Universidad de San Andrés

Escuela de Administración y Negocios
Magister en Finanzas

Descubrimiento no supervisado y análisis estadístico de patrones de velas para pronosticar retornos futuros de activos financieros

Autor: Guido Falcucci
DNI: 35229430
Director del Trabajo Final de Graduación: Gabriel Basaluzzo

Buenos Aires, 21 de Diciembre de 2020

Índice

Resumen
En este trabajo se demuestra que ciertos patrones observables en las series de precios de activos financieros tienen impacto sobre sus retornos. Las contribuciones originales del trabajo son las siguientes: a) modelar los patrones mediante algoritmos de aprendizaje no supervisados[1] ; b) definir una entrada apropiada para los modelos que incorpore información acerca de la forma del patrón, tendencia y volatilidad; c) analizar el impacto de los patrones identificados sobre los retornos de los activos mediante pruebas no paramétricas; d) predecir órdenes que maximicen la probabilidad de éxito.

1. Introducción

La forma más común de visualizar la serie de precios de un activo es mediante el gráfico de velas japonesas (Candlestick chart, 2020). Como su nombre lo indica, el método fue desarrollado en Japón durante el siglo XVIII por comerciantes de arroz y fue incorporado en el mundo occidental un siglo después. Más aún, una rama importante del *trading* llamada análisis técnico[2] se basa en el análisis patrones de velas japonesas para decidir en qué activos invertir (Morris, 2006). A lo largo de las últimas décadas y debido al sostenido incremento en el poder de procesamiento de las computadoras, se han desarrollado diversos algoritmos que operan automáticamente en los mercados mediante el reconocimiento de patrones de velas japonesas (Sengcho T. Chou and Lai, 1997; Martiny, 2012; Leslie C.O. Tiong and Lee, 2014; Jingpei Dan and Zhang, 2014).
En este trabajo se usan modelos que descubren de manera automática los distintos patrones, es decir, no requieren de conocimiento previo acerca de los mismos ni la interacción activa de un humano. Además, se prueba de manera estadística mediante una prueba no paramétrica que algunas de las secuencias de velas descubiertas tienen impacto *a posteriori* sobre los retornos de los activos. En cualquiera de los tres casos: impacto positivo, negativo, o sin impacto, cuando los estadísticos de los *tests* lo indiquen, se pueden armar estrategias de *trading* que maximicen la probabilidad de éxito.
Este trabajo está estructurado de la siguiente manera: en la sección 2 se explica cómo está formada una vela japonesa; en la sección 3 se describe el agrupamiento no supervisado y los dos algoritmos que se usaron, incluyendo

[1] Es un método de aprendizaje automático ajustado a las observaciones que no requiere conocimiento *a priori*.

[2] Es el estudio de un activo del mercado a través de gráficos de series de precios, con el propósito de predecir futuras tendencias.

las medidas de distancia más comunes; en la sección 4 se define la entrada de los modelos; en la sección 5 se detalla el procedimiento para analizar los resultados obtenidos, incluyendo la prueba no paramétrica empleada; en la sección 6 se especifican los datos utilizados, incluyendo criterios para filtrar activos ilíquidos y en la sección 7 se presentan los resultados, incluyendo los patrones descubiertos por los modelos, *ranking* según su significancia y tasa de éxito en la predicción de *trades*. Finalmente, en la sección 8 se presentan las conclusiones y se proponen lineamientos para investigaciones futuras.
Además, se incluyen los anexos A y B. El A contiene el enlace al código *python* desarrollado con el que se automatizaron todos los pasos presentados a lo largo del escrito y una breve descripción de sus módulos; el B lista los identificadores de las acciones analizadas.

2. Velas Japonesas

Una vela japonesa es una barra que refleja la situación de un activo del mercado en un periodo de tiempo determinado. Muestra los siguientes cuatro datos acerca del precio del activo durante ese periodo: el precio de apertura, de cierre, máximo y mínimo. El color de la barra refleja cuál es la tendencia. Si la barra es blanca o verde, la tendencia es alcista, es decir, el precio de cierre es mayor al de apertura. Por el contrario, si la barra es negra o roja, la tendencia es bajista, es decir, el precio de cierre es menor al de apertura.
En general, para analizar tendencias de corto plazo se utilizan periodos de un minuto, una hora y un día. Para tendencias de más largo plazo, lo común es una semana, un mes, o tres meses. En la Figura 1 se puede observar una típica vela alcista y bajista.
De acuerdo a las combinaciones que se den con los cuatro atributos mencionados, se pueden encontrar velas de gran, poco o sin cuerpo. En las Figuras 2 a 5 se muestran algunas de ellas.

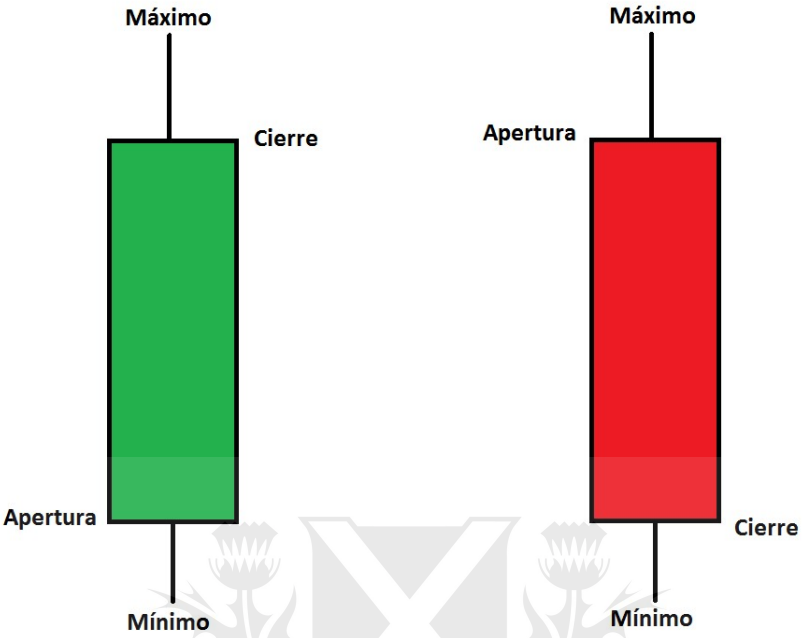


Figura 1: Vela japonesa alcista y bajista

(a) Martillo **(b)** Martillo invertido

Figura 2: Velas martillo

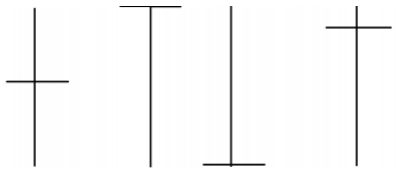


Figura 5: Vela doji[3]

[3]Vela que refleja incertidumbre en el mercado ya que no define una tendencia clara.

(a) Envolvente alzista **(b)** Envolvente bajista

Figura 3: Velas envolventes

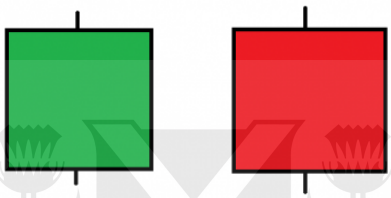


Figura 4: Vela de cuerpo chico

En la figura 6 se muestra un gráfico de velas japonesas para el Recibo de Depósito Americano[4] o *ADR (American Depositary Receipt)* del Grupo Galicia con frecuencia diaria, es decir, cada vela muestra el precio de apertura, máximo, mínimo y de cierre para cada día.

Figura 6: ADR Grupo Galicia. Velas de frecuencia diaria

[4]Es un título físico que respalda el depósito en un banco estadounidense de acciones de compañías cuyas sociedades fueron constituidas fuera de aquel país, de manera de poder operar las acciones de la compañía como si fueran cualquiera otra de ese mercado.

3. Agrupamiento no supervisado

3.1. Agrupamiento

El agrupamiento o *clustering* es el proceso por el cual se clasifican individuos en grupos mediante alguna medida de similitud. La clasificación de objetos es un proceso que realizan los humanos desde temprana edad. Por ejemplo, un niño realiza este proceso cuando observa una fotografía y reconoce personas, vehículos, edificios, plantas, animales, etc. Más aún, la clasificación jerárquica es una parte importante de las ciencias naturales. Biólogos han dedicado años para agrupar a los seres vivos en reinos, clases, familias y especies y de esta forma lograr un mejor entendimiento del comportamiento de los distintos individuos (Pang-Ning Tan and Kumar, 2018).

3.2. Agrupamiento no supervisado

El agrupamiento no supervisado o *unsupervised clustering* es el proceso por el cual se clasifican individuos en grupos conociendo únicamente datos de los mismos y alguna medida que establezca cuán parecido es un individuo a otro. Es la forma de aprendizaje más natural que tienen los humanos. Por ejemplo, un niño luego de observar un eucalipto, un olmo y un jacaranda puede agruparlos en un grupo (sin saber que son árboles) y distinguirlo de otro compuesto por vehículos mediante alguna medida de similitud mental establecida por el niño que incorpora color, forma, movimiento, etc. Cabe diferenciar este método del aprendizaje supervisado[5]; siguiendo el ejemplo, consistiría en mostrarle al niño una foto de un perro, y pedirle que identifique otros animales similares.
En este trabajo se utilizaron dos algoritmos de *clustering* no supervisados que se describen en la Sección 3.4. En la Figura 7 se muestra el resultado de un agrupamiento no supervisado con perros y gatos. Es importante observar que si el *input* del modelo no captura diferencias entre ambas especies, se resuelve la agrupación incorrecta de la Figura 8. Esto ocurriría, por ejemplo, si la entrada del modelo contemplara únicamente cantidad de patas, ojos y orejas (ambos animales poseen la misma cantidad y no hay manera de distinguirlos meramente con esa información).

[5]El modelo no sólo recibe datos de entrada sino también datos de salida.

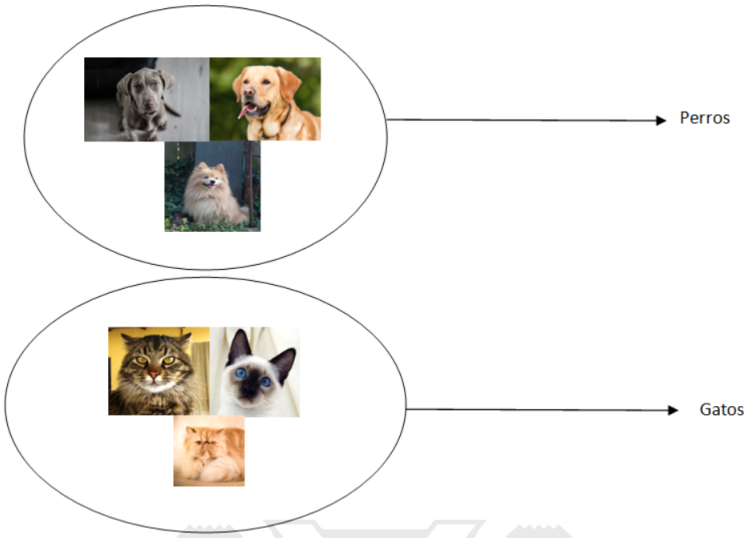


Figura 7: Agrupamiento

Figura 8: Agrupamiento indeseable

3.3. Medidas de similitud

Para poder realizar el agrupamiento es necesario definir una medida que establezca cuán similares son dos individuos.
En general, se utilizan medidas de distancia $d(P, Q)$ que cumplen con las siguientes condiciones:

$$d(P,Q) = d(Q,P),$$
$$d(P,Q) > 0 \text{ si } P \neq Q,$$
$$d(P,Q) = 0 \text{ si } P = Q,$$
$$d(P,Q) \leq d(P,R) + d(R,Q).$$

donde $R = (r_1, r_2, ..., r_n)$ es un punto intermedio y $P = (p_1, p_2, ..., p_n)$ y $Q = (q_1, q_2, ..., q_n)$ son datos que caracterizan al individuo P y Q respectivamente. Las distancias más comunes (Shraddha Pandit, 2011) son la distancia euclidiana, Manhattan, *bit-vector* y los índices de Jaccard, coseno y Dice.
La más habitual y la que se usó en este trabajo es la euclidiana definida como $D_E(P,Q)$ en el espacio euclídeo n-dimensional donde $P = (p_1, p_2, ..., p_n)$ y $Q = (q_1, q_2, ..., q_n)$.

$$D_E(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}. \quad (1)$$

Para $n \leq 3$ esta magnitud se puede graficar. En la Figura 9 se muestra el caso para $n = 2$.

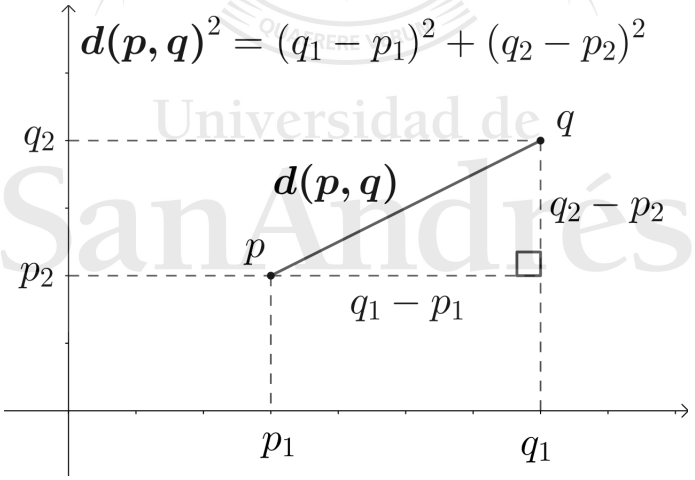


Figura 9: Distancia euclidiana en R^2

3.4. Algoritmos de agrupamiento no supervisado

Los algoritmos de agrupamiento se pueden clasificar mediante diversos enfoques. El más conocido se basa en el resultado que producen. Así, se pueden distinguir dos tipos: divisivos o de particionamiento y jerárquicos (Kaufman

and Rousseeuw, 1990). A continuación se describen cuatro algoritmos: *K-MEANS*, de agrupamiento jerárquico, basado en densidad o *DBSCAN*(*Density-based spatial clustering of applications with noise*) y *HDBSCAN* (Ricardo J. G. B. Campello and Sander, 2013). En este trabajo se optó por *K-MEANS* y *HDBSCAN*. Este último se basa en los otros dos.

