Entrega avance del proyecto semana 5

Repositorio de GitHub:
https://github.com/Caanar/Proyecto-Final-Clustering-del-Desarrollo-Paises

1. Estadísticas descriptivas

Todas las variables son numéricas y continuas (a excepción de la columna 'country', la cual es categórica y representa a cada país). Para verificar la integridad de los datos se revisó que no hubiera datos duplicados. Asimismo, se verificó que todos los países contaran con información en todas las variables. No se encontraron datos perdidos. A partir de los descriptivos se revisó que los valores dentro de cada variable se encuentren en el rango de lo esperado. A continuación se presentan en qué consiste cada variable y se presentan los descriptivos. Los coeficientes del correlograma son de Pearson.

country (nombre de los 167 países)

child_mort (mortalidad de niños menores a 5 años por cada 1000 nacimientos): encontramos que en promedio mueren 38 niños menores de 5 años por cada 1000 nacimientos, con una desviación estándar (DE) de 40.33 y un rango que va de 2.6 muertes hasta 208 muertes de menores por cada 1000 nacimientos. En el histograma vemos una distribución sesgada a la derecha, es decir, los valores se agrupan en el rango más bajo. En el correlograma observamos que esta variable está fuertemente correlacionada de manera negativa (-0.89) con la variable life_expec. Asimismo, esta variable se correlaciona fuertemente (0.85) con total_fer. Esto posiblemente se deba a que los paises con fertilidad más alta son los más pobres.

exports (exportaciones de bienes y servicios per cápita, recogido como % de GDP per capita): Encontramos que las exportaciones se encuentran en promedio en el 41.11% del GDP per capita, con una DE de 27.41% y un rango que varía del 0.11% hasta el 200% del GDP per capita. El histograma de esta variable nos muestra una distribución aproximadamente normal, con algunos valores extremos en la parte superior del rango. Por otro lado, encontramos en el correlograma que esta variable tiene una correlación media-alta (0.74) con la variable imports.

health (inversión en salud, recogido como % de GDP per capita): aquí observamos que la inversión en salud promedio consiste en el 6.82% del GDP per capita, con una desviación estándar de 2.75% y un rango que va desde 1.81% hasta 17.9% del GDP per capita. Esta variable muestra una distribución aproximadamente normal en su histograma.

imports (importación de bienes y servicios, recogido como % de GDP per capita): La importaciones en promedio son del 46.89% del GDP per capita, con una desviación estándar del 24.21% y un rango que va desde 0.07% hasta el 174% del GDP per capita. Vale la pena notar que en promedio las importaciones son mayores que las exportaciones. Si bien no es por mucho, esto puede estar reflejando las tendencias de mercado global actuales. Esta variable también muestra una distribución aproximadamente normal, con algunos datos extremos en el rango superior. Como se mencionó esta variable parece estar correlacionada con exports.

income (ingresos netos por persona): aquí podemos ver que en promedio los ingresos netos por persona son de 17144.69 dólares anuales, con una desviación estándar de 19278.07 dólares y un rango que varía desde 609 hasta 125000 dólares anuales. El ingreso anual muestra una distribución sesgada hacia la derecha, es decir que los datos tienden a acumularse en la parte baja del rango, con varios valores extremos en la parte superior del rango.
inflation (medida de la tasa de crecimiento anual del GDP total): al revisar la inflación vemos que en promedio los países cuentan con un 7.78% de crecimiento anual del GDP total, con una desviación estándar de 10.57% y un rango que va desde -4.21 hasta 104%. Es notable que hay países en los que la tasa de inflación tiende a disminuir, si bien en general tiende a aumentar. Esta variable muestra una distribución aproximadamente normal, con algunos datos extremos en la parte superior del rango.
life_expec (número promedio de años que un recién nacido viviría):para la expectativa de vida vemos que es en promedio de 70.56 años, con una desviación estándar de 8.89 años y un rango que va desde los 32.1 hasta los 82.8 años. En este caso el promedio se encuentra en la parte superior del rango, con algunos países muy por debajo de este. Esta variable muestra una distribución aproximadamente normal, algo sesgada a la izquierda, es decir que parece que hay más datos agrupados en la parte superior del rango, con algunos datos extremos en la parte inferior. Por otro lado, la expectativa de vida parece estar negativamente correlacionada (-0.76) con la fertilidad total. Esto, al igual que con la mortalidad, probablemente se debe a que los países más pobres tienen las tasas de fertilidad más altas.
total_fer (número de niños que nacerían por mujer si las tasas de edad-fertilidad se mantienen): en cuanto a fertilidad, vemos que en promedio hay 2.95 hijos por mujer, con una desviación estándar de 1.51 hijos y un rango que va desde 1.15 hasta 7.49 hijos por mujer. Esta variable muestra una distribución ligeramente sesgada a la derecha, es decir, con más concentración de datos en la parte inferior del rango.
gdpp (GDP per capita): por último, encontramos que en promedio el GDP per cápita es de 12964.16 dólares, con una desviación estándar de 18328.7 y un rango que va desde 231 hasta 105000 dólares. Esta variable también cuenta con una distribución sesgada a la derecha, es decir, con bastante acumulación en la parte inferior del rango y con algunos valores extremos en la parte superior.

