Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Análisis Inteligente de Datos

Trabajo Práctico N° 1

Análisis Univariado de Datos

Descripción de la base de datos

1. Nombre de la base de datos: Conjunto de datos de calidad del vino
2. Área/Tema: Negocios
3. Link de la base: https://archive.ics.uci.edu/ml/datasets/wine+quality
4. Fecha de publicación: 2010-10-07
5. Fecha de relevamiento de los datos: mayo 2004 – febrero 2007
6. Fuente /Responsable/Propietarios de la base de datos: Paulo Cortez, Universidad de Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez. A. Cerdeira, F. Almeida, T. Matos y J. Reis, Comisión de Viticultura de la Región del Vino Verde (CVRVV), Oporto, Portugal
7. Base de datos abierta (sí/no): sí
8. Tipo de archivo (matriz/no matriz): matriz
9. Formato del archivo/s (cvs/xlsx/txt/zip/otro (especificar)): xlsx
10. Tipo de datos (multivariado/univariado/serie de tiempo/texto/otro): multivariado
11. Características del atributo (categórico/numérico/mixto): numérico
12. Información de los atributos:

Variables de entrada (basadas en pruebas fisicoquímicas):
1 - acidez fija
2 - acidez volátil
3 - ácido cítrico
4 - azúcar residual
5 - cloruros
6 - anhídrido sulfuroso libre
7 - anhídrido sulfuroso total
8 - densidad

Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Análisis Inteligente de Datos

9 - pH
10 - sulfatos
11 - alcohol
Variable de salida (basada en datos sensoriales):
12 - calidad (puntuación entre 0 - muy malo y 10 - excelente)
13 - variedad (blanco 1 / tinto 2)

Consignas:

1.- Para la base de datos seleccionada genere una muestra aleatoria estratificada y balanceada por variedad de vino de tamaño n = 2000 utilizando como semilla los últimos tres dígitos del DNI/PASAPORTE.

2.- Realice un análisis estadístico de cada una de las variables numéricas para cada variedad de vino. 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 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 calidad de vino según la variedad (tinto – blanco).

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

6.- Elija dos variables continuas, 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 otra variable continua y estime la diferencia de medias según la variedad del vino con un nivel de confianza del 95%. Interprete el resultado obtenido.

8.- Según el resultado obtenido en el punto 7.- 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.

Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Análisis Inteligente de Datos

9.- ¿Se puede afirmar que hay diferencias significativas en la calidad del vino tinto respecto al vino blanco? Elija un test de hipótesis adecuado. Trabaje con una significación del 5%.

10.- Decida si existen diferencias significativas en las proporciones de alcohol entre los vinos de calidad baja, media y alta. Justifique.

11.- 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: 14 de mayo antes de las 9 hs.

Formato de entrega: archivo en formato.xlsx. Si lo hacen en R: código en R, archivo Rmd y archivo.html. Si lo hacen en Python entregar el archivo en Python y el informe en un documento de Word o el Colab completo en caso de usar Google Colaboratoy. La entrega se realiza con el nombre “Apellido_nombre_TP1_AID.” En Entrega Trabajo Práctico 1 en Aula Virtual.