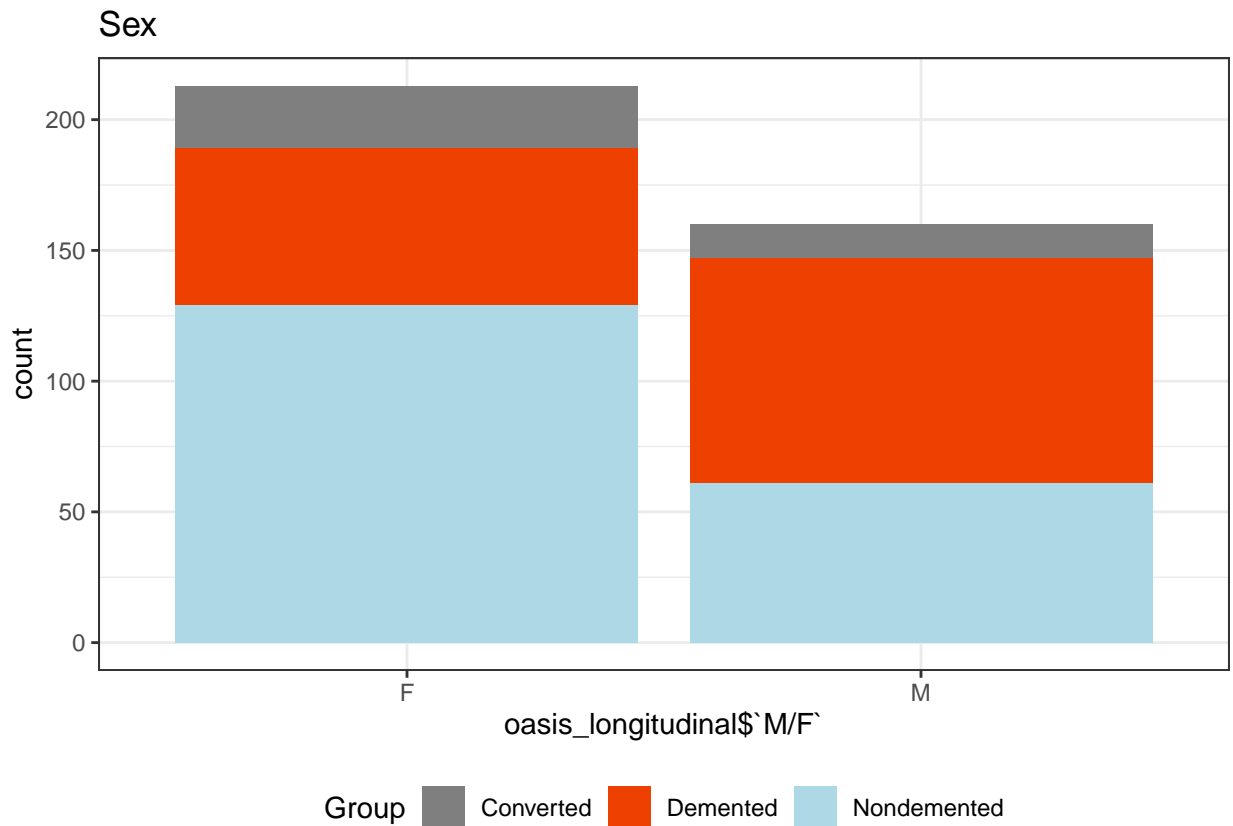
```
round(prop.table(table(oasis_longitudinal$SES, oasis_longitudinal$Group)),2)
```

```
##
##      Converted Demented Nondemented
```

```
## 1      0.06      0.07      0.12
## 2      0.02      0.07      0.20
## 3      0.02      0.09      0.12
## 4      0.01      0.11      0.10
## 5      0.00      0.01      0.01
```

```
#Sexo
ggplot(data = oasis_longitudinal, aes(x = oasis_longitudinal$`M/F`, y = ..count.., fill = Group)) +
  geom_bar() +
  labs(title = "Sex") +
  scale_fill_manual(values = c("gray50", "orangered2", "lightblue")) +
  theme_bw() +
  theme(legend.position = "bottom")
```

```
## Warning: Use of 'oasis_longitudinal$`M/F`' is discouraged. Use 'M/F' instead.
```



```
round(prop.table(table(oasis_longitudinal$`M/F`, oasis_longitudinal$Group)),2)
```

```
##
##      Converted Demented Nondemented
## F      0.06      0.16      0.35
## M      0.03      0.23      0.16
```

Random forest

Para terminar el apartado del análisis exploratorio, y complementar el último punto, vamos a realizar un análisis random forest con el que descubriremos que variables predicen mejor la variable respuesta:

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.0.4
```

```
## randomForest 4.6-14
```

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

