Parcial II-Práctico

Introducción a Ciencia de Datos
Rafael Dávila Bugarín

Como médicos, queremos determinar si un paciente es propenso a tener o no diabétes. En el Google Drive del curso, podrán encontrar un dataset que deberán entrenar para determinar la probabilidad si un paciente tiene o no diabétes. El dataset tiene la siguiente descripción:

The Diabetes prediction dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information. This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans. Additionally, the dataset can be used by researchers to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes.

Instrucciones: Cada inciso debe de ire en una celda de su notebook, al terminar envíe el notebook al profesor con su nombre completo como nombre del archivo, exámenes después de las 11am no serán revisados.

1. Convierta la variable **gender** en binaria.
2. Grafique un correlation heatmap[1] para ver cómo se relacionan las variables utilizando la librería **Seaborn**. ¿Qué variables a primera vista explican mejor la diabetes?
3. A continuación, entrene 2 modelos usando solo las 2 variables que más se correlacionaron con la salida, uno de regresión logística y uno de árboles de decisión (Training set 80% en ambos casos, profundidad=2 para el árbol, y random_state=0 en ambos casos, escriba estos modelos en celdas distintas y añada títulos de cuál modelo se entrenará). De acuerdo al accuracy score, cuál modelo fue mejor en el devset?
4. ¿Cuántos falsos positivos y falsos negativos hubo para cada modelo (usar la matriz de confusión)?
5. Llegaron nuevos pacientes (base Nuevos_pacientes) determine si tienen o no Diabetes?

Challenge (opcional):
En el drive están la incidencia delictiva para los estados de México. Haga una gráfica dónde x= años y Y=íncidencia delictiva. Debe haber tres líneas, el estado con menor índice delictivo (para escoger el mínimo utilice el año 2010), el estado con mayoríndice delictivo (referencia el año 2010) y l promedio durante los años disponibles.

[1] Tome como base este código: https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e