Universidad de Buenos Aires

Facultad de Ciencias Exactas y Nsturales

Características basadas en agrupamiento de series temporales para predicción de valores de bolsa.

Tesis presentada para optar al título de Magíster en Explotación de Datos y Descubrimiento del Conocimiento

Tesista: Lic. Javier Vásquez Sáenz
Director : Dr. Facundo Manuel Quiroga
Director : Dr. Aurelio Fernández Bariviera
Buenos Aires, 3 de agosto de 2023

Agradecimientos

Quiero agradecer en primer término a mi familia, Carol, Malena y Victoria, quienes brindaron su apoyo durante la maestría y luego durante la tesis, alentándome permanentemente. Sin ellas este hito no hubiera sido posible

Por el otro lado a mis directores de tesis, Aurelio y Facundo, por acompañarme, guiándome y dándome su apoyo durante el desarrollo de la misma y por desafiarme en el transcurso de la misma para generar un artículo con ella y emprender simultáneamente el reto de lograr su publicación en una revista internacional.

Por último a Lindor Martin, por las largas charlas donde nació la semilla de esta tesis, por su ayuda en obtener los datos iniciales y por servir de abogado del diablo a las ideas más descabelladas.

A todos ellos mi enorme agradecimiento por sus roles en este trabajo.

Índice general

Resumen

En esta tesis proponemos la hipótesis que incluir información de acciones relacionadas a la acción a predecir mejora la calidad de dicha predicción. Estas acciones relacionadas pueden obtenerse mediante distintos métodos de agrupamiento sobre información de precios, rendimientos y ratios derivados de los balances trimestrales.

En la figura 1.1 se representa el proceso completo.

Nuestro conjunto de datos se basó en las acciones de las empresas de que cotizan en los mercados de valores de los Estados Unidos, dentro de las 3000 empresas con mayor capitalización bursátil, listadas en el índice Russell 3000. Dentro de dichas empresas, se han impuesto dos condiciones: (i) que hayan cotizado y (ii) que haya presentado balances en el período de enero de 2017 a junio de 2022, representando un total de 2359 acciones para trabajar.

Utilizamos los balances trimestrales para generar 10 ratios financieros y los convertimos en series temporales. También utilizamos las series habituales de precios de cierre y rendimientos diarios. Con estos conjuntos de series temporales agrupamos las acciones representadas con estas características utilizando K-Means o K-Medoids con distancias basadas en Dynamic Time Warping, Extended Frobenius Norm y en la distancia euclidiana de los coeficientes generados por transformadas rápidas de Fourier.

La información derivada de estos agrupamientos permitió seleccionar las características a proporcionar a modelos de ARIMA y Redes Neuronales, utilizando no solo los datos de la acción a predecir, sino la correspondiente al conjunto de acciones presentes en el grupo al cual pertenece la acción, para cada tipo de agrupamiento. El objetivo de este proceso es mejorar la predicción de precios o rendimientos y generar a partir de ello recomendaciones de inversión.

Para validar la estrategia propuesta, comparamos el efecto de los diferentes modelos de agrupamiento y predicción sobre los rendimientos generados por una estrategia de inversión. Comparamos asimismo estos resultados con los obtenidos a través de métodos tradicionales de inversión en los mercados de valores.

Nuestros resultados muestran una ventaja en el uso de la información adicional proporcionada por los métodos de agrupación. Además, los métodos basados en Redes Neuronales resultan en una mayor tasa de rendimiento que aquellos basados en métodos tradicionales como ARIMA.

Las principales contribuciones de esta tesis fueron introducir las series temporales derivadas de los balances financieros para realizar agrupamientos de datos y usar los resultados de estas agrupaciones para seleccionar así la información más útil para los modelos de ARIMA y Redes Neuronales.

Capítulo 1

Introducción

El comportamiento de las bolsas de valores ha sido objeto de debate durante más de 100 años. El modelo seminal de Bachelier (1900) podría considerarse el primer intento teórico de modelar los precios de los bonos, sin embargo, no fue hasta mediados del siglo XX que se desarrollaron los modelos formales. Osborne (1959) y Osborne (1962) encuentran evidencia de que el movimiento del precio sigue un recorrido aleatorio simple y poco después, Samuelson (1965) demostró que los precios adecuadamente anticipados deberían seguir un camino aleatorio. Tal situación se cristalizó en la Hipótesis del Mercado Eficiente (Efficient Market Hipothesis o EMH), que, tal como la define Fama (1970), describe un mercado donde los precios reflejan toda la información disponible. Fama propone que la eficiencia de la información se clasifica en tres grandes categorías:

- (i) la eficiencia débil corresponde al mercado donde los precios reflejan la información contenida en los precios pasados;
- (ii) la eficiencia semifuerte refleja un mercado donde los precios incorporan toda la información pública (más allá de los precios);
- (iii) la fuerte eficiencia afirma que los precios reflejan toda la información pública y privada.

El EMH permaneció indiscutible hasta la década de 1980, a partir de cuando los datos y el poder computacional se volvieron más accesibles. Luego, los artículos empíricos encontraron inconsistencias entre las regularidades empíricas y los modelos teóricos, resultando en el trabajo de Lo (2004) quien propuso la Hipótesis de los Mercados Adaptativos. Bajo esa hipótesis, más realista que lo anterior, los mercados pueden tener pequeñas desviaciones temporales del teórico comportamiento estocástico, que si son adecuadamente detectados y explotados, pueden ser utilizados para estrategias de inversión rentables.

Hoy existe consenso en que el mercado de valores es un sistema complejo y dinámico, con variables difíciles de predecir, datos desconocidos y una gran sensibilidad a eventos impredecibles que pueden causar disrupciones. Esta complejidad se ve incrementada por la reacción emocional y de manada de muchos actores del mercado, que amplifican las tendencias subyacentes y pueden generar corridas de mercado.

En busca de esta ventaja se han generado numerosos mecanismos para predecir el comportamiento de los mercados y en consecuencia encontrar estrategias

[1] https://www.investopedia.com/articles/technical/112601.asp
[2] https://tradingstrategyguides.com/chart-pattern-trading-strategy/

de inversión con una rentabilidad superior a la media. Así, se han desarrollado innumerables modelos cuantitativos y cualitativos para predecir tendencias de mercado y precios, tales como indicadores estadísticos, técnicos, morfológicos o gráficos. Algunos de estos métodos tienen una base metodológica sólida, mientras que otros obedecen más a supersticiones o costumbres de antaño. Entre los métodos más populares se encuentran las medias móviles, las medias móviles exponenciales, las estructuras temporales de precios, los ratios de volumen diario o los métodos que determinan, basándose en patrones gráficos [1] ,[2] , la posibilidad de una suba o baja del precio de una determinada acción.

Otros métodos más sofisticados utilizan correlaciones entre diferentes indicadores económicos con grupos de acciones, como el precio del petróleo y las acciones energéticas, la producción de petróleo o el transporte, para encontrar diferencias que puedan arbitrarse con éxito. Esta correlación puede extenderse a la asociatividad con indicadores no tan obvios o grupos de acciones o datos que no necesariamente pertenecen a industrias similares, apoyada en técnicas como el clustering. Finalmente, se han propuesto numerosos modelos estadísticos, de aprendizaje automático y de aprendizaje profundo para poder predecir tendencias y precios. En este contexto, estos modelos basados en datos tienen la ventaja de que se pueden ejecutar de forma continua.

Si bien aún no se ha encontrado y aceptado una respuesta definitiva sobre el comportamiento del mercado, los investigadores e inversionistas han seguido explorando el tema con diversos niveles de éxito, para capturar la información disponible y así generar modelos y estrategias de inversión exitosas.

Dada la dificultad, y en muchos casos el objetivo de desarrollar métodos que generen rentabilidades superiores a la media, existe una gran cantidad de estudios y trabajos de predicción de precios o tendencias de los distintos mercados bursátiles. Como dato de interés, Google Scholar encuentra más de 1 millón de artículos que hacen referencia a *stock price prediction* de 1950 a 2022, y 100.000 de 2018 a 2022. En el caso de *stock price clustering*, el tema es menos explorado, con solo 580.000 y 50.000 en los mismos periodos. De los artículos que mencionan ambos términos, solo hay 21.000 desde 1950 y 17.000 desde 2018, lo que indica que se trata de una línea de investigación más reciente.

1.1. Aportes de la tesis

En este trabajo exploramos la hipótesis de que un agrupamiento de los datos utilizado por un método de predicción puede mejorar los resultados obtenidos por una estrategia de inversión, en comparación con un modelo de predicción que no tiene en cuenta las relaciones entre diferentes acciones utilizando los datos de las acciones en cada grupo. La figura 1.1 muestra un diagrama de proceso que incluye no solo los datos propios de una acción sino que agrega un paso de agrupación para identificar datos adicionales relevantes antes de entrenar modelos de pronóstico.

Agrupamos las acciones en términos de sus ratios financieros construidos a partir de los informes financieros trimestrales, la serie de precios de acciones y la de rendimientos diarios. Dado que estos tipos de datos son series de tiempo, pero de diferente naturaleza, exploramos varias características y medidas de distancia para realizar el agrupamiento.

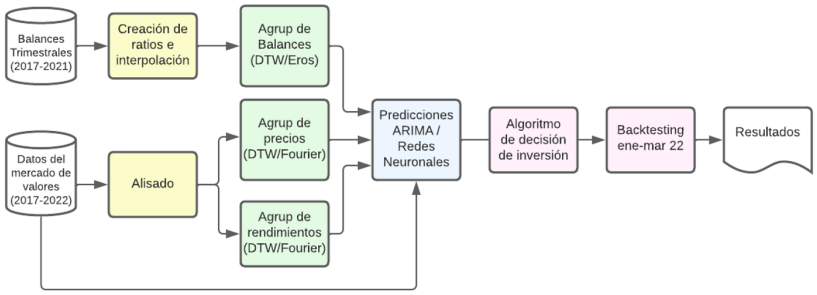

Figura 1.1: Diagrama de nuestro modelo. Utilizamos datos tanto de informes trimestrales como de series temporales de precios y rendimientos de acciones. Estos se agrupan de forma independiente, ya que requieren medidas de distancia especializadas. Utilizamos K-Means como un algoritmo de agrupamiento, con una prueba de Silhouette para determinar la cantidad óptima de agrupamientos que proporcionarán suficiente diversidad. Las diferentes medidas de distancia comparadas se basaron en Dynamic Time Warping (DTW), descomposición de Fourier y EROS, esta última una medida de similitud que combina PCA y la norma de Frobenius. Usando la información de agrupamiento, entrenamos diferentes modelos de pronóstico basados en ARIMA y Redes Neuronales. Finalmente, usamos estas predicciones para realizar una prueba con un algoritmo de inversión utilizando datos históricos.

Con estos resultados de agrupación, entrenamos modelos de predicción para cada agrupación para aislar la dinámica de submercados específicos. Comparamos modelos estadísticos y de aprendizaje profundo para predecir la tendencia de precios y rendimientos de un conjunto representativo de empresas de los principales mercados de valores de Estados Unidos. De esta manera validamos si este esquema nos permitió obtener una mayor ganancia en comparación con los métodos tradicionales.

En general los balances trimestrales han sido poco utilizados para agrupar empresas, usando habitualmente el último balance publicado, ya sea a través de la generación de diversos ratios o mediante el análisis semántico de los mismos. Sin embargo, no hemos encontrado antecedentes donde utilicen la historia de los balances de cada compañía para generar series temporales de ratios que permitan agrupar empresas basados en la evolución histórica de sus resultados financieros. Esta comparación puede proporcionar información útil para la predicción, ya que permitiría detectar empresas que, siendo similares, se encuentren en distintos estadios de su evolución.

Un punto importante de esta tesis es el volumen de datos empleado. Habitualmente los trabajos, que pueden verse en la sección 2 se realizan con conjuntos pequeños de acciones, usualmente menos de 10. En este trabajo el piloto se efectuó con 264 acciones y la ejecución completa con 2359 acciones, representando más del 95 % del volumen de los mercados analizados.

1.2. Estructura de la tesis

La tesis está organizada de la siguiente manera. La sección 2, revisa los puntos teóricos más relevantes de los temas tratados, tanto en el aspecto financiero como del procesamiento de datos. Luego, la sección 3 describe el conjunto de datos y las características utilizadas y la sección 4 se adentra en el procesamiento de datos, la exploración realizada con los métodos de agrupamiento, predicción e inversión en un subconjunto de 264 acciones, evaluando sus resultados. La sección 5 presenta la ejecución y evaluación de los métodos más eficientes en el conjunto completo de 2359 acciones y analiza los resultados respectivos. Finalmente, la sección 6 resume las conclusiones de nuestro trabajo y delinea futuras líneas de investigación. En los apéndices hemos colocado aquellos datos y resultados que podrían ser de interés al lector como referencia.

Comenzaremos entonces en la próxima sección, describiendo características de los principales mercados de valores de los Estados Unidos, las operaciones más comunes en dichos mercados, las características de la información presentada por las empresas que en ella cotizan y algunos métodos habituales de inversión. Luego describiremos algunos algoritmos de agrupación, similitud, predicción y de evaluación de resultados, para terminar describiendo trabajos que han explorado en parte el tema de esta tesis y que sirven como punto de partida para su desarrollo.

1.3. Publicación

En base a los hallazgos del piloto escribimos junto a los Directores de Tesis un artículo llamado *Data vs. information: Using clustering techniques to enhance stock returns forecasting* que fuera publicado en el volumen 88 de julio de 2023 en la *International Review of Financial Analysis* [3] . Este artículo describe más detalladamente los resultados del piloto y sus conclusiones.

[3] https://doi.org/10.1016/j.irfa.2023.102657

Capítulo 2

Marco Teórico

Esta sección desarrollará las técnicas utilizadas a lo largo de esta tesis, con foco mayor en aquellas no tan conocidas o que correspondan al mercado financiero o de inversión. Por último, expondremos algunos antecedentes en trabajos similares o trabajos previos que resultan de utilidad para el desarrollo de la tesis.

2.1. Mercado Financiero

El mercado de valores de los Estados Unidos está dominado por dos grandes bolsas de valores, el New York Stock Exchange (NYSE) [1] con más de 230 años de antigüedad y unas 2400 acciones comercializadas con un valor de mercado de 26 billones de dólares a 2021 y el National Association of Securities Dealers Automated Quotations (NASDAQ) [2] con 50 años de antigüedad y unas 3500 acciones listadas de 19 billones de dólares de valor de mercado a 2021. En ambas bolsas cotizan empresas de todo el mundo y representan el mayor mercado de acciones mundial.

Estos mercados son muy dinámicos y constituyen objeto de atención hace años con distintos niveles de sofisticación, cotizando a través de distintos instrumentos, empresas con valores de mercado desde 2,5 billones de dólares hasta empresas con una cotización de 1 a 5 millones de dólares. Esta diversidad y liquidez genera un atractivo para los inversores, quienes pueden realizar operaciones a través de varios instrumentos como la compra y venta de acciones y opciones de compra o venta, la negociación de índices y valores que apuesten a la baja o alza de acciones u otros índices subyacentes.

