

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

MÁSTER UNIVERSITARIO EN INGENIERÍA INDUSTRIAL

Reconocimiento y predicción de patrones de velas japonesas en el mercado financiero

Autor
Marta Villagrán Prieto

Dirigido por
Juan Luis Zamora Macho

Madrid
Junio 2022

Marta Villagrán Prieto, declara bajo su responsabilidad, que el Proyecto con título **Reconocimiento y predicción de patrones de velas japonesas en el mercado financiero** presentado en la ETS de Ingeniería (ICAI) de la Universidad Pontificia Comillas en el curso académico 2020/21 es de su autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: ..... Fecha: 26 / 06 / 2023

Autoriza la entrega:

El Director del Proyecto

Juan Luis Zamora Macho

Fdo.: ..... Fecha: 26 / 06 / 2023

V. B. del Coordinador de Proyectos

Nombre del Coordinador

Fdo.: ..... Fecha: ..... / ..... / .....

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor Dª. Marta Villagrán Prieto **DECLARA** ser el titular de los derechos de propiedad intelectual de la obra: , que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, los derechos de digitalización, de archivo, de reproducción, de distribución y de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

(a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.

(b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.

(c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.

(d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.

(e) Asignar por defecto a estos trabajos una licencia Creative Commons.

(f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente*).

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

(a) Que la Universidad identifique claramente su nombre como autor de la misma

(b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.

(c) Solicitar la retirada de la obra del repositorio por causa justificada.

(d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

(a) El autor se compromete a:

(b) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.

(c) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.

(d) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.

(e) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

6º. Fines y funcionamiento del Repositorio Institucional.

La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.

- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.

- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.

- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a 26 de 06 de 2023

ACEPTA

Fdo.: .....

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA
(ICAI)

MÁSTER UNIVERSITARIO EN INGENIERÍA INDUSTRIAL

Reconocimiento y predicción de patrones de velas japonesas en el mercado financiero

Autor
Marta Villagrán Prieto

Dirigido por
Juan Luis Zamora Macho

Madrid
Junio 2022

Reconocimiento y predicción de patrones de velas japonesas en el mercado financiero

Autor: Villagrán Prieto, Marta.

Director: Zamora Macho, Juan Luis.

Entidad Colaboradora: ICAI – Universidad Pontificia Comillas.

RESUMEN DEL PROYECTO

1. Introducción

La predicción de los índices bursátiles es uno de los problemas más importantes y extremadamente desafiantes en el entrono del análisis de las series temporales financieras tanto para los inversores como para los investigadores. Esto se debe principalmente a que el mercado de valores es esencialmente un sistema inestable, no lineal y complejo de cambios dinámicos que se ve afectado por numerosos factores, como las principales políticas económicas, los decretos gubernamentales, el cambo de la situación política, la psicología de los inversores, la economía futura, etc. Es por ello que el estudio de las predicciones de los precios de las acciones a corto, medio y largo plazo es uno de los temas de investigación más atractivos y significativos en el campo de la inversión y su aplicación. Los beneficios que conlleva una predicción precisa han motivado a los investigadores a desarrollar herramientas y métodos más avanzados.

La predicción de los movimientos de los precios en el mercado de valores se considera generalmente una tarea muy difícil. Una hipótesis muy conocida entre los académicos, la Hipótesis de los Mercados Eficientes (EHM) desarrollada por Fama (1970) [14], sugiere que los precios reflejan inmediatamente toda la información disponible y que lo único que hace cambiar su valor es la llegada de nueva información. Como dicha actualización es completamente imprevisible, los precios en el mercado parecen ser generados de forma aleatoria. De este modo, no sería posible obtener sistemáticamente un exceso de rentabilidad utilizando cualquier información disponible. Según esta hipótesis, la mejor opción en el mercado sería seguir la estrategia de comprar y mantener, que consiste en no intentar predecir

los movimientos de los precios, sino simplemente indenxarse al índice en cuestión.

No obstante, varias investigaciones debaten sobre la credibilidad de la EHM, entre ellas, cabe destacar los estudios realizados por Haugen (1999) [16] y Los (2000) [27]. Haugen presenta una revisión sobre las deficiencias de esta hipótesis mediante la realización de diversos experimentos en los seis principales mercados asiáticos. Ningunos de los resultados pusieron de manifiesto esa supuesta eficiencia de los mercados financieros. Además de estos trabajos, la experiencia práctica en el trading sugiere la existencia de tendencias en los precios y que los intentos de predecir los cambios futuros en los mismos pueden generar buenos rendimientos. En la práctica, se suelen utilizar dos enfoques para analizar dichos movimientos y, por lo tanto, para predecir la evolución de los precios de las acciones. El primero es el análisis fundamental, que utiliza factores económicos para estimar los valores intrínsecos de las empresas. El segundo enfoque, conocido como análisis técnico, se basa en los principios de la Teoría de Dow y utiliza la historia de los precios para predecir los movimientos futuros del mercado. Esta visión es la que se ha adoptado en el presente trabajo.

Debido a la creciente complejidad de los mercados de valores y a los muchos factores interrelacionados que intervienen en su comportamiento, los inversores se han visto obligados a evolucionar de las técnicas convencionales a nuevas estrategias computacionales avanzadas para aumentar los beneficios de las operaciones. En décadas, no se ha escatimado esfuerzos para mejorar la predicción de los mercados, aplicando herramientas cada vez más sofisticadas y nuevas técnicas que últimamente provienen del campo de la Inteligencia Artificial, con el fin de ayudar al desarrollo de sistemas de decisión financiera. Precisamente, el objetivo de este trabajo consiste en la aplicación del algoritmo de *K-Nearest Neighbors (K-NN)* para la detección de patrones de velas japonesas.

A pesar de la gran ventaja de predecir, con mayor o menor precisión, el comportamiento futuro del mercado mediante un sistema automático de inversión, éste sólo representa una pequeña parte alícuota de un proceso de inversión completo. Las estrategias de inversión son, principalmente, sistemas de apoyo a la decisión que aprovechan las predicciones generadas por los modelos para mejorar la rentabilidad en los mercados financieros. Por lo tanto, este trabajo no se centra únicamente en el desarrollo de un modelo de predicción estricto y riguroso, sino que implementa, al mismo tiempo, una estrategia de inversión holística que pueda ser aplicada en la vida real, para así reducir al máximo la fricción entre las

investigaciones académicas y la gestión real de un portfolio. Para ello, el núcleo fundamental de este documento lo compone el desarrollo de un backtesting *walk-forward* que se ejecuta de forma dinámica, y permite simular el comportamiento de la cartera frente a distintos escenarios temporales.

2. Estado de la técnica

El análisis técnico se concibe como el conjunto de reglas que permiten predecir la rentabilidad de una acción mediante el estudio de datos históricos, pues el propio precio de la acción lo descuenta todo. Se apoya en la psicología de masas alegando que la naturaleza humana tiende al equilibrio y que, por ello, la historia sigue tendencias que se repiten. Según los analistas técnicos los precios del mercado reflejan no sólo los distintos miedos, conjeturas y estados de ánimo, racionales e irracionales, de los cientos de compradores y vendedores, sino que también reflejan sus necesidades y recursos. Estos son factores adicionales que retan al análisis y para los que no hay estadísticas que se puedan obtener. Todos ellos están, sin embargo, sintetizados, ponderados y finalmente expresados en una cifra precisa a la que un comprador, un vendedor llegan juntos y a la que hacen un trato. El precio resultante es la única cifra que importa.

El vertiginoso desarrollo de la tecnología de los mercados financieros se ha producido en un doble sentido. Por una parte, los economistas han flexibilizado su percepción de la eficacia del mercado, aceptando la posibilidad de cierta capacidad predictiva de los rendimientos del mercado por medio de algoritmos no lineales. Por otra parte, ha experimentado una transformación sustancial desde el punto de vista de las herramientas empleadas, recabando avances registrados en otras disciplinas. Tradicionalmente los procedimientos utilizados para analizar los mercados de valores se basaban en métodos estadísticos como el uso de medias móviles, los filtros de Kalman, el suavizado exponencial, la media móvil autorregresiva (ARMA) y la media móvil integrada autorregresiva (ARIMA). Sin embargo, en los últimos años se ha demostrado que no funcionan bien a la hora de predecir los índices bursátiles con precisión. En contraste con las técnicas estadísticas, los métodos de Inteligencia Artificial pueden manejar los datos aleatorios, caóticos y no lineales del mercado de valores y se han utilizado ampliamente para la predicción exacta de los precios de las acciones. Estos métodos buscan extraer patrones aprendidos de datos históricos, en un proceso conocido como entrenamiento o aprendizaje, para posteriormente aplicarlos sobre nuevos datos. Los estudios empíricos que utilizan el aprendizaje automático suelen tener dos fases principales. La primera, aborda

la selección de las variables y los modelos relevantes para la predicción, separando una parte de los datos para el entrenamiento y otra para la validación de los modelos. La segunda fase aplica los modelos optimizados a los datos de test midiendo así el rendimiento predictivo.

En este sentido, el algoritmo de *K-Nearest Neighbors* es uno de los métodos más antiguo, sencillo y preciso para los modelos de clasificación y regresión de patrones. El K-NN fue propuesto por Fix & Hodges (1951) [17] y luego fue modificado por Cover & Hart (1967) [11]. El K-NN ha sido reconocido como uno de los diez mejores métodos en el *data mining* según Wu et al. (2008) [50]. En consecuencia, ha sido estudiado en la últimas décadas y ampliamente aplicado en muchos campos. Así, se trata del algoritmo de referencias en muchos problemas como reconocimiento de patrones [51], categorización de textos [28], modelos de clasificación [18], reconocimiento de objetos [5] y detección de anomalías [41]. El K-NN utiliza un término denominado *similitud de características* para predecir los valores de cualquier punto de datos nuevos. Esto significa que a dicho punto se le asigna un valor basado en su parecido con las instancias del conjunto de entrenamiento. Para ello, computa la distancia del valor a predecir con respecto a los puntos disponibles, las ordena de forma descendiente, y selecciona los K valores más cercanos al punto en cuestión para computar el valor predicho. Aunque el algoritmo del K-NN ha sido ampliamente utilizado durante décadas debido a su simplicidad, eficacia y robustez, existen cuatro deficiencias principales a las que se enfrenta: la función de distancia empleada, la determinación del tamaño de su vecindad, la estimación de la predicción en función de las distancias calculadas y la complejidad del espacio.

En consecuencia, se han desarrollado distintas investigaciones que pretenden mejorar la eficacia del mismo. En cuanto a la revisión de las funciones de distancia, Surya et al. (2019) [39] evalúan el rendimiento del K-NN utilizando siete familias de distancia diferentes sobre distintos *datasets* reales y obtienen que la métrica de *Hassanat* [1] es la más efectiva en la mayoría de los casos. Con respecto a la fijación del valor de K existen dos vertientes de estudios: los que tratan de optimizarlo

[1]

$$\text{HasD}(x,y) = \sum_{i=1}^{n} D\left(x_i, y_i\right)$$

donde

$$D(x,y) = \begin{cases} 1 - \frac{1+\text{mín}(x_i,y_i)}{1+\text{máx}(x_i,y_i)}, & \text{mín}\left(x_i,y_i\right) \geq 0 \\ 1 - \frac{1+\text{mín}(x_i,y_i)+|\text{mín}(x_i,y_i)|}{1+\text{máx}(x_i,y_i)+|\text{mín}(x_i,y_i)|}, & \text{mín}\left(x_i,y_i\right) < 0 \end{cases}$$

de forma fija mediante diferentes tipos de algoritmos como Hyunchul & Kyoung (2008) [2] y el diseño de un algoritmo genético o Wang et al. (2005) [48] y su abogacía por el concepto de confianza estadística; y los que tratan de ajustarlo de forma dinámica para cada instancia que se predice, como Wettschereck (1994) [49] y su algoritmo K-NN local adaptativo.

Por otra parte, también se han ido introduciendo nuevas variantes del modelo K-NN original que pretenden hacer frente a las principales limitaciones de las que adolece el modelo primario. Entre ellas merece la pena destacar el diseño de un K-NN ajustado por pesos por Han et al. (2001) [20], la implementación del K-NN con vecinos más cercanos compartidos propuesta por Zheng et al. (2015) [54], el empleo conjunto del K-NN con el K-means desarrollado por Wira et al. (2012) [8], el diseño también híbrido del K-NN con SVM propuesto por Zang et al. (2006) [52], la implementación del K-NN informativo realizada por Yang et al. (2007) [42] y el K-NN Bayesiano estudiado por Giuseppe (2017) [36].

Para aplicar el algoritmo K-NN en un entorno de predicción de series temporales univariantes las variables explicativas son valores desplazados en el tiempo de la variable explicada. En el caso particular del estudio de series temporales financieras cabe destacar el análisis desarrollado por Chen & Hao (2017) [9] en el que combinan el algoritmo SVM junto con el K-NN para la predicción de dos índices chinos (Shanghai y Shenzhen). La principal motivación de emplear dicho enfoque híbrido es poder discretizar el *pool* de vecinos donde se encuentran los patrones según la tendencia del valor a predecir. Además, proponen un modelo de SVM, que basado en la ponderación de características según la ganancia de información, permita que no todas las variables explicativas contribuyan con la misma intensidad en la predicción de la tendencia del valor.

Por otra parte, también se resalta el estudio llevado a cabo por Teixeira & de Oliveira (2010) [46] en el que desarrollan un método para automatizar el trading algorítmico combinando el análisis técnico con el algoritmo K-NN. A diferencia del trabajo anterior, en este los autores emplean el algoritmo como clasificador, en lugar de como un regresor. Para ello, toman 15 acciones del índice de la Bolsa de Sao Paulo (Bovespa) durante el periodo comprendido entre el 1 de abril de 1998 y el 9 de marzo de 2009 y se sirven de los algunos indicadores técnicos para representar las características de entrada del modelo. La estrategia introducida generó unos beneficios (*out-sample*) considerablemente superiores a los de la estrategia de Buy & Hold para la mayoría de las empresas generando pocas operaciones de

compra y, en consecuencia, minimizando el riesgo de exposición al mercado. Sin embargo, existen claras limitaciones en el estudio de este documento. En primer lugar, sólo se consideran 15 acciones como referencia de un único mercado. Esto no es necesariamente erróneo, pero sí cuestiona la generalidad del enfoque propuesto a una cartera más amplia. En segundo lugar, el espacio de acciones que el modelo puede realizar es muy limitado. En tercer lugar, no se detalla nada el cálculo de los vecinos ni la metodología empleada para imputar las distancias. Por último, no se realiza ningún tipo de normalización a los indicadores de entrada y, precisamente, el K-NN se trata de un algoritmo muy sensible a las escalas.

De este modo, el artículo propuesto por Li et al. (2018) [24] supera las principales limitaciones de las que adolece este enfoque. Para explorar patrones rentables en los precios históricos de las acciones, normalmente se selecciona unos prometedores (e.g. dobles suelos) y luego se verifica su rendimiento comparando la rentabilidad condicional del patrón con la de referencia. Sin embargo, los autores proponen una filosofía alternativa, sin seleccionar ningún patrón en particular, el método propuesto explora el abanico entero de opciones para detectar los que tengan claros rendimientos predictivos, no sólo los patrones conocidos, sino también los desapercibidos y potencialmente informativos. En consecuencia, el método de predicción puede beneficiarse del tratamiento separado entre dos categorías de patrones. Para evaluar el rendimiento de la estrategia propuesta recopilan datos de 108 valores americanos seleccionados entre los componentes del índice Dow Jones y el NASDAQ en el periodo comprendido entre el 2 de enero de 1994 y el 31 de diciembre de 2014. Para ello, emplean como métricas el rendimiento anualizado y el ratio de Sharpe. Adicionalmente, realizan un análisis *out-of-sample* para estudiar la sensibilidad del modelo frente a la memoria interna empleada para detectar el patrón del K-NN (*embedding dimension, m*), el horizonte de predicción de la estrategia (*T step*) y el número de vecinos empleado en cada caso. Los autores revelan que el mejor resultado se alcanza considerando un horizonte temporal de 7 días y $K = 500$.

Un estudio muy novedoso en este espectro es el realizado por Naranjo & Santos (2020) [34]. En él proponen un sistema de recomendación basado en la lógica difusa. Esta herramienta de decisión emplea velas japonesas difusas e incluye el efecto de la devaluación de la moneda en la predicción. Para ello, primero computan la siguiente sesión de mercado mediante el sistema de predicción difuso diseñado. A continuación, se compara con el real mediante el algoritmo del K-NN. Por último, construyen una estrategia de asignación de capital basada en el todo o nada con control del riesgo, pero sin incluir la capitalización. La ventaja competitiva de este

método es que no sólo implementa un sistema de predicción, sino que desarrolla un método holístico de inversión, determinando los puntos de entrada y salida del mercado, así como, el capital a invertir en cada acción. Los autores afirman que los resultados son muy prometedores, sin embargo, no tienen en cuenta los costes de transacción, los cuales tienen un impacto primordial en este tipo de estrategias a corto plazo tan agresivas. Por lo tanto, para validar la estrategia sería necesario incluir algún escenario donde fuesen contemplados.

Otro estudio muy interesante es el realizado por Zhang et al. (2017) [53] en el que proponen una nueva metodología de dos etapas que combina la descomposición empírica modal con ensamblaje (EEMD) con un modelo multidimensional basado en el K-NN, con el objetivo de predecir el precio de cierre y máximo de cuatro índices (Nasdaq-100, S&P 500, DJI y STI) a lo largo de un periodo de 20 años comprendido entre el 25 de julio de 1996 y 2l 25 de julio de 2016. Además, implementan un K-NN muy singular, teniendo en cuenta no sólo la distancia euclídea como métrica, sino también los cambios del vector patrón de la serie temporal. Para evaluar el rendimiento de la estrategia híbrida implementada, comparan el método con 3 modelos diferentes: ARIMA, un K-NN tradicional y un K-NN con EMD que supera significativamente a los dos primeros.

Finalmente, el último estudio analizado es el propuesto por Nakagawa et al. (2018) [33] en el que utilizan el K-NN con la técnica de la deformación temporal dinámica o *indexing dynamic time warping* (IDTW) para evaluar la similitud entre dos series temporales. La aplicación de dicha técnica supone la realización de una correspondencia óptima con respecto a dos series temporales mediante la expansión y contracción no lineal del eje temporal. Para validar la eficacia de la estrategia se ejecuta el modelo sobre los índices TOPIX, S&P 500, FTSE 100, DAX 30 y CAC 40, a lo largo del periodo comprendido entre enero de 1989 y diciembre de 2005 y revelan que la estrategia conjunta ofrece predicciones mucho más fiables que el modelo original.

3. Metodología

El mercado financiero es un sistema complejo, no estacionario, muy ruidoso y dominado por la no linealidad, pero a pesar de ello no sigue un camino fortuito subordinado al gusto del azar. Dado que la Bolsa es una de las arterias más importantes del corazón financiero, existen multitud de factores que pueden provocar fluctuaciones en el movimiento de estos mercados, lo que dificulta la predicción

del precio de las acciones que cotizan en ellas. No existe una fórmula estática y cerrada capaz de batir al mercado en cualquier circunstancia, ya que el propio mercado la haría inútil en muy poco tiempo. Por ello, la dificultad de establecer un modelo que se ajuste a los cambios y a la volatilidad del mercado es la piedra angular del éxito de cualquier estrategia de inversión. Para ello, la herramienta clave que disponen la mayoría de los profesionales del sector es el empleo de un análisis retrospectivo que permita analizar lo que hubiera ocurrido de haber aplicado el método sugerido. Es por esto que el núcleo fundamental de la estrategia de inversión propuesta lo constituye el diseño conjunto de dos backtesting que actúan en dos niveles complementarios:

- En primer lugar, existe uno que se ejecuta en una capa inferior. Este se encarga de calcular para cada día la rentabilidad predicha, en un horizonte de predicción fijo, del K-NN. Para ello, toma para cada instante temporal una ventana deslizante de un año para encontrar los patrones de las velas japonesas más similares al valor a predecir, computa la rentabilidad predicha en función de la rentabilidad de los vecinos más cercano y establece una ordenación de los valores de acuerdo con el triple criterio de la tasa de acierto y la rentabilidad media y mediana. Dicho backtesting se ha implementado haciendo uso de las técnicas de la computación en paralelo.

- En segundo lugar, el backtesting de la capa superior, toma la información histórica generada por el otro y ejecuta las señales de compraventa para así determinar los instantes temporales en los que se compra el mejor valor y poder calcular la rentabilidad final del método.

A continuación, se explica de forma más detallada los entresijos fundamentales de cada uno de ellos.

3. 1. Explicación del backtesting del nivel inferior

El método diseñado en el backtesting del nivel inferior se puede estructurar en torno a tres ejes principales: la construcción de los indicadores de entrada, la búsqueda de patrones similares en una ventana deslizante y la ordenación de los valores según un criterio de clasificación específico. En este sentido, se toman como entradas la evolución temporal de tres velas japonesas de un día, una semana y un mes, con el fin de sintetizar las fluctuaciones del mercado en diferentes horizontes temporales. Dichas señales se generan mediante la parametrización de los datos anteriores a lo largo de la dimensión de la ventana en cuestión, tal y como se puede

apreciar en la Figura 1.

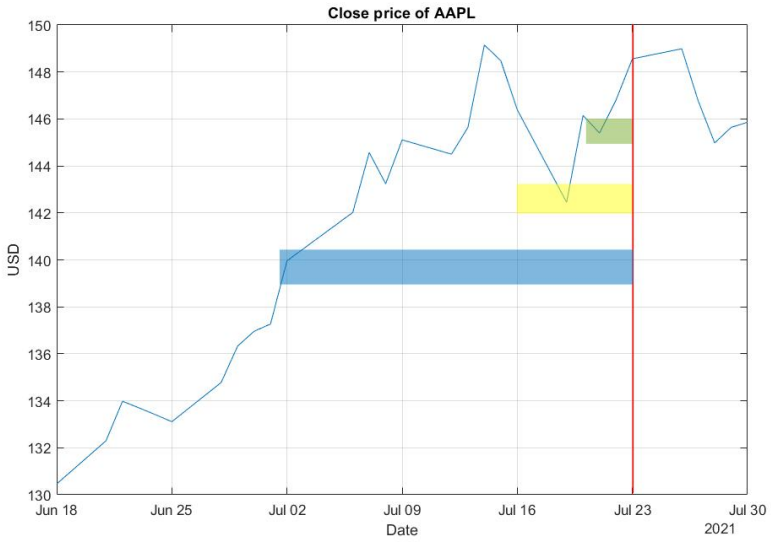


Figura 1. *Construcción del input*

El resultado de aplicar esta técnica de transformación de los datos se puede visualizar en la Figura 2.

En cuanto a la búsqueda de patrones, se utiliza el algoritmo K-NN para detectar comportamientos similares de los activos. Para su adaptación resulta esencial entender qué datos se emplean como conjunto de entrenamiento, qué se entiende por similitud en los puntos de entrenamiento, y cómo se efectúa la predicción de la rentabilidad del valor a predecir. En relación con el conjunto de entrenamiento, se utiliza una ventana deslizante de un año, que recorre el subperíodo seleccionado entre los 10 años disponibles. De este modo, se construye una ventana de 1 año hacia atrás que contiene la evolución temporal de las 15 coordenadas explicadas en el apartado anterior para aquellos valores, entre los 2555 disponibles, de los que se dispone de información suficiente. Así, el algoritmo busca los vecinos más cercanos para un valor determinado, entre un conjunto total de 643860 datos (252 días por 2555 valores). En cuanto a la cercanía, hay que tener en cuenta dos factores principales: el número de vecinos más cercanos, K, y la distancia utilizada para evaluar cada punto. En cuanto a K, se calcula la distancia por pares considerando las siguientes opciones $[100, 150, 250, 300, 350, 400, 450, 500, 750, 1000, 1500, 2000]$ de ve-

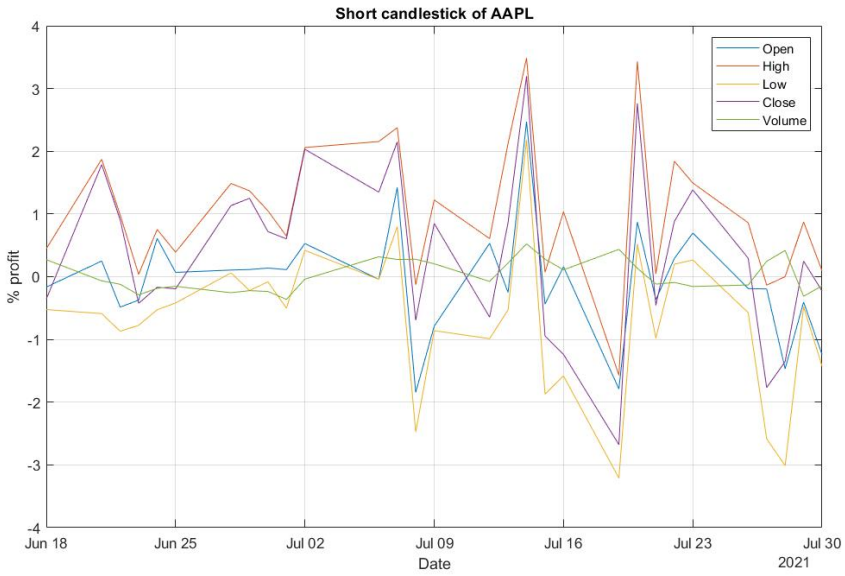


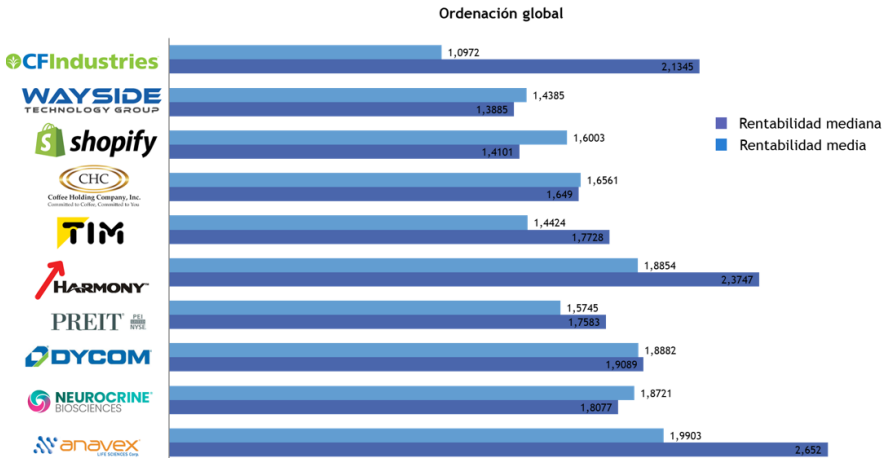
Figura 2. *Velas japonesas en la ventana corta para Apple*

cinos. Además, se utiliza la distancia *Manhattan* [2] como métrica, ya que utiliza el valor absoluto en lugar de la suma de cuadrados. Al tratarse de una normalización L_1, da menos peso a los outliers y es, por tanto, un indicador más robusto. Por último, con respecto a la predicción, se toma como rentabilidad predicha la media y mediana de las rentabilidades de los K vecinos más cercanos calculados anteriormente. Además, también se calcula la tasa de acierto del algoritmo como el porcentaje de valores cuya rentabilidad predicha es positiva.

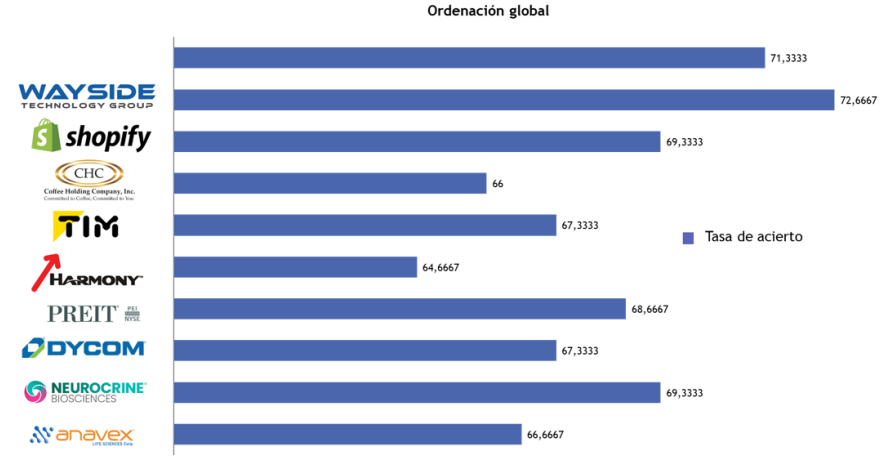
Una vez que se tiene la predicción de la rentabilidad de los valores, se procede con la ordenación de los mismos. Para ello, se construyen tres índices diferentes compuestos por la tasa de acierto, la rentabilidad media y la rentabilidad mediana predicha. De este modo, se generan tres ordenaciones descendientes distintas según cada uno de los criterios anteriores. A continuación, se construye un índice equiponderado a partir de los tres anteriores. Finalmente, se tiene por resultado la Figura 3 para cada día del periodo simulado. De este modo, se obtiene una ordenación de los valores diariamente que es lo que permite seleccionar las mejores acciones a comprar en cada sesión bursátil.

[2] $Distancia = \sum_{i=1}^{15} |x_i - y_i|$

Por último, según el criterio de ordenación, se ponderan tres métricas (la tasa de acierto y la rentabilidad media y mediana predicha) en un índice para evaluar la robustez de la rentabilidad de cada predicción. Con respecto a la duración de cada operación, se trata de una estrategia de inversión ciega, ya que tiene un horizonte de predicción fijo de una semana. Por otra parte, para sofisticar las señales de compraventa de los valores, las predicciones generadas por el K-NN se someten a umbrales para que el modelo sea más selectivo en cuanto a las acciones a comprar.



(a) Rentabilidad media y mediana



(b) Tasa de acierto

Figura 3. *Ordenación global*

La secuencia del cálculo del índice global se repite diariamente a lo largo del periodo simulado, dando lugar a una evolución de posiciones en el tiempo. El mejor valor encontrado diariamente se somete a unos umbrales en la tasa de acierto y en la rentabilidad mediana, tal y como se puede observar en las Figuras 4 y 5. En caso de estar dentro de los rangos delimitados por dichas cotas, se genera una señal de compra.

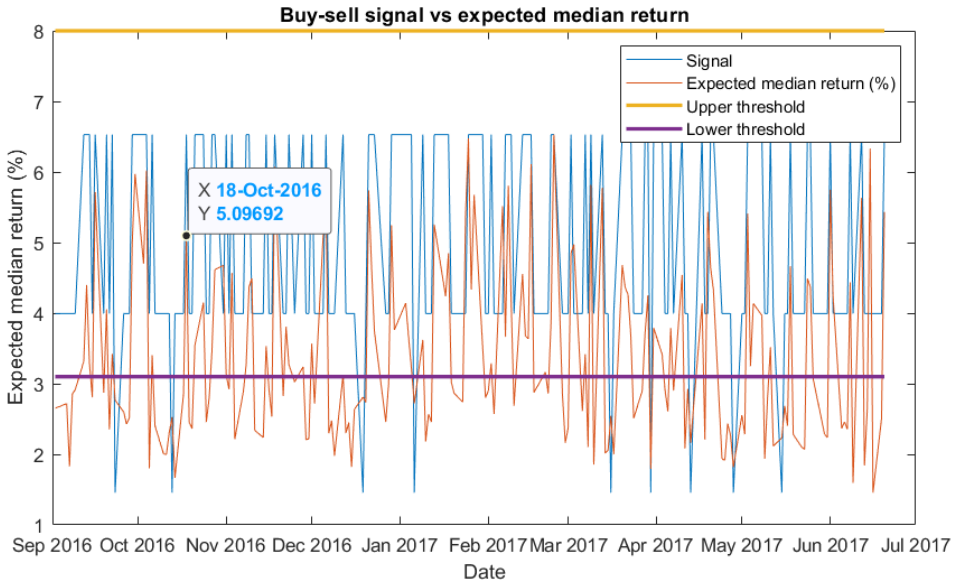


Figura 4. *Aplicación de los umbrales sobre la rentabilidad mediana con zoom*

Para ilustrar mejor cada una de las fases seguidas se incluye en la Figura 6 un esquema del procedimiento descrito.

3. 2. Explicación del backtesting dinámico del nivel superior

Si se desarrolla una estrategia de inversión teniendo en cuenta únicamente en cuenta el backtesting descrito en el apartado anterior, se puede llegar a falsos descubrimientos y a resultados sesgados por el sobreajuste de los parámetros del modelo. La razón fundamental de ello, es que en dicho método todos los hiperparámetros de la estrategia permanecen fijos a lo largo del periodo simulado. Por lo tanto, si no se actualizan cada cierta frecuencia se deja completamente de lado el dinamismo del mercado y el modelo en conjunto pierde su capacidad de adaptarse a periodos de distintas tendencias y volatilidades. Por ello, urge la necesidad de diseñar un backtesting que tenga en cuenta este detalle tan importante. Para ello se hace uso

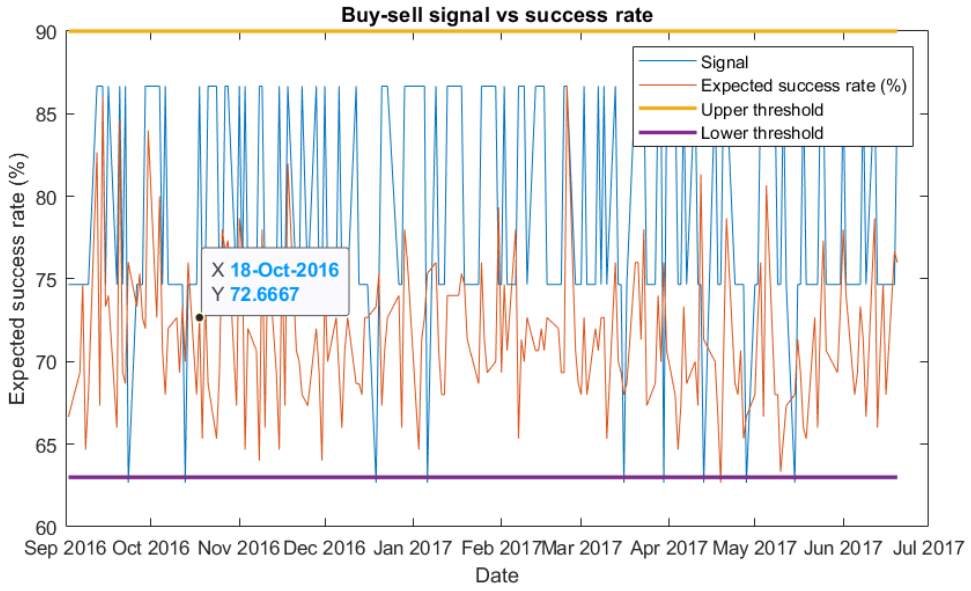


Figura 5. *Aplicación de los umbrales sobre la tasa de acierto con zoom*

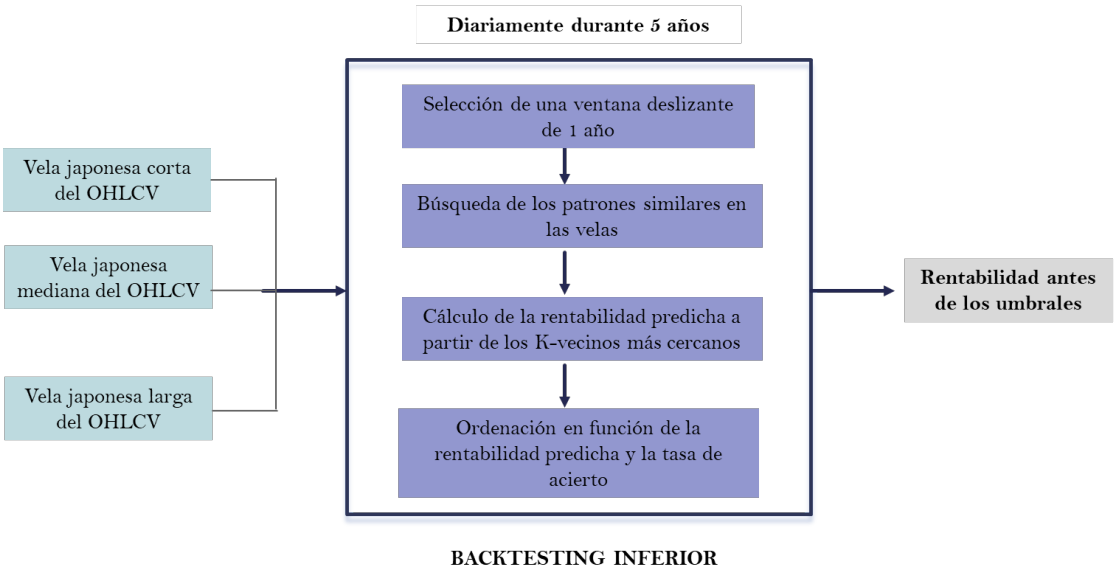


Figura 6. *Esquema del backtesting inferior*

de dos conceptos claves: la frecuencia de actualización, que se trata de un parámetro que sirve como *trigger* para la actualización de los parámetros del modelo (los

umbrales de corte descritos en la sección anterior) y, la memoria de los datos que consiste en una ventana deslizante que agrupa los datos más recientes contenidos en ella, desempeñando el papel de un búfer de información.

De este modo, la secuencia implementada es la siguiente. En primer lugar, se parte de los hiperparámetros preestablecidos y, cuando se activa la bandera de actualización, se toman los datos más recientes, generados a partir de la ordenación del K-NN, en una ventana deslizante hacia detrás para optimizar los hiperparámetros en dicha ventana, tal y como se puede observar en la Figura 7. En ella, se representa la frecuencia de actualización a través de líneas rojas verticales y, la ventana deslizante mediante rectángulos verdes. De este modo, en este escenario, el algoritmo actualizaría los parámetros cada año y emplearía un búfer de medio año para optimizar los valores. Posteriormente, se computa la estrategia de inversión con dichos valores hasta que se alcance el próximo periodo de frecuencia establecido, como se muestra en la Figura 8. En esta gráfica se muestran con líneas horizontales de diferentes colores los umbrales de corte que se obtendrían en la optimización del periodo anterior. Es decir, los primeros, los rojos, hacen referencia a los valores tomados por referencia y, por ejemplo, los verdes, son los que se obtienen tras completar la optimización cuando salta el trigger en el tercer cuatrimestre del año 2017.

Finalmente, se adjunta en la Figura 9 un esquema de las fases implementadas por el backtesting dinámico del nivel superior.

Para determinar los valores óptimos de la frecuencia y la memoria de datos, se simulan diferentes combinaciones parámetros y se obtienen los resultados adjuntos en la Tabla 1.

Frecuencia	**Memoria**	**Rendimiento anualizado (%)**
Mensual	125	17,8
Semanal	115	34,2
Diaria	115	28,4

Tabla 1. *Resultados de las simulaciones*

Adicionalmente, para optimizar el resto de hiperparámetros de la estrategia de inversión global se conduce un análisis basado en el *feature importance* que permite analizar la sensibilidad de la rentabilidad final del modelo frente a los distintos

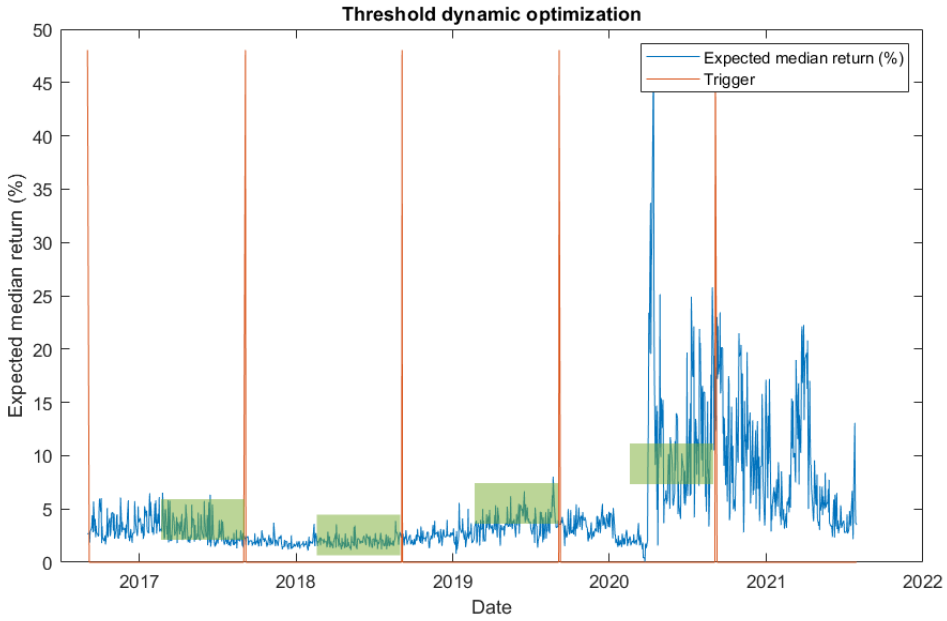


Figura 7. *Frecuencia de actualización y memoria de repetición para la optimización dinámica*

parámetros que lo configuran. Entre ellos se encuentra el número de vecinos K, la duración de la estrategia de inversión, el valor de las ventanas de las velas, la ponderación empleada en la parametrización de las mismas y el empleo de un K-NN sectorial. Los resultados experimentados revelan que se debe optar por escoger 500 vecinos, diseñar una estrategia de inversión con horizonte semanal, emplear un K-NN distinto para cada sector y normalizar las velas con el precio de apertura de la vela de la semana anterior. Estas mejoras introducidas tienen un impacto considerable sobre la rentabilidad real del K-NN, tal y como se puede observar en la Figura 10, que compara la rentabilidad original del modelo y la obtenida tras optimizar los hiperparámetros.

4. Resultados

Para analizar la validez del método propuesto se parte de la evolución temporal del precio de 2604 valores americanos que han cotizado a partir de 2011 hasta la actualidad. La fuente de datos está proporcionada por *Yahoo Finance* e incluye la información del precio de apertura, cierre, máximo y mínimo de cada una de las sesiones del intervalo muestreado, así como el volumen de las transacciones efec-

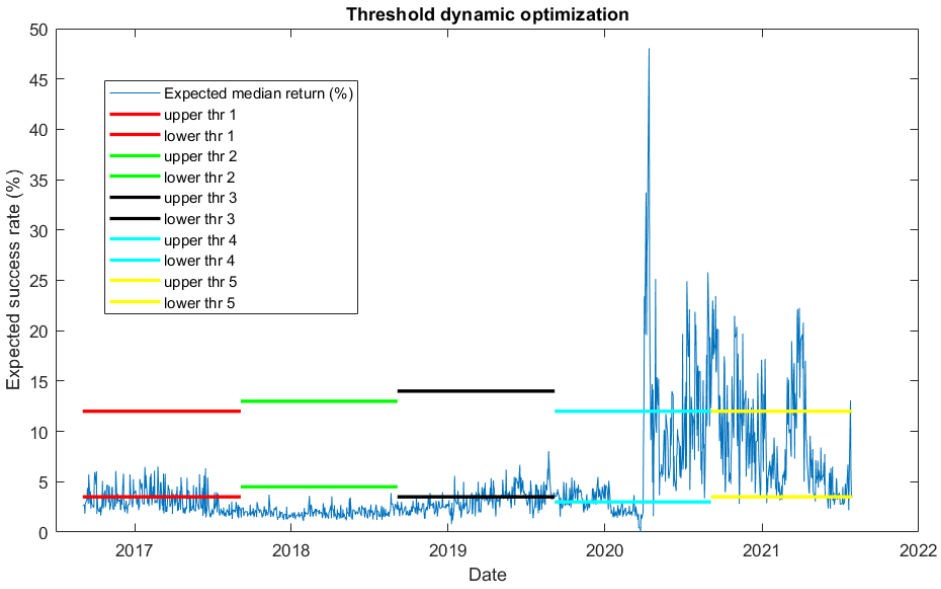


Figura 8. *Actualización de los umbrales dinámicos en la optimización*

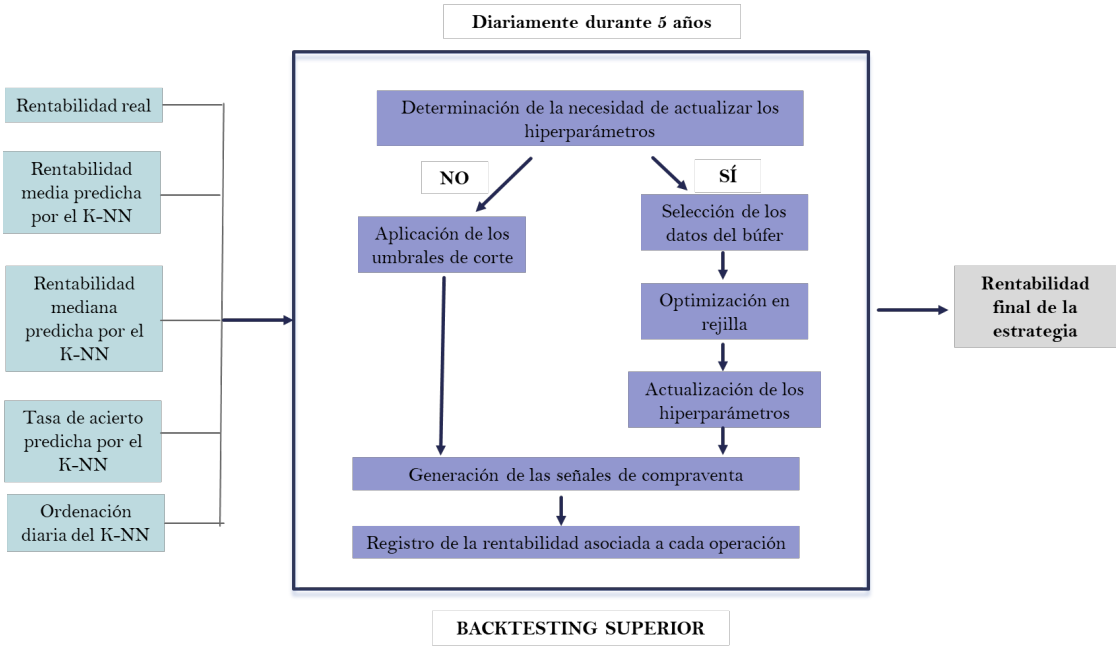


Figura 9. *Esquema del backtesting superior*

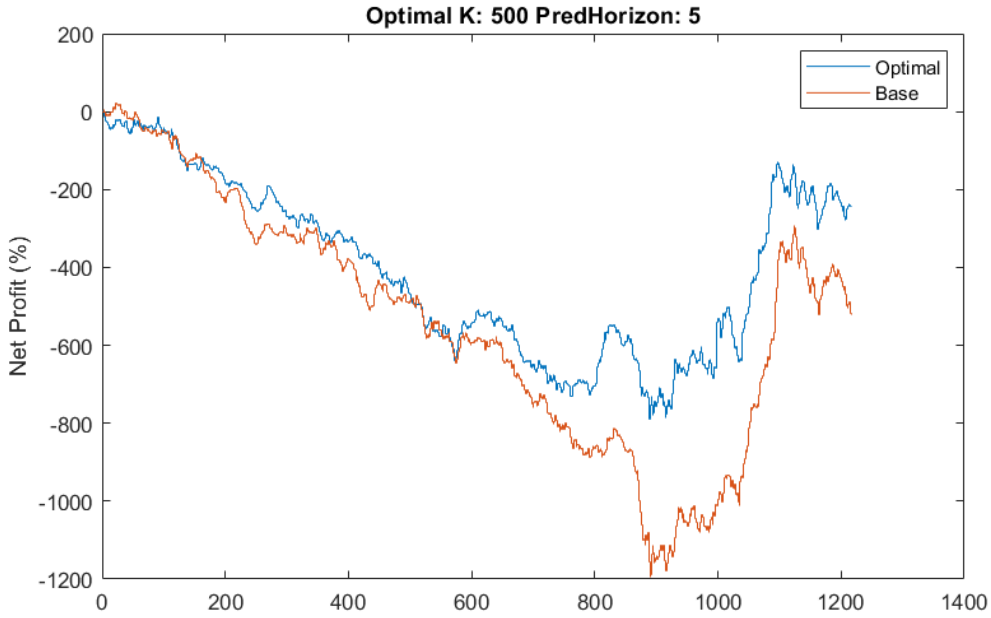


Figura 10. *Mejora de la rentabilidad con K y horizonte de predicción óptimos*

tuadas cada día. El objetivo de esta sección consiste en comparar el rendimiento del algoritmo frente a la tradicional estrategia de Buy & Hold aplicada a tres índices americanos: Dow Jones Industrial, Standard & Poor's y Russell 3000. Con ello se pretende desafiar la Hipótesis de los Mercados Eficientes.

Para proceder con este análisis comparativo se computan dos tipos de rendimientos diferentes: *trailing returns* y *rolling returns*. Los *trailing returns* se remontan a una fecha concreta para obtener la rentabilidad anualizada de un fondo/-portfolio durante un periodo de tiempo específico. En cambio, los *rolling returns* muestran los rendimientos en ciclos superpuestos que comienzan el primer día del mes, remontándose a los datos que hay disponibles. El objetivo es mostrar la frecuencia y la magnitud de los periodos buenos y malos de una inversión. La mayoría de los expertos en el mundo financiero afirman que los *rolling returns* proporcionan una imagen más fidedigna del rendimiento de un fondo de inversión [13].

A continuación, se adjuntan en las Figuras 11 y 12 los *trailing* y *rolling returns* del método y los benchmarks. Tal y como se puede apreciar, en ambos casos se bate a todos los índices al final del periodo simulado. Es innegable que una clara

limitación del método es la heterogeneidad de la distribución de la rentabilidad en el tiempo. Esto es una consecuencia directa del propio algoritmo diseñado. Al fin y al cabo, lo que realmente busca el K-NN son patrones de series temporales muy rentables a corto plazo, y esto es lo que provoca que se encuentren valores más atípicos.

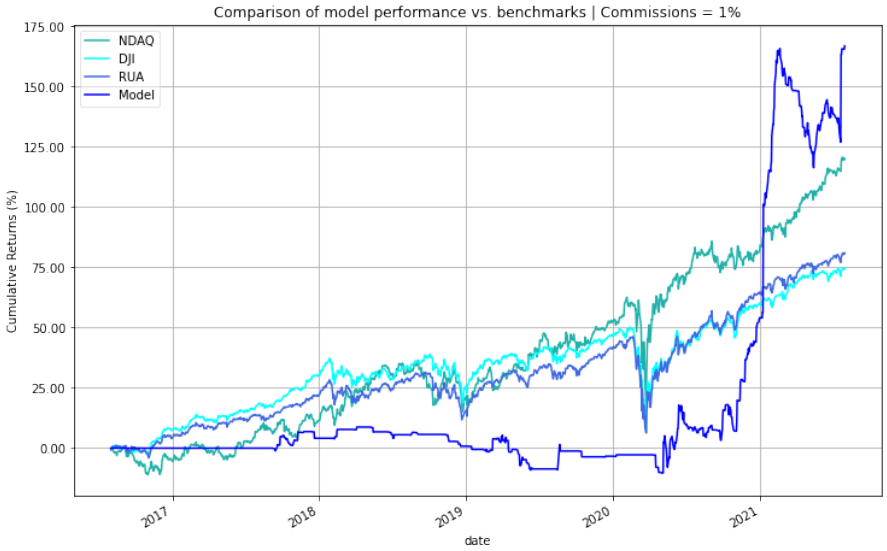


Figura 11. *Comparación de los trailing returns*

Por último, se comparan algunos ratios financieros empleados por los profesionales del sector y se adjuntan en la Tabla 2. Tal y como se puede observar en la Tabla 6.7, el método propuesto bate al mercado en todos los indicadores analizados. No obstante, el ratio de volatilidad entre ambos es de 1.5x, aunque dicho factor es menor que el presente en la comparación de ambas rentabilidades (4x), y por eso el ratio de Sortino es casi 3 veces mayor en el caso del método propuesto. Este es el indicador más fidedigno a emplear en el caso de estudio, ya que resulta crucial establecer una distinción entre la volatilidad *buena* (la presente en los últimos periodos) y la *mala*.

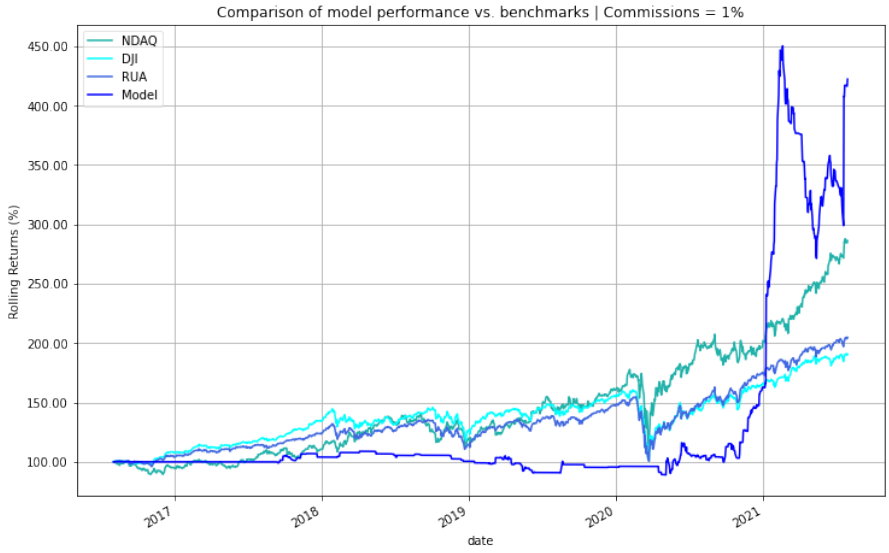


Figura 12. *Comparación de los rolling returns*

	Model	RUA
Annual return	0,414	0,158
Cumulative returns	4,645	1,048
Annual volatility	0,327	0,194
Sharpe ratio	1,201	0,855
Calmar ratio	1,108	0,451
Stability	0,401	0,840
Max drawdown	−0,374	−0,351
Omega ratio	1,913	1,200
Sortino ratio	3,805	1,174
Skew	13,270	NaN
Kurtosis	257,457	NaN
Tail ratio	1,361	0,900
Daily value at risk	−0,040	−0,024

Tabla 2. *Comparación de las estadísticas del portfolio*

5. Conclusiones

El análisis técnico siempre ha sido concebido como una práctica arcana y sus métodos han sido calificados como utopías que se alejan de la realidad. Sin embargo, con este proyecto se pone de manifiesto la eficacia de dicha técnica cuando se emplea de forma racional. Mediante la aplicación del algoritmo del K-NN a la predicción de las series temporales financieras se revela que existen ciertos patrones informativos muy rentables que permiten la generación de retornos muy positivos para el inversor.

El método propuesto no se centra únicamente en la predicción del precio de las acciones, sino que desarrolla un entorno íntegro para la implementación de estrategias de inversión bursátiles, disminuyendo completamente la fricción existente entre las investigaciones académicas y su aplicación directa para la construcción de métodos que permitan desarrollar el trading algorítmico.

A lo largo del proyecto se ha realizado un exhaustivo estudio de las principales variables que tienen un impacto sobre la rentabilidad final del método. Asimismo, tras ser optimizadas, y para validar la eficacia del método propuesto, se ha enfrentado la cartera generada por el mismo a la estrategia de comprar y mantener aplicada a tres índices americanos. Los resultados experimentales revelan que el método bate a todos los índices al final del periodo simulado.

Pattern recognition and prediction of japanese candlestick in stock markets

Author: Villagrán Prieto, Marta.

Director: Zamora Macho, Juan Luis.

Collaborating Entity: ICAI – Universidad Pontificia Comillas.

PROJECT ABSTRACT

1. Introduction

Predicting stock market indices is one of the most important and extremely challenging problems in financial time series analysis for both investors and researchers. This is mainly because the stock market is essentially an unstable, non-linear and complex system of dynamic changes that is affected by numerous factors, such as major economic policies, governmental decrees, changing political situation, investors' psychology, future economy, etc. That is why the study of short, medium and long term stock price forecasting is one of the most attractive and significant research topics in the field of investment and its application. The benefits of accurate forecasting have motivated researchers to develop more advanced tools and methods.

Predicting price movements in the stock market is generally considered a very difficult task. A well-known hypothesis among academics, the Efficient Markets Hypothesis (EHM) developed by Fama (1970) [14], suggests that prices immediately reflect all available information and that the only thing that causes stock prices to change is the arrival of new information. Since such updating is completely unpredictable, prices in the market appear to be randomly generated. Thus, it would not be possible to systematically obtain excess returns using any available information. According to this hypothesis, the best option in the market would be to follow the Buy-and-Hold strategy, which consists of not trying to predict price movements, but simply indenxing oneself to the index in question.

However, several research studies debate the credibility of the EHM, including

studies by Haugen (1999) [16] and Los (2000) [27]. Haugen presents a review of the shortcomings of this hypothesis by conducting several experiments in the six major Asian markets. None of the results revealed the alleged efficiency of financial markets. In addition to this work, practical experience in trading suggests that price trends exist and that attempts to predict future price changes can generate good returns. In practice, two approaches are commonly used to analyze such movements and, therefore, to predict the evolution of stock prices. The first is fundamental analysis, which uses economic factors to estimate the intrinsic values of companies. The second approach, known as technical analysis, is based on the principles of Dow Theory and uses price history to predict future market movements. This view is the one adopted in this paper.

Due to the increasing complexity of stock markets and the many interrelated factors involved in their behavior, investors have been forced to evolve from conventional techniques to new advanced computational strategies to increase trading profits. In decades, no effort has been spared to improve market prediction, applying increasingly sophisticated tools and new techniques that lately come from the field of Artificial Intelligence, in order to help the development of financial decision systems. Precisely, the aim of this work consists in the application of the *K-Nearest Neighbors (K-NN)* algorithm for the detection of Japanese candlestick patterns.

Despite the great advantage of predicting, with greater or lesser accuracy, the future behavior of the market by means of an automatic investment system, this only represents a small aliquot part of a complete investment process. Investment strategies are mainly decision support systems that take advantage of model-generated predictions to improve returns in financial markets. Therefore, this work does not focus solely on the development of a strict and rigorous prediction model, but implements, at the same time, a holistic investment strategy that can be applied in real life, thus minimizing the friction between academic research and actual portfolio management. To this end, the fundamental core of this document is the development of a backtesting system that is executed dynamically and allows simulating the behavior of the portfolio against different time scenarios.

2. Review of literature

Technical analysis is conceived as a set of rules for predicting the profitability of a stock by studying historical data, since the stock price itself discounts everything.

It is based on mass psychology, arguing that human nature tends towards equilibrium and that history therefore follows repeating trends. According to technical analysts market prices reflect not only the various fears, conjectures and moods, rational and irrational, of the hundreds of buyers and sellers, but also reflect their needs and resources. These are additional factors that challenge analysis and for which there are no statistics to be obtained. All of them are, however, synthesized, weighted and finally expressed in a precise figure at which a buyer, a seller arrive together and make a deal. The resulting price is the only figure that matters.

The dizzying development of financial market technology has been twofold. On the one hand, economists have relaxed their perception of market efficiency, accepting the possibility of some predictability of market returns through nonlinear algorithms. On the other hand, it has undergone a substantial transformation from the point of view of the tools used, drawing on advances in other disciplines. Traditionally, the procedures used to analyze stock markets were based on statistical methods such as the use of moving averages, Kalman filters, exponential smoothing, autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA). However, in recent years it has been shown that they do not perform well in predicting stock market indices accurately. In contrast to statistical techniques, Artificial Intelligence methods can handle the random, chaotic and nonlinear data of the stock market and have been widely used for accurate stock price prediction. These methods seek to extract learned patterns from historical data, in a process known as training or learning, and then apply them to new data. Empirical studies using machine learning usually have two main phases. The first deals with the selection of variables and models relevant for prediction, separating a part of the data for training and another for model validation. The second phase applies the optimized models to the test data, thus measuring predictive performance.

In this regard, the *K-Nearest Neighbots* algorithm is one of the oldest, simplest and most accurate methods for pattern classification and regression models. The K-NN was proposed by Fix Hodges (1951) [17] and then modified by Cover Hart (1967) [11]. The K-NN has been recognized as one of the top ten methods in *data mining* according to Wu et al. (2008) [50]. Consequently, it has been studied in the past decades and widely applied in many fields. Thus, it is the benchmark algorithm in many problems such as pattern recognition [51], text categorization [28], classification models [18], object recognition [5], and anomaly detection [41]. The K-NN uses a term called *feature similarity* to predict the values of any new

data point. This means that such a point is assigned a value based on its resemblance to instances in the training set. To do this, it computes the distance of the value to be predicted with respect to the available points, sorts them in descending order, and selects the K values closest to the point in question to compute the predicted value. Although the K-NN algorithm has been widely used for decades due to its simplicity, efficiency and robustness, there are four main shortcomings it faces: the distance function employed, the determination of its neighborhood size, the estimation of the prediction as a function of the computed distances, and the space complexity.

Consequently, several researches have been developed to improve its efficiency. Regarding the review of distance functions, Surya et al. (2019) [39] evaluate the performance of K-NN using seven different distance families over different real *datasets* and obtain that the metric of *Hassanat* [3] is the most effective in most cases. With respect to setting the value of K there are two strands of studies: those that try to optimize it in a fixed way using different types of algorithms such as Hyunchul Kyoung (2008) [2] and the design of a genetic algorithm or Wang et al. (2005) [48] and their advocacy for the concept of statistical confidence; and those that try to adjust it dynamically for each instance being predicted, such as Wettschereck (1994) [49] and his local adaptive K-NN algorithm.

On the other hand, new variants of the original K-NN model have also been introduced to address the main limitations of the primary model. Among them, it is worth highlighting the design of a weight-adjusted K-NN by Han et al. (2001) [20], the implementation of the K-NN with shared nearest neighbors proposed by Zheng et al. (2015) [54], the joint employment of the K-NN with the K-means developed by Wira et al. (2012) [8], the also hybrid design of the K-NN with SVM proposed by Zang et al. (2006) [52], the implementation of the informative K-NN by Yang et al. (2007) [42] and the Bayesian K-NN studied by Giuseppe (2017) [36].

To apply the K-NN algorithm in a univariate time series forecasting environ-

[3]

$$\text{HasD}(x,y) = \sum_{i=1}^{n} D\left(x_i, y_i\right)$$

where

$$D(x,y) = \begin{cases} 1 - \frac{1+\text{mín}(x_i,y_i)}{1+\text{máx}(x_i,y_i)}, & \text{mín}\left(x_i,y_i\right) \geq 0 \\ 1 - \frac{1+\text{mín}(x_i,y_i)+|\text{mín}(x_i,y_i)|}{1+\text{máx}(x_i,y_i)+|\text{mín}(x_i,y_i)|}, & \text{mín}\left(x_i,y_i\right) < 0 \end{cases}$$

ment the explanatory variables are time-shifted values of the explained variable. In the particular case of the study of financial time series it is worth highlighting the study developed by Chen Hao (2017) [9] in which they combine the SVM algorithm together with the K-NN for the prediction of two Chinese indices (Shanghai and Shenzhen). The main motivation of employing such a hybrid approach is to be able to discretize the *pool* of neighbors where the patterns are located according to the trend of the value to be predicted. In addition, they propose an SVM model, which based on the weighting of features according to the information gain, allows that not all explanatory variables contribute with the same intensity in the prediction of the value trend.

On the other hand, it is also highlighted the study carried out by Teixeira & de Oliveira (2010) [46] in which they develop a method to automate algorithmic trading by combining technical analysis with the K-NN algorithm. Unlike the previous work, in this one the authors employ the algorithm as a classifier, rather than as a regressor. To do so, they take 15 stocks of the Sao Paulo Stock Exchange (Bovespa) index during the period from April 1, 1998 to March 9, 2009 and use some technical indicators to represent the input characteristics of the model. The strategy introduced generated profits (*out-sample*) considerably higher than those of the Buy & Hold strategy for most of the companies by generating few buy trades and, consequently, minimizing the risk of exposure to the market. However, there are clear limitations to the study in this paper. First, only 15 stocks are considered as a benchmark for a single market. This is not necessarily wrong, but it does call into question the generality of the proposed approach to a broader portfolio. Second, the space of stocks that the model can realize is very limited. Third, nothing is detailed about the calculation of neighbors or the methodology used to impute distances. Finally, there is no normalization of the input indicators and, precisely, the K-NN is a very scale-sensitive algorithm.

Thus, the paper proposed by Li et al. (2018) [24] overcomes the main limitations that this approach suffers from. To explore profitable patterns in historical stock prices, one typically selects promising ones (e.g., double floors) and then verifies their performance by comparing the conditional return of the pattern with the benchmark. However, the authors propose an alternative philosophy, without selecting any particular pattern, the proposed method explores the entire range of options to detect those with clear predictive returns, not only the known patterns, but also the unnoticed and potentially informative ones. Consequently, the prediction method can benefit from separate treatment between two categories of

patterns. To evaluate the performance of the proposed strategy they collect data from 108 US stocks selected from the Dow Jones and NASDAQ index components over the period from January 2, 1994 to December 31, 2014. To do so, they employ annualized return and Sharpe ratio as metrics. In addition, they perform an Ôemphout-of-sample analysis to study the sensitivity of the model to the internal memory used to detect the K-NN pattern (Ôemphembedding dimension, m), the prediction horizon of the strategy (ÔemphT step) and the number of neighbors used in each case. The authors reveal that the best result is achieved considering a time horizon of 7 days and $K = 500$.

A very novel study in this spectrum is the one carried out by Naranjo & Santos (2020) [34]. In it they propose a recommendation system based on fuzzy logic. This decision tool employs fuzzy Japanese candlesticks and includes the effect of currency devaluation in the prediction. To do this, they first compute the next market session using the designed fuzzy forecasting system. Then, they compare it with the actual using the K-NN algorithm. Finally, they construct a capital allocation strategy based on all-or-nothing with risk control but without including compounding. The competitive advantage of this method is that it not only implements a prediction system, but develops a holistic investment method, determining the entry and exit points of the market, as well as, the capital to invest in each stock. The authors state that the results are very promising, however, they do not take into account transaction costs, which have a major impact on such aggressive short-term strategies. Therefore, in order to validate the strategy it would be necessary to include a scenario where they are taken into account.

Another very interesting study is the one conducted by Zhang et al. (2017) [53] in which they propose a new two-stage methodology that combines empirical modal decomposition with ensemble (EEMD) with a multidimensional model based on the K-NN, with the objective of predicting the closing and maximum price of four indexes (Nasdaq-100, S&P 500, DJI and STI) over a 20-year period from July 25, 1996 to 2l July 25, 2016. Furthermore, they implement a very unique K-NN, taking into account not only the Euclidean distance as a metric, but also the changes of the pattern vector of the time series. to evaluate the performance of the implemented hybrid strategy, they compare the method with 3 different models: ARIMA, a traditional K-NN and a K-NN with EMD that significantly outperforms the first two.

Finally, the last study analyzed is the one proposed by Nakagawa et al. (2018)

[33] in which they use the K-NN with the dynamic time warping technique or *indexing dynamic time warping* (IDTW) to evaluate the similarity between two time series. The application of this technique involves the realization of an optimal matching with respect to two time series by nonlinear expansion and contraction of the time axis. To validate the effectiveness of the strategy, the model is run on the TOPIX, S&P 500, FTSE 100, DAX 30 and CAC 40 indices over the period January 1989 to December 2005 and reveals that the joint strategy provides much more reliable predictions than the original model.

