Pontificia Universidad Católica de Chile

Escuela de Ingeniería

Departamento de Ciencia de la Computación IIC2433 - Minería de Datos

Segundo semestre de 2022

Cluster S&P500 según fundamentos financieros

GRUPO 8 - SECCIÓN 1

Integrantes:

Natalia De La Barra

Nicolás Estévez

Juan Ignacio García Forteza

Francisco Hortal Correa

Profesor:

Marcelo Mendoza

12 de diciembre de 2022

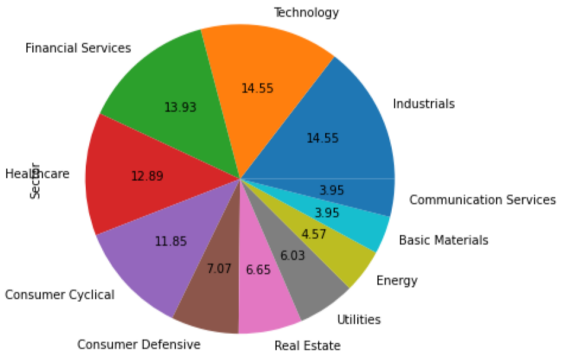
1. Introducción a la problemática abordada

El Standard and Poor´s es el benchmark financiero más famoso del mundo. Este índice bursátil rastrea el desempeño de 500 grandes empresas que cotizan en la bolsa de valores de EEUU. Al 31 de diciembre del 2020 se invirtió más de $5.4 trillones en activos atados a este índice.
Este indicador incluye múltiples clases de acciones, en realidad hay 505 tipos de acciones en el indicador.
Estos componentes cotizan en la Bolsa de Nueva York o en el Nasdaq. Las personas no analizan a las empresas desde lo financiero. *Es por esto que creemos necesario crear una herramienta basada en clustering que permita agrupar las acciones con los indicadores financieros más adecuados y que las personas no arriesguen su dinero en empresas sin fundamentos económicos.

2. Análisis del dataset a trabajar y exploración en profundidad.

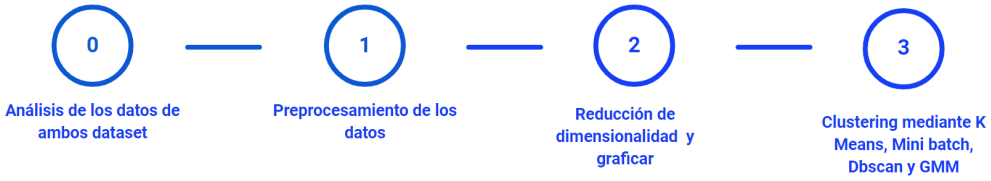
Para nuestro proyecto utilizaremos dos datasets, los cuales se componen de la siguiente manera:

Dataset 1: Se compone de 16 columnas + índice y 495 filas. sp500_companies.csv. Son datos cualitativos, como el sector industrial de la empresa, número de empleados y país de origen.



Dataset 2: Se compone de 9 columnas y 2.441 filas. 5yr_fundamentals_S&P500.txt. Aquí se encuentran datos financieros entre los años 2017 y 2021 de manera trimestral. Algunos de estos son ingresos, deuda, marketcap, utilidades y free cash flow.

3. ¿Cómo se abordó la temática?:



<u>Fundamentación de los métodos utilizados:</u>

PCA: Condensar información, seleccionar información más relevante y mejor desempeño al clusterizar.

K-Means y MiniBatch: Ampliamente utilizados. permiten encontrar grupos ocultos y confirmar o descartar teorías.

DBSCAN: Clasificar grupos de distinto tamaño, no requiere número de cluster previamente y es resistente al ruido.

GMM: Permite detectar clusters con formas complejas y nuevas etiquetas de clasificación.

4. Tratamiento/preprocesamiento que se le ha dado a los datos.

Se chequeó que no existieran datos nulos, y se eliminaron acciones que no estén en ambas bases de datos o que no esté su información para todos los años. Mediante una API se recolectó información de los fundamentos de la empresa durante 5 años. Del 2017 al 2021.

Posteriormente con los datos del dataset se hicieron 8 pilares, los cuales son 8 indicadores financieros que te dicen cómo está la empresa.

8 pilares creados por el millonario y profesor financiero estadounidense Paul Gabrail. Estos pilares nos serán útiles para segmentar las acciones. Tienen leves modificaciones para poder hacerlos más óptimos para los algoritmos de clustering.

