Práctica 2: Estudio sobre la diabetes

Asignatura: Tipología y Ciclo de Vida de los Datos (Máster en Ciencia de Datos)

Autores: Diego Alberto López Herrera, Pablo Rivas Castellanos

Mayo de 2021

Contents

---

1 Dependencias

A continuación se cargan la librerías empleadas en el presente trabajo.

```
library(ggplot2)
library(gridExtra)
library(ggcorrplot)
library(ggdendro)
library(GGally)
library(VIM)
library(dplyr)
library(pROC)
```

---

2 Problema de estudio

2.1 Objeto

Se pretende realizar varios análisis que permitan conocer mejor la diabetes. Se plantean varias cuestiones que se abordarán en detalle en la sección de análisis:

- Identificación de varios grupos de personas en base a los distintos parámetros de estudio.
- Evaluación de de la asunción de que la genealogía de la diabetes en el diagnostico de la enfermedad para comprobar si es realmente un factor relevante.
- Generación de un modelo que permita determinar la probabilidad de diagnóstico de la diabetes en base a unos parámetros de entrada.

2.2 Importancia

Para dar respuesta a estas interrogantes, se empleará el dataset *Diabetes Data Set*, disponible en *https://www.kaggle.com/mathchi/diabetes-data-set* (Mehmet 2020). Se considera un conjunto de datos de interés por dos motivos principales: el origen es una fuente altamente contrastada (*National Institute of Diabetes and Digestive and Kidney Diseases*) y se ofrece información detallada sobre múltiples variables recogidas de las personas de estudio (número de embarazos, presión sanguínea distólica, doblez de piel, nivel de insulina, índice de masa corporal, valor de genealogía de diabetes, edad, diabetes diagnosticada). El conjunto de datos está centrado en un grupo de personas que comparten características comunes: mujeres nativas americanas Pima de al menos 21 años de edad.

---

3 Integración y selección de los datos

3.1 Carga de los datos

Se procede a la carga del juego de datos, que se encuentran en el fichero "../res/diabetes.csv," en formato CSV con encabezados y coma (,) como separador de campos.

```
dt <- read.table(
   "../res/diabetes.csv",
   header = TRUE,
   sep = ",",
   stringsAsFactors = TRUE
   )
```

3.2 Atributos

Seguidamente se estudian las variables y su contenido. Para ello en primer lugar se muesta la representación como *String* del juego de datos.

```
str(dt)
```

```
## 'data.frame':    768 obs. of  9 variables:
##  $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : int  148 85 183 89 137 116 78 115 197 125 ...
##  $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : int  35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : int  0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
##  $ Outcome                 : int  1 0 1 0 1 0 1 0 1 1 ...
```

Se observa que el dataset tiene 9 variables, todas de tipo numérico excepto *Outcome*, que ha sido incorrectamente interpretada como variable numérica. A continuación se corrigen los tipos.

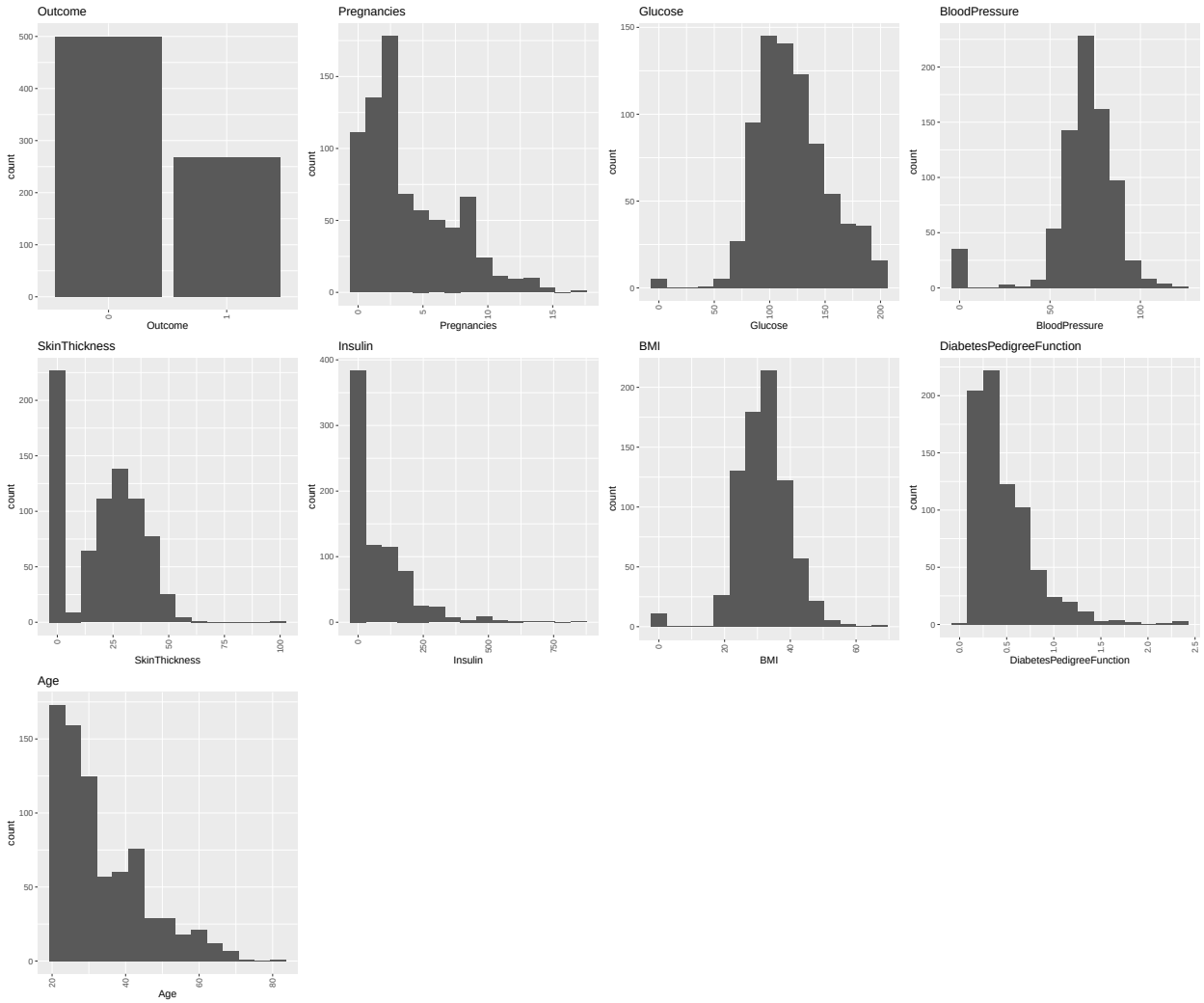
```
dt$Outcome <- as.factor(dt$Outcome)
```

Variable	Tipo	Descripción
Pregnancies	Numérica entera	Número de embarazos.
Glucose	Numérica entera	Niveles de glucosa en plasma del examen de glucosa posprandial de 2 horas ($\mu g/l$).
BloodPressure	Numérica entera	Presión arterial diastólica (mmHg).
SkinThickness	Numérica entera	Espesor del pliegue cutáneo del triceps (mm).
Insulin	Numérica entera	Niveles de insulina del examen de glucosa posprandial de 2 horas ($\mu l\ U/ml$).
BMI	Numérica entera	Índice de masa corporal (kg/m^2).
DiabetesPedigreeFunction	Numérica entera	Índice de predisposición a la diabetes por antecedentes genéticos.
Age	Numérica entera	Edad de la paciente
Outcome	Categórica nominal	Desarrollo de la diabetes.

3.3 Exploración previa

Para obtener una imagen preliminar del conjunto de datos se representa gráficamente su contenido:

```
bar.plots <- list()
cols <- c(names(dt)[sapply(dt, is.numeric)], "risk")
cols <- setdiff(colnames(dt), cols)

for(var in setdiff(colnames(dt), names(dt)[sapply(dt, is.numeric)])){
  p <- ggplot(dt, aes_string(x = var)) +
       theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
       geom_bar() +
       ggtitle(var)
  bar.plots[[var]] <- p
}

for(var in names(dt)[sapply(dt, is.numeric)]){
  p <- ggplot(dt, aes_string(x = var)) +
       theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
       geom_histogram(bins = 15) +
       ggtitle(var)
  bar.plots[[var]] <- p
}


do.call(grid.arrange, c(bar.plots, ncol=4))
```



De la exploración preliminar de las distribuciones se extraen las siguientes conclusiones:

- Existe desbalance en las clases de la variable de *Outcome*.
- Parece observarse la existencia de valores testigo en 0 en las variables *Glucose*, *BlodPressure*, *SkinThickness*, *Insulin* y *BMI*.

