Introducción a la Estadística y Ciencia de Datos - Segundo cuatrimestre 2021

Acerca de los conjuntos de datos de la Práctica 1

Estudiante: sea indulgente con las traducciones brindadas, quisimos alivianarle el trabajo de que usted traduzca pero las docentes de las materias no somos traductoras.

1. **Biomarcadores urinarios para el cáncer de páncreas**. El archivo `Debernardi.csv` contiene un conjunto de datos de un artículo, ambos de acceso abierto, publicado el 10 de diciembre de 2020. La cita es

   Antecedentes

   El cáncer de páncreas es un tipo de cáncer extremadamente mortal. Una vez diagnosticado, la tasa de supervivencia a los cinco años es inferior al 10 %. Sin embargo, si el cáncer de páncreas se detecta temprano, las probabilidades de sobrevivir son mucho mejores. Desafortunadamente, muchos casos de cáncer de páncreas resultan asinomáticos hasta que el cáncer se haya diseminado por todo el cuerpo. Una prueba de diagnóstico para identificar a las personas con cáncer de páncreas podría ser de gran ayuda.

   Artículo

   En el artículo de Silvana Debernardi y colegas, publicado en 2020 en la revista PLOS Medicine, un equipo multinacional de investigadores buscó desarrollar una prueba de diagnóstico para el tipo más común de cáncer de páncreas, llamado adenocarcinoma pancreático ductal o ACPD. Recogieron una serie de biomarcadores de la orina de tres grupos de pacientes: control, con enfermedades del páncreas no cancerígenas y con ACPD. El objetivo era desarrollar una forma precisa de identificar a los pacientes con cáncer de páncreas.

   Variables del conjunto de datos

   - sample_id: código que identifica a cada paciente participante del estudio.
   - patient_cohort: cohort 1, muestras usadas previamente; cohort 2, muestras agregadas para este estudio.
   - sample_origin: BPTB, Barts Pancreas Tissue Bank, Londres, Reino Unido; ESP: Spanish National Cancer Research Centre, Madrid.
   - age: edad en años.
   - sex: sexo identificando con F al sexo femenino y con M al masculino.
   - diagnosis: diagnóstico (1 = Control, 2 = Benigno, 3 = ACPD), 1 = Control (sin enfermedad pancreática); 2 = enfermedad hepatobiliar benigna (119 de las cuales son pancreatitis crónica); 3 = adenocarcinoma pancreático ductal, es decir, cáncer de páncreas.
   - stage: estadío del cáncer en pacientes que lo sufran, uno de los siguientes: IA, IB, IIA, IIIB, III, IV.
   - benign_sample_diagnosis: para aquel/la cuyo diagnóstico es no canceroso, ¿cuál es el diagnóstico?
   - plasma_CA19_9: niveles en plasma sanguíneo, en U/ml, de anticuerpo monoclonal CA 19-9 que a menudo está elevado en pacientes con cańcer de páncreas.

- creatinine: biomarcador urinario de función renal, en mg/ml.
- LYVE1: niveles urinarios, en ng / ml, del receptor 1 de hialuronato linfático endotelial, una proteína que puede desempeñar un rol en la metástasis tumoral.
- REG1B: niveles urinarios de una proteína que puede estar asociada con la regeneración del páncreas, en ng/ml.
- TFF1: los niveles urinarios de factor trefoil 1, que pueden estar relacionados con la regeneración y reparación del tracto urinario, en ng / ml.
- REG1A: niveles urinarios, en ng / ml, de una proteína que puede estar asociada con la regeneración del páncreas. Solo se evaluó en 306 pacientes (uno de los objetivos del estudio fue evaluar REG1B frente a REG1A).

2. El archivo `datos_titanic.csv` contiene información sobre una muestra seleccionada al azar de personas, no tripulantes, que viajaban en el barco tristemente célebre *Titanic*, al momento de su hundimiento en el Océano Atlántico.

   Variables del conjunto de datos

   - PassengerId: número natural para identificar la cantidad de pasajeras/os.
   - Survived: toma los valores $\{0, 1\}$, el valor 0 indica que la persona no sobrevivió al hundimiento y el 1 indica que se salvó.
   - Pclass: toma los valores $\{1, 2, 3\}$, indicando la clase del ticket (1 representa primera clase, 2 representa segunda clase y 3 representa tercera clase).
   - Name: nombre.
   - Sex: sexo, "female"para el femenino y "male"para el masculino.
   - Age: edad en años.
   - SibSp: cantidad de hermanas/os o cónyuges a bordo.
   - Parch: cantidad de padres o hijos a bordo.
   - Ticket: número de ticket.
   - Fare: tarifa del ticket.
   - Cabin: identificación de la cabina.
   - Embarked: puerto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton).

3. El archivo `Islander_data.csv` contiene los datos referentes a un experimento sobre los efectos de los medicamentos contra la ansiedad en la recuperación de la memoria al estar estimulado con recuerdos felices o tristes. Las y los participantes se seleccionaron sobre nuevos isleños. Medicamentos de interés (conocidos como) [Dosis 1, 2, 3]:

   A - Alprazolam (Xanax, a largo plazo) [1 mg / 3 mg / 5 mg]

   T - Triazolam (Halcion, a corto plazo) [0,25 mg / 0,5 mg / 0,75 mg]

   S - Tableta de azúcar (placebo) [1 / 2 / 3]

   - Las dosis siguen una proporción de 1:1 para garantizar la validez.

- Los recuerdos felices o tristes se prepararon 10 minutos antes de la prueba.
- Las y los participantes hicieron la prueba todos los días durante una semana para imitar la adicción.

Preguntas que quiere responder el estudio

¿Cómo afectan los medicamentos contra la ansiedad de manera diferente según la edad? ¿Existe un nivel de estancamiento en la efectividad de los medicamentos contra la ansiedad? De ser así, ¿en cuál momento? ¿Cuál es el efecto de la medicina contra la ansiedad sobre la recuperación de la memoria? ¿Y la eficacia de los placebos en un entorno de prueba?

Variables del conjunto de datos

- first_name: nombre
- last_name: apellido
- age: edad en años.
- Happy_Sad_group: H indica que participó en el grupo de recuerdos felices y S en el de los tristes.
- Dosage: dosis, toma los valores $\{1, 2, 3\}$.
- Drug: droga brindada (A, T o S según se indica más arriba).
- Mem_Score_Before: el tiempo, en segundos, que tomó terminar una prueba de memoria antes de la exposición a la droga.
- Mem_Score_After: el tiempo, en segundos, que tomó terminar una prueba de memoria después de que se logró la adicción.
- Diff: diferencia entre las variables Mem_Score_After y Mem_Score_Before (Diff=Mem_Score_After-Mem_Score_Before).