Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Análisis Inteligente de Datos

Trabajo Práctico N° 1 Obligatorio

Descripción de la base de datos

1. Nombre de la base de datos: ML Marathon Dataset by Azure Developer Community
2. Año de publicación: 2022
3. Fuente /Responsable/Propietarios de la base de datos: Microsoft Azure
4. Base de datos abierta (sí/no): sí
5. Tipo de archivo (matriz/no matriz): matriz
6. Formato del archivo/s: csv
7. Tipo de datos (multivariado/univariado/serie de tiempo/texto/otro): multivariado
8. Información de los atributos: [Data card]

Consignas:

1. Para la base de datos seleccionada genere una muestra aleatoria estratificada y balanceada por “depósito” de tamaño n = 2000 utilizando como semilla los últimos tres dígitos del DNI/PASAPORTE. Guarde los datos en un archivo y realice todo el trabajo práctico con la muestra generada.
2. Realice un análisis estadístico de cada una de las variables numéricas para cada valor de depósito. Presente la información en forma tabular y conteniendo las siguientes medidas descriptivas: Cantidad de datos, mínimo, máximo, media, mediana, moda, varianza, desviación estándar, coeficiente de variación, cuartil 1, cuartil 3, rango intercuartílico, MAD, asimetría, curtosis.
3. Represente gráficamente cada variable numérica eligiendo el gráfico que considere apropiado. Considere la posibilidad de generar rangos de datos para su análisis y representación gráfica de las variables.

4. Presente una tabla de frecuencias y porcentaje para la variable “marital” (estado civil) según el nivel de la variable “deposit”.

5. Realice un gráfico para representar la tabla construida en el punto 4.

6. Elija dos variables ~~continuas~~ numéricas, establezca rangos que representen distintos niveles de cada una y defina nuevas variables categóricas. Aplique un test adecuado para entender si existe asociación entre ambas. Utilice un nivel de significación del 5%.

7. Seleccione la variable “education” y elija otra variable categórica. Aplique un test adecuado para entender si existe asociación entre ambas. Utilice un nivel de significación del 5%.

8. Seleccione otra variable continua y estime la diferencia de medias según el valor de la variable “deposit” con un nivel de confianza del 95%. Interprete el resultado obtenido.

9. Según el resultado obtenido en el punto ~~7.-~~ 8 realice un test de hipótesis apropiado para determinar la diferencia de medias de la variable en estudio. Trabaje con una significación del 5%. Presente el planteo de hipótesis adecuado, la resolución y la decisión a tomar.

10. Seleccione una muestra de 30 elementos estratificada según la variable “deposit”. ¿Se puede afirmar que hay diferencias significativas en el balance de los que realizaron el depósito respecto a aquellos que no lo hicieron? Elija un test de hipótesis adecuado. Trabaje con una significación del 5%.

11. Decida si existen diferencias significativas en la duración respecto los niveles de educación (“secondary”, “tertiary”, “primary”, “unknown”). Justifique. Utilice un test adecuado. Realice las pruebas necesarias para comprobar los supuestos. Trabaje con una significación del 5%.

12. Elija dos variables cuantitativas, determine la variable explicativa y la variable explicada. Encuentre la ecuación de la recta de regresión lineal que explique la relación entre las variables elegidas. Escriba conclusiones acerca de la significatividad del modelo aplicado. Puede acompañar el modelo de un gráfico adecuado.

13. Presente un informe final con un mínimo de 500 y un máximo de 800 palabras del análisis de la base de datos, describiendo la base de datos, indicando la presencia de valores atípicos y las conclusiones a las que se abordó luego del análisis.

Fecha de entrega: 15 de mayo antes de las 23.55 hs.

Formato de entrega: **código en R** (.Rmd y .html) e **informe** (.doc/.pdf). La entrega se realiza con el nombre “Apellido_nombre_TP1_AID.” en “Entrega Trabajo Práctico 1” en Aula Virtual.