Como requisito para la cotización de sus acciones en estas bolsas de valores, las empresas están obligadas a comunicar formalmente hechos relevantes en las compañías, tales como compra y venta de otras compañías, emisiones de acciones, compra o venta de acciones por parte del personal ejecutivo, entre otros. La comunicación formal más relevante que deben realizar, es la publicación de balances trimestrales, los cuales deben contener información contable y financiera sobre el estado de la compañía. Estos balances trimestrales contienen diversa información, tal como ventas, ganancias, costos, impuestos, deudas, inventario, activos o flujo de caja, entre otros indicadores relevantes. Estos indicadores están

[1] www.nyse.com
[2] www.nasdaq.com

estandarizados por la Comisión de la Bolsa de Valores (SEC) y basados en los principios contables aceptados en los Estados Unidos.

El mercado de valores suele operar principalmente en días hábiles entre las 9:30 hs y 16 hs, período en el cual ocurre la mayor parte de las transacciones. Los precios de las acciones van variando en el tiempo como respuesta a la oferta y demanda existente, por factores intrínsecos (un buen o un mal resultado de una empresa), factores externos (un anuncio de crecimiento económico o una crisis que pueda eventualmente impactar en las empresas cotizantes) o factores de colusión de los participantes (donde fuercen alzas o bajas en búsqueda de beneficios). Existe también una actividad menor fuera del horario principal del mercado, extendiéndose unas 5 horas antes de la apertura y 4 horas posteriores al cierre.

2.1.1. Operaciones de los mercados de valores

La actividad más simple es la compra de instrumentos, con el objetivo de venderlas más tarde con un beneficio, esto también es conocido como compra en largo. Existe asimismo la posibilidad de alquilar un instrumento a un cierto costo diario para proceder a su venta, apostando a la baja del mismo y de esta forma recomprarlo a un valor menor y así obtener una ganancia, una vez restado el costo del alquiler del instrumento. Esta operación es conocida como venta en descubierto (*Short Sale*) y permite obtener ganancias con la baja del precio de un instrumento financiero.

En general, para realizar una operación, ya sea de compra o venta, existen varias formas distintas, siendo las más populares las siguientes

- A valor de mercado (Market Price): La operación se realiza al valor existente al momento de concretar la transacción, sin restricción alguna.
- Limitada (Limit): La operación tiene un precio límite al que debe realizarse, por ejemplo compro acciones a un precio máximo de 10 dólares o vendo a un precio mínimo de 5 dólares. Si la cotización está fuera de ese límite, la transacción no se realiza sino hasta que se cumpla la condición límite.
- Detención (Stop): La operación se ejecuta cuando el instrumento alcanza un precio determinado, por ejemplo compro una acción si su valor supera los 10 dólares, pero no antes.
- Detención móvil (Trailing Stop): Fijo un límite a partir del cual la transacción se realiza, el cual es móvil de acuerdo a un determinado porcentaje del precio o monto. Por ejemplo, tengo una acción y determino la venta de la misma si su valor disminuye en 1 dólar. Si la misma vale 5 dólares, esto significará que la venderé si baja a 4 dólares, pero si la misma sube a 7 dólares, su precio de venta cambiará a 6 dólares. Esta transacción se mueve en un solo sentido, es decir, si la misma baja a 6,5 dólares, el precio límite seguirá siendo de 6 dólares.

Este tipo de transacciones y métodos se utilizan en forma combinada para minimizar pérdidas o maximizar ganancias, y está dentro del conjunto de herramientas básicas de un inversor profesional o aficionado.

[3] http://sec.gov

2.1.2. Reportes trimestrales

Los mercados de valores exigen a las compañías que cotizan en ellos seguir ciertas normativas y reglas, establecidas por los diversos organismos reguladores. En el caso de los Estados Unidos, dicho organismo es la *Securities and Exchange Commission - SEC* [3] .

Entre las reglas impuestas a las compañías, se encuentra la obligatoriedad de publicar trimestralmente información financiera que debe incluir al menos el balance, los ingresos y el flujo de caja, pero que habitualmente también incluye un resumen ejecutivo, metas y objetivos futuros, la situación de la compañía y datos de interés para el público inversor. Estos estados financieros deben estar auditados por una firma independiente y las cifras que exponen suelen ser una imagen fiel del valor y las operaciones de la empresa.

Esto permite al mercado y a los inversores, el conocer periódicamente, el estado de la compañía y poder establecer el futuro de la misma, lo cual tienen un claro impacto en su cotización. Así, existen empresas que intentan predecir el resultado del reporte trimestral y cuando una empresa supera dichas predicciones, el precio de la acción suele tener un salto positivo, y cuando no alcanza los mismos, deriva en una reducción del precio de la acción y por ende del valor de la compañía.

Algunos de los datos trimestrales que se presentan, más conocidos y valorados por los inversores, son las ventas, los gastos, el ingreso neto, las ganancias por acción, las deudas y los activos. Los valores a informar son seleccionados basándose en regulaciones contables y tradiciones históricas y al ser valores absolutos en su gran mayoría, no es fácil comparar dichos valores entre distintas empresas, ya que su variabilidad es extremadamente alta, como podrá verse en la figura 3.2. Los datos más comunes de un balance trimestral pueden verse en la sección A.

2.1.3. Ratios Financieros

Para intentar estandarizar estos valores y volverlos comparables, desde el principio se intentaron generar diversos ratios de datos, por ejemplo, ingresos sobre ventas, o deudas netas sobre capital. Estos ratios facilitan el analizar la situación real de la empresa y compararlas con otras similares. Dada la cantidad de información disponible, cada analista desarrolló sus propios ratios, aunque un subconjunto de ellos son los usualmente aceptados por tradición y efectividad.

Existen varios trabajos que han analizado estos ratios y varios trabajos reducen este número a unos 40 ratios, financieros y contables, que permiten determinar las variables más importantes de cada balance. Sin embargo, varios de estos se superponen entre sí, por ejemplo *Deuda total/valor de la empresa*, *Activos totales/valor de la empresa* o *Deuda total/Activos totales* las cuales representan variantes de la ecuación *Activos totales = Deuda Total + Valor Neto*

O'Connor (1973) encontró que 10 de estos ratios tenían una diferencia estadísticamente significativa cuando se comparaban los mismos entre aquellas acciones que habían tenido un retorno alto en 5 años y un retorno bajo en el mismo período. Estos ratios se componían de la relación entre

- Deudas totales a Valor neto
- Capital de trabajo a Ventas

- Ventas a Activos Totales
- Ingresos disponibles para accionistas a Valor Neto
- (Ingresos por acción menos dividendos por acción) a Ingresos por acción
- Ingreso no operativo antes de impuestos a ventas
- (Ingresos netos antes de impuestos menos impuestos) a Ingresos netos antes de impuestos
- Flujo de caja a cantidad de acciones
- Deuda Corriente a Inventario
- Ingresos por acción a precio por acción

Chen and Shimerda (1981) investigó la representatividad de estos ratios y encontró, en combinación con trabajos precedentes, que un universo de entre 14 y 48 ratios puede ser reducido a través de componentes principales a entre 5 y 7 factores representando en general el 80 % o más de la variación.

Dichos factores representan las siguientes dimensiones

- Rentabilidad y Retorno de la Inversión: Involucrando datos de ventas, activos totales, ingresos antes de impuestos e intereses o ingresos netos, valor neto y activos totales.
- Actividad y la rotación de cuentas por cobrar y capital: Representadas principalmente por dos ratios, ventas a activos totales y ventas a activos rápidos, la primera con foco en la rotación del capital y la segunda en la rotación de las cuentas por cobrar.
- Liquidez, balance de activos y rotación del capital: En este grupo se encuentran ratios que miden la relación entre activos corrientes y activos totales o entre capital de trabajo neto y activos totales.
- Apalancamiento Financiero: Estos representan la relación entre deuda y capital o activos
- Posición de efectivo: Muestran la relación entre el efectivo y las ventas, activos o deudas
- Rotación de Inventario: Indicando la relación entre activos corrientes o inventario y ventas

Por último, el análisis de Wang and Lee (2008) sobre 24 ratios financieros en 4 categorías distintas, solvencia, rentabilidad, retorno sobre la inversión y rotación de activos y deuda, donde agruparon los mismos utilizando una relación difusa y en cada grupo, identificar los ratios más representativos. El resultado encontró 6 grupos de ratios y el indicador más representativo de cada uno de ellos, siendo los mismos los siguientes:

- Ratio de deuda - Solvencia: Deudas totales sobre activos totales
- Ratio de ganancia bruta - Rentabilidad: (Ingresos operacionales - Costo de la operación) sobre Ingresos operacionales

- Ratio de Ingresos antes de impuestos - Rentabilidad: Resultados antes de impuestos sobre Ingresos operacionales
- Retorno sobre capital de los accionistas - Retorno sobre la inversión: Resultado neto sobre el capital de los accionistas
- Rotación de los activos corrientes - Rotación de activos y deuda: Ingresos operacionales / Activos corrientes
- Rotación de todos los activos - Rotación de activos y deuda: Ingresos operacionales / Activos totales

Como consecuencia estos trabajos, es posible consolidar la información disponible en más de 80 datos financieros y contables en unos 10 ratios que reflejen el estado de la empresa, que permitan ver la evolución temporal de la misma y como se compara con otras empresas.

2.1.4. Precios históricos

Las series históricas de precios y volúmenes de acciones se encuentran disponibles en varios sitios y en general cuentan con los datos de precios de apertura, cierre, máximo y mínimo, además del volumen y eventualmente dividendos, para distintos períodos que van desde 1 minuto hasta 1 mes o más.

A lo largo del tiempo las acciones pueden sufrir cambios en su composición, siendo la más común los *splits* o *reverse splits* (divisiones o consolidaciones) en donde una acción se divide en una determinada cantidad para mantener el precio accesible a los inversores, o un determinado número de acciones se consolidan en una sola para mantener el precio por encima de 1 dólar, que suele ser una condición de los mercados para cotizar. En consecuencia, los precios netos puedan variar fuertemente, como sucedió por ejemplo con la acción de Amazon (AMZN) que en mayo de 2022 decidió dividir una acción en veinte nuevas, por lo que su cotización pasó de 2447 USD/acción a 125 USD/acción. Esto, que parece una caída de casi un 95 % del precio de la acción, representa, sin embargo, una alza de la cotización, y por ende de cualquier inversión, de un 1,02 % de un día a otro.

En general, los datos históricos disponibles ya están ajustados por cambios en la composición accionaria, de forma tal de no tener saltos abruptos debido a esto y que la evolución del precio sea comparable a lo largo de la historia.

2.2. Métodos de inversión

El objetivo de los diversos métodos de inversión es obtener un rendimiento favorable. En pos de esta meta se han desarrollado y se sigue innovando en diversos métodos que son utilizados tanto por inversores pequeños, intentando resguardar su capital, como por grandes inversores, tratando de generar una ganancia.

Describiremos a continuación tres métodos que utilizaremos en esta tesis y que reflejan estrategias de inversión reales.

2.2.1. Comprar y Mantener

Este método es conceptualmente el más sencillo de todos y utilizado en algún momento por todo inversor. En general se refiere a él por el nombre en inglés *Buy and Hold* y consiste en comprar un instrumento de inversión o activo para venderlo pasado cierto tiempo con la esperanza que su precio suba. El caso más común en Argentina probablemente sea el dólar estadounidense, que es comprado como refugio de ahorros bajo la premisa que el 'dólar siempre sube'. En la práctica, esta premisa ha causado pérdidas notables, dado que los instrumentos financieros no siempre tienen una curva ascendente, pudiendo reducir su precio considerablemente, siendo el mejor ejemplo el peso argentino.

2.2.2. Media Móvil de Convergencia/Divergencia

En búsqueda de métodos más sofisticados, en 1970 Appel and Dobson (2007) desarrollo un indicador que muestra cambios en fortaleza, dirección, momento y tendencia del precio de un activo. A este indicador lo llamó *Media Móvil de Convergencia/Divergencia* o más conocido por las siglas de su nombre en inglés *Moving Average Convergence Divergence (MACD).*

Este indicador se compone de tres series temporales calculadas a partir de los precios históricos del activo. Estas tres series son

- Línea MACD: que consiste en la diferencia entre una media móvil exponencial (EMA) de corto plazo y una media móvil exponencial de largo plazo (Ec 2.1)
- Línea de Señal: que consiste en la media móvil exponencial del MACD (Ec 2.2)
- Histograma: que consiste en la diferencia entre el MACD y la Señal (Ec 2.3)

Este indicador cuenta con tres parámetros y se denota *MACD(s,l,p)* donde s es el período de la media móvil exponencial de corto plazo, l es el período de la media móvil exponencial de largo plazo y p es el período de la media móvil que genera la señal. Es costumbre utilizar MACD(12,26,9) como valores estándar, y en general se calcula sobre los valores diarios de cierre.

$$macd(x) = EMA(x, s) - EMA(x, l) \tag{2.1}$$

$$signal(x) = EMA(macd(x), p) \tag{2.2}$$

$$divergence(x) = macd(x) - signal(x) \tag{2.3}$$

La idea de este indicador es que una EMA de corto plazo responde más rápidamente a fluctuaciones en el precio que una EMA de largo plazo y al comparar dichos medias móviles puede obtenerse indicadores de tendencia.

Dado que este indicador está basado en medias móviles, es un indicador retrasado y por ende puede ser lento para detectar tendencias súbitas, sin embargo, por su simplicidad es muy utilizado por inversores, aficionados y profesionales.

Este indicador señala oportunidades de compra y venta cuando las señales de MACD y la Señal se cruzan, si MACD alcanza a la Señal y la supera, es indicador de compra, si MACD es alcanzado y superado por la Señal, es oportunidad de venta como puede verse en la Figura 2.1.

Figura 2.1: Cotización del Euro contra el dólar estadounidense con los distintos componentes del MACD calculados y la localización de las señales de compra (flechas verdes) y venta (flechas rojas) - Gráfico tomado de Avantrade (http://avantrade.es).

2.2.3. Inversión intradiaria

En este método de inversión los activos no se conservan de un día a otro, la apertura y cierre de las operaciones deben ocurrir durante el día, a veces medidas en segundos o minutos, tratando de explotar cambios rápidos en el valor de un activo como consecuencia de una noticia o suceso. En general, estas estrategias de inversión se apoyan en el uso de órdenes de limitación de pérdidas o protección de ganancias llamadas *Stop Loss*, para impedir que una reversión rápida de la tendencia esperada genere pérdidas no deseadas. Estos inversores suelen operar con la compra o la venta en descubierto de activos y otros instrumentos financieros, especulando con una tendencia determinada. Por ejemplo, si creo que Apple (AAPL) tendrá un alza en el precio de su acción en los próximos minutos, compraré una determinada cantidad de acciones, especificando cuanto estoy dispuesto a perder y cuál es mi expectativa de ganancia. Si APPL cotiza a 100 USD, creo que va a subir a 101 USD y no quiero arriesgar más del 50 % de mi potencial ganancia, colocaré un Stop Loss en 99,50 USD. Si compro así 100 acciones de AAPL y estoy en lo cierto, mi ganancia será de 100 USD. Si estoy equivocado, mi pérdida máxima teórica será de 50 USD. Si bien este método no está blindado contra todas las condiciones, suele cubrir dichas pérdidas.

2.2.4. Pruebas históricas o backtesting

En general, para probar un método de inversión o predicción sobre los mercados financieros se suele realizar lo que se conoce como pruebas históricas o retrospectivas, o por su nombre en inglés *backtesting*. Esto se basa en correr los distintos métodos a probar con los datos históricos de un período temporal sin proporcionar al método el resultado del mismo, para ejecutarlos y luego compa-

rarlos. Esto es similar a lo que se hace en todo método de predicción, donde se reservan datos para la prueba y se comparan los resultados contra dichos valores reales.

2.3. Agrupamientos

El agrupamiento es la técnica de generar grupos de objetos basados en características similares, esto puede ser usado para diferentes tareas, entre ellas el aprendizaje automático y el reconocimiento de patrones. Existen diversos algoritmos que pueden asociar objetos, entre ellos los basados en conectividad, en centroides, en distribuciones o en densidad.

Para realizar el agrupamiento de las compañías utilizamos dos métodos relacionados entre sí basados en centroides, según contemos con coordenadas absolutas sobre las cuales pudiéramos calcular distancias euclidianas o sólo esté disponible la matriz de distancias. Para la primera opción utilizamos k-means, mientras que para la segunda utilizamos k-medoids. Dado que son algoritmos muy conocidos, haremos solo una breve descripción de los mismos, pero básicamente se basan en que cada grupo es representado por un vector central que puede o no coincidir con un elemento del grupo y al cual debe indicarse el número de grupos a generar como dato externo. Describiremos también OPTICS que está basado en una distribución de densidad ya que, fue utilizado aunque descartado debido a que deja observaciones sin asignar a un grupo determinado, aunque determina la cantidad óptima de grupos a diferencia de k-means y k-medioids que requieren de este dato como valor de entrada.

2.3.1. k-means

Este es un método, propuesto independientemente por Lloyd (1957) y por Forgy (1965) que busca distribuir las observaciones en k grupos, en los cuales la suma del cuadrado de las distancias euclidianas entre las observaciones se minimice para cada grupo. El método es un método iterativo que parte de k posiciones tomadas dentro del conjunto a agrupar o asigna las observaciones a distintos grupos al azar y usa los centroides como punto de partida. Con estos puntos el algoritmo busca generar grupos alrededor de ellos, para luego tomar los centroides de cada grupo como semilla para buscar nuevamente los grupos que minimicen dicha varianza, deteniéndose cuando las suma de cuadrados dentro de cada grupo no se altere más.

Formalmente, dada una serie de observaciones (x_1,...,x_n) donde cada observación es un vector real de dimensión d, k-means busca partir estas observaciones en k conjuntos S_1,...S_k, con $k \leq n$ tales que se minimice la suma de cuadrados dentro del grupo, definida como:

$$arg_S min \sum_{i=1}^{k} \sum_{x \in S} \|x - \mu_i\|^2 = arg_S min \sum_{i=1}^{k} |S_i| Var S_i \tag{2.4}$$

Uno de los inconvenientes de este método es que no garantiza encontrar la distribución óptima, sino que puede quedarse en un mínimo local, dependiendo de las posiciones iniciales. Sin embargo, es un método suficientemente sencillo y preciso para ser una buena alternativa.

2.3.2. k-medoids

Este algoritmo fue generado por Kaufman and Rousseeuw (1990). Utiliza un método de agrupamiento similar al de k-means, solo que k-medoids toma una observación como centro del cada grupo. Esta característica permite utilizar matrices de distancias entre puntos y no vectores, que provengan de medidas de disimilitud arbitrarias, mientras que k-means generalmente requiere distancias euclidianas. K-medoids requiere como k-means un valor k prefijado para generar los grupos, por ende la bondad de ese valor debe ser validada por otros métodos tales como Silhouette.

El medoide de cada grupo es definido como la observación cuya distancia promedio a todos los objetos es mínima. Tal como ocurre en k-means, esto puede llevar a soluciones no óptimas y es iterativo, ya que toma k observaciones al azar y asigna las restantes observaciones a cada grupo de forma tal que la disimilitud sea mínima, intercambiando observaciones en búsqueda de un estadio donde ya no pueda minimizarse ese costo.

2.3.3. OPTICS

Este algoritmo, propuesto por Ankerst et al. (1999) buscan encontrar grupos basados la densidad espacial del conjunto a agrupar. Para ello los ordena de forma tal que los puntos espacialmente más cercanos queden como vecinos en el ordenamiento. Almacena asimismo una distancia especial que representa la densidad que debe ser aceptada en un grupo para que dos observaciones pertenezcan al mismo grupo.

El algoritmo requiere dos parámetros ϵ que describe la distancia máxima a considerar y *Puntos Mínimos* que describe la cantidad mínima de puntos a considerar en un grupo. Un punto p será un punto núcleo si al menos la cantidad mínima de puntos indicada está dentro de su vecindario definido por ϵ, incluyendo al punto núcleo. Con esto busca las distancias del grupo al punto núcleo y entre los elementos del grupo.

El problema de este algoritmo es que puede haber puntos que no pertenezcan a un grupo dado que no alcanzan a llegar a la cantidad mínima de puntos requeridos y la distancia ϵ sea menor a la distancia a todos los puntos núcleos detectados, siendo un problema si estas observaciones huérfanas son un grupo grande dentro del conjunto de observaciones.

2.4. Medidas de similitud de series temporales

La comparación entre dos o más series temporales o secuencias de datos, tiene muchas formas posibles, dado que las mismas pueden ser analizadas desde el punto de vista temporal estricto, es decir, los valores en un momento t_i de una de las series contra el mismo momento en otra de las series, comparar las formas de las series, independientemente del marco temporal, comparar características de cada curva, etc. Esto puede ser realizado para la serie completa, o para porciones de cada serie, y por último, estas series pueden ser univariadas o multivariadas. Veremos algunas de las técnicas utilizadas para analizar la similitud.

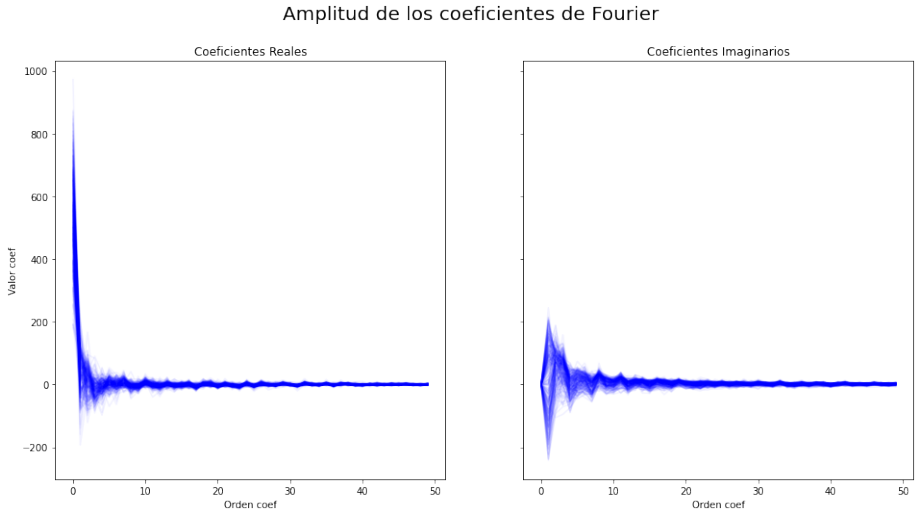

Figura 2.2: Amplitud de los distintos coeficientes obtenidos a través de una Transformación Discreta de Fourier para las 2359 curvas de precios, mostrando la variabilidad de los coeficientes para cada orden.

2.4.1. Transformada Discreta de Fourier

Si bien esta no es estrictamente una distancia, la combinación de la misma con la distancia euclidiana nos permite obtener similitudes entre series temporales. Uno de los primeros trabajos que propuso este mecanismo fue el realizado por Agrawal et al. (1993) donde utilizó una transformada discreta de Fourier (DFT), para extraer k coeficientes que representaran a la curva temporal y utilizando los más representativos generar árboles y encontrar similitudes. Demostró luego, utilizando el teorema de Parseval, que cuando la energía en el dominio temporal es la misma que la energía en el dominio de frecuencias, la distancia euclidiana entre dos señales x e y en el dominio temporal es la misma a la distancia en el dominio de frecuencias. Esto implica que podemos reducir fuertemente la dimensionalidad de las distintas curvas, e inclusive comparar curvas de distinto tamaño, tomando los valores más representativos obtenidos por la DFT. El otro punto importante es que los valores obtenidos a través de DFT son invariantes con traslaciones temporales, lo que permite obtener la similitud de curvas a pesar de que dicha similitud se encuentre en distintos momentos.

La lógica de utilizar este método de tomar solo aquellos coeficientes más relevantes, tiene sentido si consideramos que los dichos coeficientes modelan las variaciones más grandes de las curvas, y los coeficientes de orden mayor modelan las pequeñas variaciones que podemos definir como ruido.

Para el trabajo de exploración, donde nuestras curvas de precios de acciones contenían 1250 puntos, realizamos la DFT para cada curva y en la figura 2.2 podemos ver que efectivamente los coeficientes de orden menor son los que tienen mayor variabilidad y por ende poder de discriminador, más allá del orden 50 la variabilidad de los mismos no es significativa, como veremos luego nos permitió utilizarlo como una de las medidas de similitud.

Como desventaja de esta medida de similitud, la misma no permite encontrar similitudes entre curvas similares, pero donde una de ellas está comprimida o expandida temporalmente (ej. no detecta que *sen(x)* y *sen(2x)* son similares en forma, mientras que si detecta adecuadamente la similitud de curvas como *sen(x)*, *sen(x+b)* o *sen(x)+b*).

La complejidad de esta medida es *O(N.log(N))*.

2.4.2. Deformación Dinámica del Tiempo

En búsqueda de medidas de similitud que sean más apropiadas para resolver el problema de curvas similares pero, que no solo estén desplazadas en el tiempo sino también comprimidas o expandidas, se desarrolló el *Dynamic Time Warping o DTW* que intenta resolver este problema.

Las bases de esta medida fueron esbozadas por Vintsyuk (1968) y Sakoe and Chiba (1978) y formalizadas por Berndt and Clifford (1994) para resolver el problema de reconocimiento de habla, en especial cuando el lenguaje hablado puede ser hecho a distinta velocidad entre personas o inclusive con distinta duración de los fonemas individuales de acuerdo a las variantes locales o interlocutor. En estos casos, el algoritmo trata de encontrar un apareamiento óptimo entre dos secuencias considerando que,

- Cada punto en una secuencia tiene que estar apareado con uno o más puntos cercanos en la otra secuencia, lo mismo a la inversa.
- El primer punto en una secuencia tiene que coincidir con el primer punto en la otra, así como los últimos puntos deben coincidir.
- El mapeo debe ser ordenados monotónicamente, es decir, un punto en la secuencia no puede mapearse a un punto anterior en la otra secuencia comparado con el punto anterior en la secuencia original.

La combinación que satisface todas estas restricciones y tiene un valor mínimo, tomando como valor la suma de los valores absolutos de las diferencias entre puntos es la distancia propuesta.

Un detalle de esta medida de similitud es que no garantiza la desigualdad triangular y que la complejidad de la misma es de *O(NM)* donde N y M son los largos de las curvas.

2.4.3. Norma extendida de Frobernius

Uno de los problemas al medir distancias o similitudes entre secuencias multivariadas es que los métodos anteriores calculan la distancia entre los componentes respectivos de cada serie y agregan los mismos para obtener la similitud, esto deja de lado la relación entre los distintos componentes de una misma serie como información valiosa. Dado esto Yang and Shahabi (2004) proponen una medida de similitud que tome en cuenta dicha relación para obtener una distancia a la cual denominan *EROS* (Extended Frobenius Norm).

Este método aplica el análisis de componentes principales a una matriz generada a partir de las series multivaluadas y con ella obtener los componentes principales y los eigenvalores, los cuales son utilizados para comparar la similitud entre las distintas series.

Para obtener dicha distancia, dadas dos series **A** y **B** con tamaños m_A x n y m_B x n y dos vectores V_A y V_B las matrices derechas de eigenvectores que se obtienen al aplicar la descomposición en valores singulares sobre las matrices de covarianza M_A y M_B respectivamente. Suponiendo que $V_A = [a_1,...,a_n]$ y $V_B = [b_1,...,b_n]$ donde a_i y b_i son vectores ortonormales de tamaño n.

La similitud *EROS* entre **A** y **B** entonces está definida por:

$$EROS(\mathbf{A}, \mathbf{B}, w) = \sum_{t=1}^{n} w_i | < a_i, b_i > | \tag{2.5}$$

donde ¡a_i,b_i¿es el producto interno de a_i y b_i y w es un vector de pesos basado en los eigenvalores de las series de datos basado en el promedio y normalización de los mismos para que la suma de todos los pesos sea igual a 1 y mayor o igual a 0.

Con esto es posible obtener una matriz de distancias entre las distintas series multivaluadas.

Esta medida de similitud tampoco respeta la desigualdad triangular pero es útil para explorar otra característica de las series multivaluadas, que es la relación entre las distintas series univaluadas que la componen.

2.5. Predicción

Para realizar la predicción de valores no observados usamos dos métodos diferentes, ARIMA y Redes Neuronales, haremos solo una breve introducción a los mismos dado que son de uso común en el área de Minería de Datos y en particular con el procesamiento de series temporales.

2.5.1. ARIMA

Este método, utilizado especialmente en el análisis de series temporales, permite predecir en ciertas condiciones los puntos futuros de la serie. El mismo se compone de tres partes que forman su nombre por la abreviatura en inglés:

- AR - Componente Autorregresivo: Usa los valores pasados en la ecuación regresiva para la serie temporal. El parámetro p determina la cantidad de periodos u orden del componente.
- I - Integración: Es el grado de diferenciación, es decir, la cantidad de diferencias usadas para convertir la serie en estacionaria. El parámetro d representa la cantidad de diferencias usadas.
- MA - Medias Móviles: Consiste en el error del modelo como una combinación de los términos de error ϵ_t anteriores. El orden q representa la cantidad de términos a incluir en el modelo.

Esta combinación genera un modelo ARIMA(p,d,q) no estacional donde el componente $\hat{y_t}$ se puede escribir como la siguiente combinación lineal,

$$\hat{y_t} = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 \epsilon_{t-1} - \ldots \theta_q \epsilon_{t-q} \tag{2.6}$$

donde y_t es la serie diferenciada d veces, ϕ son los coeficientes de los p términos de la autorregresión y ϵ son los q términos de los errores de las observaciones pasadas.

