Reporte_Parcial2

Harold Romero

2023-10-09

Para el presente proyecto se usaran las siguientes librerias:

```
library (tidyverse)
library(caret)
library(class)
library(gmodels)
library(psych)
library(dplyr)
library(rpart)
```

En el presente documento se pretende desarrollar los puntos propuestos en el Parcial del Segundo Corte, Metodos supervisados en aprendizaje automatico kNN, regresión lineal y regresión multilineal

Trabajaremos con el conjunto de datos **diabetes_012_health_indicators.csv**

Este conjunto de datos contiene las siguientes variables:

- Diabetes_012: Describe la persona en que estado de Diabetes se encuentra
  - 0 = No diabetes - 1 = Prediabetes - 2 = Diabetes
- HighBP:
  - 0 = no high BP - 1 = high BP
- HighChol:
  - 0 = no high cholesterol - 1 = high cholesterol
- CholCheck:
  - 0 = no cholesterol check in 5 years - 1 = yes cholesterol check in 5 years
- BMI: Body Mass Index
- Smoker: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
  - 0 = no - 1 = yes
- Stroke: (Ever told) you had a stroke.
  - 0 = no - 1 = yes
- HeartDiseaseorAttack: coronary heart disease (CHD) or myocardial infarction (MI)
  - 0 = no - 1 = yes
- PhysActivity: physical activity in past 30 days - not including job
  - 0 = no - 1 = yes
- Fruits: Consume Fruit 1 or more times per day
  - 0 = no - 1 = yes

- Veggies: Consume Vegetables 1 or more times per day
  - 0 = no - 1 = yes
- HvyAlcoholConsump: (adult men >=14 drinks per week and adult women>=7 drinks per week)
  - 0 = no - 1 = yes
- AnyHealthcare: Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc.
  - 0 = no - 1 = yes
- NoDocbcCost: Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?
  - 0 = no - 1 = yes
- GenHlth: Would you say that in general your health is: scale 1-5
  - 1 = excellent - 2 = very good - 3 = good - 4 = fair - 5 = poor
- MentHlth: days of poor mental health scale 1-30 days
- PhysHlth: physical illness or injury days in past 30 days scale 1-30
- DiffWalk: Do you have serious difficulty walking or climbing stairs?
  - 0 = no - 1 = yes
- Sex:
  - 0 = female - 1 = male
- Age: 13-level age category (_AGEG5YR see codebook)
  - 1 = 18-24 - 9 = 60-64 - 13 = 80 or older
- Education: Education level (EDUCA see codebook) scale 1-6
  - 1 = Never attended school or only kindergarten - 2 = elementary etc.
- Income: Income scale (INCOME2 see codebook) scale 1-8
  - 1 = less than $10,000 - 5 = less than $35,000 - 8 = $75,000 or more

Cargar el Conjunto de Datos

Para cargar el Conjunto de Datos en R, crearemos una carpeta llamada *Data* en la carpeta del proyecto, en su interior se guardara el archivo con extensión `.cvs`

Posteriormente usaremos el siguiente comando para encontrar la carpeta padre

```
folder <- dirname(rstudioapi::getSourceEditorContext()$path)
parentFolder <- dirname(folder)
```

Posteriormente se cargara el dataset con el siguiente comando, donde el conjunto de datos los llamaremos *diabetes_012*:

```
diabetes_012 <-
  read_csv(paste0(parentFolder
                  ,"/Data/diabetes_012_health_indicators.csv"))
```

Analisis Exploratorio de Datos

Los datos que se observan en el Conjunto de Datos son de Tipo Numerico, las variables se encuentran binarizados, a excepción de `Diabetes_012, BMI, GenHlth, MenHlth, MentHlth, PhysHlth, Age, Education e Income`

Vemos las primeras observaciones del dataset asi:

```
head(diabetes_012)
```

```
## # A tibble: 6 x 22
##   Diabetes_012 HighBP HighChol CholCheck   BMI Smoker Stroke
##          <dbl>  <dbl>    <dbl>     <dbl> <dbl>  <dbl>  <dbl>
## 1            0      1        1         1    40      1      0
## 2            0      0        0         0    25      1      0
## 3            0      1        1         1    28      0      0
## 4            0      1        0         1    27      0      0
## 5            0      1        1         1    24      0      0
## 6            0      1        1         1    25      1      0
## # i 15 more variables: HeartDiseaseorAttack <dbl>, PhysActivity <dbl>,
## #   Fruits <dbl>, Veggies <dbl>, HvyAlcoholConsump <dbl>, AnyHealthcare <dbl>,
## #   NoDocbcCost <dbl>, GenHlth <dbl>, MentHlth <dbl>, PhysHlth <dbl>,
## #   DiffWalk <dbl>, Sex <dbl>, Age <dbl>, Education <dbl>, Income <dbl>
```

