

Instructivo de Uso Código

Elección Variables más Significativas

1. Cree una carpeta para realizar el proceso de forma ordenada, en esta copie los siguientes archivos:
 - Normalizacion.py
En este script se realiza la normalización de los datos ingresados
 - KMeans_Graficos.py
Script donde se realiza la ejecución del algoritmo KMeans, se almacenan los datos y se crean las gráficas de distribución de desertores por cluster
 - Analisis_Codo.py
Script donde se seleccionan las variables más relevantes, mediante la diferencia entre centroides y se encuentra el punto de inflexión
 - BaseDatos.xlsx
Archivo excel donde se encuentran los 804 estudiantes analizados con sus 203 variables analizadas
 - Desertores.xlsx
Archivo excel donde se tiene almacenada la situación del estudiante (Desertor ->1, no desertor ->0)






Nombre	Fecha de modificación	Tipo	Tamaño
 Normalizacion.py	17/01/2022 3:24 p. m.	Python File	1 KB
 KMeans_Graficos.py	28/01/2022 2:08 p. m.	Python File	10 KB
 Analisis_Codo.py	28/01/2022 2:21 p. m.	Python File	2 KB
 Desertores.xlsx	26/07/2021 8:43 a. m.	Hoja de cálculo d...	97 KB
 BaseDatos.xlsx	14/06/2021 8:51 p. m.	Hoja de cálculo d...	625 KB

Figura 1

2. Ejecute el script “**Normalizacion.py**”, este debe crear un archivo .xlsx con la información normalizada de la base de datos.







Nombre	Fecha de modificación	Tipo	Tamaño
 Normalizacion.py	17/01/2022 3:24 p. m.	Python File	1 KB
 KMeans_Graficos.py	28/01/2022 2:08 p. m.	Python File	10 KB
 Analisis_Codo.py	28/01/2022 2:21 p. m.	Python File	2 KB
 Desertores.xlsx	26/07/2021 8:43 a. m.	Hoja de cálculo d...	97 KB
 BaseDatos_Normalizada.xlsx	28/01/2022 2:00 p. m.	Hoja de cálculo d...	504 KB
 BaseDatos.xlsx	14/06/2021 8:51 p. m.	Hoja de cálculo d...	625 KB

Figura 2

3. Ejecute el script “**KMeans_Graficos.py**”, este genera:

- Una carpeta donde se almacenan archivos con los centroides y labels encontrados durante la ejecución del algoritmo KMeans. (Fig. 3)
- Una carpeta donde se almacenan las gráficas de la distribución de desertores en cada uno de los clusters. (Fig. 3)

Nombre	Fecha de modificación	Tipo	Tamaño
img	28/01/2022 2:12 p. m.	Carpeta de archivos	
excel	28/01/2022 2:12 p. m.	Carpeta de archivos	
Normalizacion.py	17/01/2022 3:24 p. m.	Python File	1 KB
KMeans_Graficos.py	28/01/2022 2:08 p. m.	Python File	10 KB
Analisis_Codo.py	17/01/2022 11:49 a. m.	Python File	2 KB
Desertores.xlsx	26/07/2021 8:43 a. m.	Hoja de cálculo d...	97 KB
BaseDatos_Normalizada.xlsx	28/01/2022 2:00 p. m.	Hoja de cálculo d...	504 KB
BaseDatos.xlsx	14/06/2021 8:51 p. m.	Hoja de cálculo d...	625 KB
Silhouette.png	28/01/2022 2:12 p. m.	Archivo PNG	14 KB
Indices.png	28/01/2022 2:12 p. m.	Archivo PNG	18 KB

Figura 3

- Una lista con las agrupaciones más discriminativas en orden descendente. (Fig. 4)
Nota: La lista está compuesta por el número de clusters, es decir, si aparece el numero 25, es por que cuando se agrupó en 25 clusters se llegó a la distribución más discriminativa

Las 3 agrupaciones con la proporción mas discriminativa en orden son: [25, 23, 22]

Figura 4

- Gráficas de los índices Silhouette y DaviesBouldin-Dunn, estos dan la calidad de la agrupación

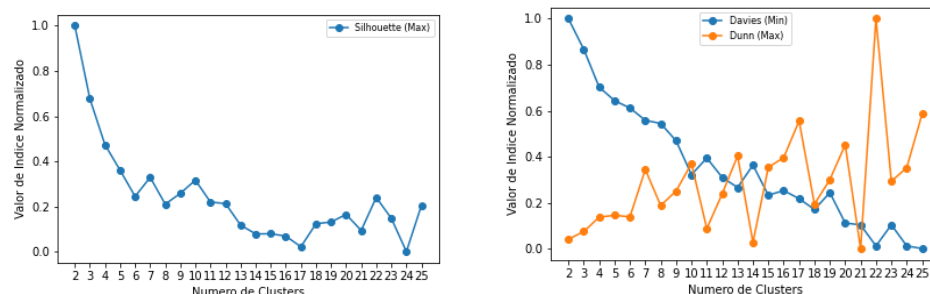


Figura 5.

- Analice cual es la mejor agrupación, para esto básiase en:
 - El orden dado por la lista, pues es ese orden se tiene una mayor discriminación entre desertores y no desertores
 - Los índices Silhouette y DaviesBouldin-Dunn,
Elija el número de Clusters que mejor cumpla con ambos criterios, Silhouette (alto), Davies (Bajo), Dunn (Alto).

