Universidad Torcuato Di Tella
Maestría en Economía & Finanzas
Segundo Trimestre 2025

Machine Learning para Finanzas
Problem Set 4: Clustering de Fondos Comunes

Alumna: Paola Nuñez | Profesor: Lionel Modi

---

1 Introducción

Este trabajo aplica **clustering no supervisado** a un panel de cincuenta y ocho Fondos Comunes de Inversión (FCI) etiquetados originalmente como `MM` (money market), `RF` (renta fija) y `EQ` (equity). Para ello, se utilizan tres variables: volatilidad del último mes y el retorno de los últimos tres y seis meses. Por último, se compara la clasificación original con la obtenida mediante herramientas de clustering.

2 Análisis exploratorio

A cada fondo se le asignó un ID único (concatenando tipo y posición en el dataset). Visualmente (ver Figura 1), los fondos de renta variable y money market aparecerían bien condensados sobre su propio centroide por niveles de volatilidad y retornos, mientras que los de renta fija muestran mayor dispersión. Esto sugiere que una partición con $K>3$ podría, principalmente, capturar heterogeneidad intraclase en RF.

Por otro lado, podemos percibir que, en esa heterogeneidad de comportamiento de los fondos de renta fija, hay un fondo *outlier* que presenta retornos negativos y se encuentra bastante alejado del centroide aparente del cluster.

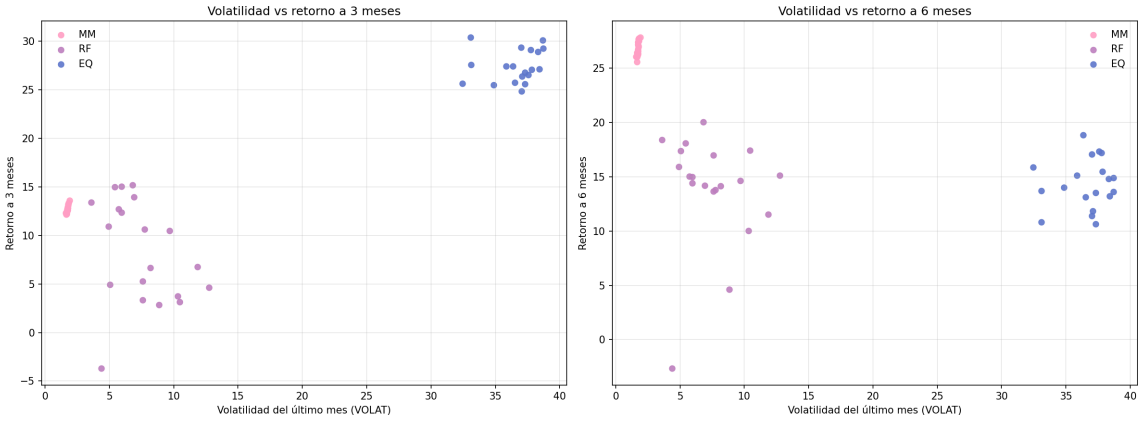


Figure 1: A la izquierda, volatilidad del último mes (`VOLAT`) vs retorno de los últimos 3 meses (`r3m`). A la derecha, volatilidad del último mes (`VOLAT`) vs retorno de los últimos 6 meses (`r6m`). A su vez, se segmenta por `TIPO` original (MM, RF o EQ).

3 Metodología

Los pasos a seguir son:

1. Preprocesar las variables `VOLAT` (volatilidad del último mes), `r3m` (retorno de los últimos 3 meses), `r6m` (retorno de los últimos seis meses). Debido a que los algoritmos de *clustering*, por usar distancias euclideanas, son muy sensibles a las diferentes escalas y a la presencia de outliers, se trabaja sobre estos dos puntos usando RobustScaler y un filtro de intercuantiles.

2. Estimar la cantidad de clusters óptima (K^*_{kmeans} y K^*_{agglo}) con el criterio de **Calinski–Harabasz** (CH).

3. Utilizar **KMeans** y **Agglomerative** (enlace Ward) para obtener la categorización de los fondos asociada a los K^* clusters óptimos.