Este conjunto de datos tiene un numero total de observaciones de 253680, donde 213703 observaciones no tienen diabetes, 4631 tienen pre diabetes y 35346 tienen diabetes asi:

```
table(diabetes_012$Diabetes_012)
```

```
##
##      0      1      2
## 213703   4631  35346
```

Realizamos un resumen estadistico de las variables con el comando `summary` de nuestro dataset

```
summary(diabetes_012)
```

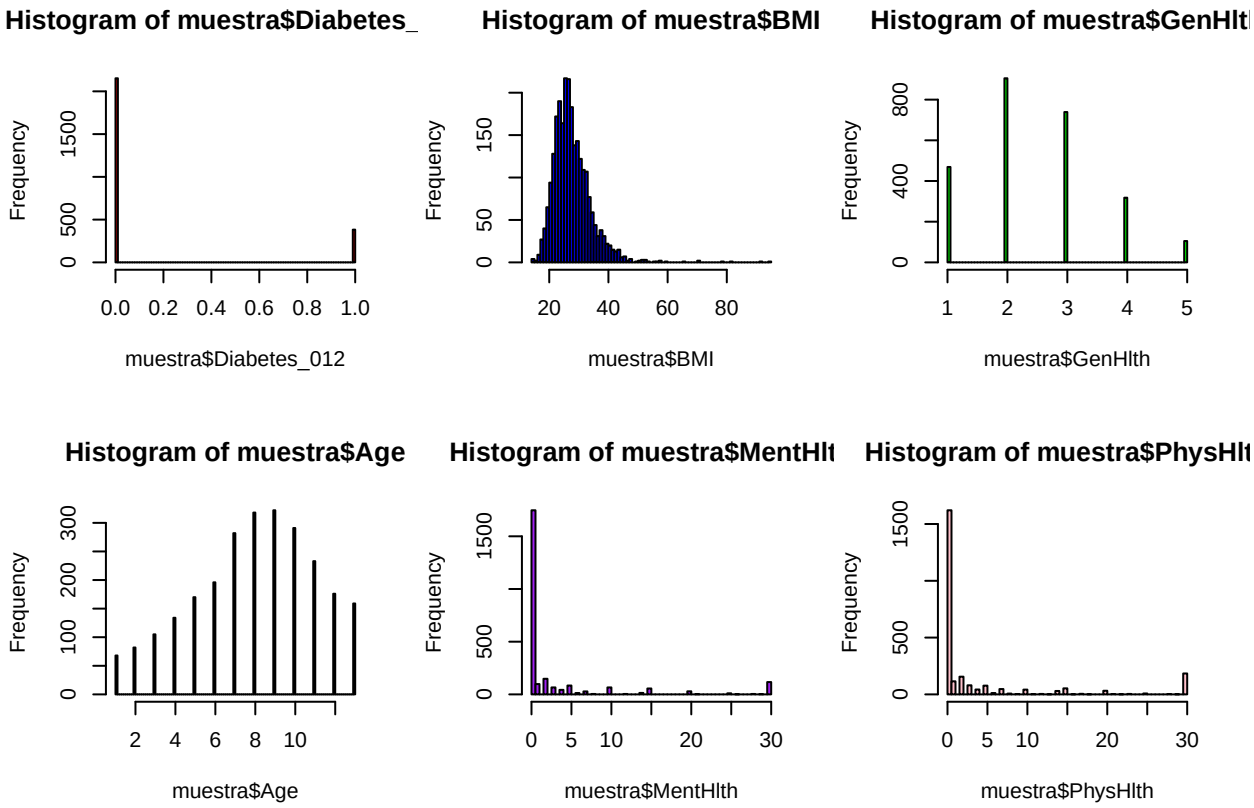
```
##   Diabetes_012        HighBP         HighChol        CholCheck
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.0000
##  Median :0.0000   Median :0.000   Median :0.0000   Median :1.0000
##  Mean   :0.2969   Mean   :0.429   Mean   :0.4241   Mean   :0.9627
##  3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :2.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##       BMI            Smoker           Stroke        HeartDiseaseorAttack
##  Min.   :12.00   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:24.00   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :27.00   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :28.38   Mean   :0.4432   Mean   :0.04057   Mean   :0.09419
##  3rd Qu.:31.00   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :98.00   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##   PhysActivity        Fruits          Veggies       HvyAlcoholConsump
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :0.0000
##  Mean   :0.7565   Mean   :0.6343   Mean   :0.8114   Mean   :0.0562
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
##  AnyHealthcare     NoDocbcCost         GenHlth         MentHlth
##  Min.   :0.0000   Min.   :0.00000   Min.   :1.000   Min.   : 0.000
##  1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 0.000
##  Median :1.0000   Median :0.00000   Median :2.000   Median : 0.000
##  Mean   :0.9511   Mean   :0.08418   Mean   :2.511   Mean   : 3.185
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:3.000   3rd Qu.: 2.000
##  Max.   :1.0000   Max.   :1.00000   Max.   :5.000   Max.   :30.000
##     PhysHlth         DiffWalk           Sex              Age
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   : 1.000
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 6.000
##  Median : 0.000   Median :0.0000   Median :0.0000   Median : 8.000
##  Mean   : 4.242   Mean   :0.1682   Mean   :0.4403   Mean   : 8.032
##  3rd Qu.: 3.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:10.000
##  Max.   :30.000   Max.   :1.0000   Max.   :1.0000   Max.   :13.000
##    Education        Income
##  Min.   :1.00   Min.   :1.000
##  1st Qu.:4.00   1st Qu.:5.000
##  Median :5.00   Median :7.000
##  Mean   :5.05   Mean   :6.054
##  3rd Qu.:6.00   3rd Qu.:8.000
##  Max.   :6.00   Max.   :8.000
```