---

4 Limpieza de los datos

En este apartado se procesarás los datos con el objetivo de obtener la mejor representación posible de los mismos. Con este fin, se comprobará la presencia de datos faltantes y de *outliers* y se decidirá el tratamiento más apropiado para ambos. Adicionalmente se estudiará la posibilidad de reducción de la dimensionalidad y la correlación entre las distintas variables.

4.1 Datos faltantes

Para estudiar la presencia de registros incompletos, se comprueba en primer lugar que las observaciones pertenezcan al dominio de las respectivas variables:

```r
summary(dt)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##  Outcome
##  0:500
##  1:268
##
##
##
##
```

Como se detectó en la representación preliminar gráfica, existen observaciones con valor 0 en las variables *Glucose*, *BlooPressure*, *SkinThickness*, *Insulin* y *BMI*. Al no pertenecer 0 al dominio de las variables mencionadas, se asume que se trata de un valor testigo empleado para indicar los valores faltantes. Al no disponer de información relativa a su significado, a continuación se sustituirán dichos valores por *NA*, para su tratamiento posterior.

```r
columns <-c ("Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI")
for(column in columns) {
  dt[dt[,column] == 0, column] <- NA
}
summary(dt)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :29.00
##  Mean   : 3.845   Mean   :121.7   Mean   : 72.41   Mean   :29.15
##  3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##                   NA's   :5       NA's   :35       NA's   :227
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0780           Min.   :21.00
##  1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437           1st Qu.:24.00
##  Median :125.00   Median :32.30   Median :0.3725           Median :29.00
##  Mean   :155.55   Mean   :32.46   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##  NA's   :374      NA's   :11
```

```
##  Outcome
##  0:500
##  1:268
##
##
##
##
##
```

A continuación se calcula el número de registros incompletos en el dataset.

```
nrow(dt[is.na(dt),])
```

```
## [1] 652
```

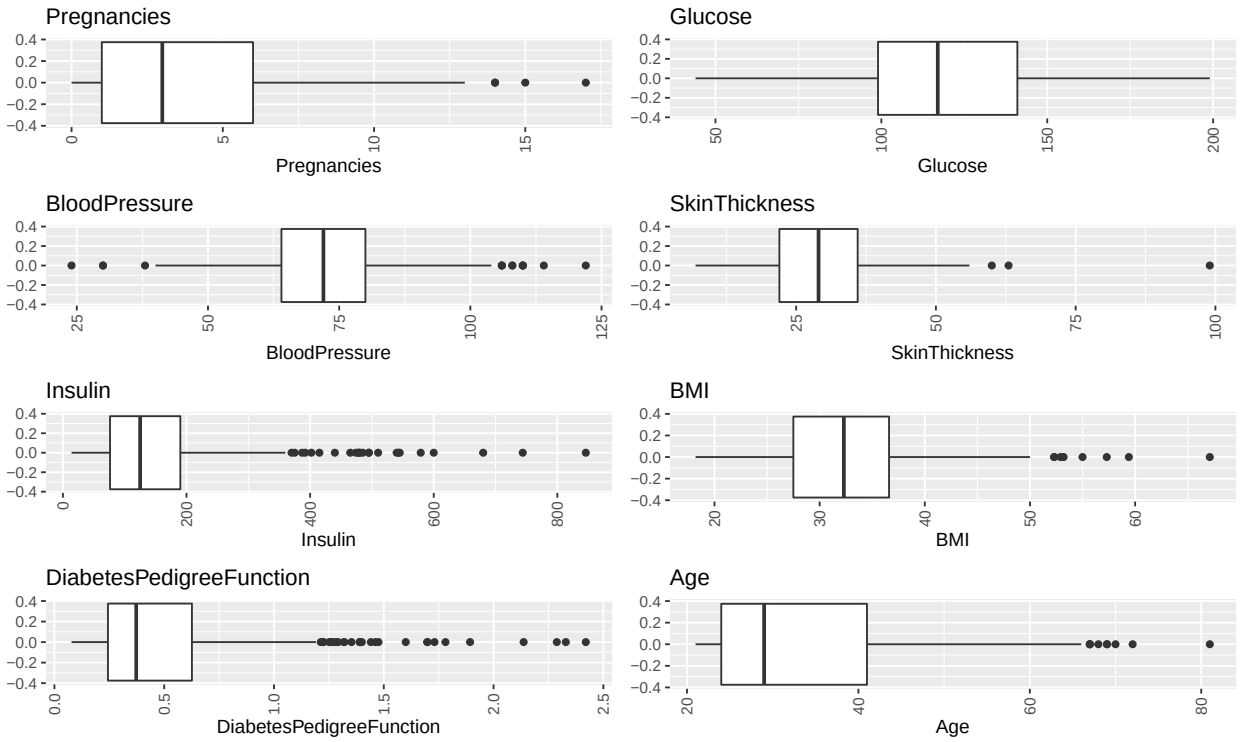
Se observa que hay un total de 425 registros incompletos entre los cuales se encuentran:

- 5 valores faltantes en la variable *Glucose*.
- 35 valores faltantes en la variable *BloodPressure*.
- 374 valores faltantes en la variable *Insulin*.
- 11 valores faltantes en la variable *BMI*.

4.2 Outliers

Para estudiar la presencia de outliers se recurre al empleo de diagramas de caja de las distintas variables numéricas:

```
box.plots <- list()
cols <- names(dt)[sapply(dt, is.numeric)]

for(var in cols){
  p <- ggplot(dt, aes_string(x=var)) +
       theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
       geom_boxplot() +
       ggtitle(var)
  box.plots[[var]] <- p
}

do.call(grid.arrange, c(box.plots, ncol=2))
```



Se observan valores extremos en las variables *Pregnancies*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *DiabetesPedigreeFunction* y *Age*. En este apartado se discute su legitimidad, y, cuando procede, su tratamiento.

1. **Número de embarazos**: Existen valores extremos en 14, 15 y 17 embarazos. Se considera que puede tratarse de observaciones plausibles y no se requiere ninguna acción correctiva.
2. **Presión sanguínea**: Se observan valores entre 24 y 122 mmHg. Atendiendo a los valores de presión sanguínea distólica aceptados, se tiene que los valores normales de personas saludables son inferiores a 80mmHg, con un valor media entre 80 y 40 mmHg ("Hipertensión y Presión Arterial" s.f.). Igualmente, niveles superiores a los 110 mmHg son considerados casos de hipertensión grado 3 ("Hipertensión y Presión Arterial" s.f.). Considerando la cercanía de los valores extremos con los puntos indicados, se considera que las observaciones son válidas y no es necesario ningún tratamiento especial.
3. **Doblez de piel**: Se observa un valor muy alejado en la prueba, con el valor 99. Se considera que podría ser un *outlier* y será será imputado conjuntamente con los valores faltantes.
4. **Nivel de insulina**: Los valores para personas saludables se encuentran entre los 16 y 166 $\mu l\ U/ml$. ("Insulin" 2019). Se considera que los valores observados son correctos al ser menores a 5 veces el valor máximo para personas saludables.
5. **Índice de masa corporal**: Se considera que índices de masa corporal superiores a $60 kg/m^2$ son sospechosos de ser outliers, por lo que se decide imputarlos conjuntamente con los valores faltantes.
6. **DiabetesPedigreeFunction**: Esta variable califica numéricamente la predisposición genética a la diabetes en función de los antecedentes familiares. Al no disponerse de información acerca de su método de cálculo, su dominio o su interpretación, se toma la decisión de considerar todos sus valores correctos y legítimos.
7. **Edad**: Se observa especial acumulación de observaciones entre los 25 y 40 años, estando todas las edades comprendidas entre 20 y 81 años. Se considera que la distribución puede ser correcta y no son necesarias correcciones adicionales.

```
dt$SkinThickness[dt$SkinThickness == 99] <-NA
dt$BMI[dt$BMI > 60] <- NA
summary(dt)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :29.00
##  Mean   : 3.845   Mean   :121.7   Mean   : 72.41   Mean   :29.02
##  3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :63.00
##                   NA's   :5       NA's   :35       NA's   :228
##     Insulin            BMI        DiabetesPedigreeFunction      Age
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0780           Min.   :21.00
##  1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437           1st Qu.:24.00
##  Median :125.00   Median :32.30   Median :0.3725           Median :29.00
##  Mean   :155.55   Mean   :32.41   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.00   Max.   :59.40   Max.   :2.4200           Max.   :81.00
##  NA's   :374      NA's   :12
##  Outcome
##  0:500
##  1:268
##
##
##
##
##
```