1. **5 yr P/E ratio:** $\frac{Market\ Cap}{last\ 5\ years\ net\ income}$
   Indica si el precio de las acciones se condice con sus utilidades

2. **5 yr ROIC:** $\frac{5\ year\ cashflow}{total\ debt\ and\ equity}$
   Compara el dinero que genera la empresa frente a su deuda

3. **5 yr revenue growth:**
   $\frac{revenue\ now - revenue\ 5\ years\ ago}{revenue\ 5\ years\ ago}$
   Representa el crecimiento de los ingresos

4. **Net income growth:**
   $\frac{net\ income\ now - net\ income\ 5\ years\ ago}{net\ income\ 5\ years\ ago}$
   Representa el crecimiento de las utilidades

5. **Shares Outstanding (S.O):**
   $\frac{S.O\ now - S.O\ 5\ years\ ago}{S.O\ 5\ years\ ago}$
   Nos indica si se deprecia la acción porque la empresa está emitiendo

6. **Long Term Liabilities (L.T.L):**
   $\frac{L.T.L}{5\ years\ free\ cash\ flow}$
   Deudas a largo plazo frente a la generación de dinero

7. **Free Cash Flow (F.C.F) Growth:**
   $\frac{F.C.F\ now - F.C.F\ 5\ years\ ago}{F.C.F\ 5\ years\ ago}$
   El crecimiento de la generación de dinero

8. **Price to F.C.F:** $\frac{Market\ cap}{5\ years\ F.C.F}$
   Indica si el precio de las acciones se condice con la generación de dinero

Se realizó un PCA para reducir la dimensionalidad de 8 dimensiones a 2. Analizando este PCA se detectó que el ratio de la varianza explicada era de un 55%. Debido a lo bajo de este ratio se decide reducir la cantidad de variables y utilizar las con mayor varianza y una combinación que incluya todos los factores financieros. Se hace un análisis de la varianza de cada pilar, dando como resultado a los

pilares 2, 3, 5 y 8 como los con mayor varianza. Además nos dimos cuenta que con estos 4 pilares podríamos cubrir todos los parámetros financieros, ya que muchos de estos se repiten en los pilares.

	pilar_1	pilar_2	pilar_3	pilar_4	pilar_5	pilar_6	pilar_7	pilar_8
count	466.000000	466.000000	466.000000	466.000000	466.000000	466.000000	466.00000	466.000000
mean	0.285878	0.291536	0.189362	0.494029	0.186275	0.329866	0.40678	0.405518
std	0.070877	0.121920	0.089354	0.055401	0.093332	0.076178	0.05479	0.077681
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000
50%	0.274452	0.269393	0.169457	0.489827	0.170002	0.327991	0.40408	0.401750
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.00000	1.000000

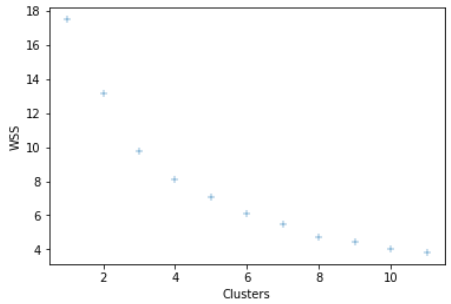
Finalmente se aplica un nuevo PCA con estas 4 variables. El ratio subió 20 puntos ubicándose en 75% en los primeros componentes. Procedimos a utilizar este PCA para el resto del análisis.

5. Demostración de funcionamiento y resultados.

Se realizaron cuatro métodos de clusterización y en función del score que nos entregó determinamos cual es el mejor método, por otra parte para mostrarlos gráficamente se realizaron scatterplots.

Para determinar el número óptimo de grupos se hicieron los siguientes análisis:

-Método Elbow:



Se puede notar que desde el número 5 los puntos comienzan a tener menos cambios.

-Silhouette Score:

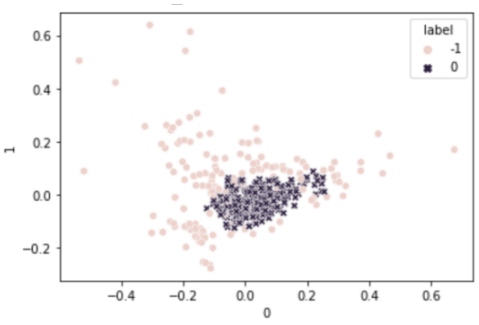
Ahora tratamos de realizar un Silhouette Method para estar seguros del número de clusters

```
Silhouette score for k(clusters) = 2 is 0.2865625875019962
Silhouette score for k(clusters) = 3 is 0.34530552426107236
Silhouette score for k(clusters) = 4 is 0.40532303243594053
Silhouette score for k(clusters) = 5 is 0.4124128468082958
Silhouette score for k(clusters) = 6 is 0.29490622259130905
Silhouette score for k(clusters) = 7 is 0.286903921684372
Silhouette score for k(clusters) = 8 is 0.3010495297861177
```

Vemos que el valor más cercano a 1 se da en 5 clusters, por lo que 5 clusters es el óptimo.