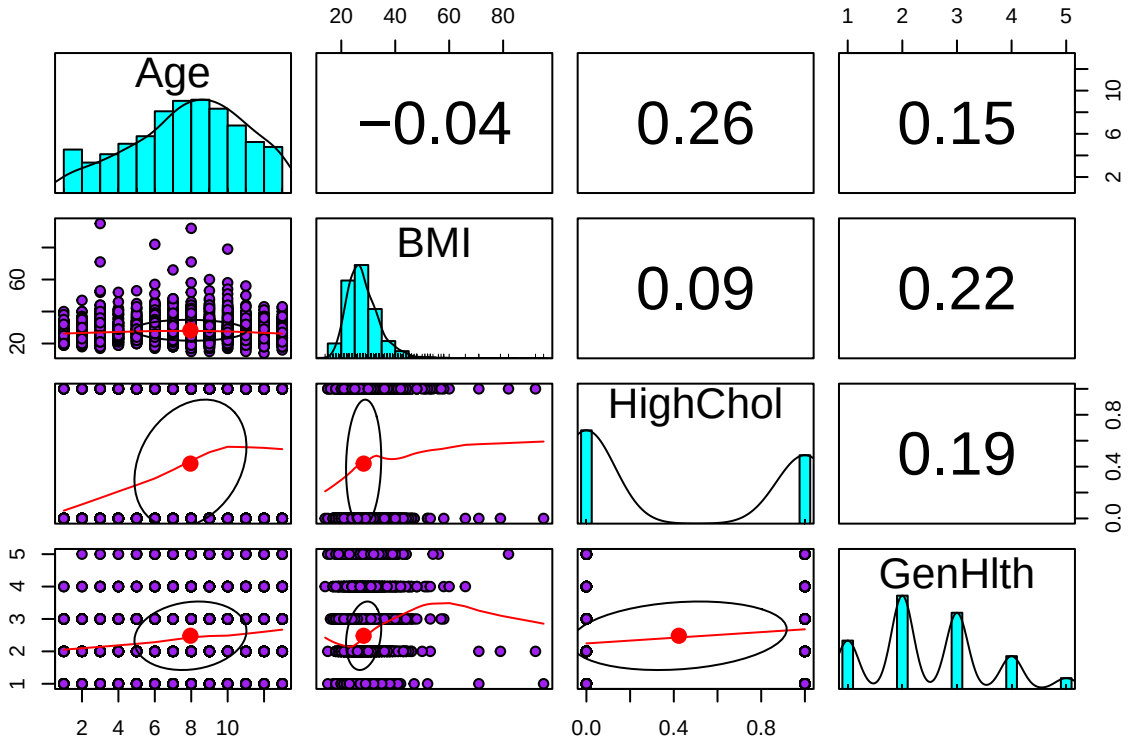
Binarizamos la variable `Diabetes_012` donde tomaremos como 0 las observaciones No diabetes y como 1 las observaciones con pre diabetes y diabetes:

```
diabetes_012$Diabetes_012 <- ifelse(diabetes_012$Diabetes_012 == 0, 0, 1)
```