4. Comparar agrupamientos respecto de las etiquetas de origen, tanto con K óptimo como con K=3.

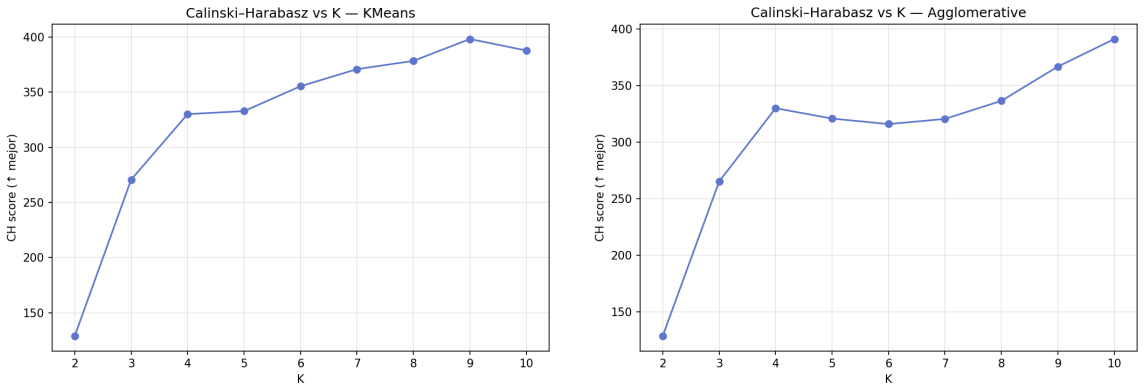
4 Estructura del proyecto

El código fuente completo está en: `https://github.com/nunezpaola/MLforfinance`. Desde la raíz, se estructura como sigue:

- `ps/ps4.py`: incorpora las funciones para preprocesamiento (escalado y filtro de outliers), scoring CH, ajuste de KMeans/Agglomerative y construcción de tablas y gráficos comparativos.
- `ps/ps4_test.py`: módulo de test que fija parámetros dentro del script, calcula el score para $K \in [2, 10]$, elige K^*, usa las herramientas elegidas para clusterizar y muestra en terminal los principales resultados obtenidos.
- `ps/ps4_logs.txt`: **logs** con los principales resultados obtenidos.

5 Resultados

En la Figura 2 se puede observar la evolución del score Calinski-Harabasz. Ambas series tienen un comportamiento monótono en K generando que el nivel óptimo de clusters sea el extremo del intervalo proporcionado ($K = 10$). Esto puede deberse principalmente a la heterogeneidad intraclase de los fondos de renta fija. Sin embargo, se eligirá como óptimo $K = 4$ debido a que el dominio de variables con las que se entrena al modelo es mucho menor a diez y el incremento en el score es notorio hasta el cuarto cluster y no logra gran mejoría a costa de incrementar hasta $K = 10$.

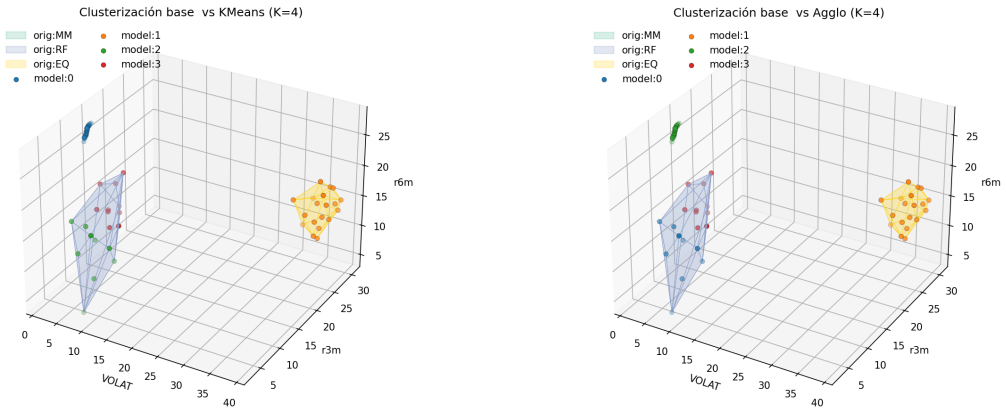


(a) KMeans

(b) Agglomerative

Figure 2: Criterio de Calinski–Harabasz por K para ambos modelos.

Los resultados de la clusterización ambos modelos se pueden observar en la Figura 3. Podemos ver que, en términos de las variables utilizadas, los fondos de renta variable y money market son fácilmente condensables entre sí, mientras que los fondos de renta fija se dividen en dos grupos. Si consideramos que los factores que explicaron la performance de los fondos son estructurales y, consecuentemente, los retornos históricos son buen *proxy* de los futuros, podemos establecer una interpretación de las dos clases de renta fija. En particular, el cuarto cluster de ambos modelos es el conjunto de fondos de renta fija que parecería estar más cerca de la frontera eficiente que el resto de los del mismo tipo. Por tal razón, la recomendación final es trabajar con la clasificación inicial pero a la hora de asignar inversiones en fondos descartar los fondos no eficientes de renta fija (tercer cluster para KMeans y primero para Agglomerative).



(a) KMeans

(b) Agglomerative

Figure 3: Clusterización de Fondos Comunes de Inversión bajo $K = 4$. El scatter 3D muestra la asignación de realizada por el analista, mientras que las áreas sólidas indican la clasificación inicial.