S&P 500

Pedro Pablo Beltranena
Andrés Martínez
Nickolas Nolte
Esteban Samayoa

Hipótesis

Descripción de datos

El dataset S&P 500 index stocks pueden ser agrupados o clusterizados basados en su rendimiento general respecto a un rango de tiempo, el cual puede ser analizados usando las variables, opening, low, high and close.

El modelo de agrupación K-mean Clustering es más eficiente para agrupar este tipo de datos respecto a su rendimiento contra otro modelo de agrupación como Hierarchical Clustering.

Analizando el conjunto de datos respecto a un modelo predictivo, podemos predecir correctamente si el precio de un stock tiende al alza o a la baja.

Descripción de los datos

Exploración

El set de datos tiene 7 variables principales

1. Date - object
2. Open - float
3. High - float
4. Low - float
5. Close - float
6. Volume - int
7. Name - object

	date	open	high	low	close	volume	Name
0	2013-02-08	15.07	15.12	14.63	14.75	8407500	AAL
1	2013-02-11	14.89	15.01	14.26	14.46	8882000	AAL
2	2013-02-12	14.45	14.51	14.10	14.27	8126000	AAL
3	2013-02-13	14.30	14.94	14.25	14.66	10259500	AAL
4	2013-02-14	14.94	14.96	13.16	13.99	31879900	AAL

```
1  df.shape
```

✓ 0.0s

```
(619029, 7)
```

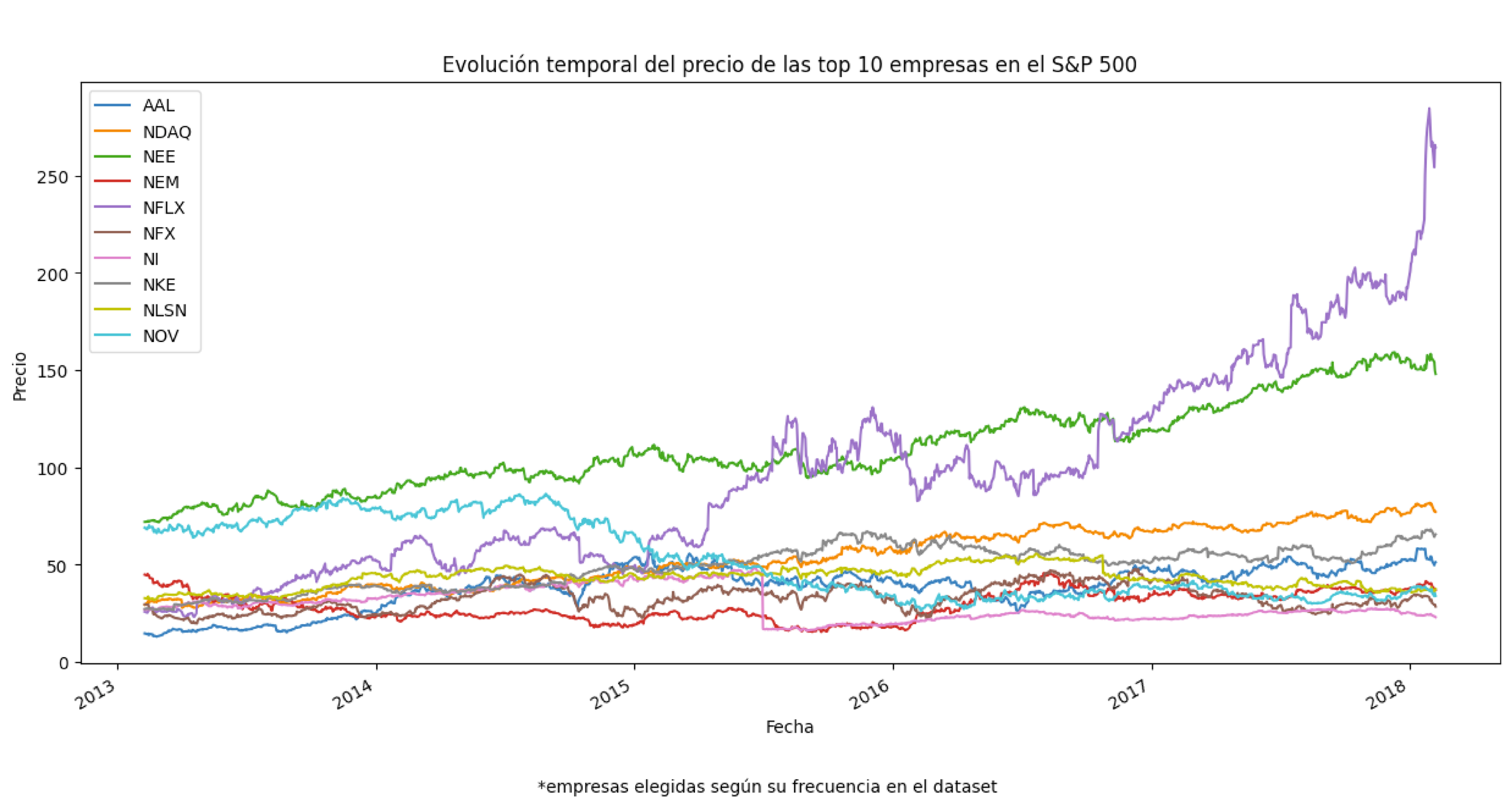
	open	high	low	close	volume
open	1.000000	0.999939	0.999928	0.999872	-0.142705
high	0.999939	1.000000	0.999903	0.999936	-0.142316
low	0.999928	0.999903	1.000000	0.999939	-0.143240
close	0.999872	0.999936	0.999939	1.000000	-0.142802
volume	-0.142705	-0.142316	-0.143240	-0.142802	1.000000

```
1  df.isnull().sum()
```

