M2.851 TIPOOGÍA Y CICLO DE VIDA DE LOS DATOS – Práctica 2

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o cualquier dataset libre disponible en Kaggle (https://www.kaggle.com). Las diferentes tareas a realizar (y justificar) son las siguientes:

Se utiliza dataset de www.kaggle.com, sobre un tema de salud al igual que en la práctica 1.

Enlace: https://www.kaggle.com/angelmm/healthteethsugar

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?
   Este dataset es importante porque busca comprender mejor el estado de salud dental de la población.
   En este caso se plantea si a los países con mayor riqueza les supone un empeoramiento de la salud dental, porque consumen más dulces, o si al contrario les representa una ventaja, por disponer de mejor asistencia sanitaria o incluso de un mejor nivel educativo.
2. Limpieza de los datos.
   a. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?
   b. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?

   Los campos más relevantes son el consumo de azúcar, el número de dientes estropeados en una población, el nivel de riqueza y los gastos en salud gubernamentales en salud durante distintos años por países.

   Como puede verse en los resultados del trabajo hay muchos elementos vacíos, ya que hay países que no tienen unos datos y otros, hay años que no se tienen datos, etc… es un volumen muy amplio en años y faltan bastantes datos, en este caso nos ajustamos a los disponibles.

   Hay valores extremos, por ejemplo, en los niveles de riqueza o PIB de los países, que por nuestro conocimiento más que una anomalía se ajustan a la realidad.

   Ver detalle en el informe del trabajo

3. Análisis de los datos.
   a. Selección de los grupos de datos que se quieren analizar/comparar.
   b. Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.
   c. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

   Ver detalle en el informe del trabajo. Por ejemplo, el caso del nivel de riqueza del país, expuesto con el PIB (GDP), hay mucha dispersión de los datos, los valores de los datos son elevados para su manejo, en ese caso interesa estandarizar y normalizar esta variable.

Se ha probado especialmente la regresión linear múltiple con diversas combinaciones de variables. Hubiera convenido disponer de más tiempo para probar modelos adicionales

4. Representación de los resultados a partir de tablas y gráficas.

   Ver detalle en el informe del trabajo

5. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

   No se está satisfecho con los resultados obtenidos como para dar una repuesta de confianza al problema, los coeficientes de bondad del modelo nos dan valores muy bajos, se debería profundizar en el análisis.

6. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

   Ver detalle en el informe del trabajo