Machine Learning

Clasificacion de tipos de clientes

Leonardo Trevizo Herrera

Maestro. Joseph Isaac Ramirez Hernandez

29 de noviembre, 2023.

Author Note

La letra ‘s’ me falla, pido disculpas.

Abstract

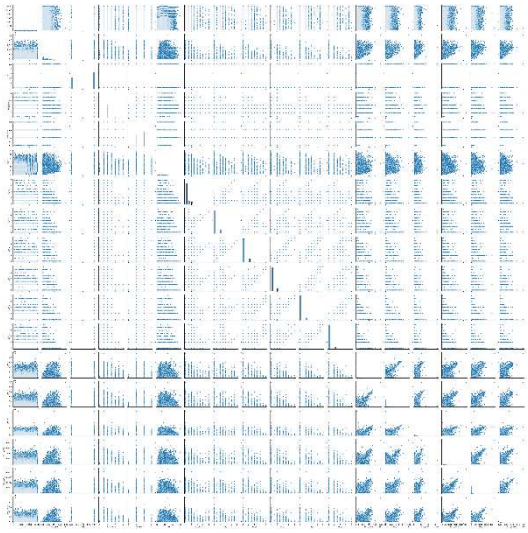
Este proyecto se centra en la implementación de K-means, un algoritmo de agrupamiento eficiente, para analizar patrones inherentes en un conjuntos de datos de la línea de crédito de clientes para un banco en Taiwan en 2005. El objetivo principal es segmentar datos no etiquetados en grupos homogéneos, facilitando así la identificación de estructuras subyacentes. El proceso comienza con la recopilación y preprocesamiento de datos, seguido de la aplicación del algoritmo K-means para crear clusters. Se exploran diferentes valores de k para determinar la cantidad óptima de grupos.

Importacion de librerias

```
import numpy as np
import matplotlib.pyplot as plt
from PIL import Image
from sklearn.cluster import KMeans
from sklearn.utils import shuffle
import pandas as pd
import seaborn as sns

import warnings
warnings.filterwarnings("ignore")
```

De las librerías qué utilicé , lo unico que puedo decir es que no llegé a utilizar a PIL, y que warnings la utilicé porque al hacer las graficas de dispersión me arrojaba una alarma que no significaba una falla ni impedía que las graficas se realizaran y que el código se ejecutara.

Preprocesamiento de datos



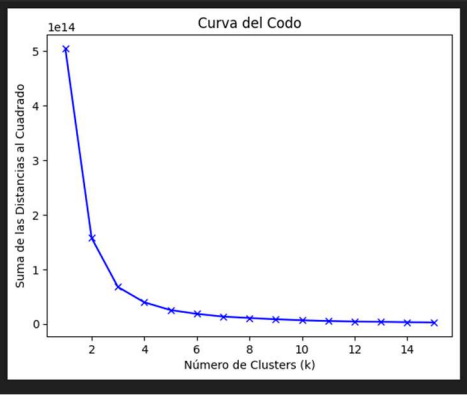
El preprocesamiento de los datos lo realicé directamente en csv, borrando columnas que no le aportaban, sino todo lo contrario, lo hacía más complejo. Al principio eran 24 columnas, y terminé usando 6 columnas.

```
data2 = pd.read_csv('data/UCI_Credit_Card_2.csv', sep=',')
X2=np.array(data.iloc[:,1:7])

print(X2.shape)
```

```
(30000, 6)
```

```
plt.figure(2) sns.pairplot(X2) plt.show()
```

```
x=elbow_method(X2, max_clusters=15)
```

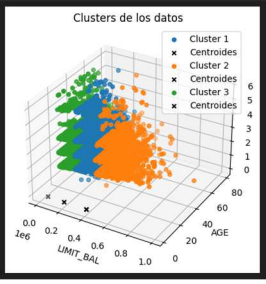


Para elegir el numero de clusters utilicé el metodo del codo, y el método mas prominente fue para un numero de clusters de 3.

Procesamiento de datos

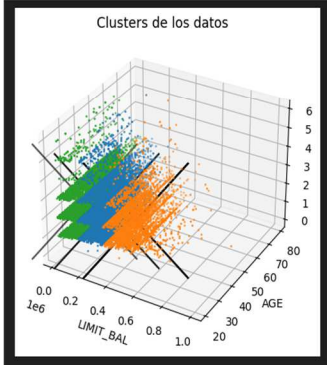
```
n2=3

kmeans_2 = KMeans(n_clusters=n2, init='random',n_init=30,tol=1e-015, max_iter=10000, algorithm='elkan')
kmeans_2.fit(X2)

yhat2 = kmeans_2.predict(X2)
```

```
az = plt.axes(projection='3d')
for i in range(n2):
    az.scatter3D(X2[yhat2==i,0], X2[yhat2==i,4], X2[yhat2==i,2], label='Cluster '+str(i+1))
    az.scatter(kmeans_2.cluster_centers_[:,0], kmeans_2.cluster_centers_[:,2], color='black', marker='x', label='Centroides')
az.legend()
plt.xlabel('LIMIT_BAL')
plt.ylabel('AGE')
#ax.set_zlabel('EDUCATION')
plt.title('Clusters de los datos')
plt.show()
```



Procesé los datos y empecé a graficar, se puede observer que los clusters parecen distantes de sus centroides, pero esto lo arreglo posteriormente eligiendo otros datos para hacer la graficación.

