Prueba - Análisis estadístico con Python

En esta prueba validaremos los conocimientos adquiridos durante este módulo..

Lee todo el documento antes de comenzar el desarrollo **individual**, para asegurarte de tener el máximo de puntaje y enfocar bien los esfuerzos.

Tiempo asociado: 2 horas cronológicas

Descripción

La OMS estima que cada año se producen 12 millones de muertes en todo el mundo debido a enfermedades cardíacas, producidas en su gran mayoría por enfermedades cardiovasculares. El pronóstico temprano de las enfermedades cardiovasculares puede ayudar en la toma de decisiones sobre cambios en el estilo de vida en pacientes de alto riesgo y reducir las complicaciones.

La base de datos **cardio.csv** contiene mediciones realizadas a pacientes hace 10 años, entre los cuales se encuentran

- Sex: M - F
- Age: edad
- Education: codificada, considerando que un valor más alto corresponde a un mayor nivel educativo.
- currentSmoker: si el paciente es o no fumador actual
- cigsPerDay: el número de cigarrillos que la persona fumaba en promedio en un día.
- BPMeds: si el paciente estaba tomando medicamentos para la presión arterial (0: No, 1: Sí)
- prevalenStroke: si el paciente había tenido un accidente cerebrovascular previamente (0: No, 1: Sí)
- prevalentHyp:: si el paciente era hipertenso o no (0: No, 1: Sí)
- diabetes: si el paciente tenía diabetes o no (0: No, 1: Sí)
- totChol: nivel de colesterol total
- sysBP: presión arterial sistólica
- diaBP_BMI: presión arterial diastólica
- BMI: Índice de masa corporal
- heartRate: : ritmo cardíaco
- glucose: nivel de glucosa

- TenYearCHD: se indica si el paciente sufrió o no una enfermedad coronaria en los últimos 10 años.(0: No, 1: Sí)

A partir de estos datos, buscaremos realizar algunas descripciones predictivos.

1. Carga los datos y explóralos. Elimina los datos nulos o incorrectos
2. Escoge tres variables cuantitativas y realiza un análisis descriptivo de ellas, utilizando indicadores y gráficos. ¿Cuaĺ presenta mayor dispersión?
3. Elimina los datos atípicos del dataset. Para las siguientes preguntas, considera el dataset "limpio"
4. ¿Qué variables cuantitativas presentan mayor correlación? Explica.
5. Si de este dataset se escoge un paciente al azar, ¿cuál es la probabilidad de que sea hombre, si se sabe que pertenece al mayor cuartil de la variable correspondiente al índice de masa corporal?
6. ¿Es razonable afirmar que, para una persona cualquiera (no necesariamente dentro del dataset), su ritmo cardiaco promedio es 75? Explica y justifica.
7. Ser hombre, ¿influye en el promedio de cigarrillos consumidos por día, dentro de los pacientes fumadores? Explica.
8. Construye un modelo de regresión lineal que permita relacionar 6 variables del dataset con el índice de masa corporal. (Debes incluir al menos dos variables cualitativas). Evalúa tu modelo y explica.
9. Construye un modelo de regresión logística para predecir el riesgo de sufrir una enfermedad coronaria en los próximos diez años, a partir de las variables descritas. Verifica el balanceo de datos y evalúa tu modelo.
10. Separa los modelos de regresión anteriores en dos distintos, respectivamente, considerando alguna variable categórica. Compara y concluye.

Requerimientos

Dentro del archivo de Jupyter Notebook debes ir ejecutando las siguientes acciones y explicar lo que estás haciendo:

1. Carga, explora y prepara datos para su análisis, utilizando las funciones correspondientes. (**2 Puntos**)
2. Analiza datos, calculando e interpretando diversos indicadores estadísticos y probabilidades. **(2 puntos)**
3. Plantea y realiza pruebas de hipótesis e inferencias, interpretando correctamente sus resultados **(3 puntos)**
4. Plantea y aplica modelos de regresión, e interpreta sus métricas de evaluación (**3 Puntos**)

😊 **¡Mucho éxito!**

Consideraciones y recomendaciones

- Debes entregar tu trabajo en un archivo de Jupyter Notebook, con todo el código y las explicaciones respectivas para desarrollar tu trabajo.