Obviamente, distintos modelos tendrán diferentes capacidades de ajuste. Para ver las bondades de cada uno de ellos y así determinar los parámetros óptimos p,d,q se suele recurrir a un criterio de información. Existen varias alternativas, como el criterio de Akaike (1974), el de Hannan and Quinn (1979), o bien el criterio Bayesiano de Schwarz (1978). Entre ellos, el Criterio de Información de Akaike es uno de los más utilizados:

$$AIC = 2k - 2ln(L) \tag{2.7}$$

donde L es el máximo valor de la función de verosimilitud para el modelo estimado y k la cantidad de parámetros del modelo estadístico (para ARIMA sería equivalente a $p+q+d$). Este índice penaliza los modelos más complejos y el mejor modelo será aquel que minimice el valor de AIC..

2.5.2. Redes Neuronales

El trabajo inicial de Redes Neuronales fue realizado por McCulloch and Pitts (1943) en los años 40, y desde allí se ha desarrollado un área de investigación y desarrollo extremadamente fructífero.

Una Red Neuronal (RN) está basada en un conjunto de nodos llamadas neuronas artificiales. Una neurona recibe señales y luego de procesar las mismas puede enviar señales a otras neuronas conectadas con ellas. Estas señales son números reales y en general existe una función no lineal que procesa las mismas. Las neuronas y conexiones suelen tener un peso asociado a cada una de ellas que se va ajustando a medida que progresa el procedimiento de aprendizaje.

Estas neuronas están agregadas en capas y la capa inicial suele tomar la entrada de datos para generar en la capa final una salida que es comparada contra la realidad y es parte del ajuste de los pesos. Varios ciclos iterativos causarán la salida a parecerse cada vez más a la realidad hasta que el proceso es terminado mediante alguna condición de finalización (ejemplo, el error es menor a un número previamente especificado o se ejecutaron una determinada cantidad de ciclos)

Dado que los pesos iniciales suelen ser generados al azar, dos modelos similares pueden obtener resultados distintos, ya que pueden terminar en mínimos locales, y el proceso no garantiza encontrar el mínimo óptimo.

A pesar de esto, una de las ventajas de las Redes Neuronales es la capacidad de aprender estructuras no lineales que la vuelven extremadamente versátil para poder predecir y separar distintas estructuras de datos, e inclusive tener la capacidad para utilizar gran cantidad de datos y aprovechar en forma eficiente, aquellos más relevantes.

Esto ha dado lugar a numerosos tipos de Redes Neuronales, desde las más simples consistiendo en neuronas que toman una o más señales y generan una señal de salida, a otras que pueden recordar y aprovechar señales pasadas. Este último tipo de neuronas son extremadamente útiles para procesar series temporales, ya que pueden utilizar entradas o resultados pasados para procesar la secuencia y así también aprender la relación de un punto temporal con puntos temporales pasados.

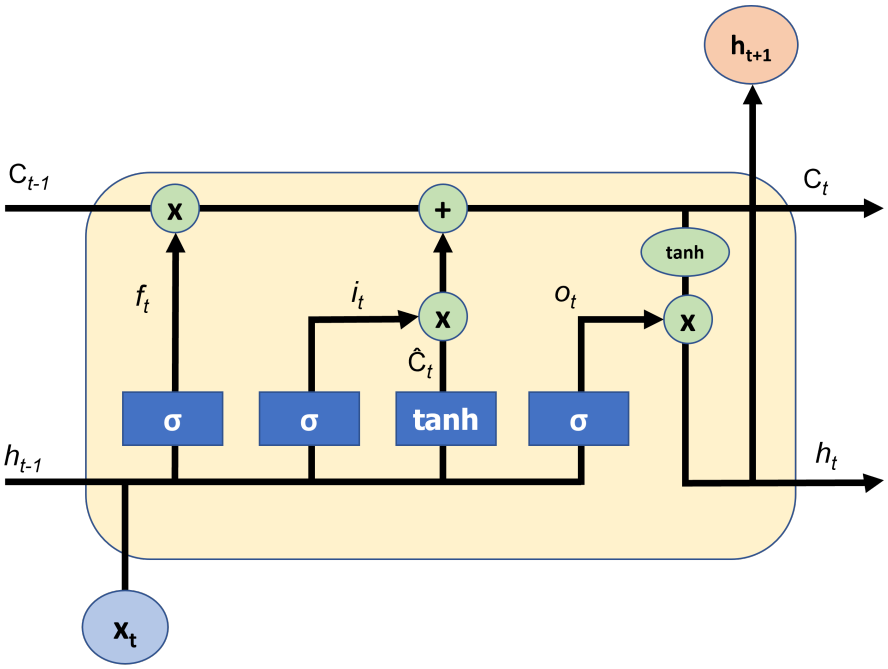

Figura 2.3: Estructura de una celda o neurona LSTM

Veremos a continuación dos tipos de capas de Redes Neuronales en particular que utilizamos en esta tesis y que resultan de interés para el desarrollo de la misma.

Capas LSTM

Uno de los tipos de capas que utilizan para generar modelos de predicción de series temporales es el llamado LSTM (Long Short Term Memory) propuesto inicialmente por Hochreiter and Schmidhuber (1997). Su nombre proviene de la capacidad de utilizar datos pasados y actuales para generar una salida (*long term memory* y *short term memory*)

Una neurona típica contendrá una puerta de entrada, una puerta de salida y una puerta de olvido y pueden recordar valores a lo largo de intervalos arbitrarios de tiempo y regular el flujo de información de entrada y salida utilizando las tres puertas.

Las Redes Neuronales compuestas por capas de este tipo se conocen como Redes Neuronales recurrentes y tienen un uso muy difundido en reconocimiento de voz, traducción automática, control robótico, juegos, salud, mercados y otros tipos de estructuras temporales o secuenciales de datos.

El funcionamiento de la celda mostrada en la figura 2.3 es el siguiente:

El primer paso es decidir que información vamos a descartar del estado de la celda, usando la primera puerta, llamada de olvido, donde combinamos el estado previo oculto h_{t-1} y a la entrada x_t y a través de una sigmoidea define si mantiene el dato o lo olvida, generando la salida f_t como puede verse en la ecuación 2.8.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.8}$$

El siguiente paso es decidir que información vamos a conservar en el estado de la celda y esto se hace combinando dos fuentes, primero una puerta de entrada usa una sigmoidea para decidir que valores vamos a actualizar con un valor i_t como puede verse en la ecuación 2.9, y por otro lado, usamos una tangente hiperbólica para generar un vector de valores candidatos $\tilde{C}_t$ que puede ser agregado al estado, como indica la ecuación 2.10.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.9}$$

$$\tilde{C}_t = tanhW_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.10}$$

Luego vemos como cambiar el estado anterior de la celda, C_{t-1} en un nuevo estado C_t, utilizando los resultados anteriores. Para ello multiplicamos el estado anterior por el resultado de la puerta de olvido f_t y a esto le añadimos el resultado de $i_t * \tilde{C}_t$, como puede verse en la ecuación 2.11.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.11}$$

Por último debemos calcular la salida de la celda, basada en este nuevo estado C_t tomando también como datos la salida anterior h_{t-1}, la entrada a la celda x_t, mediante las siguientes ecuaciones 2.12 y 2.13.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.12}$$

$$h_t = o_t * tanh(C_t) \tag{2.13}$$

Capas convolucionales

Este tipo de capas tienen la habilidad de reconocer estructuras y simplificar al mismo tiempo las secuencias de datos a procesar. De la misma forma que una capa clásica, reciben datos de entrada, lo procesan incluyendo eventualmente una función no lineal y generan una salida que alimenta otras capas de neuronas.

¿Qué las diferencia de una capa tradicional? Básicamente, la suposición que la entrada representa estructuras de algún tipo conocido, para imágenes se suelen usar convoluciones de 2 dimensiones, para series temporales, convoluciones de 1 dimensión. La otra ventaja es que permite reducir drásticamente la cantidad de pesos w_i que contiene el modelo.

Los parámetros de una capa convolucional se componen de un conjunto de filtros que pueden aprender las características de una secuencia determinada. Cada filtro es pequeño, espacialmente hablando, así por ejemplo la dimensión de un filtro de 2 dimensiones puede ser de 5x5 elementos con una profundidad variable, que depende de la entrada, en el caso de un filtro de 1 dimensión, puede tener 5 o 7 elementos de largo. A medida que vamos deslizando el filtro por la secuencia de entrada, se va activando de una forma distinta en cada posición y la red utiliza esta salida para aprender estructuras dentro de la secuencia. Las capas pueden tener varios de estos filtros en paralelo y a medida que el modelo va aprendiendo, cada uno de ellos se especializa en distintas estructuras.

La razón de la versatilidad de estas capas en reconocer estructuras locales es que se fijan en regiones adyacentes y no, como en una red totalmente conectada, la relación entre puntos distantes entre sí. Para el trabajo realizado en esta tesis,

como luego veremos, tiene sentido relacionar los puntos cercanos entre sí, dado que son los que probablemente conllevan mayor información e influencia.

2.6. Evaluación de resultados

Las siguientes son algunas de las métricas que hemos utilizado para evaluar la bondad de los agrupamientos o predicciones.

2.6.1. Agrupamientos

Una forma de evaluar si los agrupamientos obtenidos por distintos métodos son similares o tienen estructuras distintas es mediante la medida de Rand, propuesto por Rand (1971), la cual está definida de la siguiente forma: Dado un conjunto $S = s_1, ...s_n$ y dos particiones de S para comparar $X = X_1, ...X_r$ una partición de S en r subconjuntos e $Y = Y_1, ...Y_s$ una partición de S en s subconjuntos y,

- a, el número de pares de elementos de S que se encuentran en el mismo subconjunto en X y el mismo subconjunto en Y.
- b, el número de pares de elementos de S que se encuentran en subconjuntos distintos en X y en subconjuntos distintos en Y.
- c, el número de pares de elementos de S que se encuentran en el mismo subconjunto en X y en subconjuntos diferentes en Y.
- d, el número de pares de elementos de S que se encuentran en subconjuntos diferentes en X y en el mismo subconjunto en Y.

Entonces la medida de Rand es:

$$R = \frac{a+b}{a+b+c+d} \tag{2.14}$$

Donde R se encuentra entre 0 y 1, donde 1 es una similitud perfecta y 0 es una distribución disímil o al azar.

Una mejora a esta medida está dada por el ajuste por azar propuesto por Hubert and Arabie (1985) en donde dicha corrección establece una base usando la similitud esperada para todas las comparaciones de a pares entre los grupos, especificadas por un modelo estocástico. Este valor ajustado se encuentra entre -0,5 y 1, donde similarmente 1 define la similitud perfecta, 0 una distribución al azar o distinta y valores negativos implican una distribución discordante entre ambas distribuciones.

2.6.2. Pronósticos

Para comparar los resultados obtenidos por los diferentes métodos de pronóstico, es una práctica común usar alguna métrica para evaluar qué tan bien se ajusta cada método a los resultados reales. El error cuadrático medio (MSE) y el error absoluto medio (MAE) pueden ser difíciles de interpretar dada su relación directa con la escala de los valores que se predicen, y que diferentes

acciones pueden tener escalas de precios muy diferentes. Por lo tanto, utilizamos dos métricas que superan estas limitaciones: (i) el error de escala absoluto medio (MASE) y (ii) el error porcentual absoluto medio simétrico (sMAPE). MASE, definido en la Ecuación 2.15 es una medida de error sin escala que compara el error con el error promedio. sMAPE, definida en la Ecuación 2.16, es una medida que presenta los errores en términos de porcentajes (Goodwin and Lawton, 1999).

$$MASE = \frac{1}{k} \frac{\sum_{t=1}^{k} |Y_t - \widehat{Y_t}|}{\frac{1}{n-m} \sum_{t=m+1}^{n} |Y_t - Y_{t-m}|} \tag{2.15}$$

$$sMAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{2 * |y_i - \widehat{y_i}|}{|y_i| - |\widehat{y_i}|} \tag{2.16}$$

MASE compara el error de la serie pronosticada con una predicción ingenua. En nuestro caso, esta predicción es simplemente el valor de cierre del día anterior. Esta métrica indica cuánto mejor o peor es nuestra predicción en comparación con la ingenua, con valores por debajo de 1 que muestran una mejor predicción que el caso ingenuo. sMAPE, por otro lado, indica el porcentaje de desviación del valor ajustado de los datos reales, por lo que el punto de referencia no es definido por el investigador como con MASE.

2.7. Antecedentes

En esta sección, revisaremos artículos relacionados centrándonos en tres aspectos: métodos de agrupamiento para acciones, modelos generales de predicción y modelos combinados de agrupamiento/predicción.

2.7.1. Agrupamiento de datos de mercado

En la búsqueda de fuentes de información que no sean tan obvias y que permitan mejorar los modelos de predicción de precios o tendencias de mercado, varios investigadores han explorado la agrupación de datos de las distintas bolsas de valores y de las empresas que cotizan en ellas, para descubrir y obtener asociaciones que puedan aportar información adicional a los diferentes modelos predictivos.

La información adicional explorada incluye datos fundamentales de cada empresa, ya sea datos estáticos como los empleados por Dhakal (2019) o una serie temporal como los que usan Nair et al. (2017) y Basalto et al. (2005). Además, se pueden emplear informes financieros trimestrales como lo hizo Lee et al. (2010). Otras fuentes de datos alternativas incluyen opiniones de foros y noticias sobre las empresas, según lo investigado por Li and Wu (2021).

Los primeros agrupamientos se basaron en los indicadores generados por calificadoras de riesgo, como Standard and Poor, Moody's o Fitch Rating. Todavía hoy, estas agencias clasifican a las empresas que cotizan en las bolsas de valores de acuerdo con diferentes criterios, que incluyen evaluaciones de riesgo, sectores, tipos de empresas y otros indicadores que pueden utilizarse para catalogar las empresas.

Entre los trabajos realizados en los últimos años que utilizan el agrupamiento de diferentes tipos de datos estáticos para obtener correlaciones o datos no tan

obvios, podemos encontrar a Dhakal (2019) que optó por agrupar las empresas utilizando los datos estáticos obtenidos de los balances y para encontrar si el crecimiento del precio de las acciones es significativo en cualquiera de los grupos encontrados. Babu et al. (2012) utiliza un enfoque interesante, convirtiendo los balances en vectores de características y agrupando los mismos para determinar la variación de corto plazo del precio de la acción, inmediatamente después que el balance es publicado y examina las propiedades de los distintos métodos de agrupamiento en relación con estos vectores de características.

Aunque el agrupamiento de datos estáticos es un tema ampliamente explorado donde existen técnicas de agrupación y reducción de dimensionalidad muy eficientes, el agrupamiento de series temporales no está tan maduro, especialmente con series de datos multivariados. El resultado de estos agrupamientos indica, especialmente para series temporales, una contribución favorable de los datos para realizar predicciones de tendencias y precios, tal y como analizan Nair et al. (2017), Lee et al. (2010) o Wang et al. (2021).