Realizamos el Histograma de las variables *Diabetes_012, BMI, GentHlth, Age, MentHlth, PhysHlth* con el fin de ver la cantidad de observaciones realizadas por cada rango y con esto poder indicar la probabilidad que existe que una nueva observacion caiga en cada valor



Realizamos una correlación de los datos, poniendo en pares y en paneles con información de histogramas, variable clase Diabetes_012:



Parte 2 KNN

A continuación de implementaran modelos predictivos utilizando el metodo KNN al considerar las variables como variables clase 1. Diabetes_012 Versión Binaria 2. HeartDiseaseorAttack 3. Sex

Crear versiones adecuadas del conjunto de datos

Crearemos versiones adecuadas del conjunto de datos para cada modelo subdividiendo el conjunto de datos de modo que la variable clase esté equilibrada y corresponda al 1% del conjunto de datos, para ello realizaremos un muestreo estratificado:

Primer conjunto de datos para la variable Clase Diabetes_012 Version Binaria

Llamaremos a nuestro primer conjunto de datos `Datos_Primer_Modelo` con variable clase `Diabetes_012` realizando un muestreo estratificado, asi:

```
set.seed(1)
Datos_Primer_Modelo <- diabetes_012 %>%
  group_by(Diabetes_012) %>%
  sample_n(1268, replace = TRUE) %>%
  ungroup()
```

Posteriormente sacaremos los valores estadisticos del nuevo conjunto de datos con el comando `summary(Datos_Primer_Modelo)`, obteniendo los siguientes datos:

```
summary(Datos_Primer_Modelo)
```

```
##   Diabetes_012     HighBP          HighChol        CholCheck
##  Min.   :0.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000
##  Median :0.5   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :0.5   Mean   :0.5461   Mean   :0.5181   Mean   :0.9767
##  3rd Qu.:1.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       BMI            Smoker          Stroke        HeartDiseaseorAttack
##  Min.   :17.00   Min.   :0.000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:25.00   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :29.00   Median :0.000   Median :0.00000   Median :0.0000
##  Mean   :29.73   Mean   :0.483   Mean   :0.06151   Mean   :0.1258
##  3rd Qu.:33.00   3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :75.00   Max.   :1.000   Max.   :1.00000   Max.   :1.0000
##   PhysActivity        Fruits          Veggies       HvyAlcoholConsump
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :0.00000
##  Mean   :0.7224   Mean   :0.6278   Mean   :0.7882   Mean   :0.04298
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##  AnyHealthcare     NoDocbcCost         GenHlth         MentHlth
##  Min.   :0.0000   Min.   :0.00000   Min.   :1.000   Min.   : 0.000
##  1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 0.000
##  Median :1.0000   Median :0.00000   Median :3.000   Median : 0.000
##  Mean   :0.9507   Mean   :0.08991   Mean   :2.792   Mean   : 3.655
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:4.000   3rd Qu.: 2.250
##  Max.   :1.0000   Max.   :1.00000   Max.   :5.000   Max.   :30.000
##     PhysHlth         DiffWalk           Sex              Age
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   : 1.000
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 7.000
##  Median : 0.000   Median :0.0000   Median :0.0000   Median : 9.000
##  Mean   : 5.595   Mean   :0.2358   Mean   :0.4523   Mean   : 8.513
##  3rd Qu.: 5.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:11.000
##  Max.   :30.000   Max.   :1.0000   Max.   :1.0000   Max.   :13.000
##    Education         Income
##  Min.   :1.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:4.000
##  Median :5.000   Median :6.000
##  Mean   :4.935   Mean   :5.714
##  3rd Qu.:6.000   3rd Qu.:8.000
##  Max.   :6.000   Max.   :8.000
```