3. Metodology

The financial market is a complex system, non-stationary, very noisy and dominated by non-linearity, but despite this it does not follow a fortuitous path subordinated to the whim of chance. Given that the stock market is one of the most important arteries of the financial heart, there are a multitude of factors that can cause fluctuations in the movement of these markets, making it difficult to predict the price of the shares listed on them. There is no static, closed formula capable of beating the market under any circumstances, as the market itself would render it useless in a very short time. Therefore, the difficulty of establishing a model that adjusts to the changes and volatility of the market is the cornerstone of the success of any investment strategy. For this, the key tool available to most professionals in the sector is the use of a retrospective analysis to analyze what would have happened if the suggested method had been applied. This is why the fundamental core of the proposed investment strategy is the joint design of two backtesting methods that act on two complementary levels:

- First, there is one that runs on a lower layer. This one is in charge of calculating for each day the predicted return, in a fixed prediction horizon, of the K-NN. To do so, it takes for each time instant a sliding window of one year to find the Japanese candlestick patterns most similar to the value to be predicted, computes the predicted return based on the return of the nearest neighbors and establishes a ranking of the values according to the triple criteria of the success rate and the average and median return. This backtesting has been implemented using parallel computing techniques.

- Secondly, the backtesting of the upper layer takes the historical information generated by the other and executes the buy and sell signals in order to determine the time instants in which the best security is bought and to be able to calculate the final profitability of the method.

The fundamental intricacies of each are explained in more detail below.

3. 1. Explanation of the backtesting of the lower level.

The method designed in the lower level backtesting can be structured around three main axes: the construction of input indicators, the search for similar patterns in a sliding window and the sorting of stocks according to a specific ranking criterion. In this sense, the time evolution of three Japanese candlesticks of one day, one week and one month are taken as inputs, in order to synthesize market fluctuations over different time horizons. These signals are generated by parameterizing the above data along the dimension of the window in question, as shown in Figure 13.



Figure 13. *Input construction*

The result of applying this data transformation technique can be seen in Figure 14.

Regarding the search for patterns, the K-NN algorithm is used to detect similar behavior of the assets. For its adaptation, it is essential to understand what data are used as training set, what is understood by similarity in the training points, and how the prediction of the return of the security to be predicted is performed. In relation to the training set, a sliding window of one year is used,

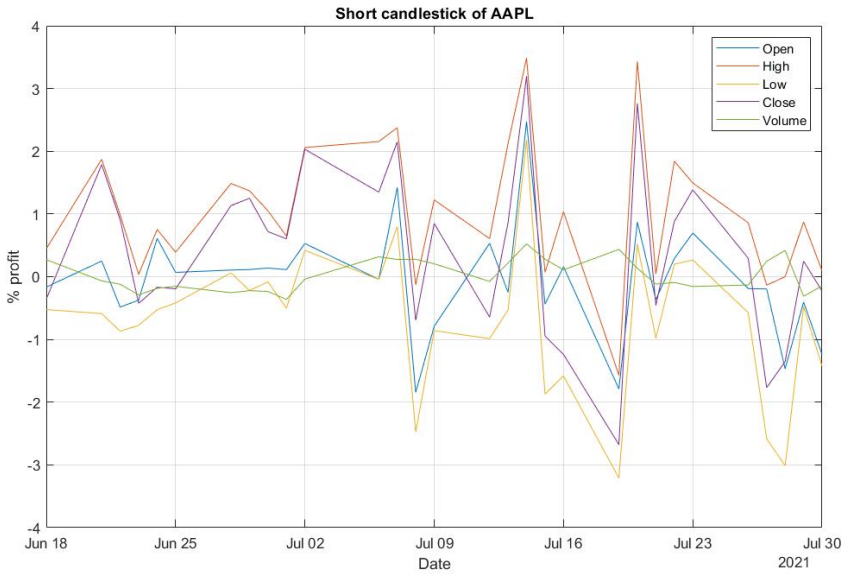


Figure 14. *Short window Japanese candlestick for Apple*

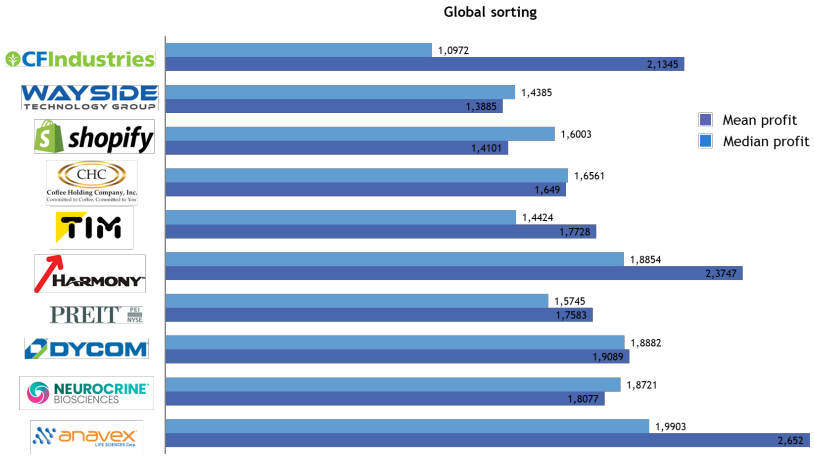
which runs through the subperiod selected among the 10 years available. In this way, a 1-year backward window is constructed containing the time evolution of the 15 coordinates explained in the previous section for those values, among the 2555 available, for which sufficient information is available. Thus, the algorithm searches for the nearest neighbors for a given value, among a total set of 643860 data (252 days for 2555 values). As for the closeness, two main factors must be taken into account: the number of nearest neighbors, K, and the distance used to evaluate each point. For K, the pairwise distance is calculated by considering the following choices $[100, 150, 250, 300, 350, 400, 450, 450, 500, 750, 1000, 1500, 2000]$ of neighbors. In addition, we use the distance *Manhattan* [4]as a metric, since it uses the absolute value instead of the sum of squares. Since it is a L_1 normalization, it gives less weight to outliers and is therefore a more robust indicator. Finally, with respect to prediction, the mean and median of the returns of the K nearest neighbors calculated above is taken as the predicted return. In addition, the success rate of the algorithm is also calculated as the percentage of values whose predicted return is positive.

Once the profitability of the securities has been predicted, we proceed with

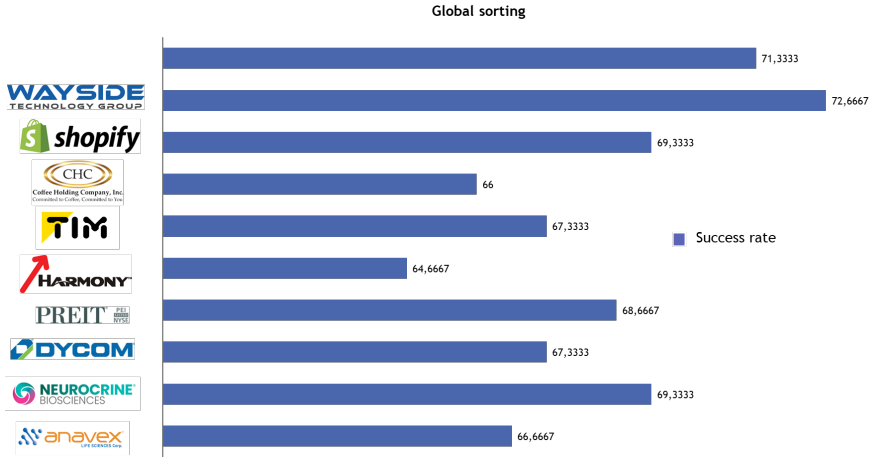
[4] $Distance = \sum_{i=1}^{15} |x_i - y_i|$

the ranking of the securities. To do this, three different indexes are constructed, composed of the success rate, the average return and the predicted median return. In this way, three different descending rankings are generated according to each of the above criteria. Next, an equal-weighted index is constructed from the three previous ones. Finally, the result is the Figure 15 for each day of the simulated period. In this way, an ordering of the stocks is obtained on a daily basis, which is what allows the best stocks to be selected for purchase in each trading session.

Finally, according to the ranking criterion, three metrics (success rate and the mean and median predicted return) are weighted in an index to evaluate the robustness of the return of each prediction. With respect to the duration of each trade, this is a blind investment strategy, as it has a fixed prediction horizon of one week. In addition, in order to make the trading signals more sophisticated, the predictions generated by the K-NN are subjected to thresholds so that the model is more selective in terms of the stocks to buy.



(a) Mean and median profit



(b) Success rate

Figure 15. *Global sorting*

The sequence of the global index calculation is repeated daily throughout the simulated period, resulting in an evolution of positions over time. The best value found daily is subjected to thresholds in the success rate and median return, as shown in Figures 16 and 17. A buy signal is generated if it is within the ranges delimited by these values.

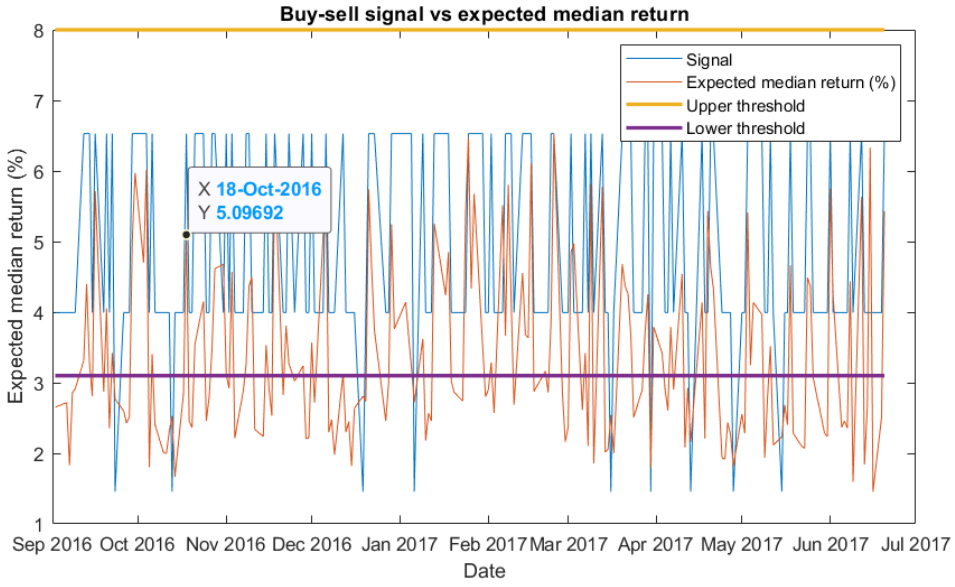


Figure 16. *Thresholds application on median profitability - Zoomed*

To better illustrate each of the steps followed, a schematic of the procedure described is included in Figure 18.

3. 2. Explanation of the dynamic backtesting of the upper level

If an investment strategy is developed with only the backtesting described in the previous section in mind, it can lead to false discoveries and biased results due to overfitting of the model parameters. The fundamental reason for this is that in such a method all the hyperparameters of the strategy remain fixed throughout the simulated period. Therefore, if they are not updated every certain frequency, the dynamism of the market is completely ignored and the model as a whole loses its ability to adapt to periods of different trends and volatilities. Therefore, there is an urgent need to design a backtesting system that takes into account this important detail. To this end, two key concepts are used: the update rate, which is a parameter that serves as a trigger for updating the model parameters (the cut-off

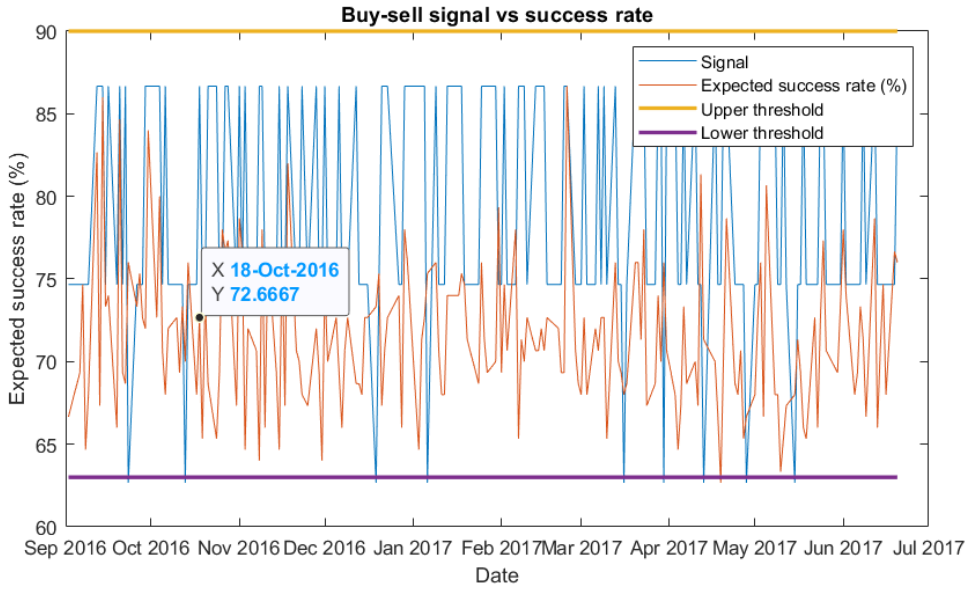


Figure 17. *Thresholds application on success rate - Zoomed*

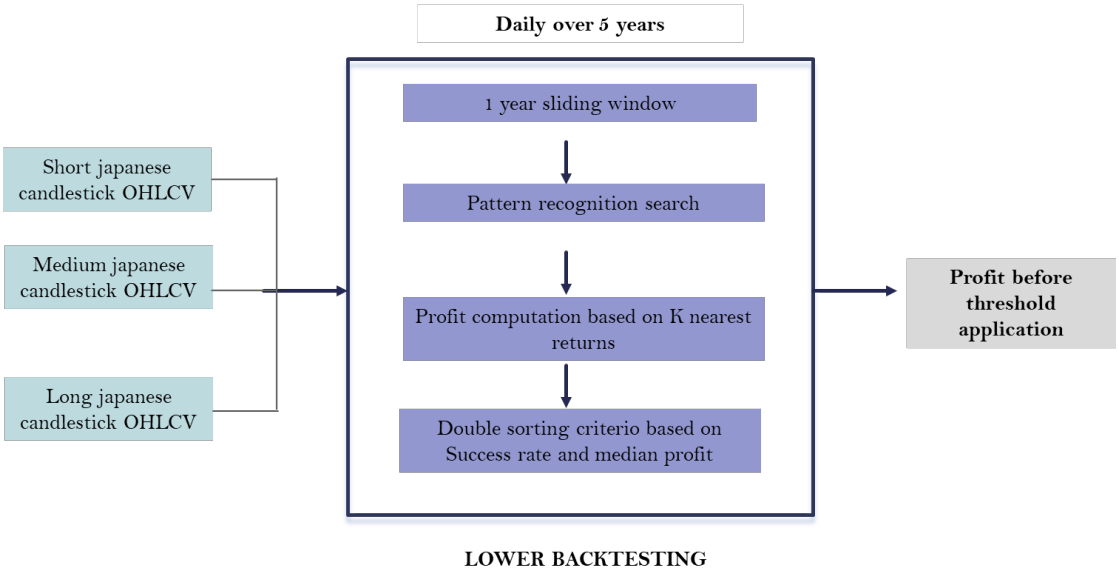


Figure 18. *Lower backtesting diagram*

thresholds described in the previous section), and the replay memory, which consists of a sliding window that groups the most recent data contained in it, playing

the role of an information buffer.

Thus, the implemented sequence is as follows. First, it is started from the preset hyperparameters and, when the update flag is set, the most recent data is taken, generated from the K-NN sorting, in a sliding window backward to optimize the hyperparameters in that window, as can be seen in Figure 19. In it, the update frequency is represented by vertical red lines and the sliding window by green rectangles. Thus, in this scenario, the algorithm would update the parameters every year and use a half-year buffer to optimize the values. Subsequently, the inversion strategy is computed with those values until the next set frequency period is reached, as shown in Figure 20. This plot shows with horizontal lines of different colors the cutoff thresholds that would be obtained in the optimization of the previous period. That is, the first ones, the red ones, refer to the values taken for reference and, for example, the green ones, are those obtained after completing the optimization when the trigger is triggered in the third quarter of 2017.

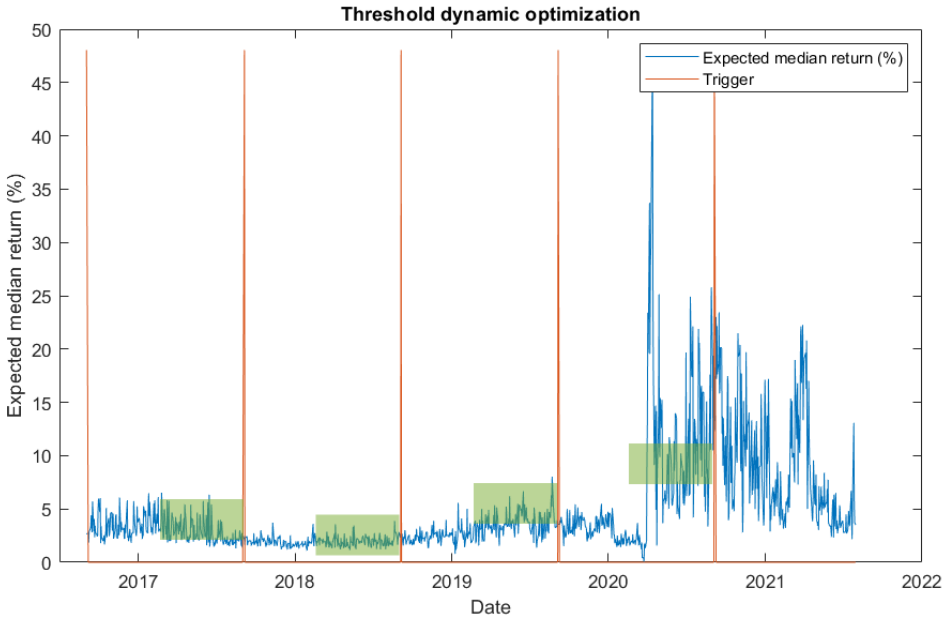


Figure 19. *Update frequency and replay memory for dynamic optimization*

Finally, a schematic of the phases implemented by the dynamic backtesting of the top level is attached in Figure 21.

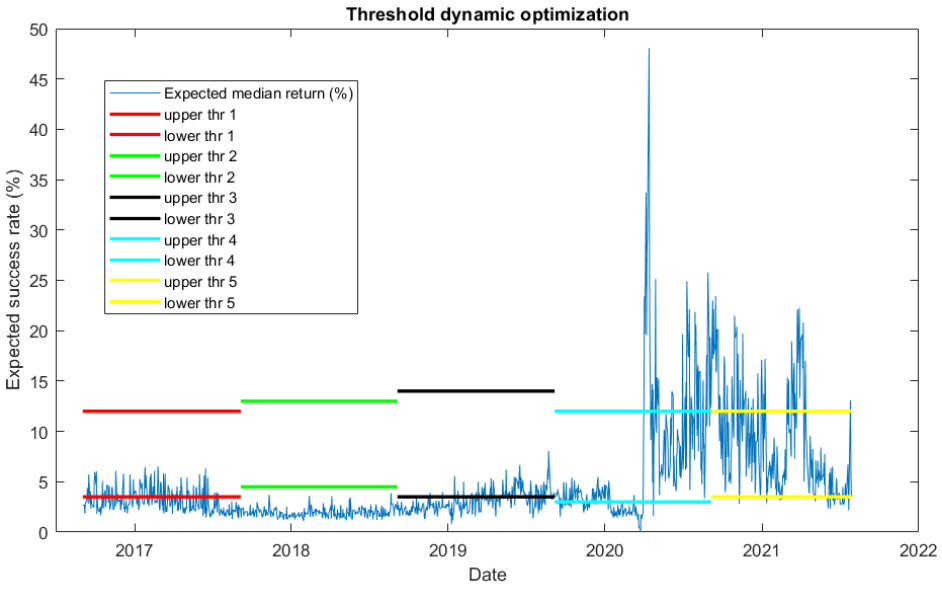


Figure 20. *Updating dynamic thresholds in optimization*

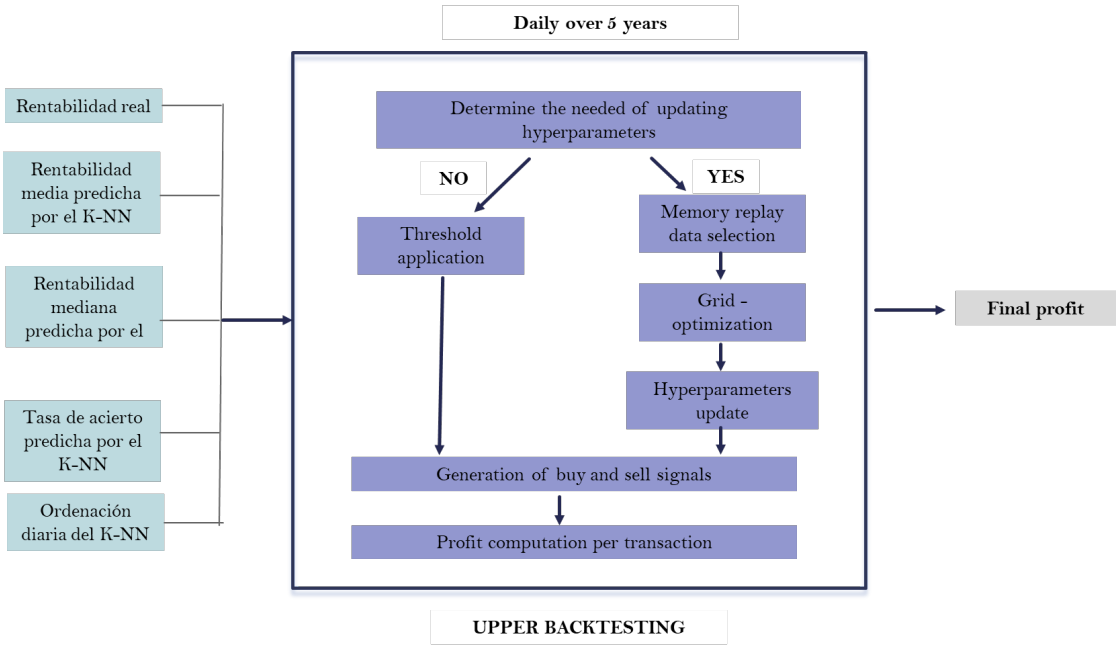


Figure 21. *Upper backtesting diagram*

To determine the optimal values for frequency update rate and memory replay, different parameter combinations are simulated and the results are shown in Table 3.

Update frequency	Memory replay	Annual Profit (%)
Monthly	125	17,8
Weekly	115	34,2
Daily	115	28,4

Table 3. *Simulation results*

Additionally, in order to optimize the rest of the hyperparameters of the global investment strategy, an analysis based on the *feature importance* is carried out to analyze the sensitivity of the final profitability of the model to the different parameters that configure it. These include the number of neighbors K, the duration of the investment strategy, the value of the candlestick windows, the weighting used in the parameterization of the candlesticks and the use of a sectoral K-NN. The experimental results reveal that one should choose 500 neighbors, design an investment strategy with a weekly horizon, use a different K-NN for each sector and normalize the candlesticks with the opening price of the previous week's candle. These introduced improvements have a considerable impact on the actual K-NN return, as can be seen in Figure 22, which compares the original model return and the one obtained after optimizing the hyperparameters.

4. Results

In order to analyze the validity of the proposed method, we start from the time evolution of the price of 2604 American stocks that have traded from 2011 to the present. The data source is provided by *Yahoo Finance* and includes the opening, closing, high and low price information for each of the sessions of the sampled interval, as well as the volume of transactions carried out each day. The objective of this section is to compare the performance of the algorithm against the traditional Buy Hold strategy applied to three American indexes: Dow Jones Industrial, Standard Poor's and Russell 3000. This is intended to challenge the Efficient Markets Hypothesis.

To proceed with this comparative analysis, two different types of returns are computed: *trailing returns* and *rolling returns*. Trailing returns go back to a spe-

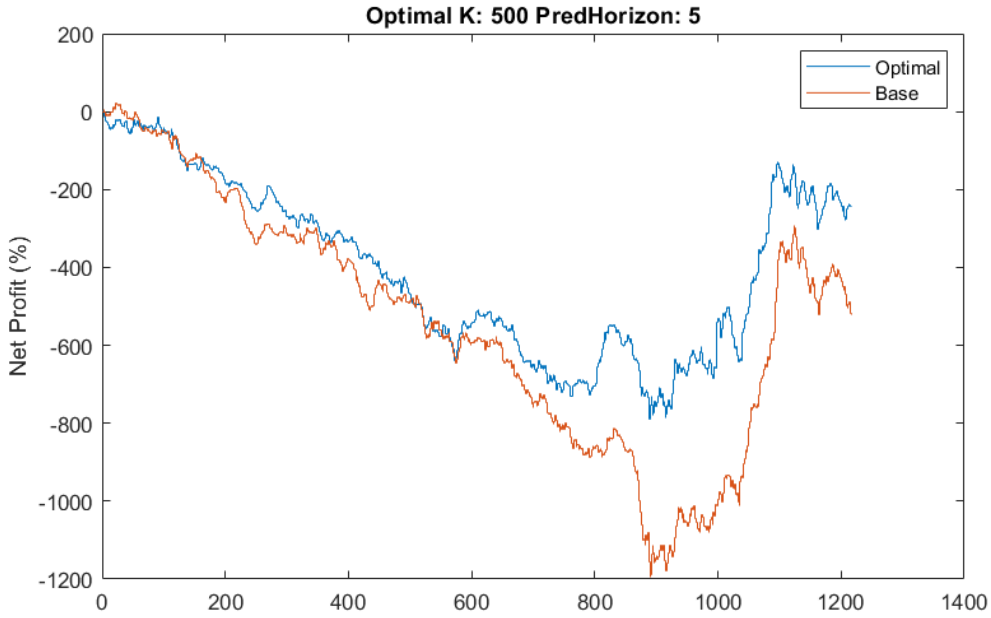


Figure 22. *Improved profitability with optimal K and prediction horizon*

cific date to obtain the annualized performance of a fund/portfolio over a specific period of time. In contrast, *rolling returns* show returns in overlapping cycles starting on the first day of the month, going back to the data that is available. The objective is to show the frequency and magnitude of good and bad periods for an investment. Most experts in the financial world claim that rolling returns provide a more reliable picture of an investment fund's performance [13].

Figures 23 and 24 show the *trailing* and *rolling returns* of the method and the benchmarks. As can be seen, in both cases all the indexes are beaten at the end of the simulated period. It is undeniable that a clear limitation of the method is the heterogeneity of the return distribution over time. This is a direct consequence of the algorithm itself. After all, what K-NN is really looking for are very profitable time series patterns in the short term, and this is what causes more outliers to be found.

Finally, some financial ratios used by industry professionals are compared and attached in Table 4. As can be seen in Table **??**, the proposed method beats the market in all the indicators analyzed. However, the volatility ratio between the

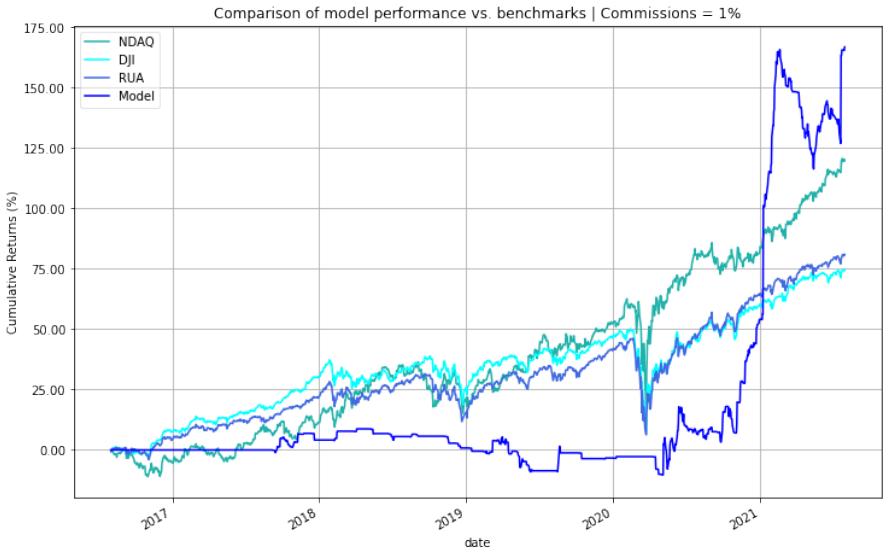


Figure 23. *Comparison of trailing returns*

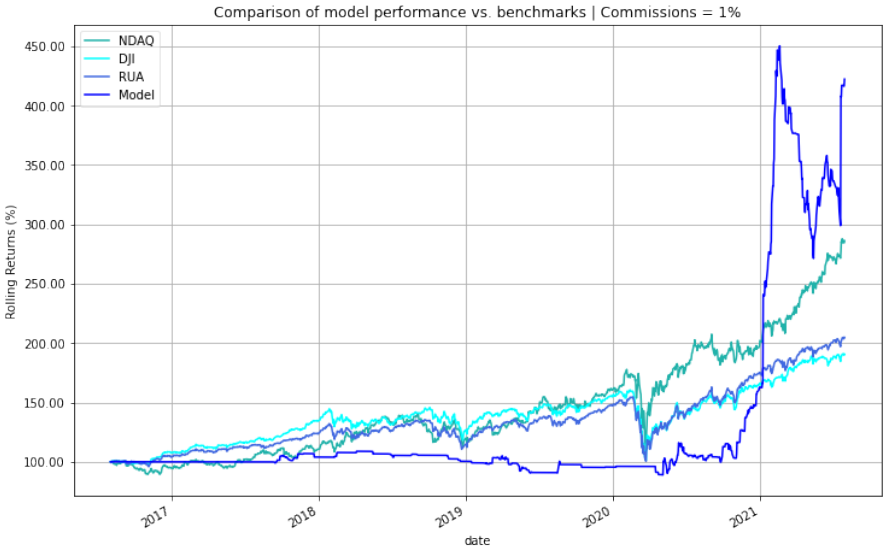


Figure 24. *Comparison of rolling returns*

two is 1.5x, although this factor is lower than the one present in the comparison of both returns (4x), and that is why Sortino's ratio is almost 3 times higher in the case of the proposed method. This is the most reliable indicator to be used in the case study, since it is crucial to distinguish between *good* volatility (that present

in the last periods) and *bad* volatility.

	Model	RUA
Annual return	0,414	0,158
Cumulative returns	4,645	1,048
Annual volatility	0,327	0,194
Sharpe ratio	1,201	0,855
Calmar ratio	1,108	0,451
Stability	0,401	0,840
Max drawdown	−0,374	−0,351
Omega ratio	1,913	1,200
Sortino ratio	3,805	1,174
Skew	13,270	NaN
Kurtosis	257,457	NaN
Tail ratio	1,361	0,900
Daily value at risk	−0,040	−0,024

Table 4. *Comparison of portfolio statistics*

5. Conclusions

Technical analysis has always been conceived as an arcane practice and its methods have been described as utopias far removed from reality. However, this project demonstrates the effectiveness of this technique when used rationally. By applying the K-NN algorithm to the prediction of financial time series, it is revealed that there are certain very profitable informative patterns that allow the generation of very positive returns for the investor.

The proposed method does not only focus on stock price prediction, but develops a complete environment for the implementation of stock market investment strategies, completely diminishing the existing friction between academic research and its direct application for the construction of methods that allow the development of algorithmic trading.

Throughout the project, an exhaustive study of the main variables that have an impact on the final profitability of the method has been carried out. Likewise, after being optimized, and in order to validate the effectiveness of the proposed method, the portfolio generated by it has been confronted with the buy-and-hold strategy applied to three American indexes. The experimental results reveal that the method beats all the indexes at the end of the simulated period.

A mis padres

Índice general

1. Introducción **1**
1.1. Motivación 1
1.2. Objetivos 2
1.3. Metodología 3
1.4. Recursos 4
1.5. Estructura de la memoria 4

2. Estado del arte **5**
2.1. Los mercados financieros 6
2.1.1. Introducción 6
2.1.2. Hipótesis de los Mercados Eficientes (EHM) 7
2.1.3. Análisis fundamental 9
2.1.4. Análisis técnico 11
2.2. Algoritmo K-NN 15
2.2.1. Introducción 15
2.2.2. Revisión de las funciones de distancia 19
2.2.3. Revisión del valor del número de vecinos K 26
2.2.4. Revisión de las adaptaciones del algoritmo 27
2.2.5. Adaptación del algoritmo para la predicción de series temporales 29
2.3. Aplicación del algoritmo K-NN a los mercados financieros 30

3. Implementación de la estrategia de inversión **49**
3.1. Construcción de las velas japonesas 50
3.2. Implementación del algoritmo K-NN 53
3.3. Criterio para la ordenación de los valores 59
3.4. Umbrales para generar las señales de compraventa 64

4. Diseño del backtesting dinámico **71**
4.1. Motivación 71
4.2. Presentación del paradigma actual 77
4.3. Explicación de la solución adoptada 85
4.3.1. Hiperparámetros a optimizar 88
4.3.2. Método de optimización 89
4.3.3. Principales dificultades 89
4.4. Implementación de las simulaciones 90
4.5. Análisis de los resultados 93

5. Optimización de la estrategia de inversión **97**
5.1. Optimización de las entradas 99
5.1.1. Análisis de las ponderaciones de las velas 101
5.2. Análisis de la granularidad de las entradas 109
5.2.1. Motivación 109
5.2.2. Implementación de las modificaciones en la estrategia 114
5.2.3. Análisis de los resultados 119
5.3. Optimización del número de vecinos 124
5.3.1. Impacto de K en el error de predicción 124
5.3.2. Impacto de K en la rentabilidad final 127
5.4. Optimización del horizonte de predicción 130

6. Resultados **133**
6.1. Comparación del modelo con los benchmarks 133
6.2. Análisis del impacto de las comisiones 136
6.3. Análisis financiero de la estrategia de inversión 142

7. Conclusiones y futuros desarrollos **151**

Apéndices **153**

A. Invirtiendo con un fin **153**

B. Informe generado para el análisis estadístico **157**

C. Informe generado para la estrategia de inversión **161**

Bibliografía **167**

Índice de figuras

1. *Construcción del input* XI
2. *Velas japonesas en la ventana corta para Apple* XII
3. *Ordenación global* XIII
4. *Aplicación de los umbrales sobre la rentabilidad mediana con zoom* . XIV
5. *Aplicación de los umbrales sobre la tasa de acierto con zoom* XV
6. *Esquema del backtesting inferior* XV
7. *Frecuencia de actualización y memoria de repetición para la optimización dinámica* XVII
8. *Actualización de los umbrales dinámicos en la optimización* XVIII
9. *Esquema del backtesting superior* XVIII
10. *Mejora de la rentabilidad con K y horizonte de predicción óptimos* . XIX
11. *Comparación de los trailing returns* XX
12. *Comparación de los rolling returns* XXI
13. *Input construction* XXX
14. *Short window Japanese candlestick for Apple* XXXI
15. *Global sorting* XXXIII
16. *Thresholds application on median profitability - Zoomed* XXXIV
17. *Thresholds application on success rate - Zoomed* XXXV
18. *Lower backtesting diagram* XXXV
19. *Update frequency and replay memory for dynamic optimization* . . . XXXVI
20. *Updating dynamic thresholds in optimization* XXXVII
21. *Upper backtesting diagram* XXXVII
22. *Improved profitability with optimal K and prediction horizon* XXXIX
23. *Comparison of trailing returns* XL
24. *Comparison of rolling returns* XL

2.1. *Representación de la altura frente a la edad* 17
2.2. *Cálculo de los pares de distancia* 18
2.3. *Selección de los K (3) valores más cercanos* 18

2.4. *Predicción one-step forward con $K = 2$* 29
2.5. *Resultados de la estrategia FWSVM-FWKNN [9]* 32
2.6. *Parámetros y partes de una vela japonesa [34]* 40
2.7. *Funciones de pertenencia para Lupper y Llower[34]* 41
2.8. *Funciones de pertenencia para Lbody [34]* 41
2.9. *Evolución del capital en el periodo de validación para las tres estrategias con el Nasdaq-100 [34]* 42
2.10. *Evolución del capital en el periodo de validación para las tres estrategias con el IBEX-35 [34]* 43
2.11. *Descomposición empírica modal de una serie temporal financiera [53]* 44
2.12. *Comparación de la predicción del precio máximo por distintos modelos [53]* 46
2.13. *Cálculo de la correspondencia de datos en diferentes series temporales [33]* 47
2.14. *Retornos acumulados para cada índice según el método [33]* 48

3.1. *Velas japonesas para AAPL en el último trimestre* 51
3.2. *Construcción del input* 52
3.3. *Velas japonesas en la ventana corta para Apple* 53
3.4. *Velas japonesas para el cierre del valor de Apple* 54
3.5. *Scatter plot de las componentes del open y volume para diferentes horizontes temporales* 56
3.6. *Scatter plot de los horizontes temporales corto y mediano para el OHLC* 57
3.7. *Rentabilidad predicha vs. real* 58
3.8. *Implementación de un bucle parfor en Matlab* 60
3.9. *Ordenación según la rentabilidad mediana* 61
3.10. *Ordenación según la rentabilidad media* 61
3.11. *Ordenación según la tasa de acierto* 62
3.12. *Ordenación global* 63
3.13. *Evolución de la posición* 64
3.14. *Aplicación de los umbrales sobre la rentabilidad mediana* 65
3.15. *Aplicación de los umbrales sobre la tasa de acierto* 66
3.16. *Aplicación de los umbrales sobre la rentabilidad mediana con zoom* . 67
3.17. *Aplicación de los umbrales sobre la tasa de acierto con zoom* 67
3.18. *Beneficio neto acumulado de la estrategia de inversión* 68
3.19. *Problema con la rentabilidad acumulada* 69
3.20. *Esquema de la estrategia de inversión propuesta* 70

4.1. *Medias móviles para IBEX 35* 73
4.2. *Medias móviles para el DJI* 74
4.3. *Medias móviles para Telefónica* 74
4.4. *Medias móviles para Telepizza* 75
4.5. *Medias móviles para Zardoya* 75
4.6. *Medias móviles para Bund* 76
4.7. *Medias móviles para el Euro Dólar* 77
4.8. *Esquema evolutivo del AG* 79
4.9. *Evolución del AG* 81
4.10. *Rendimientos netos tras la optimización de los hiperparámetros* . . 82
4.11. *Análisis de las correlaciones por año* 83
4.12. *Frecuencia de actualización y memoria de repetición para la optimización dinámica* 86
4.13. *Actualización de los umbrales dinámicos en la optimización* 87
4.14. *Esquema del backtesting inferior* 87
4.15. *Esquema del backtesting superior* 88
4.16. *Actualización mensual y memoria de 125 días* 91
4.17. *Actualización semanal y memoria de 115 días* 91
4.18. *Actualización diaria y memoria de 115 días* 92
4.19. *Sección del informe de resultados* 94

5.1. *Rentabilidad real del K-NN en el caso base* 100
5.2. *Evolución de las componentes de las velas de AAPL para diferentes horizontes temporales* 102
5.3. *Evolución de las componentes de las velas de MSFT para diferentes horizontes temporales* 103
5.4. *Evolución de las ventanas según el OHLC de las velas - Tendencia alcista* 104
5.5. *Evolución de las ventanas según el OHLC de las velas - Tendencia bajista* 105
5.6. *Comparación de las estrategias de ponderación de las entradas* . . . 107
5.7. *Comparación de las estrategias de ponderación de las entradas sin outliers* 108
5.8. *Rendimiento del sector en un periodo de recesión (Bartolini. M & Dong. A (2019)[29])* 111
5.9. *Rendimiento del sector en un periodo de recuperación (Bartolini. M & Dong. A (2019)[29])* 112
5.10. *Rendimiento del sector en un periodo de expansión (Bartolini. M & Dong. A (2019)[29])* 113

5.11. *Rendimiento del sector en un periodo de desaceleración (Bartolini. M & Dong. A (2019)[29])* 113
5.12. *Ejemplo de la selección de valores para la búsqueda de vecinos* . . . 115
5.13. *Modificación sectorial del backtesting inferior* 115
5.14. *Comparación de los K-vecinos para la estrategia global y sectorial - Open* 116
5.15. *Comparación de los K-vecinos para la estrategia global y sectorial - High* 116
5.16. *Comparación de los K-vecinos para la estrategia global y sectorial - Low* 117
5.17. *Comparación de los K-vecinos para la estrategia global y sectorial - Close* 117
5.18. *Comparación de los K-vecinos para la estrategia global y sectorial - Volume* 118
5.19. *Gráfico de barras del mejor valor según el sector - sesgado* 120
5.20. *Gráfico de barras del mejor valor según el sector* 121
5.21. *Rentabilidad media anual del mejor valor por sector* 121
5.22. *Rentabilidad en bruta de los 5 mejores valores - Caso base* 123
5.23. *Rentabilidad en bruta de los 5 mejores valores - Caso sectorial* . . . 123
5.24. *Error en la predicción vs K* 125
5.25. *Error en la predicción vs K por años* 125
5.26. *Análisis de correlaciones por año con* $K = 150$ 126
5.27. *MAE en la predicción vs K por años* 127
5.28. *MAE en la predicción vs K para distintas empresas* 128
5.29. *Influencia de K sobre la rentabilidad del K-NN* 128
5.30. *Influencia de K sobre la rentabilidad del K-NN sin outliers* 129
5.31. *Influencia del horizonte de predicción sobre la rentabilidad del K-NN sin outliers* 131
5.32. *Influencia del horizonte de predicción sobre la rentabilidad del K-NN sin outliers* 131
5.33. *Mejora de la rentabilidad con K y horizonte de predicción óptimos* . 132

6.1. *Comparación de los trailing returns* 136
6.2. *Comparación de los rolling returns* 136
6.3. *Comparación de los trailing returns - Comisiones totales del 0.5 %* . 137
6.4. *Comparación de los rolling returns - Comisiones totales del 0.5 %* . 138
6.5. *Comparación de los trailing returns - Sin costes de transacción* . . . 138
6.6. *Comparación de los rolling returns - Sin costes de transacción* . . . 139
6.7. *Underwater plot - Efecto de las comisiones* 140

6.8. *Top 3 Max Drawn del modelo* 142
6.9. *Comparación del ratio de Sharpe móvil del modelo frente a B&H del RUA* 146
6.10. *Comparación de la beta del modelo frente a B&H del RUA* 146
6.11. *Comparación de la volatilidad del modelo frente a B&H del RUA* . . 147
6.12. *Distribución de los rendimientos anuales* 148
6.13. *Mapa de calor de los retornos mensuales* 148
6.14. *Distribución de los rendimientos mensuales* 149

Índice de tablas

1. *Resultados de las simulaciones* XVI
2. *Comparación de las estadísticas del portfolio* XXI
3. *Simulation results* XXXVIII
4. *Comparison of portfolio statistics* XLI

2.1. *Dataset de ejemplo para la explicación del K-NN* 16
2.2. *Lp métricas de distancia Minkowski* 20
2.3. *L_1 métricas de distancia* 21
2.4. *Métricas de distancia del producto interior* 21
2.5. *Métricas de distancias al cuadrado* 21
2.6. *Métricas de distancia L_2 al cuadrado* 22
2.7. *Medidas de distancia de entropía de Shannon* 23
2.8. *Medidas de distancia de vicisitud* 23
2.9. *Datasets empleados por [39] para la medir la sensibilidad de la función de distancias* 24
2.10. *Promedio del accuracy, recall y precision sobre los datasets [39]* . . 25
2.11. *Comparación del rendimiento del SZSE entre FWSVM-FWKNN y SVM-KNN en MAPE y RMSE [9]* 33
2.12. *Resultados medios para la mejor configuración [46]* 35
2.13. *Rendimiento del portfolio según el método propuesto [24]* 38
2.14. *Rendimiento del portfolio con una estrategia de cruce de medias móviles [24]* 38
2.15. *Rendimiento del portfolio con una estrategia momentum [24]* 39
2.16. *Comparación del NMSE, MASE y MAPE para la predicción del precio máximo del S&P 500* 46
2.17. Rendimientos out-of-sample de cada método [33] 47

3.1. *Umbrales para la generación de las señales de compraventa* 65

4.1. *Umbrales para los hiperparámetros del AG* 80

4.2. *Umbrales para los hiperparámetros del backtesting dinámico* 89
4.3. *Resultados de las simulaciones* 93

5.1. *Rango de valores de los hiperparámetros del modelo* 101
5.2. *Tamaño de la ventana deslizante empleada en cada estrategia* 110
5.3. *Porcentajes de empresas según el sector* 119

6.1. *Trailing returns para el fondo de inversión Edelweiss Large Cap* . . 135
6.2. *Rolling returns para el fondo de inversión Edelweiss Large Cap* . . . 135
6.3. *Worst down periods for 1 %* 140
6.4. *Worst down periods for 0.5 %* 140
6.5. *Worst down periods for 0 %* 141
6.6. *Worst down periods for RUA* 141
6.7. *Comparación de las estadísticas del portfolio* 145

Acrónimos

K-NN	K-Nearest Neighbors
RUA	Russell 3000
DJI	Dow Jones Industrial
DJIA	Dow Jones Industrial Average
DJTA	Dow Jones Transportation Average
NDAQ	Nasdaq
ML	Machine Learning
ML	Hipótesis de los Mercados Eficientes
OHLCV	Open, High, Low, Close & Volume
ARMA	Media Móvil Autorregresiva
ARIMA	Media Móvil Autorregresiva Integrada
ANN	Redes Neuronales Artificiales
SVM	Máquinas de Soporte de Vectores
FWSVM	Máquinas de Soporte de Vectores con ajuste de pesos
EMD	Descomposición Empírica Modal
IDTW	Deformación Temporal Dinámica
EEMD	Descomposición Empírica Modal con Ensamblaje
IMF	Función Modal Intrínseca
RF	Random Forests
GA	Algoritmos Genéticos
PER	Price earning ratio
EV	Entreprise value

EBITDA	Earnings before Interest Taxes Depreciation & Amortization
CF	Cash Flows
MA	Medias Móviles
EMA	Medias Móviles Exponenciales
RSI	Índice de Fuerza Relativa
MACD	Media Móvil de Convergencia Divergencia
MSE	Error Cuadrático Medio
MAE	Error Medio Absoluto
EONIA	Euro Overnight Index Average

Capítulo 1

Introducción

1.1. Motivación

En el año 2007, el médico y economista estadounidense Michael Burry advertía que una crisis hipotecaria azotaría a la entonces mayor economía del mundo. Sin embargo, la audiencia general lo miró con recelo y escepticismo e ignoró sus advertencias bajo el precepto de que era imposible pronosticar el futuro del mercado hipotecario. Un año después, como en una reacción en cadena, la economía de los EE.UU. y por ende, la del mundo entero, se vio sumergida en el mayor cisma económico ocurrido hasta entonces.

Para entender la naturaleza de la incredulidad que abrazó al anuncio hecho por Burry habría que volver la mirada a la teoría clásica financiera que establece que es imposible predecir el comportamiento de un activo, pues según la Hipótesis de los Mercados Eficientes, las fluctuaciones de los mercados son meramente fortuitas y aleatorias. Sin embargo, el vertiginoso desarrollo de la tecnología ha permitido que los economistas flexibilicen su percepción de la eficiencia del mercado, aceptando cierta capacidad predictiva de los rendimientos financieros por medio de algoritmos más sofisticados. Este comportamiento está sustentado por la Hipótesis de los Mercados Adaptativos, según la cual, así como la selección natural garantiza la supervivencia del individuo más apto, los mercados financieros se comportan como un modelo evolutivo cuyos participantes se someten a entornos cambiantes y sólo aquellos que saben adaptar sus estrategias de inversión a las condiciones del mercado, pueden sobrevivir en un ambiente tan feroz como el del mundo de las finanzas.

Debido a la creciente complejidad de los mercados de valores y a los muchos factores interrelacionados que intervienen en su comportamiento, los inversores se han visto obligados a evolucionar de las técnicas convencionales a nuevas estrategias computacionales avanzadas para aumentar los beneficios de las operaciones. En décadas, no se ha escatimado esfuerzos para mejorar la predicción de los mercados, aplicando herramientas cada vez más sofisticadas y nuevas técnicas que últimamente provienen del campo de la Inteligencia Artificial, con el fin de ayudar al desarrollo de sistemas de decisión financiera. Precisamente, el objetivo de este trabajo consiste en la aplicación del algoritmo de *K-Nearest Neighbors (K-NN)* para la detección de patrones de velas japonesas.

A pesar de la gran ventaja de predecir, con mayor o menor precisión, el comportamiento futuro del mercado mediante un sistema automático de inversión, éste sólo representa una pequeña parte alícuota de un proceso de inversión completo. Las estrategias de inversión son, principalmente, sistemas de apoyo a la decisión que aprovechan las predicciones generadas por los modelos para mejorar la rentabilidad en los mercados financieros. Por lo tanto, este trabajo no se centra únicamente en el desarrollo de un modelo de predicción estricto y riguroso, sino que implementa, al mismo tiempo, una estrategia de inversión holística que pueda ser aplicada en la vida real, para así reducir al máximo la fricción entre las investigaciones académicas y la gestión real de un portfolio. Para ello, el núcleo fundamental de este documento lo compone el desarrollo de un backtesting *walk-forward* que se ejecuta de forma dinámica, y permite simular el comportamiento de la cartera frente a distintos escenarios temporales.

1.2. Objetivos

Por todo lo expuesto anteriormente el trabajo propuesto se enfoca en la consecución de los siguientes objetivos:

1. Diseñar una estrategia de inversión que, tomando como núcleo central el algoritmo K-NN, sea capaz de detectar patrones de comportamiento en los activos.

2. Construir un backtesting dinámico para validar la eficacia del método propuesto.

3. Optimizar los hiperparámetros de la estrategia de inversión diseñada.

4. Establecer un marco comparativo entre el rendimiento derivado del portfolio y el obtenido mediante la aplicación de una estrategia de Buy & Hold sobre distintos índices del mercado americano.

1.3. Metodología

A continuación, se presenta el procedimiento desarrollado para la consecución de cada uno de los objetivos descritos anteriormente. Mediante la revisión del estado del arte, se profundizará en las investigaciones que se han desarrollado en este entorno. Con ello, se podrá estudiar la viabilidad de los distintos avances alcanzados y tomarlos como semillas para poder generar los frutos deseados.

El diseño de estrategias de inversión bursátiles es el núcleo de los debates sobre la gestión financiera de las últimas décadas. Por ello, es necesario conocer los entresijos que se esconden tras la toma de decisiones económicas. Para ello, se estudiarán las bases que sustentan la predicción de las series temporales, la generación de señales de compraventa y la asignación de capital de un portfolio.

Una vez que se disponga de un conocimiento exhaustivo del problema a resolver se procederá con la implementación del modelo basado en el algoritmo del K-NN. En primer lugar, se construirá una base de datos con el histórico de precios de 2604 valores que cotizan en distintos índices americanos, incluyendo la información de precios diarios, así como distintos indicadores técnicos a lo largo de 10 años. A continuación, se implementará el algoritmo del K-NN para la predicción de las rentabilidades del modelo. Posteriormente, se construirán dos backtestings, uno que actúa en una capa superior y otro que se ejecuta en un nivel inferior, que permitan validar la eficacia de la estrategia de inversión planteada. Adicionalmente, se realizarán estudios que, basados en las técnicas de *feature importance*, permitan valorar la sensibilidad de la rentabilidad final según distintos hiperparámetros. Esto permitirá obtener la mejor combinación de los mismos y facilitará la comparación del rendimiento del modelo con la estrategia de comprar y mantener (Buy & Hold) de un ETF del Dow Jones, del Nasdaq y del Russell 3000. Con esto se tendrá una excelente representación de la validez del método propuesto en distintos escenarios y simulaciones propuestas.

1.4. Recursos

La automatización de las decisiones financieras se codificará empleando *Matlab* como principal lenguaje de programación. Adicionalmente se hará uso de *Python* para llevar a cabo los análisis exploratorios explicativos. Para el desarrollo del algoritmo del K-NN se empleará como código madre la función preestablecida de *Matlab*, adaptándola posteriormente al problema en cuestión. Asimismo, se hará uso de la *Toolbox, Parallel Computing,* de este lenguaje para la implementación en paralelo del citado algoritmo. El resto de elementos que conforman la estrategia de inversión son de propia autoría. Por último, se dispondrá de un conjunto de datos obtenidos a partir de *Yahoo Finance* que proporcionan la evolución temporal de los precios máximos, mínimos, de apertura y cierre y el volumen de transacciones para 2604 valores que cotizan a lo largo de 10 años en distintos índices americanos.

1.5. Estructura de la memoria

El presente trabajo está configurado atendiendo al siguiente esquema. En el segundo capítulo, se pretende revisar el estado del arte en lo que se refiere a las investigaciones sobre la aplicación de las técnicas de Machine Learning, y en particular del algoritmo K-NN, en el mercado financiero. En el tercer capítulo se introduce el paradigma actual para desarrollar la estrategia de inversión, así como, sus principales limitaciones. En el cuarto capítulo se expone cómo se ha construido el backtesting dinámico que permite simular diferentes escenarios de inversión. Posteriormente, en el quinto capítulo se incluyen diversas técnicas que, basadas en el análisis de importancia de características, permiten optimizar la estrategia diseñada. A continuación, en el sexto capítulo se establece una comparación entre los resultados finales obtenidos y los que se lograrían de aplicar una estrategia de comprar y mantener. Finalmente en el séptimo capítulo, se adjuntan las conclusiones y los futuros desarrollos necesarios para mejorar el método.

Capítulo 2

Estado del arte

La utilización de algoritmos para operar en el mercado financiero ha ido ganando protagonismo en la industria como resultado del desarrollo tecnológico. El trading algorítmico se ha convertido en un elemento básico del mercado financiero actual, puesto que la mayoría de las operaciones están completamente automatizadas. Los métodos derivados del Machine Learning (ML) han demostrado ser una fuerza a tener en consideración en el tratamiento de series temporales. De este modo, el objetivo fundamental de este capítulo es brindar una completa cobertura de las investigaciones desarrolladas en materia del ML, en el subdominio del algoritmo del K-NN, situando el foco en las estrategias cuantitativas de corto plazo y baja frecuencia.

Para ello, en primer lugar, se comenzará presentando los actuales paradigmas que permiten estudiar el dinamismo de los mercados financieros, así como, las dos corrientes de pensamientos desarrolladas para el diseño de las estrategias de inversión (análisis fundamental y técnico). A continuación, se analizarán las investigaciones desarrolladas en relación con el algoritmo K-NN en general, para, posteriormente, examinar la adaptación de este método al estudio de los mercados financieros. Muchas de las investigaciones revisadas sólo disponen de pruebas de concepto, con experimentos realizados en entornos poco realistas y sin ninguna aplicación directa en tiempo real. En las siguientes secciones se expondrán en detalle cada uno de los puntos anteriores, así como, las principales limitaciones de los estudios analizados.

2.1. Los mercados financieros

En esta sección se comenzará realizando una introducción al estudio de los mercados financieros. Posteriormente, se discutirá el término de la eficiencia del mercado cuyo pilar teórico divide el panorama bursátil en dos vertientes de pensamiento: los partidarios del análisis fundamental y los defensores del técnico; que serán analizados en los últimos apartados de esta sección.

2.1.1. Introducción

Los mercados financieros son un conjunto de operadores financieros y reglas de funcionamiento que se emplean para realizar la contratación de diversos activos financieros, fundamentalmente, dinero o renta fija y variable. Esta clase de activos pueden ser examinados como títulos-valores o derechos sobre bienes fácilmente convertibles en liquidez. Los mercados financieros se han expandido exponencialmente y ahora ofrecen multitud de variantes, entre ellas, los mercados de capital, derivados, criptomonedas y divisas. El enfoque principal de este trabajo se sitúa en el mercado de valores de renta variable. Estos son fórums donde sus participantes negocian el precio de las acciones, que son partes alícuotas del patrimonio de una organización. Las Bolsas de Valores forman parte de una de las arterias más importantes del corazón financiero, por ello, suponen un engranaje fundamental en la creación de un sistema financiero que refleje el ambiente donde operan los negocios. De este modo, proporcionan la estructura central y el mecanismo del sistema financiero, desempeñando dos funciones principales: servir de red del flujo de dinero y, al mismo tiempo, actuar como red de flujo de la información.

El precio de dichos títulos se negocia en las Bolsas que son lugares en los que compradores y vendedores se reúnen para pactar el valor de las acciones. Las operaciones financieras se basan en un paradigma similar al mercado de subastas, en el que un comprador potencial puja por el precio con el que desea entrar al mercado. El principio básico sobre el que se asienta el funcionamiento de este mercado es un concepto carente de dificultad alguna: comprar a precios bajos y vender a altos. Sin embargo, la complejidad del mercado de valores no debe de subestimarse nunca ya que es un reflejo de la economía y la política global. Los precios en el mercado son intrínsecamente dinámicos, volátiles y tienen gran sensibilidad. Ello ha propulsado el desarrollo de un amplio abanico de investigaciones que se centran en la predicción del precio de las acciones. Dichos estudios se pueden dividir en

dos vertientes fundamentales: aquellos que apoyan la Hipótesis de los Mercados Eficientes (EHM) y quienes la desafían.

2.1.2. Hipótesis de los Mercados Eficientes (EHM)

La eficiencia en los mercados financiero se puede definir según tres criterios:

1. Eficiencia asignativa: se estudia para analizar las consecuencias que generan las asignaciones de recursos al mercado de equilibrio en el bienestar de los seres humanos.

2. Eficiencia transaccional: hace referencia a los riesgos derivados de la participación de losagentes en los mercados financieros.

3. Eficiencia informativa: se encarga de estudiar la dependencia de los precios del mercado y su información.

La EHM pertenece al tercer tipo. En consecuencia, los mercados serán eficientes en tanto que los precios reflejen la información que se dispone. Fama (1970) [14] fue el pionero en acuñar este término y, posteriormente, Merton (1985) [30] fue quien desarrolló el marco teórico que sostiene dicha teoría. Dyckman y Morse (1986) [47] establecieron una definición en la que alegan que un mercado es eficiente si:

- Los precios de los valores negociados actúan reflejando completamente toda la información disponible.

- Esos precios reaccionan instantáneamente y de manera imparcial ante nueva información.

La dificultad en esta definición aparece cuando se pretende precisar y concretar qué se entiende por *reflejar completamente*, *información actual* y *desde un punto de vista imparcial*. Para ello, se enuncian las siguientes premisas a tener en consideración:

1. Competencia perfecta: ningún agente tiene suficiente poder como para provocar un *shock* en el mercado.

2. Mercado sin fricción: los activos son perfectamente divisibles, ya que no existen costes ni restricciones en el libre comercio.

3. Creencias homogéneas: como los agentes tienen la misma información, su racionalidad les lleva a poseer los mismos pensamientos.

4. Racionalidad individual: todos los que operan en los mercados tienen como objetivo maximizar su grado de utilidad.

Estas hipótesis tienen un conjunto de consecuencias que influencian las transacciones bursátiles:

- Gran parte de los agentes operan de forma completamente independiente.
- Todos los agentes del mercado actualizan con rapidez los precios con el fin de reflejar la información de última hora.
- La nueva información llega de forma completamente aleatoria e independiente.

De forma general, la EHM está sustentada en la teoría del paseo aleatorio, que establece que los precios de las acciones tienen a fluctuar de manera aleatoria. Por lo tanto, ello implica que los retornos de las mismas revelan un grado de correlación muy pobre. El pilar de dicha teoría se encuentra en el escrito de Cootner (1964) *El carácter aleatorio de los precios del mercado de valores* [10] donde establece las siguientes consideraciones:

- Estocasticidad: el precio está idénticamente distribuido y de manera independiente
- Varianza finita: el volumen y la varianza son finitos y, además poseen una importancia relativa.
- Correlación precio-valor: el precio refleja su valor fundamental.
- Expectativas: el rendimiento esperado es un promedio ponderado de la probabilidad invariante en el tiempo.
- Equilibrio general: el equilibrio de precios se obtiene cuando los agentes llegan a un mutuo acuerdo.
- Eficiencia informativa: los agentes negocian el precio según la información disponible
- Normalidad: la distribución de los cambios del precio sigue una normal de media estable y varianza finita.

Estas consideraciones tienen una serie de consecuencias prácticas. En primer lugar, cualquier tipo de patrón que sea detectado debe de ser eliminado por los operadores, para que no exista la tentación de predecir el futuro de la serie en base a información pasada. Esto es así, ya que toda la información relevante contenida en las series de precios pasadas ya se ha reflejado en el precio actual. No obstante, sí que existen anomalías y claros ciclos que no pueden ser considerados puramente aleatorios; por ejemplo, los efectos especiales del mes de enero, de los fines de semana o del lunes. De este modo, Sharpe (1964) [40], Lintner (1965) [25] y Mossin (1966) [31] notaron que era necesario considerar la racionalidad del agente en el modelo de los supuestos aleatorio. Dicha modificación fue autoría de Fama (1970) [14]

De este modo, la dificultad de establecer un modelo que se ajuste a los cambios y a la volatilidad del mercado es la piedra angular que supone el éxito de cualquier estrategia de inversión. Son precisamente la incertidumbre y la presunta ineficiencia de los mercados las que han dividido el panorama de los especuladores en dos principales corrientes de pensamiento: los partidarios del análisis fundamental y los del análisis técnico. El presente trabajo se ha desarrollado teniendo como referencia este último.

2.1.3. Análisis fundamental

El análisis fundamental emplea factores económicos para estimar el valor intrínseco del precio de una acción. Mediante la examinación de elementos como el resultado del ejercicio, los tipos de interés, los dividendos, la administración de la compañía y las rotaciones en el sector, enfoca las fuerzas de la oferta y la demanda para determinar el valor del mercado en cuestión. En general, el análisis fundamental tiene su piedra angular en la esencia del trabajo de Graham y en su libro titulado *"Inversor Inteligent"*, en el que formula las directrices de las recomendaciones de compraventa de acciones.

La decisión de inversión en este tipo de método se basa en la relatividad del valor intrínseco frente al precio: si dicho valor está por debajo del precio del mercado actual, entonces el mercado se encuentra sobrevalorado y, por lo tanto, se debería proceder con la venta del título. Para proceder con el cálculo de dicho valor se analizan variables tanto microeconómicas como macroeconómicas. Esta distinción es la que da lugar a dos enfoques diferentes:

- Método Top-Down: partiendo de lo general hasta lo particular, estudia pri-

mero las variables macroeconómicas y después las microeconómicas. Por ejemplo, se comienza analizando la situación económica mundial, posteriormente se estudian los países con mayor potencial atractivo para invertir y, finalmente, decide a qué sector asignar mayor cantidad de recursos.

- Método Bottom-Up: partiendo de lo particular a lo general, analiza en primer lugar las variables internas de la empresa y luego se centra en las variables globales. Por ejemplo, se empieza estudiando las empresas que tengan mayor potencial de crecimiento, posteriormente, se analizan los sectores en los que operan dichas compañías y, finalmente, se examina la situación económica del país donde cotizan.

Independientemente del tipo de enfoque que adopte el analista, existen diferentes métodos para valorar una empresa. Los métodos conceptualmente correctos para valorar las empresas con expectativas de continuidad son aquellos basados en el descuento de flujos. Es decir, aquellos que conciben a las compañías como entes generadores de flujos de fondo, y por ello, sus acciones y deudas son variables perfectamente tasables como el resto de los activos financieros que poseen en sus balance. La mayoría de los investigadores en este ámbito (Fisher & Jordan (1995) [12], Bernard (1994) [6], Penman (1991) [38] y Brown (2002) [22]) defienden que el valor de una determinada empresa reside esencialmente en su capacidad futura para generar flujos de caja. No obstante, también son comúnmente empleadas las técnicas basadas en valoraciones relativas mediante la construcción de ratios procedentes de la cuenta de resultados de las empresas.

En los métodos de descuentos de flujos se realiza un pronóstico detallado y cuidadoso de cada una de las partidas financieras vinculadas a la generación de flujos de cajas correspondientes a las operaciones de la empresa. Fernández (2008) [15] establece que la pieza clave de estas valoraciones es la tasa de descuento empleada que, depende fundamentalmente del riesgo y las volatilidades históricas. Por otra parte, también se puede obtener una estimación del valor de una empresa tomando como referencia no sólo los flujos de caja que genera, sino también los dividendos que reparte. Este método se conoce como la técnica de los dividendos descontados. Así, Bodie et al. (2009) [7] establecen que el valor real de una acción se computa considerando el rendimiento que se espera obtener mediante el usufructo de las mismas.

En los métodos basados en las valoraciones relativas se estima el valor intrínseco de una compañía en base a ratios comparativos con otras empresas. Algunos de

los más comunes son el PER (*Price earnings ratio*, EV-EBITDA (*Entreprise value - Beneficio antes de amortización, intereses e impuestos*) y EV-CF (*Entreprise value - Cash flows*).

2.1.4. Análisis técnico

El análisis técnico se concibe como el conjunto de reglas que permiten predecir la rentabilidad de una acción mediante el estudio exclusivo de datos históricos, pues considera que la propia acción lo descuenta todo. Se apoya en la psicología de masas alegando que la naturaleza humana tiende al equilibrio y que, por ello, la historia sigue tendencias que se repiten. Desde la primera publicación de Robert Edwards y el clásico *"Análisis técnico de las tendencias de acciones"* de John Magee (1948) [21], el análisis técnico ha pasado de considerarse una práctica arcana a un método globalmente aceptado. Posteriormente fue John Murphy (1999) [32] quien brindó una cobertura completa de los conceptos y técnicas disponibles.

Según los analistas técnicos los precios del mercado reflejan no sólo los distintos miedos, conjeturas y estados de ánimo, racionales e irracionales, de los cientos de compradores y vendedores, sino que también reflejan sus necesidades y recursos. Estos son factores adicionales que retan al análisis y para los que no hay estadísticas que se puedan obtener. Todos ellos están, sin embargo, sintetizados, ponderados y finalmente expresados en una cifra precisa a la que un comprador y un vendedor llegan juntos, y hacen un trato. El precio resultante es la única cifra que importa.

La piedra angular de dicho análisis la constituye los manuscritos desarrollados por Charles Dow a finales del siglo XIX. Dicha teoría se fundamenta en seis principios fundamentales:

1. El mercado lo descuenta todo: la teoría de Dow se basa en la Hipótesis de los Mercados Eficientes, según la cual los precios de los activos incorporan toda la información disponible.

2. El mercado sigue tres tendencias principales: los mercados experimentan tendencias primarias que duran un año o más, como un mercado alcista o bajista. Dentro de estas tendencias más amplias, experimentan tendencias secundarias, que a menudo van en contra de la tendencia primaria, como un retroceso dentro de un mercado alcista o un repunte dentro de un mercado bajista;

estas tendencias secundarias duran de tres semanas a tres meses. Por último, hay tendencias menores que duran menos de tres semanas y que son en gran medida ruido.

3. Las tendencias primarias tienen tres fases: según la teoría de Dow, una tendencia primaria pasa por tres fases. En un mercado alcista, son la fase de acumulación, de participación del público (o gran movimiento) y de exceso. En un mercado bajista, se denominan fase de distribución, de participación del público y de pánico (o desesperación).

4. Los índices Dow Jones Industrial y de Transporte deben confirmarse: para que se establezca una tendencia, los índices o medias de mercado postulados por Dow deben confirmarse mutuamente. Esto significa que las señales que se producen en un índice deben coincidir o corresponderse con las señales del otro. Si un índice, como el Promedio Industrial Dow Jones (DJIA), está confirmando una nueva tendencia primaria alcista, pero el otro índice (Promedio de Transportes Dow Jones (DJTA)) permanece en una tendencia primaria bajista, los operadores no deben asumir que se ha iniciado una nueva tendencia.

