Práctica Análisis Exploratorio de Datos

Dataset: Pima Indian

Se utilizará la Pima Indians Diabetes Database (PIDD), cuya propiedad original pertenece al National Institute of Diabetes and Digestive and Kidney Diseases, y los datos fueron obtenidos del UCI Machine Learning Repository - Pima Indians Diabetes Data Set (2016). Las unidades de análisis consistieron en 768 mujeres residentes cerca de Phoenix, Arizona, EEUU, pertenecientes a la etnia Pima y con al menos 21 años de edad.

En ellas fueron registradas 9 variables.

1. Concentración de glucosa plasmática a las 2hs de una prueba de tolerancia oral a la glucosa (G120 mg/dl)

2. Concentración de insulina sérica a las 2hs de una prueba de tolerancia oral a la glucosa (I120 mU/ml)

3. Presión arterial diastólica (PAD mmHg.)

4. Grosor del pliegue de la piel del tríceps (GPPT mm)

5. Índice de masa corporal (IMC= peso /altura al cuadrado= kg/m²)

6. Antecedentes Familiares o función de pedigrí de diabetes (FPD)

7. N° de embarazos (nE)

8. Edad (Edad años)

9. Variable clasificatoria (0 – 1, donde 1 es interpretado como positivo para diabetes).

El diagnóstico estuvo basado en el criterio de la OMS (i.e.: G120 ≥ 200 mg/dl en cualquier examen o evaluación de rutina médica).

En este práctica, les pedimos que analicen esta base de datos a través de un notebook de Jupyter.

1. Crear un notebook utilizando *Markdown*

   Utilizar celdas tipo Markdown para indicar el comienzo de cada uno de los siguientes apartados. Utilizar celdas tipo code para obtener los resultados.

2. Cargar las librerías para análisis de datos, visualización y Machine Learning.

3. Cargar el dataset “Pima Indian” (desde una url, desde un archivo).

   Pima Indian url: https://goo.gl/vhm1eU

4. Describir los datos:
   4.1. Indicar el nombre de los features del dataset con un Markdown para tabla.
   4.2. Indicar el tamaño de la base de datos a analizar.
   4.3. Mostrar las 5 primeras y 5 últimas muestras del dataset.
   4.4. Indicar los tipos de datos que tenemos en el dataset. ¿Qué variables son de tipo categórico? ¿Cuáles son numéricas?
   4.5. Describir ¿qué columnas contienen blank, nulos o empty values?
   4.6. Indicar el número de observaciones del target o salida (variable Outcome).
   4.7. Convertir variable Outcome a tipo numérico.

5. Visualizar el dataset:
   5.1. Para cada variable, represente su histograma.
   5.2. Para cada variable, represente su boxplot o diagrama de cajas.
   5.3. Representa cada variable de entrada en función del target.

6. Generar un heatmap con los features.

7. Normalizar todos los features con media 0 y varianza 1.

8. Separar el dataset en conjunto de train y test (un 20% de muestras para test). Comprobar tamaño del conjunto de train y test.