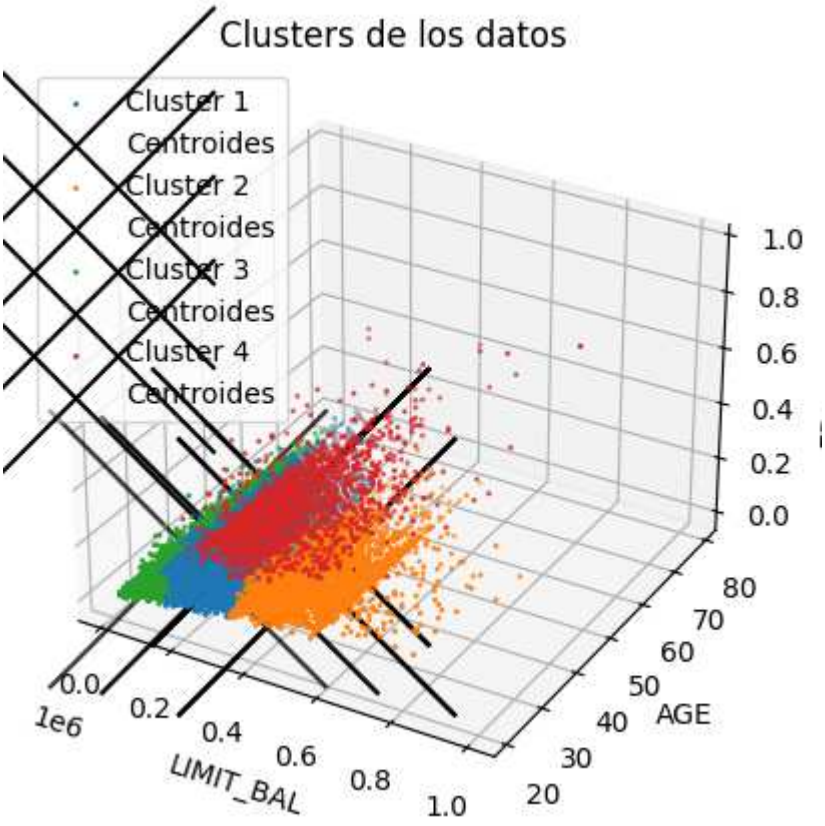
```
az = plt.axes(projection='3d')
for i in range(3):
    az.scatter3D(X2[yhat2==i,0], X2[yhat2==i,4], X2[yhat2==i,2], label='Cluster '+str(i+1), zorder=1, s=1,)
    az.scatter3D(kmeans_2.cluster_centers_[:,0], kmeans_2.cluster_centers_[:,4], kmeans_2.cluster_centers_[:,2], color='black', marker='x', label='Centroides',s=10000, zorder=50000)
ax.legend()
plt.xlabel('LIMIT_BAL')
plt.ylabel('AGE')
#ax.set_zlabel('EDUCATION')
plt.title('Clusters de los datos')
plt.show()
```



Aquí quedó solucionado, el eje de las x es el limite de la línea de crédito, la y es la edad, y la z es el nivel de educación, donde entre más pequeño el numero mayor es el nivel de excolaridad, a excepción del 0 que es desconocido. Lo unioco que no pude solucionar fueron los centroides, pues como se puede apreciar se ven detrás de los datos. Esto lo trate de solucionar estableciendo los parámetros de zorder, traté de graficar los centroides después de crear la grafica, traté también haciendosos con otro for loop mas abajo, pero nada me funcionó. Lo unico que si sirvio fue hacerlos super grandes para que se pudieran observar, y den forma de equis para que se pudiera predecir o tanteaer el centro del centroide.

Conclusión

Se pueden estimar 3 tipos de clientes, los cuales quedan bien definidos por la clase de poder adquisitivo de los clientes. Pero sucede algo sumamente interesante cuando añadimos un cluster más, surge el tipo de cliente impuntual con sus pagos que generalmente tiene pendiente o atrasada casi toda la línea de crédito como se podrá ver a continuación.



Los resultados obtenidos demuestran la eficacia de K-means en la identificación de patrones subyacentes en datos no estructurados. Este proyecto contribuye a la comprensión y aplicación práctica de técnicas de agrupamiento en el ámbito de Machine Learning, destacando la importancia de la exploración no supervisada de datos para la toma de decisiones informada.

References

Default Payments of Credit Card Clients in Taiwan from 2005

https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset