APRENDIZAJE NO SUPERVISADO - KMEANS

Clientes

Tras una importación de librerías, un preprocesamiento, donde comprobamos los nulos, eliminamos la columna CustomerID ya que es una columna que no nos aporta mucha información, y quedándonos solo con las columnas [Annual Income (k$),Spending Score (1-100)] para el estudio ya que son las columnas con mas informacion y mas relevantes, comenzamos nuestro estudio.

Con nuestra función BIC, la cual nos ayudará a la hora de detectar el número de clusters adecuado para nuestro estudio, donde a medida que aumentamos este, decrece el BIC hasta 5, que a partir de este aumenta, nos quedaremos con este.

Con esta información, nos ayudará a la hora de eliminar outliers, que en este caso eran 6.

Una vez terminada esta parte procedemos a la determinación de patrones y estudio.

Calculamos los patrones y los valores de los centroides de estos. Estos los añadimos a nuestro dataframe. para posteriormente,a partir de esta generar un gráfico donde nos mostrara nuestros 5 grupos, con valores sobre las columnas [Annual Income (k$),Spending Score (1-100)].

Y por último generamos una tabla con el número de grupos, las veces que se repite cada uno, una media de la edad, como dato informativo, y los centroides de cada columna.

Esta nos arroja una información valiosa, donde vemos que por ejemplo los dos grupos que más tienen ingresos, tienen dos puntuaciones distintas, y una edad ciertamente alejada.

Paises

Tras repetir cierta parte del proceso anterior, a la hora de elegir las columnas nos hemos guiado en las que podrían ser más importantes, quitando por ejemplo el nombre del pais.

En este caso a la hora de elegir K, nuestra función BIC, nos informa que a medida que aumentemos la K, aumentará nuestro BIC, por lo cual nos quedamos con nuestro valor inicial de 2.

A la hora de los outliers tenemos pocos, solo 3.

Repitiendo el proceso de cálculo de patrones, nos arroja dos grupos, y como se verá en el gráfico en el ejercicio tenemos el grupo 1 que es más compacto, mientras que el 0 tiene ciertos valores alejados en la gráfica de Imports vs Exports.

Y una última gráfica donde comparamos estos 2 valores, junto a la de la salud.

Cod

Repitiendo el anterior proceso, a la hora de las columnas, en este punto ha sido decisión mía la elección de columnas desde el punto de la experiencia, ya que he jugado a este juego, y como columnas como nivel, prestigio y tiempo de juego no son datos que realmente demuestren la habilidad de un jugador.

A la hora del BIC, comienza decreciendo hasta llegar a 4, a partir de este aumenta, así que nuestra K será 4.

Nuestro gráfico de los patrones nos arroja la siguiente información, el grupo 0 tiene una cierta parte compacta y otra más dispersa, el grupo 1 muy lineal ascendente, el 2 igual que el 1, y el 3 la mayoría compacta y ciertos valores dispersos.