Actividad. Estructuras y formato de los datos

Unidad	3
Entrega	Documento en Jupyter Notebook

1. Enunciado

La actividad está organizada en ejercicios y se evalúa sobre un máximo de 10 puntos. Todos los ejercicios tienen el mismo peso en la evaluación de la actividad. Un ejercicio puede contener varias preguntas o apartados, en cuyo caso la puntuación del ejercicio se repartirá de forma equitativa entre las preguntas o apartados que lo componen.

1. Utiliza el siguiente enlace para descargar el *Heart Disease Dataset*, junto con su fichero de nombres:

<https://archive.ics.uci.edu/ml/datasets/heart+disease>

   a. Utiliza un Jupyter Notebook y el paquete Pandas para abrir el archivo y presentarlo en formato DataFrame, donde el nombre de las columnas debe corresponder con el nombre real de las variables.

   b. Analizando el DataFrame, señala las variables numéricas. Entre las variables numéricas señala cuáles son continuas y cuáles discretas.

   c. Analizando el DataFrame, señala las variables categóricas. Entre las variables categóricas señala cuáles son ordinales y cuáles nominales.

   d. ¿Hay valores faltantes (*missing values*) en el dataset? ¿Qué símbolo se utiliza para marcarlos? ¿Se utiliza más de un tipo de símbolo?

   e. Utilizando métodos de Python, obtén la lista de valores posibles que puede tomar la variable "thal". ¿Se pueden ordenar estos valores en una escala de mayor a menor? ¿Podríamos considerarlos valores categóricos ordinales?

2. Utiliza el siguiente enlace para descargar el Census Income Dataset, junto con su fichero de nombres:

<https://archive.ics.uci.edu/ml/datasets/Census+Income>

   a. Utiliza un Jupyter Notebook y el paquete Pandas para abrir el archivo y presentarlo en formato DataFrame, donde el nombre de las columnas debe corresponder con el nombre real de las variables. Señala qué variables del DataFrame presentan valores faltantes (*missing values*).

   b. Analizando el DataFrame, señala las variables numéricas y categóricas, especificando si son continuas, discretas, ordinales o nominales.

   c. Utilizando métodos de Python, deduce si hay registros repetidos en el dataset. ¿Cuántos elementos hay en el DataFrame original? ¿Cuántos elementos repetidos hay? ¿Cuántos elementos únicos?

© Copyright Universidad Europea. Todos los derechos reservados

d. Utilizando métodos de Python, obtén un DataFrame llamado df_dup en el que estén los elementos repetidos del DataFrame original. ¿Cuántos de ellos tienen un país de origen indeterminado? ¿Cuál es la proporción entre mujeres y hombres? ¿Cuántos cobran más de 50 mil dólares?

e. Utilizando métodos de Python, obtén un DataFrame llamado df_uni en el que estén los elementos únicos del DataFrame original. ¿Cuántos tienen un país de origen indeterminado? ¿Cuántos tienen un tipo de trabajo indeterminado? ¿Cuántos tienen ambas variables indeterminadas a la vez? De entre quienes tienen un país de origen y un tipo de trabajo ambos indeterminados, ¿cuál es la proporción entre hombres y mujeres? ¿Qué porcentaje de ellos cobra menos de 50 mil dólares?

f. Busca en internet el significado de la variable "fnlwgt". ¿Qué representa? Conociendo su significado, ¿crees que tiene sentido, desde un punto de vista de Machine Learning, eliminar los registros duplicados en este dataset?

g. ¿Qué dificultades has encontrado a la hora de trabajar con los valores de las variables de este dataset? ¿Sería apropiado buscar una manera automática, usando métodos de Python, de eliminar esa dificultad o dificultades? ¿Se te ocurre cómo hacerlo? (En este apartado no hace falta escribir código, basta con responder en texto detallando las ideas).

3. Utiliza el siguiente enlace para descargar el conjunto de datos del Portal de Datos Abiertos del Ayuntamiento de Madrid sobre accidentes de bicicletas correspondiente al año 2019. Utiliza un Jupyter Notebook y el paquete Pandas para abrir el archivo y presentarlo en formato DataFrame, donde el nombre de las columnas debe corresponder con el nombre real de las variables.

<https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=20f4a87ebb65b510VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>

a. Analizando el DataFrame, señala las variables numéricas y categóricas, especificando si son continuas, discretas, ordinales o nominales.

b. ¿Hay valores faltantes en el dataset? Especifica las variables que presentan valores faltantes y cuántos de ellos hay en cada variable.

c. ¿Cuántos valores distintos toma la variable 'lesividad'? Especifica esos valores. ¿Qué tipo de variable es la variable 'lesividad'?

d. Ve al PDF de descripción de los datos, que se encuentra en el mismo enlace aportado en el enunciado. Busca el significado de los valores de lesividad. Explica los significados para cada uno de los valores posibles. ¿Qué tipo de variable consideras que puede ser la variable 'lesividad'? ¿Tu respuesta en este apartado es distinta a tu respuesta del apartado anterior? ¿Cómo codificarías esta variable a la hora de utilizar este DataFrame como entrada de un modelo de Machine Learning?

© Copyright Universidad Europea. Todos los derechos reservados

e. Sin tener en cuenta los valores faltantes, ¿qué proporción de los accidentes registrados presentaban un test de alcohol positivo?

2. Detalles de la entrega

- Las respuestas de la actividad se deberán entregar en un Jupyter Notebook en el que se haya respondido a cada apartado en una celda independiente, en el orden de las preguntas de este documento. El código de cada celda se debe poder ejecutar para comprobar las respuestas aportadas. Las preguntas teóricas se deben responder en una celda de formato texto.
- Subir de forma individual el documento a la actividad en el campus virtual.

© Copyright Universidad Europea. Todos los derechos reservados