Proyecto 1 (10%)
Probabilidad y Estadística
Bioingeniería
20252

Contexto

Los datos anexos sirven como un recurso valioso para explorar las complejas dinámicas de la salud del corazón y sus predictores. Los ataques cardíacos, o infartos de miocardio, siguen siendo un problema importante de salud a nivel mundial, lo que hace necesario comprender mejor sus precursores y posibles factores de mitigación. Este conjunto de datos incluye diferentes variables, como la edad, los niveles de colesterol, la presión arterial, los hábitos de tabaquismo, los patrones de ejercicio, las preferencias dietéticas y más, con el objetivo de esclarecer la compleja interacción de estas variables en la determinación de la probabilidad de un ataque cardíaco.

Mediante el uso de analítica predictiva y aprendizaje automático en este conjunto de datos, investigadores y profesionales de la salud pueden trabajar en el desarrollo de estrategias proactivas para la prevención y el manejo de las enfermedades cardíacas. El conjunto de datos se erige como un testimonio de los esfuerzos colectivos por mejorar nuestra comprensión de la salud cardiovascular y allanar el camino hacia un futuro más saludable.

El conjunto de datos, compuesto por **8.763 registros de pacientes de todo el mundo**, culmina en una característica de **clasificación binaria** que indica la presencia o ausencia de ataque cardíaco, proporcionando así un recurso integral para el análisis predictivo y la investigación en salud cardiovascular.

Variables:

- Patient ID – Identificador único de cada paciente
- Age – Edad del paciente
- **Sex** – Género del paciente (Masculino/Femenino)
- **Cholesterol** – Niveles de colesterol del paciente
- Blood Pressure – Presión arterial del paciente (sistólica/diastólica)
- **Heart Rate** – Frecuencia cardíaca del paciente
- Diabetes – Si el paciente tiene diabetes (Sí/No)
- Family History – Antecedentes familiares de problemas cardíacos (1: Sí, 0: No)
- **Smoking** – Estado de tabaquismo del paciente (1: Fumador, 0: No fumador)
- Obesity – Estado de obesidad del paciente (1: Obeso, 0: No obeso)
- Alcohol Consumption – Nivel de consumo de alcohol del paciente (Nulo/Leve/Moderado/Alto)
- **Exercise Hours Per Week** – Número de horas de ejercicio por semana
- Diet – Hábitos alimenticios del paciente (Saludable/Promedio/Poco saludable)

- Previous Heart Problems – Problemas cardíacos previos del paciente (1: Sí, 0: No)
- Medication Use – Uso de medicamentos por parte del paciente (1: Sí, 0: No)
- Stress Level – Nivel de estrés reportado por el paciente (1-10)
- Sedentary Hours Per Day – Horas de actividad sedentaria por día
- Income – Nivel de ingresos del paciente
- **BMI** – Índice de Masa Corporal (IMC) del paciente
- Triglycerides – Niveles de triglicéridos del paciente
- Physical Activity Days Per Week – Días de actividad física por semana
- **Sleep Hours Per Day** – Horas de sueño por día
- Country – País del paciente
- Continent – Continente donde reside el paciente
- Hemisphere – Hemisferio donde reside el paciente
- **Heart Attack Risk** – Ataque cardiaco (1: Sí, 0: No)

Procedimiento (el proyecto solo se hará con las variables en negrita)

- **(20%)** Realice un análisis exploratorio de las variables de interés. Para esto:
    - Determine la presencia de datos atípicos usando un diagrama de cajas y bigotes.
    - Calcule los valores de media y mediana de cada variable.
    - En el caso de haber datos ausentes, elimine estas filas, o reemplácelos por la media o la mediana.
    - Describa el tipo de distribución que siguen las variables.
    - Determine para cada variable si los datos presentan variabilidad alta, moderada o baja.
    - Realice una descripción de todo lo anterior.
- **(70%)** Determine para las variables numéricas, si existe diferencia estadística entre, hombres y mujeres, entre fumadores y no fumadores, y entre personas con riesgo de infarto y sin riesgo. Para eso: **(cada ítem debe estar acompañado de su descripción, no solo es código)**
    - (5%) Elija una prueba estadística paramétrica.
    - (5%) Plantee las hipótesis de la prueba, en términos del contexto específico (esto se hace para todas las pruebas que harán)
    - (10%) Corrobore los supuestos necesarios para la prueba paramétrica seleccionada.
    - (20%)Si no se cumple el supuesto de normalidad, realice una transformación apropiada. Recuerde que la transformación box-cox no es exitosa para variables con distribución uniforme, entonces en ese caso se recomienda usar una transformación diferente, use el siguiente código como guía para esto.

```
from sklearn.preprocessing import QuantileTransformer

QT = QuantileTransformer(output_distribution="normal")

T_obesos_t_QT = QT.fit_transform(T_obesos.values.reshape(-1,1))

T_no_obesos_t_QT = QT.fit_transform(T_no_obesos.values.reshape(-1,1))
```

En el código anterior, T_obesos y T_no_obesos, son variables que deseamos comparar, y que tienen una distribución uniforme, por lo cual no cumplen el supuesto de normalidad, y no funciona una transformación box-cox. Entonces aplicamos una transformación diferente, usando la clase QuantileTranformer de SciKit-Learn, indicándole que esperamos una distribución de salida de tipo normal (*output_distribution* = "*normal*"). Nótese que se debe hacer un reshape de la variable, dado que no puede ser unidimensional.

- (10%) Corrobore los supuestos luego de la transformación
- (20%) Una vez Cumplidos los supuestos, aplique la prueba. Construya matrices donde guarde los resultados de valor P de las diferentes pruebas hechas, conviértalas en Dataframes e imprímalas. Esto para cada set de comparaciones (según sexo, según si fuman, y según si tiene riesgo de infarto), es decir son tres matrices.

	Cholesterol	Heart Rate	...	Sleep hours
Cholesterol		p-value	...	p-value
Heart Rate	p-value		p-value	p-value
...	p-value	p-value		p-value

Sleep hours	p-value	p-value	p-value	

- (10%) Saque las conclusiones del análisis.

Entregable

Deberá entregar un archivo tipo notebook (.ipybn) con el desarrollo de la actividad. **Sea muy organizado**, haga uso de celdas de texto y de código de forma separada para **mostrar paso a paso lo que hace y lo que significa cada resultado para finalmente llegar a una conclusión**. Lo más importante del proyecto es la interpretación de los resultados, por lo que se insta a no ser muy escueto en las descripciones, de lo contrario se verá reflejado en el desempeño del trabajo. Procure no realizar códigos muy largos, es preferible hacer códigos cortos, he ir analizándolos uno a uno en celdas de texto independientes.