Karem Hernández Lechuga

Mayo 2024

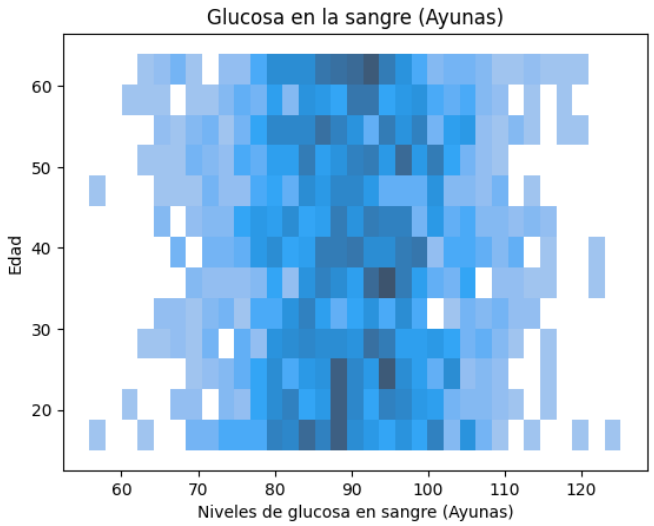
Diplomado en Ciencia de Datos con Python (UAEM)

Estudio sobre posibles indicadores de diabetes

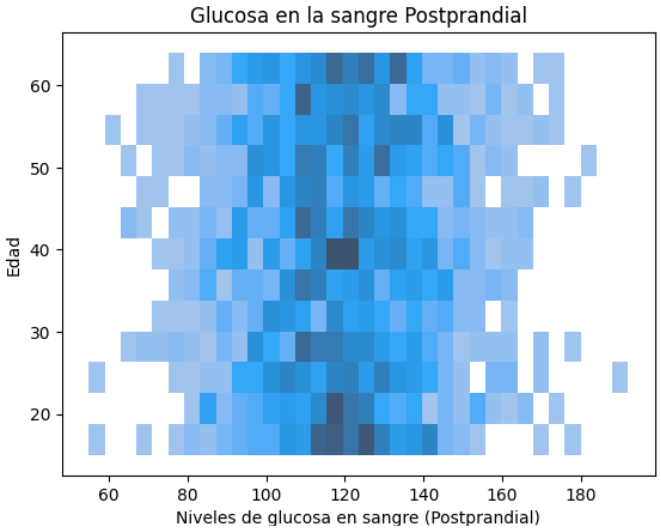
El primer paso para realizar este estudio es obtener o generar los datos a considerar. Dado que es un ejemplo de proyecto de clase dichos datos fueron generados utilizando las herramientas de las librerías de numpy y pandas de Python, la mayoría de ellos mediante la creación de una distribución normal con una media y desviación estándar previamente dadas.

Después de generar los datos correspondientes (aleatorios) se procedió al análisis de ciertas relaciones entre ellos, se consideraron solo pacientes de 15 a 65 años de edad.