```
##
```

```
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
```

```
##
```

```
##      margin
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## -- Attaching packages ----- tidyverse 1.3.0 --
```

```
## v tibble  3.0.3    v dplyr    1.0.2
```

```
## v tidyr   1.1.2    v stringr 1.4.0
```

```
## v readr   1.3.1    v forcats 0.5.0
```

```
## v purrr   0.3.4
```

```
## Warning: package 'tidyr' was built under R version 4.0.3
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::combine()      masks randomForest::combine()
```

```
## x dplyr::filter()       masks stats::filter()
```

```
## x dplyr::lag()          masks stats::lag()
```

```
## x randomForest::margin() masks ggplot2::margin()
```

```
datos_rf <- oasis_longitudinal %>%
```

```
  select(-'Subject ID', -'MRI ID', -'MR Delay', -Visit, -Hand) %>% na.omit()
```

```
datos_rf <- map_if(.x = datos_rf, .p = is.character, .f = as.factor) %>% as.data.frame()
```

```
modelo_randforest <- randomForest(formula = Group ~ . ,
```

```
  data = na.omit(datos_rf),
```

```
  mtry = 5,
```

```
  importance = TRUE,
```

```
  ntree = 1000)
```

```
importancia <- as.data.frame(modelo_randforest$importance)
```



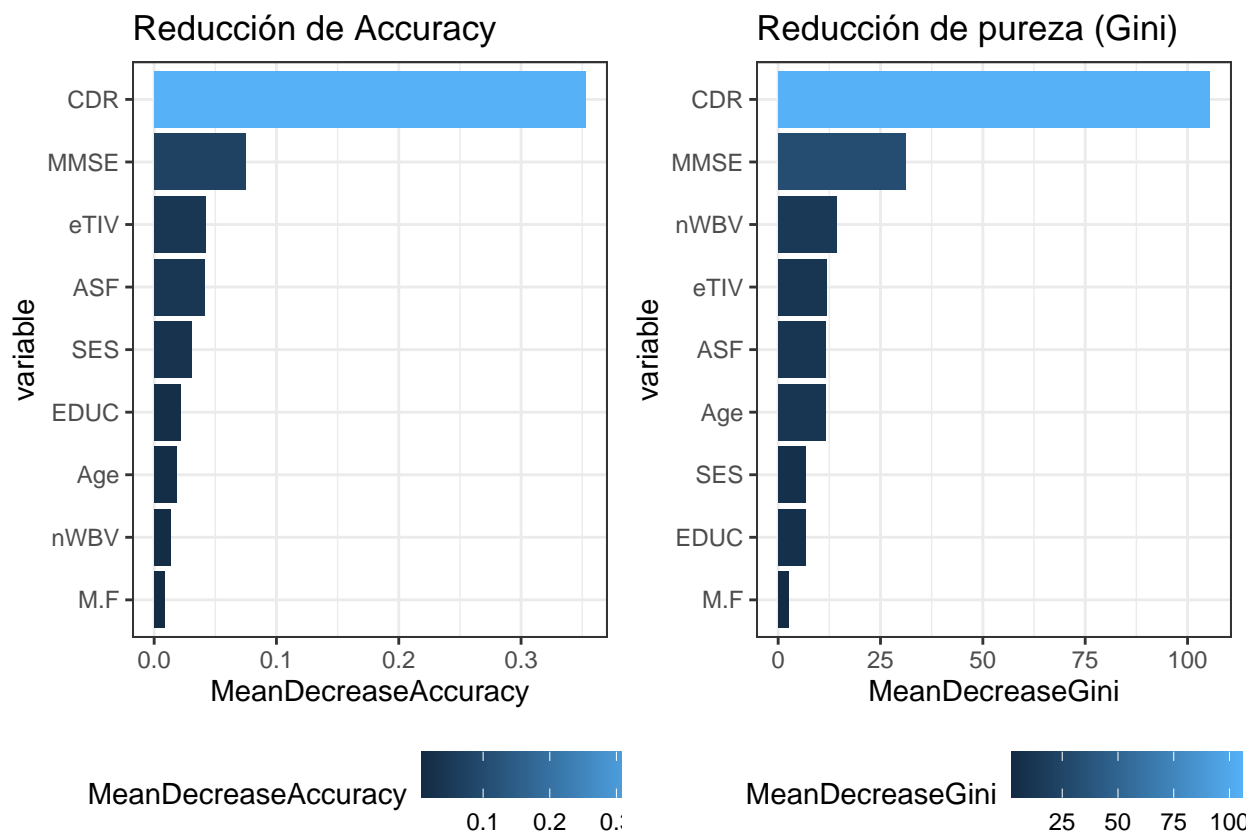
```