4.3 Imputación de valores

Se considera apropiada la imputación de los valores faltantes conjuntamente con los outliers mediante el método *k Nearest Neighbors*, con un k de 4.

```
dtbk <- dt

numeric_columns <- colnames(dt[,sapply(dt, is.numeric)])
dt[numeric_columns] <- as.data.frame(
  scale(dt[numeric_columns])
)

dt <- kNN(dt, variable = colnames(dt)[colSums(is.na(dt))>0],
          k = 4, imp_var = FALSE)


for(col.name in numeric_columns) {
  dt[,col.name] <- sd(dtbk[,col.name], na.rm = TRUE)*dt[,col.name]+
                   mean(dtbk[,col.name], na.rm = TRUE)
}


summary(dt)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.0
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:23.0
##  Median : 3.000   Median :117.0   Median : 72.00   Median :29.5
```

```
##  Mean   : 3.845   Mean   :121.6   Mean   : 72.53   Mean   :29.1
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:35.0
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :63.0
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   : 14.0   Min.   :18.20   Min.   :0.0780           Min.   :21.00
##  1st Qu.: 86.0   1st Qu.:27.50   1st Qu.:0.2437           1st Qu.:24.00
##  Median :125.0   Median :32.30   Median :0.3725           Median :29.00
##  Mean   :146.8   Mean   :32.44   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:180.0   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :59.40   Max.   :2.4200           Max.   :81.00
##  Outcome
##  0:500
##  1:268
##
##
##
##
```

4.4 Estudio reducción de la dimensionalidad por SVD

A continuación se estudia la posibilidad de reducir la dimensionalidad mediante el método *Single Value Decomposition*. La descomposición en valores singulares es un método lineal de factorización de matrices. Sea X una matriz de orden mxn, se puede descomponer en la forma:

$$X = U\Sigma V^T$$

donde:

- U es una matriz ortogonal de orden m cuyas columnas son los autovectores de AA^T.
- Σ es una matriz diagonal de orden mxn que contiene la raíz de los valores propios ordenados de AA^T de manera decreciente en la diagonal principal, llamados valores singulares.
- V es una matriz ortogonal de orden n cuyas columnas son los autovectores de A^TA.

Las columna de la matriz U conforman una base del espacio de las observaciones.

El cálculo de los vectores y valores propios se puede realizar mediante la función *svd()* de *R*. Dado que los valores singulares se encuentran ordenados de manera creciente, los vectores conformados por las columnas de menor índice de U y V tienen un efecto mayor. En el caso de que algún valore singular fuese 0, tendríamos que la dimensión asociada a ese vector no forma parte de la base del espacio de observaciones, lográndose una representación más compacta (con menor dimensionalidad) del problema. Adicionalmente, si un conjunto de valores singulares tomasen valores comparativamente pequeños, tendríamos que sus dimensiones asociadas no tiene demasiada influencia, por lo que se podría generar una representación alternativa de los datos que no las incluya. (Brunton 2020)

A continuación se convierten los datos en una matriz y se obtienen U, Σ y V con la función *svd()*. Seguidamente se muestran los valores singulares obtenidos.

```
dt1 <- dt
dt1$Outcome <- as.numeric(dt1$Outcome)
dt1 <- scale(dt1)

mat.dt1 <- as.matrix(dt1)
mdl <- svd(mat.dt1)
```

```
mdl$d
```

```
## [1] 47.05654 34.09415 30.71729 26.47081 24.11367 21.42015 17.86701 17.07176
## [9] 15.20030
```

Tras comparar los valores propios obtenidos se observa que pertenecen al mismo orden de magnitud. Por este motivo se considera apropiado mantener la base original al tener una interpretación más inmediata.

4.5 Publicación de datos preprocesados

Tras completarse la limpieza de los datos, se procede a exportarlos para permitir su reutilización posterior.

```
write.csv(dt,"../res/diabetes_clean.csv", row.names = FALSE)
```

4.6 Estudio bivariante

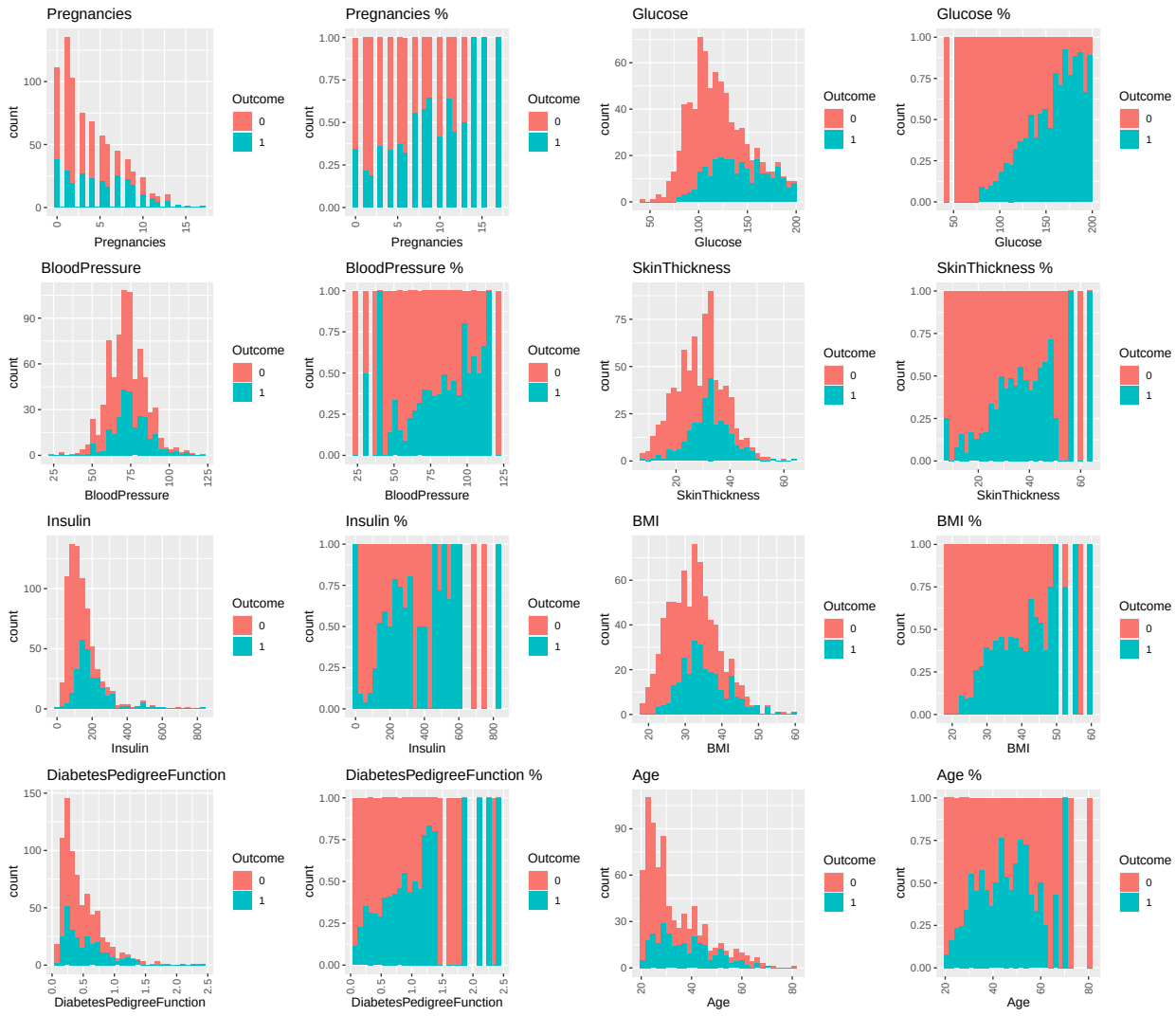
En este apartado se trata de desarrollar una mayor comprensión de la distribución de las distintas variables y la intensidad del efecto de correlación entre ellas y con la variable dependiente.