3.4.1. K-MEANS

Es uno de los algoritmos más comunes de clasificación. Su objetivo es la partición de un conjunto de n observaciones en k grupos, en donde cada observación o_i pertenece al grupo cuyo valor medio es más cercano.
En la Tabla 1 se incluye el pseudocódigo. Notar que la asignación al centroide más cercano se hará de acuerdo a la función distancia elegida. Además, cabe mencionar que la cantidad de *clusters* K es un parámetro de entrada del algoritmo.

---

1. Seleccionar K puntos aleatorios como centroides.
2. **Repetir**
3. Formar K grupos asignando cada punto al centroide más cercano.
4. Re-computar los centroides de cada grupo
5. **Hasta que los centroides no cambien**

Tabla 1: Algoritmo de K-medias

En la Figura 10 se incluye como ejemplo la salida del algoritmo de *K-MEANS* para datos aleatorios. En este caso, cada observación es un par ordenado (x, y). Se usaron distintos colores para identificar los *clusters* y se marcó el centro de cada uno con una "X".

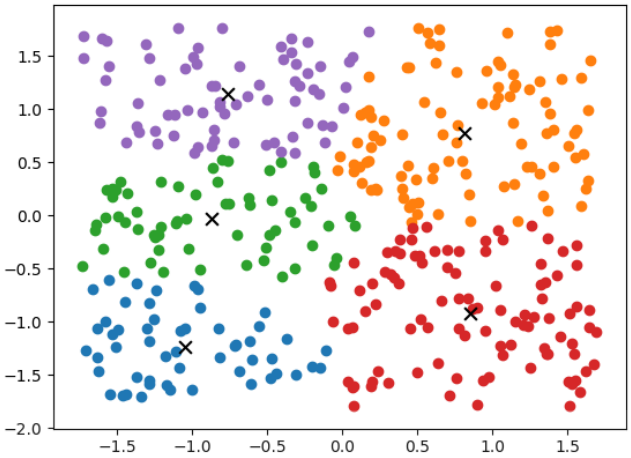


Figura 10: Resultado del algoritmo para $K = 5$

3.4.2. Algoritmo de agrupamiento jerárquico

Este algoritmo es una de las bases de HDBSCAN. La forma más sencilla de entender el modelo es gráficamente, mediante un dendrograma (figura 11), que muestra no sólo la relación grupo-subgrupo sino también el orden en el que los grupos se unieron.
Existen diversos algoritmos modernos para armar la estructura jerárquica (Muller, 2011) pero la gran mayoría comparte el siguiente enfoque general: comenzar con grupos de puntos individuales y unir sucesivamente los dos *clusters* más cercanos hasta que sólo quede un único *cluster* padre.

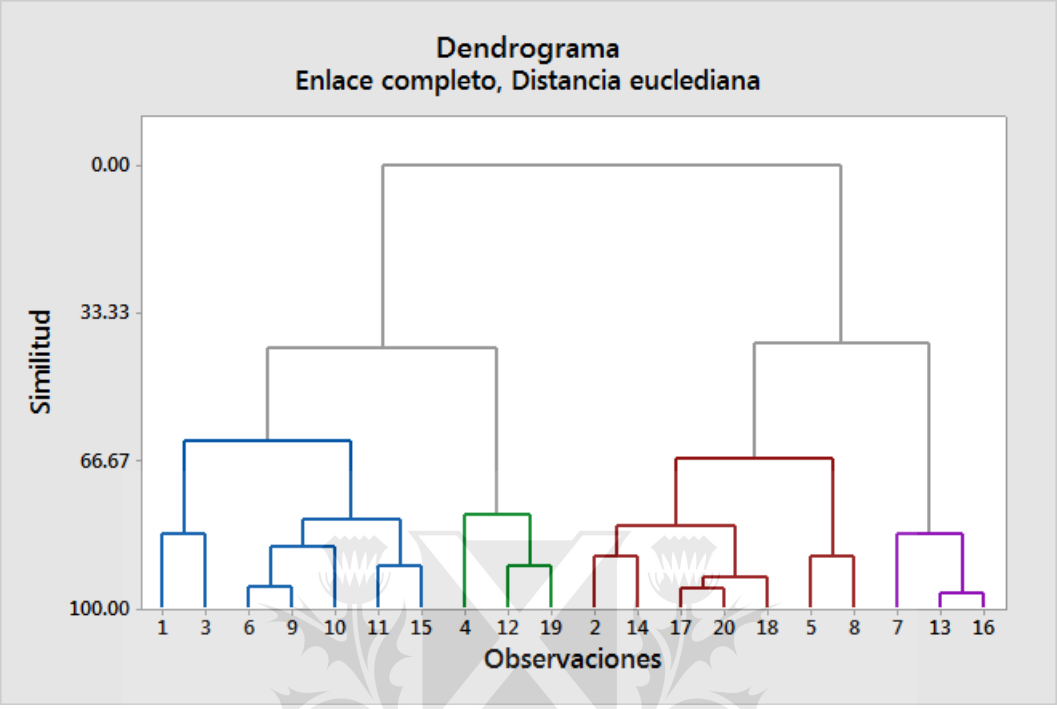
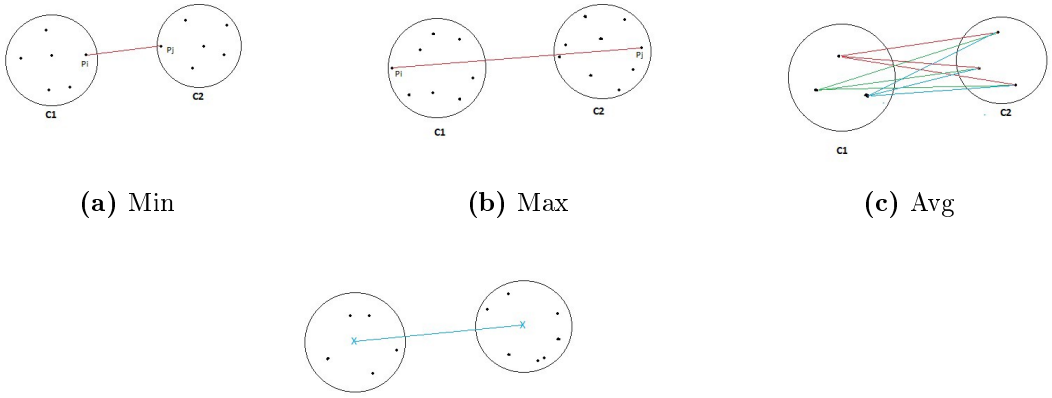


Figura 11

Proximidad entre grupos

En la tabla 2 se incluye el pseudocódigo del algoritmo. Su operación más importante es la de proximidad entre dos *clusters*. Existen cuatro operaciones comunes para computarla: la distancia entre las dos puntos más cercanos (**MIN**); la distancia entre los dos puntos más lejanos (**MAX**), un promedio entre la distancia de todos los puntos de ambos *clusters* (**AVG**) y finalmente la de distancia entre centroides (**CENT**). Se representan las operaciones en la figura 12.



(a) Min **(b)** Max **(c)** Avg

(d) Cent

Figura 12: Operaciones de proximidad

---

1. Computar la matriz de proximidad.
2. **Repetir**
   2.1. Unir los dos *clusters* mas cercanos (con función de proximidad).
   2.2. Actualizar la matriz de proximidad para reflejar la cercanía entre el nuevo *cluster* y los originales.
3. **Hasta que quede un único *cluster***

Tabla 2: Algoritmo de agrupamiento jerárquico

3.4.3. Algoritmo DBSCAN

DBSCAN es el otro fundamento de HDBSCAN y se basa en la densidad de los *clusters*. El modelo mide la densidad de un grupo mediante la cantidad de observaciones o_i que se encuentran en un determinado radio Eps de su centro. Por ejemplo, en la figura 13 hay seis puntos a una distancia $\leq Eps$, incluyendo a A.
Con este enfoque, cada punto se clasifica en alguna de las siguientes tres categorías: núcleo (*core*), alcanzables (*border*) y ruido (*noise*).

- ***Core***: Si existen al menos $MinPts$ puntos que se encuentran a una distancia $D(p_i, p_j) \leq Eps$, donde D es una función como la de la ecuación (1).
- ***Border***: No es *core* pero se encuentra a una distancia $\leq Eps$ de algún punto *core*.
- ***Noise***: No es ni *core* ni *border*.

En la figura 14 se muestran las tres clasificaciones de puntos para $MinPts = 7$ y $Eps = \epsilon$.

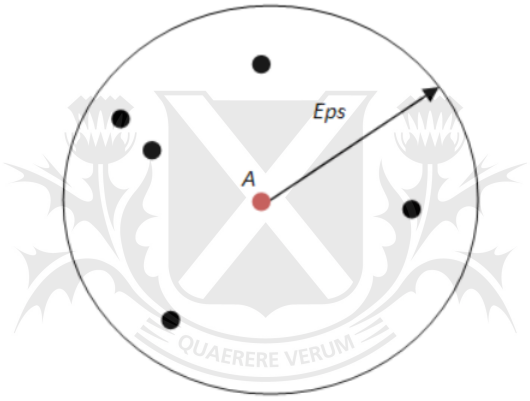


Figura 13: Densidad basada en centros

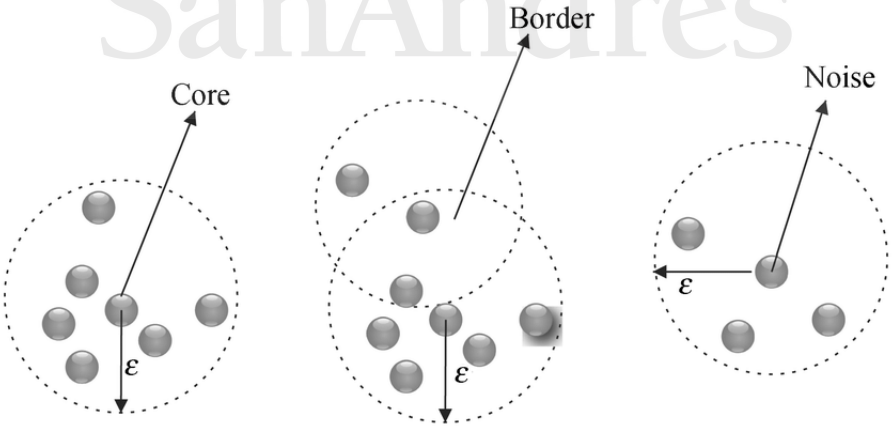


Figura 14: Clasificación de puntos para el enfoque basado en centros

Se incluye en la tabla 3 el pseudocódigo de *DBSCAN* y en la figura 15 una salida ejemplo del mismo.

1. Clasificar todos los puntos como *core*, *border* o *noise*.
2. Eliminar los puntos *noise*.
3. Asignar un *cluster* a cada punto *core*.
4. Asignar cada punto *border* al *cluster* más cercano.

Tabla 3: Algoritmo DBSCAN

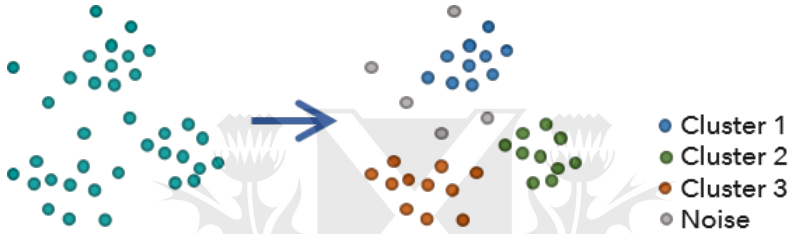


Figura 15: Ejemplo clusters hallados por el algoritmo DBSCAN

Selección de los parámetros

Una de las etapas más importantes de este algoritmo es la selección de los parámetros Eps y $MinPts$. Una forma simple de elegirlos es analizar el comportamiento de $K_{dist}(p) = D(p, k^{ésimo})$. Para puntos $k^{ésimo}$ pertenecientes al *cluster* de p, $K_{dist}(p)$ será un valor chico en comparación a la misma medida para puntos *noise*. Por lo tanto, si se computa la magnitud $K_{dist}(p_i)$ de todas las observaciones o_i para algún k, se ordenan de manera creciente y se grafican, se notará un cambio brusco en K_{dist} para un valor apropiado de Eps. Si se selecciona esta distancia como Eps y $MinPts = k$, entonces serán puntos *core* el siguiente conjunto $\{p_i : K_{dist}(p_i) \leq Eps\}$. Además, $border \cup noise = \{p_i : K_{dist}(p_i) > Eps\}$. En la figura 16 se evidencia que 0.15 resulta adecuado para $Eps\, cuando\, k = 5$.

