## Proceso:

## Objetivo: El objetivo de este trabajo grupal es realizar un aprendizaje no supervisado, este se realizará utilizando simple k menas de weka (clustering)

## Dataset:

## Link: <https://www.kaggle.com/datasets/jillanisofttech/market-segmentation-in-insurance-unsupervised>

## Descripción: En marketing, la segmentación del mercado es el proceso de dividir un amplio mercado de consumidores o negocios, que normalmente consiste en clientes existentes y potenciales, en subgrupos de consumidores en función de algún tipo de características compartidas.

El conjunto de datos de muestra resume el comportamiento de uso de aproximadamente 8950 titulares de tarjetas de crédito activos durante los últimos 6 meses. El archivo está a nivel de cliente con 18 variables de comportamiento.

## Cantidad de ejemplos y de atributos:

El dataset cuenta con 18 columnas y 8950 ejemplos

|  |  |
| --- | --- |
| **ATRIBUTO** | **TIPO** |

|  |  |
| --- | --- |
| Cust\_id | categorico |
| Balance | numerico |
| Balance\_frequency | numerico |
| Purchases | numerico |
| Oneoff\_purchases | numerico |
| Installments\_purchases | numerico |
| Cash\_advance | numerico |
| Purchases\_frequency | numerico |
| Oneoff\_purchases\_frequency | numerico |
| Purchases\_installments\_frequency | numerico |
| Cash\_advance\_frequency | numerico |
| Cash\_advance\_trx | numerico |
| Purchases\_trx | numerico |
| Credit\_limit | numerico |
| Payments | numerico |
| Minimum\_payments | numerico |
| Prc\_full\_payments | numerico |
| Tenure | numerico |

## Tareas:

La tarea a realizar es de aprendizaje no supervisado

## Limpieza:

El único atributo eliminado fue CUST\_ID el cual es único para cada uno de los ejemplos del dataset

## Selección:

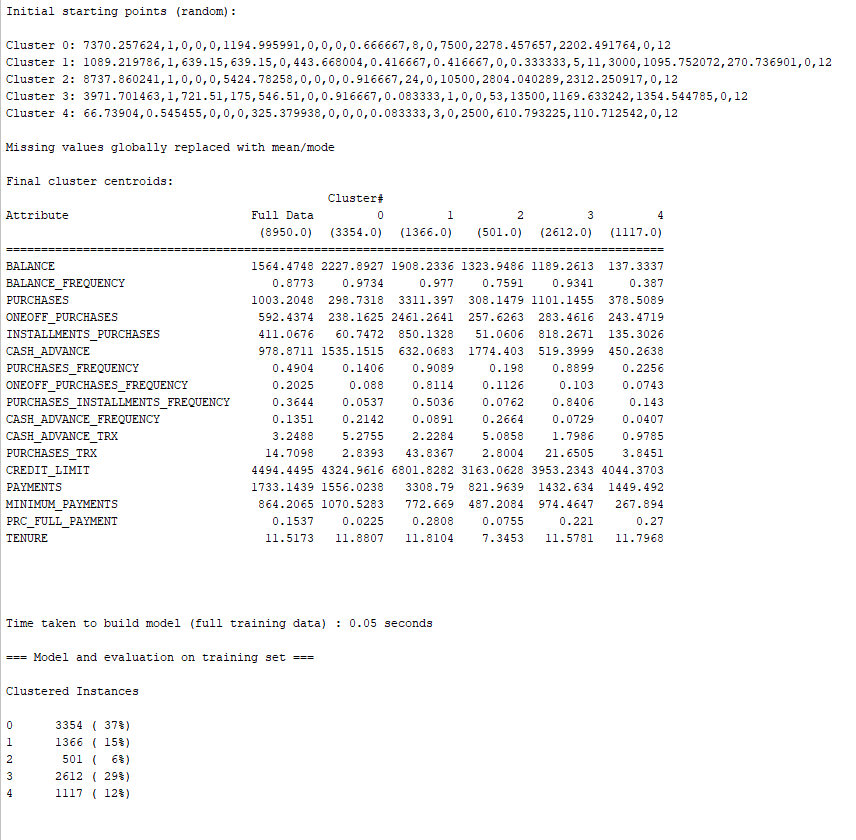
El atributo clasificador es TENURE se utiliza para indicar la cantidad de meses que cada cliente ha tenido activo su cuenta de tarjeta de crédito durante los últimos 6 meses.