4.6.1 Representación gráfica

En primer lugar se representan gráficamente las distintas variables independientes en función de la variable dependiente.

```
dep.var <- "Outcome"
bar.plots <- list()
cols <- c(names(dt)[sapply(dt, is.numeric)], dep.var)
cols <- setdiff(colnames(dt), cols)

for(var in cols){
  p <- ggplot(dt, aes_string(x=var, fill=dep.var)) +
       theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
       geom_bar() +
       ggtitle(var)
  bar.plots[[var]] <- p
  p <- ggplot(dt, aes_string(x=var, fill=dep.var)) +
       theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
       geom_bar(position = "fill") +
       ggtitle(paste(var,"%"))
  bar.plots[[paste(var,"%")]] <- p
}

cols <- names(dt)[sapply(dt, is.numeric)]
for(var in names(dt)[sapply(dt, is.numeric)]){
  p <- ggplot(dt, aes_string(x=var, fill=dep.var)) +
       theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
       geom_histogram(bins=30) +
       ggtitle(var)
  bar.plots[[var]] <- p
```

```
  p <- ggplot(dt, aes_string(x=var, fill=dep.var)) +
       theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
       geom_histogram(bins=30, position = "fill") +
       ggtitle(paste(var,"%"))
  bar.plots[[paste(var,"%")]] <- p
}

do.call(grid.arrange, c(bar.plots, ncol=4))
```



A partir del estudio visual de las gráficas, se plantean las siguiente hipótesis:

1. La probabilidad de desarrollar diabetes aumenta con el número de embarazos.
2. Niveles altos de glucosa o de insulina en el examen de glucosa posprandial de 2 horas son un indicador de la enfermedad. También son indicadores de la enfermedad o factores de riesgo niveles de presión sanguínea alta, espesor del pliegue cutáneo del triceps altos o índices de masa corporal elevado.
3. Parece observarse una tendencia creciente del recuento de diabéticos con el aumento de la función de pedigree.

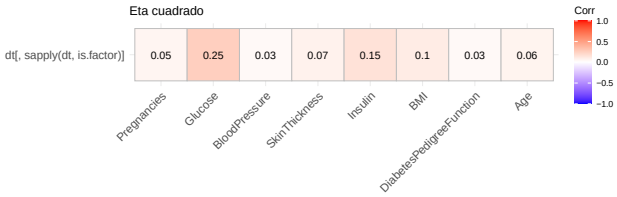
4.6.2 Estudio de correlaciones

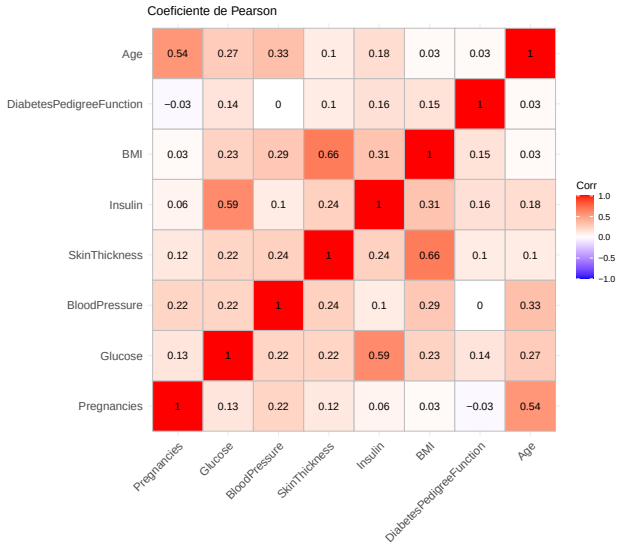
Para complementar el estudio visual de la distribución de las variables se realizará un estudio de correlaciones empleando los estadísticos *eta cuadrado* η^2 y *coeficiente de Pearson*. Nótese que el objetivo de calcular ambos coeficientes es la de disponer de una estimación aproximada de la intensidad del efecto de asociación/correlación entre variables, es decir, de la significancia práctica, independientemente de la significancia estadística. Por este motivo, tomando en consideración la robustez del método ANOVA frente a la falta de normalidad (que se observa en *Pregnancies*, *Age* o *DiabetesPedigreeFunction*), se considera aceptable el uso del coeficiente.

Para la interpretación del coeficiente se pueden emplear los valores de la tabla siguiente:

Coeficiente	Efecto débil	Efecto moderado	Efecto fuerte
η^2 ("Eta-Squared" 2020)	[0.01, 0.06]	(0.06, 0.14]	(0.14, 1]
abs(coef. Pearson) ("Pearson's Correlation Coefficient" s.f.)	[0, 0.3]	(0.3, 0.5]	(0.5, 1]

```r
# Crames's V correlation
aux <-as.data.frame(dt[, sapply(dt, is.factor)])

# Create eta-squared correlation plot.
numeric.cols <- names(dt)[sapply(dt, is.numeric)]

assoc_mat <- matrix(nrow = length(numeric.cols),
                    ncol = length(aux),
                    dimnames=list(numeric.cols, names(aux)))

for (c in seq(ncol(assoc_mat))){
   for (r in seq(length(numeric.cols))){
      anova_res <- summary(aov(dt[,numeric.cols[r]] ~ aux[, c]))[[1]][, "Sum Sq"]
      assoc_mat[[r, c]] <- anova_res[1] / sum(anova_res)
   }
}
p1 <- ggcorrplot(assoc_mat, lab = TRUE) +
      ggtitle("Eta cuadrado")

# Pearson correlation
p2 <- ggcorrplot(cor(dt[names(dt)[sapply(dt, is.numeric)]]), lab = TRUE) +
      ggtitle("Coeficiente de Pearson")


grid.arrange(p1, p2, layout_matrix=rbind(c(1,2)))
```





Se concluye que existen fuertes asociaciones entre las siguientes parejas de variables:

1. *Age* Vs. *Pregnancies*: esta asociación se puede explicar dada la naturaleza acumulativa de las dos variables: hasta ciertas edades, un número elevado de embarazos indicará edades más avanzadas.
2. *Insulin* Vs. *Glucose*: estas dos variables están asociadas al tratarse de los resultados de una misma prueba y ser la insulina una hormona que interviene directamente ne el control de la glucemia.
3. *BMI* Vs. *SkinThickness*: la relación entre estas dos variables también es explicable, dado que mayores índices de masa corporal resultarán en medidas mas altas de espesor del pliegue cutáneo del triceps.
4. *Glucose* y *Insulin* Vs. *Outcome*: Los niveles de glucosa y de insulina tras el examen de glucosa posprandial de 2 horas están fuertemente relacionado con la variable dependiente.

---

5 Análisis

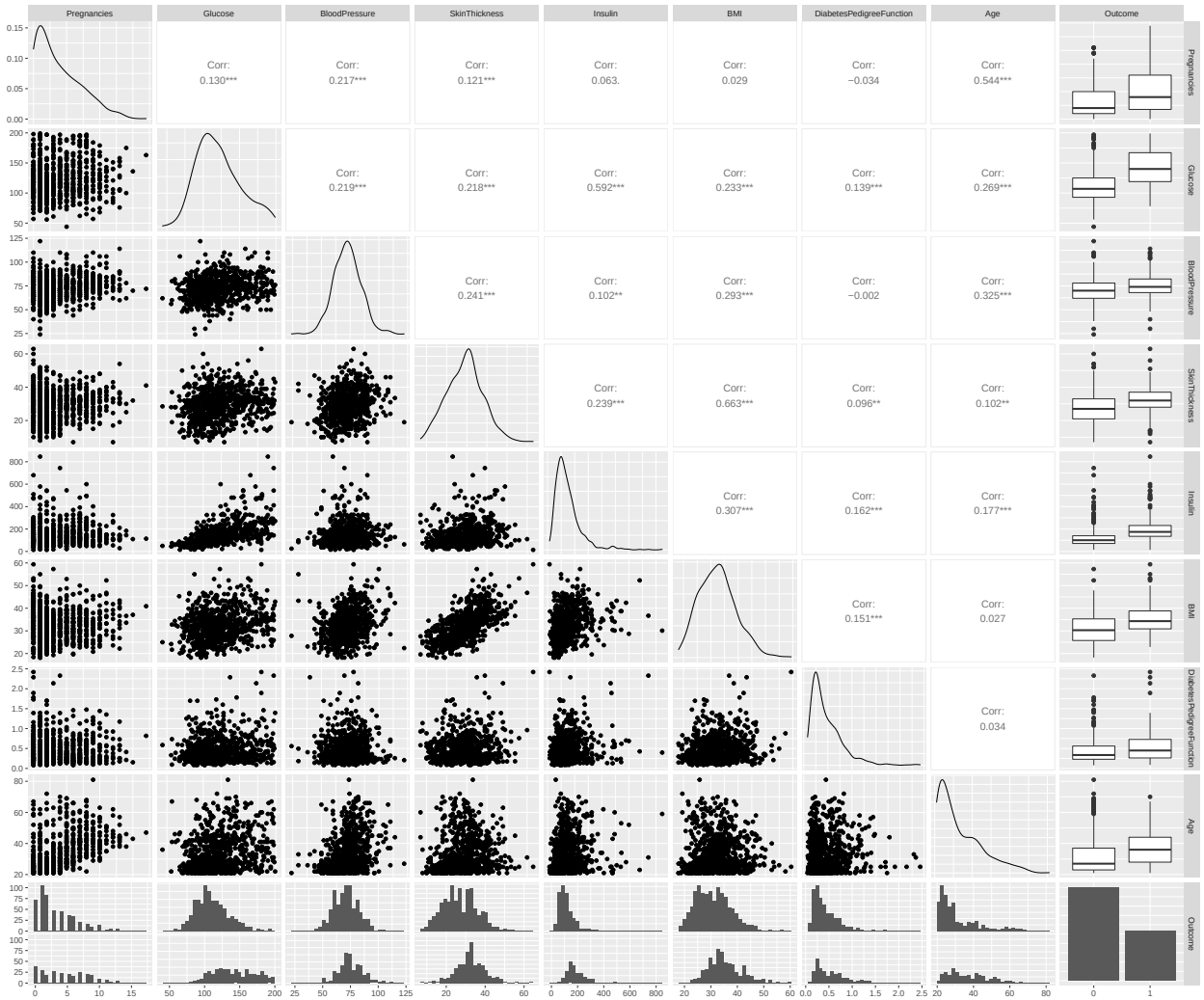
5.1 Clustering

Para obtener una comprensión más completa acerca de la composición del espacio de soluciones se plantea el empleo de modelos de aprendizaje automático no supervisado de *clustering*.