5. El volumen debe confirmar la tendencia: el volumen debe aumentar si el precio se mueve en la dirección de la tendencia principal y disminuir si se mueve en contra de ella. Un volumen bajo indica una debilidad en la tendencia. Por ejemplo, en un mercado alcista, el volumen debería aumentar cuando el precio está subiendo, y caer durante los retrocesos secundarios. Si en este ejemplo el volumen aumenta durante un retroceso, podría ser una señal de que la tendencia se está invirtiendo a medida que más participantes del mercado se vuelven bajistas.

6. Las tendencias persisten hasta que se demuestre lo contrario: las reversiones de las tendencias primarias pueden confundirse con las tendencias secundarias. Es difícil determinar si una subida en un mercado bajista es una inversión o un repunte de corta duración al que seguirán mínimos aún más bajos. En este escenario la teoría de Dow aboga por la cautela, insistiendo en que se confirme una posible inversión.

En los últimos años se han publicado numerosos estudios en prestigiosas revistas científicas cuya trama principal la ocupa el estudio del análisis técnico. Para validar quienes defienden dicho método Park e Irwin (2004) [37] diseñaron un análisis íntegro con el objetivo de investigar 137 estudios realizados sobre diferentes

técnicas de inversión. Entre las principales herramientas operativas destacan el empleo de indicadores técnicos: medias móviles (MA), índice de fuerza relativa (RSI) y el MACD; modelos estadísticos y de machine learning, y gráficos chartistas. Los autores dividieron el espectro total en dos espacios temporales: los estudios tempranos (1960-1987) y los estudios modernos (1988-2004); y abordaron en cada uno de ellos las principales limitaciones de las investigaciones desarrolladas. De forma general, probaron que el éxito del análisis técnico es relativo pues depende fuertemente del espacio temporal y del mercado en el que se desarrolle. Tras realizar una revisión exhaustiva y sistemática de la literatura existente, enunciaron los principales retos a los que se enfrentan los operadores cuando se trata de operaciones financieras:

- La frecuencia de adquisición de los datos, así como la rapidez con la que estos llegan, y el tiempo que el algoritmo tarda en incorporarlos al proceso de decisión para ejecutar una acción óptima.
- La ineficiencia del mercado. En el sentido en que los precios pueden estar afectados por *shocks* transitorios que no tienen que ver con su valoración.
- Por último, y más importante, la optimización en base a sucesos del pasado. De forma muy intuitiva, sólo se pueden optimizar los modelos tomando por referencia sucesos que han ocurrido y de los que se disponen datos. Se puede diseñar que el algoritmo sepa que cuando se da X situación lo óptimo es tomar Y decisión. Pero si el algoritmo nunca se ha enfrentado a una situación X, entonces no puede aprender del pasado en un sentido estricto. Esto refleja el clásico problema *in-sample vs. out-of-sample*.

En relación con esta última cláusula, los campos que están cobrando especial protagonismo en las investigaciones más recientes son la aplicación de las técnicas de Machine Learning y Deep Learning en la predicción de las series temporales financieras. Tradicionalmente los procedimientos utilizados para analizar los mercados de valores se basaban en métodos estadísticos como el uso de medias móviles, los filtros de Kalman, el suavizado exponencial, la media móvil autorregresiva (ARMA) y la media móvil integrada autorregresiva (ARIMA). Sin embargo, en los últimos años se ha demostrado que no funcionan del todo bien a la hora de predecir los índices bursátiles con precisión. En contraste con las técnicas estadísticas, los métodos de Inteligencia Artificial pueden manejar los datos aleatorios, caóticos y no lineales del mercado de valores y se han utilizado ampliamente para la predicción exacta de los índices bursátiles. Estos métodos buscan extraer patrones aprendidos de datos históricos en un proceso conocido como entrenamiento o

aprendizaje para, posteriormente, aplicarlos sobre nuevos datos. Los estudios empíricos que utilizan el aprendizaje automático suelen tener dos fases principales. La primera, aborda la selección de las variables y los modelos relevantes para la predicción, separando una parte de los datos para el entrenamiento y otra para la validación de los modelos. La segunda fase aplica los modelos optimizados a los datos de test midiendo así el rendimiento predictivo. Las técnicas más usadas en la literatura son las redes neuronales artificiales (ANN), las máquinas de soporte de vectores (SVM) y los bosques aleatorios o *random forests* (RF).

En general, las redes neuronales modelan procesos biológicos, concretamente las conexiones que se establecen en las neuronas del cerebro humano. La unidad básica de estas redes, la neurona, emula el cerebro humano compuesto por las dendritas que reciben las variables de entrada y emiten un valor de salida que puede servir de entrada para otras neuronas. Las capas de procesamiento de las redes están interconectadas, y se atribuyen pesos para cada conexión que se ajustan en el proceso de aprendizaje de la red, en la primera fase de mencionada anteriormente. Esta etapa no sólo optimiza las interconexiones entre las capas de neuronas, sino que también optimiza los parámetros de las funciones de transferencia entre una capa y otra minimizando, así, los errores cometidos. Finalmente, la última capa de la red neuronal se encarga de sumar todas las señales de la capa anterior en una sola variable de salida: la respuesta de la red a determinados datos de entrada.

Mientras que las redes neuronales tratan de minimizar los errores de sus respuestas empíricas en la etapa de entrenamiento, el algoritmo SVM busca minimizar el umbral superior del error de sus clasificaciones. Para ello, toma las muestras de entrenamiento y las transforma desde su espacio de dimensión original, a otro con un número mayor de dimensiones, en el que se aproxima una separación lineal mediante un hiperplano. Este algoritmo, que se suele utilizar para clasificar los datos en función de las variables de entrada, busca minimizar el margen de dicho hiperplano durante la etapa de entrenamiento del modelo. La transformación del espacio de dimensiones originales al espacio en en el que se realizan las clasificaciones se realiza con la ayuda de funciones kernel, a partir de la parametrización estimada en el entrenamiento del modelo.

Al igual que las ANNs y SVMs, los árboles de decisión se utilizan a menudo en la literatura de Machine Learning. Este método implica la subdivisión de los datos en subconjuntos separados por los valores de las variables de entrada hasta obtener la unidad básica de clasificación, de acuerdo con las muestras de entrenamiento.

Las clasificaciones consensuadas de los árboles más precisos se combinan en una sola, que comprende el algoritmo de *random forest*. La combinación de árboles de decisión en la técnica RF puede utilizarse en regresiones o clasificaciones, y genera resultados muy satisfactorios en el ámbito de la predicción de los mercados financieros.

Sin embargo, a pesar de ser los algoritmos más utilizados en la práctica en el ámbito financiero tienen una serie de limitaciones. La primera de ellas es su interpretabilidad al considerarse como *cajas negras*. La segunda de ellas, y más importante en este trabajo, es su complejidad en términos computacionales. Esta es una de las razones por la que se ha decidido emplear el algoritmo K-NN en su lugar. Este algoritmo brinda un excelente equilibrio entre la precisión en sus soluciones, su interpretabilidad y su coste computacional.

2.2. Algoritmo K-NN

El objetivo fundamental de esta sección es formular la teoría del algoritmo K-NN. Para ello, en primer lugar, se expondrá mediante un ejemplo cuáles son las fases seguidas por el modelo a la hora de establecer una predicción. A continuación, se enunciarán sus principales limitaciones y, posteriormente, se explicarán los avances recabados en la literatura existente para hacer frente a estas flaquezas. Por último, se analizarán las diferentes variantes del modelo original propuestas en las principales investigaciones académicas.

2.2.1. Introducción

El algoritmo de K-Nearest Neighbors (K-NN) es uno de los métodos más antiguo, sencillo y preciso para los modelos de clasificación y regresión de patrones. El K-NN fue propuesto por Fix & Hodges (1951) [17] y modificado por Cover & Hart (1967) [11]. El K-NN ha sido reconocido como uno de los diez mejores métodos en el *data mining* según Wu et al. (2008) [50]. En consecuencia, ha sido estudiado en la últimas décadas y ampliamente aplicado en muchos campos. Así, se trata del algoritmo de referencia en muchos problemas como reconocimiento de patrones [51], categorización de textos [28], modelos de clasificación [18], reconocimiento de objetos [5] y detección de anomalías [41].

Se considera un método de aprendizaje supervisado no paramétrico [1]. Asimismo, se etiqueta como un *lazy learner*, ya que utiliza todos los datos del entrenamiento para realizar la predicción de un nuevo valor sin tener que crear un modelo de aprendizaje. Dadas algunas características (variables explicativas) de una nueva instancia que debe ser predicha, el K-NN encuentra las K instancias de entrenamiento que más se acercan a la nueva según una métrica de distancia en concreto. A pesar de que se puede utilizar para resolver problemas tanto de regresión como de clasificación, la mayoría de los estudios lo emplean como clasificador [45].

Para ilustrar el mecanismo del algoritmo se van a explicar cada una de las fases del mismo mediante un ejemplo. El K-NN utiliza un término denominado *similitud de características* para predecir los valores de cualquier punto de datos nuevos. Esto significa que a dicho punto se le asigna un valor basado en su parecido con los puntos del conjunto de entrenamiento. Considérese la Tabla 2.1, constituida por la altura, edad y peso de una serie de individuos. El objetivo consiste en predecir el peso de una persona determinada a partir de las características restantes.

ID	Height	Age	Weight
1	5	45	77
2	5,11	26	47
3	5,6	30	55
4	5,9	34	59
5	4,8	40	72
6	5,8	36	60
7	5,3	19	40
8	5,8	28	60
9	5,5	23	45
10	5,6	32	58
11	5,5	38	?

Tabla 2.1. *Dataset de ejemplo para la explicación del K-NN*

Para un mejor entendimiento del problema, se adjunta en la Figura 2.1 la representación de la altura frente a la edad. En el gráfico anterior, el eje y representa la altura de una persona (en pies) y el eje x representa la edad (en años). Los puntos

[1] No paramétrico significa que no hay parámetros o número fijo de parámetros independientemente del tamaño de los datos. En cambio, los parámetros estarían determinados por el tamaño del conjunto de datos de entrenamiento. Por lo tanto, no hay suposiciones que deban hacerse sobre la distribución de los datos subyacentes.

están numerados según los valores de ID. El punto amarillo (ID11) es el valor a predecir.

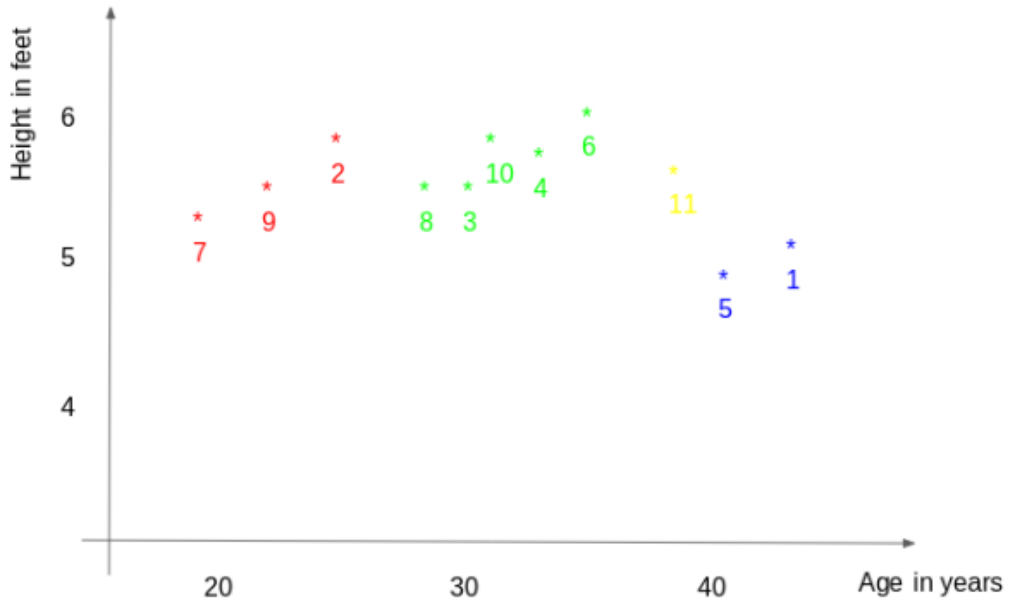


Figura 2.1. *Representación de la altura frente a la edad*

Teniendo en cuenta la definición establecida anteriormente, el K-NN asignaría el peso del ID11 teniendo en cuenta principalmente los valores del ID1 e ID5. A continuación, se adjunta una explicación paso a paso de la ejecución del algoritmo.

1. En primer lugar, se computa la distancia entre el valor a predecir y el conjunto de datos de entrenamiento disponible, tal y como se puede observar en la Figura 2.2.

2. En segundo lugar, se seleccionan los K puntos más cercanos basados en esa métrica de distancia. En este ejemplo, como se ilustra en la Figura 2.3, los puntos 1, 5 y 6 son los seleccionados si K es igual a 3.

3. Finalmente, se computa la media de dichos valores como la predicción final.

$$\text{ID } 11 = (77 + 72 + 60)/3 = 69{,}66kg$$

A continuación, se adjunta el pseudocódigo del algoritmo.

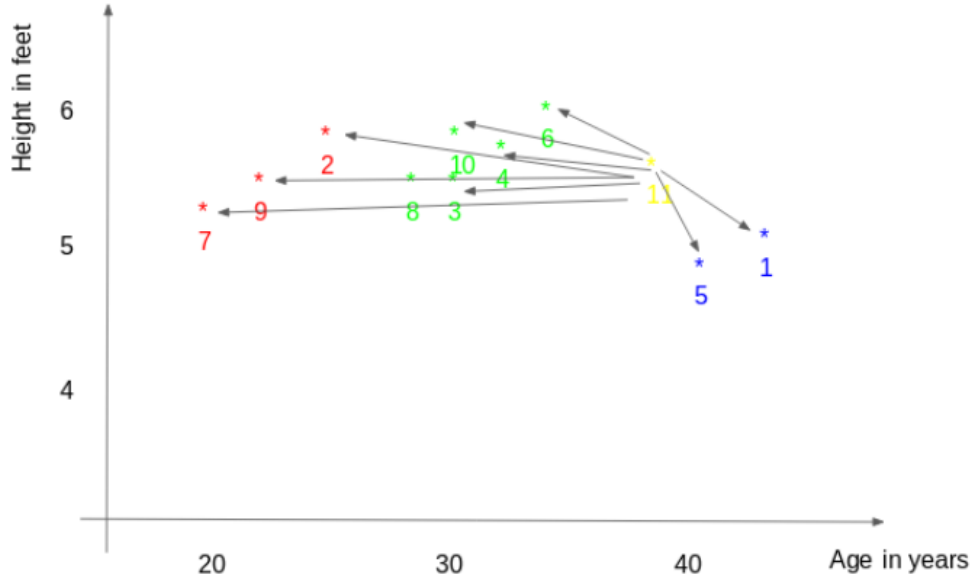


Figura 2.2. *Cálculo de los pares de distancia*

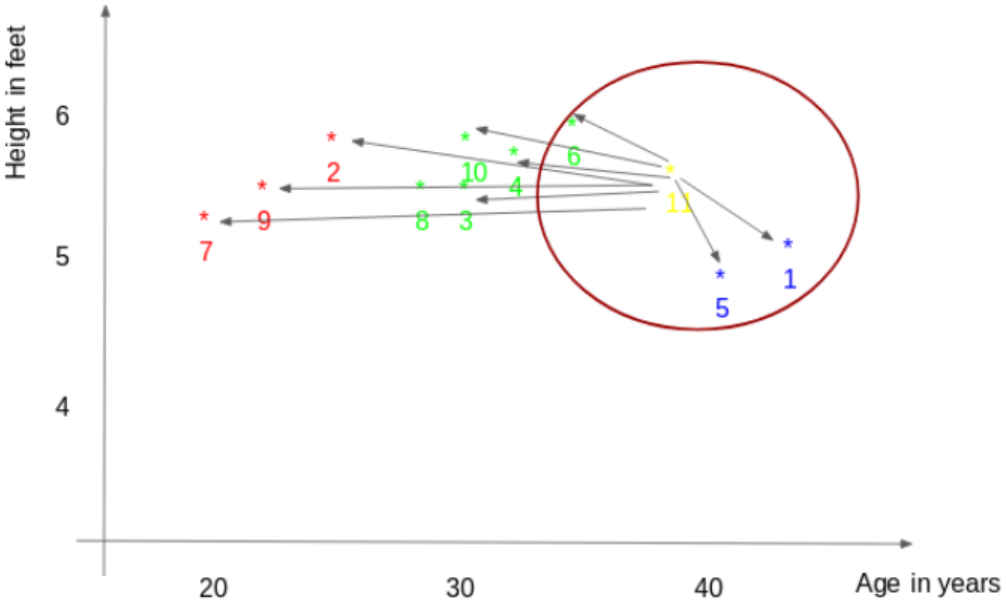


Figura 2.3. *Selección de los K (3) valores más cercanos*

Algorithm 1: Basic KNN algorithm
Input: Training samples D, Test sample d, K
Output: Class label of test sample
1: Compute the distance between d and every sample in D
2: Choose the K samples in D that are nearest to d; denote the set by P($\in$ D)
3: Assign d the mean of the K nearest samples.

Aunque el algoritmo del K-NN ha sido ampliamente utilizado durante décadas debido a su simplicidad, eficacia y robustez, existen cuatro deficiencias principales a las que se enfrenta:

1. La función de distancia para medir la diferencia o similitud entre dos instancias. Normalmente, se opta por emplear la distancia euclídea. En este caso se trata de un algoritmo con un alto coste computacional, ya que tiene que computar la distancia entre cada muestra de prueba y todas las muestras de entrenamiento. Es decir, para cada ejemplo del test se necesita una complejidad de tiempo $O(nm)$, donde n es el número de ejemplos en los datos de entrenamiento, y m es el número de características de cada ejemplo [2].

2. El tamaño de vecindad (K) que se asigna, que frecuentemente suele ser un valor fijo.

3. La estimación de la predicción en función de dichas distancias. Habitualmente se computa como la media de la variable explicativa de los K vecinos más cercanos.

4. La complejidad del espacio es $O(nm)$, ya que al ser un *lazy learner* tiene que guardar en memoria todas las muestras del entrenamiento.

El objetivo de los siguientes apartados es revisar los avances disponibles en la literatura para hacer frente a las limitaciones anteriores.

2.2.2. Revisión de las funciones de distancia

Para ello, Surya et al. (2019) [39] evalúan el rendimiento del K-NN utilizando siete familias de distancia diferentes sobre distintos *datasets* reales. Para ello, explican cada una de las funciones utilizando como ejemplo la distancia entre los

[2]La complejidad de tiempo hace referencia al número de operaciones necesarias a ejecutar por el algoritmo.

siguientes vectores $v1 = \{5{,}1, 3{,}5, 1{,}4, 0{,}3\}$, $v2 = \{5{,}4, 3{,}4, 1{,}7, 0{,}2\}$. A continuación, se describen cada una de las familias seleccionadas y se presentan las clases de métricas dentro de ellas.

1. **Lp Minkowski**: esta familia de distancias incluye tres métricas de distancia, que son casos especiales de la distancia de Minkowski y corresponden a diferentes valores de p para esta potencia. Esta distancia, que también se conoce como norma Lp, es una métrica generalizada que se define como:

$$D_{\text{Mink}}(x, y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \tag{2.1}$$

   donde p es un valor positivo. Cuando $p = 2$ la distancia se convierte en la euclídea. Cuando $p = 1$, se trata de la distancia de Manhattan. A su vez, la distancia de Chebyshev es una variante de la de Minkowski donde $p = \infty$. A continuación, se adjunta en la Tabla 2.2 cada una de las particularizaciones de la ecuación anterior.

Abrev.	**Nombre**	**Definición**	**Resultado**
MD	Manhattan	$\sum_{i=1}^{n} \|x_i - y_i\|$	0,8
CD	Chebyshev	$\text{máx}_i \|x_i - y_i\|$	0,3
ED	Euclidean	$\sqrt{\sum_{i=1}^{n} \|x_i - y_i\|^2}$	0,4472

Tabla 2.2. *Lp métricas de distancia Minkowski*

2. **Distancias** L_1: esta familia de distancias depende principalmente de encontrar la diferencia absoluta. Dicho criterio incluye las distancias Lorentziana, Canberra, Sorensen, Soergel, Kulczynski, Carácter Medio, No Intersección. A continuación, se incluyen en la Tabla 2.3 las ecuaciones para su cálculo.

3. **Medidas de distancia del producto interior**: las medidas de distancia pertenecientes a esta familia se calculan mediante el producto de pares de vectores. Este tipo de distancias incluye: Jaccard, Coseno, Dados, Acordes, tal y como se puede observar en la Tabla 2.4.

4. **Medidas de distancias al cuadrado**: las distancias que pertenecen a esta familia se obtienen calculando la suma de las geométricas. La media geométrica de dos valores es la raíz cuadrada de su producto. Las distancias en

Abrev.	Nombre	Definición	Resultado
LD	Lorentzian	$\sum_{i=1}^{n} \ln(1 + \lvert x_i - y_i \rvert)$	0,7153
CanD	Canberra	$\sum_{i=1}^{n} \frac{\lvert x_i - y_i \rvert}{\lvert x_i \rvert + \lvert y_i \rvert}$	0,0381
SD	Sorensen	$\frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{\sum_{i=1}^{n} (x_i + y_i)}$	0,0381
SoD	Soergel	$\frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{\sum_{i=1}^{n} \max(x_i, y_i)}$	0,0734
KD	Kulczynski	$\frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{\sum_{i=1}^{n} \min(x_i, y_i)}$	0,0792
MCD	Mean Character	$\frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{n}$	0,2
NID	Non Intersection	$\frac{1}{2} \sum_{i=1}^{n} \lvert x_i - y_i \rvert$	0,4

Tabla 2.3. *L_1 métricas de distancia*

Abrev.	Nombre	Definición	Resultado
JacD	Jaccard	$\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} x_i y_i}$	0,0048
CosD	Cosine	$1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$	0,0016
DicD	Dice	$1 - \frac{2 \sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2}$	0,9524
ChoD	Chord	$\sqrt{2 - 2 \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$	0,0564

Tabla 2.4. *Métricas de distancia del producto interior*

esta familia no pueden utilizarse con vectores de características de valores negativos. Esta familia engloba a las distancias de Bhattachayya, Squared Chord, Matusita, distancias de Hellinger, adjuntas en la Tabla 2.5.

Abrev.	Nombre	Definición	Resultado
BD	Bhattacharyya	$-\ln \sum_{i=1}^{n} \sqrt{x_i y_i}$	−2,34996
SCD	Squared Chord	$\sum_{i=1}^{n} \left(\sqrt{x_i} - \sqrt{y_i}\right)^2$	0,0297
MatD	Matusita	$\sqrt{\sum_{i=1}^{n} \left(\sqrt{x_i} - \sqrt{y_i}\right)^2}$	0,1722
HeD	Hellinger	$\sqrt{2 \sum_{i=1}^{n} \left(\sqrt{x_i} - \sqrt{y_i}\right)^2}$	0,2436

Tabla 2.5. *Métricas de distancias al cuadrado*

5. **Medidas de distancias L_2 al cuadrado**: en esta familia, el cuadrado de la diferencia en cada punto de ambos vectores se considera la distancia total.

En esta categoría se incluyen la distancia euclídea al cuadrado, Clark, Neyman, χ^2, Pearson χ^2, χ^2 al cuadrado, probabilístico simétrico χ^2, divergencia, simétrica aditiva χ^2, media, promedio y distancias euclídea censuradas. En la Tabla 2.6 se incluyen sus fórmulas matemáticas.

Abrev.	**Nombre**	**Definición**	**Resultado**
SED	Squared Euclidean	$\sum_{i=1}^{n} (x_i - y_i)^2$	0,2
ClaD	Clark	$\sqrt{\sum_{i=1}^{n} \left(\frac{\lvert x_i - y_i \rvert}{x_i + y_i}\right)^2}$	0,2245
NCSD	Neyman χ^2	$\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i}$	0,1181
PCSD	Pearson χ^2	$\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{y_i}$	0,1225
SquD	Squared χ^2	$\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i}$	0,0591
PSCSD	Probabilistic Symmetric χ^2	$2\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i}$	0,1182
DivD	Divergence	$2\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{(x_i + y_i)^2}$	0,1008
ASCSD	Additive Symmetric χ^2	$2\sum_{i=1}^{n} \frac{(x_i - y_i)^2 (x_i + y_i)}{x_i y_i}$	0,8054
AD	Average	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - y_i)^2}$	0,2236
MCED	Mean Censored Euclidean	$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$	0,2236
SCSD	Squared Chi-Squared	$\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\lvert x_i + y_i} \rvert$	0,0591

Tabla 2.6. *Métricas de distancia L_2 al cuadrado*

6. **Medidas de distancia de entropía de Shannon**: esta categoría, que incluyen las distancias de Kullback-Leibler, Jeffreys, la divergencia K, Topsoe, JensenShannon y la diferencia de Jensen, están relacionadas con la entropía de Shannon [3]. A continuación, se adjunta en la Tabla 2.7 la definición matemática de dichas métricas.

7. **Medidas de distancia de vicisitud**: la familia de distancias de vicisitudes consta de cuatro distancias: Vicis-Wave Hedges, Vicis Simétrica, Max Simétrica χ^2 y Min Symmetric χ^2. Estas distancias se generaron a partir de relación sintáctica para las medidas de distancia mencionadas. La Tabla 2.8 incluye la definición de las mismas.

[3] En la teoría de la información la entropía mide la incertidumbre de dicha fuente.

Abrev.	Nombre	Definición	Resultado
KLD	Kullback-Leibler	$\sum_{i=1}^{n} x_i \ln \frac{x_i}{y_i}$	−0,3402
JefD	Jeffreys	$\sum_{i=1}^{n} (x_i - y_i) \ln \frac{x_i}{y_i}$	0,1184
KDD	K divergence	$\sum_{i=1}^{n} x_i \ln \frac{2x_i}{x_i+y_i}$	−0,1853
TopD	Topsoe	$\sum_{i=1}^{n} x_i \ln \left(\frac{2x_i}{x_i+y_i}\right) + \sum_{i=1}^{n} y_i \ln \left(\frac{2y_i}{x_i+y_i}\right)$	0,0323
JSD	Jensen-Shannon	$\frac{1}{2}\left[\sum_{i=1}^{n} x_i \ln \frac{2x_i}{x_i+y_i} + \sum_{i=1}^{n} y_i \ln \frac{2y_i}{x_i+y_i}\right]$	0,014809
JDD	Jensen difference	$\frac{1}{2}\left[\sum_{i=1}^{n} \frac{x_i \ln x_i + y_i \ln y_i}{2} - \left(\frac{x_i+y_i}{2}\right) \ln \left(\frac{x_i+y_i}{2}\right)\right]$	0,0074

Tabla 2.7. *Medidas de distancia de entropía de Shannon*

Abrev.	Nombre	Definición	Resultado
VWHD	Vicis-Wave Hedges	$\sum_{i=1}^{n} \frac{\lvert x_i - y_i \rvert}{\min(x_i, y_i)}$	0,8025
VSDF1	Vicis Symmetric1	$\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\min(x_i, y_i)^2}$	0,3002
VSDF2	Vicis Symmetric2	$\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\min(x_i, y_i)}$	0,1349
VSDF3	Vicis Symmetric3	$\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\max(x_i, y_i)}$	0,1058
MSCD	Max Symmetric χ^2	$\max\left(\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i}, \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{y_i}\right)$	0,1225
MiSCSD	Min Symmetric χ^2	$\min\left(\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i}, \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{y_i}\right)$	0,1181

Tabla 2.8. *Medidas de distancia de vicisitud*

A continuación, se adjunta la Tabla 2.9 donde se describen los *datasets* empleados por los autores para realizar el estudio. La columna $\#E$ hace referencia al número de muestras, $\#F$ el número de características y $\#C$ el número de clases.

A continuación, se adjunta en la Tabla 2.10 los *scores* medios conseguidos por cada dataset.

En vista a los resultados obtenidos, los autores establecen una serie de conclusiones:

1. El rendimiento del K-NN depende significativamente de la distancia utilizada, puesto que los resultados mostraron grandes diferencias entre los valores de las diferentes distancias. No obstante, se obtiene que la distancia *Hassa-*

Name	#E	# F	#C	Data type	Min	Max
Heart	270	25	2	real& integer	0	564
Balance	625	4	3	positive integer	1	5
Cancer	683	9	2	positive integer	0	9
German	1000	24	2	positive integer	0	184
Liver	345	6	2	real& integer	0	297
Vehicle	846	18	4	positive integer	0	1018
Vote	399	10	2	positive integer	0	2
BCW	699	10	2	positive integer	1	13454352
Haberman	306	3	2	positive integer	0	83
Letter rec.	20000	16	26	positive integer	0	15
Wholesale	440	7	2	positive integer	1	112151
Australian	690	42	2	positive real	0	100001
Glass	214	9	6	positive real	0	75,41
Sonar	208	60	2	positive real	0	1
Wine	178	13	3	positive real	0,13	1680
EEG	14980	14	2	positive real	86,67	715897
Parkinson	1040	27	2	positive real	0	1490
Iris	150	4	3	positive real	0,1	7,9
Diabetes	768	8	2	real& integer	0	846
Monkey1	556	17	2	binary	0	1
Ionosphere	351	34	2	real	−1	1
Phoneme	5404	5	2	real	−1,82	4,38
Segmen	2310	19	7	real	−50	1386,33
Vowel	528	10	11	real	−5,21	5,07
Wave21	5000	21	3	real	−4,2	9,06
Wave40	5000	40	3	real	−3,97	8,82
Banknote	1372	4	2	real	−13,77	17,93
QSAR	1055	41	2	real	−5,256	147

Tabla 2.9. *Datasets empleados por [39] para la medir la sensibilidad de la función de distancias*

nat [4] es la mejor cuando se aplica a la mayoría de los conjuntos de datos en comparación con las demás distancias probadas.

[4]
$$\mathrm{HasD}(x,y) = \sum_{i=1}^{n} D\left(x_i, y_i\right)$$

Distance	Accuracy	Recall	Precision	Distance	Accuracy	Recall	Precision
ED	0,8001	0,6749	0,6724	PSCSD	0,6821	0,5528	0,5504
MD	0,8113	0,6831	0,681	DivD	0,812	0,678	0,6768
CD	0,7708	0,656	0,6467	ClaD	0,8227	0,6892	0,6871
LD	0,8316	0,6964	0,6934	ASCSD	0,6259	0,4814	0,4861
CanD	0,8282	0,6932	0,6916	SED	0,8001	0,6749	0,6724
SD	0,7407	0,6141	0,6152	AD	0,8001	0,6749	0,6724
SoD	0,7881	0,6651	0,6587	SCSD	0,8275	0,693	0,6909
KD	0,6657	0,5369	0,5325	MCED	0,7973	0,6735	0,6704
MCD	0,8113	0,6831	0,681	TopD	0,6793	0,461	0,4879
NID	0,8113	0,6831	0,681	JSD	0,6793	0,461	0,4879
CosD	0,803	0,6735	0,6693	JDD	0,7482	0,5543	0,5676
ChoD	0,7984	0,662	0,6647	JefD	0,7951	0,6404	0,6251
JacD	0,8024	0,6756	0,6739	KLD	0,513	0,3456	0,3808
DicD	0,8024	0,6756	0,6739	KDD	0,5375	0,3863	0,4182
SCD	0,8164	0,65	0,4813	KJD	0,6501	0,4984	0,5222
HeD	0,8164	0,65	0,6143	TanD	0,7496	0,5553	0,5718
BD	0,4875	0,3283	0,4855	AvgD	0,8084	0,6811	0,6794
MatD	0,8164	0,65	0,5799	HamD	0,6413	0,5407	0,5348
VWHD	0,6174	0,4772	0,5871	MeeD	0,415	0,1605	0,3333
VSDF1	0,7514	0,6043	0,5125	WIAD	0,812	0,6815	0,6804
VSDF2	0,6226	0,4828	0,5621	HauD	0,5967	0,4793	0,4871
VSDF3	0,7084	0,5791	0,5621	CSSD	0,4397	0,2538	0,332
MSCSD	0,7224	0,5876	0,3769	SPeaD	0,8023	0,6711	0,6685
MiSCSD	0,6475	0,5137	0,5621	CorD	0,803	0,6717	0,6692
PCSD	0,6946	0,57	0,5621	PeaD	0,759	0,6546	0,6395
NCSD	0,6946	0,57	0,5621	MotD	0,7407	0,6141	0,6152
SquD	0,6821	0,5528	0,5504	HasD	0,8394	0,7018	0,701

Tabla 2.10. *Promedio del accuracy, recall y precision sobre los datasets [39]*

2. Se obtienen resultados muy similares cuando se emplean distancias de la misma familia. Por lo tanto, en estos casos, y dado que el K-NN compara los ejemplos utilizando la misma distancia, los vecinos más cercanos serán iguales

donde

$$D(x,y) = \begin{cases} 1 - \frac{1+\text{mín}(x_i,y_i)}{1+\text{máx}(x_i,y_i)}, & \text{mín}\,(x_i,y_i) \geq 0 \\ 1 - \frac{1+\text{mín}(x_i,y_i)+|\text{mín}(x_i,y_i)|}{1+\text{máx}(x_i,y_i)+|\text{mín}(x_i,y_i)|}, & \text{mín}\,(x_i,y_i) < 0 \end{cases}$$

si todas las distancias se multiplican o se dividen por la misma constante.

3. No existe una métrica de distancia óptima que pueda utilizarse para todos los tipos de conjuntos de datos, ya que los resultados muestran que cada conjunto de datos favorece una métrica de distancia específica, y este resultado cumple con el teorema del no-free-lunch [5].

4. El rendimiento (medido por el accuracy, la precisión y el recall) del K-NN sólo se degrada en un 20 % cuando el de ruido alcanza el 90 %. Esto pone de manifiesto que el K-NN se trata de un algoritmo que tolera de forma satisfactoria las perturbaciones.

5. Algunas distancias se ven menos afectadas por el ruido añadido en comparación con otras distancias, por ejemplo, se aprecia cómo la distancia *Hassanat* es la que mejor funciona cuando se aplica a la mayoría de los conjuntos de datos con diferentes niveles de ruido.

2.2.3. Revisión del valor del número de vecinos K

El K-NN es un algoritmo relacionado con el valor de K. De tal modo, su precisión de clasificación es extremadamente sensible a dicho valor. Por lo tanto, una idea natural es diseñar algún método de aprendizaje para buscar automáticamente el mejor valor de K según cada caso. En la práctica dicho parámetro se suele ajustar mediante el *cross-validation*. No obstante, en la literatura existen otros enfoques más singulares para abordar este problema.

En primer lugar, cabe mencionar el estudio realizado por Wang et al. (2005) [48] en el que proponen un nuevo método para la selección del número de vecinos que se basa en el concepto de confianza estadística. En primer lugar, definen la confianza asociada a una decisión, tomada por la regla de la mayoría, a partir de un número finito de observaciones y la utilizan como criterio para determinar el valor de K. Además, el método ajusta localmente el número de vecinos más cercanos hasta llegar a un nivel satisfactorio. De este modo, la confianza estadística proporciona una forma natural de equilibrar la compensación entre la tasa de rechazo y la tasa de error, excluyendo los patrones que tienen un nivel de confianza más bajo.

[5] El teorema afirma que todos los algoritmos de optimización funcionan igual de bien cuando su rendimiento se promedia en todos los problemas posibles.

Por otra parte, Hyunchul & Kyoung (2008) [2] proponen emplear un algoritmo genético (GA) para optimizar el número de vecinos. En este estudio aplican el modelo a dos problemas específicos: la predicción de precios en el mercado de valores y la predicción de las compras en línea. Los resultados exprimentales demuestran que el enfoque adoptado puede superar el tradicional algoritmo K-NN.

2.2.4. Revisión de las adaptaciones del algoritmo

Como se ha comentado anteriormente, la eficiencia del algoritmo puede mejorarse introduciendo cambios en los factores que lo rigen. Existen muchas variantes del K-NN que se han estudiado antes para dotar al algoritmo de una eficacia superior. Entre ellas se encuentras las siguientes:

- K-NN local adaptativo: propuesto por Wettschereck (1994) [49], elige el valor de K que debe usarse para clasificar una entrada comparando los resultados de los cálculos de validación cruzada en la vecindad local de los datos no etiquetados.

- K-NN ajustado por pesos: el ajuste de los pesos se ha dado atendiendo a un doble enfoque: por una parte, para ponderar la distancia, y por otra, para ponderar la predicción. El primero de ellos, introducido por Han et al. (2001) [20] sugiere que las distancias se transformen en medidas ponderadas según la influencia de cada atributo sobre la variable a predecir. Este tipo de algoritmos es especialmente útil en el caso de que un conjunto de datos tenga muchas características siendo alguna de ellas innecesarias. El segundo de ellos, desarrollado por Bailey & Jain (2000) [1] permite ser más precisos en la predicción de los valores.

- K-NN adaptativo: el K-NN en su versión original identifica el mismo número de vecinos más cercanos para cada nueva entrada. Por contra, el K-NN adaptativo, introducido por Sun & Huang (2010) [43], optimiza el valor de K para cada muestra a predecir.

- K-NN con vecinos más cercanos compartidos: propuesto por Zheng et al. (2015) [54], trata de igualar la densidad de distribución de los datos de entrenamiento. Muchos investigadores han descubierto que este algoritmo sufre mucho en su precisión debido a desigualdades en la función de densidad de las características. Por ello, en vista de este defecto, se presenta este algoritmo que puede calcular la similitud entre las muestras de entrenamiento con las muestras del vecino más cercano.

- K-NN con K-Means: otro enfoque improvisado del algoritmo es el descrito por Wira et al. (2012) [8]. Este algoritmo híbrido trata de dividir el conjunto de puntos en K conjuntos o clústers de manera que los puntos en cada clúster estén más cerca unos de otros. Los centros de dichos clústers se toman como las nuevas muestras de entrenamiento. De este modo, para predecir la clasificación de los datos no etiquetados se calcula su distancia al centro de entrenamiento recién encontrado, y el centro que comparta la distancia mínima es la que se asigna a esa clase. A diferencia del K-NN estándar, no hace falta predefinir el parámetro de entrada K en el cálculo de los vecinos, pero sí en el número de clústers a implementar. Realmente el enfoque propuesto se basa en la utilización del algoritmo K-Means como clasificador. Un planteamiento más interesante sería, una vez designado los clústers, implementar un K-NN diferente para cada uno de ellos.

- K-NN con SVM: esta versión compuesta fue empleada para el reconocimiento de categorías visuales por Zang et al. (2006) [52]. En este algoritmo el entrenamiento del SVM se realiza con la ayuda del K-NN. Para ello, en primer lugar, se determinan los puntos de datos más cercanos. A continuación, se computa la matriz de distancia entre esos K puntos y se emplea dicha matriz como Kernel en el algoritmo SVM.

- K-NN informativo: la experiencia pasada ha demostrado que la elección óptima de K depende de los datos, lo que hace que sea laborioso ajustar el parámetro para diferentes aplicaciones. Para ello, Yang et al. (2007) [42] introducen una nueva métrica que evalúa la informatividad de los objetos a predecir. De este modo, a partir del conjunto total de los datos, se discrimina entre los que son informativos y los que no. Al seleccionar un subconjunto de objetos más informativos, los métodos muestran estabilidad ante el cambio de los parámetros de entrada y el número de vecinos K.

- K-NN condensado: como se ha indicado anteriormente, el K-NN tiene una complejidad tanto espacial como temporal $O(nm)$, por lo que puede resultar interesante reducir el tamaño del *pool* de entrenamiento sin perder información útil. Para ello, Gowda & Krishna (1979) [19] proponen una modificación del K-NN en el que se eliminan las entradas del *dataset* que no aportan ninguna información adicional.

- K-NN Bayesiano: como ya se ha demostrado anteriormente, el algoritmo K-NN es muy sensible al valor de K. Para ello, Giuseppe (2017) [36] propone un nuevo enfoque para determinar su valor. Para ello, presenta un algoritmo

bayesiano para calcular la distribución de probabilidad posterior para K dado un punto objetivo dentro de un conjunto de datos, de manera eficiente y sin el uso de métodos de Monte Carlo. La idea central es que los puntos de alrededor del objetivo son generados por la misma distribución de probabilidad, extendiéndose hacia fuera sobre el número apropiado (y desconocido) de vecinos. Una vez que los datos se proyectan sobre una métrica de distancia elegida, se puede transformar la elección de K en un problema de detección de puntos de cambio, para el que existe una solución eficiente: calcular recursivamente la probabilidad del último punto de cambio a medida que se acerca al objetivo y, así, estimar la distribución de probabilidad posterior sobre K.

2.2.5. Adaptación del algoritmo para la predicción de series temporales

Para aplicar el algoritmo K-NN en un entorno de predicción de series temporales univariantes, las variables explicativas son valores desplazados en el tiempo de la variable explicada. La Figura 2.4 muestra un ejemplo de una predicción a un día utilizando una memoria de 3 días de la serie temporal.

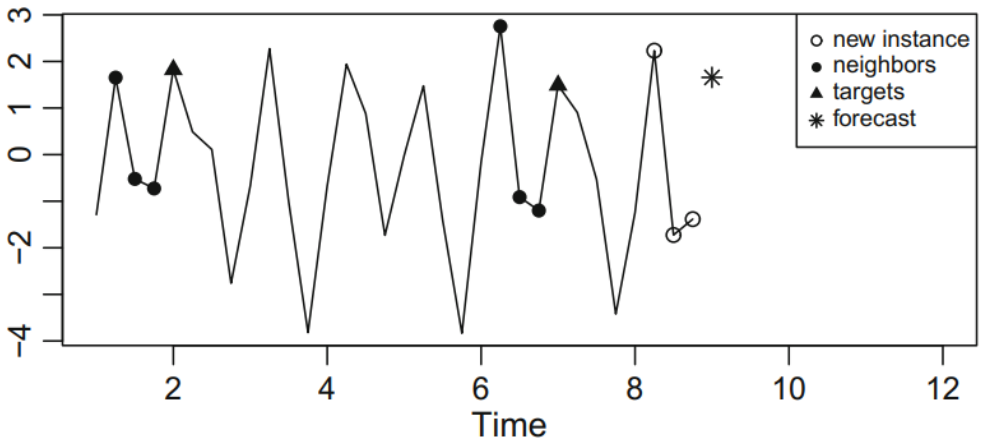


Figura 2.4. *Predicción one-step forward con $K = 2$*

Los últimos tres valores de la serie temporal son la nueva instancia sobre la que se hará la regresión y los dos conjuntos de puntos negros consecutivos son los 2 vecinos más cercanos, cuyos objetivos son los triángulos que se promedian para producir el valor predicho: el asterisco. La intuición subyacente al uso de K-NN

en series temporales es que cualquier serie temporal contiene patrones repetitivos, por lo que se pueden encontrar pautas de comportamiento anteriores similares a la estructura de la serie actual y utilizar sus patrones posteriores para predecir el comportamiento futuro.

2.3. Aplicación del algoritmo K-NN a los mercados financieros

Para concluir con el presente capítulo, se incluyen las investigaciones que han empleado el algoritmo del K-NN en el entorno financiero. La mayor parte de las mismas se centran únicamente en el análisis predictivo de la series temporales de los precios.

El primero de los estudios abordados es el realizado por Chen & Hao (2017) [9] en el que combinan el algoritmo de Máquina de Vectores de Soporte o *Support Vector Machines* (en adelante SVM) junto con el K-NN para la predicción de dos índices chinos (Shanghai y Shenzhen). La principal motivación de emplear dicho enfoque híbrido es poder discretizar el *pool* de vecinos donde se encuentran los patrones según la tendencia del valor a predecir. De este modo, el hecho de tener valores más similares hace que las predicciones generadas sean más precisas.

Para ello, en primer lugar, construyen un *dataset* con nueve indicadores técnicos diferentes, incluyendo entre ellos, medias móviles (MA), medias móviles exponenciales (EMA), MACD y el RSI a lo largo del periodo comprendido entre el 31 de octubre de 2008 y el 31 de diciembre de 2014, y lo dividen en dos (train y test). Una vez construido el *dataset* clasifican los valores según la tendencia del valor empleando para ello el algoritmo SVM, pero introduciendo una mejora significativa para solventar una de sus limitaciones. Es bien sabido que este clásico algoritmo de Machine Learning se basa en la suposición de que todas las características de las muestras contribuyen de la misma manera al valor objetivo. Sin embargo, esta suposición no es siempre cierta en muchos problemas aplicados en la vida real. En consecuencia, proponen un nuevo modelo de algoritmo SVM, denominado FWSVM (*Feature Weighted SVM*) basado en la ponderación de características de las muestras según la ganancia de información.

Para ello se parte del *dataset* de training $T_{train} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$,

que incluye las características de los indicadores técnicos, y el vector de la variable dependiente $y_i \in \{C_{-1}, C_{+1}\} = \{-1, +1\}$ que representa la tendencia alcista o bajista del valor en cuestión. Sea $C_{(i,D)}$ el subconjunto D perteneciente a la clase C_i donde $i = -1 + 1$, se obtienen las siguientes relaciones: $C_{[-1,D\}} \cup C_{\{+1,D\}} = D$ y $C_{[-1,D]} \cap C_{[+1,D]} = \emptyset$. Por lo tanto, la probabilidad de una muestra de pertenecer a la clase C_i se puede calcular como la probabilidad a priori de forma: $\left|C_{[i,D]}\right| / |D|$. De este modo, la información para clasificar el conjunto de datos puede expresarse mediante la siguiente fórmula:

$$Info(D) = -\sum_{i\in\{-1,+1\}} \frac{\left|C_{\{t,D]}\right|}{|D|} \log\left(\frac{\left|C_{\{i,D\}}\right|}{|D|}\right) \tag{2.2}$$

Supóngase que el conjunto de datos T_{train} se divide según la característica A en algunos subconjuntos indicados por $D_1, D_2, \ldots, D_v$. De acuerdo con la fórmula anterior la ganancia esperada es la siguiente:

$$\begin{aligned} \text{In}\, fo_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{In}\, fo\,(D_j) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \sum_{i\in\{-1,+1\}} \\ \times \frac{\left|C_{(i,D_j|}\right|}{|D_j|} \log\left(\frac{\left|C_{[i,D_j j}\right|}{|D_j|}\right) \end{aligned} \tag{2.3}$$

Así, la ganancia de información con respecto a cada característica se define como la raíz cuadrada de la resta entre la información original D y la nueva:

$$\text{InfoGain}(A) = \sqrt[2]{\text{In}\, fo(D) - \text{Info}(D)} \tag{2.4}$$

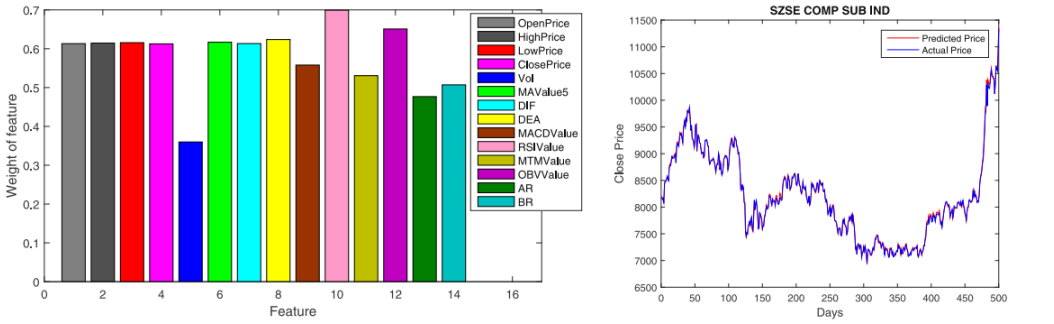
Si esta variable es mayor, la característica correspondiente es más importante y tiene una mayor contribución de cara a la clasificación. Por consiguiente, la ganancia de información puede aplicarse para medir la importancia de características, quedando la matriz de ponderación P definida de la siguiente forma:

$$P = \begin{Bmatrix} \text{InfoGain}\,(f_1) & 0 & \cdots & 0 \\ 0 & \text{InfoGain}\,(f_2) & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \text{InfoGain}\,(f_n) \end{Bmatrix} \tag{2.5}$$

Por lo tanto, una vez definida la matriz de ponderaciones, el SVM se encarga de separar los valores mediante un hiperplano $\omega \cdot \phi(\chi P) + b = 0$. A continuación, se implementa el algoritmo K-NN, en cada clase, para predecir la rentabilidad futuro.

Al igual que el SVM, el algoritmo K-NN es muy sensible a la función de distancia debido a la perceptibilidad inherente a las características irrelevantes. Por lo tanto, para computar la distancia entre dos pares de valores ponderan, también, dicha métrica mediante la matriz P anterior.

A la hora de calcular la rentabilidad futura, simulan diferentes escenarios según distintos horizontes de predicción. A continuación, se adjunta en la Figura 2.5 los resultados de la estrategia aplicada. A la izquierda, se representan la importancia de cada característica y a la derecha, se comparan el precio de cierre predicho y real.



(a) The distribution of the feature weight for 1 day ahead

(b) 1 day ahead prediction of closing price

Figura 2.5. *Resultados de la estrategia FWSVM-FWKNN [9]*

Finalmente, establecen una comparativa entre el rendimiento de una estrategia híbridas sin tener y teniendo en cuenta los pesos según la importancia de características. Dichos resultados se incluyen en la Tabla 4.3. En vista a los resultados, los autores ponen de manifiesto la importancia de aplicar las técnicas de *feature importance* y cómo dicha estrategia híbrida funciona mejor en estrategias a muy corto plazo.

No obstante, la principal limitación que tiene dicho estudio es que se trata de una investigación muy teórica centrada únicamente en la parte de la predicción de los precios, sin llegar a diseñar ningún tipo de estrategia de inversión a posteriori.

El siguiente estudio revisado es el llevado a cabo por Teixeira & de Oliveira (2010) [46] en el que desarrollan un método para automatizar el trading algorítmico combinando el análisis técnico con el algoritmo K-NN. A diferencia del trabajo anterior, en este los autores emplean el algoritmo como clasificador, en lugar de como un regresor. Para ello, toman 15 acciones del índice de la Bolsa de Sao Paulo

Metric	Method	1 day	5 days	10 days	15 days	20 days	30 days
MAPE	SVM-KNN	0.20	0.22	0.22	0.24	0.26	0.31
MAPE	FWSVM-FWKNN	0.18	0.22	0.21	0.23		
RMSE	SVM-KNN	0.0051	0.0056	0.0061	0.0062	0.0072	0.0105
RMSE	FWSVM-FWKNN	0.0050	0.0056	0.0061	0.0060	0.0070	0.0105

Tabla 2.11. *Comparación del rendimiento del SZSE entre FWSVM-FWKNN y SVM-KNN en MAPE y RMSE [9]*

(Bovespa) durante el periodo comprendido entre el 1 de abril de 1998 y el 9 de marzo de 2009 y se sirven de los siguientes indicadores técnicos para representar las características de entrada del modelo:

$$
\begin{aligned}
I_1 &= RSI(t) \\
I_2 &= \left(x(t) - \text{ bollinger }_{\text{upper}}\right) / \text{ bollinger }_{\text{upper}} \\
I_3 &= \left(x(t) - \text{ bollinger }_{\text{lower}}\right) / \text{ bollinger }_{\text{lower}} \\
I_4 &= \%K(t) \\
I_5 &= \%D(t) \\
I_6 &= \%K(t) - \%K(t-1) \\
I_7 &= \%D(t) - \%D(t-1) \\
I_8 &= (x(t) - x(t-1))/x(t-1) \\
I_9 &= (x(t) - x_l(t)) \,/\, (x_h(t) - x_l(t)) \\
I_{10} &= (PMA_s(t) - PMA_s(t-1)) \,/PMA_s(t-1) \\
I_{11} &== (PMA_l(t) - PMA_l(t-1)) \,/PMA_l(t-1) \\
I_{12} &= (PMA_s(t) - PMA_l(t-1)) \,/PMA_l(t-1) \\
I_{13} &= (x(t) - PMA_l(t)) \,/PMA_l(t) \\
I_{14} &= (x(t) - \text{mín}(x(t), \ldots, x(t-5))/ \,\text{mín}(x(t), \ldots, x(t-5)) \\
I_{15} &= (x(t) - \text{máx}(x(t), \ldots, x(t-5))/ \,\text{máx}(x(t), \ldots, x(t-5)) \\
I_{16} &= (v(t) - v(t-1))/v(t-1) \\
I_{17} &= (VMA_s(t) - VMA_s(t-1)) \,/VMA_s(t-1) \\
I_{18} &= (VMA_l(t) - VMA_l(t-1)) \,/VMA_l(t-1) \\
I_{19} &= (VMA_s(t) - VMA_l(t-1)) \,/VMA_l(t-1) \\
I_{20} &= (v(t) - VMA_l(t)) \,/VMA_l(t) \\
I_{21} &= (v(t) - \text{mín}(v(t), \ldots, v(t-5))/ \,\text{mín}(v(t), \ldots, v(t-5)) \\
I_{22} &= (v(t) - \text{máx}(v(t), \ldots, v(t-5))/ \,\text{máx}(v(t), \ldots, v(t-5))
\end{aligned}
$$

El conjunto de datos lo dividen en 10 subconjuntos donde cada uno de ellos contiene aproximadamente un año de cotización. A continuación, desarrollan un backtesting *walk-forward*, en el que seleccionan tres subconjuntos posteriores para el entrenamiento y el siguiente como test. Desarrollan una estrategia en la que consideran un modelo de negociación diario, definiendo para los experimentos un saldo inicial de efectivo y para cada día del backtesting t emplean la salida del K-NN como clasificador para iniciar una operación de compra (larga) o de venta (corta). Dichas operaciones las simulan empleando todo el saldo de caja disponible y respetando el número mínimo de 100 acciones por operación, pues se trata de una norma en el mercado de valores brasileño, y considerando los costes de transacción habituales de dicha Bolsa. Además, también consideran unos *stop loss* y

Stock	Classifier precision (%)	Difference to Buy & Hold (%)	Buy operations per year
AMBV4	40,78	+12,20	27,85
ARCZ6	34,47	+108,43	44,57
BBAS3	36,12	+45,34	47,28
BBDC4	36,81	−36,78	44,14
CMIG4	36,99	+1,10	48,28
CRUZ3	36,86	−8,09	38,57
CSNA3	39,16	+18,86	48,28
ELET 6	37,34	−10,90	54,28
ITAU4	37,40	+115,99	44,85
ITSA4	37,20	+76,83	37,14
NETC4	37,87	+51,98	52,00
PETR4	34,59	+8,13	42,14
TNLP4	35,96	+95,20	37,71
USIM5	40,00	+47,18	43,28
VALE5	36,55	+33,12	40,14

Tabla 2.12. *Resultados medios para la mejor configuración [46]*

stop gain para controlar mejor el riesgo asumido.

Para evaluar el potencial uso del método propuesto, comparan los resultados obtenidos con los que se alcanzarían adoptando una estrategia de comprar y mantener. La medida de rendimiento clave en dicha comparativa es la rentabilidad conseguida. La estrategia introducida generó unos beneficios considerablemente superiores a los de la estrategia de Buy & Hold para la mayoría de las empresas, como se puede observar en la Tabla 2.12, ejecutando pocas operaciones de compra y, en consecuencia, minimizando el riesgo de exposición al mercado.

Sin embargo, existen claras limitaciones en el estudio de este documento. En primer lugar, sólo se consideran 15 acciones como referencia de un único mercado. Esto no es necesariamente erróneo, pero sí cuestiona la generalidad del enfoque propuesto a una cartera más amplia. En segundo lugar, el espacio de acciones que el modelo puede realizar es muy limitado. En tercer lugar, no se detalla nada del cálculo de los vecinos ni la metodología empleada para imputar las distancias. Por último, no se realiza ningún tipo de normalización de los indicadores de entrada y, precisamente, el K-NN se trata de un algoritmo muy sensible a las escalas.

De este modo, el artículo propuesto por Li et al. (2018) [24] supera las principales limitaciones de las que adolece este enfoque. Para explorar patrones rentables en los precios históricos de las acciones, normalmente se selecciona unos prometedores (e.g. dobles suelos) y luego se verifica su rendimiento comparando la rentabilidad condicional del patrón con la de referencia. Sin embargo, los autores proponen una filosofía alternativa. Sin seleccionar ningún patrón en particular, el método propuesto explora el abanico entero de opciones para detectar los que tienen claros rendimientos predictivos, no sólo los patrones conocidos, sino también los desapercibidos y potencialmente informativos. Este enfoque es muy similar al propuesto en el presente trabajo.

La principal motivación es que hay una gran diferencia entre el análisis técnico y el método clásico del K-NN. Como ya se ha comentado anteriormente, el núcleo de dicho algoritmo reside en encontrar a los vecinos más cercanos. Dado que el concepto de *"más parecido"* es relativo al valor de K, siempre se pueden encontrar patrones más similares. Supóngase que cada patrón tiene la misma oportunidad de convertirse en uno de los vecinos más cercanos de los demás. En este sentido, ello implica que cada patrón es igualmente informativo, lo cual parece contradictorio. Para los defensores del análisis técnico, aparentemente, no todos los patrones, sino algunos muy específicos pueden tener el poder de predecir el futuro movimiento del mercado. Por contra, los partidarios de la Hipótesis de los Mercados Eficientes (EMH) sostienen que ninguno de ellos son informativos. Los autores defienden que la explicación correcta es probablemente una sinergia de ambos puntos de vista y, por ello, desarrollan un modelo que pretende heredar las ventajas de ambos. En consecuencia, sostienen que hay dos tipos de patrones fundamentales: los informativos y los no informativos. Los segundos son los que tienen menos poder de predicción y son posiblemente rastros de un *random walk*, mientras que los primeros poseen la información con poderes de predicción, y son estadísticamente significativos en el conjunto histórico. En consecuencia, el método de predicción puede beneficiarse del tratamiento separado entre dos categorías de patrones.

De este modo, para detectar el grado de información del vector de vecinos más cercanos formulan las siguientes hipótesis. El coste de transacción se denota mediante la variable c, y sólo se abren las posiciones cuando los valores evolutivos de los vecinos más cercanos sugieren que el rendimiento esperado puede cubrir el coste.

$$Hipotesis\ nula: \ \tilde{y} < c$$
$$Hipotesis\ alternativa: \ \tilde{y} > -c$$
$$siendo\ \tilde{y} = \frac{1}{K}\sum_{i=1}^{K}\tilde{y}_i.$$

Para ver si se puede refutar la hipótesis nula, se conduce un análisis *t-test*. De confirmarse, el patrón no se puede considerar significativo y no se tiene en cuenta.

$$t = \frac{\mu_1 - \mu_0}{\sqrt{\sigma^2/N}} \tag{2.6}$$

donde N es el número de acciones del portfolio (igual a 108 en el estudio); μ_1 es la media diaria de los rendimientos logarítmicos de dichas acciones; μ_0 es igual a 0; y σ^2 es la varianza de los retornos de las acciones en el portfolio.

Para evaluar el rendimiento de la estrategia propuesta recopilan datos de 108 valores americanos seleccionados entre los componentes del índice Dow Jones y el Nasdaq en el periodo comprendido entre el 2 de enero de 1994 y el 31 de diciembre de 2014. Para ello, emplean como métricas el rendimiento anualizado y el ratio de Sharpe. Adicionalmente, realizan un análisis *out-of-sample*, adjunto en la Tabla 2.13, para estudiar la sensibilidad del modelo frente a la memoria interna empleada para detectar el patrón del K-NN (*embedding dimension, m*), el horizonte de predicción de la estrategia (*T step*) y el número de vecinos empleado en cada caso. Los autores revelan que el mejor resultado se alcanza considerando un horizonte temporal de 7 días y $K = 500$.

Por último, para contrastar el rendimiento del método utilizan como benchmarks dos estrategias empleadas tradicionalmente en el análisis técnico: medias móviles y momentum [6]. Los resultados indican que la estrategia basada en las medias móviles no consigue garantizar un beneficio positivo de forma estadísticamente significativa (Tabla 2.14).

[6]El método de cruce de medias móviles es una de las estrategias de trading más utilizadas, en la que una media móvil (MA) a corto plazo se abre paso a través de una MA a más largo plazo para constituir una señal de compra o venta. Por contra, seguir una estrategia momentum consiste en invertir en aquellos activos financieros que estén mostrando un potencial incremento en su precio. Se compran aquellos activos que están siguiendo tendencias fuertemente alcistas esperando que esta fortaleza alcista se prolongue en el tiempo.

		Embedding dimension m					
		5	7	9	11	13	15
T Sharpe ratio	250	0,056	0,041	0,037	0,030	0,024	0,007
	300	0,061	0,058	0,036	0,030	0,024	0,015
	350	0,070	0,070	0,049	0,032	0,027	0,030
	400	0,070	0,063	0,047	0,033	0,035	0,049
	450	0,072	0,072	0,048	0,047	0,046	0,061
T annualized	500	0,074	0,077	0,048	0,058	0,043	0,045
return	250	7,47	8,45	7,42	6,18	4,89	1,32
	300	8,87	12,33	7,56	6,46	5,23	3,27
	350	11,02	15,59	10,96	6,9	6,02	6,76
	400	11,31	14,98	10,8	7,43	7,97	11,51
	450	12,14	17,37	11,33	11	10,47	14,83
	500	15,31	19,33	11,5	13,85	10,03	10,81

Tabla 2.13. *Rendimiento del portfolio según el método propuesto [24]*

(s m a, I m a, h, b)	Sharpe ratio	Maximum drawdown	t-Value
$(5, 100, 0, 0{,}01)$	0,0090	−0,0884	2,0570
$(5, 100, 0, 0{,}005)$	0,0092	−0,0785	2,1320
$(5, 100, 0, 0)$	0,0103	−0,0684	2,3798
$(7, 100, 0, 0{,}01)$	0,0104	−0,0805	2,4350
$(7, 100, 0, 0{,}005)$	0,0116	−0,0466	2,7397
$(7, 100, 0, 0)$	0,0125	−0,0297	2,9752
$(9, 100, 0, 0{,}005)$	0,0133	−0,0311	3,0581
$(9, 100, 0, 0{,}01)$	0,0137	−0,0307	3,1731
$(9, 100, 0, 0)$	0,0147	−0,0254	3,3236

Tabla 2.14. *Rendimiento del portfolio con una estrategia de cruce de medias móviles [24]*

Lookback period	Sharpe ratio	Maximum drawdown	t-Value
10	0,0420	−0,6931	8,6234
20	0,0461	−1,2029	9,3847
30	0,0470	−1,6171	9,5087
40	0,0451	−1,4991	9,2852
50	0,0432	−1,3900	9,3225
60	0,0442	−1,2541	9,4239
70	0,0453	−1,0409	9,5471
80	0,0446	−0,7921	9,5853
90	0,0451	−0,5712	9,5369
100	0,0443	−0,3496	9,5307

Tabla 2.15. *Rendimiento del portfolio con una estrategia momentum [24]*

Por otra parte, la estrategia de momentum obtiene rendimientos estables en todas las reglas y los ratios de Sharpe de la cartera son de aproximadamente 0.04 (Tabla 2.15). Sin embargo, estas precisiones no son lo suficientemente altas como para batir al mercado a o a los métodos K-NN de la cartera. La razón fundamental es que los métodos no lineales, como el K-NN, dominan a los métodos lineales porque los rendimientos del mercado son sustancialmente no lineales.

El siguiente artículo analizado es el desarrollado por Naranjo & Santos (2020) [34]. En él proponen un novedoso sistema de recomendación basado en la lógica difusa. Esta herramienta de decisión emplea velas japonesas difusas e incluye el efecto de la devaluación de la moneda en la predicción. Para ello, primero computan la siguiente sesión de mercado mediante el sistema de predicción difuso diseñado. A continuación, se compara con el real mediante el algoritmo del K-NN. Por último, construyen una estrategia de asignación de capital basada en el todo o nada con control del riesgo, pero sin incluir la capitalización.

Su enfoque no sólo emplea la lógica difusa [7] en el reconocimiento de los patrones, sino que también trabaja con la construcción de velas japonesas difusas. Para ello, toman como semilla el trabajo realizado por Lee et al. (2006) [23], y distinguen tres partes fundamentales en la vela en cuestión: el cuerpo, la sombra superior

[7]La lógica difusa es una lógica paraconsistente multivaluada en la cual los valores de verdad de las variables pueden ser cualquier número real comprendido entre 0 y 1, permitiendo tomar decisiones más o menos intensas en función de grados intermedios de cumplimiento de una premisa

y la inferior, tal y como se puede observar en la Figura 2.6. Dichas entradas se parametrizan aplicando las siguientes ecuaciones:

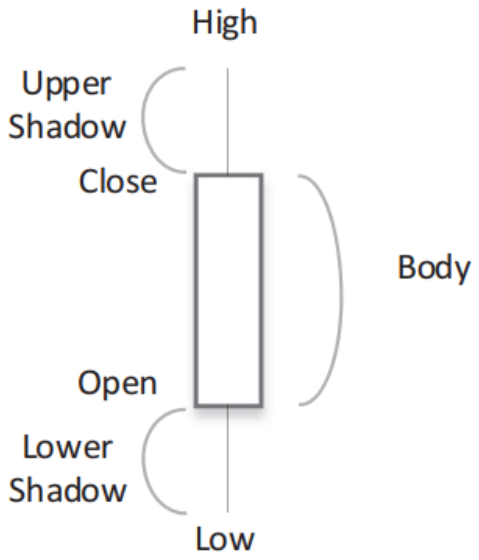


Figura 2.6. *Parámetros y partes de una vela japonesa [34]*

$$\begin{aligned} \mathrm{L_{upper}}\ (\mathrm{t}) &= 100 \cdot \frac{\mathrm{high(t)} - \mathrm{máx(open(t),\ close\ (t))}}{\mathrm{open\ (t)}} \\ \mathrm{L_{lower}}\ (\mathrm{t}) &= 100 \cdot \frac{\mathrm{mín(open(t), close(t))} - \mathrm{low(t))}}{\mathrm{open\ (t)}} \\ \mathrm{L_{body}}\ (\mathrm{t}) &= 100 \cdot \frac{\mathrm{close(t)} - \mathrm{open(t)}}{\mathrm{close(t)}} \end{aligned} \tag{2.7}$$

A continuación, asignan cuatro funciones de pertenencia trapezoidales (NULL, SHORT, MIDDLE Y LONG) a las variables difusas *Lupper* y *Llower* para cada día de bolsa, tal y como se puede observar en la Figura 2.7. Por contra, para la variable de entrada *Lbody* se emplean siete funciones de pertenencia triangulares y trapezoidales en cada sesión de negociación (BLACK LONG, BLACK MIDDLE, BLACK SHORT, NULL, WHITE SHORT, WHITE MIDDLE, WHITE LONG), adjuntas en la Figura 2.8.

Una vez designadas dichas funciones, diseñan las reglas de la lógica difusa teniendo en cuenta dos outputs diferentes: *Rsize* y *Rpos*. Estas variables se refieren a

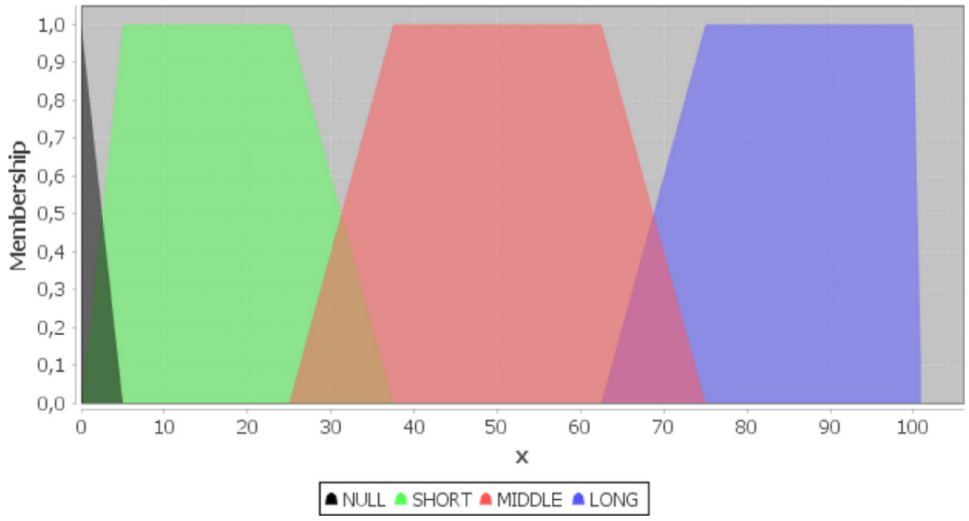


Figura 2.7. *Funciones de pertenencia para Lupper y Llower[34]*

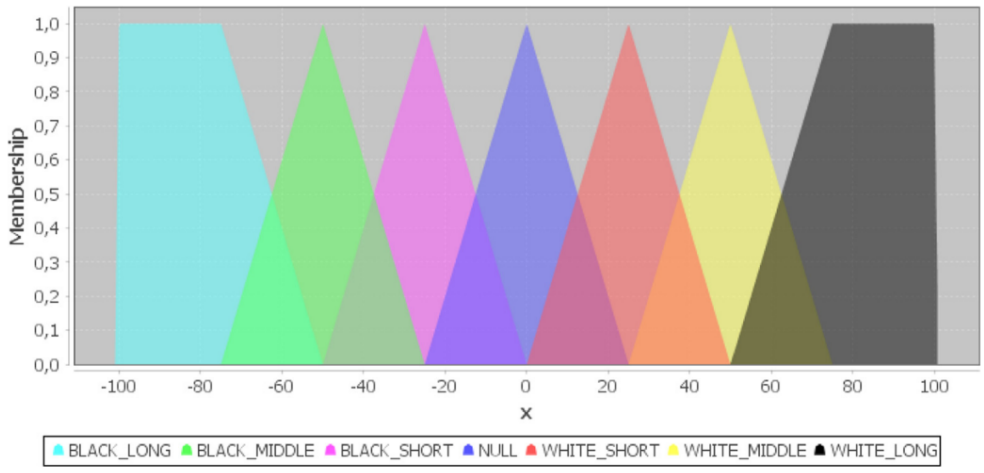


Figura 2.8. *Funciones de pertenencia para Lbody [34]*

la relación existente entre el tamaño del cuerpo y el tamaño total de la vela (*Rsize*) y a la posición del cuerpo dentro de la vela (*Rpos*). Con ello, designan un total de 224 reglas difusas a partir de la combinación de las entradas y salidas definidas anteriormente. Una vez que se han calculado los resultados de las dos variables de salida difusa (*Rsize* y *Rpos*) se usan como entradas para el clasificador del K-NN. El sistema de predicción consiste, básicamente, en comparar la vela de la sesión

anterior disponible, caracterizada por las variables anteriores, con un grupo de n velas, formado con la vela anterior conocida y las $n-1$ anteriores. Posteriormente, comparan la diferencia empleando la distancia euclídea.

La ventaja competitiva de este método es que no sólo implementa un sistema de predicción, sino que desarrolla un método holístico de inversión, determinando los puntos de entrada y salida del mercado, así como, el capital a invertir en cada acción. Para evaluar el rendimiento de la estrategia aplican el método a 15 compañías del Nasdaq-100 y al IBEX-35, y comparan el performance del modelo frente al artículo propuesto por ellos mismos en 2016 [35]. Los resultados se incluyen en la Figuras 2.9 y 2.10. Los autores afirman que los resultados son muy prometedores, sin embargo, no tienen en cuenta los costes de transacción, los cuales tienen un impacto primordial en este tipo de estrategias a corto plazo tan agresivas. Por lo tanto, para validar la estrategia sería necesario incluir algún escenario donde fuesen contemplados.

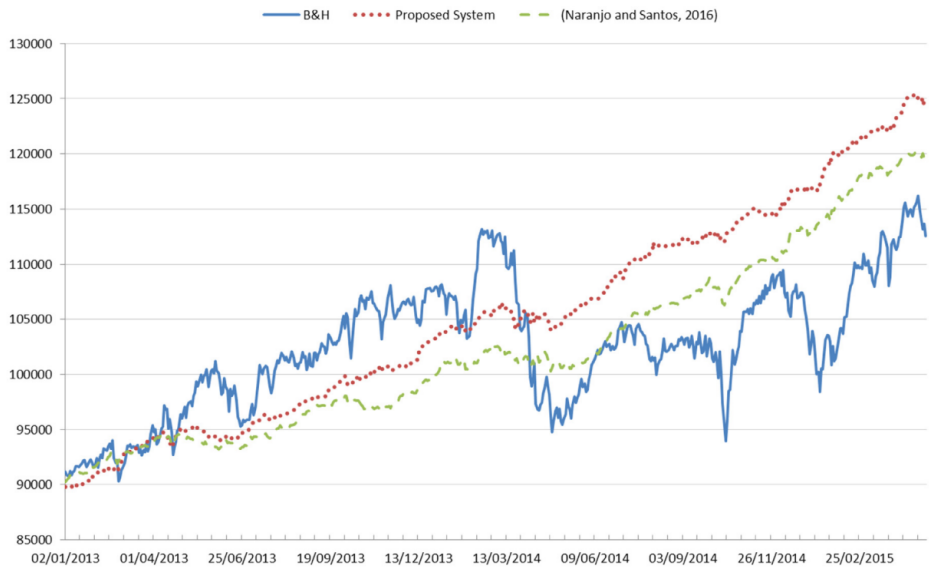


Figura 2.9. *Evolución del capital en el periodo de validación para las tres estrategias con el Nasdaq-100 [34]*

Otro estudio muy interesante es el realizado por Zhang et al. (2017) [53] en el que proponen una nueva metodología de dos etapas que combina la descomposición empírica modal con ensamblaje (EEMD) con un modelo multidimensional

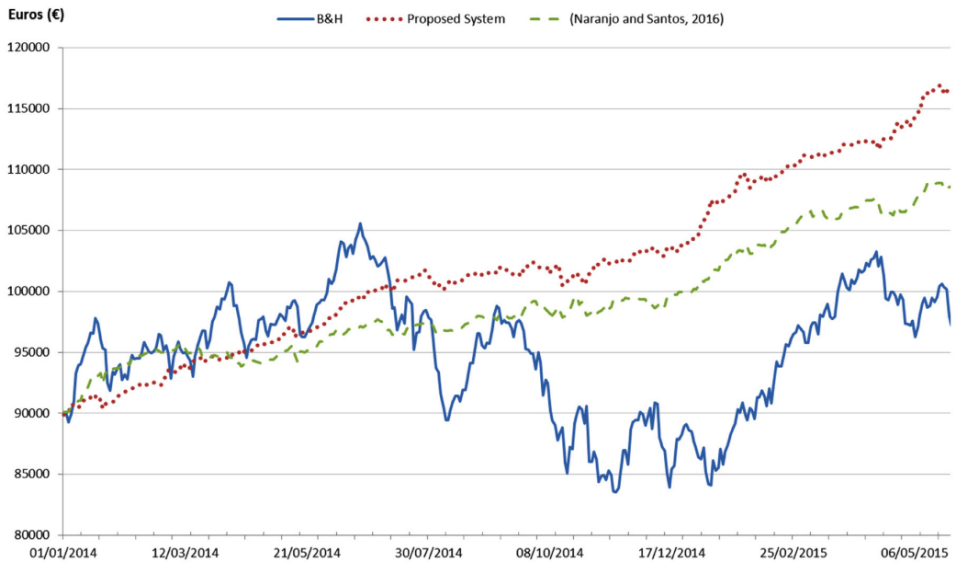


Figura 2.10. *Evolución del capital en el periodo de validación para las tres estrategias con el IBEX-35 [34]*

basado en el K-NN, con el objetivo de predecir el precio de cierre y máximo de cuatro índices (Nasdaq-100, S&P 500, DJI y STI) a lo largo de un periodo de 20 años comprendido entre el 25 de julio de 1996 y 2l 25 de julio de 2016.

La descomposición modal empírica (EMD) descompone una señal no lineal y no estacionaria en una serie de funciones modales intrínsecas (IMF). No obstante, no puede revelar la información característica de la señal con mucha precisión como resultado de la mezcla de modos. Por ello, los autores introducen la descomposición modal empírica conjunta (EEMD) como método para resolver las debilidades de EMD añadiendo ruido blanco a los datos originales. En la Figura 2.12 se adjunta una representación de las señales resultantes. Cada IMF capta las fluctuaciones de la serie temporal original ocurridas en distintas escalas de tiempo; a mayor orden, mayor escala temporal.

Por otra parte, el desarrollo del algoritmo K-NN lo implementan teniendo en cuenta un enfoque muy singular. El K-NN tradicional sólo tiene en cuenta la distancia euclídea, o cualquier otra, para predecir los valores más cercanos, pero ignora los cambios del vector patrón de la serie temporal. Para ello, proponen como alternativa que el K-NN no sólo considere dicha métrica, sino que también tenga en cuenta la similitud de los vectores de patrones, para mejorar así su capacidad pre-

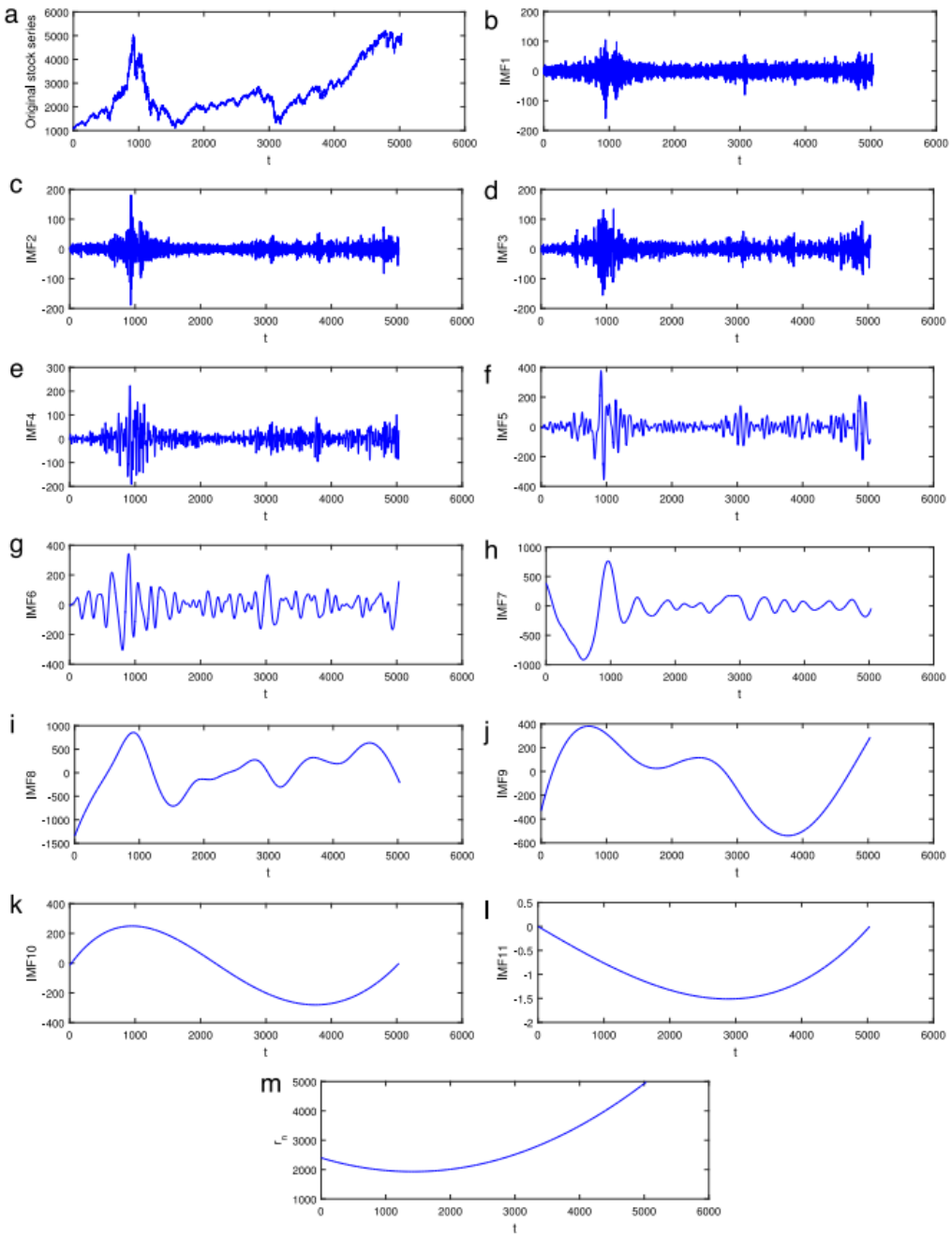


Figura 2.11. *Descomposición empírica modal de una serie temporal financiera [53]*

dictiva. Para ilustrarlo, supóngase que se parte de un vector de una serie temporal X $\{X = x_1, x_2, \ldots, x_n\}$ y otra serie Y $\{Y = y_1, y_2, \ldots, y_n\}$. Con ellos se define una matriz de diferencias $Q = \begin{pmatrix} q^x_{n-1} & q^x_{n-l+1} & \cdots & q^x_n \\ q^y_{n-1} & q^x_{n-l+1} & \cdots & q^y_n \end{pmatrix}$ donde $l(1 \leq l \leq n-1)$ es el tamaño del patrón; $q^x_i = x_{i+1} - x_i$; $n - l \leq i \leq n-1$ y $q^y_j = y_{j+1} - y_j$. Posteriormente, se mapean esas diferencias definiendo una variable binaria según la dirección del precio -1, 0 o 1, donde $d_i = -1$ si $q_i < 0$; $d_i = 1$ si $q_i > 0$ y $d_i = 0$ si $q_i = 0$.

De este modo, el algoritmo trata de buscar los K-vecinos más cercanos en la matriz $P'_d = \begin{pmatrix} d^x_{j-l} & \cdots & d^x_{j-1} \\ d^y_{j-1} & \cdots & d^y_{j-1} \end{pmatrix}$, que se construye a partir de $Q'_d = \begin{pmatrix} q^x_{j-1} & \cdots & q^x_{j-1} \\ q^y_{j-1} & \cdots & q^y_{j-1} \end{pmatrix}$. Posteriormente computa el valor estimado $\hat{X}_{n+1}$ e $\hat{y}_{n+1}$ basándose en en las diferencias anteriores:

$$\begin{aligned} \hat{x}_{n+1} &= x_n + \hat{Q}_m \quad \text{donde } \hat{Q}_m = \sum_{h=1}^{k} \frac{\hat{Q}^h_j}{k} \\ \hat{y}_{n+1} &= y_n + \tilde{Q}_m \quad \text{donde } \tilde{Q}_m = \sum_{h=1}^{k} \frac{\tilde{Q}^h_j}{k}. \end{aligned} \tag{2.8}$$

En el caso particular de estudio $\hat{X}_{n+1}$ e $\hat{y}_{n+1}$ lo representan el precio de cierre y el de apertura de los cuatro índices anteriores.

Para evaluar el rendimiento de la estrategia híbrida implementada, comparan el método con 3 modelos diferentes: ARIMA, un K-NN tradicional y un K-NN con EMD. La Figura 2.12 permite visualizar dicho estudio. Adicionalmente, se adjunta la Tabla 2.16 en la que se incluyen los errores cometidos en cada caso. Dichos experimentos confirman la validez del método propuesto. No obstante, a pesar del potencial predictivo que revela el método, se trata de una investigación muy académica y sin ninguna simulación real y práctica en el ámbito financiero.

Finalmente, el último estudio analizado es el propuesto por Nakagawa et al. (2018) [33] en el que utilizan el K-NN con la técnica de la deformación temporal dinámica o *indexing dynamic time warping* (IDTW) para evaluar la similitud entre dos series temporales. En la literatura se han propuesto muchos métodos para medir las similitudes entre los datos de las series temporales; los coeficientes de correlación y la distancia euclídea se utilizan a menudo como métricas para dicha aproximación. Sin embargo, no son los más adecuados para evaluar la semejanza

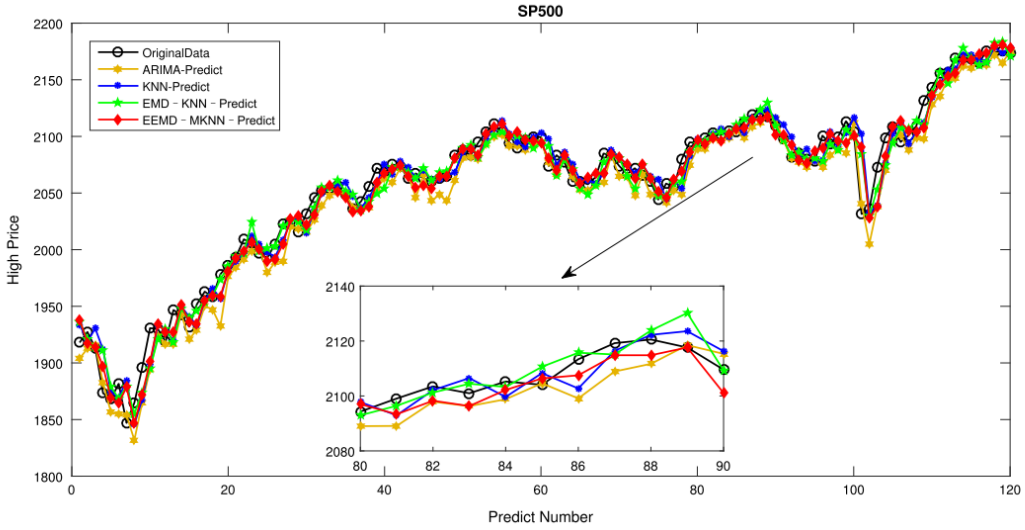


Figura 2.12. *Comparación de la predicción del precio máximo por distintos modelos [53]*

Methods	MAPE	MASE	NMSE
EEMD-MKNN	0,4201	0,8336	0,1634
EMD-KNN	0,4171	0,8273	0,1649
KNN	0,5009	0,9929	0,1907
ARIMA	0,5732	1,1433	0,2079

Tabla 2.16. *Comparación del NMSE, MASE y MAPE para la predicción del precio máximo del S&P 500*

entre los datos de las series temporales financieras. Los coeficientes de correlación sólo capturan la relación lineal entre los datos de las series temporales, y la distancia euclídea es muy sensible a pequeñas distorsiones en el eje temporal. Además, no es aconsejable su uso cuando las escalas de dos series temporales son diferentes. En este sentido la distancia DTW es una medida que puede emplearse para superar ambos problemas.

La aplicación de dicha técnica supone la realización de una correspondencia óptima con respecto a dos series temporales mediante la expansión y contracción no lineal del eje temporal. La Figura 2.14 muestra un ejemplo de correspondencia de datos de series temporales basándose en la distancia euclídea y la DTW.

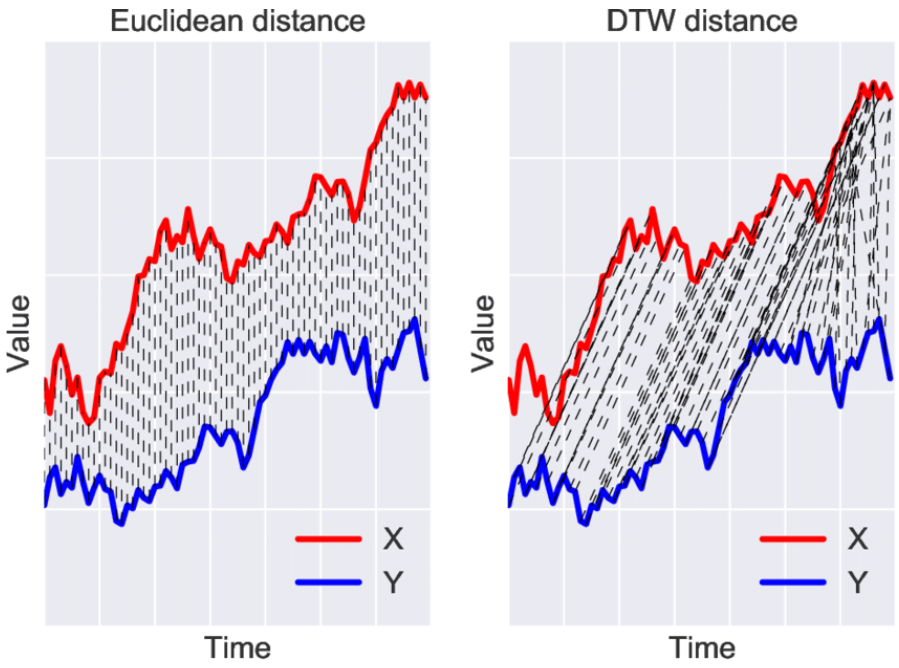


Figura 2.13. *Cálculo de la correspondencia de datos en diferentes series temporales [33]*

Para validar la eficacia de la estrategia se ejecuta el modelo sobre los índices TOPIX, S&P 500, FTSE 100, DAX 30 y CAC 40, a lo largo del periodo comprendido entre enero de 1989 y diciembre de 2005. Los resultados (out-of-sample) se pueden visualizar gráficamente en la Figura 2.14 y cuantitativamente en la Tabla 2.17. Todos ellos revelan que las predicciones del precio de las acciones utilizando la estrategia conjunta del $IDTW + k^* - NN$ son superiores a los métodos en término de rentabilidad.

		CAC40	DAX30	FTSE100	S&P500	TOPIX	Avg.
DTW	$k-$NN	0,98 %	146,03 %	34,32 %	148,15 %	146,79 %	95,25 %
	k^*-NN	122,22 %	164,03 %	66,98 %	97,12 %	152,67 %	120,60 %
DTW	$k-$NN	157,14 %	74,38 %	98,21 %	185,76 %	142,32 %	131,56 %
	k^*-NN	93,53 %	154,90 %	114,29 %	139,60 %	156,98 %	131,86 %
IDTW	$k-$NN	126,59 %	124,26 %	**142,10** %	212,87 %	174,20 %	156,00 %
	k^*-NN	**222,24** %	**212,91** %	140,74 %	**235,29** %	**234,53** %	**209,14** %

Tabla 2.17. Rendimientos out-of-sample de cada método [33]

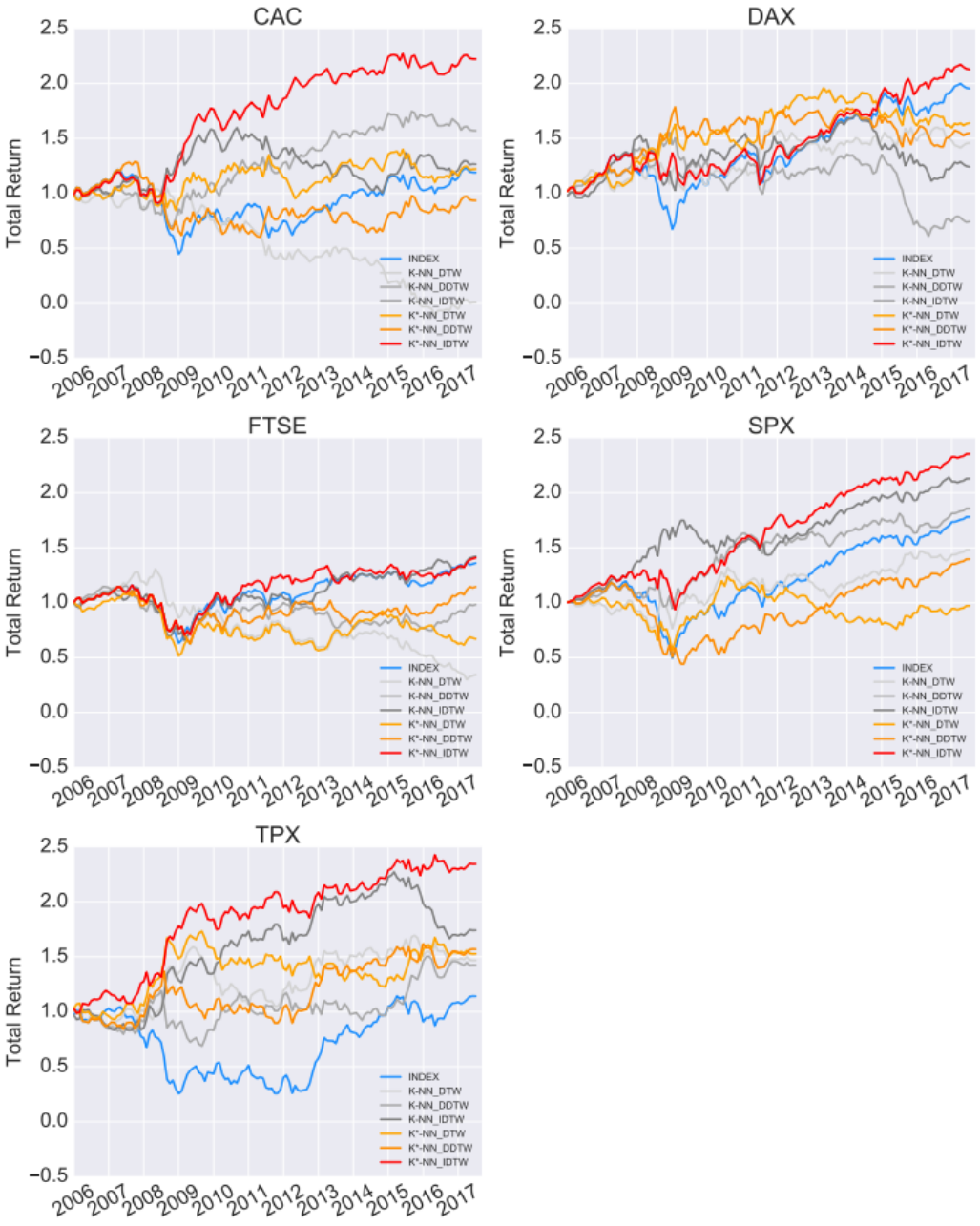


Figura 2.14. *Retornos acumulados para cada índice según el método [33]*

Capítulo 3

Implementación de la estrategia de inversión

El mercado financiero es un sistema complejo, no estacionario, muy ruidoso y dominado por la no linealidad, pero a pesar de ello no sigue un camino fortuito subordinado al gusto del azar. Dado que la Bolsa es una de las arterias más importantes del corazón financiero, existen multitud de factores que pueden provocar fluctuaciones en el movimiento de estos mercados, lo que dificulta la predicción del precio de las acciones que cotizan en ellas. Tanto es así que la correcta predicción de los mercados financieros es una antigua quimera perseguida desde sus inicios por sus inversores, siendo el diseño de estrategias de inversión uno de los mayores epicentros de controversias en el mundo financiero en las últimas décadas. No existe una fórmula estática y cerrada capaz de batir al mercado en cualquier circunstancia, ya que el propio mercado la haría inútil en muy poco tiempo. Por ello, la dificultad de establecer un modelo que se ajuste a los cambios y a la volatilidad del mercado es la piedra angular del éxito de cualquier estrategia de inversión.

El método diseñado en el presente trabajo se puede estructurar en torno a tres ejes principales: la construcción de los indicadores de entrada, la búsqueda de patrones similares en una ventana deslizante y la ordenación de los valores según un criterio de clasificación específico. En este sentido, se toman como entradas la evolución temporal de tres velas japonesas de un día, una semana y un mes, con el fin de sintetizar las fluctuaciones del mercado en diferentes horizontes temporales. En cuanto a la búsqueda de patrones, se utiliza el algoritmo K-NN para detectar comportamientos similares de los activos. Por último, según el criterio de ordenación, se ponderan tres métricas (la tasa de acierto y la rentabilidad media y mediana predicha) en un índice para evaluar la robustez de la rentabilidad de

cada predicción. Con respecto a la duración de cada operación, se trata de una estrategia de inversión ciega, ya que tiene un horizonte de predicción fijo de una semana. Por otra parte, para sofisticar las señales de compraventa de los valores, las predicciones generadas por el K-NN se someten a umbrales para que el modelo sea más selectivo en cuanto a las acciones a comprar.

Una vez explicado, en términos generales, el núcleo de la estrategia de inversión, conviene profundizar en cada uno de los ejes mencionados anteriormente.

3.1. Construcción de las velas japonesas

Tras describir de manera unívoca el problema a tratar, conviene presentar los datos con los que se va a trabajar. Se parte de la evolución temporal del precio de 2604 valores americanos que han cotizado a partir de 2011 hasta la actualidad. La fuente de datos está proporcionada por *Yahoo Finance* e incluye la información del precio de apertura, cierre, máximo y mínimo de cada una de las sesiones del intervalo muestreado, así como el volumen de las transacciones efectuadas cada día. Es precisamente este banco de información el que se utiliza para alimentar el modelo de búsqueda de los patrones. Todo modelo de predicción necesita ser entrenado con una serie de datos, y al ser el K-NN un algoritmo de Machine Learning de tipo supervisado, esta fuente de información se compone tanto de las variables independientes, que pretenden explicar el comportamiento de la variable de salida, como de la propia variable de salida, que se ha obtenido de manera real y que posteriormente se contrastará con la predicha para valorar la capacidad predictiva del modelo. A continuación, se adjunta en la Figura 3.1 una gráfica que ilustra la evolución de las velas japonesas para Apple en el último trimestre.

En este sentido, dichas señales, tanto las de entrada como la de salida, se generan mediante la parametrización de los datos anteriores a lo largo de una ventana de días. Para el caso de las variables de entrada se dispone de tres ventanas de diferentes longitudes (1, 5 y 20 días), como se puede observar en la Figura 3.2, y para la salida, se parte de una única ventana de igual valor al horizonte de predicción empleado en la estrategia de inversión. En este caso en particular, para la caracterización de esas ventanas temporales se emplean las velas japonesas, pues son capaces de resumir en detalle toda la información relevante de una sesión.

En este punto resulta esencial introducir un matiz en la construcción de estas



Figura 3.1. *Velas japonesas para AAPL en el último trimestre*

velas. Como se operan con valores que poseen rangos de cotización muy dispares, resulta esencial normalizar los valores para así garantizar que los datos sean comparables entre sí. De esta forma, se computa un ratio de rentabilidad normalizada con el precio de cierre del inicio de cada ventana temporal. Como todos los valores pertenecen al mismo mercado, no es necesario calcular la rentabilidad procedente de la evolución del tipo de cambio. De este modo, la construcción de las entradas del modelo para un día cualquiera k se puede sintetizar de forma matemática atendiendo a las siguientes ecuaciones:

$$u_{open}[k] = 100 \times \left(\frac{O[k-W+1]}{C[k-W]} - 1 \right) \tag{3.1}$$



Figura 3.2. *Construcción del input*

$$u_{high}[k] = 100 \times \left(\frac{max\{H[k-W+1] : H[k]\}}{C[k-W]} - 1 \right) \tag{3.2}$$

$$u_{low}[k] = 100 \times \left(\frac{max\{L[k-W+1] : L[k]\}}{C[k-W]} - 1 \right) \tag{3.3}$$

$$u_{close}[k] = 100 \times \left(\frac{C[k]}{C[k-W]} - 1 \right) \tag{3.4}$$

$$u_{volume}[k] = \frac{\sum_{i=k-W+1}^{k} \frac{V[k]}{W}}{\sum_{i=k-60+1}^{60} \frac{V[k]}{60}} - 1 \tag{3.5}$$

Donde :

- W: representa la ventana móvil empleada para construir las velas (corta, mediana y larga)
- $OHLCV$: representa el open, high, low, close y volume de una vela japonesa.

El resultado de aplicar estas fórmulas se puede visualizar gráficamente mediante las siguientes ilustraciones. La Figura 3.3 muestra la evolución del OHLCV para la ventana de 1 día construida parametrizando las ventanas en el horizonte temporal de la Figura 3.2. Por otro lado, la Figura 3.4 ilustra la evolución de la componente de cierre en función del tamaño de la ventana de la vela en cuestión.

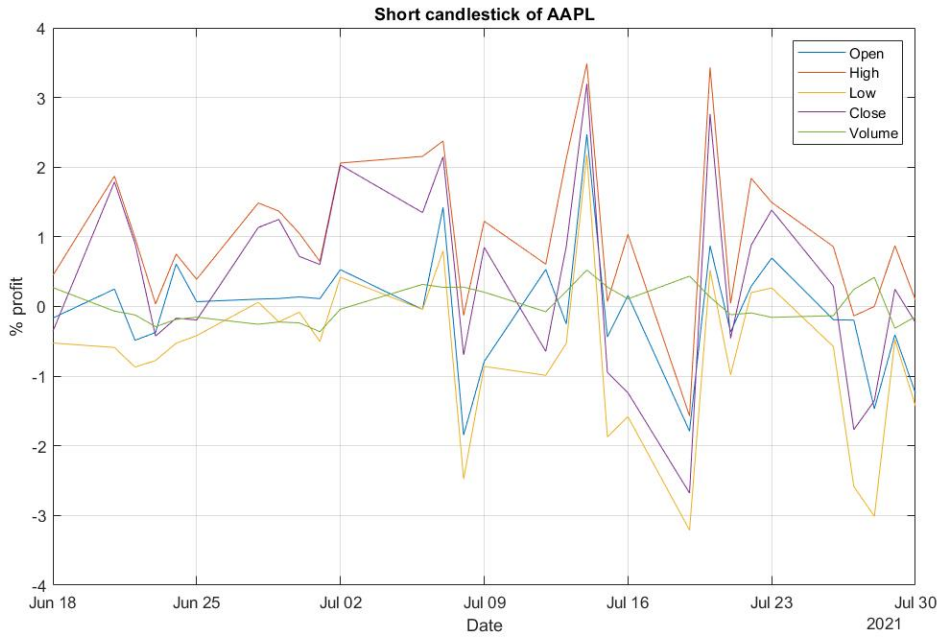


Figura 3.3. *Velas japonesas en la ventana corta para Apple*

Teniendo en cuenta todo lo anterior, las variables de entrada del modelo se construyen con un total de 15 coordenadas diferentes compuestas por el OHLCV calculado en tres horizontes temporales diferentes.

3.2. Implementación del algoritmo K-NN

Los modelos de predicción no dejan de ser herramientas que tienen como premisa fundamental el estudio de datos del pasado para la extracción de pautas de comportamiento del futuro. En este contexto surge la necesidad de simplificar la

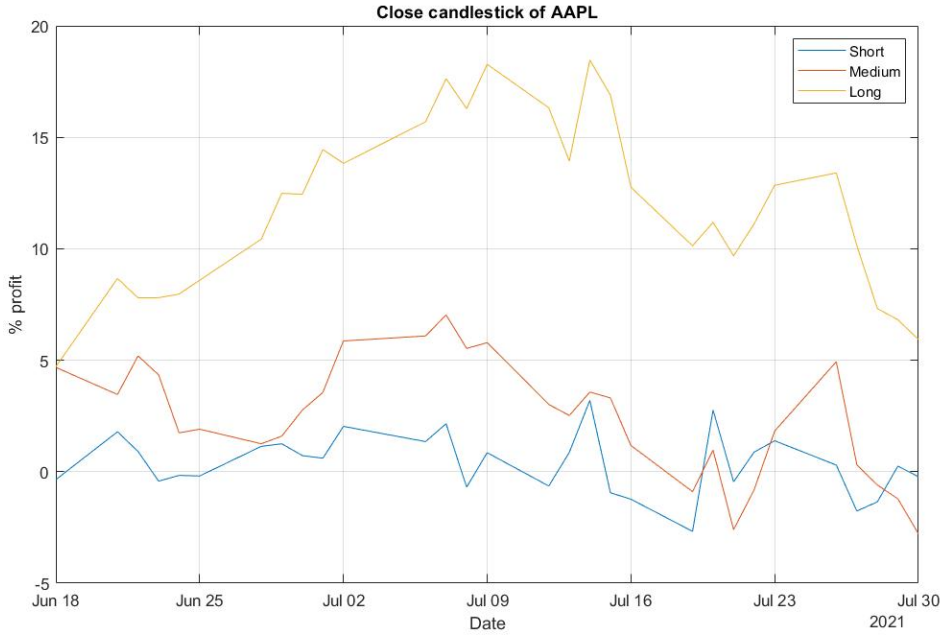


Figura 3.4. *Velas japonesas para el cierre del valor de Apple*

caótica realidad que impregna el mercado financiero a través de modelos que permitan capturar su esencia general, obviando los ruidos y perturbaciones. De este modo, resulta fundamental el desarrollo de modelos de predicción que encuentren un equilibrio entre el realismo y la complejidad del mismo, y que sean capaces de crear abstracciones fidedignas del dinamismo del mercado.

En este escenario, se pretende simplificar la realidad compleja del mercado bursátil a través de 15 variables y mediante ellas, generar una predicción sobre la rentabilidad futura. Se emplea el algoritmo K-NN por el equilibrio que manifiesta entre la robustez de las soluciones y su simplicidad. En este punto, resulta conveniente matizar que una predicción, en el ámbito bursátil, es considerada satisfactoria si contribuye a generar pautas coherentes respecto al comportamiento del mercado. Así pues, la predicción no debe de ser considerada un fin en sí mismo, sino un medio necesario para la generación de estrategias de inversión eficientes.

Intuitivamente, se trata de un algoritmo sencillo, sin embargo, su adaptación al problema en cuestión es lo que determinará su éxito. De la definición introducida en el capítulo anterior se pueden resaltar tres palabras fundamentales: training-set,

cercanía y predicción. Son precisamente estos tres parámetros los que se emplean para adecuar el algoritmo K-NN al ámbito financiero.

- En relación con el conjunto de entrenamiento, se utiliza una ventana deslizante de un año, que recorre el subperíodo seleccionado entre los 10 años disponibles. De este modo, se construye una ventana de 1 año hacia atrás que contiene la evolución temporal de las 15 coordenadas explicadas en el apartado anterior para aquellos valores, entre los 2555 disponibles, de los que se dispone de información suficiente. Así, el algoritmo busca los vecinos más cercanos para un valor determinado, entre un conjunto total de 643860 datos (252 días por 2555 [1] valores).

- En cuanto a la cercanía, hay que tener en cuenta dos factores principales: el número de vecinos más cercanos (K) y la distancia utilizada para evaluar cada punto. En cuanto a K, se calcula la distancia por pares considerando las siguientes opciones $[100, 150, 250, 500, 750, 1000, 1500, 2000]$ de vecinos. Además, se utiliza la distancia *Manhattan* como métrica, ya que utiliza el valor absoluto en lugar de la suma de cuadrados. Al tratarse de una normalización L_1, da menos peso a los outliers y es, por tanto, un indicador más robusto.

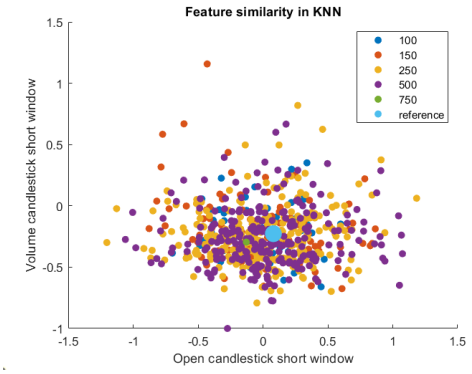
$$Distancia = \sum_{i=1}^{15} |x_i - y_i| \tag{3.6}$$

- Con respecto a la predicción, se toma como rentabilidad predicha la media y mediana de las rentabilidades de los K vecinos más cercanos calculados anteriormente. Además, también se calcula la tasa de acierto del algoritmo como el porcentaje de valores cuya rentabilidad predicha es positiva.

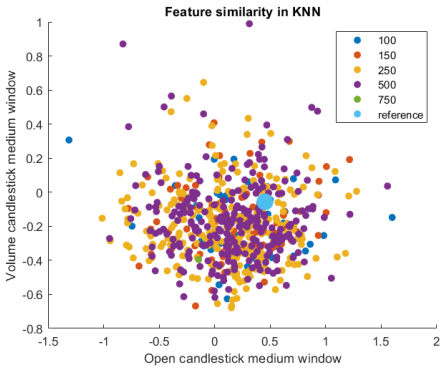
Una vez que se ha explicado de forma exhaustiva el método implementado, resulta interesante representar de forma gráfica cada una de los pasos mencionados anteriormente. En primer lugar, con las gráficas adjuntas en las Figuras 5.18a, 5.18b y 5.3c, se pretende ilustrar una representación de cómo el algoritmo K-NN computa los vecinos más cercanos. Para ello, se han proyectado 2 de las 15 componentes anteriores en 2-D atendiendo a dos criterios fundamentales: la proyección según una componente en concreto, y la proyección de acorde con una ventana deslizante en particular. El objetivo fundamental de este análisis es vislumbrar

[1] De los 2604 valores de partida se toman sólo los 2555 que siguen cotizando hasta la fecha actual

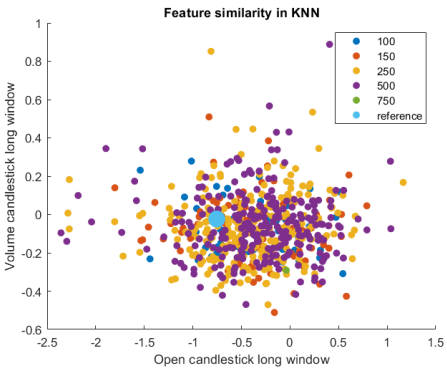
si existe alguna coordenada que tenga más peso en la asignación de los vecinos cercanos, así como, entender la distribución de la nube de puntos de los vecinos encontrados según el valor de K.



(a) Ventana corta



(b) Ventana mediana

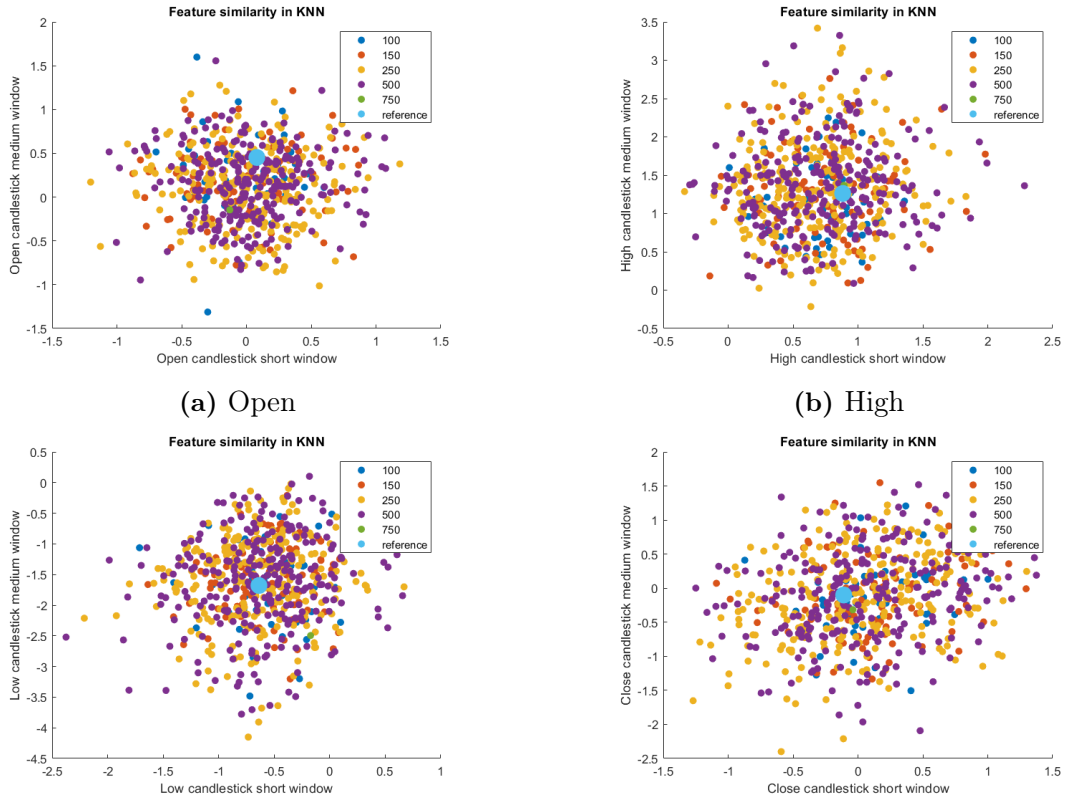


(c) Ventana larga

Figura 3.5. *Scatter plot de las componentes del open y volume para diferentes horizontes temporales*

La Figura 3.5 muestra la influencia de la ventana deslizante sobre el cálculo de los vecinos más cercanos. El punto de referencia denota la proyección del vector de 15 componentes a predecir en el subespacio mencionado ,y la nube de puntos simboliza los 100 a 750 vecinos más cercanos encontrados para dicho punto. Tal y como se puede observar en la primera gráfica (Figura 5.18a), para el caso de la ventana corta, el punto de referencia está en el centro, lo que implica que ambas componentes tienen el mismo peso para el cálculo de las distancias, pero la variabilidad de la componente del *open* es notoriamente superior que la del volumen.

Conforme se va aumentando el tamaño de la ventana deslizante, el punto de referencia deja de estar centrado, sin embargo, la variabilidad de las componentes se va ecualizando. Este fenómeno se discutirá en capítulos posteriores cuando se introduzcan las técnicas diseñadas para optimizar las entradas del modelo.



(a) Open **(b)** High

(c) Low **(d)** Close

Figura 3.6. *Scatter plot de los horizontes temporales corto y mediano para el OHLC*

Por otra parte, las gráficas adjuntas en la Figura 3.6 muestran el efecto del valor de cada coordenada del OHLC dadas dos ventanas deslizantes fijas. Con ello se pretende estudiar si, la variabilidad y el peso en el cálculo de los vecinos, también depende de cada componente del precio de la vela. Tal y como se puede observar por los límites de los ejes, ocurre un hito bastante curioso: el precio máximo de la sesión posee más variabilidad en torno al límite superior y el precio mínimo refleja el comportamiento diametralmente opuesto. Por otra parte, ninguno de los puntos de referencias mostrados parecen estar centrados, lo que pone de manifiesto que

hay componentes que tienen mayor importancia en el cálculo de los vecinos. Por último, merece la pena comentar que el comportamiento esperado de estas gráficas era una representación en forma de anillos con diámetros de tamaño creciente conforme se aumenta el número de vecinos, sin embargo, esto no se manifiesta en los gráficos anteriores. La hipótesis subyacente es que al ser un gráfico en 2-D, sólo se proyectan 2 de las 15 coordenadas disponibles y, por lo tanto, esto indica que alguna de ellas tiene una ponderación mayor en el calculo de la distancia.

Por último, en relación con la rentabilidad predicha por el modelo, resulta interesante adjuntar la Figura 3.7 en la que se comparan dichos beneficios con los reales. Se puede observar, que a lo largo de los 2 primeros años del periodo simulado, la rentabilidad predicha tiene una gran correlación negativa con respecto a la rentabilidad real. Esta es la razón fundamental por la que hay que implementar un método más sofisticado que comprar de forma ciega el mejor valor, como se discutirá en la última sesión del presente capítulo.

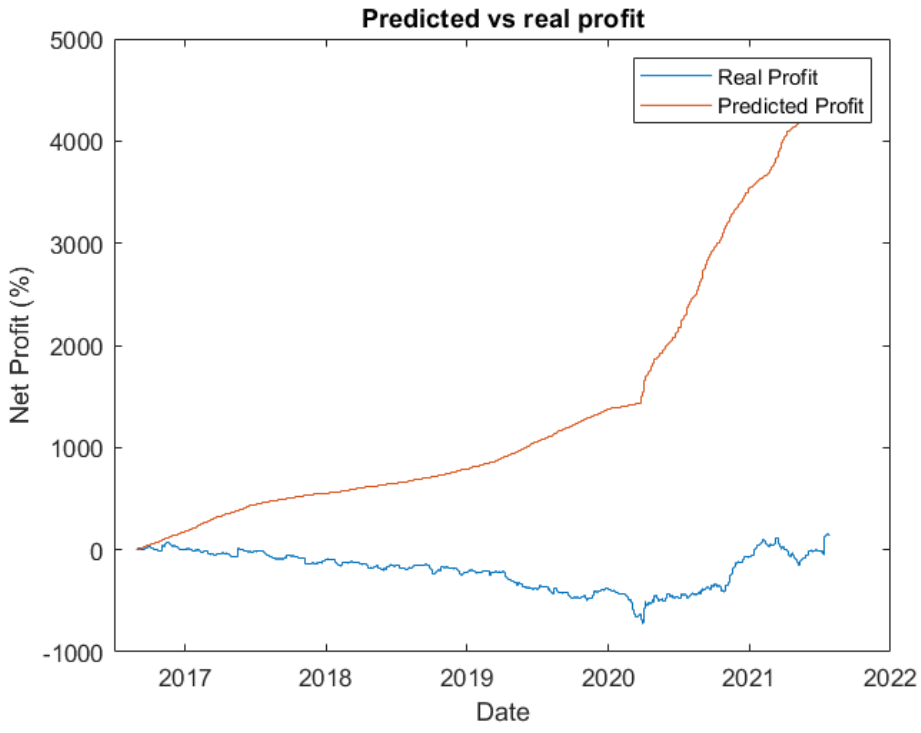


Figura 3.7. *Rentabilidad predicha vs. real*

En relación con la forma de codificar dicho backtesting, merece la pena comentar que se ha hecho un esfuerzo por habilitar la computación en paralelo para la simulación de cada uno de los días del periodo analizado. Para ello, se ha hecho uso del *Parallel Computing Toolbox* de *Matlab* [2] que soporta la computación paralela interactiva y permite acelerar el flujo de trabajo ejecutando distintas iteraciones en múltiples trabajadores de un pool paralelo.

Para su implementación, se ha empleado el comando *parfor* que se trata de un bucle que ejectuta una serie de sentencias en el cuerpo del bucle en paralelo. El cliente de Matlab emite el comando *parfor* y se coordina con los *workers* para ejecutar las iteraciones del bucle en paralelo. El cliente envía los datos necesarios sobre los que manda a los *workers*, donde se ejecuta la mayor parte del cálculo y estos envían de vuelta al cliente el resultado para poder ensamblarlo. Un bucle *parfor* puede proporcionar un rendimiento significativamente mejor que su bucle *for* análogo porque varios *workers* de Matlab pueden calcular simultáneamente operaciones dentro del mismo bucle. Cada ejecución del cuerpo es una iteración, y los *workers* evalúan las iteraciones sin ningún orden en particular y de forma independiente. Dado que cada iteración es independiente, no hay garantía de que las iteraciones estén sincronizadas de ninguna manera, ni hay necesidad de ello.

Para que la modificación implementada sea realmente efectiva es necesario disminuir al mínimo el *overhead*, es decir, la información que el cliente envía a cada *workers*, cómo se puede observar en la Figura 3.8. Esto supuso una reformulación completa del problema en cuestión y fue un reto muy importante a resolver. Gracias a ello, se consiguió reducir el tiempo diario de ejecución cerca de un 67 %, lo disminuye enormemente la carga computacional del algoritmo.

3.3. Criterio para la ordenación de los valores

Una vez que se tiene la predicción de la rentabilidad de los valores, se procede con la ordenación de los mismos. Para ello, se construyen tres índices diferentes compuestos por la tasa de acierto, la rentabilidad media y la rentabilidad mediana predicha. De este modo, se generan tres ordenaciones descendientes distintas según cada uno de los criterios anteriores. Con ello, se pretende garantizar la informatividad de los patrones, tal y como se revisó en el estado del arte, para evitar las

[2] https://es.mathworks.com/products/parallel-computing.html

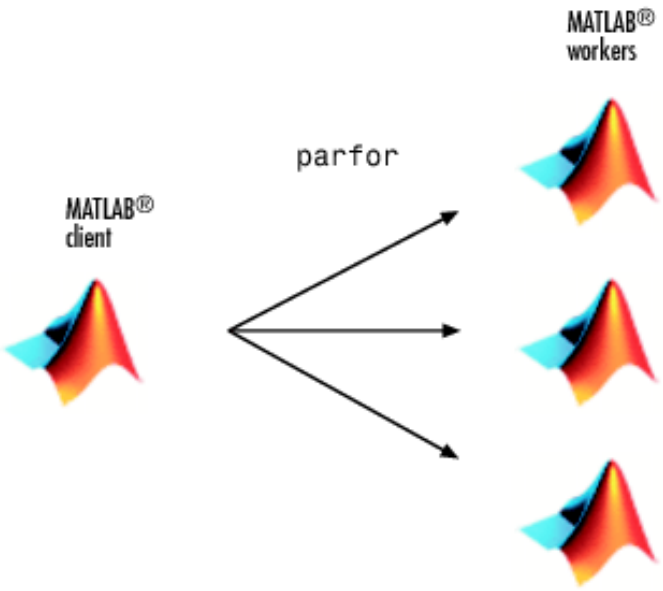


Figura 3.8. *Implementación de un bucle parfor en Matlab*

secuencias series temporales que se hayan generado debido a *random walks*

En las Figuras 3.9, 3.10 y 3.11 se incluyen el resultado de la clasificación para los 10 mejores valores según cada criterio para un día específico del periodo simulado.

A continuación, se construye un índice equiponderado a partir de los tres anteriores. Finalmente, se tiene por resultado la Figura 3.12 para cada día del periodo simulado. De este modo, se obtiene una ordenación de los valores diariamente que es lo que permite seleccionar las mejores acciones a comprar en cada sesión bursátil.

En vista a los resultados presentados se pueden resaltar una serie de hechos:

- Los límites superiores e inferiores del índice construido para la rentabilidad media son mayores que el computado para la rentabilidad mediana. Sin embargo, los valores finales del índice global no muestran unas diferencias tan significativas. Esto pone de manifiesto que, en términos prácticos, la ponderación de la rentabilidad media tiene un peso inferior que la mediana.
- La dispersión de valores es de 5.55 % para la tasa de acierto, 34.17 % para la mediana y 42.78 % para la media.

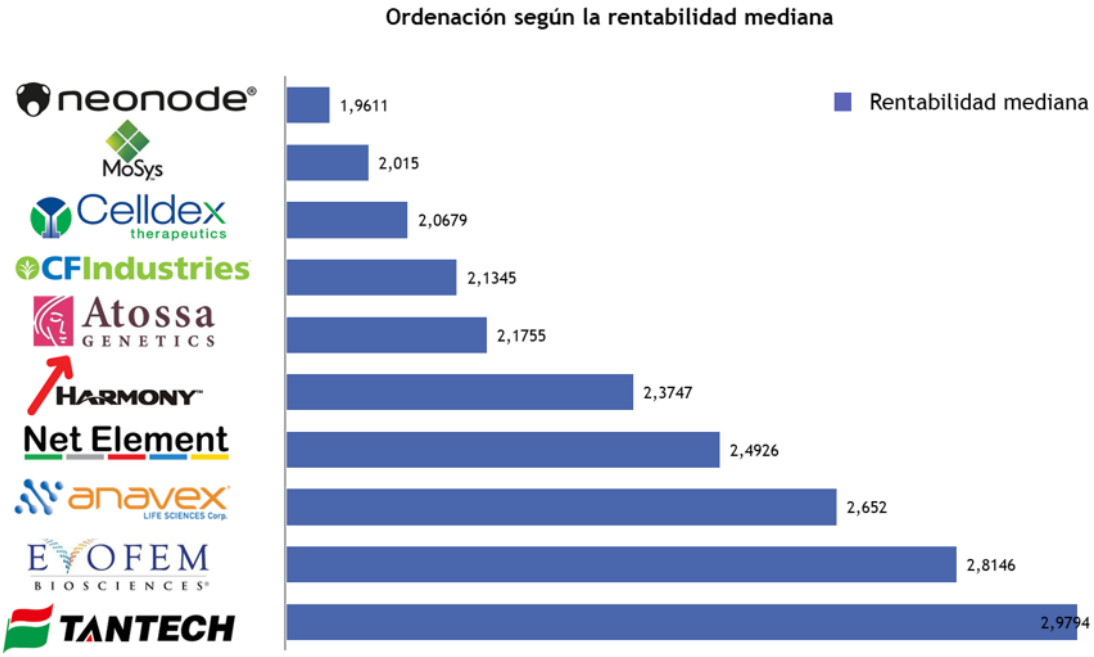


Figura 3.9. *Ordenación según la rentabilidad mediana*

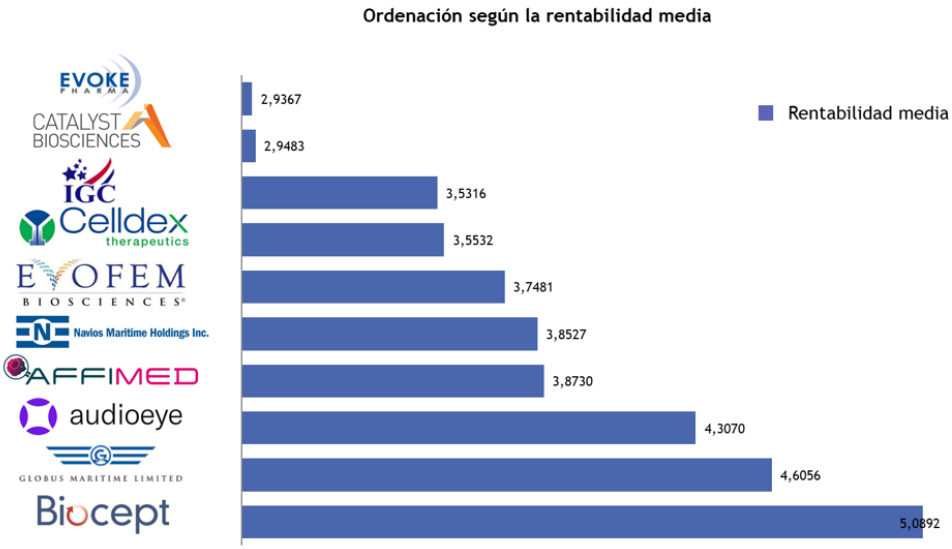
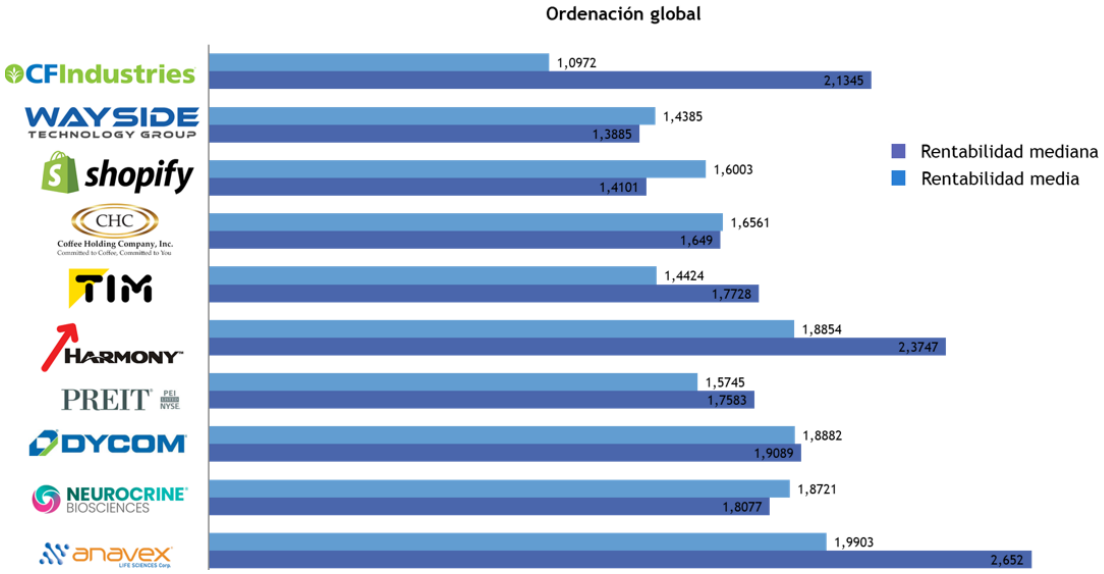


Figura 3.10. *Ordenación según la rentabilidad media*

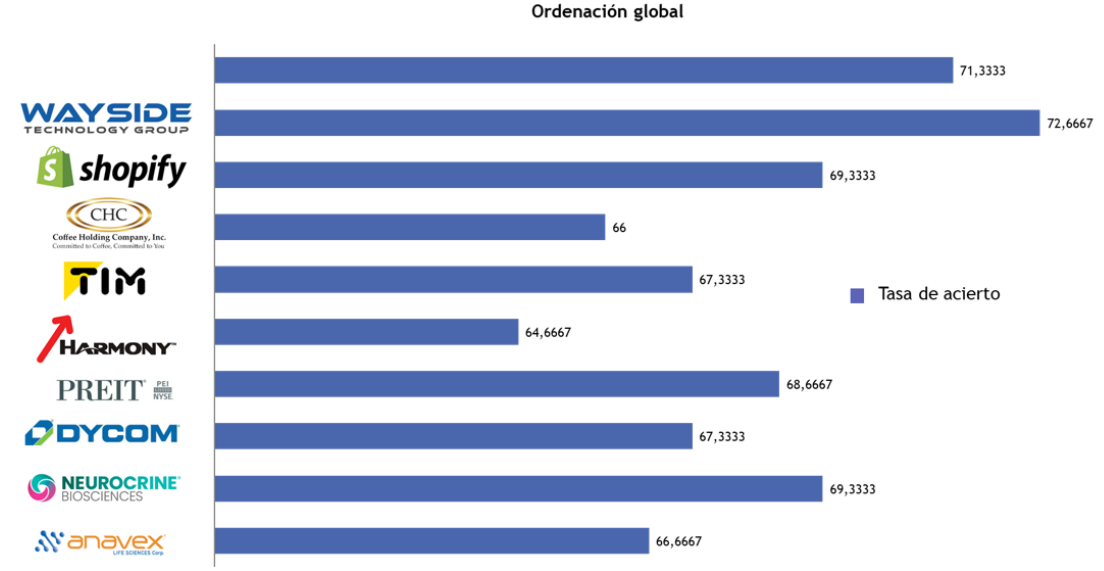


Figura 3.11. *Ordenación según la tasa de acierto*

- Con respecto a las compañías presentes en varias clasificaciones de forma simultánea, se puede afirmar que 5 de los 10 de los valores finales también lo están en el índice de la tasa de acierto y 4 de 10 de los presentes en el índice de la rentabilidad mediana también forman parte del índice global. Por contra, ningún valor del índice de la rentabilidad media pertenece simultáneamente al conjunto de los 10 mejores valores finales.



(a) Rentabilidad media y mediana



(b) Tasa de acierto

Figura 3.12. *Ordenación global*

A pesar de ser capaces de establecer juicios racionales en base a los resultados anteriores, se debería realizar un análisis estadístico más sofisticado para poder contrastar la validez de las hipótesis previamente señaladas.

Por último, la secuencia del cálculo del índice global se repite diariamente a lo largo del periodo simulado, dando lugar a una evolución de posiciones en el tiempo, como se puede observar en la Figura 3.13.



Figura 3.13. *Evolución de la posición*

3.4. Umbrales para generar las señales de compraventa

Como se ha señalado anteriormente, no basta únicamente con seleccionar al mejor valor cada día, sino que es necesario someterlo a una serie de restricciones para evitar sucesos como el de la Figura 3.7. Para ello, se establece que sólo se compra diariamente el mejor valor de la ordenación, si tanto la rentabilidad mediana predicha como la tasa de acierto, se encuentran dentro de unos rangos determinados por unos umbrales de corte. Por defecto, dichos hiperparámetros toman los valores representados en la Tabla 3.1. De esta forma, se pretende implementar una

protección similar a la que ofrecen la aplicación de los famosos *take-profit* y *stop-loss* en las estrategias de inversión.

Parámetro	Umbral superior	Umbral inferior
Tasa de acierto	90	63
Rentabilidad mediana	8	3,1

Tabla 3.1. *Umbrales para la generación de las señales de compraventa*

A continuación, se adjuntan en las Figuras 3.14 y 3.15 el resultado de aplicar dichos umbrales. Se representan las señales de compraventa como pulsos de amplitud similar al máximo rango observado en la serie temporal. Tal y como se puede observar, las operaciones se concentran en los dos primeros y en los dos últimos años.

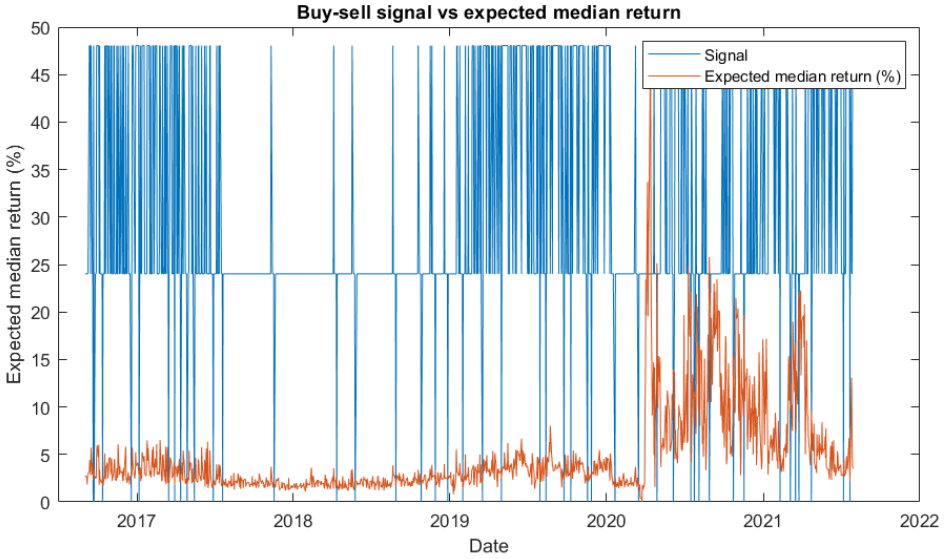


Figura 3.14. *Aplicación de los umbrales sobre la rentabilidad mediana*

Para verificar que se ha implementado de forma satisfactoria la estrategia deseada, se incluyen en las Figuras 3.16 y 3.17, las gráficas anteriores pero considerando un horizonte temporal menor. Se puede visualizar como el 18 de octubre de 2016 se produce una compra, ya que la rentabilidad media predicha es de 5.097 (menor

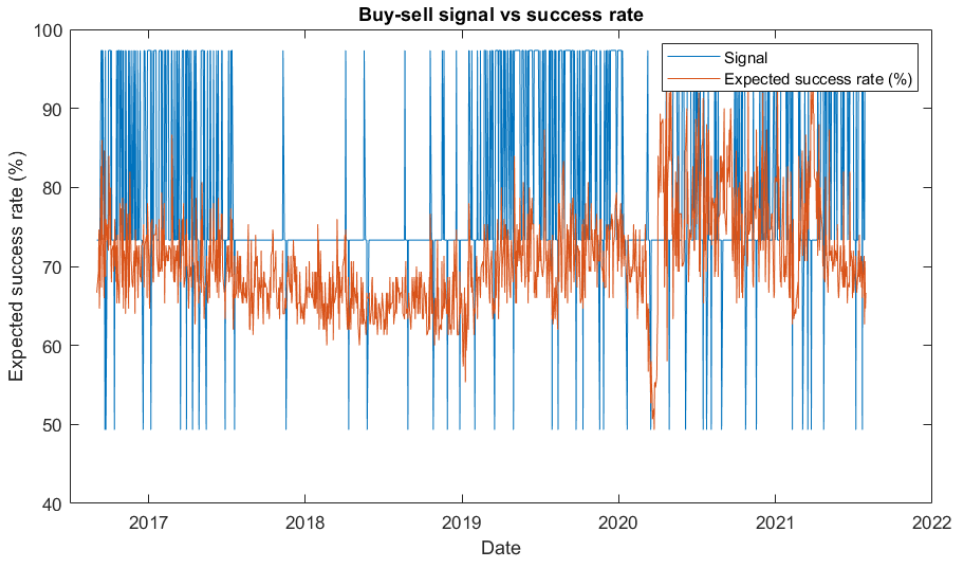


Figura 3.15. *Aplicación de los umbrales sobre la tasa de acierto*

que 8 y mayor que 3.1) y la tasa de acierto predicha es de 72.67 (menor que 90 y mayor que 63). Resulta muy interesante comentar como los umbrales de la tasa de acierto no son limitantes para el periodo seleccionado, por contra, si que lo es el inferior de la rentabilidad media predicha. Esto pone claramente de manifiesto como es necesario ir actualizando los valores de dichos hiperparámetros según las condiciones del mercado de forma que se adapten dinámicamente a las tendencias y volatilidades de cada periodo.

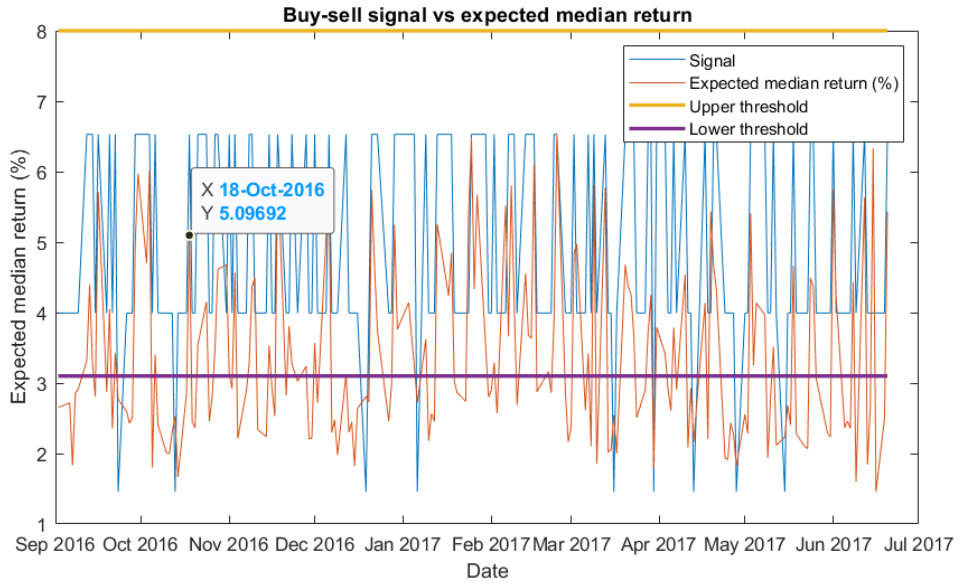


Figura 3.16. *Aplicación de los umbrales sobre la rentabilidad mediana con zoom*

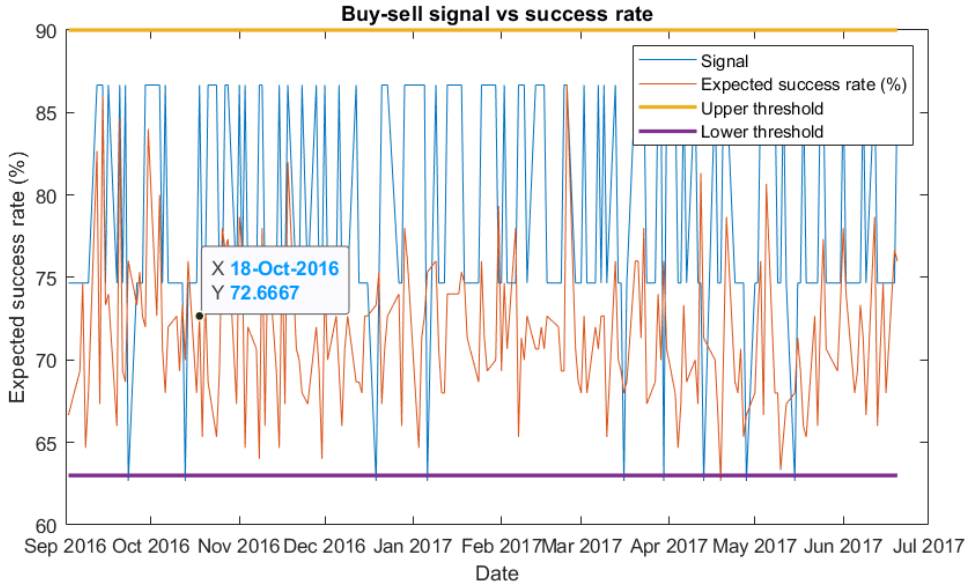


Figura 3.17. *Aplicación de los umbrales sobre la tasa de acierto con zoom*

Una vez que se han generado las señales de compraventa se puede calcular el beneficio derivado de las mismas. Dicho resultado se muestra en la Figura 3.18. Se representan con triángulos las transacciones efectuadas y se cuantifica la rentabilidad acumulada, asociada a dichas operaciones, mediante una línea continua azul. Para el cálculo de dicha rentabilidad se tienen en cuenta unos costes de transacción del 0.5 %. Como se puede observar, se alcanza una rentabilidad final anual del 23.4 %. Dicho número se obtiene aplicando la siguiente fórmula. Se establece que, en el peor escenario posible, el número máximo de posiciones abiertas mantenidas simultáneamente es de 5 compañías.

$$\text{Annual profit (\%)} = \frac{\text{accum profit (end)}}{\text{max capital invested } \times \text{ num years}} = \frac{573{,}1}{5 \times 4{,}9} = 23{,}4\,\% \tag{3.7}$$

A pesar de que la rentabilidad final obtenida es muy satisfactoria, no lo es su distribución a lo largo del tiempo, ya que dicha serie temporal manifiesta subidas muy acusadas a comienzos del año 2021.

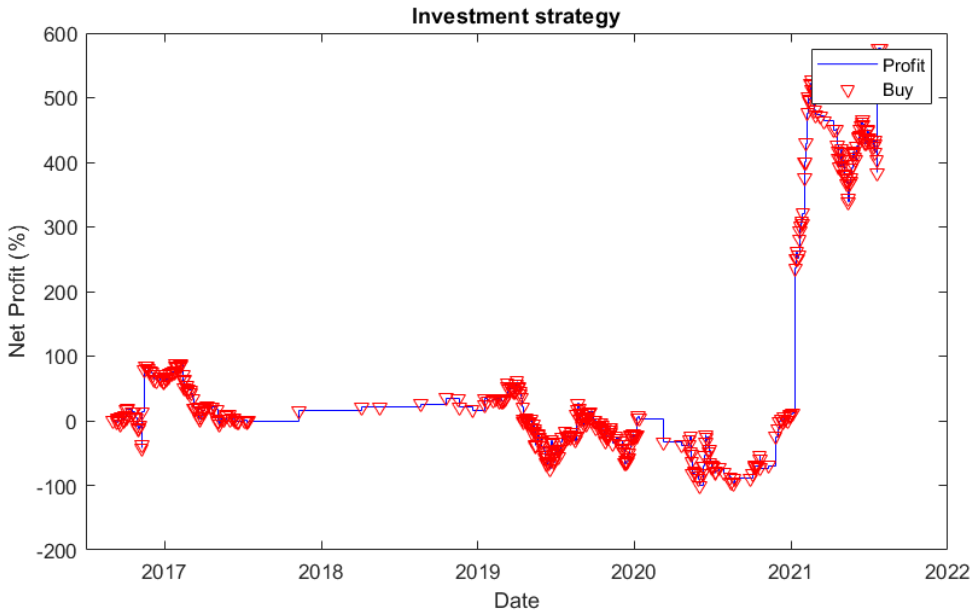


Figura 3.18. *Beneficio neto acumulado de la estrategia de inversión*

De hecho, el resultado de dicho aumento tan abrupto en la rentabilidad, es

fruto de una única operación, tal y como se puede observar en la Figura 3.19. Para remediar dicho fenómeno es necesario realizar un tratamiento de los outliers que se presentará en capítulos posteriores.

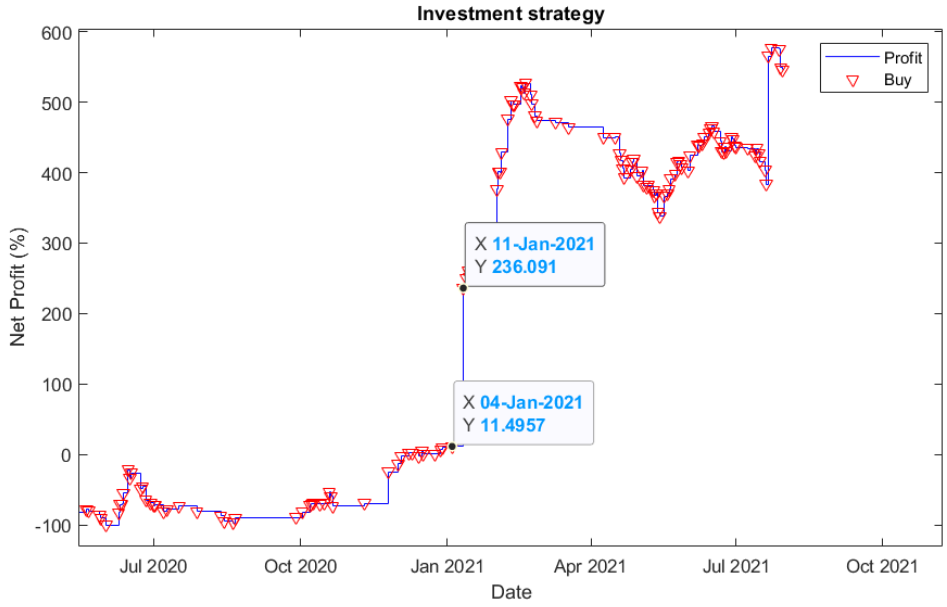


Figura 3.19. *Problema con la rentabilidad acumulada*

Para concluir con el presente capítulo, se adjunta en la Figura 3.20 un esquema explicativo de la estrategia de inversión propuesta.

Por último, se sintetizan a modo de conclusión, algunas reflexiones que se abordarán en futuras secciones.

- En relación con las entradas del modelo se ha podido observar, en la sección de la implementación del algoritmo K-NN, cómo no todas las coordenadas tienen el mismo peso en el cálculo de la rentabilidad. Por lo tanto, es necesario realizar un estudio más exhaustivo de las ponderaciones de las mismas. Esto se abordará en el capítulo de la optimización de la estrategia de inversión, concretamente en la sección de la optimización de las entradas del modelo.
- Con respecto a la comparación entre la rentabilidad real y la predicha, se ha podido notar ciertos periodos con correlaciones negativas, lo que supone un desafío para el desarrollo de las siguientes fases de la estrategia de inversión.

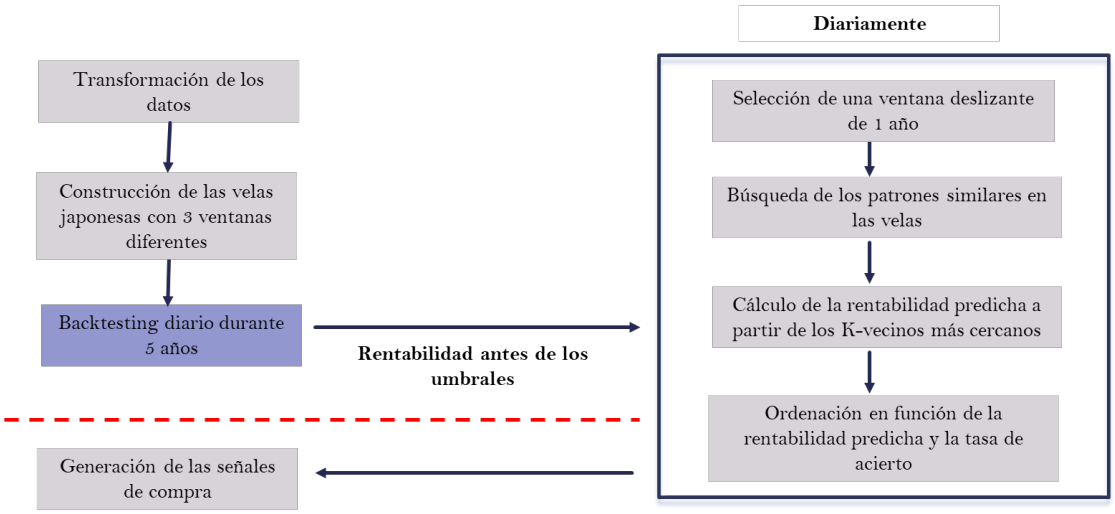


Figura 3.20. *Esquema de la estrategia de inversión propuesta*

- En términos de los índices generados para la ordenación de los valores, se ha comprobado que el construido a partir de la rentabilidad media no tiene apenas efecto sobre la ordenación global de los valores. Ello implica que dicho criterio puede ser despreciado frente a los otros dos.
- En relación con los umbrales corte, se ha puesto de manifiesto la necesidad de establecer unos umbrales dinámicos según las condiciones de mercado, ya que el hecho de delimitar en ocasiones con cotas fijas, no genera ningún efecto sobre las señales de compraventa. Dicho fenómeno será un tema de discusión en el capítulo del backtesting dinámico.

Capítulo 4

Diseño del backtesting dinámico

4.1. Motivación

En el año 2004 el comité de Basilea[1] adoptó el acuerdo de Basilea II cuyas implantaciones implicaron un notorio avance en la conquista de un sistema financiero estable y globalizado capaz de avocar el riesgo mediante el desarrollo de herramientas transparentes que sirvieran de control para las inestabilidades y perturbaciones económicas. Dicho consenso se trazó con la intención de establecer un marco integral capaz de estimular a las entidades bancarias y a los inversores para revisar y perfeccionar sus modelos evitando la generación de desequilibrios financieros.

Uno de los riesgos fundamentales que considera Basilea II es la participación en los mercados de renta variable. Desde el año 1999 los agentes del mercado comenzaron a asumir riesgos crecientes, elevando la cotización de las acciones hasta máximos nunca considerados. A partir del año 2000, el optimismo se esfumó al tiempo que el pánico comenzaba a hacer mella entre los principales inversores de-

[1]El Comité de Basilea o Comité de Supervisión Bancaria de Basilea (Basel Committee on Banking Supervisors) fue creado por los gobernadores de los bancos centrales de los países del Grupo de los Diez en 1974. Actualmente está compuesto por representantes de los bancos centrales de Bélgica, Canadá, Francia, Alemania, Italia, Japón, Luxemburgo, Países Bajos, España, Suecia, Suiza, Reino Unido y Estados Unidos. El Banco de España es miembro de pleno derecho desde 2001.

El Comité no tiene formalmente autoridad supervisora en el ámbito supranacional, por lo que sus conclusiones no tienen fuerza legal. No obstante, se trata de recomendaciones y directrices que, aún no siendo jurídicamente vinculantes, las autoridades de supervisión llevan a la práctica en su ámbito nacional, creando así una convergencia internacional sin necesidad de armonizaciones detalladas.

bido a las fluctuaciones y vaivenes del precio de los valores.

Sin embargo, dicho estado de alerta no se prolongó demasiado tiempo pues poco después los inversionistas volvieron de nuevo a la carga y el entorno financiero comenzó a recuperar su dinamismo, emprendiendo una etapa caracterizada por la clara tendencia al alza en el precio de las acciones. No obstante, durante el año 2008 llegó lo que los analistas vaticinaron como correcciones valorativas y generó el batacazo en el desinfle del precio de las acciones lo que arrastró pérdidas inconmensurables.

Tras esta crisis mundial se cuestionó la racionalidad y eficacia del criterio usado por los inversores analizando todos los movimientos y decisiones tomadas, sobre todo, aquellas relativas al riesgo asociado. En este sentido, se pusieron en tela de juicio las herramientas empeladas para el análisis del mercado de valores. Lo que evidencia este fenómeno, no es más que una de las muchas manifestaciones, que testifican la complejidad intrínseca asociada a los mercados financieros. Por ello, es necesario la construcción de algoritmos de inversión robustos que sean capaces de adaptarse con flexibilidad a las distintas situaciones del mercado.

Para ilustrar esta necesidad, se analiza un estudio en el que se presenta la interrogativa de si se puede llegar a ganar dinero en la Bolsa mediante la estrategia del cruce de medias móviles. Para ello, se utilizan como entradas las medias móviles calculadas para siete longitudes de ventanas diferentes y las siguientes acciones:

- Dos índices: el IBEX 35 y el DIJ.
- Un valor con gran volumen de transacciones: Telefónica.
- Un valor pequeño y con claro carácter especulativo: Telepizza.
- Un valor pequeño y no especulativo: Zardoya.
- Un tipo de interés: el Bund alemán.
- Un tipo de cambio: el del euro al dólar.

Se calculan los resultados para un inversor que hubiese operado diariamente con un sistema basado en el famoso cruce de medias móviles en un periodo de 8 años (1997-2005). En dicho desarrollo sólo se permiten emplear posiciones largas y, además, para otorgarle cierto realismo, se supone que el intermediario cobra unas comisiones del 0,20 % por cada transacción efectuada. Por último, merece la pena

comentar, que en este caso todos los hiperparámetros permanecían constantes en el intervalo considerado. En este sentido, el objetivo de la investigación consiste en determinar en qué supuestos la combinación de las ventanas resultan en una estrategia más rentable que la de comprar y mantener. Para representarlo, se opta por colorear de rojo las casillas en las que no se superan y, por el contrario, de verde aquellas que baten al mercado.

En primer lugar, se comienza presentando en la Figura 4.1 los resultados aplicados para el *IBEX 35*. Tal y como se puede observar, la estrategia de inversión propuesta está muy lejos de superar los rendimientos naturales del índice.



Figura 4.1. *Medias móviles para IBEX 35*

A continuación, se analiza con la Figura 4.2 el otro índice considerado. Aquí no hay ningún tipo de incertidumbre, los resultados revelan que no es aconsejable en absoluto aplicar técnicas basadas en medias móviles en este contexto.

Los resultados para Telefónica, mostrados en la Figura 4.3, parecen mejorar si se emplean un número de sesiones más reducido en la ventana corta.

En el caso de Telepizza, los resultados, presentados en la Figura 4.4 son claramente satisfactorios, quizás este fenómeno se deba al carácter especulativo que

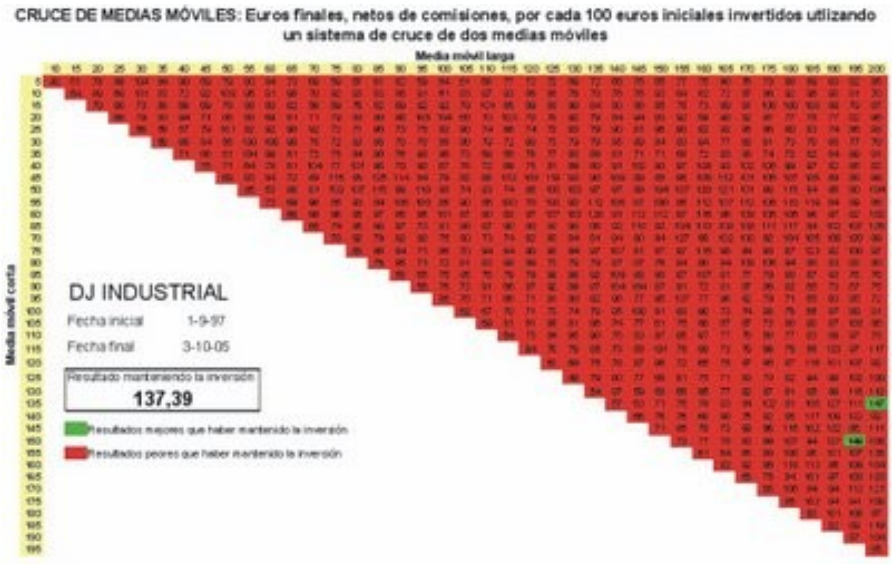


Figura 4.2. *Medias móviles para el DJI*

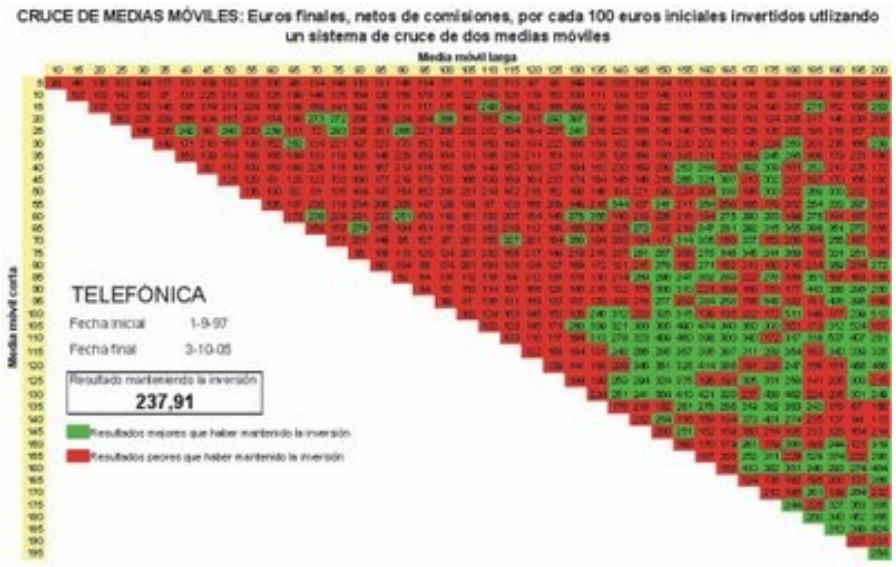


Figura 4.3. *Medias móviles para Telefónica*

impregna al valor.

En su contra, si se aplica la técnica al valor de Zardoya los rendimientos resultan ser pésimos, tal y como se puede observar en la Figura 4.5.

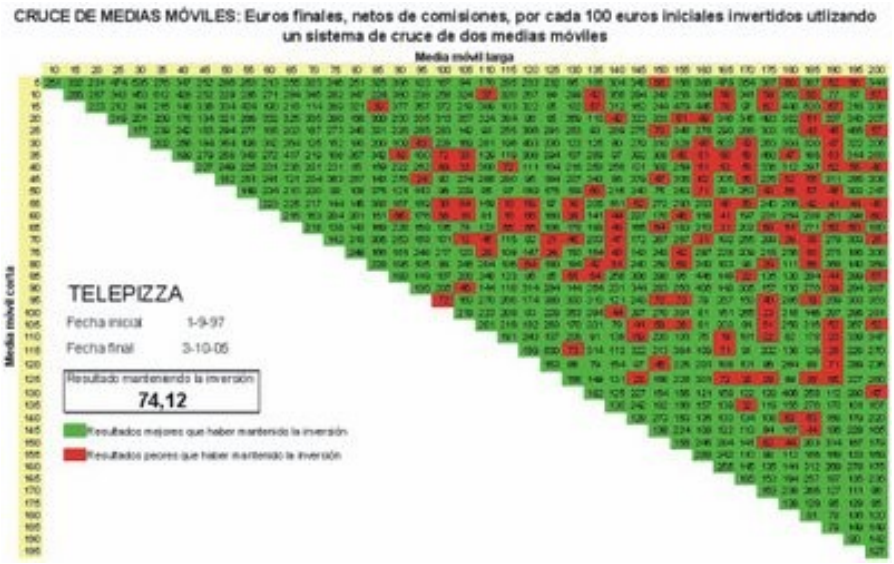


Figura 4.4. *Medias móviles para Telepizza*



Figura 4.5. *Medias móviles para Zardoya*

En el supuesto del bono alemán, como se ve en la Figura 4.6 las opciones de batir al mercado también son mínimas, a pesar de ser considerado un tipo de in-

terés que a largo plazo suele tender a oscilar en torno a un punto.

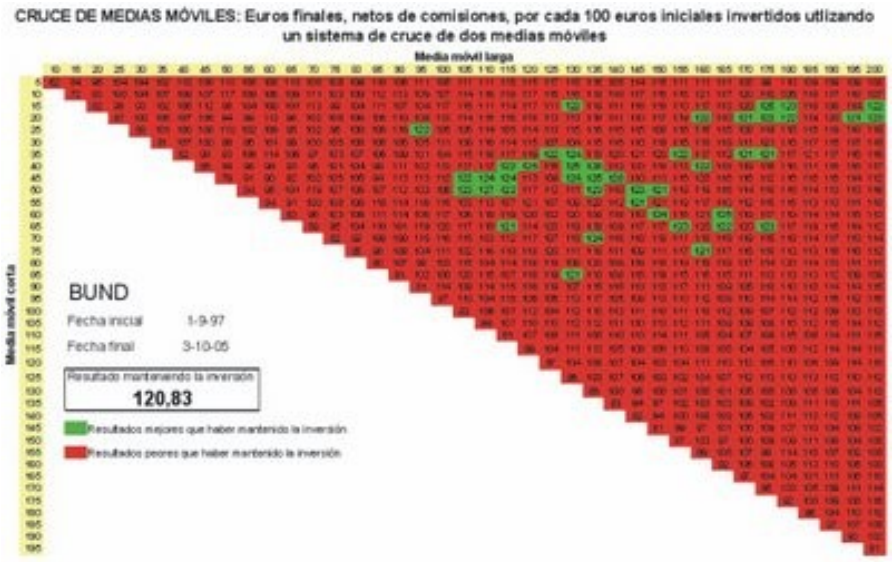


Figura 4.6. *Medias móviles para Bund*

Por último, se presentan en la Figura 4.7 los resultados tras la aplicación de esta técnica a la tasa de cambio entre el euro y el dólar. Los resultados obtenidos tampoco parecen revelar un gran avance en la aplicación de estas técnicas.

A raíz de este experimento se pueden extraer una serie de conclusiones. En general, usar técnicas basadas en medias móviles mediante la definición arbitraria de los parámetros, no conduce al diseño de estrategias de inversión eficientes. Por otra parte, no basta con la optimización de los parámetros en un intervalo de tiempo estático, sino que dicha optimización debe actualizarse cada cierta frecuencia, para así ir adaptándose a las fluctuaciones del mercado. Esto pone claramente de manifiesto la necesidad de implementar un backtesting dinámico para evaluar el rendimiento de la estrategia de inversión. En el caso de interés, los hiperparámetros del modelo no son las longitudes de las ventanas con las que se construyen las velas japonesas, puesto que se trata de un problema muy costoso en términos computacionales, sino los umbrales que permiten diseñar las señales heurísticas de compraventa.

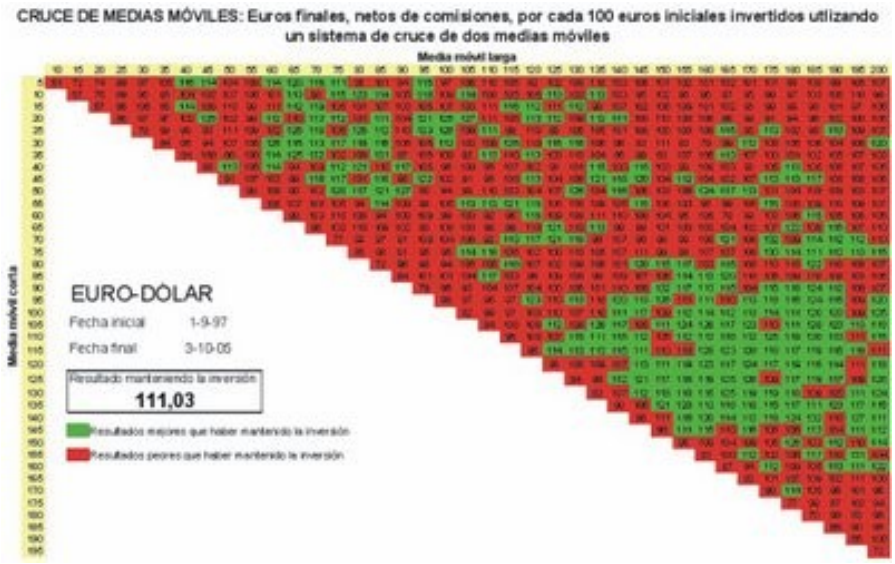


Figura 4.7. *Medias móviles para el Euro Dólar*

4.2. Presentación del paradigma actual

Como se ha discutido anteriormente, previo al desarrollo de este trabajo, se disponía de un backtesting que simulaba el histórico de operaciones que el algoritmo hubiera ejecutado. Como paso previo a la optimización dinámica de los parámetros, se diseñó un algoritmo genético, en adelante AG, que optimizaba los umbrales de corte de la estrategia de inversión, así como, el horizonte de predicción que determina la duración de la operación y el número de vecinos a considerar para el cálculo de la rentabilidad predicha.

Los algoritmos genéticos son métodos heurísticos de optimización basados en el proceso de búsqueda de una población de soluciones óptimas emulando la técnica de la evolución natural. Los dos pilares fundamentales de este proceso, la selección natural y la dinámica genética del más apto, inspiraron el desarrollo de este método. La diligencia del AG, como la de cualquier otro algoritmo de optimización, se basa en el mecanismo para equilibrar los dos objetivos básicos que se persiguen: explorar espacios de soluciones óptimas y simultáneamente, permitir la búsqueda de otros entornos prometedores. La eficacia del AG se centra en la fusión de estas dos funciones de manera óptima.

En general, se tratan de algoritmos basados en la dinámica de la selección na-

tural y en la mecánica de la genética humana. Dicha teoría tiene como premisa fundamental la supervivencia del individuo más apto y su forma de alcanzarlo reside en la adaptación del ente en su entorno. La implementación de los algoritmos genéticos se desarrolla bajo el esquema evolutivo defendido por el científico Charles Darwin, a través del cual, se crea una población inicial de individuos que son sometidos a una serie de transformaciones que combinan un intercambio de información estructurado con secuencias aleatorizadas para así fusionar en un único procedimiento una de las genialidades de las búsquedas humanas. Cada uno de estos ajustes lleva asociado una puntuación en una escala de bondad o *fitness* que será lo que determine si el gen ha de proliferar o no.

Se puede aplicar el AG a una gran variedad de problemas de optimización que no están adaptados a las condiciones de optimización clásicas (funciones no diferenciables, discontinuas o altamente no lineales.) Por otra parte, también es capaz de resolver problemas de programación de entero mixto en los que algunas de las variables están restringidas a tomar valores enteros. Esto fue un requisito fundamental en el trabajo desarrollado, pues, la mayoría de los hiperparámetros empleados debían de ser números discretos. Para simular un proceso evolutivo el algoritmo requiere de:

- Una población inicial de posibles soluciones debidamente representadas a través de individuos.
- Un proceso de selección basado en la aptitud de los individuos.
- Un proceso de transformación, es decir, un mecanismo de construcción de nuevas soluciones a partir de las disponibles en la población anterior.

Con todas estas cláusulas previas se pueden sintetizar, a modo de esquema, las etapas transitadas por el algoritmo durante la ejecución de cada una de las iteraciones:

1. Se genera la semilla aleatoria de una población inicial diversa, capaz de explorar todo el espacio del problema.
2. Se evalúa cada individuo que compone la población primaria.
3. Se seleccionan los genes que se usarán como padres para la evolución de la especie, así como, los "élite" que encabezan la generación actual. Para computar las puntuaciones se realizan dos pasos:

a) Se calcula el valor de la función de aptitud que asigna puntuaciones "crudas" del fitness de cada individuo

b) Se escalan dichas puntuaciones para acotarlas dentro de un rango más manejable, y, por último, se transforman en los "valores expectativos".

4. Se realiza el cruce y la mutación para actualizar la población.

5. Se comprueban los requisitos de parada del algoritmo. Si se cumplen, se da por finalizado el proceso, de lo contrario, se retoma el punto 2.

Para entender mejor el comportamiento del algoritmo se adjunta en la Figura 4.8 un esquema del funcionamiento del mismo.

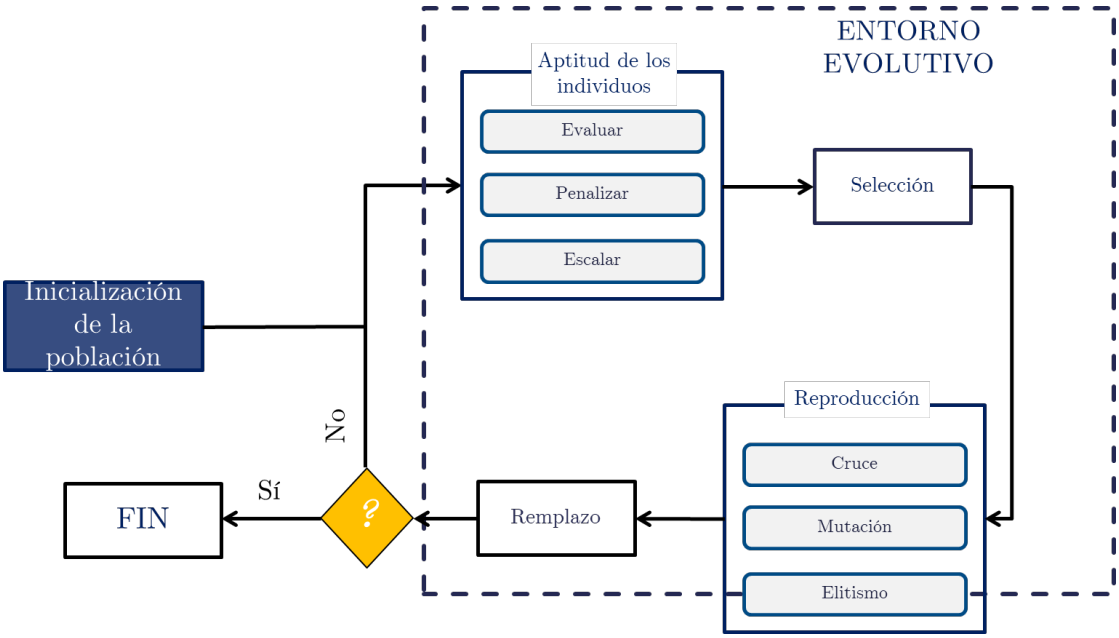


Figura 4.8. *Esquema evolutivo del AG*

Uno de los factores predominantes que determinan el rendimiento del algoritmo pasa por la selección correcta de la población inicial. Se pueden definir distintos parámetros con multitud de modalidades cuya manipulación puede desempeñar un papel fundamental en la convergencia del AG. En primer lugar, es necesario establecer la codificación a utilizar. Cada individuo que conforma la población viene representado por un cromosoma que, a su vez, está compuesto por tantos

genes como variables tenga el problema. Con respecto a la función objetivo, el AG diseñado maximiza la rentabilidad anual en el periodo analizado, no obstante, sería interesante incluir alguna medida de riesgo como el máximo *drawdown* de la estrategia.

Por otra parte, un factor crucial que puede llegar a garantizar el éxito del AG es la diversidad de la población. Si la distancia promedio entre los individuos es considerable, entonces se favorece la pluralidad de los genes. Este parámetro se puede controlar delimitando el rango inicial de la población, así como, los umbrales máximos y mínimos que pueden tomar las variables implicadas. En este sentido, si se conoce intuitivamente de antemano valores que son impracticables o que se encuentran muy alejados de la solución, conviene restringir aquellas variables involucradas, para así evitar que el AG explore zonas prohibidas y propiciar una búsqueda más profunda en aquellas que susciten mayor interés. Para ello, se limitaron los valores aplicando los umbrales inferiores y superiores adjuntos en la Tabla 4.1

Hiperparámetro	**Umbral superior**	**Umbral inferior**
Umbral superior de la tasa de acierto	90	80
Umbral inferior de la tasa de acierto	70	60
Umbral superior de la rentabilidad mediana	15	8
Umbral inferior de la rentabilidad mediana	1	7
Índice del número de vecinos [2]	8	1
Horizonte de predicción	5	1

Tabla 4.1. *Umbrales para los hiperparámetros del AG*

Para evaluar el rendimiento del AG se adjunta en la Figura 4.9 la evolución del AG. La figura superior representa el mejor y el valor medio de la función *fitness* para cada generación. En cuanto a la inferior, se muestra la distancia media entre los individuos de la población. Conforme dicha métrica tiende a 0, más cerca está el AG de lograr la convergencia. Llama la atención el peculiar aspecto escalonado que tiene la figura superior, esto se debe fundamentalmente a que el AG comienza explorando distintas zonas del espacio, y cuando encuentra en alguna de ellas un individuo muy apto, trata de explotarlo ocasionando un pulso descendente en la función objetivo.

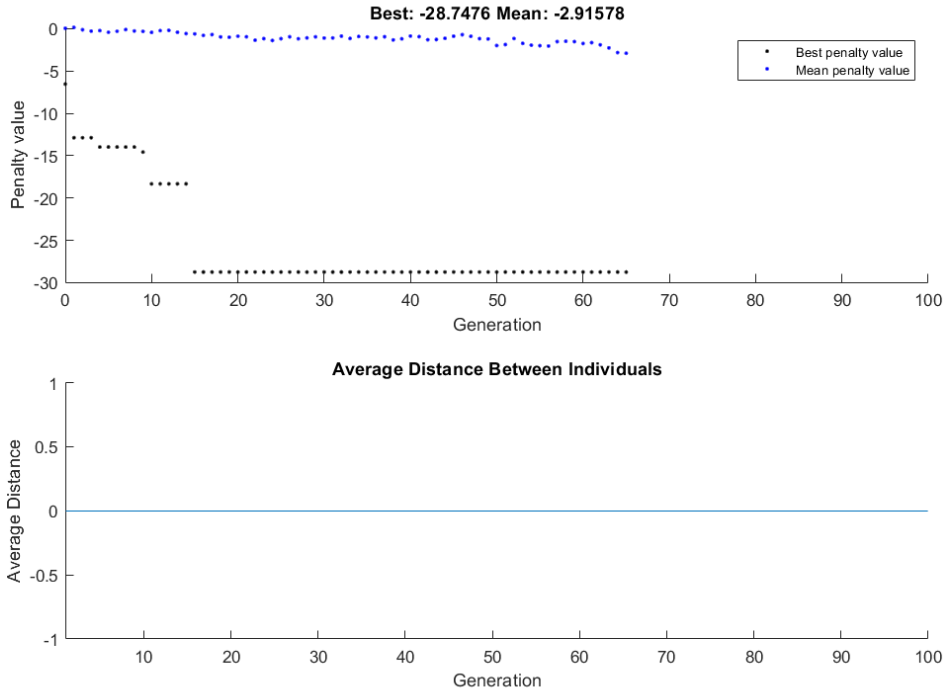


Figura 4.9. *Evolución del AG*

Por otra parte, con el fin de analizar cómo afecta la optimización de los hiperparámetros a la estrategia, se incluye en la Figura 4.10 una gráfica del rendimiento neto alcanzado para los parámetros óptimos. Se divide el periodo analizado en dos, el primero para entrenar el algoritmo en cuestión y el segundo para validar la estrategia. Se representan mediante diferentes colores el resultado final obtenido en cada subperíodo. Tal y como se puede observar, el hecho de optimizar para un intervalo temporal determinado hace que la estrategia de inversión se sobreajuste a dichas condiciones del mercado. De hecho, no es nada robusto ni reproducible que la rentabilidad final del modelo dependa fundamentalmente de las operaciones efectuadas a comienzos del año 2019.

En línea con el pensamiento anterior, se adjunta en la Figura 4.11 un análisis de correlaciones llevado a cabo año a año. A continuación, se explica brevemente lo que representan cada una de las cuatro gráficas de cada simulación [3]:

[3]No se incluyen las gráficas relativas al año 2018-2019 ya que el algoritmo no efectúa ninguna transacción

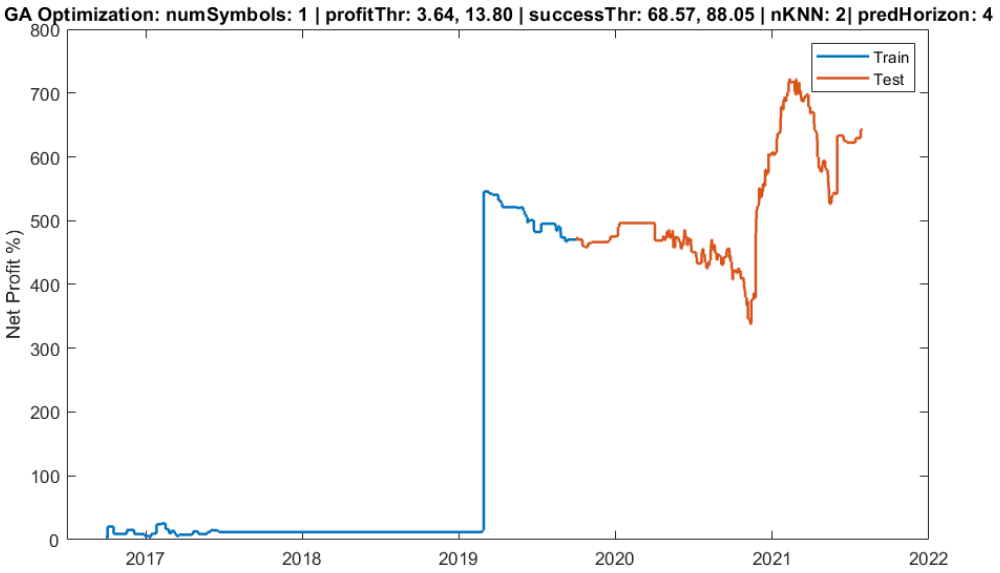


Figura 4.10. *Rendimientos netos tras la optimización de los hiperparámetros*

1. Las figuras superiores muestran la rentabilidad real frente a la rentabilidad predicha. Además, también se computa el grado de correlación en cada caso.

2. La figura de la esquina inferior izquierda representa la tasa de acierto predicha frente a la rentabilidad real.

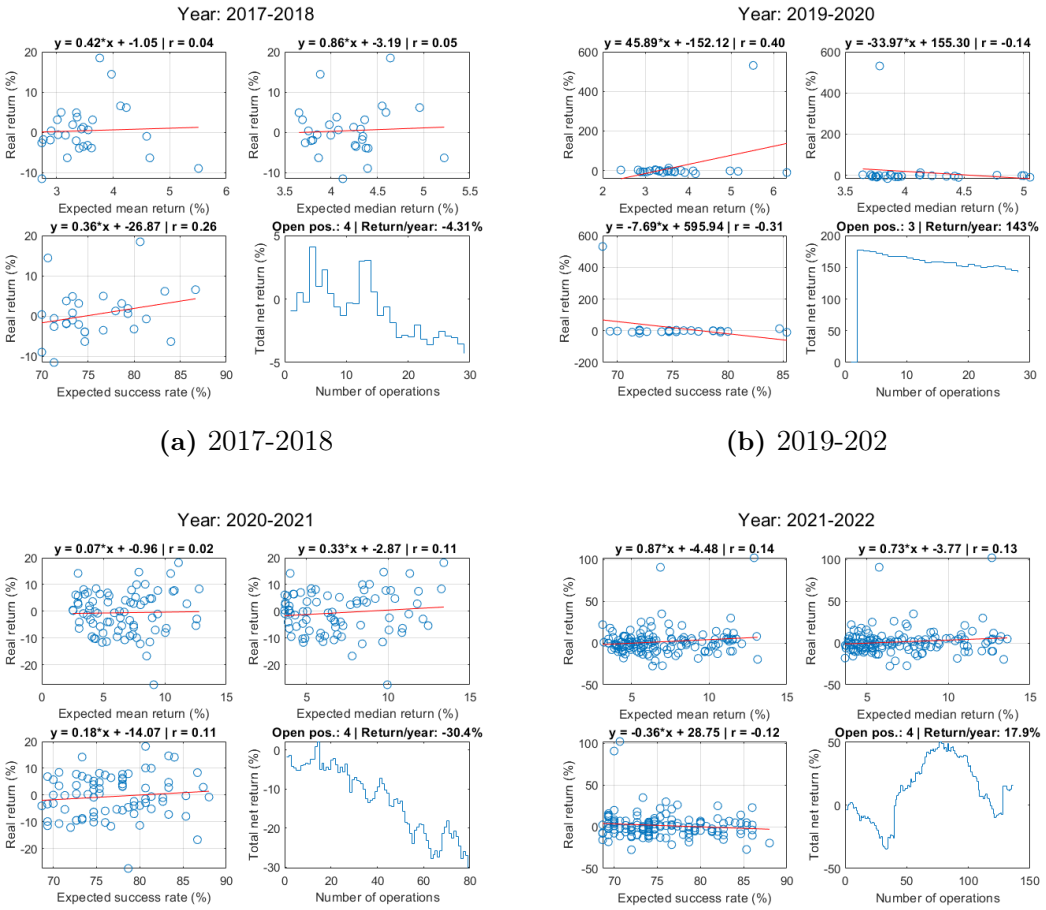
3. La gráfica de la esquina inferior derecha muestra los beneficios acumulados tras cada operación.

De forma intuitiva, el comportamiento idóneo de cada una de dichas gráficas sería el siguiente:

- De las gráficas de la rentabilidad predicha frente a la real, se espera una correlación ligeramente positiva.

- De la gráfica que compara la tasa de acierto frente la rentabilidad real, lo que parece razonable es que a mayor fiabilidad en la predicción de la rentabilidad, mayor sea la rentabilidad real de dicha transacción.

- Por último, de la figura de la esquina inferior derecha se espera que la concentración de las operaciones sea homogénea en el tiempo y que la curva muestre cierta tendencia monótona creciente.



(a) 2017-2018 **(b)** 2019-202

(c) 2020-2021 **(d)** 2021-2022

Figura 4.11. *Análisis de las correlaciones por año*

A partir de los resultados obtenidos se pueden realizar una serie de reflexiones. En primer lugar, se pone claramente de manifiesto que el hecho de optimizar los hiperparámetros para un subperíodo determinado y mantenerlos fijos en el tiempo no conduce al desarrollo de estrategias de inversión reproducibles en el tiempo. De aquí surge la necesidad de diseñar un backtesting dinámico. No obstante, merece la pena comentar una serie de sucesos interesantes que pueden ayudar a entender mejor el funcionamiento del algoritmo implementado.

- En primer lugar, se puede observar cómo en la Figura 4.11b, relativa al año 2019-2020, la correlación entre la rentabilidad media y la real es positiva, mientras que la presente entre la rentabilidad mediana y la real es negativa.

Esto revela que dicho efecto positivo se debe a la presencia de valores atípicos, tal y como se puede observar en la gráfica del rendimiento con la subida tan abrupta.

- Con respecto a esa misma gráfica (Figura 4.11b) merece la pena comentar que la correlación entre la tasa de acierto y la rentabilidad real es la más negativa de las 4. Esto es un indicio adicional de que las predicciones efectuadas no son fidedignas.

- Resulta muy curioso cómo, a pesar de que las gráficas de los últimos dos años tienen prácticamente la misma correlación entre la rentabilidad mediana y la real, una presenta rendimientos negativos y la otra no. En el caso del periodo del año 2020-2021 (Figura 4.11c) se tiene una correlación positiva en la tasa de acierto, pero prácticamente nula en la rentabilidad media. Por contra, en el caso del año 2021-2022 (Figura 4.11d) se obtiene una correlación ligeramente positiva en la rentabilidad media, pero negativa en la tasa de acierto. Esto puede ser un indicador de que la tasa de acierto no es limitante en cuanto a la generación de las señales de compraventa, como ya se había anticipado en el capítulo anterior.

- En línea con el pensamiento anterior, en la primera gráfica (Figura 4.11a) se tienen correlaciones prácticamente nulas en las rentabilidades predichas, pero positiva en la tasa de acierto, sin tener ningún impacto en el resultado anual alcanzado.

De forma adicional, sería interesante repetir el experimento analizado pero quitando los outliers de la rentabilidad del periodo de entrenamiento (con el empleo de una media móvil centrada o computando los *rolling returns* [4]). Para concluir esta sección, se ha demostrado experimentalmente que es necesaria una optimización dinámica de los parámetros de la estrategia de inversión para que el modelo sea capaz de ajustarse a las fluctuaciones del mercado. De lo contrario, el método diseñado pierde la capacidad de abstracción y no genera los rendimientos deseados en situaciones futuras, ya que los datos del entrenamiento no son suficientemente significativos.

[4]Esto será foco de discusión del Capítulo 6 de resultados

4.3. Explicación de la solución adoptada

El objetivo fundamental de esta sección es presentar cómo se ha diseñado el backtesting dinámico que elimina completamente los sesgos procedentes del método anterior. Para ello, se hace uso de dos conceptos claves:

- Frecuencia de actualización: se trata de un parámetro que sirve como *trigger* para la actualización de los parámetros del modelo.
- Memoria de los datos: consiste en una ventana deslizante que agrupa los datos más recientes contenidos en ella. Desempeña el papel de un búfer de información.

De este modo, la secuencia implementada es la siguiente. En primer lugar, se parte de los hiperparámetros preestablecidos y, cuando se activa la bandera de actualización, se toman los datos más recientes, generados a partir de la ordenación del K-NN, en una ventana deslizante hacia detrás para optimizar los hiperparámetros en dicha ventana, tal y como se puede observar en la Figura 4.12. En ella, se representa la frecuencia de actualización a través de líneas rojas verticales y la ventana deslizante mediante rectángulos verdes. De este modo, en este escenario, el algoritmo actualizaría los parámetros cada año y emplearía un búfer de medio año para optimizar los valores. Posteriormente, se computa la estrategia de inversión con dichos valores hasta que se alcance el próximo periodo de frecuencia establecido, como se muestra en la Figura 4.13. En esta gráfica se muestran con líneas horizontales de diferentes colores los umbrales de corte que se obtendrían en la optimización del periodo anterior. Es decir, los primeros, los rojos, hacen referencia a los valores tomados por referencia y, por ejemplo, los verdes, son los que se obtienen tras completar la optimización cuando salta el trigger en el tercer cuatrimestre del año 2017.

De este modo, bajo el actual paradigma se distinguen dos backtestings que actúan en niveles diferentes

- En primer lugar, el que se emplea en la capa inferior, calcula para cada día la rentabilidad predicha del K-NN en un horizonte de predicción fijo. Para ello, toma para cada instante temporal una ventana deslizante de un año para encontrar los patrones de las velas japonesas más similares al valor a predecir, computa la rentabilidad predicha en función de la rentabilidad de los vecinos más cercano y establece una ordenación de los valores de acuerdo

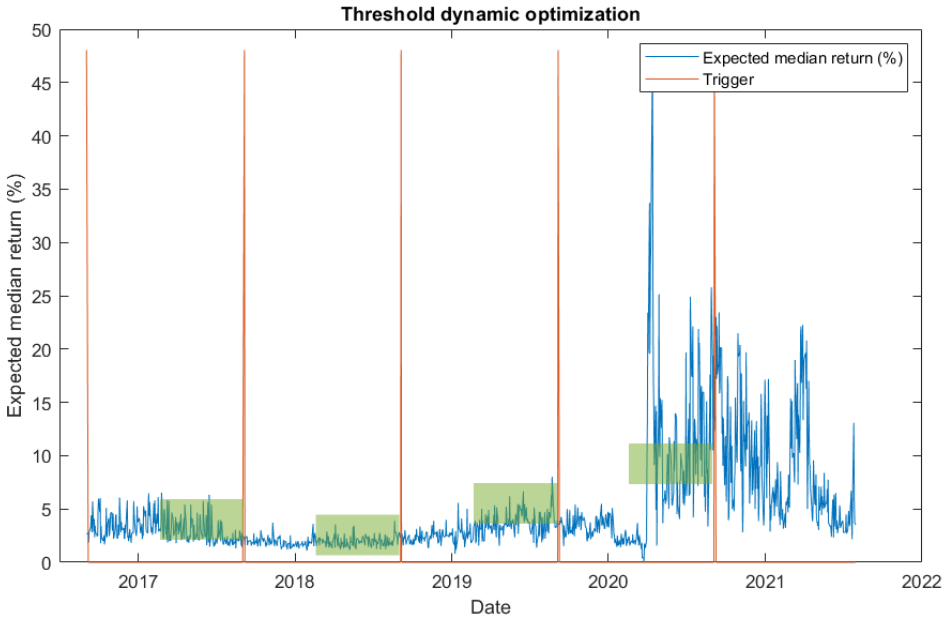


Figura 4.12. *Frecuencia de actualización y memoria de repetición para la optimización dinámica*

con el triple criterio de la tasa de acierto y la rentabilidad media y mediana. Dicho backtesting se ha implementado haciendo uso de las técnicas de la computación en paralelo.

- En segundo lugar, el backtesting de la capa superior, toma la información histórica generada por el otro y ejecuta las señales de compraventa para así determinar los instantes temporales en los que se compra el mejor valor y poder calcular la rentabilidad final del método.

A continuación, se representa en las Figuras 4.14 y 4.15 las fases seguidas por cada uno de ellos. La razón fundamental de desacoplar ambos backtestings es permitir la paralelización del backtesting que actúa en la capa inferior para así poder reducir el coste computacional. Además, esto tiene como ventaja adicional que se puede obtener una estimación de la rentabilidad bruta del algoritmo K-NN sin tener que realizar el segundo backtesting.

Una vez que se ha entendido de forma general el método implementado, conviene discutir en detalle las particularidades del mismo. Para ello, se explicará de

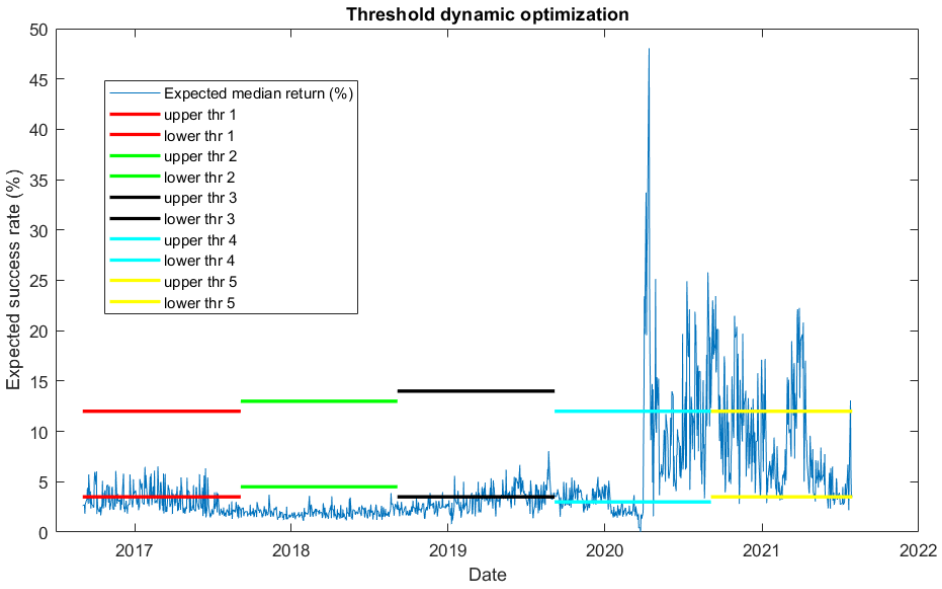


Figura 4.13. *Actualización de los umbrales dinámicos en la optimización*

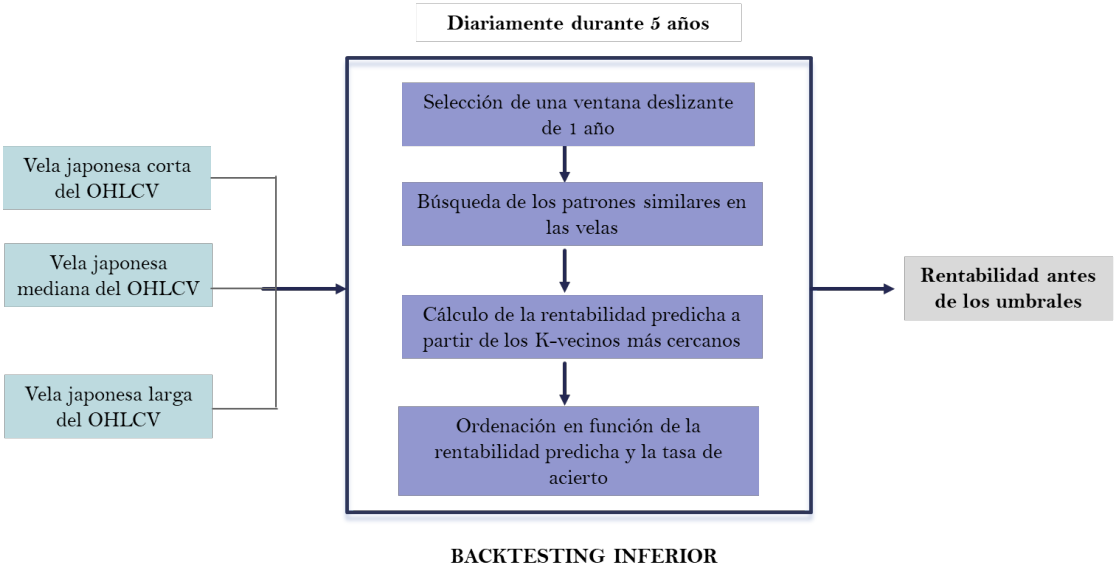


Figura 4.14. *Esquema del backtesting inferior*

forma exhaustiva las implicaciones derivadas de la elección de los hiperparámetros a seleccionar, el método de optimización y las principales dificultades abordadas.

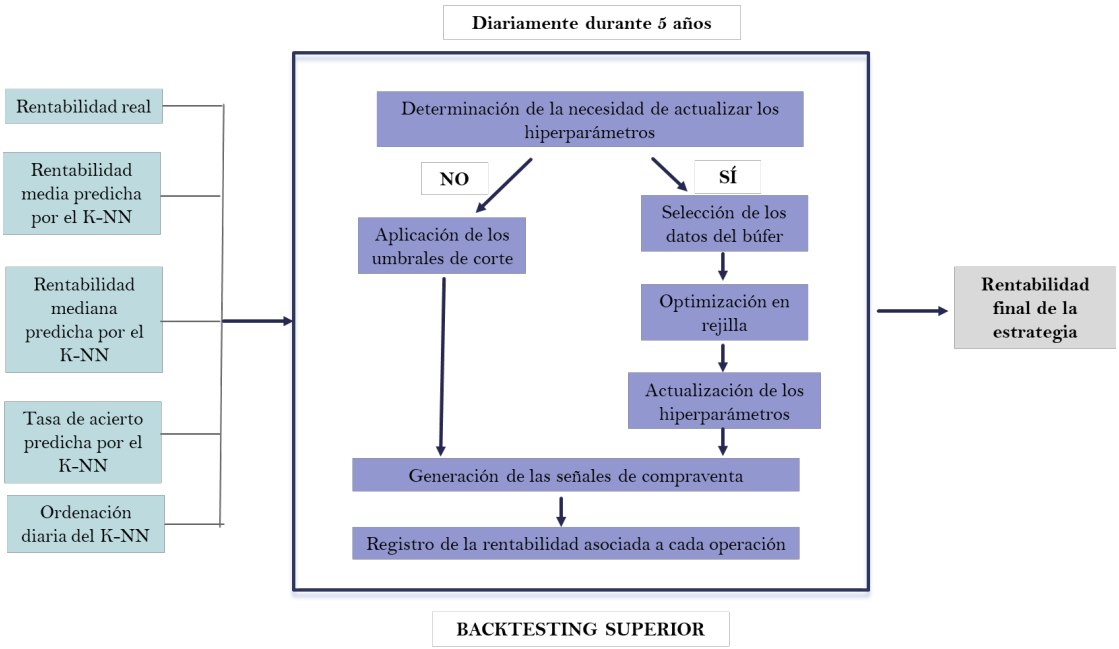


Figura 4.15. *Esquema del backtesting superior*

4.3.1. Hiperparámetros a optimizar

Con respecto a los hiperparámetros a optimizar, es muy importante entender por qué se limita el foco a los umbrales de corte de la generación de las señales de compraventa. La razón fundamental de ello es que para optimizar otros parámetros, como pueden ser el número de vecinos o el horizonte de predicción, es necesario ejecutar de nuevo todo el backtesting del algoritmo del K-NN, que tarda aproximadamente unas 8 horas, lo cual hace inmanejable probar cada combinación. En capítulos posteriores se analizará cómo se pueden afinar mejor dichos hiperparámetros.

Por lo tanto, una vez entendida la necesidad de simplificar el problema en cuestión, conviene presentar el rango de valores que se utilizan en ambos casos, adjuntos en la Tabla 4.2. Como nota aclaratoria, la columna *salto* hace referencia a la longitud de paso del vector.

Hiperparámetro	Umbral superior	Umbral inferior	Salto
Umbral superior de la tasa de acierto	90	80	2
Umbral inferior de la tasa de acierto	70	60	1
Umbral superior de la rentabilidad mediana	15	12	1
Umbral inferior de la rentabilidad mediana	1	5	0,5

Tabla 4.2. *Umbrales para los hiperparámetros del backtesting dinámico*

4.3.2. Método de optimización

En relación al método de optimización, se emplea una búsqueda en rejilla o *grid search* del mejor valor de la rentabilidad anual. Para ello, se prueban todas las combinaciones de los hiperparámetros posibles y se guarda en memoria el mejor valor encontrado hasta el momento. De este modo, es muy importante mantener un equilibrio entre la explotación y exploración de las soluciones con tal de no aumentar de forma desmedida la carga computacional. Por ello, es de suma importancia la correcta definición de los rangos de los hiperparámetros.

4.3.3. Principales dificultades

Atendiendo a los principales obstáculos encontrados, se pueden destacar los siguientes:

- En primer lugar, como ya se ha indicado, se parte de la información bursátil a lo largo de 10 años de 2604 valores. El principal problema es que no se dispone información de todos ellos en todos los instantes temporales, pues, no todos cotizan en el mismo tiempo. Siendo conscientes de este suceso, es necesario manejar diferentes índices de Tickers según los valores que estén presentes en cada instante.

- En segundo lugar, la estrategia de inversión diseñada tiene un horizonte temporal fijo. De este modo, las ventanas son completamente ciegas. El problema aparece cuando se genera una señal de compra adicional, estando el valor ya en cartera. Para remediar este hecho, se establece que si se generan varias señales de compra consecutivas, se prolonga la inversión hasta el máximo horizonte temporal posible. Para ilustrar esta casuística, supóngase que el algoritmo genera una orden de compra de Apple el 1 de octubre de 2021 y,

tres días más tarde, vuelve a activar el *flag* de compra. En esta situación, se venderá el título el 12 de octubre de 2021, 7 días hábiles más tarde.

4.4. Implementación de las simulaciones

Como se ha introducido anteriormente, los dos parámetros fundamentales que permiten la construcción del backtesting dinámico son la frecuencia de adaptación de los parámetros y el búfer de datos que se emplea para la optimización de los umbrales de corte. De este modo, el objetivo fundamental de esta sección es conducir una serie de pruebas experimentales que permitan cuantificar el impacto de cada uno de estos parámetros. Como ya se ha mencionado en secciones anteriores el principal problema con estas simulaciones es su excesiva carga computacional, por lo tanto, hay que ser muy selectivos en las pruebas pilotos a seleccionar.

Para validar la eficacia del método implementado se adjuntan las Figuras 4.16, 4.17 y 4.18. Todas ellas están realizadas con el mismo formato con el fin de que los resultados sean comparables entre sí. En ellas se incluyen la siguiente información:

- En primer lugar, las líneas azul y amarilla representan, respectivamente, los beneficios acumulados en bruto y neto, de aplicar la estrategia de inversión diseñada.
- En segundo lugar, la línea morada permite visualizar la evolución de la asignación de capital del portfolio a lo largo del tiempo.
- En tercer lugar, la línea roja muestra el dinero destinado a cubrir las comisiones de transacción.
- En cuarto lugar, con respecto a las anotaciones en la cabecera del título, el término de *Open positions* hace referencia al máximo número de títulos sostenidos de forma simultánea en la cartera. Los *thresholds* hacen referencia a los umbrales tomados por referencia para el número de símbolos a comprar, la rentabilidad mediana y la tasa de acierto. Estos valores son con los que se empieza en el primer periodo simulado del backtesting hasta que se activa el *trigger* de la actualización de los hiperparámetros. Por último, el *annual profit* hace referencia al rendimiento anualizado de la inversión.
- En quinto y último lugar, se representan en la parte inferior de cada gráfico los beneficios año a año.

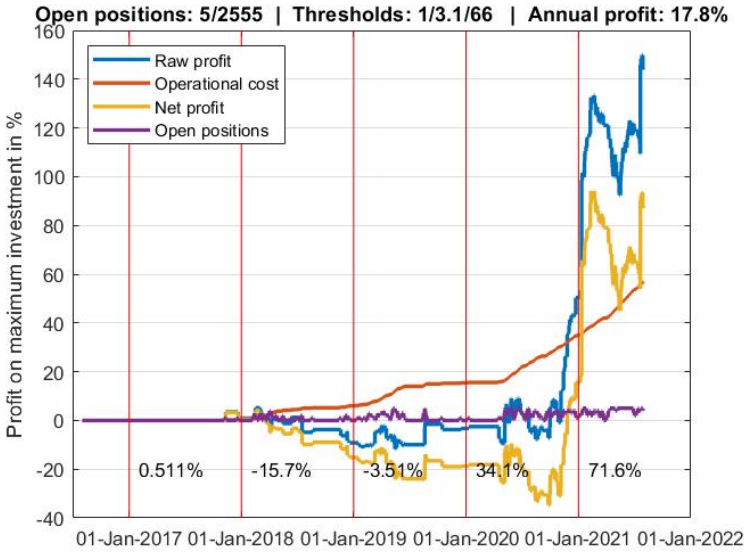


Figura 4.16. *Actualización mensual y memoria de 125 días*

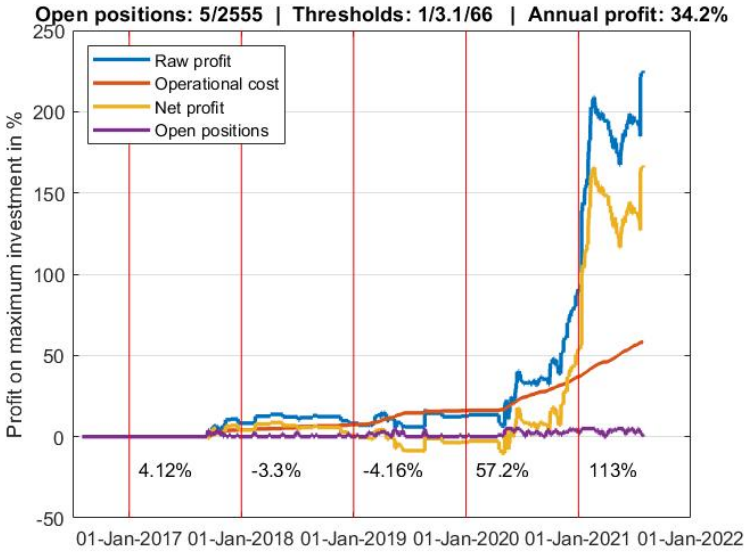


Figura 4.17. *Actualización semanal y memoria de 115 días*

En vista a los resultados obtenidos, se pueden analizar una serie de particularidades:

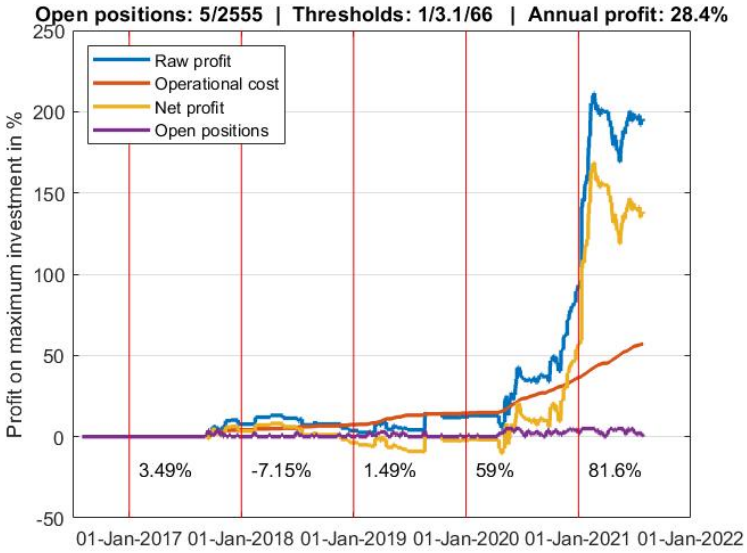


Figura 4.18. *Actualización diaria y memoria de 115 días*

1. Si se establece una comparación entre la Figura 4.16 y 4.17, se puede observar cómo al aumentar la frecuencia con la que se actualizan los hiperparámetros, se aumenta la rentabilidad anual alcanzada. Esto parece intuitivo, ya que al reducir el periodo de actualización, se obtiene una información más precisa de cada situación del mercado y ello permite al modelo adaptarse mejor a la dinámica actual.