Por ejemplo, para la Figura 5, el número de clústeres ideal sería 2, sin embargo al analizar la distribución (Figura 6) nos damos cuenta que no se llega a una distribución discriminativa, por esto se descarta,. Analizando los índices, podemos ver que otra agrupación buena (Silhouette->alto, Dunn->Alto, Davies->Bajo) se encuentra con 10 clusters, entonces se analiza la distribución con 10 (Figura 7), y se concluye que esta si es una buena distribución porque hay clusters con alto número de desertores y otros con baja cantidad (El C4 tiene un 21% y el C8 tiene un 55,6%).

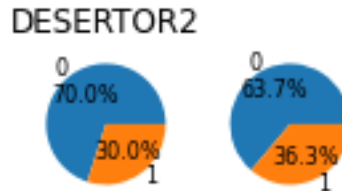


Figura 6.

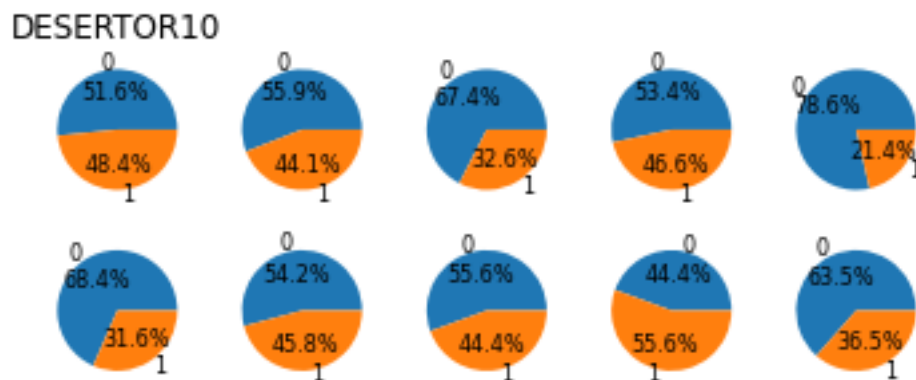


Figura 7

- En el script “**Analisis_Codo.py**”, modifique la línea 39:

```
centroids=pd.read_excel("excel/centroids2.xlsx",index_col=0)
```

cambiando el número que acompaña a ‘centroide’ por el número de clusters que usted encuentra apropiado en el ítem 4 (Fig. 8)

```
35
36 #----- Definicion de variables -----
37
38 #DataFrames
39 centroids=pd.read_excel("excel/centroids2.xlsx",index_col=0) # DF con los centroides de la agrupacion elegida
40 bd=pd.read_excel("BaseDatos_Normalizada.xlsx",index_col=0) # DF con la BD normalizada
41 bd_new= pd.DataFrame() # DF donde se almacenaran las columnas significativas
42 bd_new['a']=range(0,len(bd)) # Se especifica el numero de registros con que se trabajara
43
```

Figura 8

6. Ejecute el script “**Analisis_Codo.py**”, este genera un archivo .xlsx con las variables más significativas, en orden de relevancia (La primera columna corresponde a la variable de mayor relevancia).

Nombre	Fecha de modificación	Tipo	Tamaño
img	28/01/2022 2:12 p. m.	Carpeta de archivos	
excel	28/01/2022 2:12 p. m.	Carpeta de archivos	
Normalizacion.py	17/01/2022 3:24 p. m.	Python File	1 KB
KMeans_Graficos.py	28/01/2022 2:08 p. m.	Python File	10 KB
Analisis_Codo.py	28/01/2022 2:21 p. m.	Python File	2 KB
Desertores.xlsx	26/07/2021 8:43 a. m.	Hoja de cálculo d...	97 KB
BaseDatos_Normalizada.xlsx	28/01/2022 2:00 p. m.	Hoja de cálculo d...	504 KB
BaseDatos.xlsx	14/06/2021 8:51 p. m.	Hoja de cálculo d...	625 KB
Base_Datos20variables.xlsx	28/01/2022 2:21 p. m.	Hoja de cálculo d...	53 KB
Silhouette.png	28/01/2022 2:12 p. m.	Archivo PNG	14 KB
Indices.png	28/01/2022 2:12 p. m.	Archivo PNG	18 KB

Figura 8

7. Repita el proceso (sin realizar el ítem 2), reemplazando el archivo ‘BaseDatos’ por el archivo resultante del ítem 6.

Este proceso se repetirá hasta que se llegue a una de estas dos situaciones mencionadas a continuación, en este caso se tomará la base de datos utilizada en el proceso anterior:

- El criterio del codo no aplique (esto será cuando el codo este ubicado en los extremos, incluya a todas las variables o ninguna)

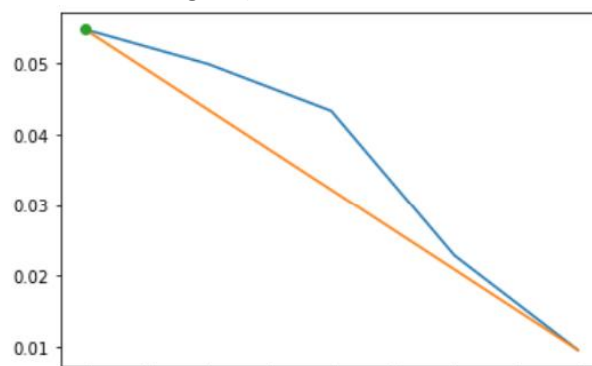


Figura 9

- Las distribuciones a las que se llegan con los diferentes número de clusters no son discriminativas: No exista una diferencia notable entre la distribución de desertores y no desertores entre los clusters.

8. Cuando encuentre las variables mas relevantes, ejecute nuevamente el algoritmo KMeans alimentándolo con la base de datos que contiene a las variables relevantes, analice la distribución de desertores y los índices de calidad de clusters; el número de clusters que tenga mejor estas características será el elegido para el modelo.