PR UD 3. APRENDIZAJE NO SUPERVISADO - KMEANS

Descripción de la tarea

Realizar estudios exploratorios de los datos usando análisis cluster y empleando el algoritmo de K-means o K-medias. Para ello habrá que:

- Preprocesar los datos. En esta etapa se seleccionarán las variables que sean relevantes, detectará los outliers y observaciones relevantes.
- Análisis cluster. Se determinará cuántos grupos significativos se pueden encontrar en los datos y se calcularán.
- Realizar una descripción semántica de los patrones encontrados. analizando la importancia de cada grupo y analizando la importancia de las variables en su definición, para ello hay que realizar una descripción estadística de los grupos encontrados.

Los datasets a tratar son los siguientes:

- 1. Segmentación de clientes
  - https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python
  - En el dataset se encuentran datos relativos a clientes de un centro comercial (género, edad, ingresos anuales, puntuación de gasto (asignada por el centro comercial))
  - El objetivo es categorizar los clientes
- 2. Segmentación de países
  - https://www.kaggle.com/datasets/hellbuoy/pca-kmeans-hierarchical-clustering
  - En el dataset se encuentran datos relativos a factores socioeconómicos y de salud que determinan el desarrollo de un país.
  - El objetivo es categorizar los países en base a esos factores socioeconómicos y de salud.
- 3. Segmentación de jugadores
  - https://www.kaggle.com/datasets/aishahakami/call-of-duty-players
  - Este dataset contiene datos sobre el comportamiento de una serie de jugadores de Call of Duty. Por ejemplo, número de victorias, nivel del jugador, derrotas, prestigio, hits, etc.
  - El objetivo es categorizar a estos jugadores (separarlos en grupos o clusters) según esos datos.

Para cada uno de ellos realizar un documento de google colab con los siguientes epígrafes y tareas:

- Importación de librerías necesarias
- Preprocesamiento y elección de variables relevantes para el estudio.
  - Se deberá realizar un pequeño análisis estadístico y argumentar qué variables se eligen para el estudio.
- Escalamiento y detección de outliers
  - Argumentar si es necesario o no realizar escalamiento de las variables, y, si es necesario, escalarlas.
  - Detectar los outliers y eliminarlos
    - Utilizando el método de Jacknife en el que ejecutaremos el algoritmo K-means eliminando una observación cada vez.
      - Almacenaremos el índice SSE de cada una de las ejecuciones en un vector. Esa información se puede obtener con la variable inertia_ del modelo
      - El siguiente código muestra un ejemplo de cómo almacenar esos valores

```
1
2 X = new_df.to_numpy()
3
4 N = X.shape[0] # Número de observaciones
5 K = 4 # Número de clusters
6
7 SSE = []
8 for i in range(0, N):
9   X_sin_i = np.delete(X, i, axis=0) # Eliminamos la observación i
10  # Aplicamos K-medias a X_sin_i y obtenemos el índice SSE
11  kmeans = KMeans(n_clusters=K, n_init=10, random_state=100).fit(X_sin_i)
12  SSE.append(kmeans.inertia_)
13
```

El siguiente ejemplo muestra cómo detectar analíticamente los outliers, para ello añadimos la posición del valor que detectamos como outlier (está fuera del umbral) a un vector llamado outliers.

```
## Detección analítica de outliers
sigma = np.std(SSE) # Desviación típica de SSE
mu = np.mean(SSE) # Media
# Umbral: 2 para distribuciones normales y
#         3 para cualquier otra distribución
umbral = 2

outliers = []
for i in range(0, N):
  if np.abs(SSE[i]-mu) > umbral*sigma:
    outliers.append(i)

print(outliers)
```

- Ejecución de análisis cluster con un valor de k concreto (número de grupos), por ejemplo 4, y determinar el número de réplicas (n_init) del algoritmo que conduce a que la repetición de dicho algoritmo genere la misma solución.
- Determinar el valor de k (número de grupos)
  - Optimizando la función BIC
- Ejecución de análisis con los datos calculados (k y n_init)
- Mostrar para el análisis anterior una tabla con los siguientes datos: número de grupo, cantidad de observaciones de ese grupo y el centroide que representa a ese grupo(valor de cada variable)
  - Esta tabla podría ser una tabla de pandas
  - A continuación se muestra un ejemplo de tabla

Número de grupo	Número de observaciones del grupo	Centroide: característica x	Centroide: característica y	Centroide: característica ...
0	n1			
1	n2			
...	...			
k-1	...	...	…	...

- Realiza una descripción semántica de los grupos
  - Explicación de las características más relevantes de cada grupo (para ayudarte a explicar los grupos realiza un análisis estadístico de los resultados)

Documentación a entregar. Se entregará en moodle centros un archivo (cuaderno de júpiter o google colab) para cada uno de los ejercicios (datasets) con el desarrollo de las tareas anteriormente mencionadas. Las explicaciones a cada tarea se pueden realizar en el mismo cuaderno de júpiter o en un documento pdf.