Realizamos un modelo de arbol con el fin de identificar las variables mas significativas para predecir la variable clase, para ello usamos la función `rpart y variable importance` incluidas dentro de la libreria `library(rpart)`

El resultado es:

```
print(importancia_caracteristicas)
```

```
##              GenHlth               HighBP                  BMI
##          108.3105368           60.1734779           22.1117355
##               Income             PhysHlth            Education
##           21.2272085           19.4283810           10.4582832
##                  Age             DiffWalk             HighChol
```

```
##            6.3791487            1.9632275            1.9340835
## HeartDiseaseorAttack         PhysActivity            CholCheck
##            1.5791959            0.8846393            0.1891840
```

Dividimos nuestro conjunto de datos en 70% para datos de entrenamiento y 30% para datos de prueba, posteriomente Vamos a tomar como predictores a todos las variables que se encuentra en el conjunto de datos a excepción de nuestra variable clase `Diabetes_012`

```
sample.index1 <- sample(1:nrow(Datos_Primer_Modelo)
                        ,nrow(Datos_Primer_Modelo)*0.7
                        ,replace = F)
predictors <-
  colnames(Datos_Primer_Modelo)[-1]

train.data1 <-
  Datos_Primer_Modelo[sample.index1
                      , c(predictors, "Diabetes_012")
                      , drop = FALSE] #muestras seleccionadas para el entrenamiento
test.data1 <-
  Datos_Primer_Modelo[-sample.index1
                      , c(predictors
                          , "Diabetes_012")
                      , drop = FALSE] #muestras seleccionadas para test
```

Entrenamos nuestro modelo con crossvalidation 10 veces y con un tuneLenght de 50 con el fin de determinar el mejor K para nuestro modelo y un reprocesamiento z score

Posterior al entrenamiento de nuestro modelo, hacemos una `confusionMatrix` con el fin de ver los resultados de nuestro modelo, donde se pueden observar los falsos positivos y los positivos positivos, asi mismo ver cual es el # de presición, la sensibilidad, especificidad y demas datos estadisticos que nos permiten determinar que tan confiable es nuestro modelo al ingresar un nuevo dato:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 272  89
##          1 118 282
##
##                Accuracy : 0.728
##                  95% CI : (0.6949, 0.7593)
##     No Information Rate : 0.5125
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.4567
##
##  Mcnemar's Test P-Value : 0.05164
##
##             Sensitivity : 0.6974
##             Specificity : 0.7601
##          Pos Pred Value : 0.7535
##          Neg Pred Value : 0.7050
##              Prevalence : 0.5125
##          Detection Rate : 0.3574
##    Detection Prevalence : 0.4744
##       Balanced Accuracy : 0.7288
```

```
##
##        'Positive' Class : 0
##
```

Retiramos de nuestro modelo las variables predictoras que consideramos que no aportan a nuestro modelo de acuerdo con el modelo de arbol y al resultado de confusionMatrix, se eliminan las variables `"Smoker", "MentHlth", "AnyHealthcare", "NoDocbcCost", "Veggies"`, reentrenamos nuestro modelo con cross validation 5 Veces, con un preprocesamiento z score, con un rendimiento:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 267  87
##          1 123 284
##
##                Accuracy : 0.724
##                  95% CI : (0.6908, 0.7555)
##     No Information Rate : 0.5125
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.4491
##
##  Mcnemar's Test P-Value : 0.01573
##
##             Sensitivity : 0.6846
##             Specificity : 0.7655
##          Pos Pred Value : 0.7542
##          Neg Pred Value : 0.6978
##              Prevalence : 0.5125
##          Detection Rate : 0.3509
##    Detection Prevalence : 0.4652
##       Balanced Accuracy : 0.7251
##
##        'Positive' Class : 0
##
```

Retiramos 5 Variables predictores que consideramos que no aportan a nuestro modelo, variables retiradas `"HvyAlcoholConsump","Fruits","Sex","Stroke","CholCheck"`Se prueba el rendimiento el modelo usando 3 validaciones cruzadas, con 10 repeticiones