Existen varios métodos para agrupar series temporales, ya sea por similitud de forma, por extracción de componentes fundamentales o por comparación punto a punto, como describen en sus trabajos Bagnall et al. (2017) o Aghabozorgi et al. (2015). En particular, mencionamos el uso de Extended Frobenius Norm (EROS) para agrupar series de tiempo multivariadas como lo propone Yang and Shahabi (2004), ya que usaremos este método para agrupar la serie de datos financieros. Este método usa PCA para reducir la dimensionalidad de la serie de tiempo multivariada, y luego usa la norma de Frobenius para comparar series de tiempo, logrando mejores resultados que simplemente comparando la serie multivariada original con la distancia euclidiana.

2.7.2. Modelos de predicción

Un escenario similar se encuentra para los modelos de predicción de tendencias o precios de acciones, ya sea con series temporales o datos estáticos, que suelen referirse a los últimos datos de balance en comparación con los resultados observados en otras empresas similares. Entre los modelos utilizados se encuentran métodos estadísticos tanto univariados como multivariados como VARIMA o Prophet Taylor and Letham (2018), combinaciones de los mismos Yenidoğan et al. (2018); Samal et al. (2019) Fang et al. (2019) o métodos basados en gradientes, árboles de decisión y Redes Neuronales Krauss et al. (2017).

El uso de Redes Neuronales para predecir el precio y las tendencias de las acciones en las bolsas de valores es un tema sobre el cual existe abundante literatura con diferentes niveles de predicción considerando la naturaleza caótica del problema. Así encontramos Redes Neuronales que combinan capas convolucionales y recurrentes, generativas antagónicas, propuestas por Liu and Motani (2021), o algoritmos genéticos realizados por Hadavandi et al. (2010) para refinar las predicciones. Existe un debate sobre si los métodos estadísticos son superiores a las Redes Neuronales en términos de precisión de predicción. Algunos investigadores, como Yamak et al. (2019) con trabajos realizados en Bitcoin, ocupan esta posición. Otros lo refutan, como Siami-Namini et al. (2018) con varios índices financieros, Adebiyi et al. (2014) o Ma (2020) ambos utilizando el precio de las acciones de Dell.

2.7.3. Modelos de predicción con agrupamiento

Varios autores han trabajado en la combinación de agrupamiento con varios métodos de predicción. Por un lado, Hadavandi et al. (2010) propuso agrupar los datos de la serie temporal de acciones utilizando mapas autoorganizados y luego alimentar estos datos en una red genética, con resultados favorables en términos de precisión de las predicciones en comparación con otros métodos. Por otro lado, podemos mencionar a Wang et al. (2021), quienes propusieron agrupar las series temporales de precios utilizando la distancia de similitud morfológica y k-medias para encontrar diferentes tipos de series. Posteriormente, entrenan diferentes Redes Neuronales para cada tipo de serie. Otros autores usan el agrupamiento para crear características, utilizando métodos de agrupamiento basados en la distancia como DTW o métodos de agrupamiento de características que utilizan características de la serie temporal como entropía, estabilidad, autocorrelación de coeficientes de transformada rápida de Fourier o coeficientes de transformada de ondas continuas para alimentar esta información junto con la serie temporal para una red neuronal como la realizada por Tadayon and Iwashita (2020), Bandara et al. (2020) o Li et al. (2022). Por último, Lin et al. (2022) combina datos del mercado y de sentimiento para la predicción de la tendencia del mercado y Long et al. (2020) utiliza los registros de transacciones del mercado y la información pública de las acciones para identificar y agrupar estructuras de inversión y añadir esta información a los datos de precios y volúmenes y predecir a través de una red neuronal basada en capas convolucionales y LSTM bidireccionales para predecir el movimiento de datos. En estos casos, los autores encuentran que las características agregadas por el agrupamiento mejoran las predicciones realizadas por las Redes Neuronales en comparación con los modelos sin agrupamiento.

Es sobre la base de estos trabajos, que muestran que los distintos mecanismos de predicción no solo se benefician de una selección previa de tipos de datos o características, como es de conocimiento público en el análisis de datos, sino también de la cantidad y calidad de los datos entregados, que postulamos la idea que, un agrupamiento previo para seleccionar aquellos datos que resulten relevantes, debería ayudar a la predicción de valores y tendencias en el mercado de acciones. Si esta idea fuera cierta debería mejorarse el resultado final, medido como un rendimiento positivo.

Nótese que aunque el agrupamiento de series de tiempo basado en precios se ha intentado anteriormente en varias ocasiones, el agrupamiento de series de tiempo basado en reportes financieros trimestrales no es tan frecuente y fue analizado de manera limitada por Tupe (2014) y Marvin (2015), entre otros. Por lo tanto, ampliamos este enfoque utilizando un conjunto de medidas de distancia más representativas para series de tiempo multivariadas.

En el próximo capítulo describimos las características principales y fuentes de los conjuntos de datos utilizados y algunas transformaciones que hemos realizado para facilitar su procesamiento.

Capítulo 3

Conjuntos de Datos

En esta sección, describimos el conjunto de datos utilizado en nuestros experimentos, las características y transformaciones utilizadas y las fuentes de los mismos.

3.1. Definición de las empresas

Para obtener el listado de empresas que utilizamos en esta tesis, el primer paso fue seleccionar que mercados de valores a utilizar. Tal como hemos mencionado en el capítulo 2.1, tomamos los dos mayores de los Estados Unidos, el NSYE y NASDAQ, donde en conjunto cotizan aproximadamente unas 6000 empresas con un valor conjunto de 50 billones de dólares. Una característica notable de la valuación de mercado de estas empresas, es que un porcentaje pequeño de ellas constituye el 80 % del valor total de los mercados, dado que la mitad de las empresas tiene valuaciones por debajo de los 100 millones de dólares, como puede observarse en la figura 3.1. Del universo de empresas cotizantes, tomamos aquellas que, en conjunto, representaran mejor el mercado y tuvieran cierta estabilidad. Por este motivo descartamos empresas de valuación baja, dado que la variabilidad de la cotización de estas es muy grande y responde en numerosas ocasiones a eventos impredecibles tales como noticias o tendencias de mercado o incluso a manipulaciones de inversores. Las empresas medianas y grandes tienen un comportamiento menos sensible a estos hechos y cuentan con una inercia mayor. Como punto de partida utilizamos el índice Russell 3000[1] . Este índice contiene aproximadamente unas 3000 acciones con la mayor capitalización de mercado y que representan en forma conjunta entre el 96 % y el 98 % del valor de los mercados de acciones de Estados Unidos. Esta lista es dinámica y va variando en el tiempo de acuerdo a los cambios en la capitalización de las distintas empresas. Debido a esta dinámica tomamos arbitrariamente la lista que componía el índice al 1^o de marzo de 2021, la cual contaba en ese momento con 3057 compañías.

Con esta lista de empresas y para obtener una serie continua suficientemente extensa, seleccionamos aquellas compañías que hubieran cotizado y presentado balances entre el 1^o de enero de 2017 y el 31 de marzo de 2022. Esta condición dejó fuera a compañías nuevas, sin suficiente historia, aquellas que fueron adquiridas o fusionadas con otra compañía o aquellas que dejaron de cotizar por

[1] https://www.ftserussell.com/russell-indexes-your-index-matters/russell-3000

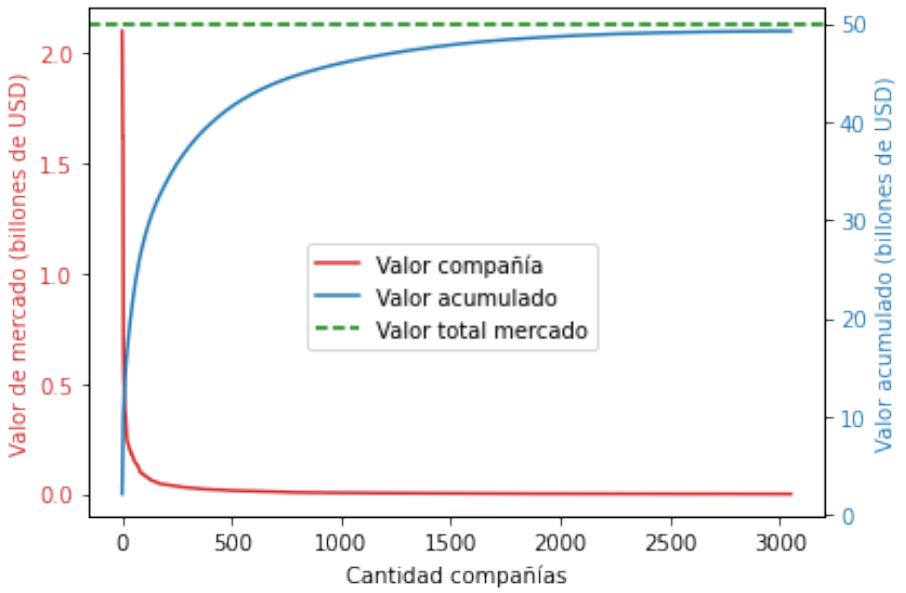

Figura 3.1: Valor de mercado y valor acumulado de las mayores 3000 compañías de NASDAQ y NSYE, ordenadas por valor de mercado.

cualquier razón, siendo las más comunes la quiebra o el retiro de la cotización pública. Luego de estas restricciones, nuestro conjunto de prueba quedó con 2359 acciones, las cuales contaban con al menos 19 balances trimestrales y el resultado de las cotizaciones para las 1315 sesiones de mercado en esos 5 años y 3 meses.

3.2. Fuentes de datos

En general los datos históricos de precios están ampliamente disponibles en varias fuentes, en nuestro caso utilizamos los servicios de Yahoo Finance[2] a través de la librería de Python *yfinance*[3] dado que esta ofrece una forma sencilla de descargar los datos. Un punto que es dable resaltar, es que estas series de precios ya están ajustadas por divisiones o consolidaciones de acciones, tal como fuera descrito en la sección 2.1.4, simplificando la tarea del usuario en ajustar los mismos manualmente.

Los reportes trimestrales en general no son tan accesibles. El último reporte trimestral puede accederse libremente en una cantidad de sitios, pero el acceso a los datos históricos suele ser un servicio pago. Para obtener los reportes trimestrales utilizamos los servicios de Alpha Vantage[4] a través de la API pública que poseen[5] y de la librería de Python *alpha-vantage*[6] dado que ofrecieron en un momento la posibilidad de obtener los datos a un costo bajo.

[2] https://finance.yahoo.com/
[3] https://pypi.org/project/yfinance/
[4] https://www.alphavantage.co/
[5] https://www.alphavantage.co/documentation/
[6] https://pypi.org/project/alpha-vantage/

3.3. Conjuntos de Datos

Con estas fuentes generamos diversos conjuntos de datos, los cuales fueron utilizados en distintas actividades, a saber:

- Balances (BAL): Serie de balances trimestrales entre el 1/ene/2017 y 31/dic/2021.
- Ratios (RAT): Serie de 10 ratios financieros mensuales entre el 1/ene/2017 y 31/dic/2021. Dependiendo de las fechas de los primeros y últimos balances, no necesariamente tenían completos los extremos, enero y febrero de 2017 y noviembre y diciembre de 2021, dependiendo del mes de publicación del balance trimestral.
- Agrupamiento (AGR): Serie de precios diarios suavizados a través de una media móvil - 3/ene/2017 a 31/dic/2021 para las todas las acciones. De estas series se derivan los rendimientos diarios.
- Pronostico e inversión piloto con datos diarios (PD): Serie de precios diarios - 1/jul/2019 a 30/jun/2022 para el conjunto utilizado en el piloto. De estas series se derivan los rendimientos diarios.
- Pronostico e inversión (CD): Serie de precios diarios - 3/ene/2017 a 31/mar/2022 para todas las acciones.

Decidimos comenzar la tarea con un piloto, considerando que el volumen de datos existente podría plantear problemas de performance en la ejecución de las distintas rutinas de agrupamiento y predicción, y para agilizar el desarrollo y exploración de los distintos métodos. El mismo se compone de un 11 % del conjunto total intentando que fuera suficientemente representativo del conjunto en general. En consecuencia, generamos un conjunto de 260 empresas y de esta forma acelerar el desarrollo, exploración y calibración de los algoritmos a usar. A este conjunto agregamos las dos acciones de Google (GOOG y GOOGL), Visa (V) y Mastercard (MA) como acciones testigo, ya que ambos pares son muy similares entre sí y nos servirán para tener una verificación básica del funcionamiento de los algoritmos de agrupamiento con un total de 264 acciones. En consiguiente generamos subconjuntos de BAL, RAT y AGR y adicionalmente generamos PD para realizar pruebas adicionales a las planificadas anteriormente.

Una vez concluida esta fase, y con los algoritmos depurados, realizamos la prueba sobre el total de datos obtenido con los conjuntos BAT, RAT, AGR y CD.

3.4. Transformación de datos

Para poder utilizar los datos obtenidos de fuentes externas realizamos algunas transformaciones sobre ellos que permitieron obtener los distintos conjuntos descritos en la sección anterior.

3.4.1. Series de precios

Con las series de precios diarios de Yahoo, generamos los conjuntos PD y CD, conteniendo el precio de apertura, el de cierre, el máximo y mínimo diario y el volumen operado dentro del horario habitual de operación (9:30 hs a 16 hs).

Dado que las series de precios ya están ajustadas por las divisiones o consolidaciones de acciones y para evitar oscilaciones de alta frecuencia, suavizamos ligeramente la serie con un promedio móvil durante un período de 5 días para el conjunto de series de precios que usamos en el agrupamiento. Con estos datos transformados generamos el conjunto AGR. Esto facilita la comparación de las series temporales de precios de cierre.

3.4.2. Balances trimestrales y series temporales derivadas

Para la obtención de las series históricas de balances consolidamos los datos obtenidos a través de las API de AlphaVantage del Estado de Resultados *(Income Statement)*, de la Hoja de Balance *(Balance Sheet)* y del Flujo de Caja *(Cash Flow)*, obteniendo casi de 90 valores descriptivos de cada balance trimestral y anual, los cuales están presentados en el Apéndice A -Datos de Reportes Trimestrales A. Con esto generamos el conjunto de datos BAL.

En general, estos balances se presentan cada tres meses, aunque en ocasiones las compañías deciden alterar su cronograma de presentación o corregir algún balance anterior, por lo que utilizamos el último balance válido para un determinado trimestre.

Una de las características de estos balances es la dispersión de valores, ya que no es lo mismo una empresa con una valuación mayor a 2 billones de dólares que una de 10 millones de dólares, para la mayor parte de los valores planteados, como puede verse en la figura 3.2. En algunos casos hay valores negativos, los cuales son legítimos como Ganancia Total, Ingreso Neto o Flujo de Caja y no constituyen outliers sino un reflejo de una situación económica de la empresa.