5.1.1 Inspección preliminar

En primer lugar, se procede a la visualización de las gráficas de dispersión de las distintas parejas de variables:

```
1 ggpairs(dt)
```

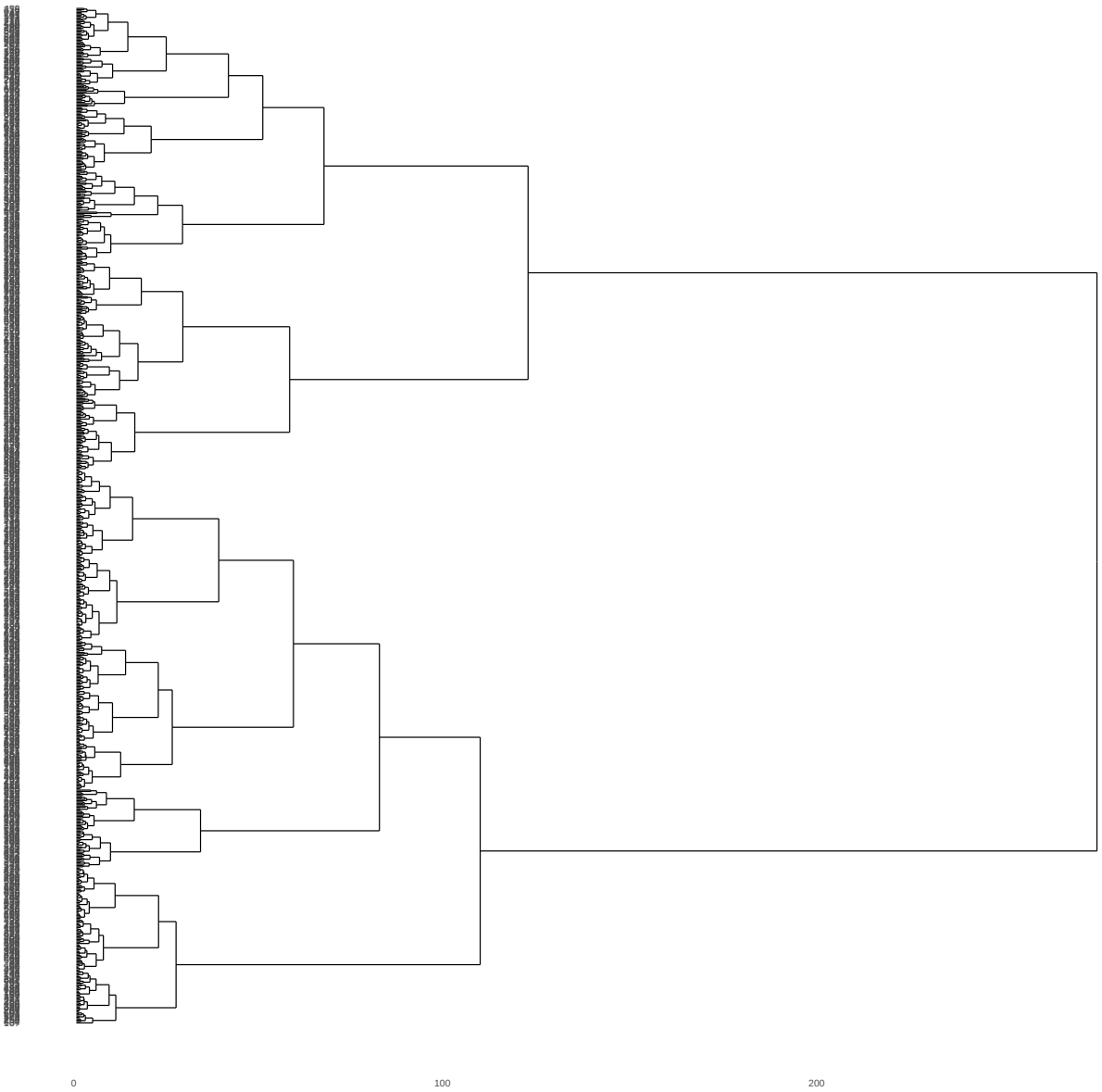


En las proyecciones bidimensionales del espacio no se observan agrupaciones de puntos claras. Si se observan distribuciones diferentes para las variables *Glucose*, *SkinThickness*, *BMI* y *Age* en función de la clase de la variable *Outcome*. Nótese que esta conclusión esta alineada con los resultados del estudio de correlaciones, que detectó importantes correlaciones entre estas variables.

5.1.2 Ajuste del modelo

Dada su simplicidad y flexibilidad, se decide ajustar un modelo de *clustering* jerárquico sobre el espacio de las variables independientes, para extrapolar los diferentes tipos de pacientes.

```
scaled.dt <- dt[,!(names(dt) %in% c("Outcome"))]
numeric.cols <- names(scaled.dt)[sapply(scaled.dt, is.numeric)]

scaled.dt[, numeric.cols] <- scale(dt[, numeric.cols])

c.hier <- hclust(dist(scaled.dt, method = "euclidean"), method = "ward")
ggdendrogram(c.hier, rotate = TRUE)
```



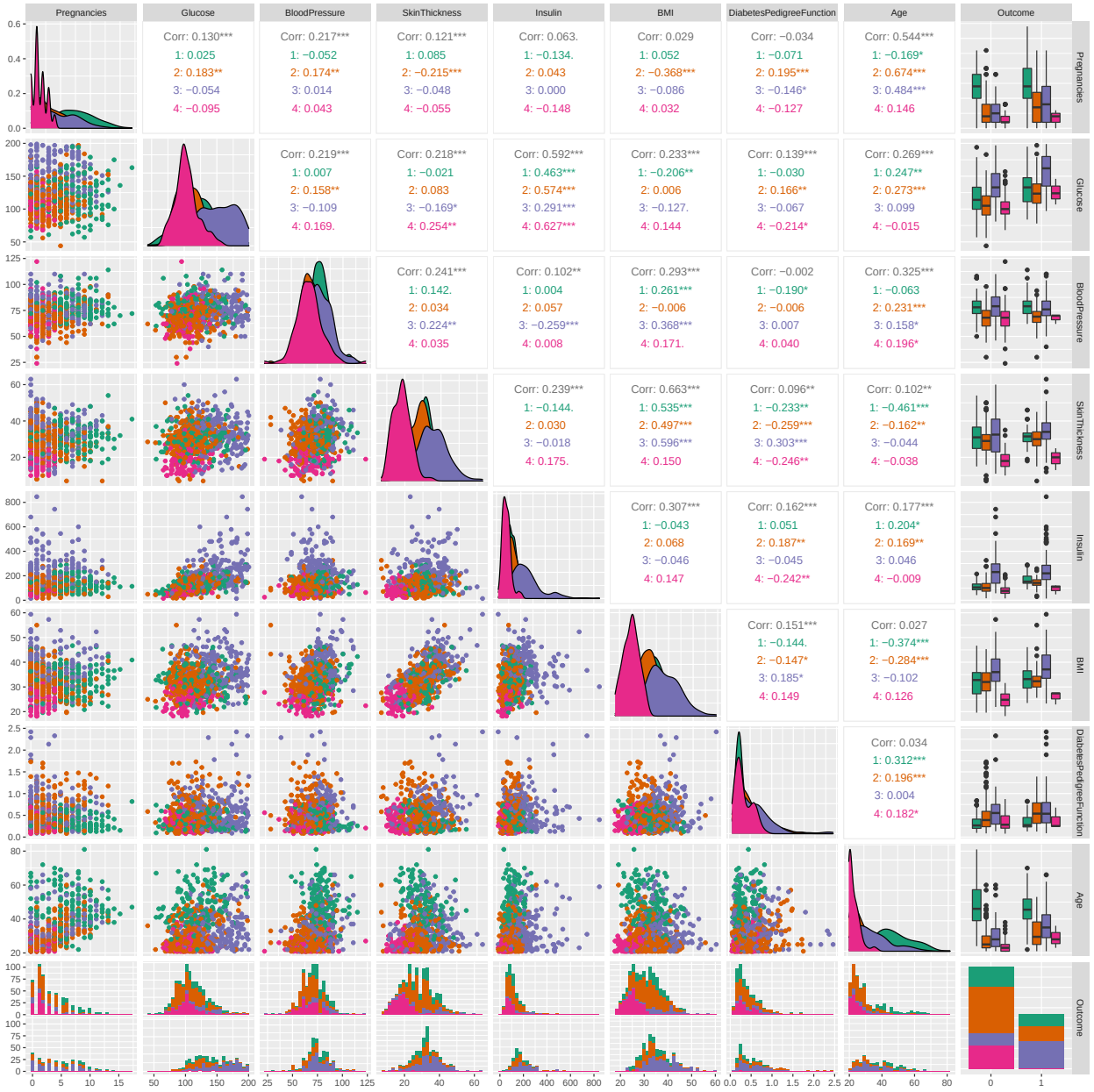
Atendiendo al dendograma obtenido, 4 clusters parece la opción más apropiada.