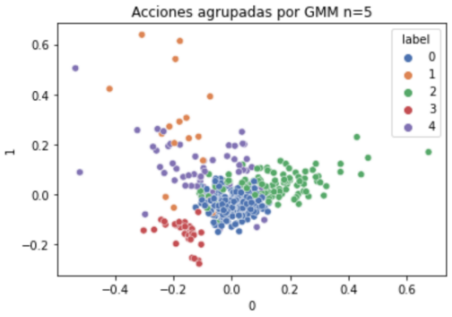
Se realizó clustering con 5 clusters y 4 métodos, obteniendo como resultado:

DBSCAN



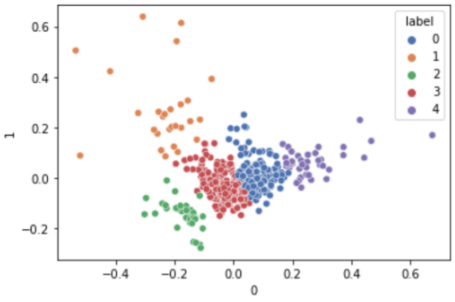
Silhouette score: 0.2948638251137764

GMM



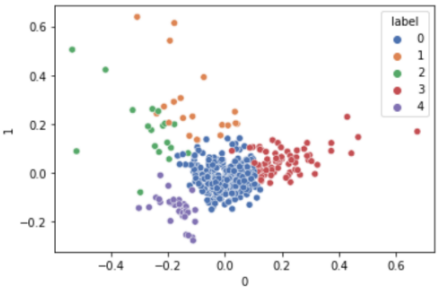
Silhouette score: 0.2826617842065334

Mini Batch K-means



Silhouette score: 0.28528898920562956

K-Means



Silhouette score: 0.41241284680829576

Método de clusterización	Score
K-means	0.41241284680829576
Mini batch	0.28528898920562956
GMM	0.2826617842065334
DBSCAN	0.2948638251137764

Como podemos observar el método de clusterización que tiene mayor score es el de K-means con 0.412 aproximadamente además su scatterplot se ve congruente con la data y con una buena separación de grupos. Este método es el que elegimos para desarrollar el análisis final.

Mediante la clusterización realizada con K-Means, podemos perfilar las empresas en 5 grupos diferenciados por los fundamentos financieros. Estos son:

Grupo 0 "Empresas promedio": En este grupo hay un total de 321 empresas. Estas son de crecimiento relativamente bajo, se caracterizan por una poca emisión de acciones y una deuda relativamente baja. Las industrias principales de este grupo son: Servicios financieros (16%), Industria (16%) y salud (14%). Ejemplos de empresas pertenecientes a este grupo son: IBM, Caterpillar y 3M.

Grupo 1 "Crecimiento": En este grupo hay un total de 20 empresas. Se caracterizan por tener un Market Cap muy alto en relación a sus utilidades y free cash flow, tienen un gran crecimiento en los ingresos y utilidades pero una deuda a largo plazo elevada en comparación al free cash flow. Las industrias principales de este grupo son: Tecnología (40%) y Salud (20%). Ejemplos de empresas pertenecientes a este grupo son: Amazon, Tesla, y Dexcom.

Grupo 2 “Diluidor”: En este grupo hay un total de 18 empresas. Estas se caracterizan por tener poca deuda en comparación al free cash flow. Poseen un mayor crecimiento de los ingresos que utilidades y una alta emisión de nuevas acciones. Las industrias principales de este grupo son: Tecnología (16%) e industria (27%). Ejemplos de empresas pertenecientes a este grupo son: L3Harris Technologies y Diamondback Energy.

Grupo 3 “Multiplicadores de la inversión”: En este grupo hay un total de 70 empresas. Estas se caracterizan por dar un muy buen retorno por el capital invertido, poco crecimiento de los ingresos y una baja emisión de acciones. Las industrias principales de este grupo son: Tecnología (31%) y consumo cíclico (17%). Ejemplos de empresas pertenecientes a este grupo son: Amazon, Microsoft y Nike.

Grupo 4 “Cash Makers”: En este grupo hay un total de 37 empresas. Se caracterizan por tener una deuda muy baja, presentar un crecimiento bajo en utilidades e ingresos (más bajo en ingresos), gran crecimiento en el free cash flow y malos retornos en el capital invertido. Las industrias principales de este grupo son: Energía/Utilities (54%) y consumo cíclico (12%). Ejemplos de empresas pertenecientes a este grupo son: Entergy Corp, Boeing y Goldman Sachs.

6. Conclusiones significativas y propuestas de trabajo pendiente.

Se concluyeron 4 recomendaciones de inversión basadas en los resultados anteriores:

En lo que respecta al trabajo pendiente a realizar, nos gustaría trabajar con datos de más empresas, y no solo con las de S&P500, esto con el fin de ampliar la cantidad de datos con los que se trabaja y así poder clusterizar de mejor manera. También nos gustaría realizar una clusterización sobre las empresas del grupo promedio, ya que al ser 321 empresas creemos que estas pueden ser sub clasificadas dentro de su mismo cluster.
Finalmente, también nos gustaría poder realizar este mismo procedimiento pero para analizar la situación de las acciones que se transan en nuestro país.

Bibliografía:

https://www.annuityexpertadvice.com/why-is-the-stock-market-down-crashing/

https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks

https://site.financialmodelingprep.com/developer/docs/

https://www.youtube.com/watch?v=1u6qvel9XnM