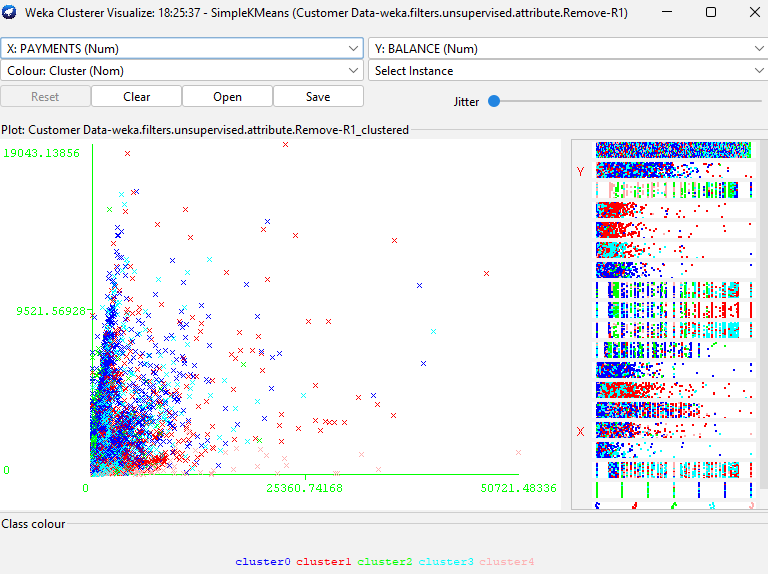
## Integración: No se integraron datos ya que el dataset proporcionó todo lo que se necesitaba para este proceso de minería de datos

## Transformación: Competition\_Date se modificó para que en vez de tener una fecha completa solo tenga año y mes, luego es una ultima corrida del modelo se volvio a modificar competition date para que esta solo tenga año de competencia

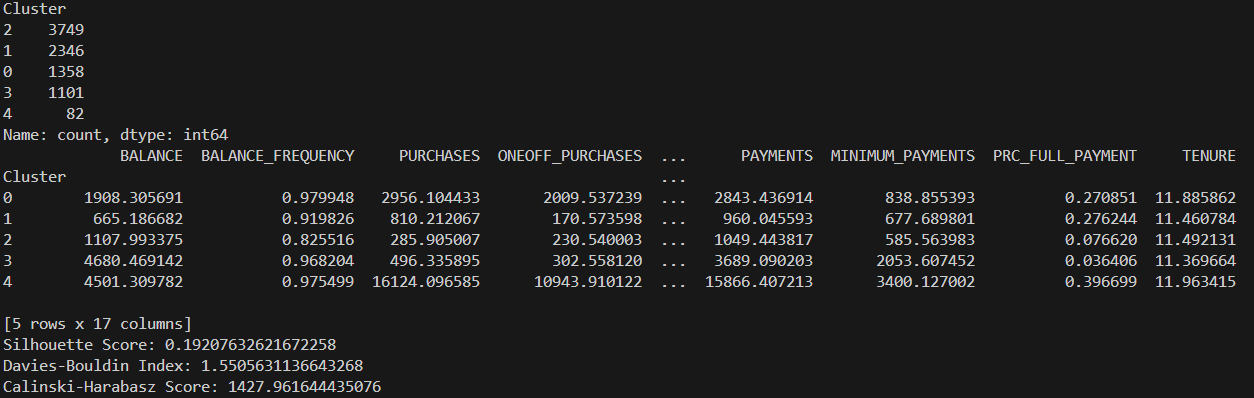
## Minería:

Se aplicará el siguiente algoritmo simple k menas de weka y kmeans de sklearn utilizando la libreria de pandas, ambos algoritmos se utilizarán con 5 clusters utilizando la distancia euclidean





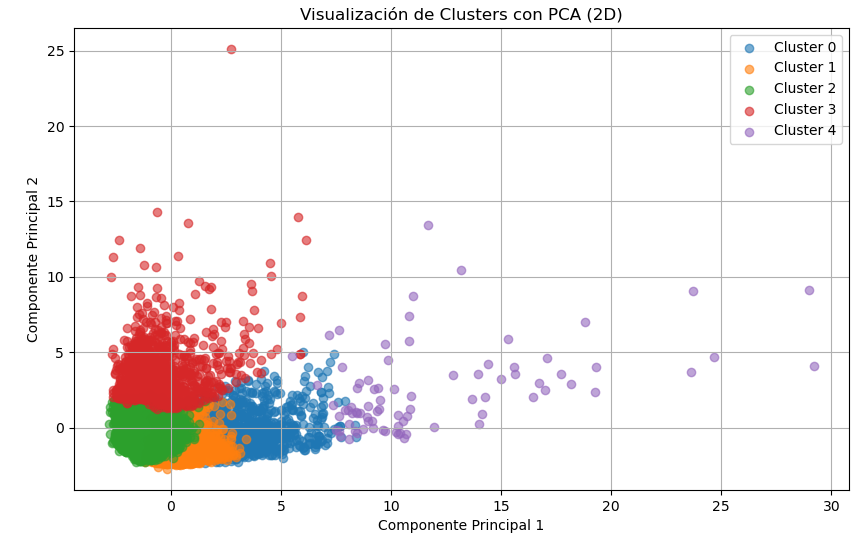
Mediante el script de python se obtuvieron los siguientes resultados, se utilizo Kmeans con la misma cantidad de clusteres que en weka