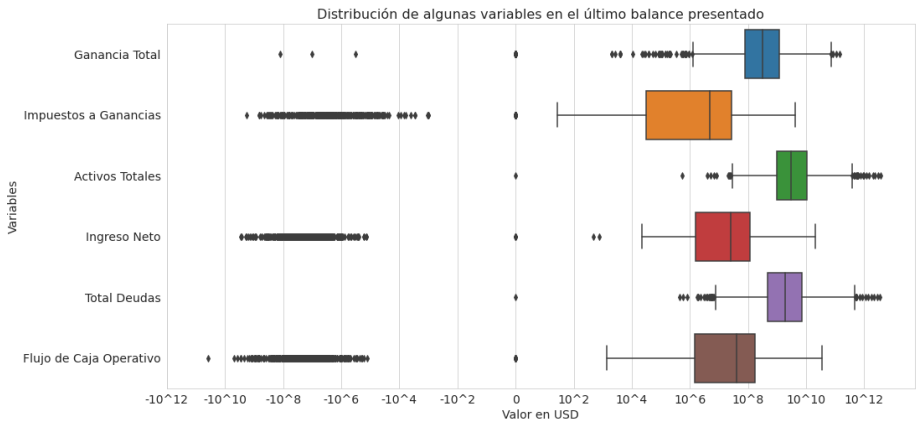

Figura 3.2: Ejemplo de distribución de los valores de seis variables representativas del balance del 4to trimestre de 2021 y utilizadas en el cálculo de ratios.

Para reducir la variabilidad de los valores, el primer enfoque fue utilizar una escala logarítmica modificada donde poder representar tanto valores negativos como positivos. Este enfoque fue el utilizado en el trabajo de especialización, pero fallaba al reconocer que existen relaciones entre los distintos datos que revelan mejor el estado de una compañía que el valor absoluto o modificado de los balances.

Por ende decidimos transformar estos datos en ratios que fueran representativos de los balances y que permitieran una mejor comparación, basándonos en los trabajos expuestos en 2.1.3 y elegimos diez ratios que cumplían estas condiciones, los cuales pueden verse en la Tabla 3.1.

En consecuencia, obtuvimos series de tiempo trimestrales de 10 ratios para los cinco años entre el 1^o de enero de 2017 y el 31 de diciembre de 2021. Sin embargo, dado que los cierres trimestrales de las empresas no necesariamente coinciden, transformamos estas series en series mensuales, interpolando los puntos faltantes mediante splines cúbicos, como puede verse en el ejemplo de la figura 3.3. Esto nos permitió generar el conjunto de datos RAT, que será utilizado para las actividades de agrupamiento. Debido al cronograma de presentación de balances, existen series que comienzan en enero de 2017, terminando en octubre de 2021, otras de febrero de 2017 a noviembre de 2021 y por último otras entre marzo de 2017 y diciembre de 2021.

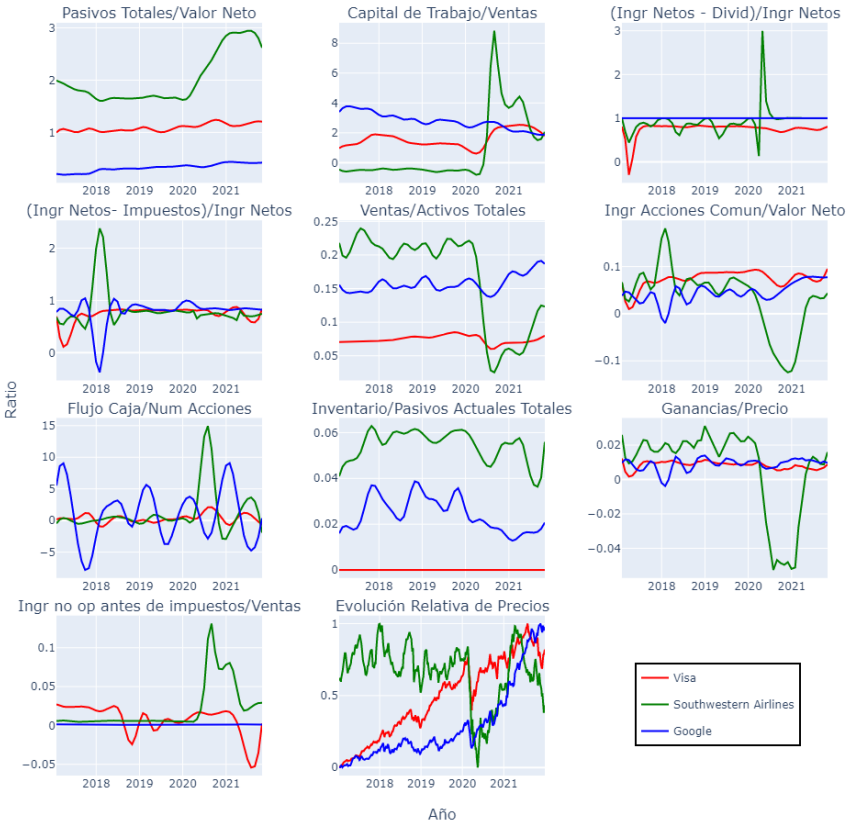

Figura 3.3: Ejemplo de los diez ratios usados para agrupar reportes trimestrales y la evolución relativa de precios para tres compañías de la muestra (Visa, Southwest Airlines and Google) durante el período 2017-2021.

Para ver la variabilidad de los ratios, presentamos en la tabla 3.2 las características de algunos de ellos.

Con una porción de estos datos, en el próximo capítulo desarrollaremos la etapa de exploración, donde ensayamos distintos métodos de agrupación, pronóstico e inversión. El objetivo de esta etapa fue medir y optimizar los algoritmos utilizados, sometiendo luego los mismos a pruebas adicionales para verificar su eficacia y preparar los algoritmos finales que correrían con el conjunto de datos completos.

Código	Ratio	Descripción
DT/VN	$\frac{DeudasTotales}{ValorNeto}$	Este es un indicador de la salud financiera de la compañía y permite ver si la empresa puede pagar sus deudas en caso de que la situación empeore, donde el Valor Neto se calcula como los Activos Totales menos las Deudas Totales.
CT/V	$\frac{CapitaldeTrabajo}{Ventas}$	Esta es una métrica usada para determinar que tan eficientemente una empresa utiliza su Capital de Trabajo para generar sus Ventas. El Capital de Trabajo se calcula como los Activos Corrientes Totales menos las Deudas Totales.
V/AT	$\frac{Ventas}{ActivosTotales}$	Esta métrica mide la eficiencia del uso de los activos de una compañía para generar Ventas. En general, un valor alto de esta métrica indica una empresa eficiente en el uso de sus activos.
IA/VN	$\frac{IngresosAccionistas}{ValorNeto}$	Esta métrica mide el ratio entre los ingresos disponibles para los inversionistas comunes contra el Valor Neto de la empresa. Una empresa con mayores ingresos respecto al valor, es muestra un posicionamiento más rentable frente a otras.
I-D/I	$\frac{Ingresos-Dividendos}{Ingresos}$	Esta métrica mide el porcentaje de ingresos que no se distribuye como dividendos, permitiendo diferenciar aquellas empresas que distribuyen dividendos de aquellas que retienen los mismos.
InO/V	$\frac{IngresosnoOperativos}{Ventas}$	Esta métrica mide ratio entre los ingresos no operativos antes de impuestos a ventas. Permite medir la incidencia de ingresos no operativos sobre las ventas totales.
IN-Imp/IN	$\frac{IngresosNetos-Impuestos}{IngresosNetos}$	Esta métrica mide el impacto de los impuestos sobre los ingresos, un porcentaje menor de impuestos redunda en mayores ingresos.
FC/NAcc	$\frac{FlujodeCaja}{Nro.Acciones}$	Esta métrica mide el flujo de caja por acción. Este ratio provee un indicador sobre la fortaleza y sustentabilidad de un modelo de negocios
I/DC	$\frac{Inventario}{DeudaCorriente}$	Este ratio mide la relación entre la deuda corriente y el inventario y expone la capacidad de repago de las deudas usando el inventario en existencia.
G/VM	$\frac{Ganancias}{ValordeMercado}$	Este ratio mide la relación entre las ganancias por acción y el precio por acción, indicando la rentabilidad de la compañía respecto al precio de la acción y el potencial de crecimiento.

Cuadro 3.1: Ratios usados para las series financieras.

Métrica	DT/VN	CT/V	V/AT	IA/VN
Promedio	3,207	3,32e+06	0,171	1010
Desviación	23,35	5,56e+07	0,376	4,89e+04
Mínimo	-163,10	-1,69e+08	-0,005	-9,76
25 %	0,676	0,145	0,0381	0,0033
50 %	1,428	0,812	0,127	0,0273
75 %	3,50	2,23	0,235	0,0536
Máximo	984,13	2,24e+09	16,50	2,37e+06
Métrica	I-D/I	InO/V	IN-Imp/IN	FC/NAcc
Promedio	7460	-1,57e+04	-44,57	0,68
Desviación	2,21e+05	7,24e+05	2203	25,76
Mínimo	-8,30e+06	-1,99e+07	-1,07e+05	-459,29
25 %	0,676	0,0032	0,752	-0,563
50 %	0,95	0,016	0,80	-0,0012
75 %	1,00	0,0553	0,987	0,636
Máximo	24,46	1,81e+07	19,04	776,92

Cuadro 3.2: Valores representativos de algunos ratios para el balance del último trimestre de 2021.

Capítulo 4

Exploración con el Piloto

Como fuera mencionado en la introducción, el objetivo de esta etapa fue ensayar las distintas técnicas de agrupamiento, predicción e inversión en un conjunto pequeño de datos, pero que a su vez fuera representativo del total de datos existentes. Por un lado, esto nos permitió acelerar los tiempos de desarrollo, pero por el otro lado también sirvió para verificar la estabilidad del modelo más allá del conjunto inicial de prueba.

De esta forma generamos subconjuntos de datos con 264 empresas a partir de los conjuntos de datos (BAL, RAT, AGR y PI). Como veremos luego, los tiempos de agrupamiento y entrenamientos de modelos en esta etapa se extendieron desde unos minutos hasta más de 12 horas para la agrupación de series de precios usando DTW+K-Means o la generación de predicciones con Redes Neuronales.

La figura 4.1 muestra cómo empleamos el conjunto de datos para agrupar, entrenar y probar los modelos. Para ello tomamos los ratios financieros para generar dos agrupaciones distintas y las series de precios entre el 1^o de enero de 2017 y el 31 de diciembre de 2021, y su derivada de rendimientos diarios, para generar tres agrupaciones adicionales. Estas cinco agrupaciones se utilizaron posteriormente para proporcionar datos adicionales a los conjuntos de datos utilizados para las predicciones con ARIMA y Redes Neuronales. Se tuvo en cuenta que el período utilizado para generar los agrupamientos no se superpusieran con ningún período de pronóstico (F) para evitar filtraciones de información del conjunto de entrenamiento.

4.1. Agrupamiento de datos

Utilizando las series temporales de precios y ratios financieros, el primer paso fue agrupar las empresas del piloto en grupos de acuerdo con las series de precios (AGR), las series derivadas de rendimientos diarios o los ratios financieros (RAT). La búsqueda generó grupos de empresas que compartían características similares de acuerdo a la distancia utilizada en cada método.

4.1.1. Ratios financieros

La serie de diez ratios financieras (RAT) tiene 58 puntos mensuales para el período 2017-2021. Exploramos no solo la relación entre las series equivalentes de

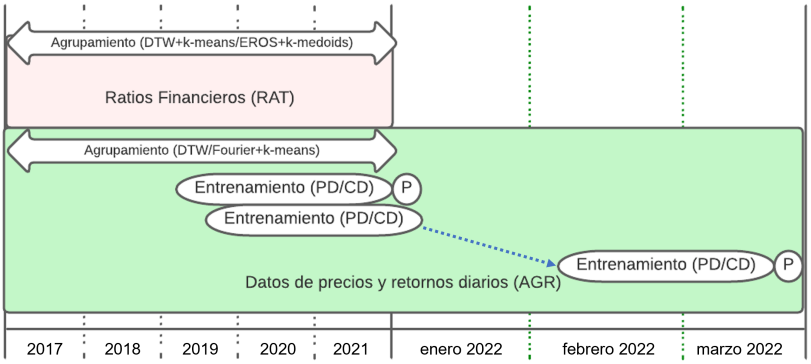

Figura 4.1: Distribución temporal de ambos conjuntos de datos y su uso a lo largo del tiempo, utilizando una ventana móvil para entrenar y ajustar el modelo y la predicción (P) del día siguiente utilizando ARIMA y Redes Neuronales.

cada acción de interés (por ejemplo, Ventas totales a activos para cada acción), sino también la relación entre las diferentes series de proporciones para la misma acción. En otras palabras, nuestro agrupamiento se basa en la comparación de proporciones tanto entre acciones como dentro de ellas. Por lo tanto, usamos dos métodos de agrupamiento diferentes para capturar diferentes conjuntos de información, ya sea en la similitud de las curvas o en la relación entre las series temporales.

Comenzamos usando DTW, una distancia comúnmente empleada para series de tiempo univariadas y multivariadas. Esta distancia encuentra similitudes entre las curvas, inclusive en escenarios de compresión, extensión o traslación de la misma, lo cual nos permitió encontrar relaciones entre empresas con estructuras de ratios similares y que estuvieran desfasadas en el tiempo. Para agrupar, utilizamos un algoritmo clásico como K-Means. El gran problema de esta medida, es que toma las distancias de series multivariadas, pero no llega a capturar la relación entre las distintas series univariadas, ya que compara series similares del conjunto multivariado, pero no la relación con otras series del conjunto.

En consecuencia, también buscamos evaluar la similitud entre estas series temporales multivariadas, teniendo en cuenta la relación entre las distintas series univariadas que la componen. Para ello utilizamos EROS dado que este método captura en la distancia la relación entre las series. Un inconveniente fue que EROS obtiene distancias entre series, pero no vectores, con lo cual el resultado del mismo es una matriz de distancia. Por ende no es posible utilizar K-Means y probamos dos métodos alternativos de agrupamiento que pueden utilizar estas matrices de distancia, OPTICS y K-Medoids. Descartamos OPTICS, el cual permite generar los grupos basados en densidad y determina la cantidad de grupos óptima, debido a que deja observaciones sin asignar a ningún grupo. En el conjunto del piloto fue un 5 % de las observaciones, pero el conjunto final superó el 20 % de las observaciones. Considerando esto nos inclinamos por el uso de K-Medoids dado que el mismo, al utilizar centroides, genera resultados similares a los de K-Means y por ende los resultados son comparables.