Evaluación

Los criterios de evaluación de la tarea son a, b, c, d del RA4. Aplica técnicas de aprendizaje no supervisado relacionándolas con los tipos de problemas que tratan de resolver..

Para evaluar la práctica se puntúa cada dataset planteado con 3,33 puntos.

Rúbrica

	100%	**75%**	**50%**	**25%**	**0%**
Preprocesamiento y elección de variables relevantes para el estudio **1 punto**	Realiza el preprocesamiento de los datos y elige las variables a considerar basándose en estudios estadísticos.	Realiza el preprocesamiento de los datos y elige las variables a considerar, pero no muestra un análisis estadístico o no explica porqué ha seleccionado esas variables.	Realiza el preprocesamiento de los datos y elige las variables a considerar, pero no muestra un análisis estadístico y no explica porqué ha seleccionado esas variables.	Realiza el preprocesamiento de los datos pero no elige las variables a considerar.	No realiza el preprocesamiento ni elige las variables para el estudio.
Escalamiento de variables y detección de outliers **1 punto**	Explica si es necesario realizar escalamiento de los datos (si lo considera necesario lo realiza), realiza la detección de outliers y elimina los que encuentra	No explica si es necesario realizar escalamiento de los datos pero realiza correctamente la detección de outliers y elimina los que encuentra	No explica si es necesario realizar escalamiento de los datos y realiza la detección de outliers con fallos menores (y elimina los que encuentra)	No explica si es necesario realizar escalamiento de los datos o no realiza correctamente la detección de outliers y elimina los que encuentra	No explica si es necesario realizar escalamiento o no de los datos y no realiza el proceso de detección de outliers
Ejecución del algoritmo y determinación del número de réplicas **2 puntos**	Realiza la ejecución del algoritmo y determina correctamente el número de réplicas necesarias (n_init)	Realiza la ejecución del algoritmo y determina el número de réplicas necesarias (n_init) con algún error menor	Realiza la ejecución del algoritmo y determina el número de réplicas necesarias (n_init) con varios errores	Realiza la ejecución del algoritmo pero no determina el número de réplicas	No realiza la ejecución del algoritmo ni determina el número de réplicas necesarias n_init
Determinación del valor de k (número de grupos) optimizando la función BIC **2 puntos**	Determina el valor del número de grupos optimizando la función BIC (esta función se ha implementado aparte). Además muestra una gráfica con el valor del BIC para cada número de grupos (k)	Determina el valor del número de grupos optimizando la función BIC (esta función se ha implementado aparte).	Determina el valor del número de grupos optimizando la función BIC (esta función no se implementa aparte).	Determina el valor del número de grupos optimizando la función BIC con algunos errores	No determina el valor del número de grupos

	100%	**75%**	**50%**	**25%**	**0%**
Ejecución de análisis con los datos calculados anteriormente **1 puntos**	Realiza la ejecución correctamente con los datos calculados	Realiza la ejecución correctamente con los datos calculados, con algún error	Realiza la ejecución correctamente con los datos calculados, con algunos errores menores	Realiza la ejecución correctamente con los datos calculados, con bastantes errores	No realiza la ejecución correctamente con los datos calculados
Mostrar una tabla con los centrides y número de elementos de cada grupo **1,5 punto**	Muestra la tabla siguiendo las especificaciones propuestas. Esta tabla se realiza usando la librería pandas.	Muestra la tabla siguiendo las especificaciones propuestas. No se usa la librería pandas.	Muestra la tabla siguiendo las especificaciones propuestas. Esta tabla contiene algunos errores	Muestra una tabla que sigue parcialmente las especificaciones propuestas.	No muestra la tabla o no sigue las especificaciones propuestas
Descripción semántica de cada grupo **1,5 punto**	Realiza una explicación por cada grupo resultante en el que destaca las características más importantes. Esta explicación está basada en un análisis estadístico.	Realiza una explicación por cada grupo resultante en el que destaca las características más importantes. No muestra ningún estudio estadístico	Realiza una explicación por cada grupo resultante en el que destaca las características más importantes. Estas explicaciones son escasas o incompletas	Se realizan explicaciones de la mayoría de grupos pero no de todos.	No se realizan las explicaciones de cada grupo o son profundamente incompletas o erróneas.

Ampliación

Realiza las mismas actividades con los siguientes datasets:

- https://archive.ics.uci.edu/ml/datasets/wholesale+customers
- https://www.kaggle.com/prasertk/healthy-lifestyle-cities-report-2021
- https://www.kaggle.com/justinas/nba-players-data
- Investiga un dataset de tu interés e intenta categorizarlo en distintos grupos