Matemáticas Computacionales
Práctica 2: Estudio de una Base Datos.

Profesor: Ángel Isabel Moreno Saucedo
Semestre Febrero - Junio 2021

1. Introducción

2. Base de datos

La base de datos Pima Indian Diabetes se encuentra en la literatura y es la más usada para experimentaciones de algoritmos de aprendizaje máquinas. Esta base datos cuenta con el registro de 768 pacientes femeninos con las 9 características:

1. Número de embarazos.
2. Concentración de glucosa a dos horas de una prueba de tolerancia oral a la glucosa. (mg/dl)
3. Presión arterial diastólica. (mm Hg)
4. Grosor del pliegue de la piel del tríceps. (mm)
5. Concentración de insulina sérica a dos horas de una prueba de tolerancia oral a la glucosa. (mu U/ml)
6. Índice de masa corporal. (kg/m^2)
7. Función pedigrí de diabetes.
8. Edad. (años)
9. Diabetes.

Una de las principales dificultades que se enfrenta al momento de trabajar con una base datos, es que en los registros pueden encontrarse datos faltantes en algunas de las características. La Figura (1) nos enseña como está compuesta la base de datos. Se cuenta con el 91 % de datos registrados teniendo 9 % de registros faltantes, además las características de 2-horas de suero insulina y grosor del pliegue de la piel del tríceps, son los que cuentan con el mayor número de registros faltantes con el 49 % y 30 % respectivamente. Las demas características cuentan con menos del 5 % de registros faltantes o ninguno.

De las caracteríticas donde no hay registros faltantes, la Figura (2) muestra unas gráficas de caja bigotes para la edad, número de embarazos y función pedigrí. La edad de los pacientes Subfigura (2a) se encuentra entre 21 y 81 años con una media de 33 años, y la mitad de los registros se encuentran entre los valores 24 a 41 años, teniendo algunos datos atípicos de entre 70 a 81 años, la Subfigura (2b) muestra la cantidad de embarazos de los registros se encuentra entre 0 y 17 con una media de 4 embarazos, y la mitad de los registros se encuentran entre

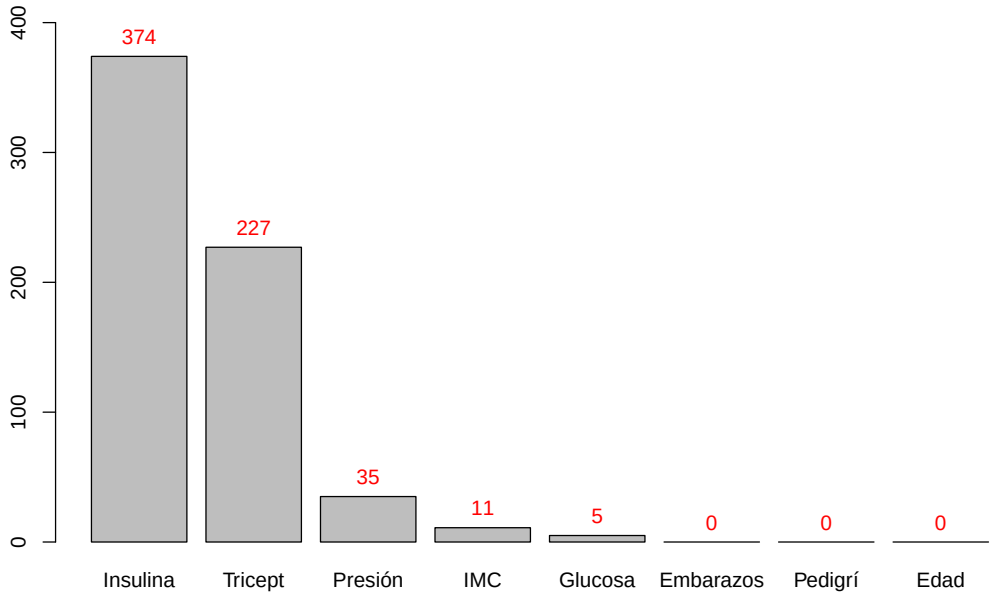


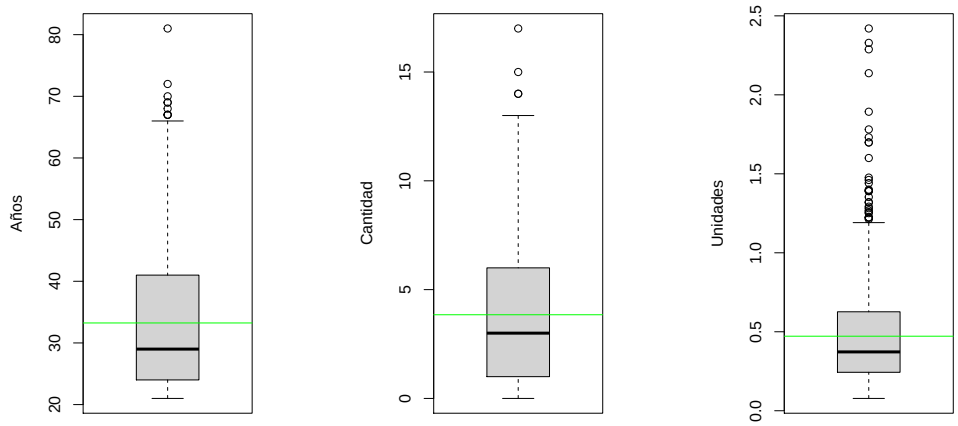
Figura 1: Gráfica de barrras que muestra la cantidad de datos faltantes en cada característica.

el rango de 1 a 6 embarazos y de la característica función pedigrí de diabetes tiene registros con valores entre 0.078 y 2.42 con una media de 0.4719, la mitad de los registros oscilan entre 0.2437 y 0.6262 esto se observa en la Subfigura (2c).

3. Tarea

3.1. Puntos Extra

Referencias



(a) Edad (b) Num. Embarazos (c) Fun. Pedigrí

Figura 2: Diagrama de caja bigotes de la edad, número de embarazos y función pedigrí.

```r
head(PimaIndiansDiabetes, n = 20)

# list types for each attribute
sapply(PimaIndiansDiabetes, class)

# distribution of class variable
y <- PimaIndiansDiabetes$diabetes
cbind(freq=table(y), percentage=prop.table(table(y))*100)

y <- iris$Species
cbind(freq=table(y), percentage=prop.table(table(y))*100)

# summarize the dataset
summary(PimaIndiansDiabetes)

# calculate standard deviation for all attributes
sapply(PimaIndiansDiabetes[,1:8], sd)

# calculate skewness for each variable
library(e1071)
skew <- apply(PimaIndiansDiabetes[,1:8], 2, skewness)
# display skewness, larger/smaller deviations from 0 show more skew
print(skew)
```

Cuadro 1: Código en python del método Monte-Carlo.