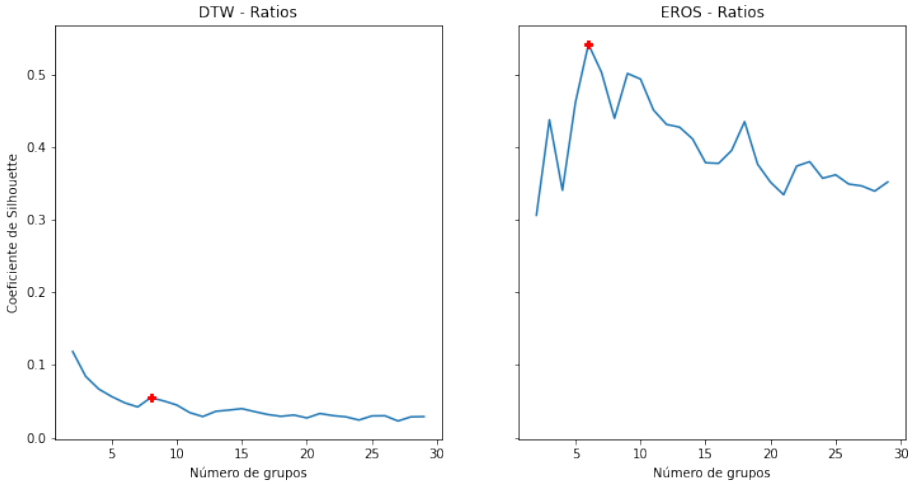

Figura 4.2: Valores de Silhouette para DTW+K-Means y EROS+K-Medoids para distintas cantidades de grupos para el agrupamiento de los Ratios Financieros. En rojo indicamos el valor considerado óptimo para el agrupamiento.

Al utilizar K-Means o K-Medioids queda pendiente la selección del número de grupos, teniendo aquí un balance entre cohesión de los grupos y por el otro lado una separación que sea relevante para proveer de información a los algoritmos de predicción.

Para determinar los óptimos en cada caso, corrimos los algoritmos de agrupamiento con distintos valores de k, variándolo entre 2 y 30, y calculando el coeficiente de Silhouette para cada caso, como puede verse en la figura 4.2. Seleccionamos los agrupamientos que, a partir de los gráficos, se encontraran en un máximo local y que tuvieran un valor de k razonable para generar más de 4 o 5 grupos distintos. Este método es arbitrario, pero esperábamos que nos diera una mayor discriminación de los datos útiles.

Basados en los resultados de la figura 4.2 planteamos un conjunto de 6 grupos para DTW y 6 grupos para EROS.

Una de las principales preguntas era si estos métodos descubrirían características distintas o si serían similares. Para ello realizamos una matriz de correspondencia entre ambos métodos para descubrir si los datos estaban bien definidos o dispersos a lo largo de la matriz. Como podemos ver en la figura 4.3 los valores están muy distribuidos en la matriz sin encontrar una agrupación que sea similar en uno y otro método. Esto indica que ambos métodos utilizaron características distintas y reflejan propiedades de ambos grupos disímiles, que era uno de los objetivos de la búsqueda. Calculamos la medida de Rand ajustada y obtuvimos un valor de 0,018, lo cual representa una muy baja correlación entre ambas distribuciones de grupos.

4.1.2. Serie de precios

Como paso siguiente utilizamos el conjunto AGR, que contiene los precios de las acciones de cierre diario para el período 2017-2021, con unos 1250 puntos de datos para cada acción. La agrupación de estos datos es más sencilla, ya que

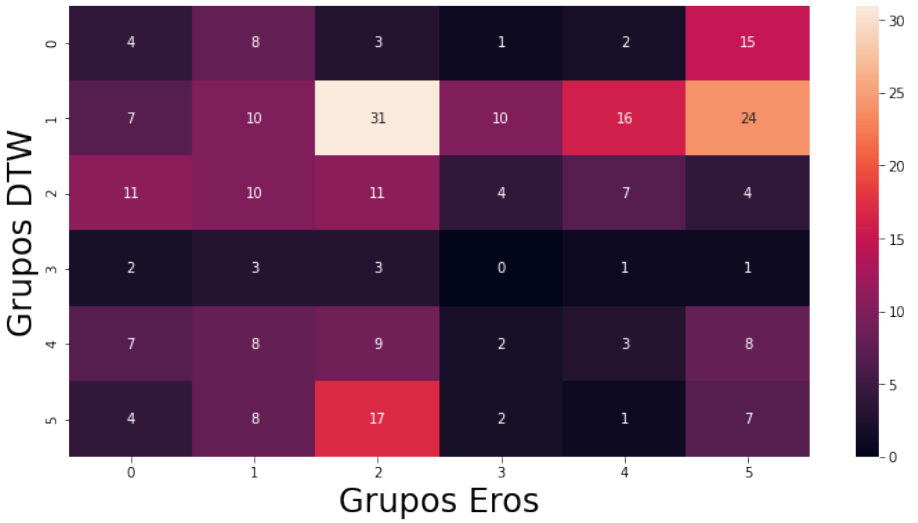

Figura 4.3: Matriz de correspondencia entre la agrupación de series de ratios financieros obtenida mediante DTW+K-Means y EROS+K-Medoids.

la agrupación de series temporales univariadas está bien analizada. Con estas series, exploramos la agrupación de dos series temporales diferentes: precios de cierre y rendimientos diarios, que también incluimos como características, ya que pueden eliminar tendencias en los datos.

Como fue descripto anteriormente, las series se alisaron aplicando una media móvil de 5 días para reducir las oscilaciones de alta frecuencia y quedarnos con las formas de las curvas principales.

Las series de precios de cierre y rendimientos diarios se agruparon utilizando K-Means con dos métricas de distancia diferentes: (a) DTW como distancia entre series temporales de precios o las de rendimientos; y (b) la distancia euclidiana sobre los 100 coeficientes reales e imaginarios de la transformada de Fourier más relevantes.

En el caso del agrupamiento mediante series de Fourier, descompusimos cada curva en sus componentes reales e imaginarios. Como puede verse en la figura 2.2, la amplitud de los mismos disminuye rápidamente, siendo mucho más significativos los primeros. En nuestro caso tomamos los 50 coeficientes más relevantes de cada tipo, para asegurar el tener información adicional, aunque un estudio de componentes principales muestra que con tres dimensiones la variabilidad explicada supera el 82 % y con cinco, el 90 %.

En todos los casos, optimizamos el número de grupos midiendo el coeficiente de Silhouette para k entre 2 y 30 y eligiendo un equilibrio entre el número de conglomerados y el valor del coeficiente de Silhouette. Como en el caso anterior, el objetivo era establecer un número de grupos que proporcionara una buena diferenciación de grupos, pero con un valor razonablemente bajo de k.

Para las series de precios, la búsqueda del agrupamiento óptimo, utilizando DTW, llevó aproximadamente 12 hs, optando por 4 grupos como la mejor selección como combinación de cantidad de grupos y valor de Silhouette, como

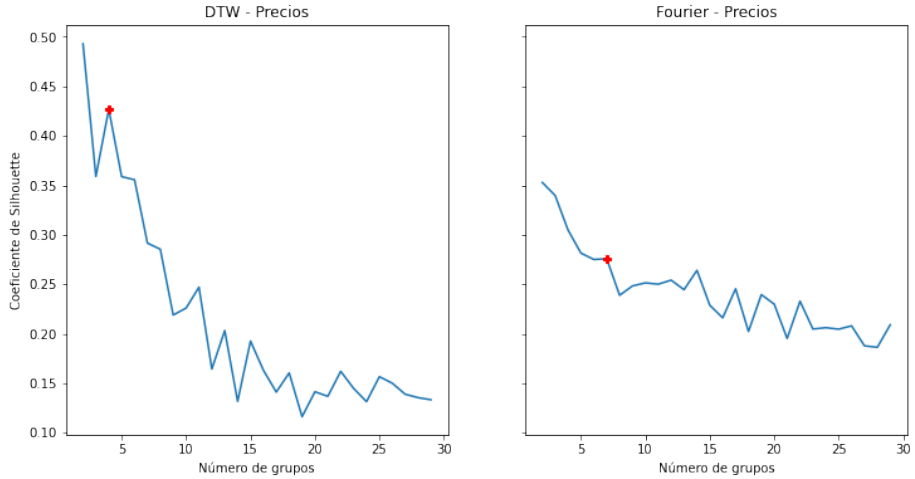

Figura 4.4: Coeficientes de Silhouette para el agrupamiento de series de precios basado en DTW y Fourier utilizando K-Means. En rojo indicamos el valor considerado óptimo para el agrupamiento.

podemos ver en la figura 4.4, mientras que con los coeficientes de Fourier dicho valor se situó en 7 grupos.

El resultado de ambos métodos también muestra una captura de características bastante similares, en especial para los primeros 3 grupos de DTW, como puede verse en la matriz de correspondencia ilustrada en la figura 4.5. La medida de Rand ajustada nos confirmó dicha interpretación, dado que el valor obtenido fue de 0,333, mucho más alta que en el caso de los ratios.

4.1.3. Series de rendimientos diarios

Por último, calculamos las series de rendimientos diarios a partir de las series de precios de cierre, utilizando la ecuación 4.1. Esta forma de calcular el rendimiento diario es la habitual en la literatura financiera, donde se utiliza el logaritmo del cociente entre los precios del día actual y del día anterior.

$$rendimiento = log(\frac{PrecioCierre_t}{PrecioCierre_{t-1}}) \tag{4.1}$$

Las curvas mostraron una gran variabilidad diaria, como se ve en la figura 4.6. Esto causó que la descomposición de Fourier no fuera tan clara, como la mostrada en la figura 2.2, sino una situación en donde todos los coeficientes tienen relevancia, como se ven en la figura 4.7. Por este motivo tuvimos que tomar toda la serie para poder agruparlos.

En ambos casos, DTW y Fourier dieron una clara señal que el agrupamiento óptimo se encontraba en 7 grupos, como podemos ver en la figura 4.8. Sin embargo, en ambos casos, la participación en estos grupos estaba muy desbalanceada, con el 87 % de las series concentradas en 1 solo grupo para DTW y el 89 % de las series concentradas en 2 grupos para Fourier, como puede verse en la matriz de confusión ilustrada en la figura 4.9. En este caso, la medida ajustada de Rand fue menor a la que esperábamos, con un valor de 0,189. Debido a

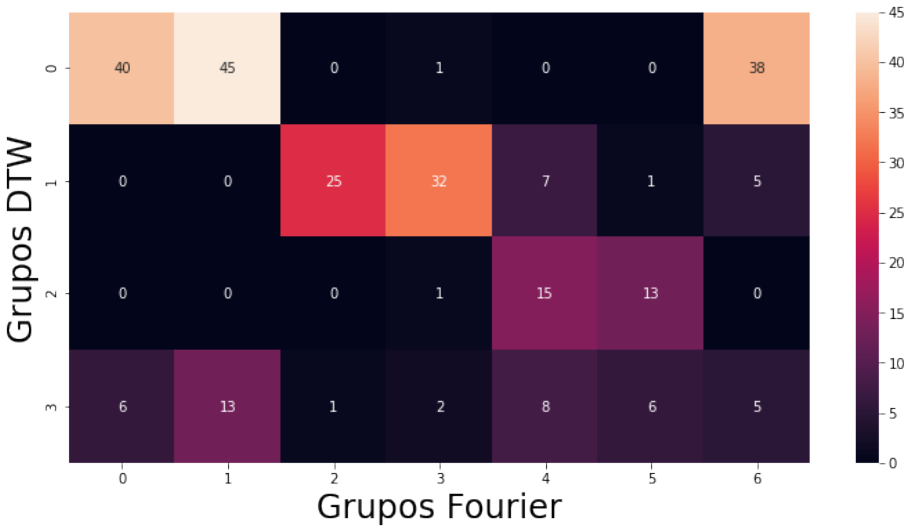

Figura 4.5: Matriz de correspondencia entre la agrupación de series de precios obtenida por DTW+K-Means y Fourier+K-Means.

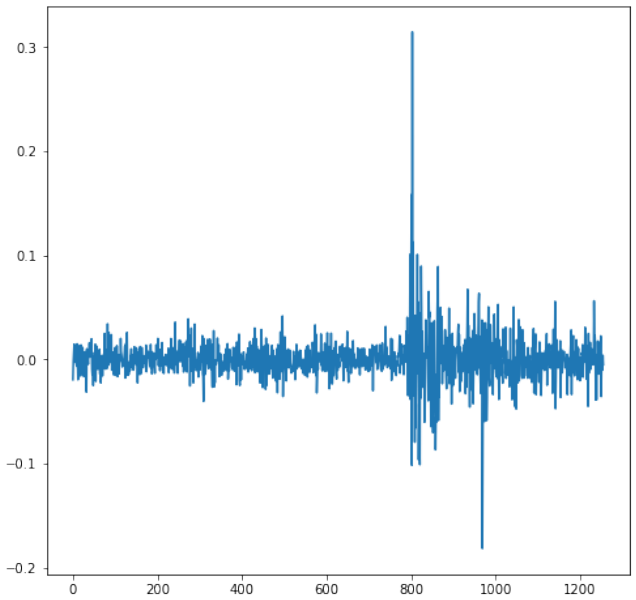

Figura 4.6: Rendimientos diarios de las distintas acciones

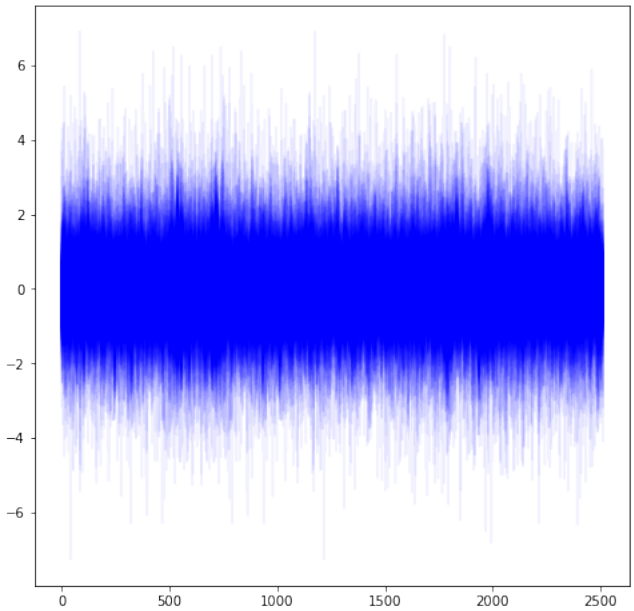

Figura 4.7: Coeficientes de Fourier para las series de rendimientos diarios.

esto, estimamos que poca información podrá obtenerse de este agrupamiento y descartamos DTW como forma de agrupación para series de rendimientos diarios, quedándonos solo con el agrupamiento obtenido mediante los coeficientes de Fourier.

En el anexo B pueden verse los grupos resultantes y las características comunes a las curvas de cada grupo.