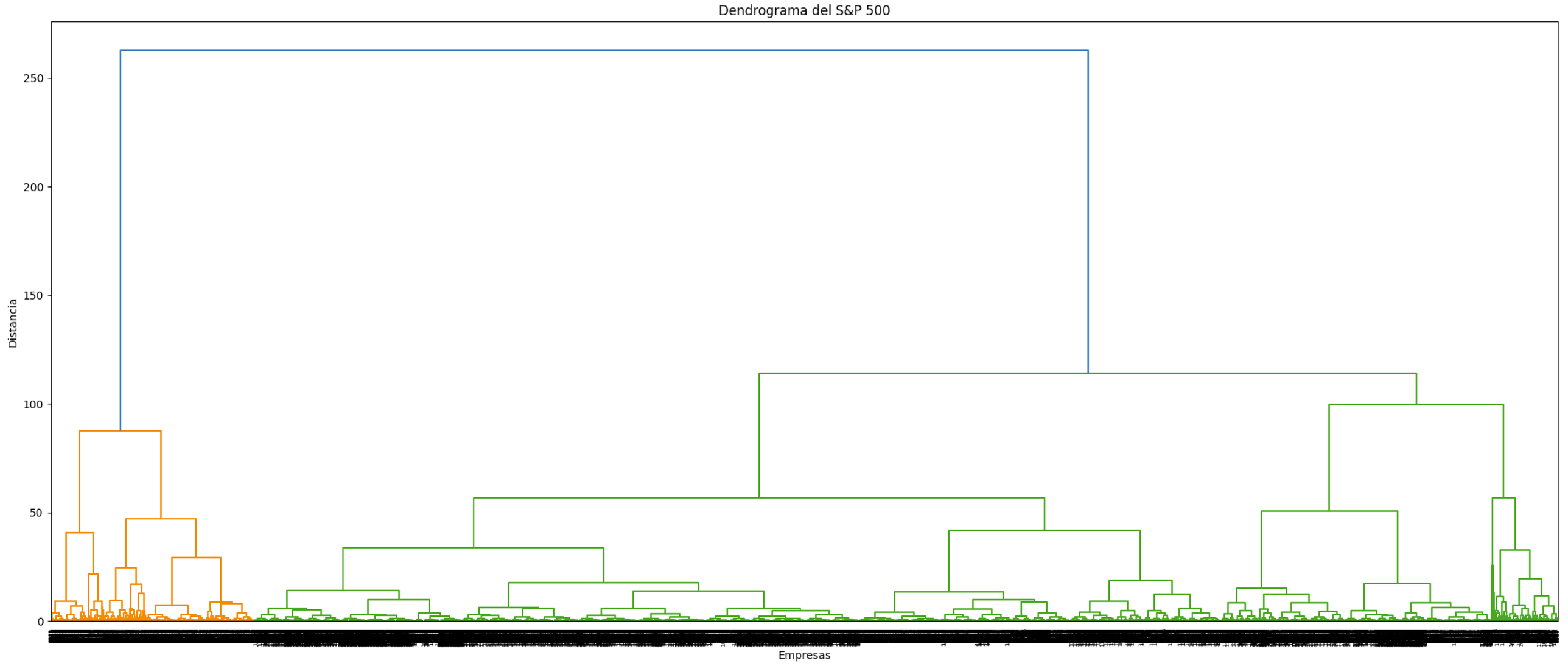
✓ 0.2s

```
date      0
open      0
high      0
low       0
close     0
volume    0
Name      0
dtype: int64
```

```
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   date    619029 non-null  object
 1   open    619029 non-null  float64
 2   high    619029 non-null  float64
 3   low     619029 non-null  float64
 4   close   619029 non-null  float64
 5   volume  619029 non-null  int64
 6   Name    619029 non-null  object
dtypes: float64(4), int64(1), object(2)
memory usage: 33.1+ MB
```