5.1.3 Visualización de resultados e interpretación.

A continuación se representan las agrupaciones obtenidos y su interpretación:

```
plot.clusters <- function(dt, clusters) {
  # Color palet extracted from:
  # https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-visualization/
  color.palette <- c("#1B9E77", "#D95F02", "#7570B3",
                     "#E7298A", "#66A61E",  "#E6AB02", "#A6761D")
  c <- as.factor(clusters)
  K=length(levels(c))
```

```
  print(ggpairs(dt, aes(colour = c)) +
    scale_color_manual(values= color.palette[1:K]) +
    scale_fill_manual(values= color.palette[1:K]))
}


cls <- cutree(c.hier, k = 4)
plot.clusters(dt, cls)
```



1. **Pacientes de riesgo muy bajo** (en magenta): Se trata de personas sin problemas de obesidad (índice de masa corporal inferior $30 kg/m^2$ y espesor de pliegue cutáneo del triceps menor a $30 mm$), con preeminencia del rango de edades de hasta 30 años y niveles de glucosa e insulina normales en el examen de glucosa posprandial de 2 horas (hasta los $125 \mu g/l$ y $200 \mu l\ U/ml$ respectivamente).

2. **Pacientes de riesgo bajo** (en naranja): De manera bastante similar al caso anterior, se trata de personas en la franja de edades bajas (principalmente hasta los 40 años) y niveles de glucosa e insulina normales en el examen de glucosa posprandial de 2 horas (hasta los $125\mu g/l$ y $200\mu l$ U/ml respectivamente). La diferencia fundamental con los pacientes de riesgo muy bajo se encuentra en la existencia de problemas de obesidad, con una distribución del espesor de pliegue cutáneo del triceps concentrada entre 25 y $50mm$, índice de masa corporal entre 25 y $45kg/m^2$.
3. **Pacientes con nivel de riesgo moderado** (en verde): Se trata de pacientes con niveles de glucosa e insulina normales o moderadamente altos en el examen de glucosa posprandial de 2 horas (hasta los $175\mu g/l$ y $250\mu l$ U/ml respectivamente), con problemas de obesidad y con una distribución del espesor de pliegue cutáneo del triceps concentrada entre 25 y $50mm$, índice de masa corporal entre 25 y $45kg/m^2$ y edades preeminentemente superiores a los 30 años.
4. **Pacientes de alto riesgo** (en morado): Se trata de pacientes con niveles de glucosa e insulina altos en el examen de glucosa posprandial de 2 horas (superiores a los $125\mu g/l$ y $100\mu l$ U/ml respectivamente), con preeminencia de pacientes con problemas de obesidad (BMI superior a $30kg/m^2$ y $SkinThickness$ superior a $20mm$) y distribución de edades amplia y preeminecia de edades hasta los 40 años.

5.2 Contraste de hipótesis

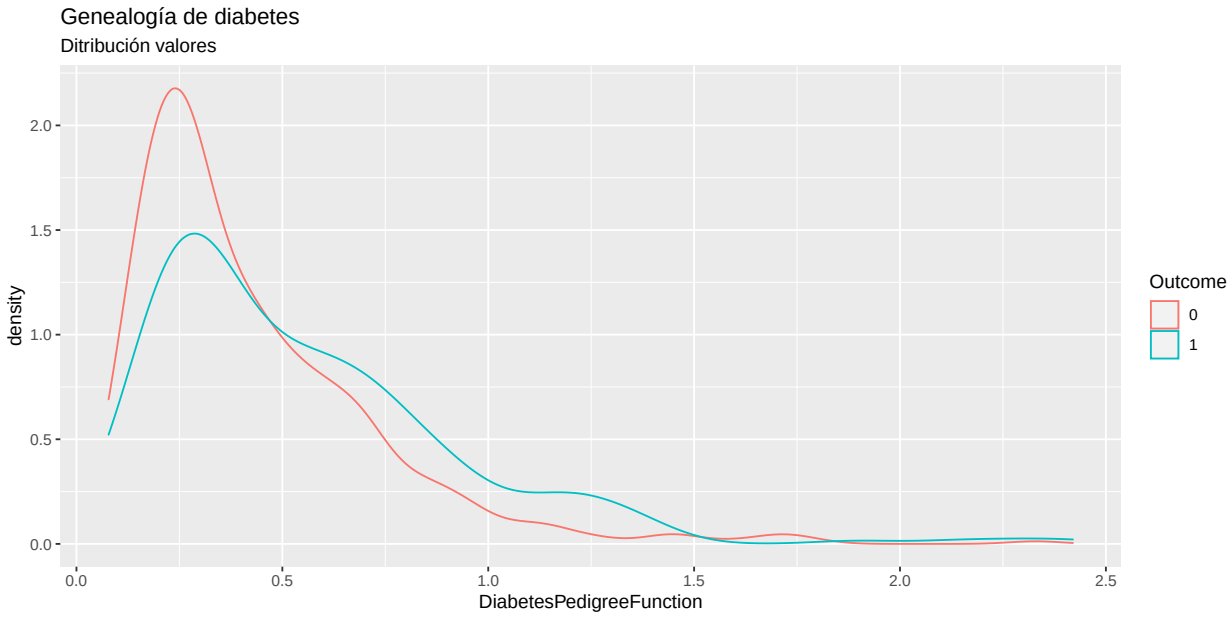
Se decide realizar un contraste de hipótesis que permita identificar si las personas que han sido diagnosticadas como diabéticas presentan un valor de genealogía de la diabetes (variable *DiabetesPedigreeFunction*) mayor que las no diagnosticadas con diabetes, con un intervalo del confianza del 95%.

5.2.1 Pregunta de investigación

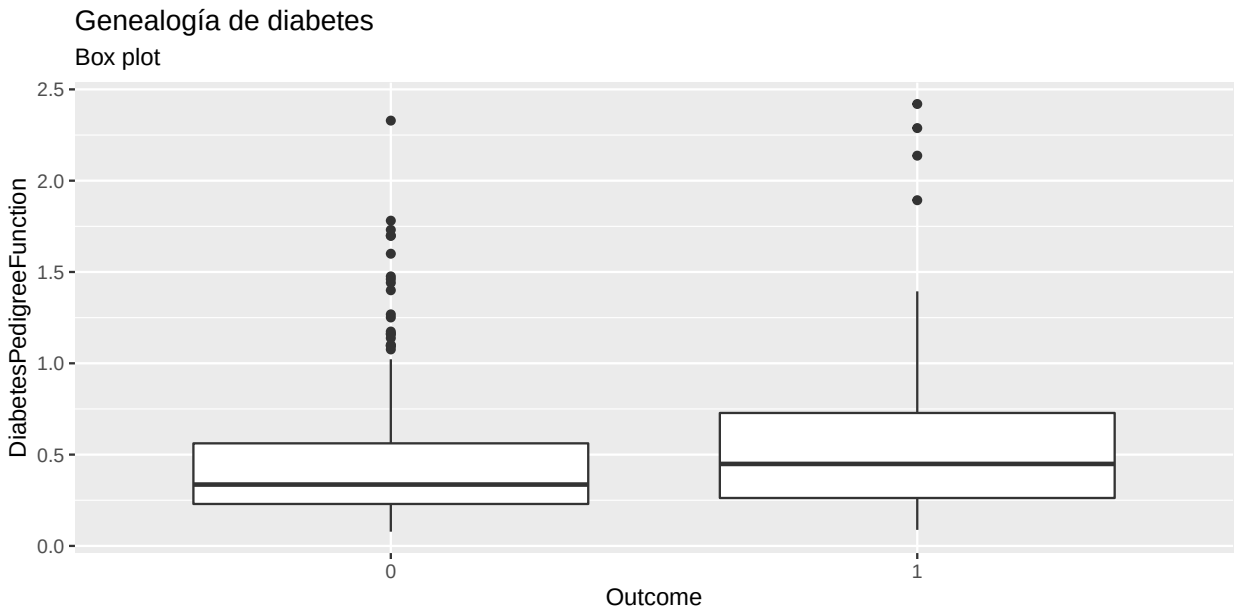
¿Las personas diagnosticadas como diabéticas presentan un genealogía de diabetes mayor que las personas no diagnosticadas como diabéticas?