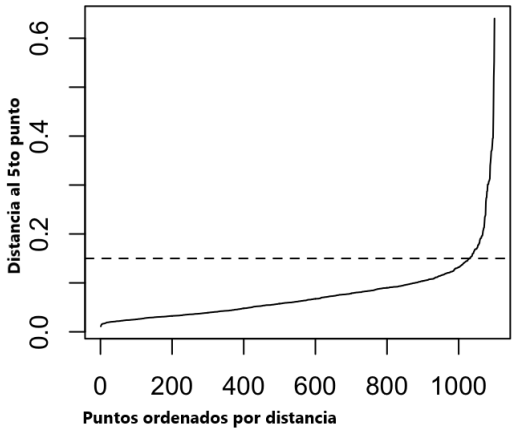


Figura 16: Distancia $K_{dist}(p_i) = D(p, k^5)$

3.4.4. Comparación: Algoritmo K-MEANS y DBSCAN

En la tabla 4 se detallan las diferencias entre los dos algoritmos. Las principales ventajas de *DBSCAN* sobre *K-MEANS* es que localiza zonas de alta y baja densidad y que no necesita el número total de *clusters* como parámetro de entrada ya que lo calcula de manera indirecta.

	K-MEANS	DHBSCAN
1	Los *clusters* tienen forma convexa y son del mismo tamaño.	Los *clusters* pueden tener forma y tamaño distinto.
2	El resultado depende de la cantidad de *clusters* especificados.	La cantidad de *clusters* no es un parámetro del modelo.
3	Es más eficiente en el manejo de gran conjunto de datos.	No es tan eficiente en el manejo de grande datos.
4	No es eficiente en el manejo de datos con mucho ruido y *outliers* .	Una de sus principales ventajas es que maneja de manera efectiva datos con ruido y *outliers*.
5	Puede agrupar anomalías en *clusters* "normales".	Otra de sus principales ventajas es que localiza zonas de alta densidad que están separadas de otras mediante zonas de baja densidad (común en anomalías)

Tabla 4: K-MEANS y DBSCAN

En la figura 17 se muestran las clasificaciones producidas por los dos algoritmos para un conjunto de puntos en R^2 de densidad variable.

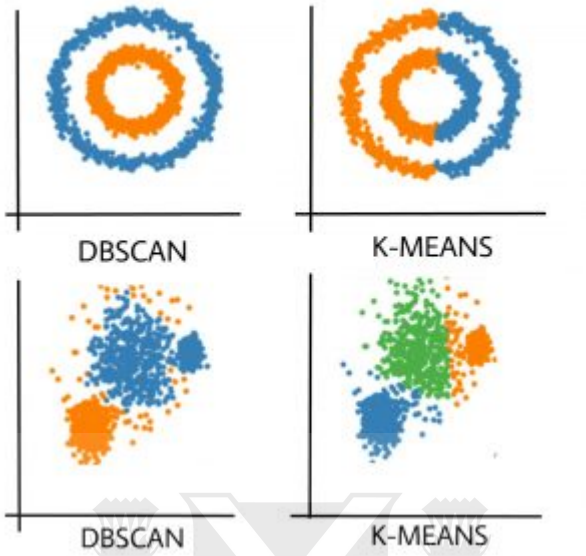


Figura 17: K-MEANS y DBSCAN

3.4.5. Comparación: Algoritmo de agrupamiento jerárquico y DBSCAN

En la tabla 5 se compara el modelo jerárquico con el basado en densidad. Una importante ventaja de ambos modelos es que la cantidad de *clusters* no es un parámetro de entrada.

	Ventajas	**Desventajas**
Agrupamiento Jerárquico	▪ La cantidad de *clusters* n no es un parámetro de entrada. ▪ Fácil de implementar.	▪ Complejidad computacional alta ($O(n^2 \log n)$). ▪ Sensible a *outliers*. ▪ Dificultad ante *clusters* de distinto tamaño y formas convexas. ▪ No existe una función objetivo que se minimice.
DBSCAN	▪ La cantidad de *clusters* n no es un parámetro de entrada. ▪ Capaz de agrupar observaciones en *clusters* de distintas formas. ▪ Identifica y elimina *outliers*	▪ Dificultad ante densidades cambiantes. ▪ Dificultad ante dimensionalidad alta en los datos.

Tabla 5: Agrupamiento jerárquico y DBSCAN

3.4.6. Algoritmo HDBSCAN

El segundo algoritmo utilizado es una versión optimizada de *DBSCAN*, desarrollado por los mismos autores (Ricardo J. G. B. Campello and Sander, 2013) que hereda los beneficios de los algoritmos jerárquicos y de densidad y que además resuelve las desventajas descriptas en la tabla 5. Extiende *DBSCAN* convirtiéndolo en un algoritmo jerárquico extrayendo luego una estructura plana de *clusters*. El algoritmo se puede descomponer en las siguientes etapas (How HDBSCAN Works, 2020):

1. Transformar el espacio según la densidad.

2. Armar el mínimo árbol de expansión basado en el grafo de distancias ponderadas.

3. Construir *clusters* jerárquicos de componentes conectados.

4. Condensar la estructura jerárquica basándose en los *clusters* de menor tamaño.

5. Extraer los clusters más "estables".

Los dos parámetros más importantes del modelo son:

- **min_cluster_size**: Tamaño mínimo de un *cluster*.
- **min_samples** : Mínima cantidad de puntos próximos a un punto para que sea considerado *core* (*MinPts* de la sub-sección 3.4.3).

4. Definición de la entrada del modelo

La entrada de los modelos explicados en la sección anterior incorpora la siguiente información de una secuencia de velas:

- Precios de apertura, máximo, mínimo y de cierre.
- Forma.
- Tendencia (alcista, bajista o sin marcada tendencia) previa al patrón.
- Volatilidad (alta, media o baja) previa al patrón.
- Volumen (alto, medio, bajo).
- Fuerza de la tendencia.
- *Momentum* del activo financiero.

Para incorporar información de los precios se denota $CS_{i\tau}$ a la vela japonesa del activo i en la observación τ

$$CS_{i\tau} = (\frac{H_{i\tau}}{O_{i\tau}}, \frac{L_{i\tau}}{O_{i\tau}}, \frac{C_{i\tau}}{O_{i\tau}}), \tag{2}$$

donde $O_{i\tau}$, $H_{i\tau}$, $L_{i\tau}$, $C_{i\tau}$ son los cuatro precios que la definen: apertura, máximo, mínimo y de cierre respectivamente. Notar que se normalizan $H_{i\tau}$, $L_{i\tau}$, $C_{i\tau}$ por $O_{i\tau}$ para poder trabajar con precios de diferentes compañías indistintamente. De esta manera se reduce una dimensión.

Para capturar información acerca de la forma del patrón se define $M_{i\tau}$ como punto medio de $CS_{i\tau}$:

$$M_{i\tau} = \frac{H_{i\tau} - L_{i\tau}}{2} + L_{i\tau}.$$

Usando los puntos medios de dos velas consecutivas, se define la posición relativa $RP_{i\tau(\tau-1)}$ de la vela $i\tau$ con respecto a la $i(\tau-1)$ como:

$$RP_{i\tau(\tau-1)} = \frac{M_\tau - M_{\tau-1}}{M_{\tau-1}}. \tag{3}$$

Se define $trend_{is}$ a la tendencia del activo i previa a la secuencia s como:

$$trend_{is} = \begin{cases} 1 & \text{si la tendencia anterior a } s \text{ es alcista,} \\ -1 & \text{si la tendencia anterior a } s \text{ es bajista,} \\ 0 & \text{en caso contrario.} \end{cases} \quad (4)$$

En lugar de establecer si la tendencia es alcista o bajista de la manera tradicional, es decir, mediante máximos y mínimos ascendentes para la tendencia alcista y descendentes para la bajista, se incorporó información de medias móviles para que la función indicadora de tendencia no sea tan estricta.
Se denota $SMA_{i\tau n}$ a la media móvil simple del activo i en la observación τ de de n periodos como:

$$SMA_{i\tau n} = \frac{1}{n} \sum_{t=\tau-n}^{\tau-1} C_{it},$$

donde C_{it} es el precio de cierre del activo i en la observación t.
La tendencia de la secuencia s del activo $i <CS_{i0}, CS_{i1}, ...CS_{im} >$ es alcista si $SMA_{i08} > SMA_{i020} > SMA_{i050} > SMA_{i0100}$ y es bajista si $SMA_{i08} < SMA_{i020} < SMA_{i050} < SMA_{i0100}$. En caso contrario no hay tendencia. Es decir, es alcista, si las medias móviles más "rápidas" (las de menor periodos), se encuentran por encima de las más lentas (las de mayor periodo) y es bajista si se encuentran por debajo. Luego, $trend_{is}$ queda definido como:

$$trend_{is} = \begin{cases} 1 & SMA_{i8} > SMA_{i20} > SMA_{i50} > SMA_{i100}, \\ -1 & SMA_{i8} < SMA_{i20} < SMA_{i50} < SMA_{i100}, \\ 0 & \text{en caso contrario.} \end{cases} \quad (5)$$

En la figuras 18 y 19 se ilustran tendencias alcista y bajista para el ADR de NIO Inc y para Exxon-Mobil respectivamente. Las medias móviles tienen periodos de 8, 20, 50 y 200.

Figura 18: Tendencia alcista NIO

Figura 19: Tendencia bajista XOM

Se denota $L\sigma_{is}^{MK}$ al nivel de volatilidad histórica de la secuencia de velas s basado en M observaciones anteriores a s y relativa a la volatilidad del activo medida a partir de K observaciones.

$$L\sigma_{is}^{MK} = \begin{cases} 1 & \text{si el grado de volatilidad es alto,} \\ -1 & \text{si el grado de volatilidad es bajo,} \\ 0 & \text{en caso contrario.} \end{cases} \tag{6}$$

Para poder establecer el nivel de volatilidad se utilizó el indicador *ATR(Average True Range)* introducido por J. Welles Wilder (1978). Se definen a continuación $TR_{i\tau}$, ATR_i^N y ΔATR_i^{MK}:

$$\begin{aligned} TR_{i\tau} &= \text{máx}(H_{i\tau} - L_{i\tau},\, |H_{i\tau} - C_{i\tau-1}|,\, |L_{i\tau} - C_{i\tau-1}|), \\ ATR_i^N &= \frac{1}{N}\sum_{n=0}^{N} TR_{in}, \\ \Delta ATR_i^{MK} &= \frac{ATR_i^M - ATR_i^K}{ATR_i^K}. \end{aligned} \tag{7}$$

donde K y M son cantidad de observaciones y $M << K$.
Se re-define entonces $L\sigma_{is}^{MK}$ en base a la ecuación (7) como:

$$L\sigma_{is}^{MK} = \begin{cases} 1 & \text{si } \Delta ATR_i^{MK} > \ \ 0{,}2, \\ -1 & \text{si } \Delta ATR_i^{MK} < -0{,}2, \\ 0 & \text{en caso contrario.} \end{cases} \tag{8}$$

Para ilustrar el cálculo de ΔATR_i^{MK} se incluye la figura 20. En el ejemplo, la cantidad de velas que constituyen un patrón es cinco ($s = 5$), $M = 10$ y $K = 30$. El ATR_{APPL}^{10} y ATR_{APPL}^{30} en la vela marcada con un punto negro son 1.56 y 0.94 respectivamente. Luego $\Delta ATR_{APPL}^{10,30} = \frac{1{,}56-0{,}94}{0{,}94} = 0{,}66$. Por lo tanto $L\sigma_{APPL,s}^{10,30} = 1$, es decir , se considera elevada la volatilidad del patrón de la figura relativa a la medida a partir de 30 observaciones.

Figura 20: Ejemplo ATR aplicado a APPL

Se denota $V_{i\tau B}$ al volumen relativo de la observación τ con respecto a B observaciones anteriores como:

$$V_{i\tau B} = \frac{V_{it}}{\frac{1}{B}\sum_{t=\tau-B}^{\tau-1} V_{it}}, \tag{9}$$

donde V_{it} es el volumen del día t. El denominador de la ecuación (9) es la media móvil simple del volumen.

Se denota también

$$ADX_{i\tau}^{L} \tag{10}$$

al Índice Direccional Medio (*Average Directional Index*, en inglés) de L periodos desarrollado por Wilder (1978).

Por último se denota

$$RSI_{i\tau}^{P} \tag{11}$$

al Índice de Fuerza Relativa (*Relative Strength Index*, en inglés) de P periodos, desarrollado por Wilder (1978).

En base a las ecuaciones (2), (3), (5), (8), (9), (10) y (11) se define el patrón P del activo i formado a partir de una secuencia de s velas como:

$$P(i,s,M,K,B,L,P) = \left(C_{i0},\, C_{i1},...,C_{is},\, RP_{i01},\, RP_{i12},...,RP_{i(s-1)s},\, trend_{is},\, L\sigma_{is}^{MK},\, V_{is200}^{B},\, ADX_{is}^{L},\, RSI_{is}^{P}\right). \tag{12}$$

Luego, para un patrón definido a partir de s velas, un punto de la entrada del modelo tiene $\mathbf{4(s+1)}$ parámetros:

$$\begin{aligned}\text{Cantidad de parámetros } &= 3s\ precios + \\ &\quad (s-1)\ deforma + \\ &\quad 1\ de\ tendencia + \\ &\quad 1\ de\ volatilidad + \\ &\quad 1\ de\ volumen + \\ &\quad 1\ de\ fuerza\ de\ tendencia + \\ &\quad 1\ de\ momentum \\ &= 3s + (s-1) + 1 + 1 + 1 + 1 + 1 \\ &= 4s + 4\end{aligned}$$

5. Análisis

La salida de los modelos descriptos en la sección anterior son los *clusters*. Dado un *cluster* compuesto por secuencias de velas japonesas, se analizó si los retornos de los precios posteriores al patrón son significativamente distintos de cero y de esa manera se estableció si el *cluster* en sí es o no significativo. Para determinar dicha significancia, se empleó una prueba no paramétrica. Finalmente, con precios de activos posteriores a la muestra con la que se entrenaron los modelos, se predijeron órdenes de compra o venta. Para hacer la predicción, primero se hallaron los *clusters* que mejor representan a los nuevos patrones y en el caso de corresponder a *clusters* significativos, se midió el retorno entre el precio de cierre del patrón y tres o siete días posteriores. La estrategia será "larga" o de compra del activo financiero si el *cluster* en cuestión tiene impacto positivo en lo retornos. Análogamente, será una estrategia de venta en corto[6] si el impacto es negativo.

5.1. Pruebas no paramétricas

Las pruebas paramétricas, usadas frecuentemente en el análisis de eventos en economía y finanzas (MacKinlay, 1997), tienen la desventaja que requieren suposiciones acerca de la distribución probabilista de los retornos de los activos financieros. En general, se asume que los retornos tienen una distribución normal o gaussiana. Para no tener que hacer tales suposiciones, se analiza la significancia de los patrones detectados por los modelos descriptos en la sección 3 mediante la prueba no paramétrica de los rangos con signo de Wilcoxon Wilcoxon (1945).

5.2. Prueba de los rangos con signo de Wilcoxon

El estadístico definido por Wilcoxon (1945) es:

$$W = Min(S^+, S^-),$$

donde S^+ es la cantidad de retornos positivos y S^- de negativos. Para $n > 30$ la distribución de W se aproxima a una distribución normal con media μ_w y desvío estándar σ_w.

$$\mu_w = \frac{n(n+1)}{4},$$

[6] Se asume que no se posee el activo financiero

$$\sigma_w = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

Se puede calcular el estadístico Z de distribución aproximadamente normal de media 0 y varianza 1 de la siguiente manera:

$$Z_0 = \frac{W - \mu_w}{\sigma_w}$$

Se utilizó un nivel de significación (α) de 0.05.

5.3. Predicción

Para hallar la estrategia óptima de *trading* para precios de activos *out-of-sample* se utilizaron funciones de predicción de *K-MEANS* y *HDBSCAN*. Tales operaciones encuentran de manera rápida los *clusters* más cercanos para una secuencia de velas no utilizada en la etapa del entrenamiento del modelo. Una alternativa sería re-computar todos los *clusters* y realizar el análisis no paramétrico descripto en la sección anterior pero tal operación tiene un costo computacional elevado. Para que esta opción sea eficiente se recomienda el uso de *cloud computing*.
En la figura 21 se resumen los distintos pasos del análisis, desde el entrenamiento de los modelos hasta la selección de la estrategia.

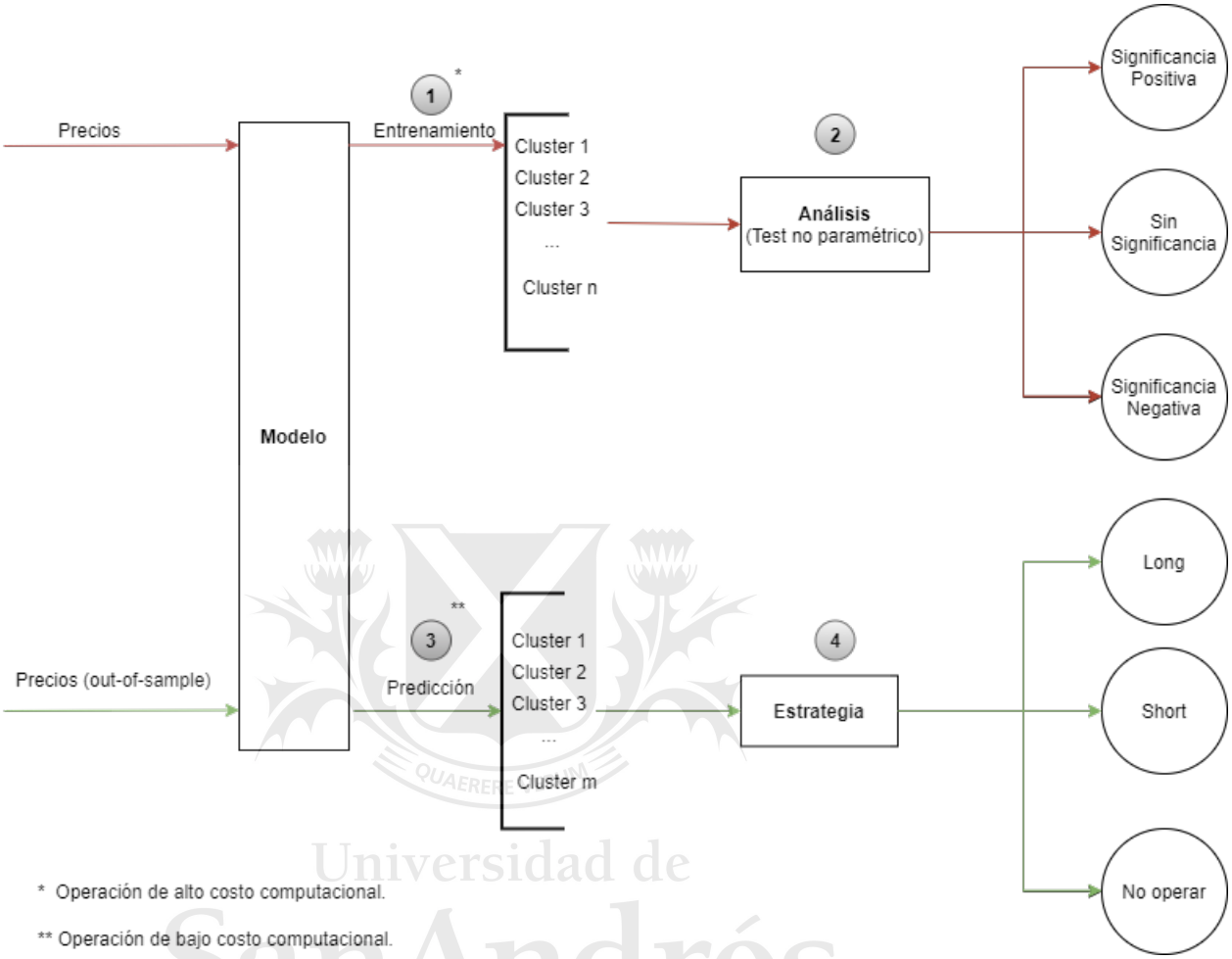


Figura 21: Pasos del análisis

6. Datos

El reconocimiento de patrones se realizó sobre los precios de los activos financieros que componen al índice *Standard & Poor's 500 o S&P500*, uno de los índices bursátiles más importantes de Estados Unidos. El periodo analizado es desde comienzos del 2000 hasta Septiembre de 2020. El análisis de patrones se hizo incorporando y excluyendo la pandemia del 2020 debida al COVID-19[7].
Se aplicaron los siguientes criterios sobre los activos que componen el índice mencionado:

- Se usaron velas de frecuencia diaria debido al elevado de tiempo de cómputo de cada modelo.
- Se eliminaron secuencias de velas en donde la acción tuvo bajo volumen en al menos un día (menor a 100,000 acciones).
- Se eliminaron secuencias de velas con datos incompletos.

En el anexo B se incluyen los *tickers* de todas las acciones candidatas luego de aplicar los filtros mencionados.
En la tabla 6 se especifican los periodos contemplados.

	Periodo de entrenamiento	**Periodo de predicción**
Excluyendo COVID-19	2000/01/03 - 2018/12/31	2019/01/02 - 2019/07/01
Incluyendo COVID-19	2000/01/03 - 2020/02/28	2020/03/02 - 2020/09/11

Tabla 6: Periodos

[7]Enfermedad infecciosa causada por el virus SARS-CoV-2. Debido a la elevada volatilidad que desencadenó la pandemia (Índice VIX en 85.5 % el 16 de Marzo de 2020) se aplicaron los modelos a periodos anteriores a la pandemia para establecer también si existen patrones que tienen impacto en los retornos de los activos financieros en condiciones normales de mercado.

7. Resultados

7.1. Entrenamiento de los modelos

Mediante precios correspondientes a los periodos de entrenamiento de la tabla 6, se computaron dos tipos de entradas con cinco y diez velas consecutivas con los siguientes parámetros definidos en la ecuación (12):

s	M	K	B	L	P
5	10	100	100	14	14
10	10	100	100	14	14

Tabla 7: Parámetros utilizados

Luego, se entrenaron los modelos *K-MEANS* y *HDBSCAN* con los parámetros que se detallan en las tablas 8 y 9.

Cantidad de clusters	**Cantidad de velas**	**Incluyendo periodo COVID-19?**
30	5	S
30	10	S
30	5	N
30	10	N
40	5	S
40	10	S
40	5	N
40	10	N
50	5	S
50	10	S
50	5	N
50	10	N
60	5	S
60	10	S
60	5	N
60	10	N
70	5	S
70	10	S
70	5	N
70	10	N
80	5	S
80	10	S
80	5	N
80	10	N
90	5	S
90	10	S
90	5	N
90	10	N
100	5	S
100	10	S
100	5	N
100	10	N
200	5	S
200	10	S
200	5	N
200	10	N

Tabla 8: Entrada para K-MEANS

Tamaño de cluster	Cantidad mínima de observaciones	Cantidad de velas	Incluyendo periodo COVID-19
10	10	5	S
10	10	10	S
10	10	5	N
10	10	10	N
20	20	5	S
20	20	10	S
20	20	5	N
20	20	10	N
30	30	5	S
30	30	10	S
30	30	5	N
30	30	10	N
40	40	5	S
40	40	10	S
40	40	5	N
40	40	10	N
50	50	5	S
50	50	10	S
50	50	5	N
50	50	10	N
60	60	5	S
60	60	10	S
60	60	5	N
60	60	10	N
70	70	5	S
70	70	10	S
70	70	5	N
70	70	10	N
80	80	5	S
80	80	10	S
80	80	5	N
80	80	10	N
90	90	5	S
90	90	10	S
90	90	5	N
90	90	10	N
100	100	5	S
100	100	10	S
100	100	5	N
100	100	10	N
200	200	5	S
200	200	10	S
200	200	5	N
200	200	10	N

Tabla 9: Entrada para HDBSCAN

Como se observa en la tabla 8, el modelo *K-MEANS* fue entrenado con distintas cantidades de *clusters*, patrones de cinco o diez velas e incluyendo o

excluyendo el periodo de la pandemia del 2020. En el caso de *HDBSCAN* (tabla 9) , como se describió en la sección 3.4.6, se necesitan de dos parámetros: Tamaño de *cluster* o *min cluster size* y cantidad mínima de observaciones o *min samples* Además para ambos algoritmos, se analizaron los retornos *a posteriori* tomando tres y siete días.

7.2. Pasos

Para cada modelo y para cada una de las entradas especificadas en las tablas 8 y 9 se ejecutaron los siguientes pasos:

1. Se entrenó el modelo encontrando así los *clusters*.
2. Para cada *cluster* se analizó el impacto en los retornos de las acciones en los precios posteriores al patrón utilizando la metodología descripta en la sección 5.
3. Se aplicó la operación de predicción sobre patrones *out-of-sample* encontrando así al *cluster* que mejor agrupa a cada patrón.
4. Se calcularon los retornos sobre cada patrón *out-of-sample* perteneciente a *clusters* significativos, descartando a los que no tienen la suficiente significancia. Los retornos se calcularon para tres y siete días posteriores a los patrones.

La obtención de los *tickers* que componen al índice S&P500, los precios de los activos, el cómputo de las entradas de los modelos (ecuación 12), la obtención de los *clusters* y el análisis de los datos, se automatizaron mediante módulos desarrollados en *Python*. El enlace al código completo se encuentra en el anexo A. En la imagen 22 se muestra la arquitectura de la solución. Cabe destacar que los datos requeridos en cada etapa del proceso se almacenaron en disco para evitar recalcularlos constantemente. De esta manera, por ejemplo, los precios se obtienen automáticamente de *Yahoo* mediante el módulo *Online Prices Provider* o se obtienen del disco de la computadora mediante el módulo *Offline Prices Provider*.