2. Por otra parte, si se comparan los resultados de las Figuras 4.17 y 4.18 se obtiene un efecto contrario a la hipótesis enunciada anteriormente. Dicha explicación se debe a la aleatoriedad fruto del día de actualización. Es decir, el resultado de la Figura 4.18 es la media de las simulaciones que se obtendrían al imponer días específicos de la semana en la actualización. De hecho, este fenómeno es muy interesante y pone claramente de manifiesto que las anomalías de calendario, entre ellas el Efecto Día de Semana, representan un enigma para los mercados financieros y una aparente contradicción para la teoría de mercados eficientes. El Efecto Día de Semana en los mercados accionarios consiste en que la distribución de los retornos de un día es diferente a la de los otros días. Específicamente se pueden destacar el Efecto Lunes y el Efecto Fin de Semana. El Efecto Lunes se caracteriza porque el día lunes tendría una rentabilidad menor que los demás días y frecuentemente

una mayor volatilidad. Por el contrario, el Efecto Fin de Semana tiene como característica que el día viernes presentaría un mayor retorno que el resto de los días de la semana.

A continuación, se adjunta en la Tabla 4.3 un resumen de los resultados obtenidos.

Frecuencia	Memoria	Rendimiento anualizado (%)
Mensual	125	17,8
Semanal	115	34,2
Diaria	115	28,4

Tabla 4.3. *Resultados de las simulaciones*

De este modo, el mejor rendimiento alcanzado es del 34.2 % anual, aunque un resultado más determinista indicaría una rentabilidad del 28.4 %, como ya se ha indicado anteriormente. A pesar de que en ambos casos se tratan de valores muy satisfactorios, no se puede afirmar lo mismo de la distribución de los rendimientos año a año, ya que el mayor volumen de operaciones se concentran en los últimos dos años. Además, una de las mayores desventajas de la implementación de un backtesting histórico es que, se prueba un único escenario (la trayectoria histórica), que puede ser fácilmente sobreajustada (Bailey et al. (2014) [4]) y que la simulación puede no ser necesariamente representativa del rendimiento futuro, ya que los resultados pueden estar sesgados por la secuencia particular de puntos de datos. De aquí surge la necesidad de optimizar la estrategia de inversión empleando otros métodos diferentes que se explicarán en el siguiente capítulo.

4.5. Análisis de los resultados

Para el análisis de los resultados se adjuntan en los Apéndices A y B dos informes generados con los datos de la estrategia de inversión.

El primero de ellos se genera para un día en concreto y presenta el índice de la ordenación de los valores según el triple criterio descrito. A su vez, también incluye información sobre el histórico de una serie de indicadores (MACD, RSI y las velas japonesas), así como información financiera de la compañía en cuestión. Entre

ellos está el balance general, el informe de flujos de caja, la cuenta de pérdidas y ganancias, indicadores usados en la estrategia de *Value investing* defendida por el ilustre Warren Buffett y las recomendaciones de los analistas. Adicionalmente se adjunta una breve descripción sobre la compañía y el sector al que pertenece. De este modo, con todo este abanico de datos se pueden realizar análisis muy completos del método diseñado. Por último, una sección muy interesante de este informe es la que se adjunta en la Figura 4.19. En ella se computan la tasa de acierto y la rentabilidad media, mediana, máxima y mínima predichas, para cada subconjunto de vecinos y distintos horizontes de inversión.

```
-------------------------------------------------------
 5 days
 SET/TOTAL  SUCCESS MEAN   MEDIAN MAX     MIN    STD
 100/627250  74.00  11.19   9.69  59.07 -35.77  16.77
 150/627250  69.33   9.40   7.77  59.07 -35.77  16.44
 250/627250  66.00   8.50   6.64  59.07 -39.62  16.63
 500/627250  62.40   7.43   5.21  72.41 -39.62  16.78
 750/627250  60.67   6.95   4.37 389.08 -39.62  21.23
1000/627250  59.00   5.99   3.64 389.08 -41.67  20.07
1500/627250  58.73   5.16   3.10 389.08 -41.67  18.46
2000/627250  58.00   4.74   2.71 389.08 -61.33  17.60
-------------------------------------------------------
 4 days
 SET/TOTAL  SUCCESS MEAN   MEDIAN MAX     MIN    STD
 100/627250  68.00   8.16   6.06  62.83 -40.63  15.60
 150/627250  65.33   6.62   4.52  62.83 -40.63  15.07
 250/627250  64.40   6.86   4.52  62.83 -40.63  15.09
 500/627250  62.00   5.73   3.60  62.83 -40.63  14.66
 750/627250  60.80   5.63   3.12 407.86 -40.63  20.61
1000/627250  58.40   4.82   2.62 407.86 -40.63  19.23
1500/627250  57.87   4.27   2.34 407.86 -40.63  17.39
2000/627250  57.10   3.85   2.14 407.86 -40.63  16.53
-------------------------------------------------------
 3 days
 SET/TOTAL  SUCCESS MEAN   MEDIAN MAX     MIN    STD
 100/627250  64.00   6.26   3.93  46.99 -28.18  13.57
 150/627250  61.33   4.72   3.34  53.01 -28.18  13.48
 250/627250  62.40   4.61   3.34  53.01 -31.97  13.46
 500/627250  59.40   4.20   2.50  58.41 -31.97  13.08
 750/627250  58.67   4.57   2.45 575.98 -31.97  24.41
1000/627250  56.60   3.77   1.95 575.98 -31.97  21.98
1500/627250  55.40   3.09   1.58 575.98 -31.97  19.03
2000/627250  54.35   2.79   1.22 575.98 -34.72  17.38
-------------------------------------------------------
```

Figura 4.19. *Sección del informe de resultados*

A la vista de los resultados se pueden realizar una serie de afirmaciones: [5]

[5]Lógicamente para poder afirmar dichas hipótesis sería necesario realizar análisis estadísticos

- Tanto la tasa de acierto como la rentabilidad media y mediana predicha, parecen estar ordenadas de forma descendiente según el número de vecinos tomados. La lógica subyacente es que al ser más selectivos en cuanto al número de datos que se categorizan como similares, los resultados obtenidos son también más fidedignos.

- En relación a las rentabilidades máximas y mínimas ocurre el fenómeno contrario. Al aumentar el pool de datos, aumenta la posibilidad de encontrar valores más dispares al actual.

- Si se observa la desviación típica de la rentabilidad se puede notar un efecto muy contraintuitivo. Como se ha indicado anteriormente, a menor número de vecinos, más similares son los datos entre sí, y por lo tanto, menor dispersión tendría que haber en la muestra. Sin embargo, se observa que en todos los casos se alcanza un máximo de desviación en 750 vecinos y luego dicha variabilidad se reduce paulatinamente.

El segundo informe realiza un exhaustivo análisis del resultado de aplicar la estrategia de inversión. Para ello, se incluyen indicadores como el número de operaciones rentables y no, las máximas rachas de rentabilidad tanto positivas como negativas, el beneficio esperado por cada operación y la duración media de cada una de ellas. Además, se completa dicho informe con incluyendo el histórico de operaciones efectuadas por el algoritmo.

más exhaustivos

Capítulo 5

Optimización de la estrategia de inversión

Uno de los errores más comunes en la investigación financiera es tomar algunos datos, pasarlos por un algoritmo de Machine Learning, como el K-NN en este caso, hacer una prueba retrospectiva de las predicciones y repetir la secuencia hasta que aparezca un backtesting que genere resultados satisfactorios. Las revistas académicas están llenas de estos pseudodescubrimientos, e incluso los grandes fondos de cobertura caen constantemente en esta trampa. No importa si el backtesting es un *walk-forward* con una muestra de *out-of sample*. El hecho de que se repita una prueba una y otra vez sobre los mismos datos conduce probablemente a un falso descubrimiento. Este error metodológico es tan notorio entre los estadísticos, que ellos mismos lo consideran un fraude científico y la American Statistical Association lo advierte en sus directrices éticas (American Statistical Association (2016) [3]). Normalmente se necesitan unas 20 iteraciones de este tipo para descubrir una estrategia de inversión (falsa) sujeta al nivel de significación estándar (tasa de falsos positivos) del 5 % (López de Prado (2018) [26]).

Para subsanar este problema se propone realizar diferentes análisis de importancia de las características o *feature importance*. Con el objetivo de ilustrar mejor la idea propuesta, supóngase que se da un par de matrices (X, y) que contienen respectivamente las características y etiquetas para un determinado instrumento financiero. A partir de estos datos se puede ajustar un modelo de Machine Learning y evaluar el error de generalización a través de sucesivas validaciones. Supóngase que se obtiene un buen rendimiento, la siguiente pregunta natural es entender qué características han contribuido a dicho suceso. Para ello, se generan algunas preguntas como, por ejemplo, si resulta favorable el hecho de añadir otras

características que refuercen la señal responsable del poder predictivo del algoritmo. Asimismo, también se podrían eliminar algunas de las características que sólo añaden ruido al sistema. Lo fundamental es que la compresión de la importancia de las características abre la proverbial *caja negra* de los modelos de ML. Esta es una de las razones fundamentales por las que el mantra de la *caja negra* es un fenómeno exagerado por los escépticos del ML. Lógicamente, el algoritmo aprende sin que ningún agente humano le dirija su búsqueda, ya que se trata de un aprendizaje automatizado, pero ello no implica que luego no se pueda analizar e interpretar la información del mismo. De tal forma, una vez que se realiza el análisis de las características y factores más importantes se puede llegar a aprender más conduciendo diferentes experimentos con el objetivo de ir cerrando algunas de las siguientes interrogativas:

- *¿Son las características importantes durante todo el tiempo o sólo en algunos entornos específicos?* Esto se abordará en la primera sección del presente capítulo. Para ello, se estudiará la correlación existente entre la rentabilidad real y cada una de las variables independientes.

- *¿Qué provoca un cambio de importancia a lo largo del tiempo? ¿Puede predecir el algoritmo dichos cambios de regímenes?* Este factor se analizará en la última sección de este apartado. Para ello, se estudiará el efecto del horizonte de predicción sobre la estrategia de inversión. La idea subyacente guarda cierta analogía con el concepto de valor de una opción. En términos generales dicho valor se puede descomponer en dos: el intrínseco y el temporal. El primero depende del precio de ejercicio y del activo subyacente, mientras que el segundo, que es el importante en este análisis, depende de la volatilidad, el tiempo a vencimiento, el tipo de interés a corto y los dividendos. Si se analiza el efecto de las dos primeras variables se intuye que a menor tiempo de vencimiento, menor es el valor de la prima. Por contra, a mayor volatilidad del mercado, mayor es el valor teórico de la opción call. Por lo tanto, el objetivo fundamental es entender si se pueden extrapolar dichos conceptos al ámbito interés, es decir, si el horizonte de predicción, que sería el equivalente al tiempo de vencimiento, tiene un impacto diferente según la volatilidad del mercado.

- *¿Son estas características importantes para todos los instrumentos financieros?* Esta pregunta se tratará de abordar en el penúltimo apartado analizando el impacto del número de vecinos que se toma para la predicción de la rentabilidad. La hipótesis inicial es que, según las características de cada acción, sería aconsejable emplear un número de vecinos diferente para

así detectar patrones más atípicos y no caer en valores medios, tal y como proponen Sun et al. (2010) [44] mediante el diseño de un K-NN adaptativo.

- *¿Cuál es el subconjunto de características con el rango más alto de correlación en todo el universo de inversión?* Se discutirá en la segunda sección. Para ello, se analizará el impacto de implementar un algoritmo K-NN sectorial en lugar de global. El término sectorial hace referencia a qué pool de datos coge el K-NN para encontrar los patrones más similares; si sólo los de empresas de un mismo sector, o los de todas las acciones. La intuición indica que la rentabilidad de la cartera puede estar muy sesgada por la composición sectorial o, incluso industrial, del conjunto de acciones.

Esta información es mucho más interesante de investigar en las estrategias. En las sucesivas secciones se tratará de resolver las interrogativas planteadas. Para ello, en lugar de analizar los rendimientos presentados en las Figuras 4.16, 4.17 y 4.18, se evaluará el impacto que tienen los parámetros anteriores sobre la rentabilidad bruta derivada de la ordenación del algoritmo K-NN, sin la aplicación de los umbrales de compraventa. De este modo, no se tiene que ejecutar el backtesting dinámico que actúa sobre la capa superior. Es decir, se ejecutan únicamente las fases señaladas en la Figura 4.14 del capítulo anterior.

Por lo tanto, en los siguientes apartados se analizará el impacto de cada uno de los hiperparámetros anteriores sobre la rentabilidad del K-NN para el caso base, adjunto en la Figura 5.1.

A continuación, se incluyen en la Tabla 5.1 los hiperparámetros del mismo, así como, la sección en la que se evaluará su impacto sobre la rentabilidad del modelo.

5.1. Optimización de las entradas

Garbage in garbage out es un tópico en los círculos de Machine Learning. Cualquiera que trabaje con Inteligencia Artificial sabe que la calidad de los datos determina en gran medida la calidad del resultado. Pero la *basura* es una categoría amplia y en expansión en la ciencia de datos: datos mal etiquetados o inexactos, datos que reflejan prejuicios humanos subyacentes, datos incompletos. Parafraseando a Tolstoi, los grandes conjuntos de datos son todos iguales, pero todos los conjuntos de datos basura son basura a su manera, única y horrible.

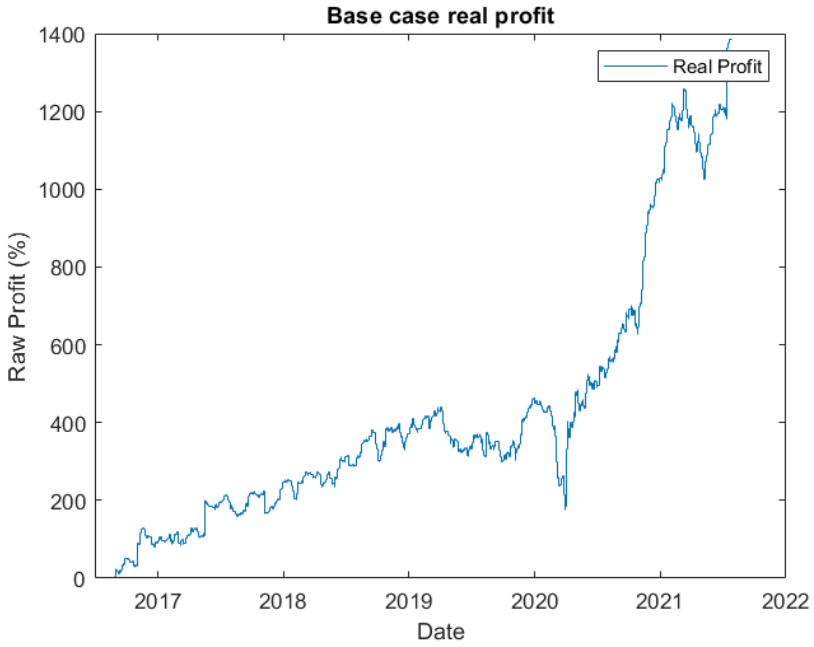


Figura 5.1. *Rentabilidad real del K-NN en el caso base*

Las personas creen en el aprendizaje automático. El filósofo e historiador israelí Yuval Noah Harrari acuñó el término *dataísmo* para describir una fe ciega en los algoritmos. Esta fe va más allá de la capacidad del aprendizaje automático para analizar datos. Muchos creen que el aprendizaje automático es capaz de predecir el futuro de forma automática. Esto es inexacto, el aprendizaje automático es excelente para identificar patrones en grandes conjuntos de datos bien etiquetados. En ciertos casos, esos patrones seguirán desarrollándose en el futuro. En otros casos, no. Sin embargo, el investigador de aprendizaje automático debe asumir la responsabilidad de la percepción de que la IA es predictiva. Mitigar los efectos de los datos basura se convierte en un imperativo moral cuando su algoritmo se utiliza para tomar decisiones de libertad condicional o para invertir miles de millones de dólares de pensiones ganados con esfuerzo.

La fe en el aprendizaje automático no está fuera de lugar. La razón por la se considera más objetivo que otros métodos de análisis de datos es porque es demostrablemente superior. El aprendizaje automático es capaz de producir modelos que pueden analizar conjuntos de datos más grandes y complejos que los métodos convencionales, y ofrecer resultados más precisos, todo ello a escala. Cuando

Hiperparámetro	Valores	Caso base	Sección
Ponderación del volumen en las velas	Sí/No	Sí	1
Normalización de las velas	Apertura/Propia	Apertura	1
K-Vecinos más cercanos	Vector K [1]	150	3
Horizonte de predicción	$[1, 2, 3, 4, 5]$ días	5	4
Ventana deslizante para la búsqueda de patrones en global	1 año	1 año	Futuros desarrollos
Ventana deslizante para la búsqueda de patrones en sectores	2 años	2 años	Futuros desarrollos
Ventana deslizante para la búsqueda de patrones en industrias	3 años	3 años	Futuros desarrollos
Ventana deslizante para la búsqueda de patrones en valores	4 años	4 años	Futuros desarrollos

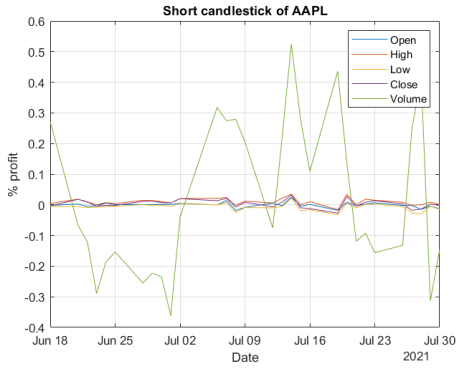
Tabla 5.1. *Rango de valores de los hiperparámetros del modelo*

se utiliza correctamente, puede ayudar a las empresas a identificar oportunidades rentables y evitar riesgos.

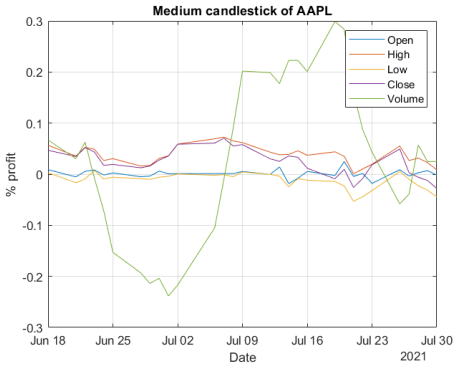
Por lo tanto, queda suficientemente probada la necesidad de optimizar las entradas que el modelo emplea para la búsqueda de los patrones de los vecinos más cercanos. Debido a su extrema importancia en el campo en cuestión, se procederá con la implementación de dos análisis complementarios. El primero de ellos evaluará la sensibilidad de la rentabilidad del K-NN frente a distintas ponderaciones de sus entradas.

5.1.1. Análisis de las ponderaciones de las velas

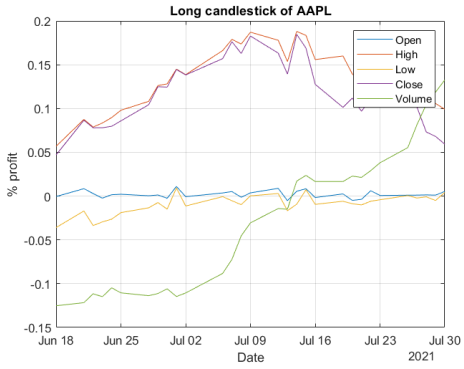
Para entender la motivación del estudio conviene analizar las Figuras 5.2, 5.3, 5.4 y 5.5. Los dos conjuntos de gráficos primeros ilustran la evolución del OHLCV de las velas según la ventana deslizante empleada para los valores de Apple y Microsoft. Por contra, los dos últimos representan justamente lo contrario, es decir, la evolución de cada una de las ventanas deslizantes según el precio de la vela en cuestión. A continuación, se analiza el efecto combinado de estos parámetros.



(a) Ventana corta



(b) Ventana mediana

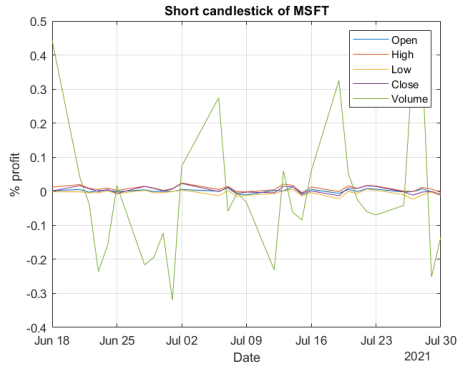


(c) Ventana larga

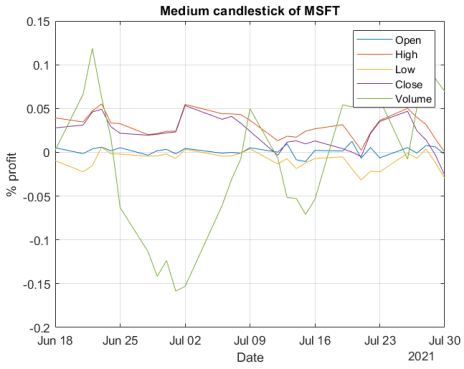
Figura 5.2. *Evolución de las componentes de las velas de AAPL para diferentes horizontes temporales*

Si se observan las Figuras 5.2 y 5.3 se puede notar como a mayor longitud de ventana, más ponderadas están las componentes de las velas entre sí, sobre todo, el volumen respecto a las 4 restantes. Esto puede ser muy interesante de cara a escalar el volumen con un factor diferente según la longitud de la ventana empleada. Conviene matizar que se ha incluido la información de dos valores diferentes para descartar que se trate de un efecto aislado de una compañía determinada.

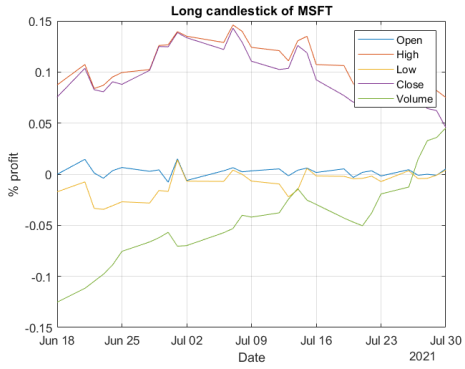
Por otra parte, si se analizan las Figuras 5.4 y 5.5 se puede destacar un efecto muy significativo. Cualquiera sea la tendencia del activo subyacente, las componentes del precio de apertura y del precio mínimo presentan mucho más ruido que las componentes del precio máximo y de cierre. Esto tiene su justificación en la



(a) Ventana corta



(b) Ventana mediana



(c) Ventana larga

Figura 5.3. *Evolución de las componentes de las velas de MSFT para diferentes horizontes temporales*

forma en la que se calculan dichas velas, y puede ser muy interesante de cara a entender qué componentes aportan información más valiosa (e.g. con menos fluctuaciones). Bajo el actual paradigma, para normalizar las componentes se emplea el precio de cierre de la semana anterior. A efectos prácticos esto implica que para el caso de la componente de apertura, se calcule realmente el rendimiento diario de la vela y de ahí que la señal resultante tenga tantas perturbaciones.

Por lo tanto, en vista a los resultados analizados se pueden establecer las siguientes afirmaciones:

- Existe una clara influencia en la ponderación del volumen frente al resto de componentes de las velas. Para ello, se explorará el efecto de no escalar el

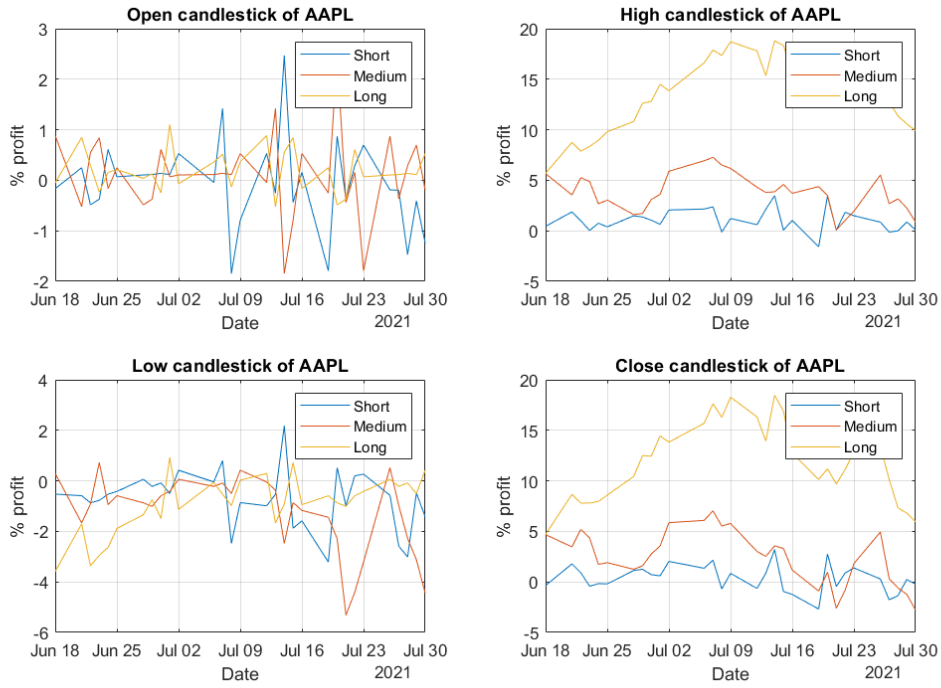


Figura 5.4. *Evolución de las ventanas según el OHLC de las velas - Tendencia alcista*

volumen.

- Se observa un impacto significativo en el método de normalización empleado. De este modo, se estudiarán otras formas para homogeneizar las componentes. En el caso base, la forma de parametrizar las entradas es mediante la aplicación de las siguientes ecuaciones:

$$u_{open}[k] = 100 \times \left(\frac{O[k-W+1]}{C[k-W]} - 1 \right) \tag{5.1}$$

$$u_{high}[k] = 100 \times \left(\frac{max\{H[k-W+1] : H[k]\}}{C[k-W]} - 1 \right) \tag{5.2}$$

$$u_{low}[k] = 100 \times \left(\frac{max\{L[k-W+1] : L[k]\}}{C[k-W]} - 1 \right) \tag{5.3}$$

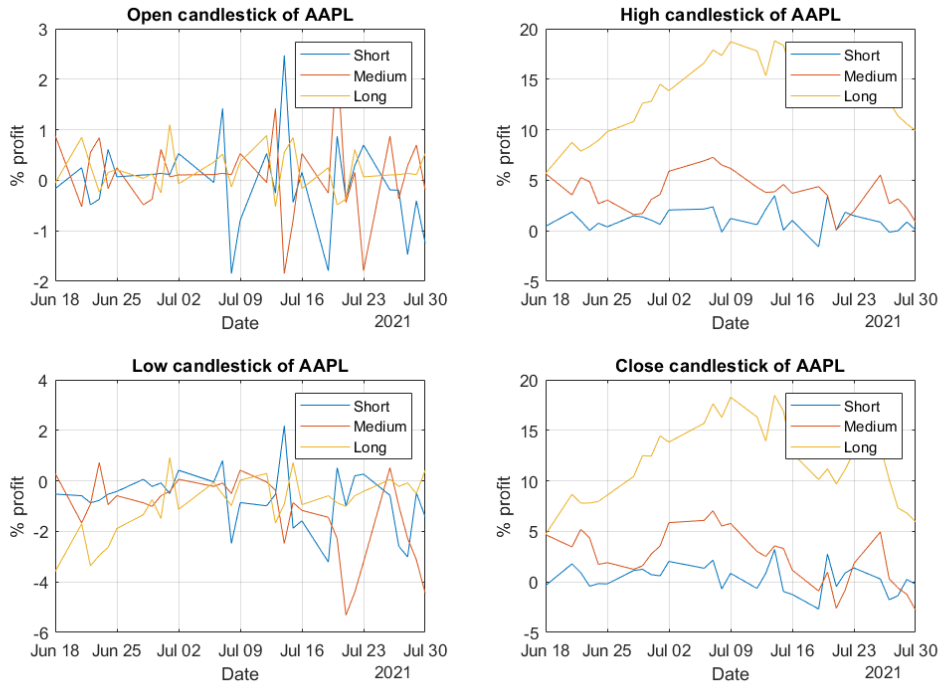


Figura 5.5. *Evolución de las ventanas según el OHLC de las velas - Tendencia bajista*

$$u_{close}[k] = 100 \times \left(\frac{C[k]}{C[k-W]} - 1 \right) \tag{5.4}$$

$$u_{volume}[k] = \frac{\sum_{i=k-W+1}^{k} \frac{V[k]}{W}}{\sum_{i=k-60+1}^{60} \frac{V[k]}{60}} - 1 \tag{5.5}$$

Como ya se ha indicado anteriormente, esto tiene un gran impacto en la coordenada del open, ya que como se puede apreciar en la Ecuación 5.1, se computan rendimientos diarios. Por lo tanto, merece la pena explorar otros métodos de normalización. Para ello, se proponen dos alternativas. En primer lugar, normalizar con el precio de apertura de la ventana anterior, aplicando las siguientes fórmulas:

$$u_{open}[k] = 100 \times \left(\frac{O[k-W+1]}{O[k-2\times W+1]} - 1 \right) \tag{5.6}$$

$$u_{high}[k] = 100 \times \left(\frac{max\{H[k-W+1] : H[k]\}}{O[k-2\times W+1]} - 1 \right) \tag{5.7}$$

$$u_{low}[k] = 100 \times \left(\frac{max\{L[k-W+1] : H[k]\}}{O[k-2\times W+1]} - 1 \right) \tag{5.8}$$

$$u_{close}[k] = 100 \times \left(\frac{C[k]}{O[k-2\times W+1]} - 1 \right) \tag{5.9}$$

$$u_{volume}[k] = \frac{\sum_{i=k-W+1}^{k} \frac{V[k]}{W}}{\sum_{i=k-60+1}^{60} \frac{V[k]}{60}} - 1 \tag{5.10}$$

En segundo lugar, también parece intuitivo normalizar cada componente consigo misma, en lugar de referenciarla al precio de cierre, es decir, para cualquier día dividirla entre el valor de la vela de la ventana anterior. Dicho ajuste se implementa mediante la ejecución de las siguientes ecuaciones.

$$u_{open}[k] = 100 \times \left(\frac{O[k-W+1]}{O[k-2\times W+1]} - 1 \right) \tag{5.11}$$

$$u_{high}[k] = 100 \times \left(\frac{max\{H[k-W+1] : H[k]\}}{max\{H[k-2\times W+1] : H[k-W+1]\}} - 1 \right) \tag{5.12}$$

$$u_{low}[k] = 100 \times \left(\frac{max\{L[k-W+1] : H[k]\}}{max\{L[k-2\times W+1] : L[k-W+1]\}} - 1 \right) \tag{5.13}$$

$$u_{close}[k] = 100 \times \left(\frac{C[k]}{C[k-W]} - 1 \right) \tag{5.14}$$

$$u_{volume}[k] = \frac{\sum_{i=k-W+1}^{k} \frac{V[k]}{W}}{\sum_{i=k-60+1}^{60} \frac{V[k]}{60}} - 1 \tag{5.15}$$

A continuación, se adjuntan en la Figura 5.6 una gráfica que permite comparar el rendimiento obtenido al aplicar las modificaciones anteriores. En ella, además de incluir la rentabilidad real del modelo, también se adjunta la predicha por el K-NN. En la leyenda las variables significan lo siguiente:

- Original: se corresponden con el caso base
- Open: representan la normalización efectuada con las Ecuaciones 5.6, 5.7, 5.7, 5.8, 5.9 y 5.10, teniendo como denominador el precio de apertura de la ventana anterior.
- NS: se corresponde con la normalización aplicando las Ecuaciones 5.11, 5.12, 5.12, 5.13, 5.14 y 5.15, sin tener en cuenta la ponderación adicional del volumen.
- N: se corresponde con la normalización aplicando las Ecuaciones 5.11, 5.12, 5.12, 5.13, 5.14 y 5.15, teniendo en cuenta la ponderación adicional del volumen.

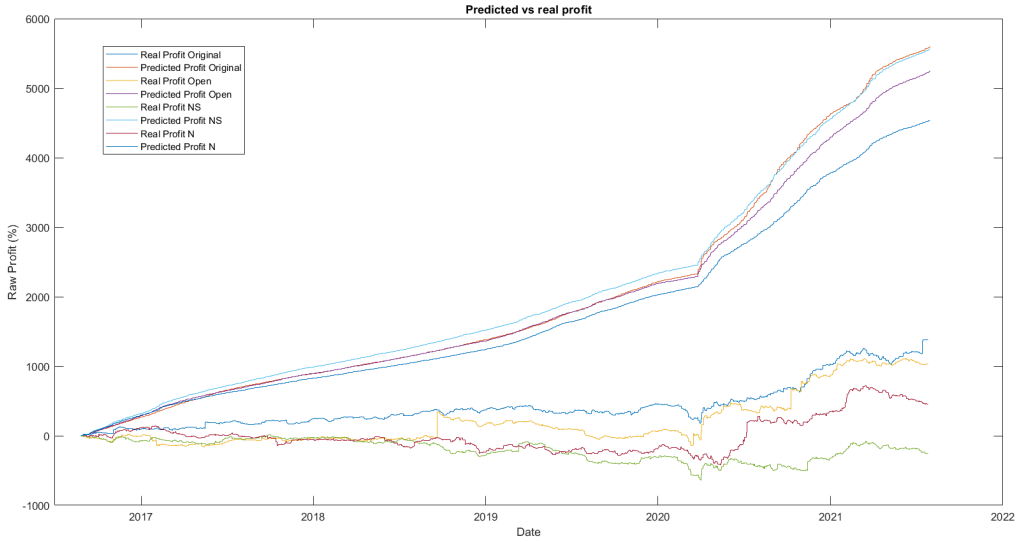


Figura 5.6. *Comparación de las estrategias de ponderación de las entradas*

Tal y como se puede observar, para el caso de la normalización con el precio de apertura, se dan subidas muy acusadas en la rentabilidad real. Por lo tanto, conviene realizar un previo tratamiento de los outliers para no sobreajustar la estrategia

de inversión a unas condicionaes determinadas. Para ello, se eliminan todos los outliers de dicha serie temporal mediante la aplicación de una media móvil de 6 meses centrada. Esta es la forma más idónea de implementarlo, ya que al ser los datos no estacionarios y recurrentes, no tiene sentido aplicar una normalización según un único valor fijo, sino en una ventana móvil según las condiciones de mercado que se dan en cada momento. Es importante matizar, que dicho tratamiento sólo aplica en el caso de la rentabilidad real, pues la rentabilidad mediana, al ser la mediana de muchos valores, no contiene valores fuera de rango como figura en la gráfica anterior. A continuación, se adjunta en la Figura 5.7 el efecto de aplicar dicho ajuste.

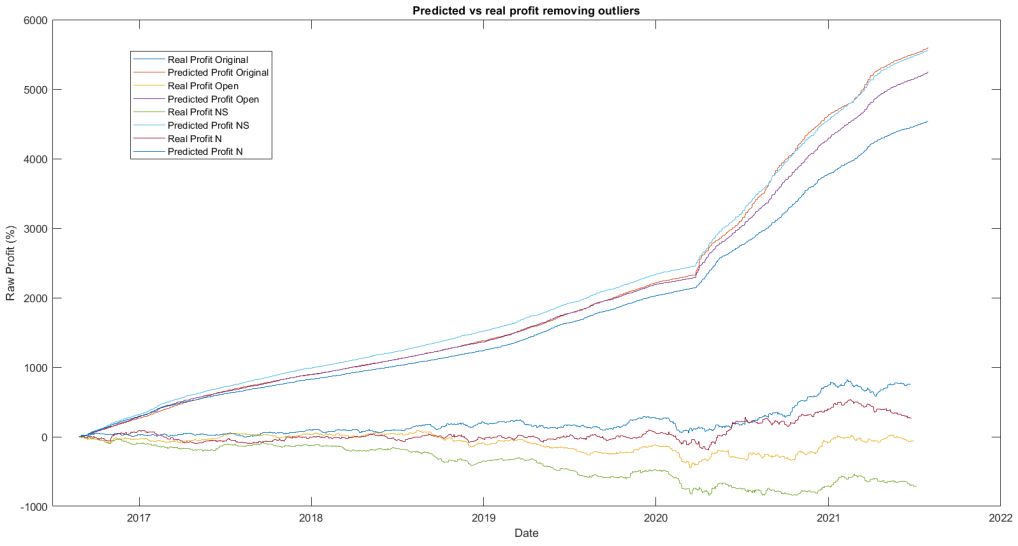


Figura 5.7. *Comparación de las estrategias de ponderación de las entradas sin outliers*

A partir de los resultados anteriores se pueden analizar una serie de sucesos:

- El factor de la ponderación del volumen resulta crucial para obtener estrategias de inversión rentables. Sin embargo, su efecto sobre la rentabilidad mediana predicha no es favorable.

- La mejor estrategia de inversión es la original, seguida de la que implementa la normalización con cada componente.

- Resulta imprescindible establecer las conclusiones bajo resultados libres de outliers. Tal y como se puede observar en la Figura 5.6, la estrategia basada en la homogeneización con el precio de apertura es la segunda mejor, mientras que en el escenario representado en la Figura 5.7 es la penúltima.

5.2. Análisis de la granularidad de las entradas

El concepto de granularidad hace referencia al nivel de detalle con el que se almacena y se analiza la información. Por ejemplo, una empresa que se dedique a la venta de zapatos puede registrar de forma diaria lo que vende y, por otro lado, mensualmente la información relativa al pago de sueldos. Por lo tanto, la posibilidad de analizar la información con mayor profundidad va a depender del nivel de detalle con el que se almacenen los datos. Si se extrapola este concepto a la estrategia de inversión descrita, el término de granularidad de las entradas se refiere al nivel de detalle con el que se localizan los patrones de los vecinos más similares.

5.2.1. Motivación

En el caso base se buscan patrones semejantes en todo el conjunto de valores a lo largo de un año. Esto tiene como ventaja el volumen de información que se maneja. Sin embargo, aquí es cuando nace la disyuntiva entre la cantidad y la calidad de los datos. Parece intuitivo pensar, que mientras más selectivo sea el *pool* de los datos de búsqueda, más fidedignas serán las predicciones del modelo. De este modo, se presentan tres estrategias adicionales para el cálculo de los vecinos:

- Sectorial: para predecir la rentabilidad de cada valor emplea únicamente las compañías pertenecientes al mismo sector. La motivación principal de esta alternativa es probar si realmente puede ser efectiva desarrollar una estrategia sectorial.

- Industrial: para estimar el rendimiento de cada valor emplea exclusivamente las empresas que se engloban en la misma industria.

- Individual: para predecir la rentabilidad de cada valor emplea únicamente el histórico de información propia. En este caso, el interés que suscita el desarrollo de esta estrategia es demostrar si patrones particulares de un activo como, por ejemplo, sus fluctuaciones y movimientos tras la presentación de los resultados, puede influir en la búsqueda de patrones rentables.

Como se ha mencionado anteriormente, el hecho de mejorar la calidad de los datos tiene un efecto en el volumen de información. Por ello, para paliar este efecto se aumenta el tamaño deslizante de la ventana de búsqueda de vecinos. De este modo, los valores empleados son los que se incluyen en la Tabla 5.2.

Estrategia	**Ventana deslizante del K-NN**
Global	1 año
Sectorial	2 años
Industrial	3 años
Individual	4 años

Tabla 5.2. *Tamaño de la ventana deslizante empleada en cada estrategia*

De las tres estrategias adicionales presentadas, merece la pena explorar la sectorial. En el mundo de las finanzas hay inversores que consideran que la estrategia de rotación de sectores puede ser rentable en ciertos instantes temporales. Lo cierto es que cada ciclo económico es diferente a su manera, pero ciertos patrones han tendido a repetirse a lo largo del tiempo. El ciclo económico, que refleja las fluctuaciones de la actividad en una economía, puede ser un determinante fundamental del rendimiento del sector de la renta variable a medio plazo. Un ciclo económico típico se caracteriza por un periodo de crecimiento económico, seguido de un periodo de ralentización del crecimiento y, a continuación, una contracción o recesión. A continuación, el ciclo se repite. De este modo, parece intuitivo buscar patrones de velas similares en empresas que se encuentren sumergidas en el mismo ciclo.

En la actualidad, la estructura del *Global Industry Classification Standard (GICS)* [2] incluye 11 sectores: consumo discrecional, consumo básico, energía, financiero, sanitario, industrial, tecnología de la información, materiales, inmobiliario, servicios de comunicación y servicios públicos. Aunque dicha clasificación se ha extendido entre los participantes del mercado, su historial de resultados es limitado, ya que sólo se remonta a 1989 y abarca únicamente tres recesiones. De este modo, para obtener una descripción completa del rendimiento de los sectores a lo largo de varios ciclos económicos se toman los datos de Kenneth R.French [3] del rendimiento de 48 carteras industriales basadas en la clasificación estándar.

[2]Global Industry Classification Standard Report
[3]Kenneth R.French Data

Para evaluar plenamente el rendimiento del sector a lo largo de los ciclos económicos, el autor valora el beneficio alcanzado y la consistencia del mismo en cada ciclo, empleando para ello las siguientes seis métricas:

1. Rendimiento medio mensual del sector
2. Alpha de la rentabilidad media mensual generada sobre el mercado general
3. Rentabilidad media del sector en cada ciclo económico
4. Alpha de la rentabilidad media generada sobre el mercado general
5. Porcentaje de meses en los que el sector superó al mercado en general
6. Porcentaje de ciclos en los que el sector superó al mercado en general

A continuación, se adjuntan en las Figuras 5.8, 5.9, 5.10 y 5.11 los resultados obtenidos.

	Cons. Disc.	Cons. Staples	Energy	Financials	Health Care	Industrials	Materials	Real Estate	Technology	Utilities
Average Monthly Return (%)	-1.5	-0.1	-0.2	-1.8	-0.5	-1.7	-1.4	-3.1	-2.5	-0.3
Average Monthly Excess Return (%)	0.1	1.5	1.4	-0.2	1.1	-0.1	0.2	-1.5	-0.9	1.3
Average Period Return (%)	-12.0	1.0	-3.5	-13.3	-2.9	-14.8	-11.5	-21.6	-20.3	-1.6
Average Period Excess Return (%)	1.0	14.0	9.0	0.0	10.0	-2.0	1.0	-9.0	-7.0	11.0
Hit Rate (% of Months Outperforming the Market)	49.3	70.4	53.5	53.5	59.2	46.5	47.9	43.7	39.4	60.6
Hit Rate (% of Periods Outperforming the Market)	29.0	86.0	71.0	43.0	86.0	29.0	43.0	14.0	14.0	100.0
Aggregated Z-Score	-2.0	8.3	4.6	-2.0	5.2	-3.4	-1.3	-8.6	-7.7	6.6

Figura 5.8. *Rendimiento del sector en un periodo de recesión (Bartolini. M & Dong. A (2019)[29])*

Durante una recesión, como se muestra en la Figura 5.8, los sectores no cíclicos, como el de productos básicos de consumo, los servicios públicos y la sanidad, se comportaron bien, ya que sus vínculos comerciales con el gasto no discrecional son menos sensibles a las fluctuaciones económicas. Superaron al mercado en general en una media de más del 10 % durante seis de los últimos años. Los sectores inmobiliario y tecnológico se encuentran entre los sectores con peores resultados en todas las métricas. Dado que su negocio está vinculado al gasto de los consumidores y de las empresas, estos sectores tienden a ser los primeros en en experimentar recortes

del gasto durante los periodos de disminución de los ingresos y de las actividades empresariales.

	Cons. Disc.	Cons. Staples	Energy	Financials	Health Care	Industrials	Materials	Real Estate	Technology	Utilities
Average Monthly Return (%)	3.4	1.9	2.8	2.4	2.3	2.9	3.0	3.6	3.0	1.6
Average Monthly Excess Return (%)	1.1	-0.3	0.6	0.1	0.0	0.6	0.7	1.3	0.7	-0.6
Average Period Return (%)	33.1	18.0	27.1	23.1	21.4	27.4	29.3	39.2	28.4	14.7
Average Period Excess Return (%)	12.0	-3.0	6.0	2.0	0.0	6.0	8.0	18.0	7.0	-7.0
Hit Rate (% of Months Outperforming the Market)	64.5	43.5	53.2	54.8	46.8	56.5	61.3	58.1	53.2	45.2
Hit Rate (% of Periods Outperforming the Market)	86.0	29.0	57.0	57.0	43.0	71.0	71.0	57.0	71.0	29.0
Aggregated Z-Score	7.3	-7.5	0.8	-1.5	-4.1	2.2	3.8	7.2	2.3	-9.1

Figura 5.9. *Rendimiento del sector en un periodo de recuperación (Bartolini. M & Dong. A (2019)[29])*

En una fase de recuperación, la mejora del mercado laboral y de la confianza de los consumidores lleva a un aumento del gasto discrecional en restaurantes, viajes y bienes duraderos, beneficiando a los sectores de consumo discrecional. Aunque las recesiones tienden a golpear con fuerza el mercado inmobiliario, los bajos tipos de interés y la flexibilización de la política monetaria tras la recesión hacen que sea más barato y fácil la compra de bienes inmuebles. La recuperación de las actividades comerciales también eleva el valor del comercial del *real estate*, lo que contribuye a los buenos resultados del sector inmobiliario. El rendimiento de los bienes de consumo discrecional es más consistente que el de los bienes inmuebles, dado su mayor índice de aciertos, tanto mensual como periódicamente. Por otra parte, los sectores que se ven favorecidos durante la recesión, como los bienes de consumo básico, los servicios públicos y la sanidad, pierden su atractivo a medida que el mercado se recupera y los inversores adoptan sectores más cíclicos para captar los repuntes del mercado.

La fase de expansión se caracteriza por una menor dispersión sectorial, ya que el crecimiento económico se acelera hasta su máximo y más sectores se benefician del auge económico. Los rendimientos del mercado son los segundos mejores durante esta fase, después de la fase de recuperación, pero la duración de esta fase tiende a ser más larga. Durante esta fase, la utilización de la capacidad tiende a aumentar significativamente, hasta alcanzar su máximo. Las empresas se sienten más

	Cons. Disc.	Cons. Staples	Energy	Financials	Health Care	Industrials	Materials	Real Estate	Technology	Utilities
Average Monthly Return (%)	1.4	0.9	1.2	1.7	1.0	1.4	1.2	1.5	1.8	0.7
Average Monthly Excess Return (%)	0.1	-0.4	-0.2	0.4	-0.3	0.1	-0.2	0.1	0.5	-0.6
Average Period Return (%)	16.6	10.6	15.5	18.7	10.8	16.2	13.1	17.8	21.0	7.6
Average Period Excess Return (%)	1.0	-5.0	0.0	3.0	-4.0	1.0	-2.0	3.0	6.0	-8.0
Hit Rate (% of Months Outperforming the Market)	50.7	47.3	47.8	58.0	47.3	51.7	46.8	51.7	51.7	41.5
Hit Rate (% of Periods Outperforming the Market)	73.0	45.0	45.0	91.0	27.0	55.0	55.0	55.0	82.0	9.0
Aggregated Z-Score	2.8	-4.7	-0.9	8.0	-5.0	2.1	-1.9	3.3	7.6	-10.3

Figura 5.10. *Rendimiento del sector en un periodo de expansión (Bartolini. M & Dong. A (2019)[29])*

seguras de sus perspectivas de crecimiento y comienzan a expandirse, destinando más capital para mejorar la productividad, como la inversión en tecnología. Los tipos de interés también empiezan a subir desde niveles muy bajos. El aumento del volumen de préstamos y de los tipos de interés tiende a beneficiar a financieros. El rendimiento superior de las empresas tecnológicas y financieras es bastante consistente, ya que superan al mercado en 10 de las 12 fases de expansión. Los sectores no cíclicos siguen estando en desventaja durante la fase de expansión económica.

	Cons. Disc.	Cons. Staples	Energy	Financials	Health Care	Industrials	Materials	Real Estate	Technology	Utilities
Average Monthly Return (%)	0.8	1.3	1.0	1.0	1.3	1.1	0.9	0.5	1.0	1.0
Average Monthly Excess Return (%)	-0.1	0.3	0.0	0.0	0.3	0.1	-0.1	-0.4	0.0	0.1
Average Period Return (%)	5.5	14.6	8.5	13.7	15.0	11.9	6.5	2.4	10.1	11.8
Average Period Excess Return (%)	-5.0	4.0	-2.0	4.0	5.0	2.0	-4.0	-8.0	0.0	2.0
Hit Rate (% of Months Outperforming the Market)	46.9	57.9	51.4	48.0	54.2	53.1	48.3	44.7	48.9	50.3
Hit Rate (% of Periods Outperforming the Market)	36.0	73.0	55.0	36.0	73.0	73.0	36.0	27	45.0	55.0
Aggregated Z-Score	-5.4	8.1	-0.3	0.5	7.3	3.8	-3.8	-10.8	-0.9	1.4

Figura 5.11. *Rendimiento del sector en un periodo de desaceleración (Bartolini. M & Dong. A (2019)[29])*

A medida que el crecimiento económico se desacelera y los costes de los insumos aumentan, el crecimiento de la rentabilidad de las empresas se ve presionado,

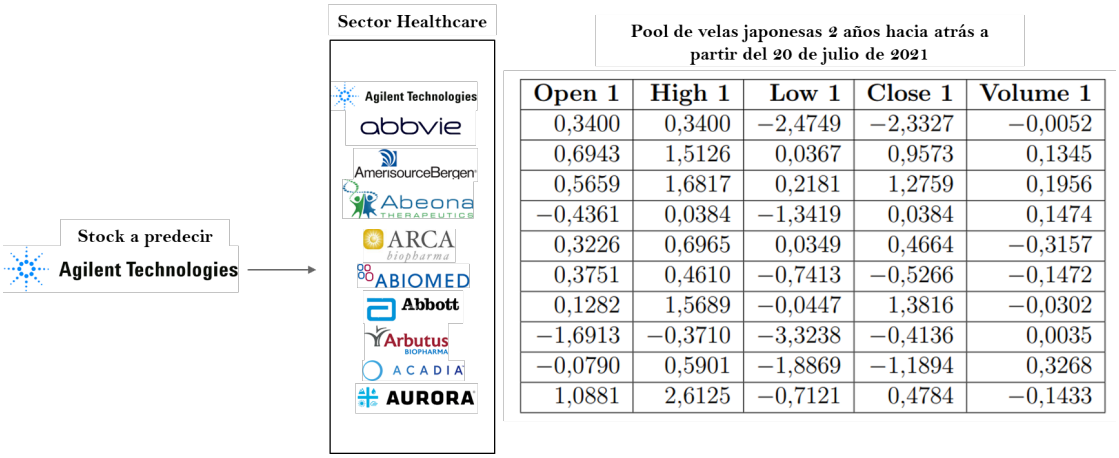
manteniéndose positivo pero ralentizando. Dadas las limitaciones de capacidad y eficiencia, las empresas gastan más en gastos de capital para satisfacer la demanda, pero suele haber un desfase entre el despliegue de CAPEX y el aumento de la productividad. El rendimiento global del mercado está en su segundo peor momento durante esta fase. Los inversores empiezan a posicionarse de forma más defensiva y a reducir su inversión. Esto hace que el rendimiento de los sectores de salud y productos básicos sea superior al de los de los sectores de consumo discrecional e inmobiliario. El sector industrial es el tercero con mejor rendimiento, ya que se beneficia del aumento de la inversión en productos de capital. Sin embargo, los aumentos de las inversiones no se produjeron o no fueron significativos cada vez que el crecimiento económico se ralentizó.

Con este análisis se revela que puede ser interesante realizar un K-NN sectorial para así tomar ventaja de los ciclos económicos entre sectores. Una vez finalizado el estudio, es importante matizar que el objetivo de este no se centra en probar la eficacia de una estrategia sectorial pura, sino en demostrar cómo las empresas pertenecientes a un mismo sector se comportan de manera similar debido a los ciclos económicos.

5.2.2. Implementación de las modificaciones en la estrategia

Como se ha introducido anteriormente, los ajustes se centran mayoritariamente en el parámetro de la ventana deslizante en dos aspectos fundamentales: tamaño y contenido. El primero de ellos ya se ha discutido con los valores adjuntos en la Tabla 5.2. En cambio, el segundo hace referencia al *pool* de compañías que se seleccionan para calcular la rentabilidad media, mediana y la tasa de acierto predicha para cada día del periodo simulado. Para ilustrar la metodología empleada se adjunta la Figura 5.12. En ella se han simulado las fases implementadas por el backtesting inferior para un día específico, el 20 de julio de 2021 y para la compañía Agilent Technologies. De este modo, como dicha empresa pertenece al sector de *Healthcare*, se genera una máscara binaria de stocks entre, los 2555 disponibles, que pertenezcan al mismo sector y, posteriormente, dichos valores son los que el K-NN recibe como inputs. Del pool total de 1287720 valores (2555 $stocks \times 2$ $years \times 252$ $days$) se seleccionan únicamente los 253000 datos que son del sector *Healthcare*.

A continuación, se adjunta en la Figura 5.13 como queda la estrategia de inversión resultante con la introducción de las modificaciones explicadas. Se señala con un recuadro de color azul oscuro la etapa en la que tiene lugar el ajuste.



Pool de velas japonesas 2 años hacia atrás a partir del 20 de julio de 2021

Open 1	High 1	Low 1	Close 1	Volume 1
0,3400	0,3400	−2,4749	−2,3327	−0,0052
0,6943	1,5126	0,0367	0,9573	0,1345
0,5659	1,6817	0,2181	1,2759	0,1956
−0,4361	0,0384	−1,3419	0,0384	0,1474
0,3226	0,6965	0,0349	0,4664	−0,3157
0,3751	0,4610	−0,7413	−0,5266	−0,1472
0,1282	1,5689	−0,0447	1,3816	−0,0302
−1,6913	−0,3710	−3,3238	−0,4136	0,0035
−0,0790	0,5901	−1,8869	−1,1894	0,3268
1,0881	2,6125	−0,7121	0,4784	−0,1433

Figura 5.12. *Ejemplo de la selección de valores para la búsqueda de vecinos*

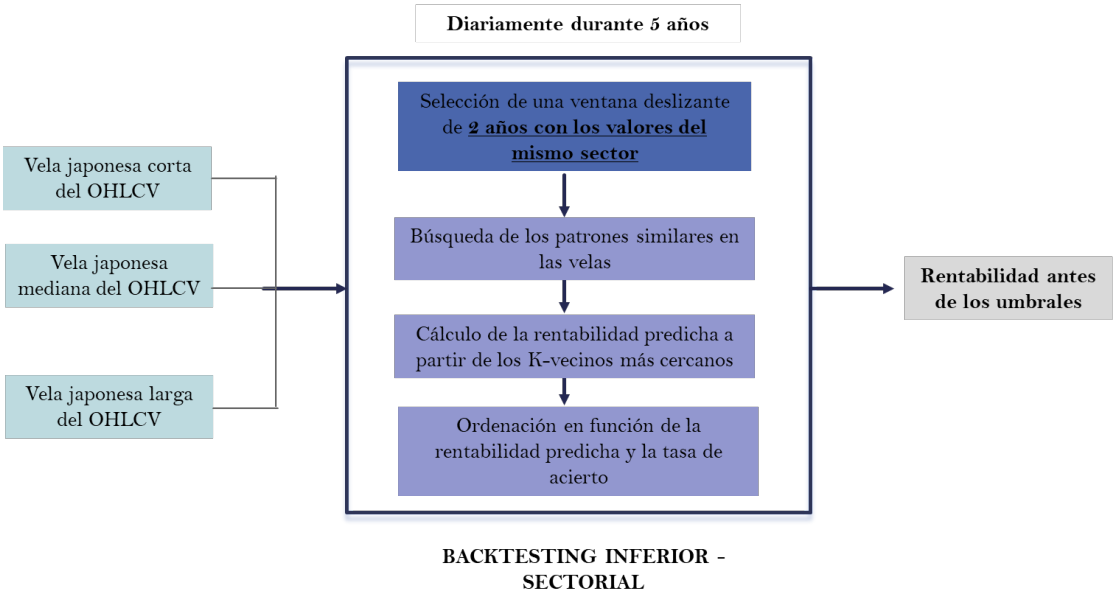
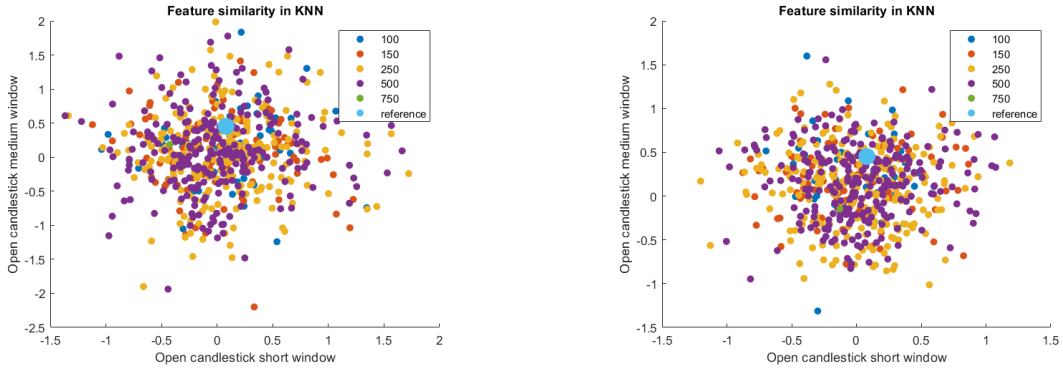


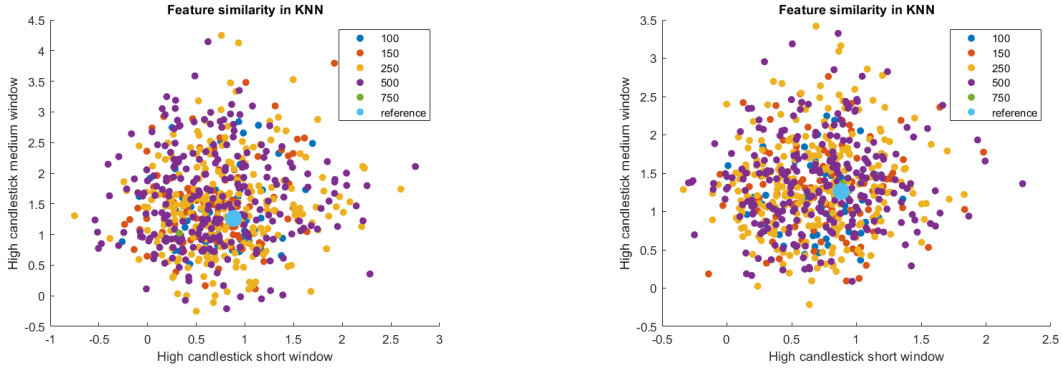
Figura 5.13. *Modificación sectorial del backtesting inferior*

Resulta interesante comparar el aspecto particular de la nube de puntos de los vecinos más cercanos con respecto al caso base. Para ello, se incluyen las Figuras 5.14, 5.15, 5.16, 5.17 y 5.18, en las que se representan dos de las 15 coordenadas disponibles en un *scatter-plot*, así como el punto a predecir.



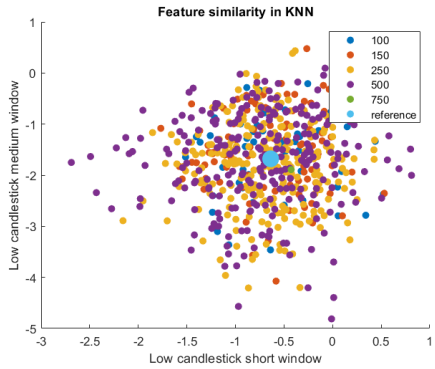
(a) Sectorial **(b)** Global (caso base)

Figura 5.14. *Comparación de los K-vecinos para la estrategia global y sectorial - Open*

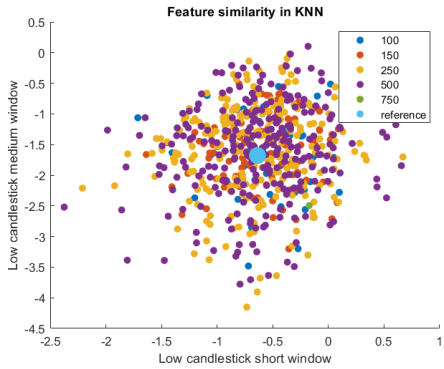


(a) Sectorial **(b)** Global (caso base)

Figura 5.15. *Comparación de los K-vecinos para la estrategia global y sectorial - High*

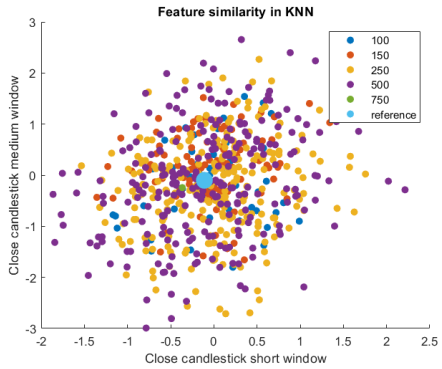


(a) Sectorial

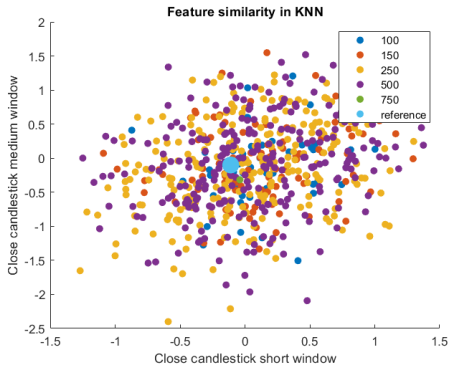


(b) Global (caso base)

Figura 5.16. *Comparación de los K-vecinos para la estrategia global y sectorial - Low*

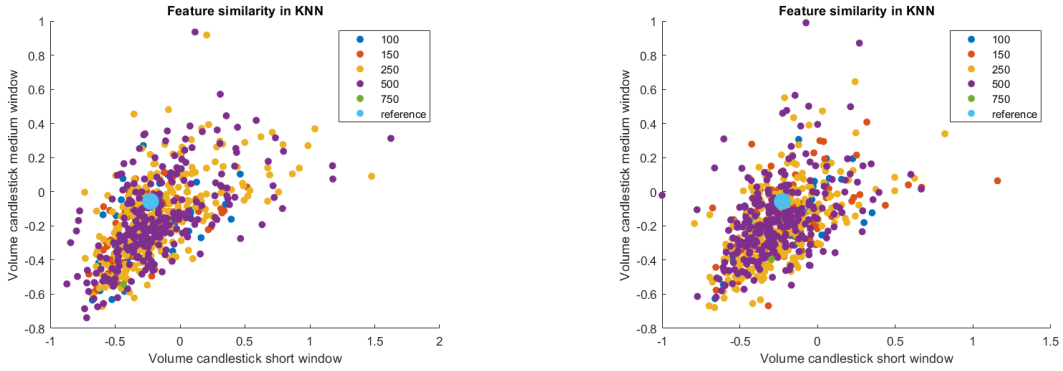


(a) Sectorial



(b) Global (caso base)

Figura 5.17. *Comparación de los K-vecinos para la estrategia global y sectorial - Close*



(a) Sectorial **(b)** Global (caso base)

Figura 5.18. *Comparación de los K-vecinos para la estrategia global y sectorial - Volume*

Tal y como se puede observar, se produce un efecto muy contraintuitivo. En todos los elementos del OHLCV de la vela, excepto en el volumen, la dispersión de los datos es mayor para el caso del análisis sectorial que para el caso base, a pesar de que en el *pool* de datos a buscar debería de ser más similar. Esto se debe a que el vector-K del número de vecinos ($[100, 150, 250, 500, 750, 1000, 1500, 2000]$; se mantiene fijo en ambas simulaciones. Para ello, conviene calcular la proporción por sectores de los valores, adjunto en la Tabla 5.3.

Sector	**Porcentaje de empresas (%)**
Basic Materials	5,09
Communication Services	11,08
Consumer Defensive	4,19
Energy	5,79
Financial Services	11,27
Healthcare	21,21
Industrials	12,21
Real Estate	5,87
Technology	15,34
Utilities	2,23

Tabla 5.3. *Porcentajes de empresas según el sector*

En vista a la tabla anterior, se puede notar como las clases no están balanceadas, existiendo claramente una categoría mayoritaria (*Healthcare*). A parte de este fenómeno, lo que realmente interesa en este punto es que en el peor escenario posible, sector de *Utilities*, se tiene un pool de datos 22 veces menor que en el caso base ($\frac{1}{0{,}023} \times 2\ years\ sliding - window\ =\ 22{,}42 \approx 22$). De esta forma, para que el algoritmo se favorezca realmente de tener unos datos más selectivos, se tendría que reducir como mínimo proporcionalmente el número de vecinos a tener en cuenta.

5.2.3. Análisis de los resultados

Para abordar esta sección se llevará a cabo un estudio bidireccional. En primer lugar, se analizará la exposición sectorial del mejor valor del K-NN año a año, así como, el rendimiento obtenido con cada transacción efectuada. En segundo lugar, se agregarán dichos datos para computar la rentabilidad derivada de la estrategia

de inversión planteada.

Para el primer estudio se incluyen las Figuras 5.19, 5.20 y 5.21. Las dos primeras muestran un histograma que representa la distribución sectorial del mejor valor de forma anual. El principal problema que presenta la primera de ellas (Figura 5.19) es que está sesgada por la presencia de la clase mayoritaria de empresas *Healthcare*, como se indicó en la Tabla 5.3. Para enmendar este error, se toman los porcentajes calculados año a año como las probabilidades de ocurrencia a priori, normalizando así los valores según su distribución. A partir de ahí, los datos de la Figura 5.20 son comparables entre sí. Por otra parte, la Figura 5.21 muestra la distribución anual de los rendimientos del mejor valor del K-NN por sectores.