5.2.2 Inspección preliminar

```
# distribución de los valores
ggplot() + geom_density(data=dt, aes(x = DiabetesPedigreeFunction, group = Outcome, colour=Outcome)) + l
```



```r
# box plot
ggplot(data = dt, aes(x=Outcome, y=DiabetesPedigreeFunction)) + geom_boxplot()  +
  labs(title = "Genealogía de diabetes", subtitle = "Box plot")
```



A nivel visual, se puede apreciar que se presentan valores más elevados sobre la genealogía de diabetes en las personas diagnosticadas como diabéticas que en las no diagnosticadas.

5.2.3 Hipótesis nula y alternativa

- **Hipótesis nula**: el valor de genealogía de diabetes de las personas con diabetes diagnosticada es menor o igual que el valor de genealogía de diabetes en personas sin diabetes diagnosticada.
- **Hipótesis alternativa**: el valor de genealogía de diabetes de las personas con diabetes diagnosticada es mayor que el valor de genealogía de diabetes en personas sin diabetes diagnosticada.

5.2.4 Método

Se hacen algunas comprobaciones previas antes de determinar todas las características del método a aplicar:

A) Tamaño de las muestras

```r
# tamaño de la muestra de personas diagnosticadas como diabéticos
print(sprintf('Tamaño muestra diabéticas: %i',
              nrow(dt[dt$Outcome == 1,])))
```

```
## [1] "Tamaño muestra diabéticas: 268"
```

```r
# tamaño de la muestra de diestros
print(sprintf('Tamaño muestra no diabéticas: %i',
              nrow(dt[dt$Outcome == 0,])))
```

```
## [1] "Tamaño muestra no diabéticas: 500"
```

B) Test de igualdad de varianzas en la genealogía de diabetes.

```
# test de homocedasticidad
var.test(dt$DiabetesPedigreeFunction[dt$Outcome == 1]
         , dt$DiabetesPedigreeFunction[dt$Outcome == 0])
```

```
##
##  F test to compare two variances
##
## data:  dt$DiabetesPedigreeFunction[dt$Outcome == 1] and dt$DiabetesPedigreeFunction[dt$Outcome == 0]
## F = 1.55, num df = 267, denom df = 499, p-value = 3.03e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.259981 1.919708
## sample estimates:
## ratio of variances
##           1.549969
```

El valor observado relacionado con las varianzas muestrales, el ratio de varianzas, es de 1.549969, estando el intervalo de confianza del 95% determinado entre los valores 1.259981 y 1.919708, por lo que se acepta la hipótesis nula (en este caso, la igualdad de las dos varianzas). En resumen, se puede asumir **homocedasticidad**.

El método a aplicar para validar la hipótesis planteada (es decir, para evaluar si hay suficiente evidencia para rechazar la hipótesis nula), depende de las siguientes consideraciones propias del estudio planteado:

1. Se trata de un **contraste de dos muestras independientes sobre la media**, con varianzas desconocidas.
2. Las muestras consideradas contienen una cantidad elevada de elementos, ya que se dispone de 268 personas diagnosticadas con diabetes y 500 no diagnosticadas. Por tanto, se puede considerar la aplicación del teorema del límite central (TLC), que establece que el contraste de hipótesis sobre la media de una muestra se aproxima a una distribución normal aunque la población original no siga una distribución normal, siempre que el tamaño de la muestra n sea suficientemente grande ($n < 30$). Se asume, de esta forma, **normalidad** en los datos.
3. El test estadístico a aplicar va a ser **parámetrico**, puesto que se asume, a través del teorema del límite central, que la población original sigue una distribución normal. Por tanto, se realizará un test paramétrico para obtener las inferencias sobre la población.
4. Se trata de un test **unilateral**, pues la hipótesis alternativa plantea tan solo si las personas diagnosticadas con diabetes presentan mayores valores de genealogía de diabetes, por lo que se evalúa tan solo la cola de la derecha, buscando valores lo suficientemente altos de las variables para rechazar la hipótesis nula.
5. En base al test de igualdad de varianzas (homocedasticidad) realizado previamente, se concluye que las dos varianzas son iguales con un nivel de confianza del 95 % (se puede asumir **homocedasticidad**)

5.2.5 Cálculos

```
diab <- dt$DiabetesPedigreeFunction[dt$Outcome == 1]
no_diab <- dt$DiabetesPedigreeFunction[dt$Outcome == 0]
# número de observaciones de cada muestra
n_1 <- length(diab)
n_2 <- length(no_diab)
```

```r
# intervalo de confianza 95%: alfa = 0.05
alfa <- 0.05
# media y desviación estándar de las diferencias
mean_1 <- mean(diab)
mean_2 <- mean(no_diab)
s_1 <- sd(diab)
s_2 <- sd(no_diab)
sd <- sqrt(((n_1-1)*s_1^2+(n_2-1)*s_2^2)/(n_1+n_2-2))
# estadístico de contraste
tobs <- (mean_1-mean_2)/(sd*sqrt(1/n_1+1/n_2))
#Región de aceptación
tcrit <- qt(1-alfa, df=n_1+n_2-2)
#Cálculo del valor p
pvalue <- pt(tobs, lower.tail=FALSE, df=n_1+n_2-2)
pvalue <- pnorm(tobs, lower.tail=FALSE)

tobs;tcrit;pvalue
```

```
## [1] 4.885826
```

```
## [1] 1.646845
```

```
## [1] 5.149794e-07
```

- Estadístico de contraste: 4.885826
- Valor crítico: 1.646845
- Valor p: 5.1497943 · 10-7

5.2.6 Interpretación

Se puede rechazar la hipótesis nula, aceptando la hipótesis alternativa. Por tanto, se infiere que la genealogía de la diabetes en las personas diagnosticadas con diabetes es mayor que las no diagnosticadas con diabetes con un intervalo de confianza del 95%. Cabe destacar que estos resultados serían extrapolables a una población con las mismas características que la muestra de estudio.

5.3 Regresión logística

Se pretende estudiar la probabilidad que tiene una persona de ser diagnosticada con diabetes en base al conocimiento previo de diferentes parámetros: número de embarazos, presión sanguínea distólica, doblez de piel, nivel de insulina, índice de masa corporal, valor de genealogía de diabetes, edad. Se generará un modelo de regresión logística que pueda proporcionar la probabilidad de ser diagnosticada con diabetes en base a dichas variables de entrada.