1. Niveles de glucosa en ayunas



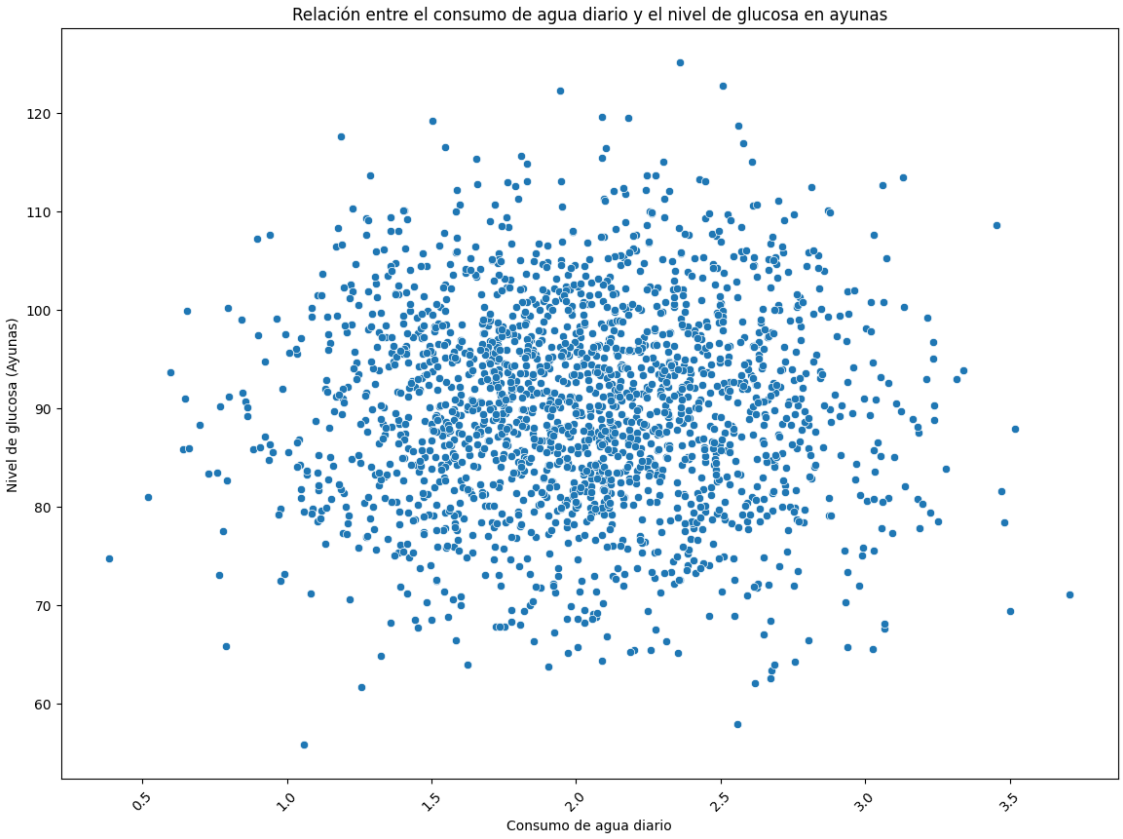
2. Niveles de glucosa después de comida.



Los gráficos anteriores muestran que la distribución de los niveles de glucosa es en general imparcial para la edad de los pacientes, tanto en los niveles en ayunas como después de comer. Esto se debe a que los datos fueron generados aleatoriamente, de otra forma, es posible que

observáramos un sesgo hacia una población de edad mayor con tendencia a un mayores niveles de glucosa en sangre.

3. Relación de nivel de glucosa en ayunas y el consumo de agua diario:



De la misma forma, el gráfico de la sección 3 muestra una distribución aleatoria, cuya media de 2 L de consumo de agua diario puede apreciarse al igual que la media de 90 para el nivel de glucosa en la sangre de mi población de estudio.

La siguiente parte del estudio consistió en entrenar un modelo de tipo Random Forest. De acuerdo a éste se obtuvieron los siguientes resultados, utilizando como variables objetivo los niveles de glucosa en sangre en ayunas y después de comer (postprandial).

Niveles de glucosa en ayunas:

Error cuadrático medio (del entrenamiento): 14.897252443021694
Error cuadrático medio (de la prueba): 119.06687542545912
Coeficiente de determinación (R^2): -0.048585419117831874

De acuerdo a este resultado, el error cuadrático es mas o menos bueno (es decir, no tan grande) en el caso del entrenamiento, sin embargo, en la prueba obtenemos un valor mayor, es decir, los datos no se ajustan con precisión al modelo creado.

Las 5 características más importantes resultaron ser:

1. Frecuencia cardíaca en reposo: 0.04775085750962274
2. Altura: 0.040744811549114246
3. Circunferencia de cadera: 0.03847190054452689
4. Circunferencia de cintura: 0.038395590968800984
5. Historial médico familiar: 0.03792668708872702

Niveles de glucosa postprandial:

Error cuadrático medio (del entrenamiento): 58.69241023333163
Error cuadrático medio (de la prueba): 355.8855910348237
Coeficiente de determinación (R^2): -0.028199084258927476

De manera muy similar al análisis de los niveles de glucosa en ayunas, , el error cuadrático es mas o menos bueno (es decir, no tan grande) en el caso del entrenamiento, sin embargo, en la prueba obtenemos un valor mucho mayor, es decir, los datos no se ajustan con precisión al modelo creado.

Las 5 características más importantes para este análisis resultaron ser:

1. Niveles de colesterol (LDL): 0.04156057667268993
2. Circunferencia de cintura: 0.03961014514485462
3. Consumo de tabaco: 0.03733485482797581
4. Consumo de cafeína: 0.035358353985773985
5. Frecuencia cardíaca en reposo: 0.03512224801569416

Una observación interesante es que en ambos un indicador parece ser la frecuencia cardiaca en reposo y la circunferencia de la cintura.

Pensando en estos resultados, considero que sería mejor hacer una análisis de datos reales, pues podría existir una relación entre las variables consideradas y posiblemente obtendríamos valores de ajuste mejores para el modelo dado.