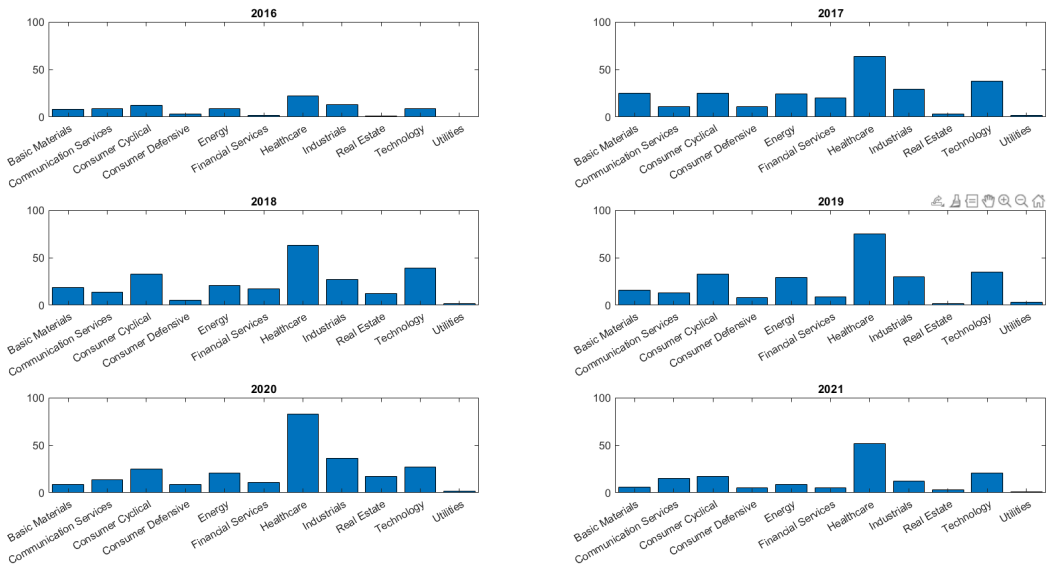


Figura 5.19. *Gráfico de barras del mejor valor según el sector - sesgado*

En vista a los resultados, se pueden realizar una serie de observaciones:

- Como se puntualizó en la sección anterior, se pone de manifiesto como en los periodos de recesión, como el ocurrido en el año 2020 derivado del COVID, se favorecen los sectores cíclicos y se castiga el de *real estate*.
- Por contra, los periodos de recuperación, como el 2021, favorecen la compra

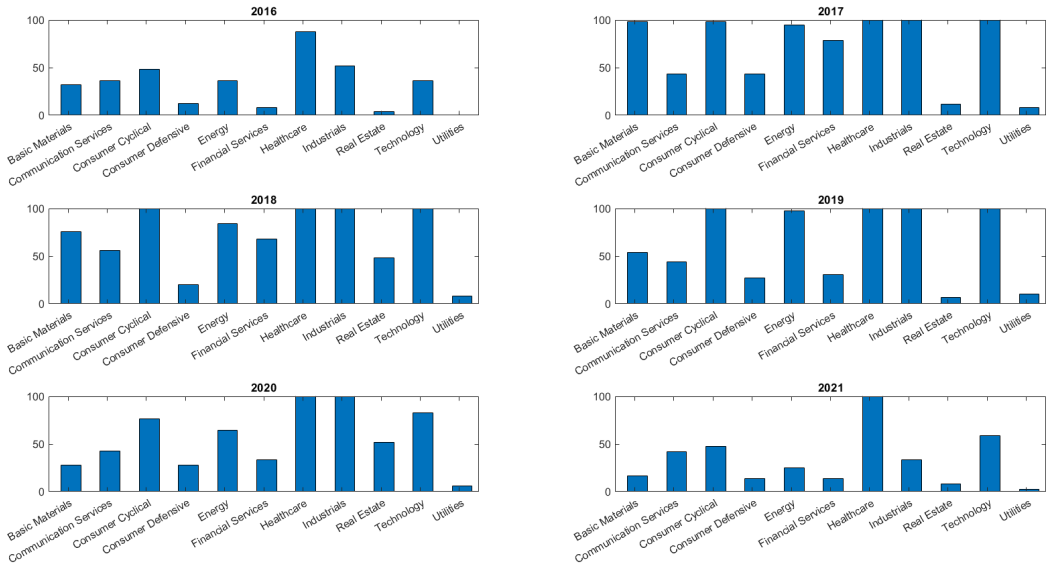


Figura 5.20. *Gráfico de barras del mejor valor según el sector*

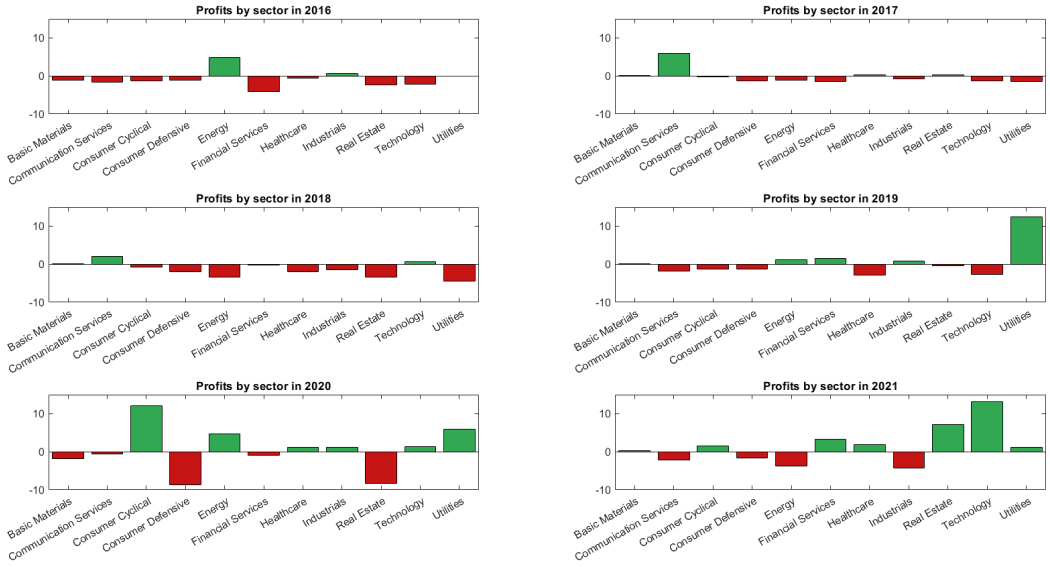


Figura 5.21. *Rentabilidad media anual del mejor valor por sector*

de bienes inmuebles a bajo precio, promoviendo así la compra de acciones de este último sector.

- Con respecto al aspecto del histograma de la Figura 5.20, se puede observar pares de patrones similares en los años 2016 y 2021, 2017 y 2019 y, 2018 y 2020.
- Finalmente, cabe destacar cómo el sector de *Utilities* es el que menos veces aparece en la clasificación siendo, sin embargo, el que mayor contribuye (proporcionalmente) a los beneficios conseguidos.

En segundo lugar, si se analizan los datos anteriores de forma agregada, se pueden comparar los resultados con respecto al rendimiento del K-NN en el caso base. Para ello, se incluyen las Figuras 5.22 y 5.23 en las que se muestra la evolución de la rentabilidad en bruto que se obtiene tras la aplicación del backtesting que actúa en el nivel inferior. Se representan diferentes curvas que hacen referencia a los beneficios derivados de la compra (sin acumular valores consecutivos) del mejor al quinto valor. Tal y como se puede observar, la modificación introducida no sólo mejora el valor final de la rentabilidad del modelo, sino también su distribución a lo largo del tiempo, reduciendo al mismo tiempo el riesgo asumido.

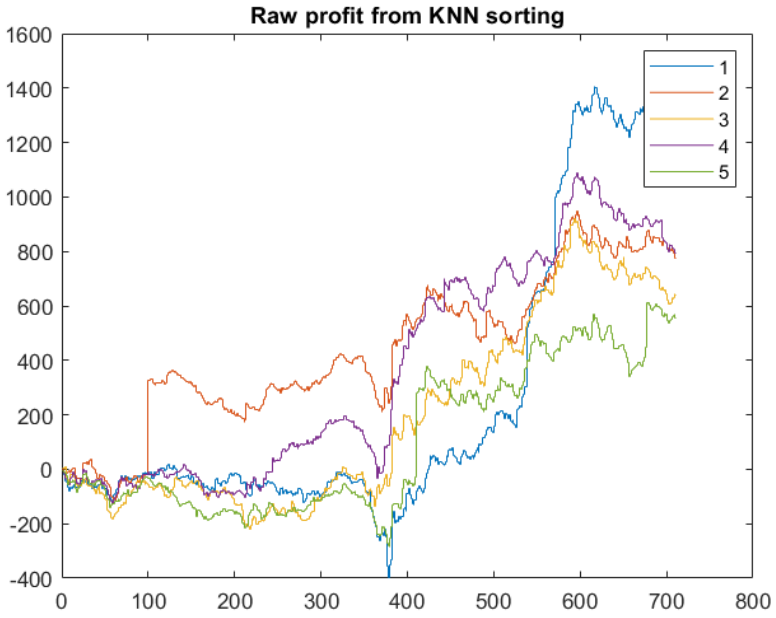


Figura 5.22. *Rentabilidad en bruta de los 5 mejores valores - Caso base*

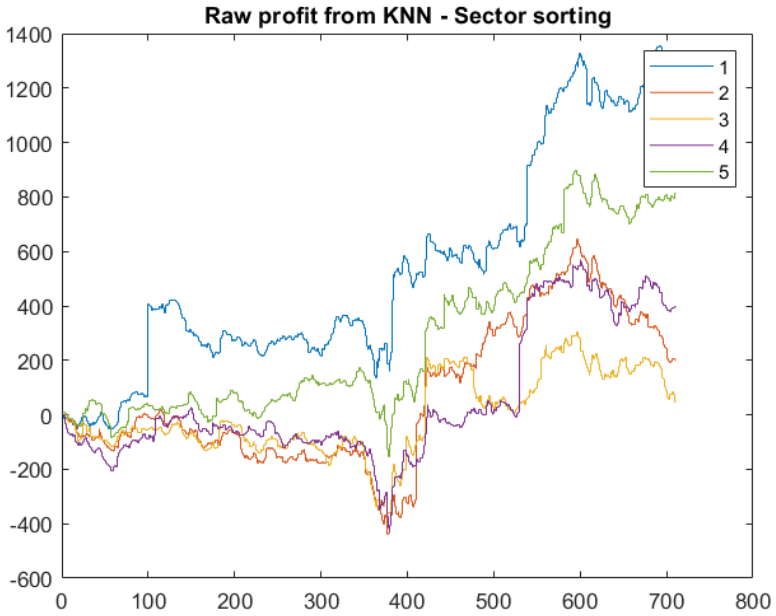


Figura 5.23. *Rentabilidad en bruta de los 5 mejores valores - Caso sectorial*

5.3. Optimización del número de vecinos

La selección del valor óptimo de K es el requisito esencial para crear un modelo K-NN decente y preciso. Si se selecciona un valor de K demasiado bajo, el modelo se vuelve demasiado específico y no consigue generalizar bien. Además, tiende a ser sensible al ruido. Si se selecciona K demasiado grande, el modelo se vuelve demasiado generalizado y no logra predecir con precisión los puntos de datos. Una forma comúnmente empleada en la práctica consiste en representar el error de predicción cometido según el número de vecinos considerado. Sin embargo, esta solución no es del todo eficaz en el caso de estudio, ya que el objetivo de este proyecto no es predecir la serie de precios temporales, sino diseñar estrategias de inversión rentables. No obstante, uno puede pensar que a mayor grado de precisión en los resultados, mayor información de calidad reciben los backtesting para el cómputo de las rentabilidades finales del modelo. De este modo, se analizará como principal punto de partida qué valor de K proporciona un menor error en cuanto a la predicción, y, posteriormente, se analizará el impacto real de dicho parámetro sobre la rentabilidad final.

5.3.1. Impacto de K en el error de predicción

Para ello, en primer lugar, se adjuntan las Figuras 5.24 y 5.25. En la primera de ellas se representa el error global a lo largo de todo el periodo, mientras que en la segunda se desagrega por años para comprobar si las características son igual de importantes durante todo el tiempo o dependen de la confluencia de unos factores específicos.

En vista a los resultados obtenidos se pueden realizar una serie de conclusiones:

- El mejor valor de K en todo el instante simulado es de 150 vecinos, tanto para el error computado de forma global como calculado de manera anual; seguido en ambos casos de $K = 500$ *vecinos*
- En la Figura 5.24 se puede observar cómo a partir de 1000 vecinos el error es prácticamente invariable.
- En la Figura 5.25, se puede apreciar cómo el error aumenta considerablemente en los años 2020 y 2021. Este fenómeno merece ser estudiado en detalle. Para ello, se incluye en la Figura 5.26 un análisis de correlaciones de la rentabilidad real frente a la mediana predicha año a año, y tomando como valor de K 150 vecinos.

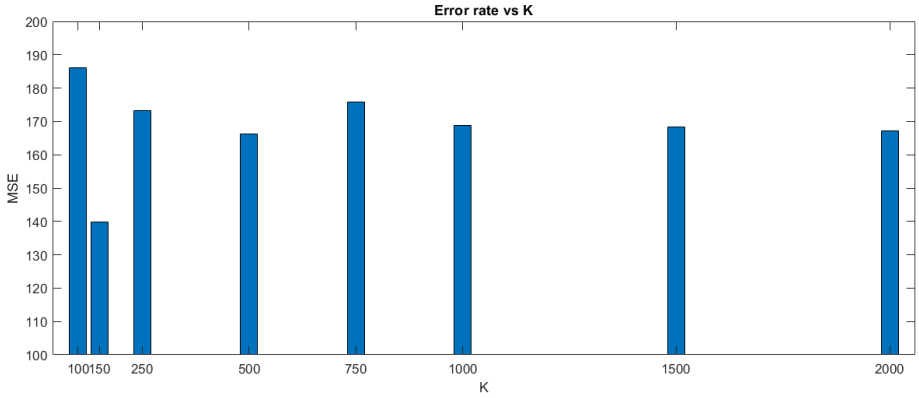


Figura 5.24. *Error en la predicción vs K*

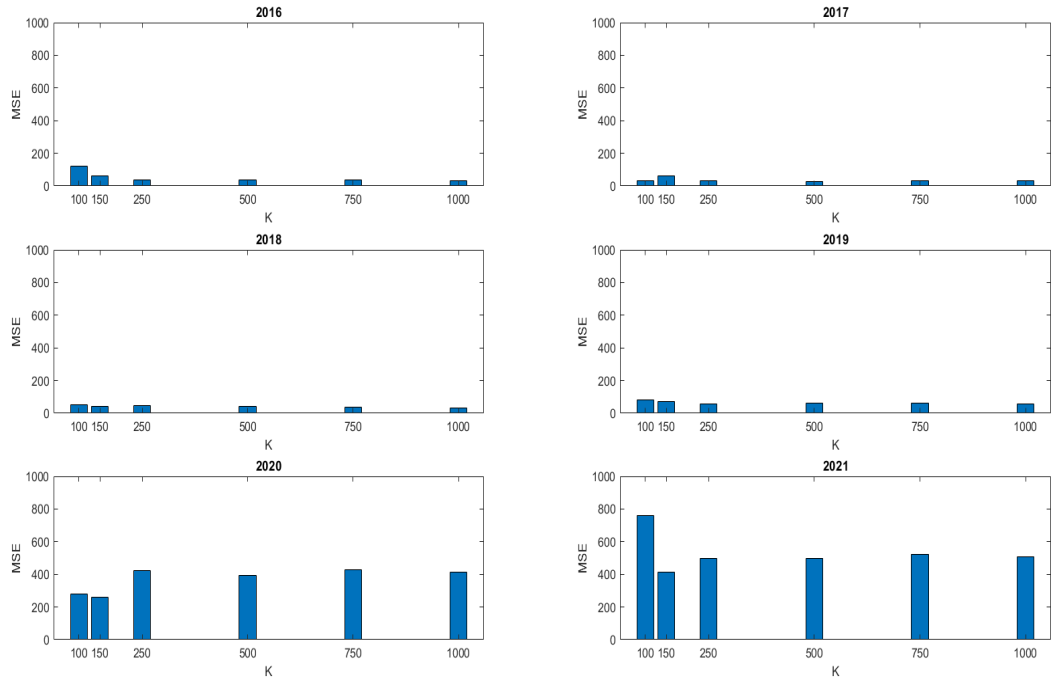


Figura 5.25. *Error en la predicción vs K por años*

Tal y como se puede observar en la Figura 5.26 la dispersión de los datos aumenta considerablemente en los últimos dos años y esto tiene un impacto enorme en la función de pérdida. El error cuadrático medio se computa como la media de

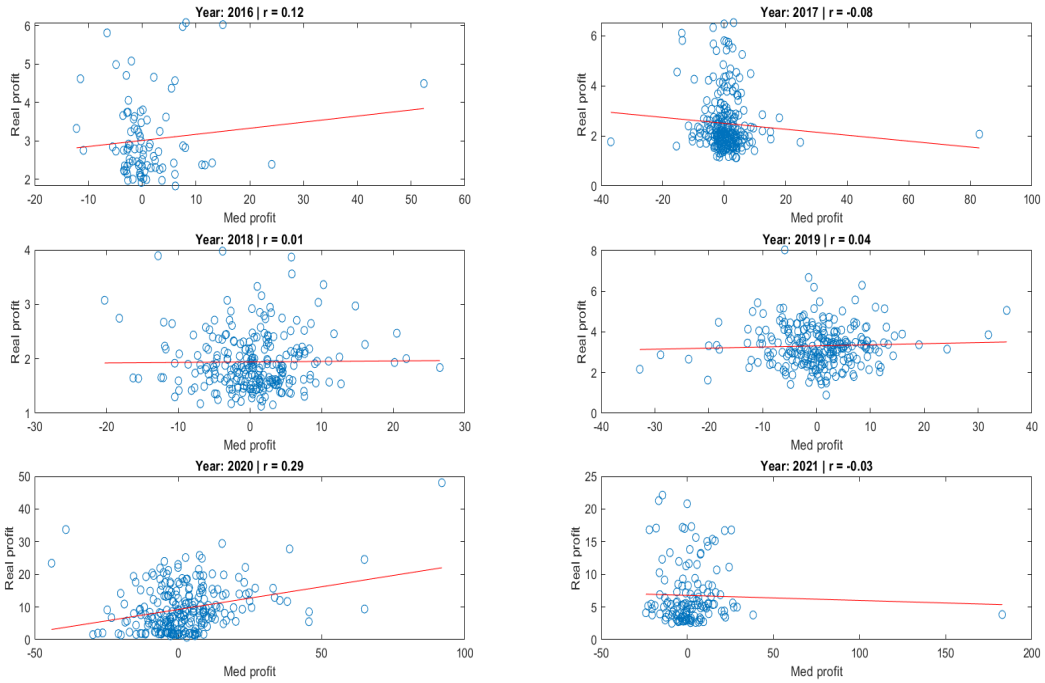


Figura 5.26. *Análisis de correlaciones por año con* $K = 150$

la suma de los cuadrados de la diferencia entre la rentabilidad real y la mediana predicha por el modelo. Por lo tanto, se trata de una métrica muy sensible en caso de valores elevados, como ocurre a lo largo de los años 2020 y 2021. Para solventar dicho efecto, se generan las gráficas de la Figura 5.28 teniendo en cuenta como métrica el error medio absoluto (MAE).

En vista a los resultados se puede apreciar cómo ahora el error es más uniforme a lo largo de los años. Además, se ponen en tela de juicio las afirmaciones realizadas sobre el número de vecinos óptimo, necesitando un número más amplio de combinaciones a observar. Por otra parte, mediante la gráfica adjunta en la Figura 5.28 se pretende responder a la pregunta planteada inicialmente en la motivación del capítulo. Es decir, se pretende contrastar si convendría aplicar un K-NN adaptativo según cada empresa, tal y como sugieren Sun et al. (2010) [44]. Para ello, se seleccionan 4 empresas americanas con gran capitalización bursátil y pertenecientes a diferentes sectores. Tal y como se puede observar, efectivamente, el número de vecinos óptimo difiere según el valor a predecir.

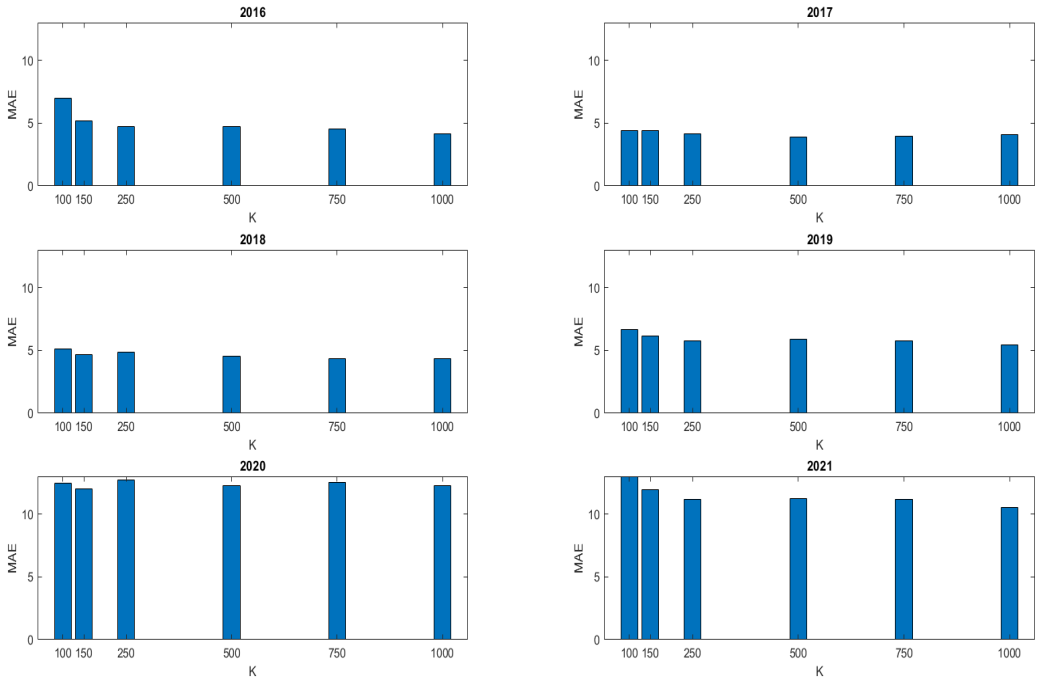


Figura 5.27. *MAE en la predicción vs K por años*

5.3.2. Impacto de K en la rentabilidad final

Como ya se ha indicado anteriormente, las predicciones en el ámbito financiero no deben de ser consideradas un fin en sí mismo, sino un medio necesario para construir estrategias de inversión rentables. Por lo tanto, de aquí surge la necesidad de completar el análisis de la influencia de K sobre la rentabilidad real obtenida tras la compra del mejor valor en la ordenación del K-NN. Para ello, se adjuntan las gráficas 5.29 y 5.30 en las que se incluyen el rendimiento neto según los vecinos considerados, dado un horizonte de predicción fijo e igual a una semana. Debido a la presencia de outliers en los resultados de la Figura 5.29, para establecer las conclusiones es necesario acudir a los obtenidos con la Figura 5.30. Se puede observar cómo el número óptimo de vecinos en el periodo simulado es de 500, seguido de 250 y 150. Por lo tanto, sería necesario ampliar la búsqueda de valores entorno a ese rango.

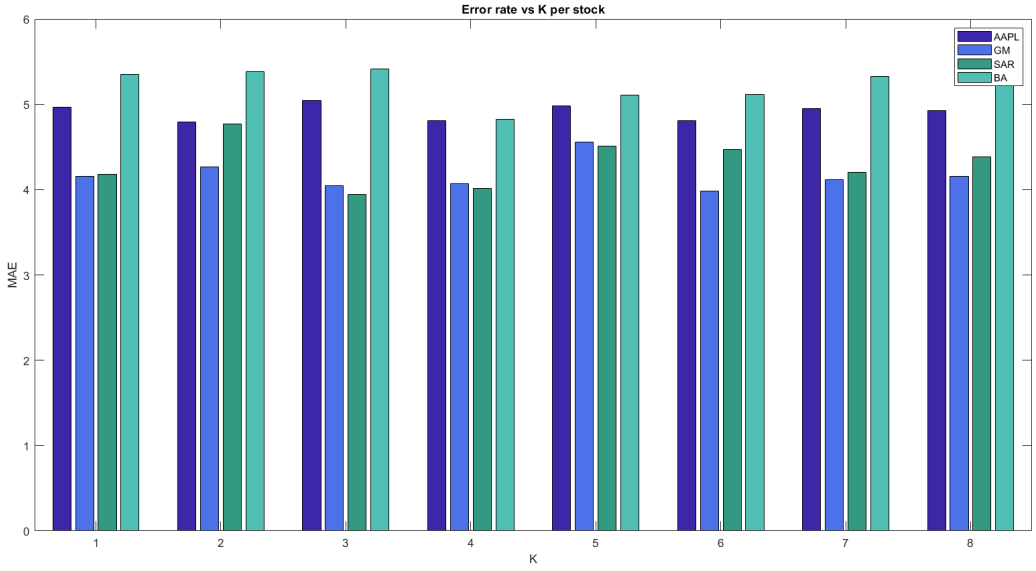


Figura 5.28. *MAE en la predicción vs K para distintas empresas*



Figura 5.29. *Influencia de K sobre la rentabilidad del K-NN*

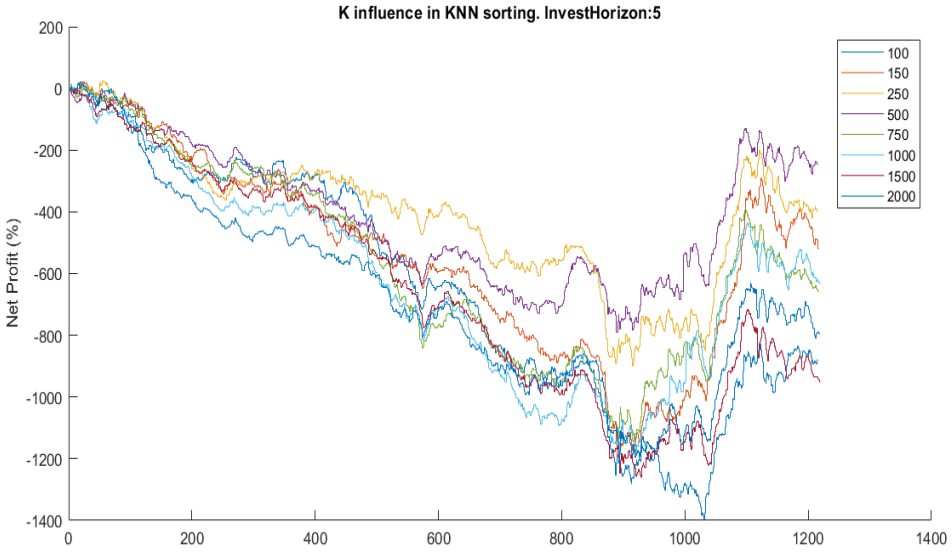


Figura 5.30. *Influencia de K sobre la rentabilidad del K-NN sin outliers*

5.4. Optimización del horizonte de predicción

En esta sección se pretende analizar el efecto que tiene el horizonte de predicción sobre la rentabilidad final del K-NN. Bajo el esquema actual, la estrategia de inversión diseñada tiene una duración fija e igual a dicho parámetro tratándose, pues, de un método completamente ciego. No obstante, también se almacena la información de la predicción de las rentabilidades medias, mediana y la tasa de acierto para el conjunto de días comprendido desde la compra hasta la semana siguiente, que es cuando se efectúa la venta del título. Gracias a ello, se puede evaluar la sensibilidad de la rentabilidad final según cada valor. Además, también se deja para futuros desarrollos la opción de diseñar una estrategia de inversión dinámica en cuanto a su duración. Para ilustrar esto último, supóngase que se produce la compra de un valor X en un día determinado k. Para generar dicha señal de compra se ha tenido que calcular la rentabilidad predicha a 5 días. De este modo, lo que se propone es que se cierre la operación, no obligatoriamente a la semana, sino en el momento en el que se produzca (si se llega a producir) un codo en la rentabilidad acumulada, ya que no tendría sentido mantener dicho valor X un día más en cartera.

Para proceder con el estudio se adjuntan las Figuras 5.31 y 5.32, en las que se representan de nuevo las rentabilidades del K-NN (tomando como número de vecinos 500) con outliers (Figura 5.31) y sin outliers (Figuras 5.32). Se puede observar claramente cómo el horizonte semanal es el mejor, por lo que, en vista de estos resultados merecería la pena explorar horizontes de inversión superiores.

Finalmente, se incluye en la Figura 5.33 una comparación de las rentabilidades del caso base y de la obtenida mediante los parámetros óptimos ($K = 500$ *y* $predHorizon = 5$). Tal y como se puede observar, se produce una mejora muy significativa de los resultados.

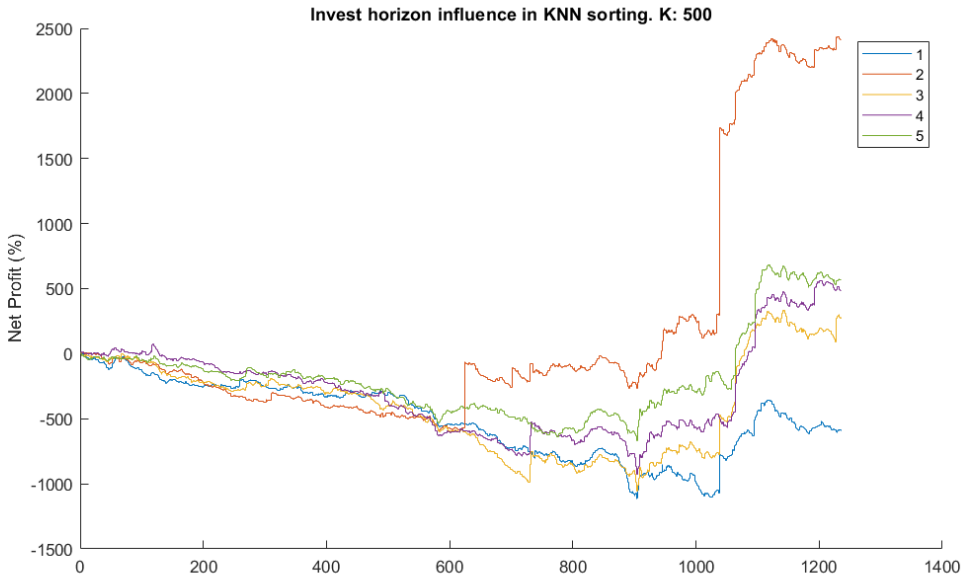


Figura 5.31. *Influencia del horizonte de predicción sobre la rentabilidad del K-NN sin outliers*

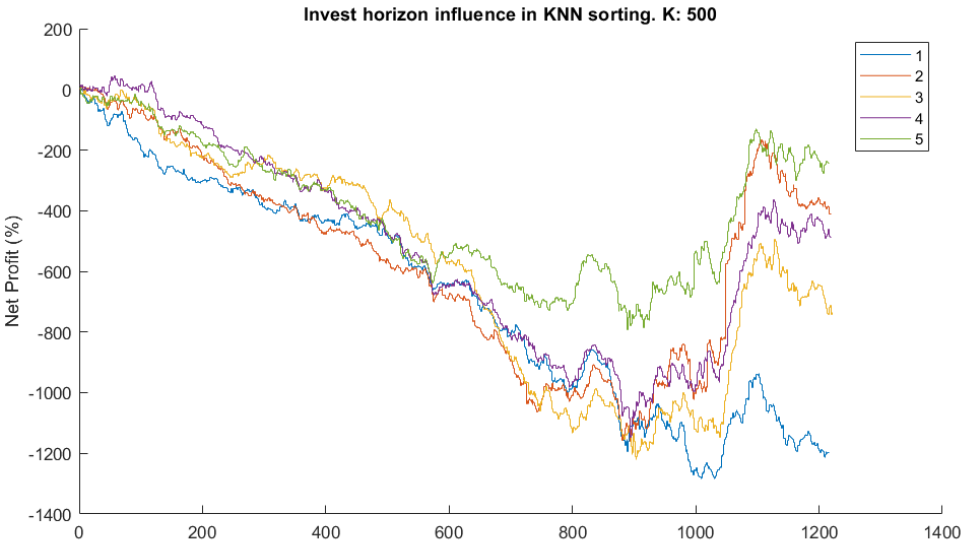


Figura 5.32. *Influencia del horizonte de predicción sobre la rentabilidad del K-NN sin outliers*

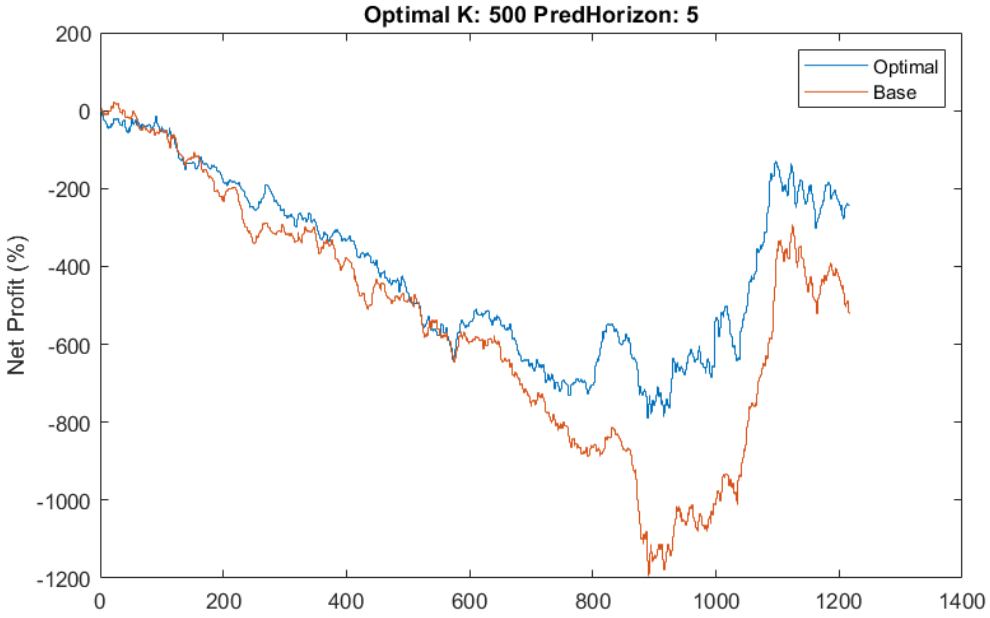


Figura 5.33. *Mejora de la rentabilidad con K y horizonte de predicción óptimos*

Capítulo 6

Resultados

El objetivo fundamental de este capítulo es exponer el procedimiento llevado a cabo para contrastar el rendimiento de la estrategia de inversión una vez que ha sido propiamente optimizada. Para ser rigurosos en la metodología implementada, se expondrá de forma detalla el análisis secuencial llevado a cabo en los sucesivos apartados.

- En primer lugar, se comparará el rendimiento del algoritmo frente a la estrategia de Buy & Hold de tres índices americanos: Dow Jones Industrial (DJI), Nasdaq (NDAQ) y Russell 3000 (RUA).
- En segundo lugar, se simularán distintos escenarios para evaluar el impacto de los costes de transacción.
- En tercer y último lugar, se estudiará el *performance* del algoritmo desde un punto de vista cuantitativo y de gestión financiera.

Para realizar estos análisis se hace uso de la librería *Pyfolio* de *Python*. Se trata de una fuente de código abierto que proporicona herramientas operativas para evaluar el rendimiento y el riesgo de diferentes carteras de valores [1].

6.1. Comparación del modelo con los benchmarks

El objetivo de esta sección es evaluar el rendimiento de la estrategia de inversión con respecto a los benchmarks descritos anteriormente. Además de los índices de referencia tradicionales (DJI y NDAQ), se utiliza el Russell 3000, ya que la

[1] https://quantopian.github.io/pyfolio/

mayor parte de las empresas que lo componen forman parte de la lista de Tickers diseñada para construir el *dataset* de este algoritmo. Para ello, se simula (con el backtesting explicado en los capítulos anteriores) cuál hubiera sido el rendimiento del algoritmo en el periodo temporal comprendido entre el 1 de agosto de 2016 y el 30 de julio de 2021. [2]

Para proceder con este análisis comparativo se computan dos tipos de rendimientos diferentes: *trailing returns* y *rolling returns*. Los *trailing returns* se remontan a una fecha concreta para obtener la rentabilidad anualizada de un fondo/-portfolio durante un periodo de tiempo específico. En cambio, los *rolling returns* muestran los rendimientos en ciclos superpuestos que comienzan el primer día del mes, remontándose a los datos que hay disponibles. El objetivo es mostrar la frecuencia y la magnitud de los periodos buenos y malos de una inversión. La mayoría de los expertos en el mundo financiero afirman que los *rolling returns* proporcionan una imagen más fidedigna del rendimiento de un fondo de inversión [13].

Los *trailing returns* pueden dar resultados variables. Como el mercado se encuentra sumergido en continuas subidas y bajadas, la rentabilidad final puede variar mucho de un punto a otro. Básicamente esto se debe a tres factores fundamentales:

- La rentabilidad tiene un sesgo de recencia, ya que se basa en una única lectura puntual de la rentabilidad del fondo o portfolio.
- La trayectoria del mercado y el rendimiento del fondo final del periodo tienen una influencia indebida en la rentabilidad final.
- Un fuerte aumento de la rentabilidad reciente hará que los rendimientos de todos los periodos anteriores parezcan saludables. Del mismo modo, un descenso en la rentabilidad reciente hará que todos los rendimientos parezcan débiles.

Para paliar estos efectos, los *rolling returns* miden el rendimiento del fondo en todas las escalas de tiempo, sin incluir ningún tipo de sesgo. Para ello, promedian las instancias de rendimientos puntuales en el periodo deseado. Esta rentabilidad no sólo mide el rendimiento a 3, 5 o 10 años, sino que también considera varios

[2]El principal problema de comparar cualquier estrategia de inversión con el comportamiento de un índice es que este último no es negociable en los mercados. No obstante, se supone que el comportamiento de un ETF del mismo (que sí es tradeable) es sustancialmente análogo.

bloques dentro de este periodo en distintos intervalos temporales. Para ilustrar este fenómeno se adjuntan en las Tablas 6.1 y 6.2 los resultados del fondo Edelweiss Large Cap según la métrica empleada [3].

Return as on	Sensex value	3 year CAGR	5 year CAGR
28-Aug-18	38,896	14,40	20,37
26-Oct-18	33,349	7,03	12,85
2-Apr-19	39,056	14,16	14,23

Tabla 6.1. *Trailing returns para el fondo de inversión Edelweiss Large Cap*

Return as on	Sensex value	3 year CAGR	5 year CAGR
28-Aug-18	38,896	13,16	14,51
26-Oct-18	33,349	13,07	14,51
2-Apr-19	39,056	12,94	14,52

Tabla 6.2. *Rolling returns para el fondo de inversión Edelweiss Large Cap*

Como se puede observar, los resultados proporcionados en la Tabla 6.2 son completamente atemporales, y están libre de sesgos. A continuación, se adjuntan en las Figuras 6.1 y 6.2 los *trailing* y *rolling returns* del método y los benchmarks.

Tal y como se puede apreciar, en ambos casos se bate a todos los índices al final del periodo simulado. Aunque sea necesario esperar hasta el último año para superar a los becnhmarks, como afirma el ilustre Warren Buffet *"la Bolsa es un mecanismo por el cual se transfiere dinero del impaciente al paciente"*, por lo que la paciencia se paga a un precio alto. No obstante, es innegable que una clara limitación del método es la heterogeneidad de la distribución de la rentabilidad en el tiempo. Esto es una consecuencia directa del propio algoritmo diseñado. Al fin y al cabo, lo que realmente busca el K-NN son patrones de series temporales muy rentables a corto plazo, y esto es lo que provoca que se encuentren valores más atípicos. Por lo tanto, este fenómeno es el causante de que se pierda robustez de forma generalizada.

[3]La tasa de crecimiento anual compuesto o Compound annual growth rate (CAGR) es la tasa media de crecimiento anual de una inversión a lo largo de un periodo de tiempo determinado superior a un año. Representa una de las formas más precisas de calcular y determinar la rentabilidad de los activos individuales, las carteras de inversión y cualquier cosa que pueda aumentar o disminuir su valor con el tiempo.

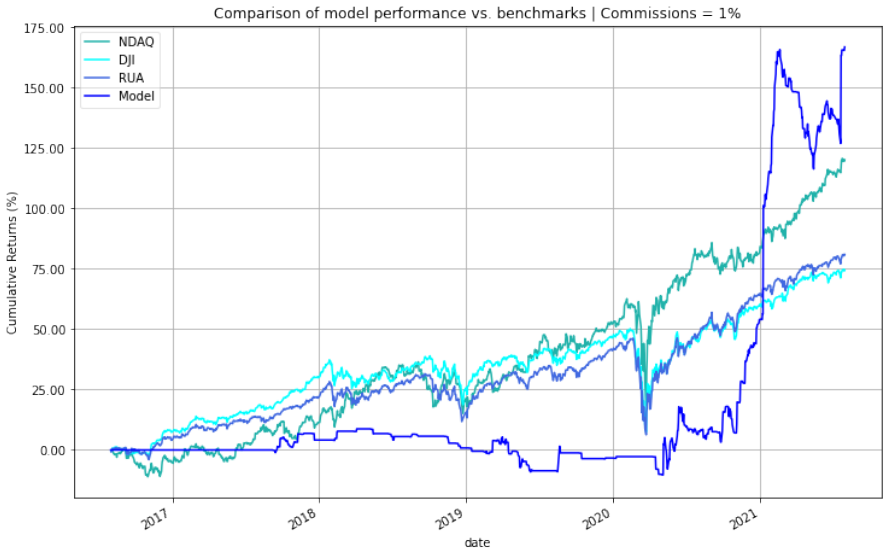


Figura 6.1. *Comparación de los trailing returns*

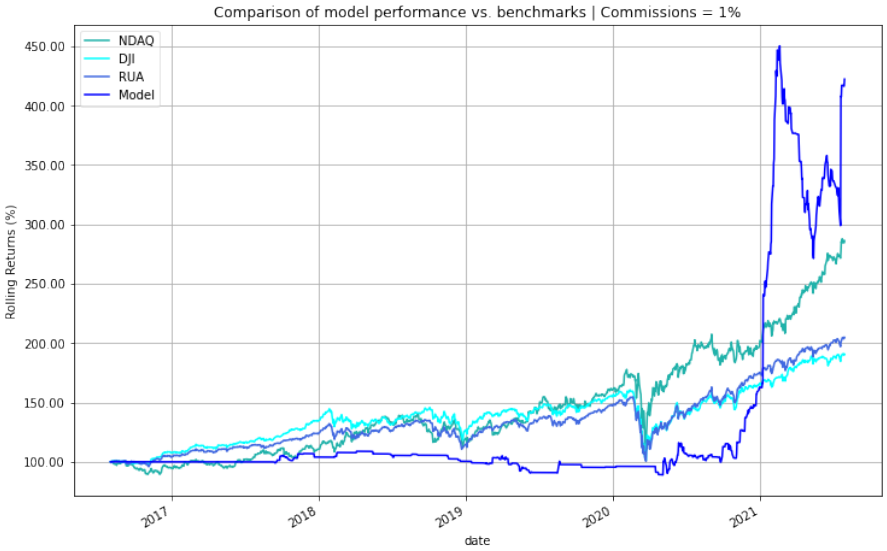


Figura 6.2. *Comparación de los rolling returns*

6.2. Análisis del impacto de las comisiones

En este tipo de algoritmos tan agresivos y a corto plazo, las comisiones determinan un papel fundamental en el rendimiento del mismo. De darse las condiciones

necesarias (superación de los umbrales de la tasa de acierto y de la mediana de la rentabilidad) el algoritmo puede llegar a ejecutar diariamente 2 operaciones (el cierre del valor de la semana anterior y la compra del nuevo título). Esto supone que, en el peor de los escenarios posibles, se puede llegar a perder en el transcurso de los 5 años alrededor de un 252 % simplemente en comisiones. Las gráficas del escenario anterior se han construido teniendo en cuenta unos costes de transacción del 0.5 % en cada operación (un 1 % en la conjunta). A continuación, se incluyen en las Figuras 6.3, 6.4, 6.5 y 6.6 los *trailing* y *rolling returns* del algoritmo cuando se consideran unos costes de transacción del 0.25 % (en cada operación) y del 0 %, respectivamente.

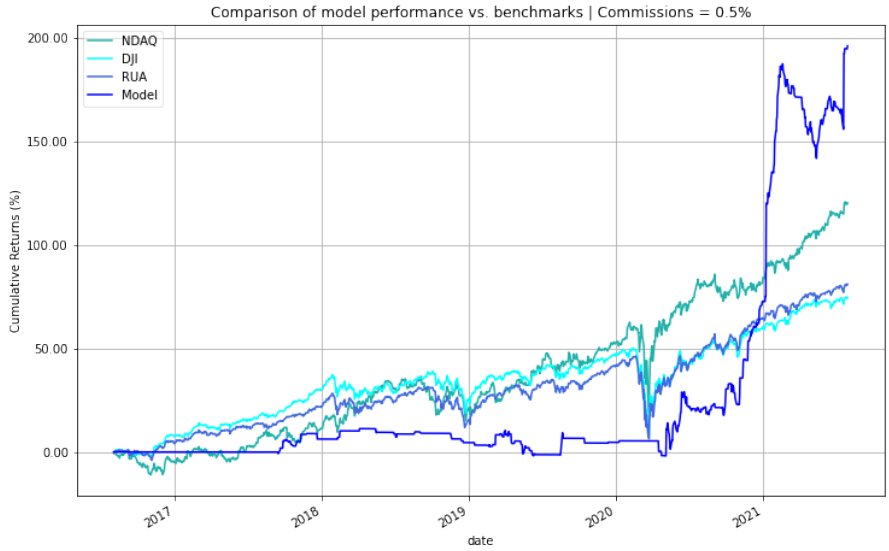


Figura 6.3. *Comparación de los trailing returns - Comisiones totales del 0.5 %*

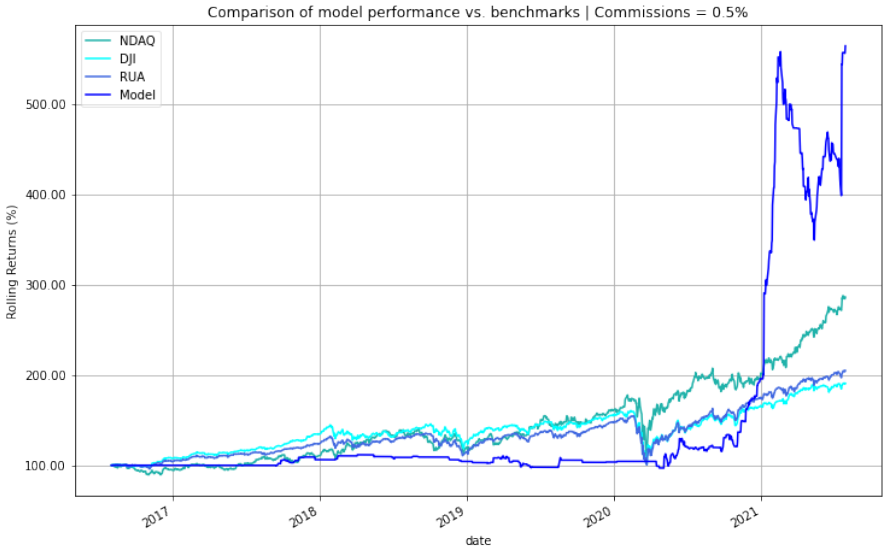


Figura 6.4. *Comparación de los rolling returns - Comisiones totales del 0.5 %*

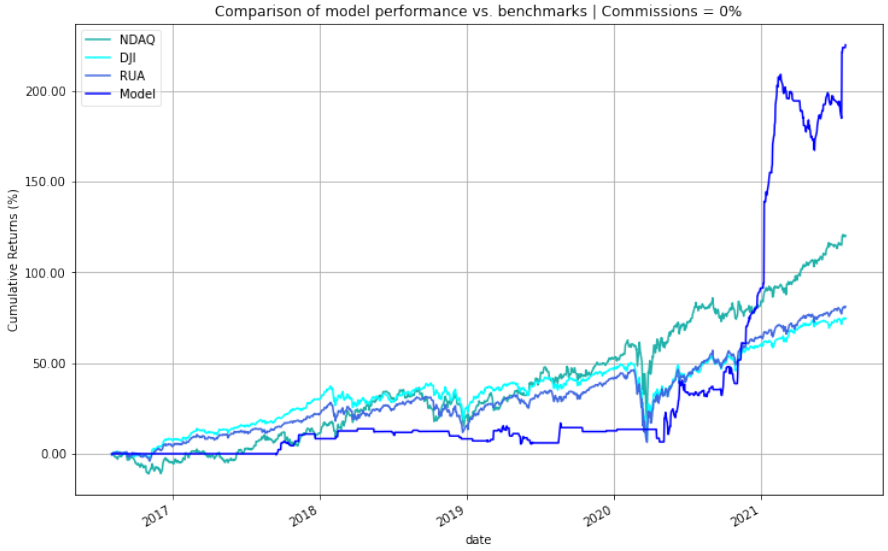


Figura 6.5. *Comparación de los trailing returns - Sin costes de transacción*

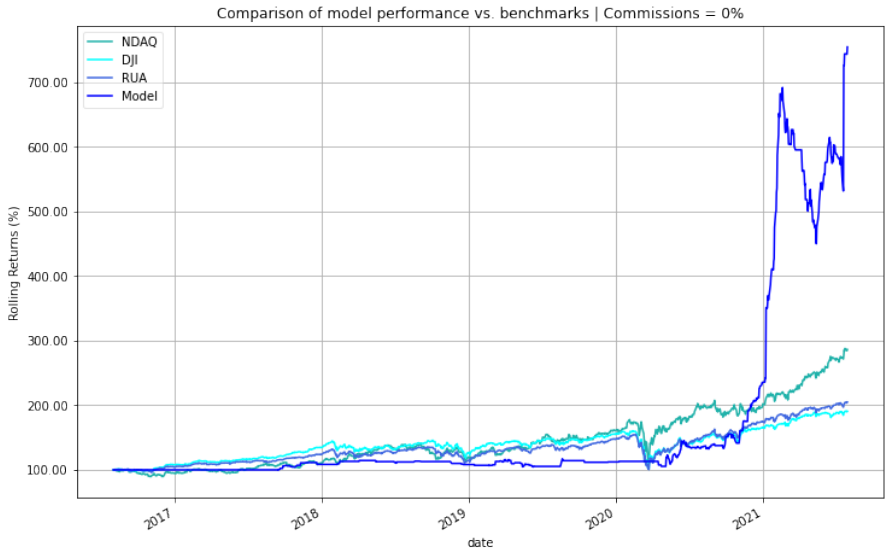


Figura 6.6. *Comparación de los rolling returns - Sin costes de transacción*

Tal y como se puede observar, la rentabilidad acumulada final se disminuye en casi un 35 % por los costes de transacción. La magnitud de este fenómeno se puede apreciar en la Figura 6.7, en la que se representa un gráfico muy usado en el ámbito financiero (*underwater plot*), que se centra únicamente en las pérdidas y muestra el tiempo que tardó el valor de la cartera en recuperar el máximo anterior tras sufrir una pérdida prolongada. Este gráfico permite distinguir fácilmente entre períodos de pérdidas normales y prolongadas. Se puede apreciar de forma muy visual qué parte de la pérdida se debe únicamente al hecho de tener unos costes de transacción más elevados y qué parte de la misma se debe a que el valor comprado no generase beneficios positivos.

En relación con la gráfica anterior, se puede realizar un análisis muy interesante. Este consiste en determinar cuáles son los tres periodos de máximos retornos negativos consecutivos según las comisiones consideradas. Para ello, se adjuntan las Tablas 6.3, 6.4, 6.5 y 6.6 que incluyen la información de la máxima racha negativa según los periodos, la duración de dicha racha y la fecha del pico, para el caso de comisiones del 1 %, 0,5 %, 0 % y el índice Russell 3000, respectivamente.

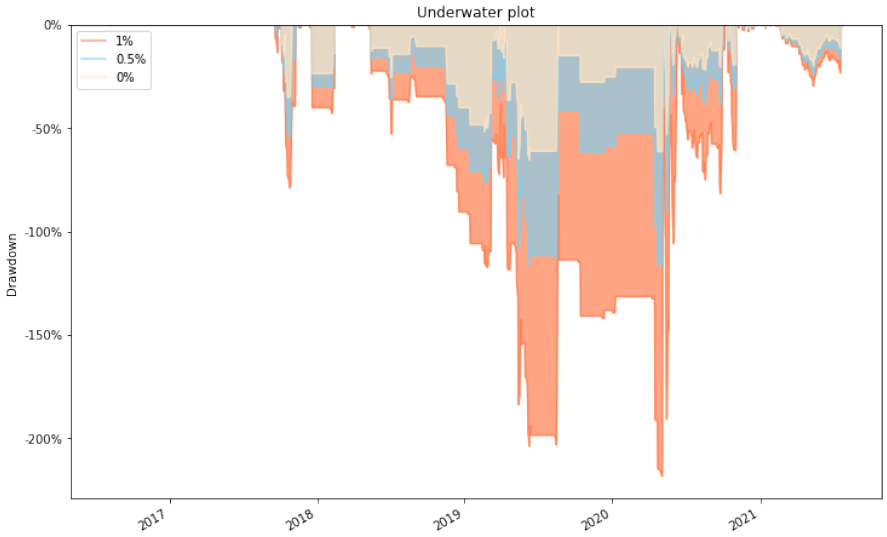


Figura 6.7. *Underwater plot - Efecto de las comisiones*

Periods	Net drawdown in %	Peak date	Valley date	Recovery date	Duration
0	39,75	2021−02−18	2021−05−13	*NaT*	*NaN*
1	18,25	2018−05−08	2020−05−04	2020−06−11	548
2	14,04	2020−06−12	2020−09−24	2020−11−04	104

Tabla 6.3. *Worst down periods for 1 %*

Periods	Net drawdown in %	Peak date	Valley date	Recovery date	Duration
0	37,39	2021−02−18	2021−05−13	2021−07−29	116
1	13,22	2018−05−08	2020−05−04	2020−05−08	524
2	12,03	2020−05−08	2020−05−15	2020−05−26	13

Tabla 6.4. *Worst down periods for 0.5 %*

Periods	Net drawdown in %	Peak date	Valley date	Recovery date	Duration
0	34,95	2021−02−18	2021−05−13	2021−07−20	109
1	11,58	2020−05−08	2020−05−15	2020−05−26	13
2	10,04	2019−08−21	2020−04−23	2020−05−05	185

Tabla 6.5. *Worst down periods for 0 %*

Periods	Net drawdown in %	Peak date	Valley date	Recovery date	Duration
0	35,09	2020−02−19	2020−03−23	2020−08−21	133
1	18,41	2018−09−20	2018−12−21	2019−07−02	204
2	9,97	2018−01−26	2018−02−08	2018−08−07	138

Tabla 6.6. *Worst down periods for RUA*

En vista a los resultados se pueden establecer una serie de conclusiones:

- El hecho de reducir los costes de transacción, mejora aproximadamente un 2 % cada periodo de máxima racha negativa. La explicación de que dicho factor sea una constante tiene sentido porque se ejecutan las mismas operaciones, y a todas las comisiones les afecta por igual.
- Si se compara la tercera peor racha en el escenario de 1 % de comisiones y el de sin costes, se puede observar un efecto muy contraintuitivo. A pesar de que se reduce la pérdida en rentabilidad en un 4 %, se prolonga la duración de la caída en el tiempo (104 frente a 185 días).
- Algo similar ocurre si se compara el segundo peor caso de las simulaciones con 1 % y 0.5 % de comisiones y el rendimiento del índice. A pesar de que en ambos escenarios la máxima caída es menor que la del Russell 3000, la duración se prolonga casi un año más en el tiempo. Esto se puede apreciar en la Figura 6.8.

Por lo tanto, con este análisis se ha revelado el enorme impacto que tienen los costes de transacción en la estrategia de inversión diseñada. En consecuencia, es muy importante para el inversor encontrar brokers que le permitan operar con comisiones muy bajas para así poder sacarle el máximo partido al método implementado. Finalmente, también pone de manifiesto que las investigaciones revisadas

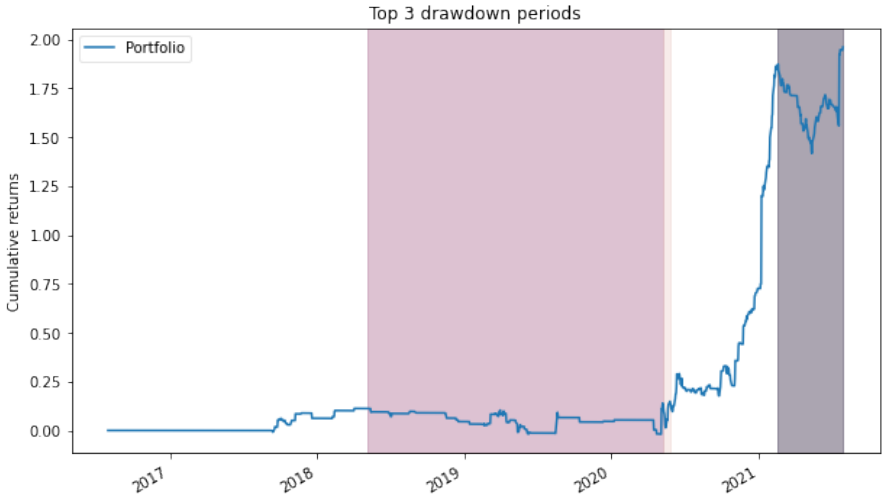


Figura 6.8. *Top 3 Max Drawn del modelo*

que no incluyen los costes de transacción, no pueden considerarse válidas.

6.3. Análisis financiero de la estrategia de inversión

Finalmente, para concluir con este capítulo se va a analizar el rendimiento de la estrategia de inversión propuesta teniendo en cuenta algunas de las gráficas e indicadores que emplean los profesionales del sector. En primer lugar, conviene definir los principales ratios empleados al medir la eficiencia de un portfolio:

- **Ratio de Calmar**

  El ratio de Calmar es un parámetro del rendimiento entre la rentabilidad anualizada y el máximo drawdown del activo en cuestión. Se computa teniendo en cuenta los datos de los últimos 36 meses. El drawdown se puede definir como la diferencia entre el máximo de referencia de una curva y el mínimo, en un período de tiempo. Es una medida de riesgo que se utiliza en el ratio de Calmar, en lugar de la volatilidad. Se computa de acuerdo a la siguiente relación:

$$\text{Ratio de Calmar} = \frac{\text{Rentabilidad anualizada}}{\mid \text{Máximo drawdown} \mid} \tag{6.1}$$

- **Ratio de Sharpe**

  Este ratio financiero fue desarrollado por el nobel de economía William F. Sharpe para saber si la rentabilidad de una inversión se debe a una decisión inteligente o, si por el contrario, es resultado de haber asumido más riesgo del debido. Es un ratio que calcula la rentabilidad ajustada según su riesgo. El ratio de Sharpe se calcula restando la rentabilidad de un activo sin riesgo (como puede ser la deuda a 10 años de Alemania o más formalmente el EONIA) a la rentabilidad de la inversión y dividendo el resultado entre el riesgo, calculado como la desviación típica de la rentabilidad de la inversión, tal y como se observa en la siguiente fórmula:

$$S = \frac{R - R_f}{\sigma} \tag{6.2}$$

- **Ratio de Sortino**

  Esta métrica muestra el exceso de rendimiento por encima de un determinado objetivo por unidad de riesgo a la baja.

$$S = \frac{R - T}{DR} \tag{6.3}$$

  donde DR es la desviación estándar de los rendimientos negativos (downside risk) representado matemáticamente por:

$$DR = \sqrt{\int_{-\infty}^{t} (T - r)^2 f(r) dr} \tag{6.4}$$

  La fórmula del ratio de Sortino, es similar a la del ratio de Sharpe. En su cálculo considera el exceso de rentabilidad esperada ($Rp \breve{} Rf$) sobre el activo sin riesgo; sin embargo, la medida de volatilidad utilizada para corregir el ratio por riesgo no tiene nada que ver con la tradicional desviación estándar (Sp), utilizándose en su lugar, una medida de dispersión de los rendimientos negativos. Este indicador fue desarrollado para diferenciar entre "la buena" y "la mala" volatilidad en el ratio de Sharpe. De esta forma, si en el período una inversión financiera tuvo un extraordinario comportamiento al alza, la mayor volatilidad o desvío estándar que mostrara por esta causa, no sería considerada en la parte del análisis de la rentabilidad que considera el riesgo, medido éste como la probabilidad de que se produzcan pérdidas. Los inversores, de acuerdo con Sortino, no están asustados por las oscilaciones de su

inversión en sí misma, sino que se sienten presionados sólo en un entorno de resultados que no lleguen a lo que ellos consideran un mínimo aceptable (en muchos casos, el activo libre de riesgo).

- **Ratio de Omega**

  Se define como la relación ponderada por la probabilidad de las ganancias frente a las pérdidas para un objetivo de rentabilidad determinado. El ratio es una alternativa al ratio de Sharpe y se basa en la información que la relación de Sharpe descarta. El Omega se calcula creando una partición en la distribución de la rentabilidad acumulada para crear un área de pérdidas y un área de ganancias en relación con este umbral. Se computa matemáticamente atendiendo a la siguiente fórmula:

$$\Omega(\theta) = \frac{\int_{\theta}^{\infty}[1 - F(r)]dr}{\int_{-\infty}^{\theta} F(r)dr} \tag{6.5}$$

  donde F es la distribución de probabilidad acumulada de la función de retornos y θ es un umbral objetivo que establece qué se considera perdida y qué ganancia. Un valor de Omega alto implica que la cartera proporciona más ganancias relativas que pérdidas en relación con el umbral establecido.

- **Ratio de cola**

  Este ratio mide la probabilidad de que una inversión se vaya a alejar más de tres desviaciones estándar de su media de lo que resultaría de una distribución normal.

A continuación, se computan dichos indicadores para la cartera de valores que se genera con el algoritmo y para el Russell 3000. Tal y como se puede observar en la Tabla 6.7, el método propuesto bate al mercado en todos los indicadores analizados. No obstante, el ratio de volatilidad entre ambos es de 1.5x, aunque dicho factor es menor que el presente en la comparación de ambas rentabilidades (4x), y por eso el ratio de Sortino es casi 3 veces mayor en el caso del método propuesto. Este es el indicador más fidedigno a emplear en el caso de estudio, ya que resulta crucial establecer una distinción entre la volatilidad *"buena"* (la presente en los últimos periodos) y la *"mala"*.

	Model	RUA
Annual return	0,414	0,158
Cumulative returns	4,645	1,048
Annual volatility	0,327	0,194
Sharpe ratio	1,201	0,855
Calmar ratio	1,108	0,451
Stability	0,401	0,840
Max drawdown	−0,374	−0,351
Omega ratio	1,913	1,200
Sortino ratio	3,805	1,174
Skew	13,270	NaN
Kurtosis	257,457	NaN
Tail ratio	1,361	0,900
Daily value at risk	−0,040	−0,024

Tabla 6.7. *Comparación de las estadísticas del portfolio*

A continuación, se procede con la representación de algunos de dichos indicadores de manera gráfica. Para ello se adjuntan las Figuras 6.9, 6.10 y 6.11 que permiten visualizar, respectivamente, el ratio de Sharpe, la beta y la volatilidad móvil del modelo frente al Russell 3000.

En primer lugar, con respecto al ratio de Sharpe se puede observar en la Figura 6.9 cómo en el periodo comprendido entre el año 2018 y 2020, la eficacia del benchmark es mejor que la del método. Esto se debe al escaso número de operaciones que el algoritmo ejecuta a lo largo de dicho periodo.

En segundo lugar, la beta de un portfolio es una medida del riesgo sistemático global de una cartera de inversiones. Se computa calculando la media ponderada del coeficiente beta de todos los valores individuales de una cartera. El mercado en general tiene un coeficiente $\beta = 1$. En este sentido, este portoflio, que tiene una $\beta < 1$, presenta un riesgo sistemático menor que el mercado; a pesar de que su volatilidad anualizada, como se puntualizó en la Tabla 6.7, sea mayor que la del índice.

En relación con la anterior, se puede observar, en la Figura 6.11, la volatilidad móvil de la cartera frente al índice. Se pone nuevamente de manifiesto que la volatilidad es significativamente menor en todos los periodos simulados salvo

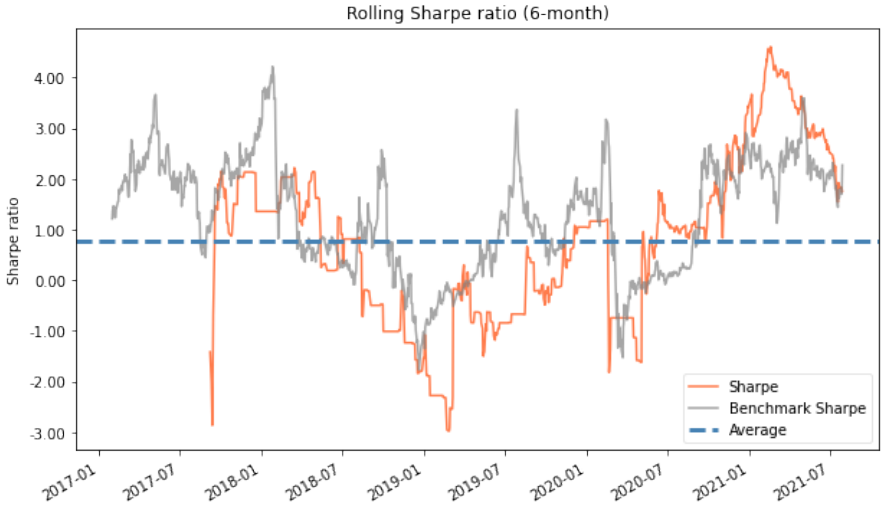


Figura 6.9. *Comparación del ratio de Sharpe móvil del modelo frente a B&H del RUA*

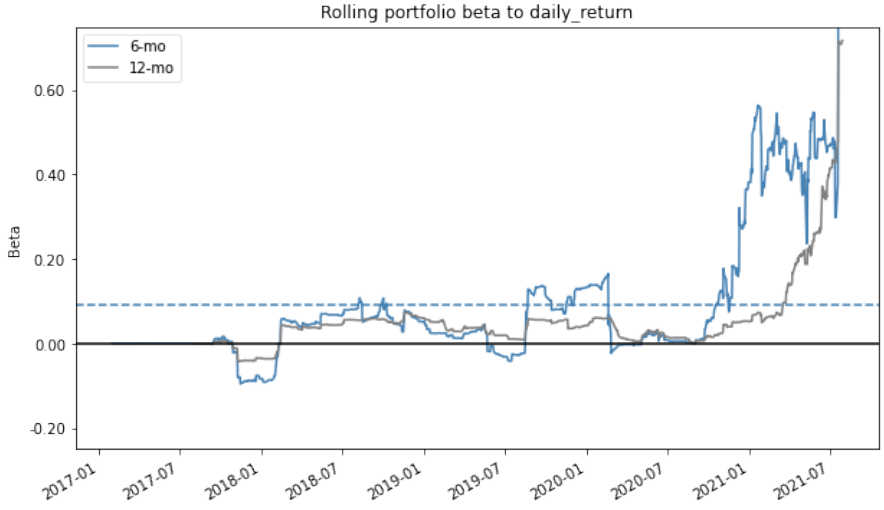


Figura 6.10. *Comparación de la beta del modelo frente a B&H del RUA*

en el último año. Esto se debe fundamentalmente al pico de rentabilidad que se produce a comienzos del año 2021, revelándose cómo un exceso de rentabilidad va acompañado, generalmente, de una exposición al riesgo mayor.

Por último, como la principal limitación que tiene el modelo es la distribución de la rentabilidad a lo largo del tiempo, se adjuntan en las Figuras 6.12, 6.13 y

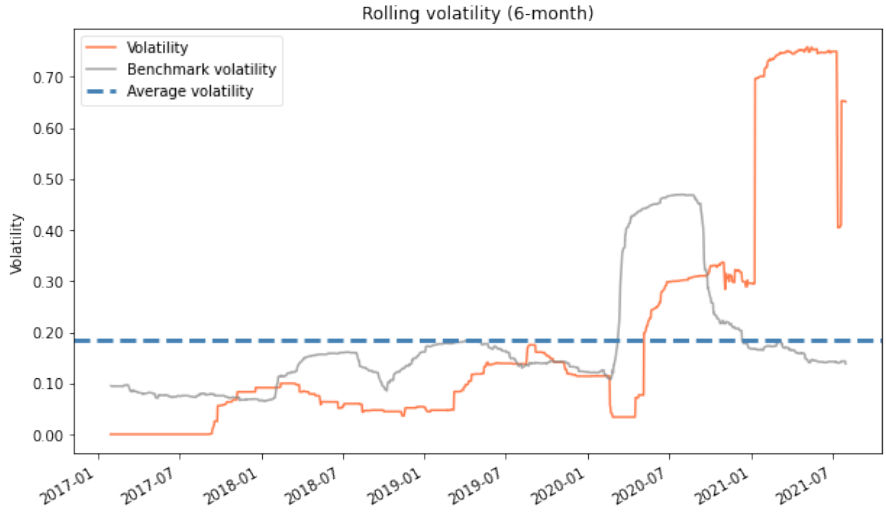


Figura 6.11. *Comparación de la volatilidad del modelo frente a B&H del RUA*

6.14, unas gráficas que permiten realizar un estudio pormenorizado de las flaquezas de las que adolece la estrategia.