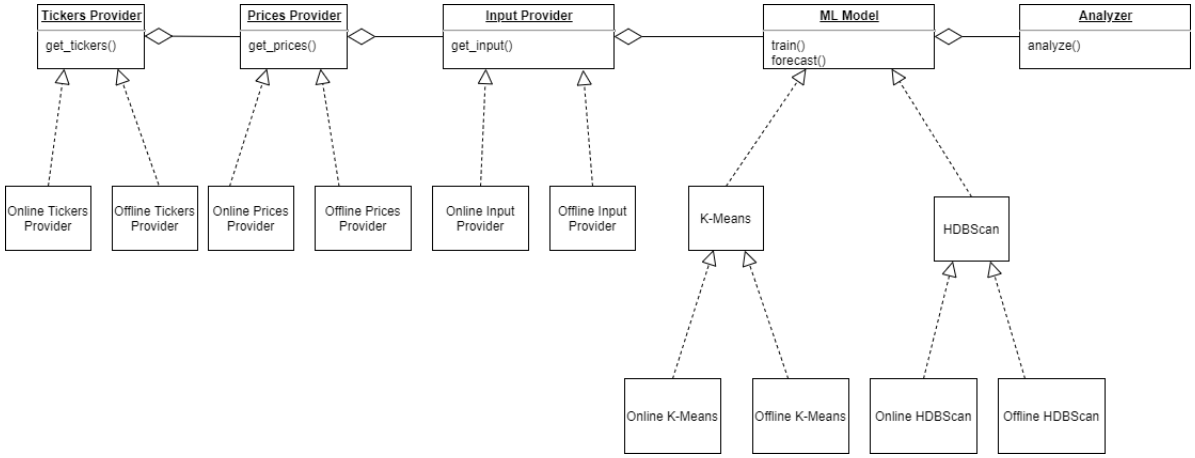


Figura 22: Diagrama de arquitectura de la solución

7.3. Resultados obtenidos

Periodo para entrenamiento: 2000/01/03 - 2018/12/31
Periodo para predicción: 2019/01/02 - 2019/07/01
M=10; K=100; B=100; L=14; P=14 (parámetros de la ecuación 12)

Total clusters	Cant velas (patrón)	Cant velas (retornos)	Tasa de éxito	Mediana	Min	Max	Media ret negativos	Media ret positivos
200	50	7	56.82 %	0.58 %	-32.86 %	55.13 %	-3.19 %	3.09 %
90	10	7	56.81 %	0.58 %	-32.86 %	55.13 %	-3.18 %	3.06 %
70	10	7	56.78 %	0.57 %	-32.86 %	55.13 %	-3.17 %	3.06 %
90	50	7	56.76 %	0.57 %	-32.40 %	55.13 %	-3.17 %	3.07 %
50	10	7	56.75 %	0.57 %	-32.86 %	55.13 %	-3.17 %	3.07 %
100	50	7	56.75 %	0.57 %	-32.86 %	55.13 %	-3.18 %	3.07 %
50	50	7	56.73 %	0.56 %	-32.86 %	55.13 %	-3.15 %	3.04 %
80	10	7	56.72 %	0.56 %	-32.86 %	55.13 %	-3.16 %	3.04 %
100	10	7	56.72 %	0.56 %	-32.86 %	55.13 %	-3.15 %	3.04 %
70	50	7	56.72 %	0.57 %	-32.86 %	55.13 %	-3.17 %	3.06 %
60	50	7	56.72 %	0.57 %	-32.86 %	55.13 %	-3.16 %	3.05 %
200	10	7	56.72 %	0.58 %	-32.40 %	55.13 %	-3.21 %	3.10 %
40	10	7	56.71 %	0.56 %	-32.86 %	55.13 %	-3.16 %	3.06 %
60	10	7	56.70 %	0.56 %	-32.86 %	55.13 %	-3.14 %	3.03 %
40	50	7	56.66 %	0.56 %	-32.86 %	55.13 %	-3.17 %	3.07 %
10	10	7	56.64 %	0.54 %	-32.86 %	55.13 %	-3.11 %	3.00 %
20	10	7	56.64 %	0.54 %	-32.86 %	55.13 %	-3.11 %	3.00 %
30	10	7	56.64 %	0.54 %	-32.86 %	55.13 %	-3.11 %	3.00 %
10	50	7	56.62 %	0.54 %	-32.86 %	55.13 %	-3.11 %	3.00 %
20	50	7	56.62 %	0.54 %	-32.86 %	55.13 %	-3.11 %	3.00 %
30	50	7	56.62 %	0.54 %	-32.86 %	55.13 %	-3.11 %	3.00 %
80	50	7	56.57 %	0.55 %	-32.86 %	55.13 %	-3.14 %	3.05 %
200	10	3	55.22 %	0.28 %	-33.06 %	39.72 %	-2.03 %	2.01 %
80	50	3	55.20 %	0.28 %	-33.06 %	39.72 %	-2.05 %	2.02 %
90	10	3	55.12 %	0.27 %	-33.06 %	39.72 %	-2.04 %	2.02 %
80	10	3	55.11 %	0.27 %	-33.06 %	39.72 %	-2.04 %	2.01 %
100	10	3	55.03 %	0.27 %	-33.06 %	39.72 %	-2.02 %	2.00 %
50	50	3	55.03 %	0.26 %	-33.06 %	39.72 %	-2.01 %	1.98 %
200	50	3	55.01 %	0.26 %	-33.06 %	39.72 %	-2.03 %	2.01 %
60	50	3	55.00 %	0.26 %	-33.06 %	39.72 %	-2.01 %	1.99 %
90	50	3	54.98 %	0.26 %	-33.06 %	39.72 %	-2.01 %	1.99 %
70	50	3	54.98 %	0.26 %	-33.06 %	39.72 %	-2.03 %	2.00 %
10	10	3	54.92 %	0.25 %	-33.06 %	39.72 %	-2.01 %	1.97 %
10	50	3	54.89 %	0.25 %	-33.06 %	39.72 %	-2.01 %	1.97 %
70	10	3	54.86 %	0.25 %	-33.06 %	39.72 %	-2.01 %	2.00 %
100	50	3	54.85 %	0.25 %	-33.06 %	39.72 %	-2.01 %	1.99 %
60	10	3	54.83 %	0.25 %	-33.06 %	39.72 %	-1.99 %	1.97 %
50	10	3	54.82 %	0.25 %	-33.06 %	39.72 %	-1.98 %	1.97 %
40	10	3	54.78 %	0.24 %	-33.06 %	39.72 %	-1.98 %	1.96 %
30	50	3	54.73 %	0.25 %	-33.06 %	39.72 %	-1.98 %	1.98 %
40	50	3	54.66 %	0.24 %	-33.06 %	39.72 %	-1.98 %	1.97 %
20	10	3	54.65 %	0.25 %	-33.06 %	39.72 %	-2.01 %	2.00 %
30	10	3	54.65 %	0.24 %	-33.06 %	39.72 %	-1.98 %	1.98 %
20	50	3	54.63 %	0.24 %	-33.06 %	39.72 %	-2.01 %	2.00 %

Tabla 10: Resultados K-MEANS excluyendo periodo COVID-19

Periodo para entrenamiento: 2000/01/03 - 2018/12/31
Periodo para predicción: 2019/01/02 - 2019/07/01
M=10; K=100; B=100; L=14; P=14 (parámetros de la ecuación 12)

Total clusters	Tamaño patrón	Cant velas ret.	Cant. *trades*	Cant. *trades* exitosos	Cant. *clusters* sig	% *clusters* sig.	Cant. estrategias long	Cant. estrategias short
200	50	7	31208	17731	157	78.50 %	156	1
90	10	7	31964	18159	74	82.22 %	74	0
70	10	7	32311	18345	60	85.71 %	60	0
90	50	7	31890	18100	76	84.44 %	75	1
50	10	7	32158	18251	43	86.00 %	43	0
100	50	7	32217	18282	86	86.00 %	84	2
50	50	7	33134	18796	45	90.00 %	45	0
80	10	7	32804	18607	68	85.00 %	68	0
100	10	7	32535	18454	84	84.00 %	83	1
70	50	7	32506	18437	61	87.14 %	61	0
60	50	7	32690	18541	52	86.67 %	52	0
200	10	7	30471	17282	154	77.00 %	153	1
40	10	7	32268	18299	35	87.50 %	35	0
60	10	7	33340	18904	53	88.33 %	53	0
40	50	7	32409	18364	35	87.50 %	35	0
10	10	7	34766	19691	10	100.00 %	10	0
20	10	7	34766	19691	19	95.00 %	19	0
30	10	7	34766	19691	29	96.67 %	29	0
10	50	7	34791	19698	10	100.00 %	10	0
20	50	7	34791	19698	19	95.00 %	19	0
30	50	7	34791	19698	29	96.67 %	29	0
80	50	7	32739	18521	69	86.25 %	67	2
200	10	3	30187	16669	143	71.50 %	134	9
80	50	3	30038	16580	62	77.50 %	59	3
90	10	3	30306	16706	70	77.78 %	66	4
80	10	3	30341	16721	62	77.50 %	58	4
100	10	3	30801	16950	76	76.00 %	70	6
50	50	3	33157	18245	43	86.00 %	41	2
200	50	3	30434	16741	147	73.50 %	134	13
60	50	3	32089	17648	49	81.67 %	46	3
90	50	3	31891	17535	73	81.11 %	67	6
70	50	3	31888	17531	57	81.43 %	54	3
10	10	3	33520	18408	9	90.00 %	9	0
10	50	3	33556	18420	9	90.00 %	9	0
70	10	3	32010	17561	57	81.43 %	52	5
100	50	3	32195	17658	82	82.00 %	74	8
60	10	3	33261	18237	51	85.00 %	47	4
50	10	3	33815	18538	43	86.00 %	40	3
40	10	3	34340	18813	35	87.50 %	32	3
30	50	3	34451	18856	27	90.00 %	25	2
40	50	3	34432	18821	36	90.00 %	33	3
20	10	3	32286	17644	17	85.00 %	16	1
30	10	3	33948	18552	26	86.67 %	24	2
20	50	3	32323	17658	17	85.00 %	16	1

Tabla 11: Resultados K-MEANS excluyendo periodo COVID-19 (cont.)

Periodo para entrenamiento: 2000/01/03 - 2020/02/28
Periodo para predicción: 2020/03/02 - 2020/09/11
M=10; K=100; B=100; L=14; P=14 (parámetros de la ecuación 12)

Total clusters	Cant velas (patrón)	Cant velas (retornos)	Tasa de éxito	Mediana	Min	Max	Media ret negativos	Media ret positivos
200	10	7	58.72 %	1.11 %	-37.50 %	84.97 %	-3.97 %	5.58 %
200	5	7	58.70 %	1.11 %	-37.50 %	84.97 %	-3.96 %	5.54 %
40	10	7	58.63 %	1.11 %	-37.50 %	84.97 %	-3.99 %	5.58 %
40	5	7	58.63 %	1.11 %	-37.50 %	84.97 %	-4.00 %	5.58 %
50	5	7	58.61 %	1.11 %	-37.50 %	84.97 %	-4.01 %	5.59 %
100	10	7	58.58 %	1.10 %	-37.50 %	84.97 %	-3.99 %	5.57 %
70	5	7	58.56 %	1.10 %	-37.50 %	84.97 %	-4.01 %	5.58 %
50	10	7	58.55 %	1.09 %	-37.50 %	84.97 %	-4.02 %	5.57 %
100	5	7	58.50 %	1.09 %	-37.50 %	84.97 %	-4.01 %	5.56 %
60	5	7	58.50 %	1.09 %	-37.50 %	84.97 %	-4.02 %	5.57 %
60	10	7	58.48 %	1.09 %	-37.50 %	84.97 %	-4.02 %	5.57 %
80	5	7	58.47 %	1.08 %	-37.50 %	84.97 %	-4.01 %	5.56 %
80	10	7	58.46 %	1.07 %	-37.50 %	84.97 %	-4.01 %	5.56 %
70	10	7	58.45 %	1.09 %	-37.50 %	84.97 %	-4.02 %	5.58 %
30	5	7	58.45 %	1.09 %	-37.50 %	84.97 %	-4.04 %	5.58 %
90	5	7	58.43 %	1.08 %	-37.50 %	84.97 %	-4.02 %	5.57 %
90	10	7	58.41 %	1.08 %	-37.50 %	84.97 %	-4.02 %	5.57 %
30	10	7	58.22 %	1.05 %	-37.50 %	84.97 %	-4.05 %	5.56 %
10	10	7	58.21 %	1.05 %	-37.50 %	84.97 %	-4.04 %	5.54 %
20	10	7	58.21 %	1.05 %	-37.50 %	84.97 %	-4.04 %	5.54 %
10	5	7	58.21 %	1.05 %	-37.50 %	84.97 %	-4.04 %	5.54 %
20	5	7	58.21 %	1.05 %	-37.50 %	84.97 %	-4.04 %	5.54 %
200	5	3	56.25 %	0.54 %	-31.83 %	71.39 %	-3.02 %	3.73 %
90	10	3	56.23 %	0.54 %	-31.83 %	71.39 %	-3.00 %	3.71 %
80	5	3	56.22 %	0.54 %	-31.83 %	71.39 %	-3.02 %	3.74 %
90	5	3	56.20 %	0.54 %	-31.83 %	71.39 %	-3.03 %	3.75 %
200	10	3	56.17 %	0.53 %	-31.83 %	65.69 %	-2.99 %	3.74 %
100	10	3	56.14 %	0.53 %	-31.83 %	71.39 %	-3.02 %	3.73 %
100	5	3	56.12 %	0.53 %	-31.83 %	71.39 %	-3.03 %	3.72 %
80	10	3	56.06 %	0.53 %	-31.83 %	71.39 %	-3.03 %	3.75 %
70	10	3	56.03 %	0.53 %	-31.83 %	71.39 %	-3.04 %	3.75 %
60	5	3	55.99 %	0.52 %	-33.11 %	71.39 %	-3.03 %	3.75 %
40	5	3	55.99 %	0.52 %	-31.83 %	71.39 %	-3.06 %	3.74 %
30	10	3	55.95 %	0.51 %	-34.86 %	71.39 %	-3.02 %	3.75 %
50	5	3	55.93 %	0.51 %	-31.83 %	71.39 %	-3.06 %	3.74 %
30	5	3	55.91 %	0.51 %	-34.86 %	71.39 %	-3.07 %	3.77 %
20	5	3	55.90 %	0.52 %	-33.11 %	71.39 %	-3.05 %	3.77 %
20	10	3	55.89 %	0.52 %	-33.11 %	71.39 %	-3.05 %	3.77 %
70	5	3	55.87 %	0.51 %	-31.83 %	71.39 %	-3.06 %	3.74 %
50	10	3	55.87 %	0.51 %	-31.83 %	71.39 %	-3.06 %	3.73 %
60	10	3	55.86 %	0.50 %	-33.11 %	71.39 %	-3.05 %	3.74 %
40	10	3	55.86 %	0.51 %	-31.83 %	71.39 %	-3.07 %	3.73 %
10	5	3	55.74 %	0.50 %	-31.83 %	71.39 %	-3.16 %	3.70 %
10	10	3	55.55 %	0.47 %	-35.78 %	71.39 %	-3.16 %	3.65 %