2. Algoritmo planteado:

Como se observó en los descriptivos, las variables se encuentran en diferentes escalas, por lo que consideramos pertinente estandarizar los datos con StandardScaler() en un comienzo, aunque también se correrá el código sin estandarizar a modo de exploración. Dado que algunas de las variables se encuentran altamente correlacionadas, consideramos pertinente una reducción de dimensiones por medio de Análisis de Componentes Principales (PCA), buscando conservar el número de componentes que expliquen al menos el 90% de la varianza. Con los datos resultado del PCA se procederá a utilizar distintos algoritmos de clustering para explorar posibles agrupaciones de países. Utilizaremos Clustering K-Medias, K-Medoides, Clustering Jerárquico y

DBSCAN con el objetivo de comparar diferentes perspectivas y complementar el análisis para la propuesta de una toma de decisión para la ONG sobre cuáles son los países que deberían tener prioridad en la inversión de recursos. Dado que todas las variables son cuantitativas, utilizaremos la distancia euclidiana en todos los algoritmos. Los hiperparámetros en cada uno de estos algoritmos serán identificados y utilizados según sea el caso. Por ejemplo, en el caso del número de clusters a elegir en K-Medias y K-Medoides se utilizará el coeficiente de Silhouette y el método del codo para elegir el número que minimice la varianza intra cluster.

3. **Implementación algoritmo:**

Como se puede ver en el script 'Primera Implementación' en la carpeta 'Códigos' del repositorio, comenzamos cargando algunas librerías básicas y cargamos los datos haciendo uso de pd.read_csv(). Luego se importaron las librerías correspondientes para correr el Análisis de Componentes Principales y para estandarizar los datos. Se procede a estandarizarlos y a correr el PCA con 5 componentes debido a que con este número se alcanza una varianza explicada del 94.53%. Luego se importan las librerías relacionadas a la implementación de K-Medias. Se realizan gráficos mostrando el coeficiente de Silhouette y la varianza intra-cluster con diferente número de clusters para K-Medias y K-Medoides. Encontramos que para K-Medias el número óptimo de clusters se encuentra alrededor de 10, pero para K-Medoides en un rango parecido.
Luego se formula un loop sobre diferentes estados aleatorios, para elegir el que minimice la varianza intra-clusters en la asignación de puntos para K-Medias y se elige el estado aleatorio con menor resultado. Con esto se formulan los clusters según K-Medias. Se imprime un DataFrame con los países que se encuentran en cada clúster y se imprimen los descriptivos para cada variable, agrupando por cada clúster.

Para realizar el modelo de K-Medoides se utilizaron también 10 clusters y se imprimieron los descriptivos de cada uno de ellos, en este caso se puede evidenciar que el cluster que más países tiene es número 8 (42 países) y en este se tienen en cuenta los países con un menor PIB per cápita, menor expectativa de vida y demás variables, lo que nos hace ver que estos son los países con menor desarrollo socio económico; dentro de este cluster 8 se encuentran países como Afganistán, Camerún, Haití, Uganda, Ruanda, etc. y podemos notar que muchos de estos son principalmente de África. Por otro lado, el cluster con países con maoy desarrollo económico parece ser el número 6, este cuenta con 24 países y dentro de ellos está Australia, Austria, Canadá, Dinamarca, Francia, Estados Unidos y demás países con niveles altos de ingresos, de PIB per cápita, etc.

Para tener un acercamiento al algoritmo de cluster jerárquico, en primera instancia quisimos agrupar los países por dos variables que, podrían ser las que mayor relevancia tienen al momento de explicar el desarrollo económico en los países, estas son ingreso e índice de salud. Posteriormente y con el fin de entender un poco cómo se adaptaría nuestro algoritmo de cluster jerárquico a nuestros datos realizamos un dendograma. El particular, nos muestra una matriz

bastante profunda por lo que para tener mayor claridad sobre los clusters a conformar en quisimos probar fijando la distancia en 80 para conformar así cuatro (4) clusters. De estos clusters conformados en nuestro algoritmo de cluster jerarquico podemos evidenciar como hay una gran concentración en el cluster de color verde el cual agrupa los paises con menor ingreso.