5.3.1 Creación del modelo.

Se crea el modelo de regresión logística.

```r
glm_logit = glm(formula= Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness
                + Insulin + BMI + DiabetesPedigreeFunction + Age
                   , data=dt, family=binomial(link=logit))
```

Se inspecciona el modelo creado:

```
summary(glm_logit)
```

```
##
## Call:
## glm(formula = Outcome ~ Pregnancies + Glucose + BloodPressure +
##     SkinThickness + Insulin + BMI + DiabetesPedigreeFunction +
##     Age, family = binomial(link = logit), data = dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2476  -0.7205  -0.3798   0.7311   2.3594
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -8.838960   0.821139 -10.764  < 2e-16 ***
## Pregnancies               0.118740   0.032084   3.701 0.000215 ***
## Glucose                   0.031959   0.004054   7.883  3.2e-15 ***
## BloodPressure            -0.006142   0.008532  -0.720 0.471568
## SkinThickness             0.019042   0.013852   1.375 0.169228
## Insulin                   0.002813   0.001290   2.180 0.029240 *
## BMI                       0.070815   0.019580   3.617 0.000298 ***
## DiabetesPedigreeFunction  0.838911   0.299469   2.801 0.005089 **
## Age                       0.012203   0.009535   1.280 0.200636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 706.26  on 759  degrees of freedom
## AIC: 724.26
##
## Number of Fisher Scoring iterations: 5
```

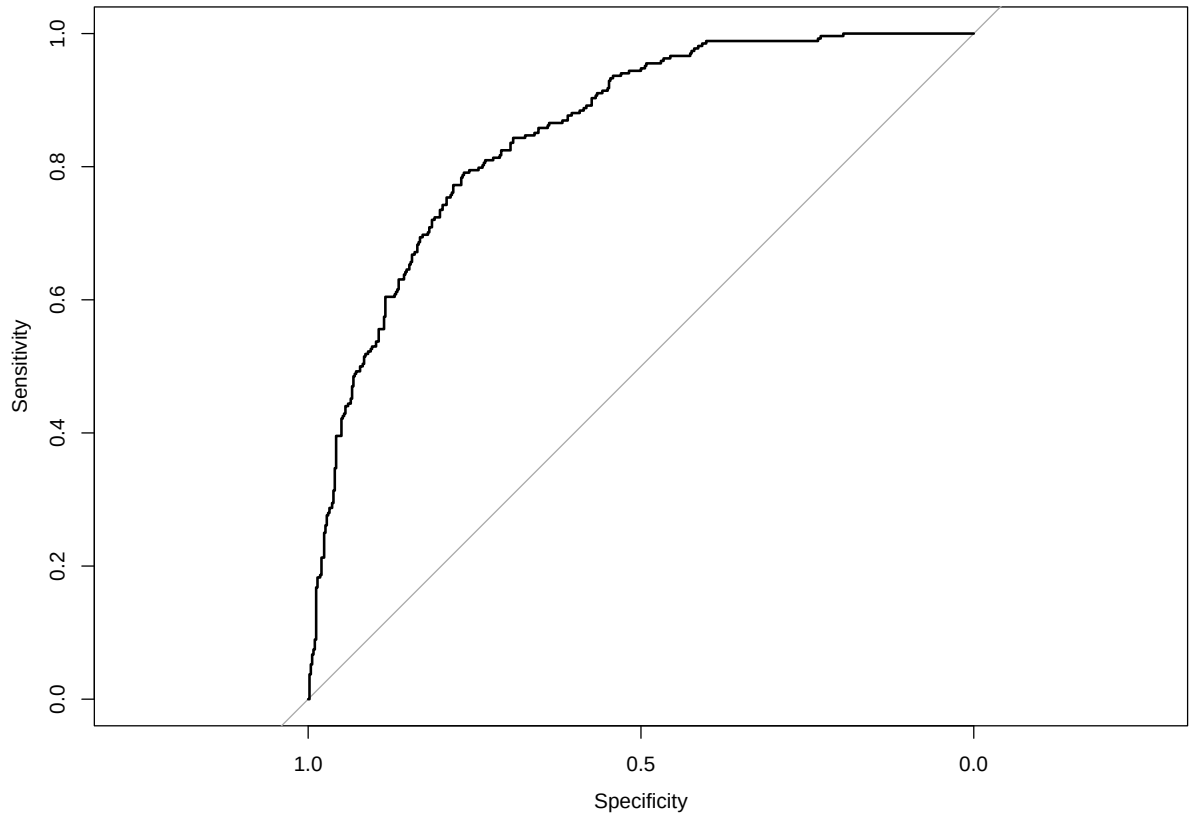
Por un lado, se puede observar que las variables con una mayor significancia en el modelo son *Pregnancies*, *Glucose* y *BMI*. Todas ellas, tienen una influencia directa (signo positivo) sobre la probabilidad de ser diagnosticada con diabetes y presentan significancia elevada para rechazar la hipótesis nula ($\Pr[>|z|] < 0.001$). Por otra parte, se tienen con significancia alta, pero menor que las comentadas previamente, las variables *DiabetesPedigreeFunction* e *Insulin* ($\Pr[>|z|] < 0.05$), que también influyen de forma directa en la probabilidad de ser diagnosticada de diabetes. Por último, no se consideran significativas para el modelo las variables *BloodPressure*, *SkinThickness* y *Age*.

5.3.2 Evaluación del modelo

Para la evaluación del modelo, se genera, representa e inspecciona la curva ROC:

```
df_vars_explic <- dt %>% select(Pregnancies, Glucose, BloodPressure, SkinThickness,
                                Insulin, BMI, DiabetesPedigreeFunction, Age)
prob <- predict(glm_logit,
                df_vars_explic,
```

```
                  type="response")
r <- roc(dt$Outcome, prob, data=df_vars_explic)
plot(r)
```



Se evalúa el área bajo la curva ROC (AUROC):

```
auc(r)
```

```
## Area under the curve: 0.8511
```

Visualmente se aprecia que la curva ROC llega a alejarse mucho a la diagonal, con lo que el área bajo la curva (AUROC) se intuye de un tamaño significativo. Al calcularla, se obtiene un valor de AUROC=0.8511 que confirma lo que se había intuido visualmente, con lo que se puede determinar que el modelo discrimina de forma excelente.

5.3.3 Visualización de datos e interpretación

Se ha generado un modelo de regresión logística que permite discriminar de forma excelente la probabilidad de ser diagnosticada con diabetes. Las variables con una mayor significancia en el modelo son *Pregnancies*, *Glucose* y *BMI*, presentando todas ellas una influencia directa (signo positivo) en la probabilidad de ser diagnosticada con diabetes.

6 Resolución del problema. Conclusiones

6.1 Clústering

Se interpreta que hay tres factores que intervienen de manera destacada en la aparición de la enfermedad:

1. La presencia de **problemas de obesidad**, cuyos indicadores son las variables *BMI* y *SkinThickness*, está asociada con la presencia de la enfermedad y valores altos (superiores a $25kg/m^2$ y $20mm$) podrían ser considerados factores de riesgo que aumenta la probabilidad de la misma.
2. **La edad**. La probabilidad de desarrollo de la enfermedad aumenta sensiblemente con la edad.
3. Los **resultados del test de glucosa posprandial de 2 horas**, disponibles en las variables *Glucose* e *Insulin*. Se trata de dos variables en que la intensidad del efecto de asociación entre ambas es fuerte (test de pearson de 0.59). Resultados con altos niveles de glucemia y/o de insulina en plasma (superiores a los $125\mu g/l$ y $100\mu l\ U/ml$ respectivamente) están habitualmente relacionados con la presencia de la enfermedad, y habitualmente son considerados un indicador de la misma, a pesar de que en este trabajo no se ha demostrado la causalidad.

6.2 Contraste de hipótesis

Se infiere que la genealogía de la diabetes en las mujeres nativas norteamericanas mayores de 21 años diagnosticadas con diabetes es mayor que las no diagnosticadas con diabetes con un intervalo de confianza del 95%.

6.3 Regresión logística

Se ha generado un modelo de regresión logística que permite discriminar de forma excelente la probabilidad de ser diagnosticada con diabetes. Las variables con una mayor significancia en el modelo son *Pregnancies*, *Glucose* y *BMI*, presentando todas ellas una influencia directa (signo positivo) en la probabilidad de ser diagnosticada con diabetes.

---

7 Contribuciones

Contribuciones	Firma
Investigación previa	DALH, PRC
Redacción de las respuestas	DALH, PRC
Desarrollo del código	DALH, PRC

Bibligrafía y agradecimientos

Brunton, S. 2020. "Singular Value Decomposition (SVD): Mathematical Overview." Youtube. 2020. https://www.youtube.com/watch?v=nbBvuuNVfco&list=PLMrJAkhIeNNSVjnsviglFoY2nXildDCcv&index=2.

"Eta-Squared." 2020. WIKIVERSITY. 2020. https://en.wikiversity.org/wiki/Eta-squared.

"Hipertensión y Presión Arterial." s.f. Infosalus. s.f. https://www.infosalus.com/enfermedades/cardiologia/hipertension/que-es-hipertension-69.html.

"Insulin." 2019. Medscape. 2019. https://emedicine.medscape.com/article/2089224-overview.

Mehmet, A. 2020. "Diabetes Data Set." 2020. https://www.kaggle.com/mathchi/diabetes-data-set.

"Pearson's Correlation Coefficient." s.f. StatisticsSolutions. s.f. https://www.statisticssolutions.com/pearsons-correlation-coefficient/#:~:text=High%20degree%3A%20If%20the%20coefficient,to%20be%20a%20small%20correlation.