Tabla 12: Resultados K-MEANS incluyendo periodo COVID-19

Periodo para entrenamiento: 2000/01/03 - 2020/02/28
Periodo para predicción: 2020/03/02 - 2020/09/11
M=10; K=100; B=100; L=14; P=14 (parámetros de la ecuación 12)

Total clusters	Tamaño patrón	Cant velas ret.	Cant. *trades*	Cant. *trades* exitosos	Cant. *clusters* sig	% *clusters* sig.	Cant. estrategias long	Cant. estrategias short
200	10	7	37543	22045	157	78.50 %	156	1
200	5	7	37814	22196	158	79.00 %	157	1
40	10	7	38265	22436	35	87.50 %	35	0
40	5	7	38353	22485	37	92.50 %	37	0
50	5	7	38312	22456	45	90.00 %	45	0
100	10	7	38410	22499	87	87.00 %	86	1
70	5	7	38444	22513	62	88.57 %	62	0
50	10	7	38735	22678	46	92.00 %	46	0
100	5	7	38570	22563	85	85.00 %	85	0
60	5	7	38675	22623	54	90.00 %	54	0
60	10	7	38694	22630	55	91.67 %	55	0
80	5	7	38973	22788	72	90.00 %	71	1
80	10	7	39038	22821	72	90.00 %	71	1
70	10	7	38697	22620	62	88.57 %	62	0
30	5	7	38708	22623	28	93.33 %	28	0
90	5	7	38979	22775	79	87.78 %	78	1
90	10	7	38909	22725	79	87.78 %	78	1
30	10	7	39299	22878	29	96.67 %	29	0
10	10	7	39654	23082	10	100.00 %	10	0
20	10	7	39654	23082	20	100.00 %	20	0
10	5	7	39655	23082	10	100.00 %	10	0
20	5	7	39655	23082	20	100.00 %	20	0
200	5	3	37372	21022	149	74.50 %	139	10
90	10	3	38272	21519	74	82.22 %	68	6
80	5	3	38115	21427	67	83.75 %	64	3
90	5	3	38218	21480	74	82.22 %	69	5
200	10	3	37468	21045	154	77.00 %	140	14
100	10	3	38066	21369	81	81.00 %	75	6
100	5	3	38957	21862	80	80.00 %	74	6
80	10	3	38396	21525	67	83.75 %	63	4
70	10	3	39517	22142	62	88.57 %	58	4
60	5	3	39353	22034	54	90.00 %	50	4
40	5	3	39265	21983	35	87.50 %	33	2
30	10	3	40179	22481	29	96.67 %	26	3
50	5	3	40035	22391	45	90.00 %	42	3
30	5	3	40254	22506	28	93.33 %	26	2
20	5	3	39335	21990	18	90.00 %	17	1
20	10	3	39313	21974	18	90.00 %	17	1
70	5	3	39358	21991	61	87.14 %	58	3
50	10	3	40307	22520	46	92.00 %	43	3
60	10	3	40371	22553	56	93.33 %	51	5
40	10	3	39858	22263	37	92.50 %	35	2
10	5	3	39706	22134	9	90.00 %	9	0
10	10	3	41205	22889	10	100.00 %	10	0

Tabla 13: Resultados K-MEANS incluyendo periodo COVID-19 (cont.)

Periodo para entrenamiento: 2000/01/03 - 2018/12/31
Periodo para predicción: 2019/01/02 - 2019/07/01
M=10; K=100; B=100; L=14; P=14 (parámetros de la ecuación 12)

Min tamaño	Min muestras	Tamaño patrón	Cant velas ret.	Tasa de éxito	Retorno efectivo	Mediana	Min	Max	Media ret. negativos	Media ret positivos
200	200	5	7	70.00 %	2087.32 %	1.65 %	-14.44 %	9.99 %	-3.03 %	3.42 %
30	30	10	7	59.77 %	594.27 %	0.83 %	-18.39 %	11.30 %	-2.75 %	2.92 %
200	200	5	3	62.50 %	496.09 %	0.63 %	-6.66 %	7.93 %	-1.85 %	2.36 %
60	60	5	7	64.42 %	294.58 %	1.25 %	-26.27 %	10.04 %	-3.18 %	3.22 %
200	200	10	7	64.10 %	223.60 %	1.12 %	-18.39 %	11.38 %	-3.12 %	3.05 %
20	20	10	3	60.00 %	201.57 %	0.61 %	-5.90 %	7.20 %	-1.53 %	1.94 %
80	80	10	3	61.95 %	133.04 %	0.59 %	-4.34 %	5.87 %	-1.20 %	1.98 %
80	80	10	7	59.20 %	117.07 %	0.81 %	-26.27 %	8.54 %	-2.53 %	2.95 %
70	70	10	3	58.57 %	105.33 %	0.28 %	-5.90 %	10.36 %	-1.40 %	1.91 %
40	40	10	7	59.01 %	104.08 %	0.83 %	-26.27 %	14.27 %	-3.16 %	2.89 %
60	60	10	7	59.04 %	103.85 %	0.62 %	-26.27 %	11.40 %	-2.88 %	2.79 %
60	60	5	3	61.21 %	100.27 %	0.62 %	-7.27 %	6.67 %	-1.81 %	2.18 %
70	70	10	7	55.37 %	90.20 %	0.47 %	-26.27 %	13.81 %	-2.87 %	3.15 %
70	70	5	7	61.45 %	87.66 %	0.68 %	-26.27 %	8.51 %	-2.94 %	2.59 %
200	200	10	3	63.78 %	87.33 %	0.58 %	-6.03 %	7.19 %	-1.91 %	1.90 %
60	60	10	3	58.87 %	80.58 %	0.64 %	-13.37 %	6.67 %	-1.80 %	2.03 %
90	90	10	7	59.26 %	75.59 %	0.68 %	-26.27 %	8.51 %	-2.81 %	2.78 %
70	70	5	3	64.71 %	74.02 %	0.59 %	-6.03 %	6.67 %	-1.75 %	1.71 %
80	80	5	3	63.29 %	73.12 %	0.52 %	-3.22 %	5.83 %	-1.16 %	1.80 %
40	40	5	7	58.67 %	55.48 %	0.75 %	-26.27 %	14.27 %	-3.39 %	2.96 %
20	20	5	7	56.93 %	55.35 %	0.39 %	-12.36 %	15.03 %	-3.13 %	2.79 %
30	30	5	7	58.49 %	52.37 %	0.71 %	-26.27 %	11.30 %	-3.40 %	2.84 %
10	10	5	3	52.73 %	48.82 %	0.16 %	-7.86 %	10.36 %	-1.52 %	1.87 %
90	90	10	3	61.22 %	46.48 %	0.46 %	-6.57 %	6.01 %	-1.77 %	1.80 %
100	100	10	7	52.60 %	32.51 %	0.46 %	-18.39 %	9.10 %	-2.74 %	2.96 %
20	20	10	7	57.35 %	25.09 %	0.42 %	-12.36 %	11.30 %	-3.08 %	2.55 %
30	30	5	3	51.91 %	20.64 %	0.16 %	-31.30 %	8.47 %	-2.05 %	2.17 %
80	80	5	7	56.03 %	19.28 %	0.48 %	-26.27 %	8.88 %	-3.45 %	3.19 %
100	100	5	7	56.67 %	14.83 %	0.85 %	-26.27 %	9.42 %	-4.00 %	3.60 %
100	100	10	3	56.36 %	13.74 %	0.28 %	-5.90 %	6.01 %	-1.79 %	1.85 %
50	50	10	3	58.19 %	13.54 %	0.44 %	-31.30 %	6.28 %	-2.15 %	1.77 %
20	20	5	3	55.45 %	11.93 %	0.31 %	-13.64 %	7.55 %	-1.99 %	1.76 %
90	90	5	7	55.75 %	10.79 %	0.53 %	-26.27 %	11.38 %	-3.51 %	3.19 %
50	50	5	3	55.69 %	10.02 %	0.44 %	-31.30 %	8.47 %	-2.25 %	2.01 %
10	10	5	7	60.83 %	9.10 %	0.60 %	-15.71 %	9.33 %	-3.35 %	2.33 %
90	90	5	3	57.14 %	8.25 %	0.40 %	-6.57 %	6.01 %	-2.22 %	1.84 %
10	10	10	7	53.88 %	6.88 %	0.41 %	-16.34 %	11.30 %	-3.14 %	2.89 %
50	50	5	7	54.35 %	3.63 %	0.64 %	-26.27 %	9.42 %	-3.22 %	2.92 %
50	50	10	7	53.82 %	1.35 %	0.26 %	-26.27 %	10.59 %	-3.01 %	2.76 %
40	40	5	3	56.97 %	-0.44 %	0.44 %	-31.30 %	7.12 %	-2.33 %	1.88 %
40	40	10	3	55.97 %	-2.53 %	0.34 %	-31.30 %	7.12 %	-2.35 %	1.95 %
100	100	5	3	57.58 %	-9.03 %	0.30 %	-31.30 %	6.01 %	-2.56 %	1.88 %
30	30	10	3	54.58 %	-22.65 %	0.25 %	-12.84 %	7.12 %	-2.22 %	1.73 %
10	10	10	3	50.82 %	-27.45 %	0.07 %	-10.53 %	8.67 %	-1.99 %	1.65 %

Tabla 14: Resultados HDBSCAN excluyendo periodo COVID-19

Periodo para entrenamiento: 2000/01/03 - 2018/12/31
Periodo para predicción: 2019/01/02 - 2019/07/01
M=10; K=100; B=100; L=14; P=14 (parámetros de la ecuación 12)

Min. tamaño	Min. muestras	Tamaño patrón	Cant. velas ret.	Cant. *trades*	Cant. *trades* exitosos	Cant. *clusters* sig	Total *clusters*	% *clusters* sig	Cant. estr long	Cant. estr short
200	200	5	7	220	154	9	21	42.86 %	9	0
30	30	10	7	343	205	137	1156	11.85 %	132	5
200	200	5	3	240	150	8	22	36.36 %	8	0
60	60	5	7	163	105	58	285	20.35 %	57	1
200	200	10	7	156	100	13	22	59.09 %	13	0
20	20	10	3	210	126	118	1628	7.25 %	111	7
80	80	10	3	113	70	31	146	21.23 %	27	4
80	80	10	7	125	74	34	132	25.76 %	34	0
70	70	10	3	140	82	40	196	20.41 %	37	3
40	40	10	7	222	131	86	655	13.13 %	86	0
60	60	10	7	188	111	59	275	21.45 %	57	2
60	60	5	3	116	71	33	305	10.82 %	31	2
70	70	10	7	177	98	41	188	21.81 %	40	1
70	70	5	7	166	102	44	206	21.36 %	43	1
200	200	10	3	127	81	9	21	42.86 %	9	0
60	60	10	3	141	83	44	282	15.60 %	43	1
90	90	10	7	135	80	31	105	29.52 %	31	0
70	70	5	3	119	77	29	212	13.68 %	29	0
80	80	5	3	79	50	27	154	17.53 %	27	0
40	40	5	7	196	115	91	700	13.00 %	90	1
20	20	5	7	274	156	138	1685	8.19 %	133	5
30	30	5	7	265	155	128	1284	9.97 %	124	4
10	10	5	3	165	87	98	1973	4.97 %	86	12
90	90	10	3	98	60	25	105	23.81 %	23	2
100	100	10	7	154	81	21	79	26.58 %	21	0
20	20	10	7	279	160	139	1568	8.86 %	133	6
30	30	5	3	235	122	112	1313	8.53 %	103	9
80	80	5	7	116	65	35	143	24.48 %	33	2
100	100	5	7	90	51	20	81	24.69 %	20	0
100	100	10	3	55	31	12	82	14.63 %	12	0
50	50	10	3	177	103	57	412	13.83 %	51	6
20	20	5	3	202	112	98	1768	5.54 %	89	9
90	90	5	7	113	63	26	113	23.01 %	26	0
50	50	5	3	167	93	52	473	10.99 %	47	5
10	10	5	7	217	132	126	1843	6.84 %	113	13
90	90	5	3	119	68	21	110	19.09 %	19	2
10	10	10	7	219	118	115	1662	6.92 %	109	6
50	50	5	7	184	100	64	453	14.13 %	61	3
50	50	10	7	249	134	76	410	18.54 %	72	4
40	40	5	3	165	94	56	729	7.68 %	53	3
40	40	10	3	159	89	60	672	8.93 %	58	2
100	100	5	3	99	57	22	91	24.18 %	21	1
30	30	10	3	251	137	108	1203	8.98 %	101	7
10	10	10	3	183	93	97	1764	5.50 %	78	19

Tabla 15: Resultados HDBSCAN excluyendo periodo COVID-19 (cont.)