Muesta los clusters con sus valores y las metricas de evaluacion, siendo Silhoutte Score de 0.19 es un valor bajo y quiere decir que los clusteres estan solapados.

Davies-Bouldin Index de 1.55 es bueno, pero si es menor es mejor

Calinski-Harabasz de 1427.96 es bueno, quiere decir que hay una buena separación entre grupos.



Componente principal 1 y 2 no hacen referencia a nada dentro del dataset son combinaciones lineales de todas las variables

## Evaluación:

**WEKA:**

**Cluster 0 (azul)**: concentra la mayoría de los puntos en la zona inferior izquierda, correspondiente a clientes con bajos pagos y bajo saldo. Se trata de un grupo con actividad financiera limitada, posiblemente clientes inactivos o con bajo uso de la tarjeta.

**Cluster 1 (rojo)**: se encuentra más disperso hacia la derecha del gráfico, en zonas de pagos altos y saldos bajos. Esto indica que son clientes que utilizan la tarjeta y luego realizan pagos elevados, posiblemente saldando su deuda mensualmente. Representan un perfil de cliente responsable.

**Cluster 2 (verde)**: aparece en regiones con saldos altos pero pagos más bajos, lo que sugiere un perfil de riesgo, con tendencia a acumular deuda o a pagar solo el mínimo.

**Cluster 3 (celeste)**: se ubica en zonas intermedias, con balances moderados y pagosregulares. Representan un comportamiento más equilibrado, pero sin extremos.

**Cluster 4 (amarillo)**: es un grupo más disperso y reducido, con clientes en distintas zonas del gráfico. Podría representar casos atípicos o perfiles híbridos, no fácilmente encasillables, y podrían requerir un análisis más específico.

**PYTHON:**

**Cluster 0 (azul)**: grupo más numeroso. Clientes con actividad financiera moderada, balances estables y comportamiento regular de pagos y compras.

**Cluster 1 (naranja)**: clientes con bajo saldo, bajo gasto y bajo pago, posiblemente usuarios poco activos o nuevos clientes.

**Cluster 2 (verde)**: comportamiento ligeramente más activo que el anterior, pero aún moderado.

**Cluster 3 (rojo)**: clientes con alta frecuencia de uso, en especial en compras en cuotas (INSTALLMENTS\_PURCHASES). Perfil de consumidores activos.

**Cluster 4 (violeta)**: grupo reducido con valores extremos en PURCHASES, PAYMENTS, y CREDIT\_LIMIT. Podría tratarse de clientes corporativos o con altísima capacidad de gasto.

## Conclusiones:

A través del algoritmo K-Means, se logró identificar patrones ocultos en los datos que reflejan diferencias relevantes en variables como BALANCE, PURCHASES, PAYMENTS y CREDIT\_LIMIT. Estos clusters ofrecen una base sólida para personalizar estrategias comerciales, como planes de ahorro, productos financieros específicos o gestión de patrimonio.

En Weka, se pudo observar claramente el perfil promedio de cada cluster, visualizar la distribución de clientes según atributos, y agrupar instancias similares de forma visual e interpretativa.

En Python, además de obtener resultados similares, se aplicaron métricas de evaluación como Silhouette Score, Davies-Bouldin Index y Calinski-Harabasz Score, las cuales indicaron una segmentación razonablemente buena. Además, el uso de PCA permitió visualizar los clusters en un espacio bidimensional, evidenciando la separación entre grupos.

En conjunto, ambas herramientas confirmaron que el modelo fue capaz de identificar grupos diferenciados de clientes, lo cual valida la utilidad del clustering en escenarios donde no se cuenta con etiquetas de clase. Estos resultados pueden ser aprovechados para la toma de decisiones basadas en datos.