En primer lugar, se incluye en la Figura 6.12 una representación del histograma de los rendimientos anuales, junto con su media y su mediana. A pesar de la poca uniformidad presente, en el peor caso se pierde alrededor de un 7 %, que se ve claramente compensado con el beneficio generado en los últimos años. La diferencia entre la media y la mediana es tan significativa debido a la presencia de valores dispersos. Como se ha comentado anteriormente, esto es un reflejo directo del uso que se le ha dado al K-NN en este particular escenario. Al emplearse, en cierto modo, como detector de patrones anómalos genera rendimientos muy positivos en ciertos instantes temporales.

Finalmente, en las Figuras 6.13 y 6.14 se analiza la distribución de los rendimientos mensuales con un mapa de calor y un histograma, respectivamente. De nuevo, se puede apreciar un efecto similar al ya discutido. En ningún mes se pierde una cantidad considerable de capital en comparación con los retornos positivos que luego el modelo genera.

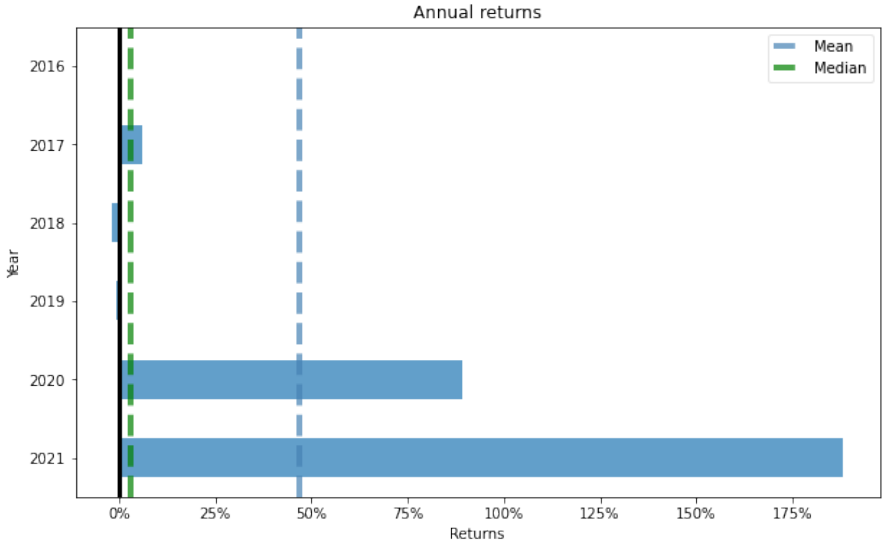


Figura 6.12. *Distribución de los rendimientos anuales*

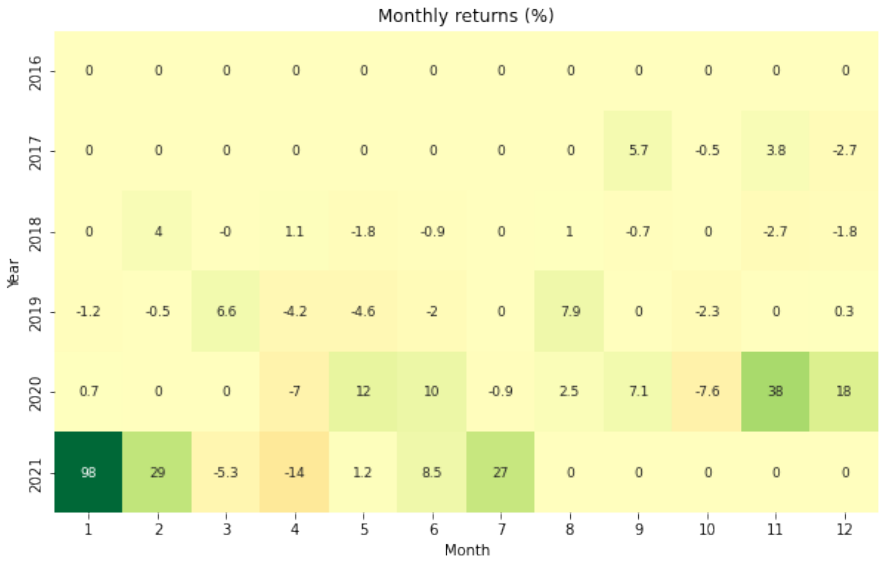


Figura 6.13. *Mapa de calor de los retornos mensuales*

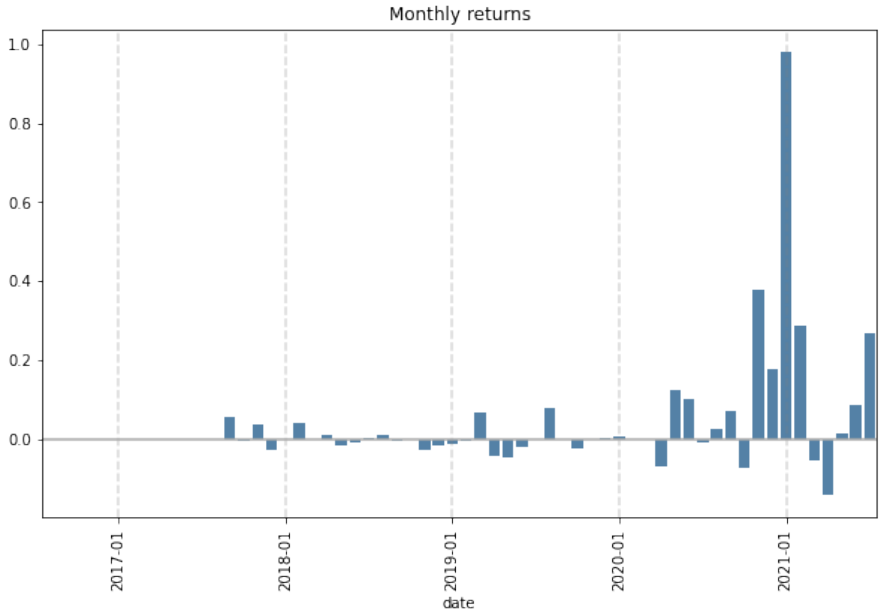


Figura 6.14. *Distribución de los rendimientos mensuales*

Capítulo 7

Conclusiones y futuros desarrollos

En el presente trabajo se ha desarrollado una herramienta que sirve como soporte funcional para desarrollar de manera holística el proceso de inversión al que se enfrentan multitud de agentes diariamente en el mercado de valores, y con la que se bate recursivamente al mercado en el periodo analizado. No se centra exclusivamente en la predicción del precio de las acciones, sino que desarrolla un marco completo para la ejecución de distintas estrategias de inversión que toman por referencia, mayoritariamente, las premisas del análisis técnico.

A lo largo del proyecto, se ha ido analizando el impacto de diferentes variables sobre la rentabilidad final del modelo. Para comparar los diferentes escenarios se han construidos dos backtesting que actúan en niveles jerárquicamente diferentes y que permiten reducir al mínimo la fricción entre los resultados experimentales de la mayoría de las investigaciones y los que se obtienen en la práctica financiera.

Las variables analizadas se podrían clusterizar en dos grupos fundamentales: el relativo a mejorar la predicción de las velas japonesas y el que se ocupa de optimizar la eficiencia de la estrategia de inversión. Con respecto a la primera clasificación, se puede afirmar la importancia de la correcta ponderación del peso de las velas y su normalización, así como, la sensibilidad de la rentabilidad final frente al número de vecinos empleados en la predicción. Las simulaciones ejecutadas revelan que sería recomendable ponderar las velas japonesas de forma diferente según la longitud de las ventanas adoptadas. Asimismo, en relación con el impacto del valor de K en la rentabilidad se observa, cómo el número óptimo de vecinos a emplear es de 500.

Por otra parte, teniendo en cuenta el segundo grupo, se pone de manifiesto la

necesidad de realizar un previo análisis de las acciones según el sector al que pertenecen. Este ajuste no sólo mejora la rentabilidad final del modelo, sino también su distribución a lo largo del tiempo. Al mismo tiempo, las diferentes simulaciones diseñadas revelan la influencia del horizonte de predicción fijo sobre la estrategia de inversión.

Son precisamente estos análisis los que plantean interesantes interrogativas que podrían resolverse en futuros desarrollos. Se deja como futura línea de investigación el empleo de un Autoencoder que permita extraer la combinación óptima de las ventanas a utilizar para predecir los movimientos de la rentabilidad. Además, también se plantea la posibilidad de diseñar una estrategia de inversión dinámica que ajuste la duración de las operaciones según las condiciones de rendimiento y volatilidad del mercado. Por último, el estudio sectorial revela que es necesario hacer un claro pretratamiento de las acciones según diferentes clúster, que pueden ser distintos a los sectoriales desarrollados en este trabajo. Por ello, merece la pena explorar otras variables como la tendencia y la volatilidad de las acciones en las que se invierte. Con todos estos cambios seguramente se adquiera un modelo más robusto y sólido en el tiempo y con rendimientos más distribuidos a lo largo de los distintos horizontes temporales, para así, poder hacer frente a una de las mayores limitaciones del modelo en cuestión.

Apéndice A

Invirtiendo con un fin



En 2005, la Asamblea General de las Naciones Unidas adoptó la Agenda 2030 para el Desarrollo Sostenible, un plan de acción a favor de las personas, el planeta y la prosperidad, que también tiene la intención de fortalecer la paz universal y el acceso a la justicia. La Agenda plantea *17 Objetivos de Desarrollo Sostenible (ODS)* con 169 metas de carácter integrado e indivisible que abarcan las esferas económica, social y ambiental.

Desde ese hito, los objetivos han estimulado todos los organismos de la sociedad a trabajar conjuntamente para su conquista. En este camino, resulta fundamental disponer de nuevas formas de financiación para propulsar cambios revolucionarios que se alineen con el plan marcado. Para ello, desde el comienzo las Bolsas de Valores Sostenibles de las Naciones Unidas (UN SSE, por sus siglas en inglés) han destinado diferentes recursos con el fin de promover la generación de riqueza sostenible a la hora de establecer la asignación del capital de los inversores.

Es precisamente el mercado financiero el engranaje fundamental que opera como vínculo de cohesión entre los distintos agentes que actúan en el mercado bursátil, que son esencialmente las empresas públicas y los operadores que invierten en ellas. Es esencialmente aquí donde surge uno de los obstáculos fundamentales para el desarrollo de estas prácticas sostenibles: la falta de transparencia y visibilidad por parte de los principales organismos de los mercados financieros. La ausencia de índices globales que incluyan las empresas que mejor se alinean con los ODS es lo que no permite que los agentes las conozcan y, por lo tanto, inviertan en ellas para financiar su labor de promulgar el desarrollo sostenible.

Por ello, en este trabajo se ha hecho un esfuerzo por unificar diferentes criterios que catalogan a empresas sostenibles para así disponer de una referencia global que permita establecer un rating según el *perfomance* de la empresa de acuerdo a los 17 objetivos establecidos.

La inversión social y responsable ha pasado de ser considerada una simple tendencia a una realidad a tener en cuenta. Durante los últimos años las bolsas globales han llegado a sufrir caídas superiores al 40 %. Sin embargo, aquellas empresas con perfiles de sostenibilidad robustos mostraron un comportamiento muy diferencial. Con esta la crisis derivada del COVID-19 se ha puesto realmente en evidencia la importancia de las prácticas sociales sostenibles. Ante noticias tan devastadoras y situaciones sin precedentes, la sociedad ha mostrado un cambio en sus preferencias hacia los activos sostenibles, pues han sido precisamente las

empresas más sostenibles, las que han sabido adaptarse con mayor facilidad a este entorno tan caótico. Con esta crisis sanitaria se ha puesto claramente de manifiesto que las empresas se estancan cuando la sociedad lo hace, por ello, la gestión del capital humano tiene que seguir siendo el pilar fundamental de todas ellas.

Apéndice B

Informe generado para el análisis estadístico

```
================================================================================
ATOS: 1 / 2509
================================================================================
COMPANY PROFILE: Atossa Therapeutics, Inc., a clinical-stage
biopharmaceutical company, discovers and develops medicines in
the areas of oncology and infectious diseases. The company's lead
program is Endoxifen, an active metabolite of tamoxifen, which is
in Phase II clinical trials to treat and prevent breast cancer.
It is also developing AT-301, a proprietary drug candidate for
nasal administration in patients diagnosed with COVID-19; AT-H201,
a drug candidate to improve lung function in COVID-19 patients;
and immunotherapy/chimeric antigen receptor therapy programs for
the treatment of breast cancer. Atossa Therapeutics, Inc. has a
research agreement with Dana-Farber Cancer Institute, Inc. to
support research of cytokine-coated nanoparticles for the potential
treatment of breast cancer. The company was formerly known as
Atossa Genetics Inc. and changed its name to Atossa Therapeutics, Inc.
in January 2020. Atossa Therapeutics, Inc. was founded in 2008
and is headquartered in Seattle, Washington.
--------------------------------------------------------------------------------
STOCK EXCHANGE: NMS
SECTOR: Healthcare
INDUSTRY: Biotechnology
--------------------------------------------------------------------------------
CURRENCY: USD
LAST CURRENCY EXCHANGE RATE (EUROS):  0.841
CURRENCY CHANGE(%): [  0.00  -0.28  -0.14  -0.21  -0.38 ]
--------------------------------------------------------------------------------
 5 days
 SET/TOTAL  SUCCESS MEAN   MEDIAN MAX     MIN    STD
 100/627250  74.00  11.19   9.69  59.07 -35.77  16.77
 150/627250  69.33   9.40   7.77  59.07 -35.77  16.44
 250/627250  66.00   8.50   6.64  59.07 -39.62  16.63
 500/627250  62.40   7.43   5.21  72.41 -39.62  16.78
 750/627250  60.67   6.95   4.37 389.08 -39.62  21.23
1000/627250  59.00   5.99   3.64 389.08 -41.67  20.07
1500/627250  58.73   5.16   3.10 389.08 -41.67  18.46
2000/627250  58.00   4.74   2.71 389.08 -61.33  17.60
--------------------------------------------------------------------------------
 4 days
 SET/TOTAL  SUCCESS MEAN   MEDIAN MAX     MIN    STD
 100/627250  68.00   8.16   6.06  62.83 -40.63  15.60
 150/627250  65.33   6.62   4.52  62.83 -40.63  15.07
 250/627250  64.40   6.86   4.52  62.83 -40.63  15.09
 500/627250  62.00   5.73   3.60  62.83 -40.63  14.66
 750/627250  60.80   5.63   3.12 407.86 -40.63  20.61
1000/627250  58.40   4.82   2.62 407.86 -40.63  19.23
1500/627250  57.87   4.27   2.34 407.86 -40.63  17.39
2000/627250  57.10   3.85   2.14 407.86 -40.63  16.53
--------------------------------------------------------------------------------
 3 days
 SET/TOTAL  SUCCESS MEAN   MEDIAN MAX     MIN    STD
 100/627250  64.00   6.26   3.93  46.99 -28.18  13.57
 150/627250  61.33   4.72   3.34  53.01 -28.18  13.48
 250/627250  62.40   4.61   3.34  53.01 -31.97  13.46
 500/627250  59.40   4.20   2.50  58.41 -31.97  13.08
 750/627250  58.67   4.57   2.45 575.98 -31.97  24.41
1000/627250  56.60   3.77   1.95 575.98 -31.97  21.98
1500/627250  55.40   3.09   1.58 575.98 -31.97  19.03
2000/627250  54.35   2.79   1.22 575.98 -34.72  17.38
--------------------------------------------------------------------------------
 2 days
 SET/TOTAL  SUCCESS MEAN   MEDIAN MAX     MIN    STD
```

```
 100/627250  61.00   3.36   2.55  40.65 -20.89  11.37
 150/627250  54.00   2.43   1.08  46.99 -25.79  11.69
 250/627250  54.00   2.28   1.08  46.99 -25.79  10.78
 500/627250  54.60   2.46   1.34  55.88 -32.55  11.16
 750/627250  53.47   2.72   1.16 413.97 -32.55  18.45
1000/627250  53.30   2.36   1.16 413.97 -32.55  16.70
1500/627250  52.40   1.86   0.78 413.97 -32.55  14.53
2000/627250  51.90   1.59   0.64 413.97 -33.72  13.31
--------------------------------------------------------------------------------
 1 days
 SET/TOTAL  SUCCESS MEAN   MEDIAN MAX     MIN    STD
 100/627250  47.00   0.99  -0.20  24.49 -12.94   7.21
 150/627250  50.00   0.93   0.05  25.68 -25.00   7.66
 250/627250  48.40   0.86   0.00  31.46 -25.00   7.73
 500/627250  51.40   1.39   0.31 114.11 -36.59   9.33
 750/627250  52.40   1.46   0.49 152.84 -36.59  10.17
1000/627250  50.60   1.26   0.23 152.84 -36.59   9.48
1500/627250  50.93   1.06   0.26 152.84 -36.59   8.47
2000/627250  50.25   0.91   0.11 152.84 -36.59   7.86
--------------------------------------------------------------------------------
LAST CLOSE:  2.990
--------------------------------------------------------------------------------
CLOSE CHANGE DAILY SEQUENCE(%):    [  -5.26   -1.01    0.77   -1.27  -23.33 ]
CLOSE CHANGE WEEKLY SEQUENCE(%):   [ -19.19  -16.87  -15.41   -7.11  -28.47 ]
CLOSE CHANGE MONTHLY SEQUENCE(%):  [ -12.90   12.04   83.47   44.14  -53.28 ]
CLOSE CHANGE YEARLY SEQUENCE(%):   [ -90.40  -44.91   -9.66   86.51  -25.44 ]
--------------------------------------------------------------------------------
CLOSE PROFIT DAILY SEQUENCE(%):    [  -5.26   -1.29    0.63   -1.47  -23.62 ]
CLOSE PROFIT WEEKLY SEQUENCE(%):   [ -18.81  -16.82  -15.17   -6.80  -29.18 ]
CLOSE PROFIT MONTHLY SEQUENCE(%):  [ -13.08   10.72   80.54   48.47  -53.43 ]
CLOSE PROFIT YEARLY SEQUENCE(%):   [ -90.86  -43.97   -7.13   77.94  -26.11 ]
--------------------------------------------------------------------------------
OPEN PROFIT DAILY SEQUENCE(%):     [   0.99   -3.46   -1.65    2.62   -7.85 ]
OPEN PROFIT WEEKLY SEQUENCE(%):    [ -15.12   -5.79   -1.62  -28.63   -9.33 ]
OPEN PROFIT MONTHLY SEQUENCE(%):   [  -6.58    8.95   71.55   60.78  -40.02 ]
OPEN PROFIT YEARLY SEQUENCE(%):    [ -90.61  -44.34  -10.74   70.92   -3.00 ]
--------------------------------------------------------------------------------
MEANS (20,60,240) [MEDIUM/LONG SHORT/LONG CLOSE/LONG]: [  1.69   1.77   1.11 ]
--------------------------------------------------------------------------------
RSI   9 (%):   [ 14.67  12.93  17.60  29.83  24.64 ]
RSI  14 (%):   [ 32.01  32.48  32.66  34.10  24.66 ]
RSI  20 (%):   [ 22.82  23.71  27.43  31.00  25.98 ]
RSI  60 (%):   [ 54.23  54.07  53.74  54.05  51.42 ]
RSI 240 (%):   [ 50.20  50.35  50.65  50.21  49.16 ]
--------------------------------------------------------------------------------
MACD           (  9, 12, 26):   [ -0.28  -0.31  -0.33  -0.35  -0.41 ]
MACD SIGNAL    (  9, 12, 26):   [ -0.06  -0.11  -0.15  -0.19  -0.24 ]
MACD HISTOGRAM (  9, 12, 26):   [ -0.23  -0.21  -0.18  -0.16  -0.18 ]
--------------------------------------------------------------------------------
MACD           ( 20, 60,240):   [  1.19   1.17   1.15   1.12   1.08 ]
MACD SIGNAL    ( 20, 60,240):   [  1.15   1.15   1.15   1.15   1.14 ]
MACD HISTOGRAM ( 20, 60,240):   [  0.04   0.01  -0.01  -0.03  -0.06 ]
--------------------------------------------------------------------------------
CANDLE STICKS   1 days - HIGH(%):   [   7.18    0.98    4.70    5.60   -4.98 ]
CANDLE STICKS   1 days - OPEN(%):   [  -2.39   -0.53   -0.90    1.06   -5.49 ]
CANDLE STICKS   1 days - CLOSE(%):  [  -5.26   -1.29    0.63   -1.47  -23.62 ]
CANDLE STICKS   1 days - LOW(%):    [  -8.13   -6.58   -2.17   -1.97  -28.48 ]
CANDLE STICKS   1 days - VOLUME(%): [ -43.43  -62.51  -65.06  -64.87   86.70 ]
--------------------------------------------------------------------------------
CANDLE STICKS   5 days - HIGH(%):   [  23.74    3.20   35.23    1.90    7.18 ]
CANDLE STICKS   5 days - OPEN(%):   [   1.14    2.26   -0.60   -8.43   -2.39 ]
CANDLE STICKS   5 days - CLOSE(%):  [ -18.81  -16.82  -15.17   -6.80  -29.18 ]
CANDLE STICKS   5 days - LOW(%):    [ -23.54  -20.20  -18.56  -17.54  -33.68 ]
```

```
CANDLE STICKS   5 days - VOLUME(%): [ 110.63   -7.18   32.35  -26.49  -30.35 ]
--------------------------------------------------------------------------------
CANDLE STICKS  20 days - HIGH(%):   [  44.24   20.96   89.89  126.28   12.49 ]
CANDLE STICKS  20 days - OPEN(%):   [  20.11   -0.90   -1.59    3.16    2.26 ]
CANDLE STICKS  20 days - CLOSE(%):  [ -13.08   10.72   80.54   48.47  -53.43 ]
CANDLE STICKS  20 days - LOW(%):    [ -22.44  -32.01  -10.28  -13.50  -56.38 ]
CANDLE STICKS  20 days - VOLUME(%): [ -53.30  -17.87   20.54   75.05  -10.92 ]
--------------------------------------------------------------------------------
FINANTIAL INFO: STOCK SHARE STRUCTURE
--------------------------------------------------------------------------------
PERCENTAGE HELD BY INSIDERS | INSTITUTIONS | FLOAT:  0.04 | 12.09 | 87.87
SHORT INTEREST RATIO (DAYS TO BE COVERED WITH DAYLY MEAN VOLUME): 2.2130
--------------------------------------------------------------------------------
FINANTIAL INFO: INCOME STATEMENT
--------------------------------------------------------------------------------
REVENUE PER SHARE(TTM) / SHARE PRICE (%):     NaN
EPS / PRICE (%) [TRAILING FORWARD GROWTH]: [ -23.81   -5.02   78.93]
EARNINGS HISTORY:
  QUARTER  ACTUAL_EP(%)  ESTIMATED_EP(%) DIFFERENCE(%) SURPRISE(%)
31-Mar-2021 -1.338         -1.672           0.334        20.00
31-Dec-2020 -29.431         -9.030          -20.401      -225.90
30-Sep-2020 -11.371         -11.371           0.000         0.00
30-Jun-2020 -14.381         -9.365          -5.017       -53.60
EARNINGS PREDICTION FOR CURRENT QUARTER:
GROWTH(%) NUM_ANALYSTS AVG_EP(%) LOW_EP(%) HIGH_EP(%) YEAR_AGO_EP(%)
  93.00          2     -1.003    -1.338    -1.003      -14.381
EARNINGS PREDICTION FOR NEXT QUARTER:
GROWTH(%) NUM_ANALYSTS AVG_EP(%) LOW_EP(%) HIGH_EP(%) YEAR_AGO_EP(%)
  91.20          2     -1.003    -1.338    -1.003      -11.371
EARNINGS PREDICTION FOR CURRENT YEAR:
GROWTH(%) NUM_ANALYSTS AVG_EP(%) LOW_EP(%) HIGH_EP(%) YEAR_AGO_EP(%)
  92.90          2     -4.682    -5.017    -4.682      -65.886
EARNINGS PREDICTION FOR NEXT YEAR:
GROWTH(%) NUM_ANALYSTS AVG_EP(%) LOW_EP(%) HIGH_EP(%) YEAR_AGO_EP(%)
  -7.10          2     -5.017    -5.017    -5.017      -4.682
MARGINS(%): [GROSS EBITDA OPERATING PROFIT]: [   0.00    0.00    0.00    0.00]
--------------------------------------------------------------------------------
FINANTIAL INFO: BALANCE SHEET AND CASH FLOW
--------------------------------------------------------------------------------
[BOOK VALUE/PRICE CASH/PRICE DEBT/PRICE]: (%) [  38.49   38.09    0.01]
[RETURN/ASSETS RETURN/EQUITY]: (%) [ -12.43  -24.46]
[CURRENT RATIO QUICK RATIO]: (%) [ 102.26  100.28]
--------------------------------------------------------------------------------
FINANTIAL INFO: TARGET PRICE AND RECOMMENDATIONS
--------------------------------------------------------------------------------
RECOMMENDATION: buy - 2 (Num. analysts: 2)
RECOMMENDATION TRENDS:
  MONTH STRONG_BUY BUY HOLD UNDERPERFORM SELL
   May      0       1    0       0         0
   Apr      0       2    0       0         0
   Mar      0       2    0       0         0
   Feb      0       1    0       0         0
TARGET PRICES [LOW MEAN MEDIAN HIGH]: [  7.00   7.50   7.50   8.00]
TARGET PROFITS(%) [LOW MEAN MEDIAN HIGH]: [134.11 150.84 150.84 167.56]
--------------------------------------------------------------------------------
FINANTIAL INFO: VALUE INVESTING
--------------------------------------------------------------------------------
PE<15 [TRAILING FORWARD]: [  -4.20  -19.93]
PBV<1.5 :  2.60
PExPBV<22.5 [TRAILING FORWARD]:  [ -10.91  -51.78]
PRICE/EARNINGS TO GROWTH (PEG) [TRAILING FORWARD]: [    NaN     NaN]
```

Apéndice C

Informe generado para la estrategia de inversión

==========
INVESTMENT STRATEGY ANALYSIS
==========
INVESTMENT STRATEGY PARAMETERS:

INITIAL DATE: 18-Jul-2016
FINAL DATE: 16-Jul-2021
PREDICTION HORIZON: 5

INVESTMENT STRATEGY RESULTS FOR OPTIMIZATION

NUMBER OF YEARS: 5.0
TOTAL PROFIT (%): +1.3
PROFIT/YEAR (%): +0.3
NUMBER OF OPERATIONS (BUY OR SELL): 576
NUMBER OF BUY OPERATIONS: 288
NUMBER OF FINAL OPEN POSITIONS: 0
NUMBER OF MAXIMUM OPEN POSITIONS: 119
NUMBER OF OPERATIONS/YEAR: 115.2
% OF POSITIVE OPERATIONS: 63.2
% OF NEGATIVE OPERATIONS: 36.8
MAXIMUM PROFIT (%): +28.7
MINIMUM PROFIT (%): -3.7
MAXIMUM PROFIT STREAK (%): +32.4
MINIMUM PROFIT STREAK (%): -6.0
AVERAGE PROFIT/OPERATION (%): +0.0
MAXIMUM RAW PROFIT/OPERATION (%): +98.8
MINIMUM RAW PROFIT/OPERATION (%): -44.9
AVERAGE OPERATION DURATION (days): 6.7
MAXIMUM OPERATION DURATION (days): 15
MINIMUM OPERATION DURATION (days): 6
COST OF BUY-SELL OPERATION (%): 1.0

OPERATIONS:

Symbol	Buy date	Sell date	Days	Buy price	Sell Price	Raw profit	Net profit
ACR	26-Mar-2020	15-Apr-2020	14	6.33	8.70	+36.3%	+35.3%
AHT	26-Mar-2020	03-Apr-2020	7	7.80	6.60	-15.0%	-16.0%
ARAY	26-Mar-2020	02-Apr-2020	6	1.61	1.72	+6.2%	+5.2%
ARI	26-Mar-2020	13-Apr-2020	12	7.43	9.15	+22.7%	+21.7%
ATI	26-Mar-2020	02-Apr-2020	6	7.53	7.75	+2.3%	+1.3%
ATRO	26-Mar-2020	09-Apr-2020	11	8.42	9.55	+13.7%	+12.7%
COOP	26-Mar-2020	08-Apr-2020	10	7.11	6.14	-13.7%	-14.7%
DOOO	26-Mar-2020	02-Apr-2020	6	16.46	15.28	-7.7%	-8.7%
EGY	26-Mar-2020	02-Apr-2020	6	0.87	0.93	+6.5%	+5.5%
ENLC	26-Mar-2020	02-Apr-2020	6	1.00	1.09	+8.4%	+7.4%
FCRD	26-Mar-2020	02-Apr-2020	6	1.72	2.26	+30.6%	+29.6%
FET	26-Mar-2020	07-Apr-2020	9	3.80	3.60	-4.5%	-5.5%
IO	26-Mar-2020	07-Apr-2020	9	1.20	1.77	+48.7%	+47.7%
IPI	26-Mar-2020	02-Apr-2020	6	8.10	7.50	-7.9%	-8.9%
IVR	26-Mar-2020	02-Apr-2020	6	3.72	2.83	-24.4%	-25.4%
KIN	26-Mar-2020	02-Apr-2020	6	3.75	3.54	-6.1%	-7.1%
NMIH	26-Mar-2020	02-Apr-2020	6	11.83	10.96	-7.9%	-8.9%
NYMT	26-Mar-2020	02-Apr-2020	6	1.50	1.30	-13.8%	-14.8%
PBF	26-Mar-2020	09-Apr-2020	11	7.29	8.53	+17.3%	+16.3%
PII	26-Mar-2020	02-Apr-2020	6	49.26	45.00	-9.2%	-10.2%
RC	26-Mar-2020	02-Apr-2020	6	5.39	5.54	+2.2%	+1.2%
SAH	26-Mar-2020	02-Apr-2020	6	13.74	12.32	-10.9%	-11.9%
SBS	26-Mar-2020	02-Apr-2020	6	7.06	6.79	-4.4%	-5.4%
SVC	26-Mar-2020	13-Apr-2020	12	6.00	6.60	+9.6%	+8.6%
TTI	26-Mar-2020	13-Apr-2020	12	0.36	0.30	-17.0%	-18.0%
TWO	26-Mar-2020	14-Apr-2020	13	5.10	4.76	-6.9%	-7.9%
STAF	27-Mar-2020	09-Apr-2020	10	2.46	2.34	-3.3%	-4.3%
AA	30-Mar-2020	08-Apr-2020	8	6.55	7.17	+11.9%	+10.9%
ALDX	30-Mar-2020	06-Apr-2020	6	2.50	2.06	-15.1%	-16.1%
ANIX	30-Mar-2020	06-Apr-2020	6	1.50	1.57	+7.8%	+6.8%
BCS	30-Mar-2020	06-Apr-2020	6	4.60	4.21	-5.7%	-6.7%
CBRL	30-Mar-2020	06-Apr-2020	6	73.17	74.00	+4.2%	+3.2%
CVEO	30-Mar-2020	07-Apr-2020	7	5.03	5.00	+2.6%	+1.6%
DX	30-Mar-2020	06-Apr-2020	6	11.85	10.01	-13.0%	-14.0%
ERF	30-Mar-2020	06-Apr-2020	6	1.30	1.83	+45.0%	+44.0%

FLS	30-Mar-2020	06-Apr-2020	6	21.60	24.71	+17.9%	+16.9%
HEES	30-Mar-2020	06-Apr-2020	6	13.59	14.13	+7.1%	+6.1%
HMHC	30-Mar-2020	20-Apr-2020	15	2.10	1.15	-43.9%	-44.9%
IBN	30-Mar-2020	06-Apr-2020	6	8.20	7.76	-2.5%	-3.5%
KEY	30-Mar-2020	06-Apr-2020	6	11.04	9.80	-8.5%	-9.5%
M	30-Mar-2020	09-Apr-2020	9	5.43	6.52	+23.1%	+22.1%
MRO	30-Mar-2020	09-Apr-2020	9	3.49	4.35	+27.8%	+26.8%
OCN	30-Mar-2020	06-Apr-2020	6	8.25	5.25	-34.4%	-35.4%
PDS	30-Mar-2020	08-Apr-2020	8	0.30	0.34	+15.8%	+14.8%
REPH	30-Mar-2020	07-Apr-2020	7	6.71	8.25	+26.7%	+25.7%
TTM	30-Mar-2020	06-Apr-2020	6	4.57	4.31	-2.8%	-3.8%
VEDL	30-Mar-2020	06-Apr-2020	6	3.40	3.45	+4.5%	+3.5%
APA	31-Mar-2020	08-Apr-2020	7	4.41	6.83	+56.8%	+55.8%
AXAS	31-Mar-2020	07-Apr-2020	6	2.60	2.80	+10.0%	+9.0%
HLX	31-Mar-2020	13-Apr-2020	9	1.86	1.82	-1.3%	-2.3%
HXL	31-Mar-2020	07-Apr-2020	6	36.32	33.74	-5.2%	-6.2%
NR	31-Mar-2020	14-Apr-2020	10	0.86	1.15	+35.1%	+34.1%
OII	31-Mar-2020	13-Apr-2020	9	3.51	4.26	+22.4%	+21.4%
PBA	31-Mar-2020	09-Apr-2020	8	17.40	20.59	+20.2%	+19.2%
PDCE	31-Mar-2020	08-Apr-2020	7	6.06	8.83	+47.5%	+46.5%
SEAS	31-Mar-2020	13-Apr-2020	9	11.48	12.13	+6.6%	+5.6%
SIG	31-Mar-2020	09-Apr-2020	8	7.03	8.07	+16.6%	+15.6%
SM	31-Mar-2020	09-Apr-2020	8	1.19	1.69	+44.2%	+43.2%
SNV	31-Mar-2020	07-Apr-2020	6	17.76	17.69	+1.7%	+0.7%
SPR	31-Mar-2020	14-Apr-2020	10	24.64	22.22	-8.9%	-9.9%
WTI	31-Mar-2020	07-Apr-2020	6	1.71	1.92	+14.6%	+13.6%
WWW	31-Mar-2020	07-Apr-2020	6	15.52	15.35	+1.0%	-0.0%
ABR	01-Apr-2020	13-Apr-2020	8	4.63	7.96	+73.4%	+72.4%
AGNC	01-Apr-2020	08-Apr-2020	6	10.09	11.14	+11.8%	+10.8%
BZH	01-Apr-2020	08-Apr-2020	6	6.10	4.75	-21.2%	-22.2%
CIM	01-Apr-2020	13-Apr-2020	8	8.50	7.90	-6.2%	-7.2%
CLDT	01-Apr-2020	08-Apr-2020	6	5.56	6.54	+19.1%	+18.1%
CLR	01-Apr-2020	09-Apr-2020	7	7.28	13.23	+84.5%	+83.5%
CVE	01-Apr-2020	08-Apr-2020	6	1.93	2.81	+47.4%	+46.4%
DVN	01-Apr-2020	08-Apr-2020	6	6.57	9.03	+39.1%	+38.1%
EC	01-Apr-2020	08-Apr-2020	6	9.25	11.51	+26.0%	+25.0%
HRB	01-Apr-2020	08-Apr-2020	6	13.33	13.84	+5.1%	+4.1%
HSKA	01-Apr-2020	13-Apr-2020	8	53.35	63.38	+19.9%	+18.9%
MAC	01-Apr-2020	13-Apr-2020	8	5.28	9.66	+84.6%	+83.6%
MEOH	01-Apr-2020	08-Apr-2020	6	11.71	13.38	+15.7%	+14.7%
NRZ	01-Apr-2020	15-Apr-2020	10	4.64	4.89	+5.8%	+4.8%
PAR	01-Apr-2020	08-Apr-2020	6	12.50	13.25	+7.3%	+6.3%
RCL	01-Apr-2020	08-Apr-2020	6	30.30	34.88	+16.5%	+15.5%
REX	01-Apr-2020	08-Apr-2020	6	44.81	47.83	+8.1%	+7.1%
ROIC	01-Apr-2020	09-Apr-2020	7	7.82	9.35	+21.4%	+20.4%
STRL	01-Apr-2020	08-Apr-2020	6	9.44	8.22	-11.9%	-12.9%
TGLS	01-Apr-2020	08-Apr-2020	6	3.78	3.14	-15.9%	-16.9%
USCR	01-Apr-2020	09-Apr-2020	7	17.24	19.22	+13.2%	+12.2%
WES	01-Apr-2020	08-Apr-2020	6	3.10	4.95	+61.6%	+60.6%
WMC	01-Apr-2020	13-Apr-2020	8	2.18	3.17	+46.7%	+45.7%
AAIC	02-Apr-2020	09-Apr-2020	6	1.95	2.75	+42.2%	+41.2%
AINV	02-Apr-2020	09-Apr-2020	6	5.98	7.73	+30.3%	+29.3%
CCL	02-Apr-2020	13-Apr-2020	7	8.01	11.77	+47.2%	+46.2%
CLNC	02-Apr-2020	13-Apr-2020	7	3.17	4.83	+52.6%	+51.6%
CNK	02-Apr-2020	13-Apr-2020	7	8.16	12.56	+54.2%	+53.2%
CPE	02-Apr-2020	09-Apr-2020	6	4.70	5.20	+11.6%	+10.6%
CPRI	02-Apr-2020	09-Apr-2020	6	8.75	14.94	+72.2%	+71.2%
CZR	02-Apr-2020	13-Apr-2020	7	11.54	18.25	+58.4%	+57.4%
EAT	02-Apr-2020	09-Apr-2020	6	11.26	15.54	+39.2%	+38.2%
EVOP	02-Apr-2020	09-Apr-2020	6	13.57	14.85	+10.3%	+9.3%
FRTA	02-Apr-2020	09-Apr-2020	6	5.71	6.36	+12.3%	+11.3%
GPRE	02-Apr-2020	13-Apr-2020	7	4.35	4.63	+6.6%	+5.6%
GPS	02-Apr-2020	13-Apr-2020	7	5.82	9.16	+57.7%	+56.7%
KSS	02-Apr-2020	13-Apr-2020	7	12.90	20.42	+58.6%	+57.6%
LGIH	02-Apr-2020	13-Apr-2020	7	38.07	54.20	+42.6%	+41.6%
LTRPA	02-Apr-2020	09-Apr-2020	6	1.58	2.16	+37.8%	+36.8%
NBRV	02-Apr-2020	16-Apr-2020	10	5.89	5.20	-11.3%	-12.3%
NOV	02-Apr-2020	09-Apr-2020	6	9.70	12.32	+28.1%	+27.1%
POLA	02-Apr-2020	09-Apr-2020	6	1.20	1.19	-0.0%	-1.0%
PRPL	02-Apr-2020	13-Apr-2020	7	5.17	7.06	+36.8%	+35.8%
PVH	02-Apr-2020	13-Apr-2020	7	35.00	49.38	+41.3%	+40.3%

SKT	02-Apr-2020	09-Apr-2020	6	4.60	6.90	+51.2%	+50.2%
SLB	02-Apr-2020	09-Apr-2020	6	13.10	18.38	+41.5%	+40.5%
SPG	02-Apr-2020	13-Apr-2020	7	46.12	68.60	+49.0%	+48.0%
TALO	02-Apr-2020	09-Apr-2020	6	5.33	7.46	+41.1%	+40.1%
TMHC	02-Apr-2020	09-Apr-2020	6	8.90	12.67	+43.5%	+42.5%
TRGP	02-Apr-2020	13-Apr-2020	7	6.01	9.16	+52.7%	+51.7%
WPG	02-Apr-2020	14-Apr-2020	8	5.82	8.32	+43.4%	+42.4%
WRI	02-Apr-2020	13-Apr-2020	7	13.48	17.59	+30.7%	+29.7%
ADS	03-Apr-2020	13-Apr-2020	6	28.41	36.87	+28.7%	+27.7%
AGS	03-Apr-2020	14-Apr-2020	7	1.64	3.30	+99.8%	+98.8%
AKR	03-Apr-2020	13-Apr-2020	6	10.36	14.47	+38.6%	+37.6%
ARLP	03-Apr-2020	13-Apr-2020	6	3.26	3.33	+1.3%	+0.3%
BBBY	03-Apr-2020	13-Apr-2020	6	3.55	5.70	+59.3%	+58.3%
BGFV	03-Apr-2020	14-Apr-2020	7	0.75	0.99	+31.1%	+30.1%
BGSF	03-Apr-2020	13-Apr-2020	6	6.59	9.09	+36.8%	+35.8%
CDR	03-Apr-2020	15-Apr-2020	8	4.64	5.89	+25.5%	+24.5%
CHS	03-Apr-2020	15-Apr-2020	8	1.03	1.54	+47.6%	+46.6%
CLNY	03-Apr-2020	13-Apr-2020	6	1.61	2.45	+51.0%	+50.0%
EBIX	03-Apr-2020	13-Apr-2020	6	12.67	16.80	+31.5%	+30.5%
EXPR	03-Apr-2020	13-Apr-2020	6	1.46	2.47	+67.8%	+66.8%
FLY	03-Apr-2020	15-Apr-2020	8	5.14	6.23	+19.7%	+18.7%
HOV	03-Apr-2020	15-Apr-2020	8	6.76	10.30	+50.4%	+49.4%
KIM	03-Apr-2020	13-Apr-2020	6	7.69	10.80	+39.3%	+38.3%
LB	03-Apr-2020	13-Apr-2020	6	10.14	14.93	+46.1%	+45.1%
MGNI	03-Apr-2020	13-Apr-2020	6	4.77	5.74	+19.4%	+18.4%
NYMT	03-Apr-2020	13-Apr-2020	6	1.20	2.39	+97.6%	+96.6%
OMF	03-Apr-2020	13-Apr-2020	6	16.94	22.83	+33.7%	+32.7%
PLCE	03-Apr-2020	13-Apr-2020	6	16.03	26.00	+60.9%	+59.9%
TDG	03-Apr-2020	13-Apr-2020	6	282.67	340.54	+19.5%	+18.5%
EYES	06-Apr-2020	14-Apr-2020	6	0.75	1.32	+74.2%	+73.2%
ARR	07-Apr-2020	15-Apr-2020	6	8.37	8.27	-2.8%	-3.8%
AZUL	07-Apr-2020	15-Apr-2020	6	10.23	9.23	-11.3%	-12.3%
DAL	07-Apr-2020	15-Apr-2020	6	25.21	24.91	-2.8%	-3.8%
EPRT	07-Apr-2020	15-Apr-2020	6	11.43	11.33	-2.5%	-3.5%
HOME	07-Apr-2020	15-Apr-2020	6	1.85	2.05	+9.0%	+8.0%
MCFT	07-Apr-2020	15-Apr-2020	6	5.64	7.47	+30.3%	+29.3%
NLY	07-Apr-2020	15-Apr-2020	6	4.67	5.75	+21.1%	+20.1%
RC	07-Apr-2020	15-Apr-2020	6	4.84	6.25	+27.0%	+26.0%
SNR	07-Apr-2020	15-Apr-2020	6	2.31	3.10	+32.0%	+31.0%
UAL	07-Apr-2020	15-Apr-2020	6	28.20	31.97	+11.5%	+10.5%
WSR	07-Apr-2020	15-Apr-2020	6	5.20	6.53	+23.5%	+22.5%
SMHI	13-Apr-2020	20-Apr-2020	6	4.19	3.17	-23.9%	-24.9%
ATRO	15-Apr-2020	22-Apr-2020	6	8.40	7.85	-5.5%	-6.5%
BPFH	15-Apr-2020	22-Apr-2020	6	6.88	7.15	+5.1%	+4.1%
DBD	15-Apr-2020	22-Apr-2020	6	3.99	3.57	-9.5%	-10.5%
FBC	15-Apr-2020	22-Apr-2020	6	20.57	21.18	+4.2%	+3.2%
AAL	24-Apr-2020	01-May-2020	6	10.40	11.41	+8.0%	+7.0%
CCL	24-Apr-2020	01-May-2020	6	12.25	14.90	+19.8%	+18.8%
FLY	24-Apr-2020	01-May-2020	6	5.09	5.78	+11.8%	+10.8%
NCLH	24-Apr-2020	01-May-2020	6	11.37	15.25	+32.1%	+31.1%
SIG	24-Apr-2020	01-May-2020	6	7.44	9.66	+27.8%	+26.8%
SPR	24-Apr-2020	04-May-2020	7	19.50	19.14	-3.5%	-4.5%
VNTR	24-Apr-2020	01-May-2020	6	1.34	1.25	-8.2%	-9.2%
WD	24-Apr-2020	01-May-2020	6	33.44	37.45	+10.3%	+9.3%
CIM	27-Apr-2020	04-May-2020	6	7.30	7.56	+2.3%	+1.3%
WES	07-May-2020	14-May-2020	6	7.51	6.82	-9.4%	-10.4%
ERJ	15-May-2020	22-May-2020	6	4.46	4.80	+6.2%	+5.2%
EYES	19-May-2020	27-May-2020	6	0.94	1.04	+10.0%	+9.0%
OCN	11-Jun-2020	18-Jun-2020	6	10.35	10.35	+1.1%	+0.1%
AER	15-Jun-2020	22-Jun-2020	6	32.28	31.67	-1.2%	-2.2%
AXL	15-Jun-2020	22-Jun-2020	6	7.30	7.35	+1.4%	+0.4%
DAN	15-Jun-2020	22-Jun-2020	6	11.70	11.87	+2.1%	+1.1%
LB	15-Jun-2020	22-Jun-2020	6	14.74	14.50	-1.0%	-2.0%
MTP	15-Jun-2020	22-Jun-2020	6	1.37	1.36	-0.1%	-1.1%
AYRO	23-Jun-2020	30-Jun-2020	6	2.73	3.00	+10.2%	+9.2%
BIOL	06-Jul-2020	13-Jul-2020	6	0.39	0.50	+27.5%	+26.5%
AAL	07-Jul-2020	14-Jul-2020	6	12.51	11.32	-9.8%	-10.8%
ACR	07-Jul-2020	14-Jul-2020	6	6.75	6.66	-1.6%	-2.6%
AHT	07-Jul-2020	14-Jul-2020	6	7.20	6.00	-16.9%	-17.9%
AXAS	07-Jul-2020	14-Jul-2020	6	4.40	4.20	-4.8%	-5.8%
FOSL	07-Jul-2020	14-Jul-2020	6	4.26	3.91	-8.5%	-9.5%

LTHM	07-Jul-2020	14-Jul-2020	6	6.33	6.14	-3.3%	-4.3%
MTDR	07-Jul-2020	14-Jul-2020	6	7.86	8.08	+2.5%	+1.5%
NCLH	07-Jul-2020	14-Jul-2020	6	15.88	15.22	-4.5%	-5.5%
NR	07-Jul-2020	15-Jul-2020	7	1.97	1.90	-4.4%	-5.4%
OKE	07-Jul-2020	14-Jul-2020	6	27.75	26.37	-5.3%	-6.3%
PLAY	07-Jul-2020	14-Jul-2020	6	12.58	11.67	-7.5%	-8.5%
SIG	07-Jul-2020	14-Jul-2020	6	10.32	10.37	+0.2%	-0.8%
SM	07-Jul-2020	14-Jul-2020	6	3.52	3.30	-6.5%	-7.5%
TALO	07-Jul-2020	14-Jul-2020	6	8.79	7.50	-14.9%	-15.9%
WLL	07-Jul-2020	14-Jul-2020	6	1.10	1.01	-8.5%	-9.5%
ICON	08-Jul-2020	15-Jul-2020	6	0.79	0.91	+13.8%	+12.8%
PBF	08-Jul-2020	15-Jul-2020	6	8.49	9.33	+8.6%	+7.6%
OBSV	13-Jul-2020	20-Jul-2020	6	2.71	2.90	+5.8%	+4.8%
AVEO	14-Jul-2020	21-Jul-2020	6	4.22	4.75	+11.5%	+10.5%
CLSN	20-Jul-2020	27-Jul-2020	6	1.09	1.24	+11.7%	+10.7%
APLT	22-Jul-2020	29-Jul-2020	6	26.63	25.13	-7.1%	-8.1%
EOLS	22-Jul-2020	29-Jul-2020	6	3.24	3.43	+4.2%	+3.2%
LQDA	24-Jul-2020	04-Aug-2020	8	6.00	5.89	-3.2%	-4.2%
NKLA	27-Jul-2020	06-Aug-2020	9	31.60	34.99	+8.7%	+7.7%
OBSV	29-Jul-2020	07-Aug-2020	8	2.63	2.66	-0.1%	-1.1%
GNUS	31-Jul-2020	07-Aug-2020	6	1.56	1.66	+6.4%	+5.4%
EOLS	03-Aug-2020	10-Aug-2020	6	3.27	3.51	+7.4%	+6.4%
NOVN	18-Aug-2020	25-Aug-2020	6	5.60	4.43	-20.3%	-21.3%
VJET	20-Aug-2020	27-Aug-2020	6	5.00	7.90	+58.0%	+57.0%
ONVO	24-Aug-2020	31-Aug-2020	6	7.75	8.72	+11.4%	+10.4%
KODK	27-Aug-2020	04-Sep-2020	7	6.17	6.40	+3.7%	+2.7%
PRPO	27-Aug-2020	03-Sep-2020	6	2.20	2.14	-2.8%	-3.8%
CODX	31-Aug-2020	11-Sep-2020	9	11.41	11.75	+3.8%	+2.8%
IBIO	01-Sep-2020	09-Sep-2020	6	2.03	2.21	+10.4%	+9.4%
SHIP	01-Sep-2020	10-Sep-2020	7	0.48	0.46	-3.1%	-4.1%
UONE	01-Sep-2020	09-Sep-2020	6	3.64	3.47	-3.3%	-4.3%
APRN	02-Sep-2020	10-Sep-2020	6	6.71	7.16	+7.8%	+6.8%
CIDM	02-Sep-2020	10-Sep-2020	6	0.81	0.67	-16.5%	-17.5%
ONTX	02-Sep-2020	14-Sep-2020	8	3.45	3.45	+0.7%	-0.3%
XSPA	02-Sep-2020	10-Sep-2020	6	1.97	2.10	+7.7%	+6.7%
SUNW	03-Sep-2020	15-Sep-2020	8	0.63	0.80	+26.8%	+25.8%
ACST	08-Sep-2020	15-Sep-2020	6	0.19	0.20	+4.8%	+3.8%
PXLW	08-Sep-2020	15-Sep-2020	6	2.12	2.27	+6.6%	+5.6%
SRNE	08-Sep-2020	15-Sep-2020	6	6.05	7.64	+25.8%	+24.8%
PAYS	11-Sep-2020	18-Sep-2020	6	5.55	5.87	+5.5%	+4.5%
BPT	15-Sep-2020	22-Sep-2020	6	1.60	1.65	+4.0%	+3.0%
RIG	15-Sep-2020	22-Sep-2020	6	1.03	1.01	-1.1%	-2.1%
PEI	16-Sep-2020	23-Sep-2020	6	0.79	0.59	-24.5%	-25.5%
CRBP	23-Sep-2020	30-Sep-2020	6	1.65	1.99	+20.2%	+19.2%
NKLA	28-Sep-2020	05-Oct-2020	6	19.66	24.52	+23.8%	+22.8%
ACB	29-Sep-2020	06-Oct-2020	6	4.71	4.65	-2.2%	-3.2%
AMRN	29-Sep-2020	06-Oct-2020	6	4.00	5.09	+26.1%	+25.1%
RIG	05-Oct-2020	12-Oct-2020	6	0.82	0.84	+1.6%	+0.6%
VRS	14-Oct-2020	21-Oct-2020	6	7.98	8.05	+0.1%	-0.9%
BTU	16-Oct-2020	23-Oct-2020	6	1.68	1.60	-5.6%	-6.6%
MICT	19-Oct-2020	26-Oct-2020	6	3.03	2.75	-10.3%	-11.3%
GLBS	26-Oct-2020	02-Nov-2020	6	7.85	7.41	-3.9%	-4.9%
AMC	29-Oct-2020	06-Nov-2020	7	2.65	2.43	-8.8%	-9.8%
DKNG	30-Oct-2020	06-Nov-2020	6	37.43	42.23	+11.5%	+10.5%
BLCM	09-Nov-2020	16-Nov-2020	6	2.89	2.91	+1.1%	+0.1%
FSLY	11-Nov-2020	18-Nov-2020	6	72.50	79.38	+9.1%	+8.1%
NVCN	12-Nov-2020	23-Nov-2020	8	0.81	0.87	+6.6%	+5.6%
OCUP	24-Nov-2020	02-Dec-2020	6	5.77	6.07	+3.2%	+2.2%
AUTO	04-Dec-2020	11-Dec-2020	6	2.80	2.79	-0.3%	-1.3%
QDEL	04-Dec-2020	11-Dec-2020	6	198.46	197.15	-0.6%	-1.6%
BCLI	08-Dec-2020	21-Dec-2020	10	5.50	5.22	-5.9%	-6.9%
CBAT	15-Dec-2020	22-Dec-2020	6	4.97	5.82	+16.2%	+15.2%
IBIO	15-Dec-2020	22-Dec-2020	6	1.18	1.21	+1.8%	+0.8%
QTT	22-Feb-2021	01-Mar-2021	6	4.36	3.81	-12.4%	-13.4%
GNUS	25-Feb-2021	04-Mar-2021	6	1.80	1.61	-9.7%	-10.7%
EBIX	01-Mar-2021	08-Mar-2021	6	24.98	25.10	+1.9%	+0.9%
ATOS	03-Mar-2021	10-Mar-2021	6	2.69	2.99	+12.9%	+11.9%
WKHS	04-Mar-2021	12-Mar-2021	7	14.31	16.08	+13.0%	+12.0%
EH	05-Mar-2021	15-Mar-2021	7	36.00	52.30	+45.5%	+44.5%
MTP	05-Mar-2021	12-Mar-2021	6	1.93	2.24	+15.9%	+14.9%
TWST	05-Mar-2021	12-Mar-2021	6	118.11	136.79	+15.7%	+14.7%

CRMD	08-Mar-2021	15-Mar-2021	6	7.35	9.40	+27.6%	+26.6%
FOLD	08-Mar-2021	15-Mar-2021	6	10.55	10.81	+2.2%	+1.2%
OTRK	08-Mar-2021	16-Mar-2021	7	25.94	37.05	+42.8%	+41.8%
AEZS	09-Mar-2021	16-Mar-2021	6	1.15	1.52	+31.3%	+30.3%
EKSO	09-Mar-2021	16-Mar-2021	6	6.11	7.21	+17.3%	+16.3%
FSLY	10-Mar-2021	17-Mar-2021	6	70.00	72.00	+2.8%	+1.8%
IMMR	10-Mar-2021	17-Mar-2021	6	10.37	10.13	-2.4%	-3.4%
TXMD	10-Mar-2021	17-Mar-2021	6	1.40	1.56	+11.4%	+10.4%
XONE	10-Mar-2021	17-Mar-2021	6	33.52	30.89	-7.9%	-8.9%
WKHS	16-Mar-2021	29-Mar-2021	10	17.86	13.38	-24.2%	-25.2%
UAVS	17-Mar-2021	24-Mar-2021	6	6.98	7.11	+2.4%	+1.4%
CNST	18-Mar-2021	25-Mar-2021	6	29.45	23.44	-19.3%	-20.3%
IOVA	18-Mar-2021	25-Mar-2021	6	32.14	28.95	-8.7%	-9.7%
NOVA	18-Mar-2021	25-Mar-2021	6	36.34	32.73	-8.7%	-9.7%
QDEL	18-Mar-2021	25-Mar-2021	6	140.44	126.10	-8.9%	-9.9%
SRAC	18-Mar-2021	25-Mar-2021	6	14.13	13.56	-2.7%	-3.7%
TWLO	18-Mar-2021	25-Mar-2021	6	380.68	318.38	-15.2%	-16.2%
NNDM	19-Mar-2021	31-Mar-2021	9	9.95	8.60	-12.1%	-13.1%
PLAN	19-Mar-2021	26-Mar-2021	6	53.61	52.22	-1.4%	-2.4%
RIDE	19-Mar-2021	31-Mar-2021	9	12.98	11.74	-8.1%	-9.1%
RIOT	19-Mar-2021	26-Mar-2021	6	61.03	49.41	-18.1%	-19.1%
TIGR	19-Mar-2021	29-Mar-2021	7	21.13	14.92	-28.6%	-29.6%
NCTY	22-Mar-2021	29-Mar-2021	6	50.31	37.75	-24.3%	-25.3%
SOS	22-Mar-2021	29-Mar-2021	6	7.58	6.20	-17.5%	-18.5%
BWEN	24-Mar-2021	31-Mar-2021	6	5.35	6.85	+29.3%	+28.3%
CRMD	24-Mar-2021	05-Apr-2021	8	9.63	10.34	+8.1%	+7.1%
FCEL	24-Mar-2021	06-Apr-2021	9	14.07	13.46	-4.2%	-5.2%
ICPT	24-Mar-2021	31-Mar-2021	6	22.18	22.74	+3.6%	+2.6%
INSG	24-Mar-2021	31-Mar-2021	6	10.15	9.75	-3.0%	-4.0%
PLUG	24-Mar-2021	31-Mar-2021	6	37.04	36.32	-1.0%	-2.0%
SOL	24-Mar-2021	31-Mar-2021	6	12.01	12.67	+6.6%	+5.6%
SSYS	24-Mar-2021	06-Apr-2021	9	26.24	25.42	-2.9%	-3.9%
TPIC	24-Mar-2021	31-Mar-2021	6	49.55	56.01	+14.2%	+13.2%
VERB	24-Mar-2021	31-Mar-2021	6	1.60	1.42	-10.4%	-11.4%
VERI	24-Mar-2021	07-Apr-2021	10	27.33	25.05	-8.6%	-9.6%
SFIX	29-Mar-2021	06-Apr-2021	6	48.67	49.65	+1.7%	+0.7%
XL	30-Mar-2021	07-Apr-2021	6	8.67	8.92	+2.0%	+1.0%
PRG	06-Apr-2021	13-Apr-2021	6	26.53	43.28	+61.9%	+60.9%
NCTY	08-Apr-2021	15-Apr-2021	6	26.13	23.52	-10.8%	-11.8%
NCTY	06-May-2021	13-May-2021	6	19.36	15.60	-19.9%	-20.9%

Bibliografía

[1] A note on distance-weighted k-nearest neighbor rules. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(4):311–313, 1978.

[2] Hyunchul Ahn and Kyoung-jae Kim. Using genetic algorithms to optimize nearest neighbors for data mining. *Annals OR*, 163:5–18, 10 2008.

[3] American Statistical Association. Ethical guidelines for statistical practice. *Professional Ethics of the American Statistical Association*, page 4, 04 2016.

[4] David Bailey, Jonathan (Jon) Borwein, Marcos Lopez de Prado, and Qiji Zhu. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61:458, 05 2014.

[5] Ferid Bajramovic, Frank Mattern, Nicholas J. Butko, and Joachim Denzler. A comparison of nearest neighbor search algorithms for generic object recognition. In *ACIVS*, 2006.

[6] Victor L. Bernard. Accounting-based valuation methods, determinants of market-to-book ratios, and implications for financial statement analysis. 1994.

[7] Zvi Bodie, Alex Kane, and Alan J. Marcus. *Investments*. McGraw-Hill/Irwin, Boston, 2009.

[8] Putu Wira Buana, D. Jannet, and I Ketut Gede Darma Putra. Combination of k-nearest neighbor and k-means based on term re-weighting for classify indonesian news. *International Journal of Computer Applications*, 50:37–42, 2012.

[9] Yingjun Chen and Yongtao Hao. A feature weighted support vector machine and k-nearest neighbor algorithm for stock market indices prediction. *Expert Systems with Applications*, 80:340–355, 2017.

[10] P.L Cootner. *The Random Character of Stock Market Prices.* MIT-Press, 1964.

[11] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[12] R. J. Jordan D. E. Fisher. *Security Analysis and Portfolio Management.* 6. Prentice-Hall,, 1995.

[13] Sanket Dhanorkar. Judging mutual fund performance: Should you look at trailing returns or rolling returns?

[14] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

[15] Pablo Fernandez. Métodos de valoración de empresas. *IEEM: Revista de Antiguos Alumnos, ISSN 1510-4214, Año 3, Nº. 2, 2000, pags. 8-16*, 01 2008.

[16] M. C. Findlay and E. E. Williams. A fresh look at the efficient market hypothesis: How the intellectual history of finance encouraged a real "fraud-on-the-market". *Journal of Post Keynesian Economics*, 23(2):181–199, 2000.

[17] Evelyn Fix and Joseph L. Hodges. Discriminatory analysis - nonparametric discrimination: Consistency properties. *International Statistical Review*, 57:238, 1989.

[18] Xiubo Geng, Tie-Yan Liu, Tao Qin, Andrew Arnold, Hang Li, and Heung-Yeung Shum. Query dependent ranking using k-nearest neighbor. pages 115–122, 07 2008.

[19] K. Chidananda Gowda and G. Krishna. The condensed nearest neighbor rule using the concept of mutual nearest neighborhood (corresp.). *IEEE Trans. Inf. Theory*, 25:488–490, 1979.

[20] Eui-Hong (Sam) Han, George Karypis, and Vipin Kumar. Text categorization using weight adjusted k-nearest neighbor classification. In David Cheung, Graham J. Williams, and Qing Li, editors, *Advances in Knowledge Discovery and Data Mining*, pages 53–65, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[21] R. Edwards J. Magee, W. Bassetti. *Technical Analysis of Stock Trends.* 2019.

[22] F. Reilly K. Brown. *Investment Analysis and Portfolio Management*. 7. South-Western, 1995.

[23] Chiung-Hon Leon Lee. Modeling personalized fuzzy candlestick patterns for investment decision making. In *2009 Asia-Pacific Conference on Information Processing*, volume 2, pages 286–289, 2009.

[24] Daye Li, Zhizhong Li, and Rongrong Li. Automate the identification of technical patterns: a k-nearest-neighbour model approach. *Applied Economics*, 50(17):1978–1991, 2018.

[25] John Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, 47(1):13–37, 1965.

[26] Marcos López de Prado. The 10 reasons most machine learning funds fail. *The Journal of Portfolio Management*, 44(6):120–133, 2018.

[27] Cornelis A. Los. Nonparametric Efficiency Testing of Asian Stock Markets Using Weekly Data. Finance 0409033, University Library of Munich, Germany, September 2004.

[28] Suneetha Manne, Sita Kumari Kotha, and S. Sameen Fatima. Text categorization with k-nearest neighbor approach. 2012.

[29] Anqi Dong Matthew Bartolini. Sector business cycle analysis.

[30] Robert C. Merton. On the current state of the stock market rationality hypothesis. Working papers 1717-85., Massachusetts Institute of Technology (MIT), Sloan School of Management, 1985.

[31] Jan Mossin. Equilibrium in a capital asset market. *Econometrica*, 34(4):768–783, 1966.

[32] J. Murphy. *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*. Prentice Hall Press, 1999.

[33] Kei Nakagawa, Mitsuyoshi Imamura, and Kenichi Yoshida. *Stock Price Prediction with Fluctuation Patterns Using Indexing Dynamic Time Warping and*

$$k^*$$

-Nearest Neighbors, pages 97–111. 06 2018.

[34] R. Naranjo and M. Santos. A fuzzy decision system for money investment in stock markets based on fuzzy candlesticks pattern recognition. *Expert Systems with Applications*, 133:34–48, 2019.

[35] Rodrigo Naranjo and Matilde Santos. Fuzzy candlesticks forecasting using pattern recognition for stock markets. In Manuel Graña, José Manuel López-Guede, Oier Etxaniz, Álvaro Herrero, Héctor Quintián, and Emilio Corchado, editors, *International Joint Conference SOCO'16-CISIS'16-ICEUTE'16*, pages 323–333, Cham, 2017. Springer International Publishing.

[36] Giuseppe Nuti. An efficient algorithm for bayesian nearest neighbours. *Methodology and Computing in Applied Probability*, 21, 12 2019.

[37] Cheol-Ho Park and Scott Irwin. The profitability of technical analysis: A review. *SSRN Electronic Journal*, 11 2004.

[38] Stephen H. Penman. An evaluation of accounting rate-of-return. *Journal of Accounting, Auditing & Finance*, 6:233 – 255, 1991.

[39] V. B. Surya Prasath, Haneen Arafat Abu Alfeilat, Ahmad Basheer Hassanat, Omar Lasassmeh, S M Fakhr Ahmad, Tarawneh, Mahmoud Bashir Alhasanat, and Hamzeh Eyal Salman. Effects of distance measure choice on knn classifier performance-a review. 2019.

[40] William Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, 1964.

[41] Dedi Candro Parulian Sinaga, Tulus, and Poltak Sihombing. Performance of distance-based k-nearest neighbor classification method using local mean vector and harmonic distance. *IOP Conference Series: Materials Science and Engineering*, 725(1):012122, jan 2020.

[42] Yang Song, Jian Huang, Ding Zhou, Hongyuan Zha, and C. Lee Giles. Iknn: Informative k-nearest neighbor pattern classification. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Knowledge Discovery in Databases: PKDD 2007*, pages 248–264, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[43] Shiliang Sun and Rongqing Huang. An adaptive k-nearest neighbor algorithm. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 1, pages 91–94, 2010.

[44] Shiliang Sun and Rongqing Huang. An adaptive k-nearest neighbor algorithm. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 1, pages 91–94, 2010.

[45] Kashvi Taunk, Sanjukta De, Srishti Verma, and Aleena Swetapadma. A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 1255–1260, 2019.

[46] Lamartine Almeida Teixeira and Adriano Lorena Inácio de Oliveira. A method for automatic stock trading combining technical analysis and nearest neighbor classification. *Expert Systems with Applications*, 37(10):6885–6890, 2010.

[47] Morse Dale Thomas RDyckman. *Efficient Capital Markets and Accounting: A Critical Analysis*. Prentice-Hall, 1986.

[48] Jigang Wang, Predrag Neskovic, and Leon N. Cooper. Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. *Pattern Recognition*, 39(3):417–423, 2006.

[49] Dietrich Wettschereck and Thomas G. Dietterich. Locally adaptive nearest neighbor algorithms. In *NIPS*, 1993.

[50] Xindong Wu, Vipin Kumar, Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, G. Mclachlan, Shu Kay Angus Ng, Bing Liu, Philip Yu, Zhi-Hua Zhou, Michael Steinbach, David Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14, 12 2007.

[51] Yue Wu, Jun-Wei Liu, Chen-Zhuo Zhu, Zhuang-Fei Bai, Qi-Guang Miao, Wen-Ping Ma, and Mao-Guo Gong. Computational intelligence in remote sensing image registration: A survey. *International Journal of Automation and Computing*, 18, 12 2020.

[52] Hao Zhang, A.C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2126–2136, 2006.

[53] Ningning Zhang, Aijing Lin, and Pengjian Shang. Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting. *Physica A: Statistical Mechanics and its Applications*, 477(C):161–173, 2017.

[54] Wei Zheng. An improved k-nearest neighbor classification algorithm using shared nearest neighbor similarity. 2015.