Periodo para entrenamiento: 2000/01/03 - 2020/02/28
Periodo para predicción: 2020/03/02 - 2020/09/11
M=10; K=100; B=100; L=14; P=14 (parámetros de la ecuación 12)

Min tamaño	Min muestras	Tamaño patrón	Cant velas ret.	Tasa de éxito	Retorno efectivo	Mediana	Min	Max	Media ret. negativos	Media ret positivos
30	30	5	7	4.36 %	63.61 %	1.44 %	-13.49 %	17.37 %	-3.02 %	4.36 %
20	20	5	7	4.46 %	58.77 %	1.17 %	-18.06 %	28.97 %	-3.13 %	4.46 %
40	40	5	7	4.72 %	61.34 %	1.38 %	-14.29 %	20.64 %	-3.07 %	4.72 %
100	100	5	7	4.55 %	64.91 %	1.52 %	-10.94 %	20.64 %	-2.40 %	4.55 %
60	60	5	7	4.16 %	59.53 %	0.89 %	-16.37 %	20.64 %	-2.67 %	4.16 %
70	70	5	7	4.03 %	59.33 %	0.99 %	-11.71 %	20.64 %	-2.40 %	4.03 %
30	30	10	7	3.92 %	61.86 %	1.44 %	-10.59 %	15.33 %	-2.90 %	3.92 %
80	80	10	7	4.08 %	62.63 %	0.94 %	-13.21 %	15.24 %	-2.87 %	4.08 %
30	30	5	3	3.00 %	62.59 %	0.97 %	-11.83 %	11.51 %	-2.32 %	3.00 %
10	10	5	7	4.65 %	60.87 %	1.18 %	-8.24 %	23.00 %	-2.93 %	4.65 %
20	20	10	7	4.08 %	58.75 %	0.94 %	-12.33 %	28.97 %	-3.07 %	4.08 %
100	100	10	7	4.04 %	61.63 %	1.44 %	-16.85 %	16.69 %	-2.68 %	4.04 %
80	80	5	7	4.40 %	55.30 %	0.88 %	-10.87 %	14.58 %	-2.91 %	4.40 %
200	200	5	7	4.54 %	59.46 %	1.43 %	-10.87 %	16.99 %	-2.75 %	4.54 %
50	50	5	7	4.16 %	58.33 %	0.95 %	-13.12 %	13.55 %	-3.09 %	4.16 %
50	50	10	7	3.79 %	59.65 %	1.08 %	-8.40 %	15.35 %	-2.60 %	3.79 %
70	70	10	7	4.15 %	64.06 %	1.10 %	-16.85 %	15.69 %	-3.16 %	4.15 %
60	60	10	7	3.61 %	61.76 %	1.04 %	-8.84 %	13.64 %	-2.90 %	3.61 %
40	40	10	7	4.07 %	57.30 %	0.59 %	-10.59 %	13.71 %	-3.01 %	4.07 %
200	200	10	7	4.16 %	60.32 %	1.05 %	-10.92 %	16.69 %	-3.10 %	4.16 %
90	90	5	7	4.74 %	51.61 %	0.20 %	-10.87 %	20.64 %	-2.92 %	4.74 %
30	30	10	3	2.50 %	67.65 %	1.26 %	-7.28 %	9.95 %	-2.27 %	2.50 %
90	90	10	7	3.99 %	57.48 %	0.53 %	-19.56 %	16.69 %	-3.18 %	3.99 %
10	10	10	7	4.43 %	53.91 %	0.60 %	-8.89 %	28.97 %	-3.14 %	4.43 %
80	80	5	3	2.91 %	60.63 %	0.79 %	-7.69 %	13.99 %	-2.30 %	2.91 %
20	20	5	3	2.74 %	58.92 %	0.64 %	-11.83 %	9.36 %	-2.80 %	2.74 %
50	50	10	3	3.11 %	59.41 %	1.00 %	-6.19 %	10.70 %	-2.18 %	3.11 %
10	10	5	3	3.03 %	60.34 %	0.87 %	-9.25 %	11.44 %	-2.50 %	3.03 %
40	40	5	3	2.75 %	54.87 %	0.27 %	-8.46 %	17.14 %	-2.27 %	2.75 %
90	90	10	3	2.53 %	63.74 %	0.98 %	-6.24 %	7.27 %	-1.98 %	2.53 %
90	90	5	3	2.66 %	59.46 %	0.61 %	-8.58 %	13.99 %	-2.07 %	2.66 %
50	50	5	3	2.77 %	54.55 %	0.29 %	-6.19 %	11.51 %	-2.02 %	2.77 %
70	70	5	3	2.76 %	55.06 %	0.36 %	-6.19 %	13.99 %	-2.19 %	2.76 %
60	60	10	3	2.59 %	60.71 %	0.99 %	-6.24 %	9.13 %	-2.35 %	2.59 %
100	100	5	3	2.71 %	54.82 %	0.46 %	-11.46 %	13.99 %	-2.51 %	2.71 %
100	100	10	3	2.39 %	60.32 %	0.60 %	-8.83 %	13.99 %	-2.36 %	2.39 %
20	20	10	3	2.46 %	59.32 %	0.64 %	-11.83 %	9.12 %	-2.68 %	2.46 %
70	70	10	3	2.87 %	54.64 %	0.33 %	-7.28 %	9.27 %	-2.11 %	2.87 %
40	40	10	3	2.71 %	55.96 %	0.70 %	-8.64 %	9.13 %	-2.49 %	2.71 %
60	60	5	3	2.54 %	55.09 %	0.29 %	-11.83 %	12.88 %	-2.56 %	2.54 %
200	200	5	3	2.31 %	53.79 %	0.48 %	-8.06 %	6.45 %	-2.44 %	2.31 %
10	10	10	3	2.46 %	52.43 %	0.25 %	-10.25 %	11.44 %	-2.54 %	2.46 %
200	200	10	3	1.95 %	52.63 %	0.46 %	-11.46 %	5.23 %	-2.15 %	1.95 %
80	80	10	3	1.93 %	45.65 %	-0.37 %	-6.19 %	7.71 %	-1.95 %	1.93 %

Tabla 16: Resultados HDBSCAN incluyendo periodo COVID-19

Periodo para entrenamiento: 2000/01/03 - 2020/02/28
Periodo para predicción: 2020/03/02 - 2020/09/11
M=10; K=100; B=100; L=14; P=14 (parámetros de la ecuación 12)

Min. tamaño	Min. muestras	Tamaño patrón	Cant. velas ret.	Cant. *trades*	Cant. *trades* exitosos	Cant. *clusters* sig	Total *clusters*	% *clusters* sig	Cant. estr long	Cant. estr short
30	30	5	7	371	236	155	1871	8.28 %	64	5
20	20	5	7	325	191	159	1749	9.09 %	83	3
40	40	5	7	238	146	98	2030	4.83 %	70	10
100	100	5	7	171	111	32	1936	1.65 %	104	3
60	60	5	7	215	128	68	95	71.58 %	24	0
70	70	5	7	209	124	53	89	59.55 %	32	1
30	30	10	7	215	133	112	105	106.67 %	30	0
80	80	10	7	190	119	39	88	44.32 %	30	2
30	30	5	3	270	169	112	1677	6.68 %	93	7
10	10	5	7	161	98	107	1550	6.90 %	123	1
20	20	10	7	240	141	124	1805	6.87 %	111	10
100	100	10	7	172	106	33	1720	1.92 %	156	3
80	80	5	7	217	120	44	18	244.44 %	8	0
200	200	5	7	148	88	13	21	61.90 %	12	0
50	50	5	7	192	112	79	23	343.48 %	9	0
50	50	10	7	171	102	78	23	339.13 %	13	0
70	70	10	7	128	82	51	1277	3.99 %	74	1
60	60	10	7	170	105	61	1216	5.02 %	110	2
40	40	10	7	178	102	87	1379	6.31 %	107	5
200	200	10	7	126	76	12	1329	0.90 %	154	1
90	90	5	7	155	80	39	708	5.51 %	61	1
30	30	10	3	136	92	75	671	11.18 %	86	1
90	90	10	7	127	73	36	787	4.57 %	74	4
10	10	10	7	128	69	86	748	11.50 %	98	0
80	80	5	3	127	77	31	456	6.80 %	43	2
20	20	5	3	241	142	121	396	30.56 %	76	2
50	50	10	3	101	60	45	502	8.96 %	58	2
10	10	5	3	116	70	80	473	16.91 %	77	2
40	40	5	3	195	107	78	327	23.85 %	35	2
90	90	10	3	91	58	26	285	9.12 %	61	0
90	90	5	3	111	66	30	340	8.82 %	47	5
50	50	5	3	143	78	60	324	18.52 %	67	1
70	70	5	3	158	87	39	221	17.65 %	26	0
60	60	10	3	112	68	37	206	17.96 %	49	2
100	100	5	3	197	108	30	252	11.90 %	37	2
100	100	10	3	126	76	24	240	10.00 %	52	1
20	20	10	3	177	105	100	157	63.69 %	28	1
70	70	10	3	97	53	26	151	17.22 %	39	0
40	40	10	3	109	61	62	178	34.83 %	30	1
60	60	5	3	167	92	52	164	31.71 %	43	1
200	200	5	3	132	71	9	125	7.20 %	26	0
10	10	10	3	103	54	69	124	55.65 %	36	0
200	200	10	3	133	70	8	141	5.67 %	30	0
80	80	10	3	92	42	29	128	22.66 %	38	1

Tabla 17: Resultados HDBSCAN incluyendo periodo COVID-19 (cont)

En las tablas 10 a 17 se muestran los resultados de los análisis de cada modelo para cada una de las distintas entradas. Los resultados están ordenados de mayor a menor, por retorno efectivo, tasa de éxito, retorno medio positivo, negativo, mínimo y máximo. Por ejemplo, en la fila cuatro de la tabla 14 y 15 se puede observar que se entrenó el modelo *HDBSCAN* para el periodo excluyendo al COVID-19 con los parámetros: Min. tamaño = 60; Min. muestras = 60; s = 5; M=10; K=100; B=100; L=14; P=14; Cant. velas ret = 7 (retornos entre el cierre de la última vela del patrón y el precio de cierre siete días después).

El resultado para esos parámetros es el siguiente: a) de 285 *clusters* encontrados, 58 (20.35 %) son significativos, es decir, tienen retornos post-patrón estadísticamente distintos de cero; b) de los 58 *clusters* significativos 57 corresponden a retornos positivos (estrategias *long*) y uno a retornos negativos (estrategia *short*); c) de todas las secuencias de velas del periodo de predicción únicamente 240 se agruparon en los *clusters* significativos de las cuales 150 resultaron exitosas; d) el retorno efectivo de las 240 operaciones generadas entre el precio de cierre de la última vela del patrón y siete días posteriores es del 294.58 %; e) la mediana, mínimo, máximo, media de los valores negativos, y media de los positivos de los retornos son: 1.25 %, -26.27 %, 10.04 %, -3.18 % y 3.22 % respectivamente.

7.4. Simplificaciones

Se realizaron las siguientes simplificaciones para entrenar los modelos y computar los *trades* candidatos:

- No se consideraron gastos en comisiones. Cabe mencionar que hoy en día *brokers* como *Interactive Brokers*, *Tasty Works* o *TD Ameritrade* ofrecen comisiones muy bajas, especialmente para acciones de EE.UU, y son por lo tanto despreciables.

- No se consideró el posible *slippage*[8] de las operaciones. Como se trabajó con activos líquidos (S&P500) y con precios de cierre[9] se lo puede despreciar.

- No se consideró el interés que se debe pagar por alquilar acciones para posiciones *short*. Si bien el interés depende de la acción, es también un componente menor a considerar, especialmente para acciones de EE.UU

- Se fijó el análisis de los retornos para precios *out-of-sample* en tres o siete días posteriores a los patrones. El horizonte de inversión para este tipo de estrategias *swing trading*[10] podría ser variable, es decir, para cada *cluster* se podría calcular el horizonte óptimo. Por lo tanto, se podría optimizar la salida de las posiciones para *trades* que no evolucionaron como se esperaba. Por ejemplo, dado un *cluster* que representa un patrón donde el precio baja en los días posteriores a la última vela (oportunidad para *trade short sell*), si la posición no resulta como el patrón anticipa, se podría minimizar la pérdida comprando el activo

[8] Es la diferencia entre la señal de precio de compra y el precio en el que efectivamente se efectuó la operación.

[9] En general, el mayor volumen diario se da en la apertura y cierre de mercado.

(cerrando la posición short) antes de los tres o siete días establecidos. Análogamente, se podría maximizar la ganancia optimizando la salida de las posiciones que resultaron exitosas.

- Se utilizaron velas japonesas diarias. Los mismos modelos y el mismo análisis se podría efectuar sobre precios de mayor frecuencia (horas, minutos).

- Se usaron compañías *large cap*[11]. Se podrían incorporar empresas *mid-cap* o *small-cap* cuyas acciones sean lo suficientemente líquidas.

7.5. Interpretación

Con ambos modelos, *K-MEANS* y *HDBSCAN*, se hallaron *clusters* cuyos patrones resultan en retornos distintos de cero (columna "Cant. *clusters* sig"). Si bien el primer modelo se utilizó como base, ya que es uno de los algoritmos más comunes, los resultados que arroja no son de gran utilidad práctica. Cómo se explica en la sección 3.4.1, el algoritmo agrupa los puntos en base a la menor distancia (euclidiana en este caso) entre el punto y los centros de los *clusters*. De esta manera, no logra captar patrones de alta densidad ni contemplar anomalías. La tasa de éxito de *K-MEANS* (tablas 10 - 13) se basa en una gran cantidad de ocurrencias que no serían factibles de operar con gran facilidad. Por ejemplo, en la tabla 11 se puede observar que para 200 *clusters* el algoritmo encontró 157 *clusters* significativos (78.5 %) y logró una tasa de éxito de 58.72 % con 31,208 operaciones. En cambio, para el mismo periodo, los mejores modelos de *HDBSCAN* (tablas 14 y 17) necesitaron no más de 300 *trades* para lograr tasas de éxito de hasta 70 %. Esto muestra que el segundo modelo generó *clusters* de distintos tamaños y formas identificando anomalías y ruido.
Una manera de lograr resultados más prácticos con *K-MEANS* es aumentar la cantidad de *clusters* y así reducir el porcentaje de conjuntos significativos y de ocurrencias. El problema de *clusters* convexos y de igual tamaño, explicado en la sección 3.4.4, no se presenta con *HDBSCAN* ya que cada *cluster* es de tamaño variable.
En la figura 23 se incluye un ejemplo de una predicción exitosa para la acción CMI . Las velas entre las dos primeras barras verticales limitan al patrón y las otras dos muestran la posición *long* con precio de entrada de US$ 162.22 y salida US$ 171.60 con un retorno de aproximadamente 5.75 %.