importancia <- rownames_to_column(importancia,var = "variable")

p1 <- ggplot(data = importancia, aes(x = reorder(variable, MeanDecreaseAccuracy),
                                     y = MeanDecreaseAccuracy,
                                     fill = MeanDecreaseAccuracy)) +
  labs(x = "variable", title = "Reducción de Accuracy") +
  geom_col() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "bottom")

p2 <- ggplot(data = importancia, aes(x = reorder(variable, MeanDecreaseGini),
                                     y = MeanDecreaseGini,
                                     fill = MeanDecreaseGini)) +
  labs(x = "variable", title = "Reducción de pureza (Gini)") +
  geom_col() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "bottom")
ggarrange(p1, p2)

```



Este análisis apunta a que las mejores variables para predecir la demencia son las CDR y MMSE.

Preprocesamiento

Tratamiento de los valores ausentes

Como vimos en la exploración de los datos hay valores ausentes, principalmente concentrados en las variables SES, EDUC y MMSE. Antes de seguir habría que eliminarlos ya que hay algoritmos de ML que no admiten estos valores. Tenemos dos opciones, eliminar las variables con valores ausentes o eliminar las observaciones con valores ausentes. Otra opción sería realizar una imputación para no perder la información de esas variables u observaciones, imputar sería estimar los valores que faltan por medio de la información que sí tenemos.

```
#Si eliminamos las observaciones se nos quedaría:  
nrow(na.omit(oasis_cross_sectional))
```

```
## [1] 216
```

```
nrow(na.omit(oasis_longitudinal))
```

```
## [1] 354
```

```
oasis_longitudinal_narm=na.omit(oasis_longitudinal)
```

```
#Imputación  
library(recipes)
```

```
## Warning: package 'recipes' was built under R version 4.0.4
```

```
##  
## Attaching package: 'recipes'
```

```
## The following object is masked from 'package:stringr':  
##  
##      fixed
```

```
## The following object is masked from 'package:stats':  
##  
##      step
```

```
objeto_recipe <- recipe(formula = Group ~ Age + 'M/F' + EDUC + SES +  
                        MMSE + CDR + eTIV + nWBV + ASF,  
                        data = oasis_longitudinal)  
objeto_recipe
```

```
## Data Recipe  
##  
## Inputs:  
##  
##      role #variables  
## outcome      1  
## predictor     9
```

```
objeto_recipe <- objeto_recipe %>% step_bagimpute(SES)
objeto_recipe
```

```
## Data Recipe
##
## Inputs:
##
##      role #variables
## outcome      1
## predictor      9
##
## Operations:
##
## Bagged tree imputation for SES
```

La segunda opción para los datos seccionales no sería opción ya que nos cargaríamos 4 variables, en cambio para los longitudinales solo eliminaríamos la variable SES. Pero es verdad que solo habían 16 observaciones con valores ausentes así que en ambas opciones parece mejor opción eliminar las observaciones con valores ausentes.

Variables con varianza cercana 0

Otra parte importante del preprocesado será eliminar variables que no aporten nada, como vimos la variable Hand no se utilizará ya que no tiene diferentes niveles, pero además hay una forma de ver si las variables pueden no aportar información y es viendo si su varianza es igual o cercana a 0. Con la función `nearZeroVars`, podemos averiguar si alguna función tiene varianza cercana a 0.

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.4
```

```
## Loading required package: lattice
```

```
##
```

```
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
```

```
##
```

```
## lift
```

```
oasis_longitudinal %>% select(Age, 'M/F', EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF) %>% nearZeroVar(saveMe
```

```
##      freqRatio percentUnique zeroVar  nzv
## Age      1.181818      10.455764  FALSE FALSE
## M/F      1.331250       0.536193  FALSE FALSE
## EDUC     1.271605       3.217158  FALSE FALSE
## SES      1.170455       1.340483  FALSE FALSE
## MMSE     1.252747       4.825737  FALSE FALSE
## CDR      1.674797       1.072386  FALSE FALSE
## eTIV     1.000000      99.463807  FALSE FALSE
## nWBV     1.000000     100.000000  FALSE FALSE
## ASF      1.000000      99.463807  FALSE FALSE
```

Entre los predictores incluidos en el modelo, no se detecta ninguno con varianza cero o próxima a cero.

Normalización

La normalización es un paso importante para ajustar el modelo, hay diferentes tipos de normalización vamos a implementar dos tipos y comprobaremos cual de ellas se puede ajustar mejor.

```
# Normalización estándar
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

oasis_longitudinal_n <- as.data.frame(lapply(oasis_longitudinal_narm[8:15], normalize))

#Estandarización por puntuación Z
oasis_longitudinal_z <- as.data.frame(scale(oasis_longitudinal_narm[8:15]))
```