*empresas elegidas según su frecuencia en el dataset


Dendrograma del S&P 500
Distancia
250
200
150
100
50
0
Empresas

Remoción de datos

Aspectos a tomar en cuenta

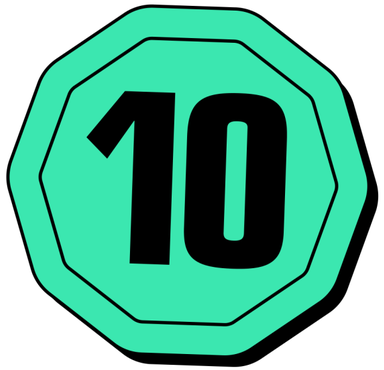
28 NAN

11 en open

8 en high

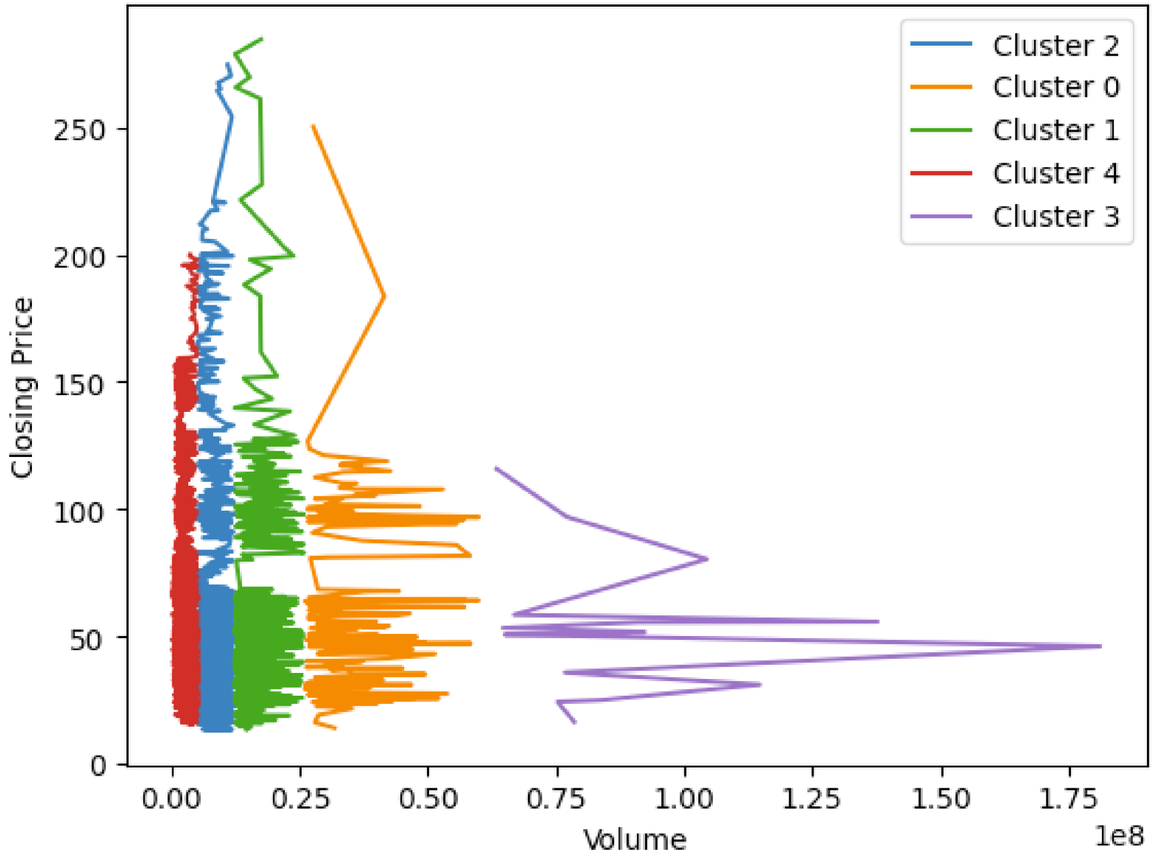
8 en low

Reducción de data set
600,000+ observaciones

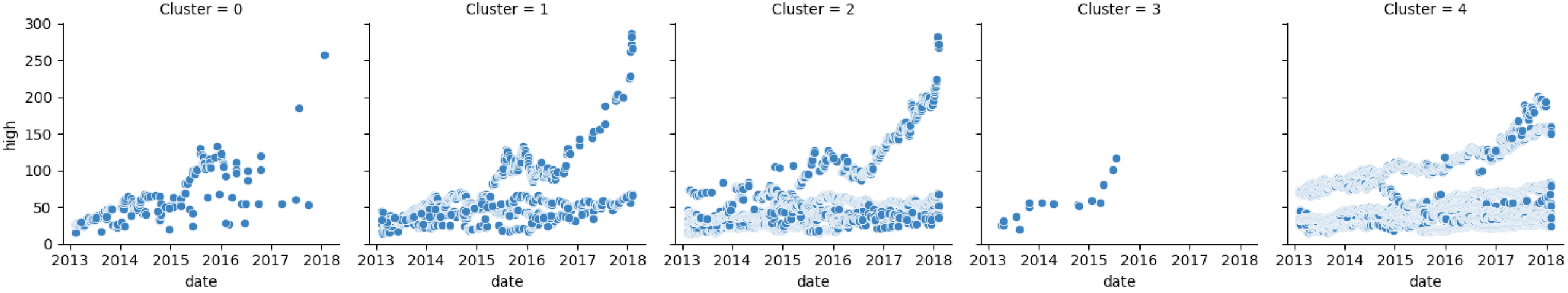


Resultados

	date	open	high	low	close	volume	Name	Cluster
0	2013-02-08	15.07	15.12	14.630	14.75	8407500	AAL	2
1	2013-02-11	14.89	15.01	14.260	14.46	8882000	AAL	2
2	2013-02-12	14.45	14.51	14.100	14.27	8126000	AAL	2
3	2013-02-13	14.30	14.94	14.250	14.66	10259500	AAL	2
4	2013-02-14	14.94	14.96	13.160	13.99	31879900	AAL	0
...	...	...	...	...	...	...	...	...
416328	2018-02-01	36.80	37.53	36.630	37.43	2517830	NOV	4
416329	2018-02-02	36.92	37.06	35.620	35.69	2658658	NOV	4
416330	2018-02-05	35.46	36.29	33.700	33.85	5453975	NOV	2
416331	2018-02-06	33.33	35.80	33.001	35.39	6492866	NOV	2
416332	2018-02-07	35.23	36.10	33.930	34.05	5037110	NOV	4




```
Precisión del modelo: 0.9712184116009456
0         False
1         False
2         False
3          True
4         False
          ...
619024     True
619025    False
619026    False
619027     True
619028     True
Length: 619029, dtype: bool
```


Cluster = 0
Cluster = 1
Cluster = 2
Cluster = 3
Cluster = 4
high
300
250
200
150
100
50
0
2013
2014
2015
2016
2017
2018
date

Conclusiones

Tuvimos que hacer una limpieza de datos eliminando datos vacíos que no nos servían para el análisis. 28 en total distribuidos como lo explicamos anteriormente.

Concluimos la hípotesis de que el *K-mean clustering* es mas eficiente para este tipo de datos que otro modelo de agrupación como el *Hierarchical Clustering*.

Aplicamos el *Random Forest Classifier,* no nos dio resultados relevantes para ver cual de los metodos era el más adecuado. El porcentaje de positivos y negativos esta casi que igual.

Curiosidades

- Empresa de energía tiene una caída bastante pronunciada y coincide en tiempo con los Acuerdo de Paris, que tienen como objetivo: *aumentar la capacidad de adaptación y reducir la vulnerabilidad frente a los impactos del cambio climático de todos los países*

- *Mientras que NEE, es la segunda mejor empresa con rendimiento a través del tiempo. Pero su enfoque energetico cumple con acuerdos internacionales o fuentes renovables.*