[10] *Trading* especulativo en donde las posiciones se toman por algunos días. Se diferencia del *day trading* en donde se comienza y termina el día sin posiciones.
[11] Empresa cuya capitalización bursátil es de al menos 10 billones de dólares.

Figura 23: Ejemplo predicción de HDBSCAN para el periodo excluyendo al COVID-19

8. Conclusión

A partir de los resultados de la sección 7 se concluye que es posible identificar secuencias de velas japonesas seguidas por probables subas o bajas de precios. El algoritmo *HDBSCAN* es más eficiente que *K-MEANS* para agrupar información de precios de acciones ya que logra clasificar patrones de distinta densidad, incluyendo anomalías y ruido que son ignorados al momento de ejecutar una estrategia. Así, dependiendo de los parámetros de configuración del algoritmo, computa entre 90 y 200 señales de compra y venta de activos financieros, con tasas de éxito que alcanzan el 70 % y con *ratios* de ganancia : riesgo mayores a 1:1.

Se puede observar también que la mayoría de los *clusters* significativos son de suba de precios en lugar de bajas. En general, las bajas de precios marcados ocurren de manera más abrupta que las subas de tal manera que son más difíciles de operar. Además, ocurren con menor frecuencia (en el 2019 el S&P500 subió el 60 % de los días) y, en consecuencia, es lógico que la mayoría de las estrategias sean *long*. Esto se puede apreciar en la tabla 15 donde las estrategias *short*, dependiendo del modelo, representan como máximo el 10 % de las estrategias totales.

8.1. Líneas de investigación futura

Se proponen las siguientes modificaciones e incorporaciones en los modelos y en el análisis presentado:

- Hacer uso de secuencia de velas de mayor frecuencia, empezando por velas con periodos de una hora y de algunos minutos. De esta manera, se podrían hacer estrategias de *day trading*[12].
- Analizar otros instrumentos financieros como futuros o bonos.[13]
- Incorporar compañías *mid-cap* y *small-cap*, como por ejemplo las que componen los índices *S&P MidCap 400* y *S&P SmallCap 600* respectivamente. La ventaja de analizar patrones en tales empresas es que al tener menos volumen son operadas en menor medida por algoritmos de HFT[14].
- Optimizar la entrada de los modelos. Se podrían modelar patrones de secuencias de velas variables en vez de los cinco y diez fijados en este trabajo e incorporar, además, indicadores adicionales a la entrada como por ejemplo cruce de medias móviles.
- Ampliar la estrategia incorporando información de opciones. En este trabajo la estrategia de *trading* se basó únicamente en una de las dimensiones del mercado: la dirección. Por lo tanto, las órdenes simuladas son posiciones estrictamente *long* o *short*. Añadiendo datos acerca de la volatilidad implícita[15], como por ejemplo *IV rank*[16], sería posible intentar predecir también la volatilidad implícita y hacer estrategias con opciones, futuros y los respectivos subyacentes.
- Incorporar otra información además de precios y volúmenes como indicadores del sentimiento de las empresas y del mercado.

[12]Forma de *trading* especulativo en donde se comienza y se termina el día sin posiciones. La compra y/o venta de los activos se hace en el mismo día.

[13]En este caso el análisis se debe realizar sobre el *yield* de los bonos

[14] *Trading* de alta frequencia o *High-frequency trading* es un tipo de algoritmo caracterizado por ejecutar un gran volumen de órdenes a alta velocidad (milisegundos) en los mercados de manera electrónica. Es más difícil identificar y explotar patrones en activos financieros operados por tales algoritmos ya que reducen la ventana de oportunidad de los patrones del análisis técnico.

[15]Es la volatilidad del subyacente de la opción que al ser ingresada al modelo de valuación de la opción devuelve el precio teórico igual al actual precio de mercado.

[16]Mide la volatilidad implícita actual en relación a la histórica reflejando su nivel (alto o bajo). Los *traders* de opciones intentarán vender opciones cuando la volatilidad se encuentra alta con pronóstico de baja y comprar opciones en el caso contrario.

- Usar otros algoritmos de aprendizaje no supervisado como *affinity propagation*[17].

La mayoría de los puntos mencionados anteriormente, como por ejemplo la incorporación de señales de precios de mayor frecuencia, conllevan a un mayor costo computacional. Por lo tanto, se recomiendan las siguientes mejoras: 1) usar algoritmos paralelizables y 2) ejecutarlos en la nube.

[17] Algoritmo basado en el pasaje de mensajes entre los distintos puntos.

Referencias

Candlestick chart. Candlestick chart — Wikipedia, the free encyclopedia, 2020. URL `https://en.wikipedia.org/wiki/Candlestick_chart`. [Online; accessed 10-Oct-2020].

How HDBSCAN Works. Scikit-learn: Machine learning in python, 2020. URL `https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html`. [Online; accessed 10-Oct-2020].

Jr. J. Welles Wilder. *New Concepts in technical trading systems.* Trend Research, 1978. ISBN 0-89459-027-8.

Weiren Shi Bin Fang Jingpei Dan, Wenbo Guo and Tingping Zhang. Deterministic echo state networks based stock price forecasting. 2014.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An introduction to Cluster Analysis.* WILEY, 1990. ISBN 978-0-471-73578-6.

David C.L. Ngo Leslie C.O. Tiong and Yunli Lee. Stock price prediction model using candlestick pattern feature. *International Journal Of Interactive Digital Media*, 1(3), 2014.

A. Craig MacKinlay. Event studies in economics and finance. *Jorunal of Economic Literature*, pages 13–19, 1997.

Karsten Martiny. Unsupervised discovery of significant candlestick patterns for forecasting security price movements. 1(3), 2012.

Gregory L. Morris. *Candlestick Charting Explained: Timeless Techniques for Trading Stocks and Futures.* McGraw-Hill, 3 edition, 2006. ISBN 978-0-07-163217-1.

Daniel Muller. Modern hierarchical, agglomerative clustering algorithms. 2011.

Anuj Karpatne Pang-Ning Tan, Michael Steinbach and Vipin Kumar. *Introduction to Data Mining.* 2 edition, 2018.

Davoud Moulavi Ricardo J. G. B. Campello and Joerg Sander. *Advances in Knowledge Discovery and Data Mining*, volume 7819. 2013.

Chau-chen Yang Seng-cho T. Chou, Hsien-jung Hsu and Feipei Lai. A stock selection dss combining ai and technical analysis. *Annals of Operations Research*, pages 335–353, December 1997.

Suchita Gupta Shraddha Pandit. A comparative study on distance measuring approaches for clustering. *International Journal of Research in Computer Science*, 2(1):29–31, December 2011.

Frank Wilcoxon. 1945.

J. Welles Wilder. 1978.

A. Código python

El código mencionado a continuación se encuentran en el siguiente repositorio: `https://github.com/falcu/unsupervised-candlesticks`
Los pasos descriptos en este trabajo fueron automatizados mediante código orientado a objetos en python. Los principales paquetes de la solución son:

- **model.data.market_provider**: Descarga automáticamente los *tickers* y precios de los mismos. Además permite persistir los datos en disco.
- **model.data.model_input_provider**: Genera automáticamente las entradas de los modelos. Todos los parámetros mencionados en la ecuación (12) son configurables.
- **model.ml.model**: Encapsula el uso de los algoritmos *K-MEANS* y *HDBSCAN*.
- **model.ml.analysis**: Analiza cada modelo mediante el test no paramétrico y predice de *clusters* con precios *out-of-sample*. Computa los *trades* y sus retornos.
- **model.utilities.utilities**: Agrupa funciones generales como la serialización de datos a disco y la subsecuente lectura.

B. Identificadores de las compañías

A	AAL	AAP	AAPL	ABBV	ABC	ABMD	ABT	ACN	ADBE	ADI	ADM
ADP	ADSK	AEE	AEP	AES	AFL	AIG	AIV	AIZ	AJG	AKAM	ALB
ALGN	ALK	ALL	ALLE	ALXN	AMAT	AMCR	AMD	AME	AMGN	AMP	AMT
AMZN	ANET	ANSS	ANTM	AON	AOS	APA	APD	APH	APTV	ARE	ATO
ATVI	AVB	AVGO	AVY	AWK	AXP	AZO	BA	BAC	BAX	BBY	BDX
BEN	BF-B	BIIB	BIO	BK	BKNG	BKR	BLK	BLL	BMY	BR	BRK-B
BSX	BWA	BXP	C	CAG	CAH	CARR	CAT	CB	CBOE	CBRE	CCI
CCL	CDNS	CDW	CE	CERN	CF	CFG	CHD	CHRW	CHTR	CI	CINF
CL	CLX	CMA	CMCSA	CME	CMG	CMI	CMS	CNC	CNP	COF	COG
COO	COP	COST	COTY	CPB	CPRT	CRM	CSCO	CSX	CTAS	CTL	CTSH
CTVA	CTXS	CVS	CVX	CXO	D	DAL	DD	DE	DFS	DG	
DGX	DHI	DHR	DIS	DISCA	DISCK	DISH	DLR	DLTR	DOV	DOW	
DPZ	DRE	DRI	DTE	DUK	DVA	DVN	DXC	DXCM	EA	EBAY	
ECL	ED	EFX	EIX	EL	EMN	EMR	EOG	EQIX	EQR	ES	
ESS	ETFC	ETN	ETR	EVRG	EW	EXC	EXPD	EXPE	EXR	F	
FANG	FAST	FB	FBHS	FCX	FDX	FE	FFIV	FIS	FISV	FITB	
FLIR	FLS	FLT	FMC	FOX	FOXA	FRC	FRT	FTI	FTNT	FTV	
GD	GE	GILD	GIS	GL	GLW	GM	GOOG	GOOGL	GPC	GPN	
GPS	GRMN	GS	GWW	HAL	HAS	HBAN	HBI	HCA	HD	HES	
HFC	HIG	HII	HLT	HOLX	HON	HPE	HPQ	HRB	HRL	HSIC	
HST	HSY	HUM	HWM	IBM	ICE	IDXX	IEX	IFF	ILMN	INCY	
INFO	INTC	INTU	IP	IPG	IPGP	IQV	IR	IRM	ISRG	IT	
ITW	IVZ	J	JBHT	JCI	JKHY	JNJ	JNPR	JPM	K	KEY	
KEYS	KHC	KIM	KLAC	KMB	KMI	KMX	KO	KR	KSS	KSU	
L	LB	LDOS	LEG	LEN	LH	LHX	LIN	LKQ	LLY	LMT	
LNC	LNT	LOW	LRCX	LUV	LVS	LW	LYB	LYV	MA	MAA	
MAR	MAS	MCD	MCHP	MCK	MCO	MDLZ	MDT	MET	MGM	MHK	
MKC	MKTX	MLM	MMC	MMM	MNST	MO	MOS	MPC	MRK	MRO	
MS	MSCI	MSFT	MSI	MTB	MTD	MU	MXIM	MYL	NBL	NCLH	
NDAQ	NEE	NEM	NFLX	NI	NKE	NLOK	NLSN	NOC	NOV	NOW	
NRG	NSC	NTAP	NTRS	NUE	NVDA	NVR	NWL	NWS	NWSA	O	
ODFL	OKE	OMC	ORCL	ORLY	OTIS	OXY	PAYC	PAYX	PBCT	PCAR	
PEAK	PEG	PEP	PFE	PFG	PG	PGR	PH	PHM	PKG	PKI	
PLD	PM	PNC	PNR	PNW	PPG	PPL	PRGO	PRU	PSA	PSX	
PVH	PWR	PXD	PYPL	QCOM	QRVO	RCL	RE	REG	REGN	RF	
RHI	RJF	RL	RMD	ROK	ROL	ROP	ROST	RSG	RTX	SBAC	
SBUX	SCHW	SEE	SHW	SIVB	SJM	SLB	SLG	SNA	SNPS	SO	
SPG	SPGI	SRE	STE	STT	STX	STZ	SWK	SWKS	SYF	SYK	
SYY	T	TAP	TDG	TDY	TEL	TFC	TFX	TGT	TIF	TJX	
TMO	TMUS	TPR	TROW	TRV	TSCO	TSN	TT	TTWO	TWTR	TXN	
TXT	TYL	UA	UAA	UAL	UDR	UHS	ULTA	UNH	UNM	UNP	
UPS	URI	USB	V	VAR	VFC	VIAC	VLO	VMC	VNO	VRSK	
VRSN	VRTX	VTR	VZ	WAB	WAT	WBA	WDC	WEC	WELL	WFC	
WHR	WLTW	WM	WMB	WMT	WRB	WRK	WST	WU	WY	WYNN	
XEL	XLNX	XOM	XRAY	XRX	XYL	YUM	ZBH	ZBRA	ZION	ZTS	

Tabla 18: Tickers de las compañías