4.1.4. Optimización de los tiempos de procesamiento

Como era de esperar, basado en la literatura previa, en el cálculo de distancias con DTW y su posterior agrupamiento enfrentamos problemas de performance. El algoritmo tiene una complejidad temporal *O(NM)* donde N y M son los largos de las series temporales, en nuestro caso 1259 puntos para las series de precios y 57/58 puntos para las series de ratios. Dado que el algoritmo completo de agrupamiento implicaba la comparación de distancias entre todos los pares, el orden de complejidad es $O(NMK^2)$, donde K es la cantidad de series a comparar. En el desarrollo del piloto, detectamos que para este caso reducido, el agrupamiento de las series basado en DTW llevó casi 12 horas. El pronóstico para el conjunto completo, al multiplicar K por 9, con el mismo equipamiento, podía durar al menos un mes.

Como consecuencia, ensayamos varias alternativas, basadas en el muestreo de datos que redujera tanto *N,M* o *K*. Para ello tomamos un conjunto de 515 acciones y los datos de precios entre el 3 de enero de 2017 y el 31 de marzo de 2022 con un total de 1316 puntos a partir del conjunto CD y probamos las siguientes alternativas:

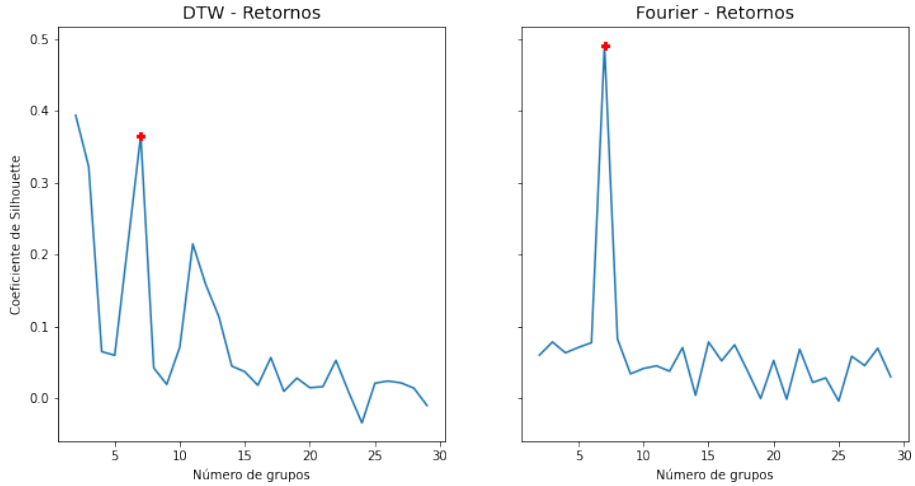

Figura 4.8: Coeficientes de Silhouette para el agrupamiento de series de rendimientos diarios basado en DTW y Fourier utilizando K-Means. En rojo indicamos el valor considerado óptimo para el agrupamiento.

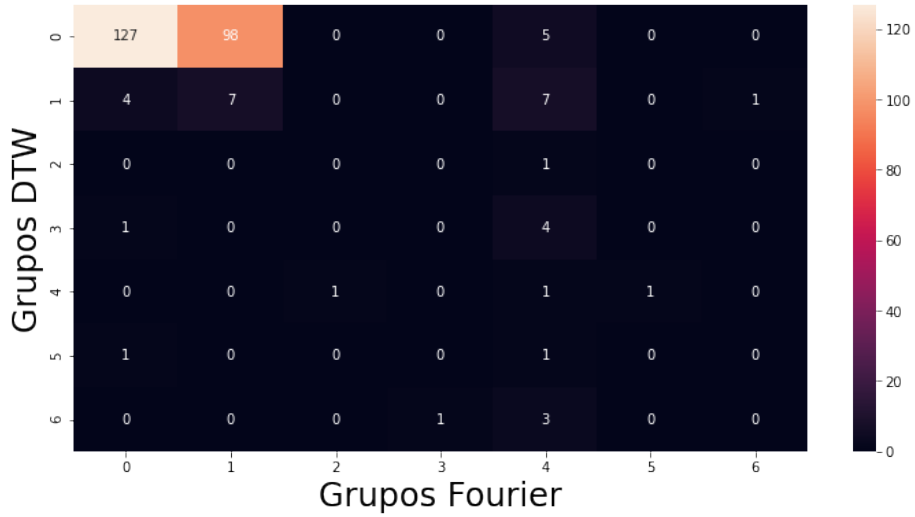

Figura 4.9: Matriz de correspondencia entre la agrupación de series de rendimientos diarios obtenida por DTW+K-Means y Fourier+K-Means.

- a.- Completo: Agrupamiento de las series de precios diarios de las 515 acciones, con 1316 puntos cada una, que constituyó el caso base sobre el cual comparar el resto de los muestreos.
- b.- Subconjunto Aleatorio 50 % A: Agrupamiento de 257 acciones, tomadas al azar con precios diarios, con 1316 puntos cada una y realizando la predicción de grupos de las acciones remanentes, buscando estudiar el impacto de reducir K.
- c.- Subconjunto Aleatorio 50 % B: Agrupamiento de 258 acciones, con las acciones no utilizadas en el punto b con precios diarios, con 1316 puntos cada una y realizando la predicción de grupos de las acciones remanentes, buscando estudiar el impacto de reducir K.
- d.- Día por Medio A: Agrupamiento de 515 acciones con precios día por medio, con 658 puntos cada, una buscando reducir N y M.
- e.- Día por Medio B: Agrupamiento de 515 acciones con precios día por medio, con 658 puntos cada una, pero en los días no utilizados por el punto d buscando reducir N y M.
- f.- Semana: Agrupamiento de 515 acciones con precios semanales, con 263 puntos cada una, buscando reducir N y M.
- g.- Quincena: Agrupamiento de 515 acciones con precios quincenales, con 132 puntos cada una, buscando reducir N y M.
- h.- Últimos 18 meses: Agrupamiento de 515 acciones con precios diarios para los últimos 18 meses (01/jul/2020 a 31/mar/2022), con 376 puntos cada una, buscando reducir N y M de otra forma y dándole mayor importancia a los eventos más recientes.
- i.- Últimos 3 meses: Agrupamiento de 515 acciones con precios diarios para los últimos 3 meses (01/sep/2021 a 31/mar/2022), con 66 puntos cada una, buscando reducir N y M y colocando más énfasis en los últimos eventos.

Para cada caso corrimos el agrupamiento con valores de k entre 2 y 11, y calculamos el tiempo requerido para ejecutarlo, los coeficientes de Silhouette, la distancia euclidiana punto a punto entre la curva completa y las curvas alternativas y para un valor arbitrario (k=7) calculamos las matrices de confusión resultantes de cada caso.

Como puede verse en los resultados (tabla 4.1), los métodos que replicaron mejor los coeficientes de Silhouette del caso completo (figura 4.10) y cuyas matrices de confusión tenía la menor dispersión (figura 4.11), resultaron aquellos muestreos temporales donde se tomaba una muestra a lo largo de toda la serie, ya sea día por medio o semanales.

Esto resultó buenas noticias a futuro porque significó que podríamos reducir N y M en un factor de 2 o 5 confortablemente y aun así estar cerca del resultado óptimo en el agrupamiento. Cualquiera de estos dos métodos representó una disminución notable del tiempo de corrida, en un factor de 4 para el día por medio y de 50 para el muestreo semanal, mostrando la menor diferencia de todos los métodos. Asimismo, tomando en cuenta los casos de los últimos 3 y 18 meses,

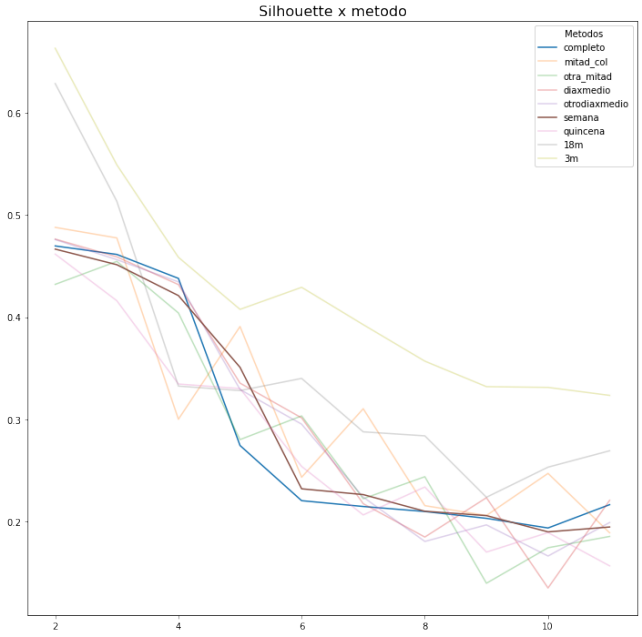

Figura 4.10: Curvas de Silhouette para los distintos métodos de muestreo con k entre 2 y 11

.

estos resultados sugieren que el comportamiento a corto plazo de las acciones es muy diferente al del largo plazo, al menos basándonos en la representación elegida.

4.1.5. Resumen

El resultado de estas tareas generó cinco esquemas de agrupamiento diferentes, dos para la serie de precios, dos para las ratios financieras y uno para los rendimientos diarios, como se puede ver en la Tabla 4.2 según el tipo de distancia utilizada.

Por el otro lado, medimos la consistencia de los grupos mediante los pares de acciones añadidos a la muestra, en particular Google en sus dos acciones (GOOG y GOOGL), Visa (V) y Master Card (MA). Podemos ver en la tabla 4.3 que las dos compañías de tarjetas de crédito están en los mismos grupos, pero es notable que las dos de Google no lo están siempre. En particular, al usar Fourier, el algoritmo encuentra alguna diferencia entre el comportamiento de ellas en series de precios y series de rendimientos diarios.

4.2. Métodos de pronóstico

Los distintos grupos generados en el paso anterior se utilizaron para ejecutar varios métodos de pronóstico aplicados tanto a las series temporales de precios como a las de rendimientos. Usamos los datos entre el 1^o de julio de 2019 al 31 de marzo de 2022 del conjunto PD como un conjunto de entrenamiento

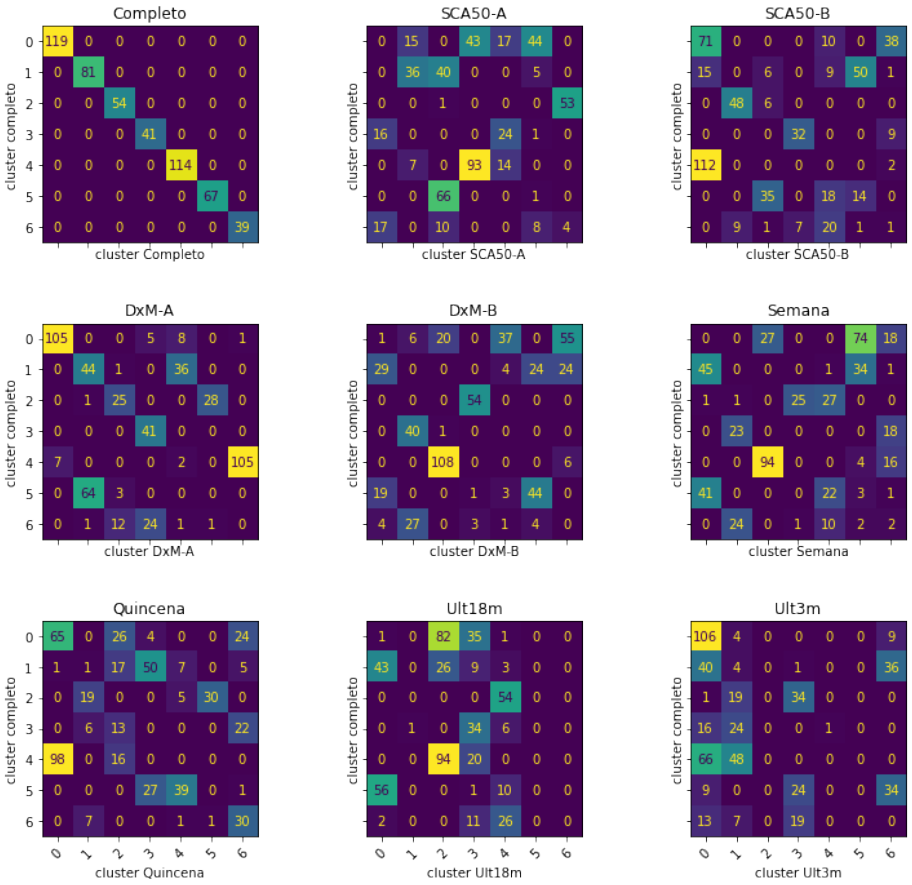

Figura 4.11: Matriz de confusión de los grupos resultantes con k=7 en los distintos métodos de muestreo. Las matrices se generaron considerando en cada caso el agrupamiento con el conjunto de exploración de 264 acciones.

Método	Acciones	Puntos	Tiempo (s)	Δ Silh	Rand Aj.
Completo	515	1316	44501	0	1
Subc Aleat 50 % A	257	1316	15563	0,215	0,442
Subc Aleat 50 % B	258	1316	16391	0,127	0,442
Día por medio A	515	658	9923	0,122	0,689
Día por medio B	515	658	9979	0,104	0,519
Semanal	515	263	892	0,084	0,430
Quincenal	515	132	543	0,149	0,381
Últimos 18 meses	515	376	2885	0,272	0,333
Últimos 3 meses	515	62	337	0,455	0,215

Cuadro 4.1: Características, tiempos, distancia euclidiana entre las curvas de Silhouette y medida de Rand ajustada para cada método de muestreos de puntos en comparación con la curva completa.

Serie temporal	Distancia		
	DTW	EROS	Fourier
Ratios Financieros	X	X	
Precios	X		X
Rendimientos diarios			x

Cuadro 4.2: Distancias usadas para cada tipo de series temporales.

para predecir precios y rendimientos para las 61 sesiones de negociación del mercado entre el 3 de enero de 2022 y el 31 de marzo de 2022. Ejecutamos los escenarios usando una ventana móvil, por lo que agregamos los datos de cierre o rendimientos del día anterior a nuestro conjunto histórico para predecir el día siguiente, tomando siempre la misma cantidad de datos de entrenamiento, como se muestra en la figura 4.1.

Diseñamos varios escenarios (representados en la figura 4.12), utilizando datos agrupados y no agrupados para predecir el precio de cierre o rendimiento de cada acción. Estos escenarios describen los datos de entrenamiento (acciones y características) utilizados para entrenar un modelo para predecir el precio de cierre o los rendimientos diarios de una acción específica:

- Predicciones con datos sin agrupar:
  - Simple: Usa solo los precios de cierre o rendimientos diarios de la acción a predecir.
  - Múltiple (intradiario): Usa los datos de apertura, cierre, máximo, mínimo y volumen o los porcentajes de rendimientos o cambios diarios de la acción a predecir.
  - Múltiple (todas las acciones): Usa el precio de cierre o los rendimientos diarios de todo el conjunto de acciones sin discriminar.
- Predicciones con los datos agrupados

Acción	DTW Ratios	EROS Ratios	Fourier Precios	DTW Precios	Fourier Rendimientos
GOOG	5	2	1	0	1
GOOGL	5	2	6	0	0
MA	1	2	4	1	1
V	1	2	4	1	1

Cuadro 4.3: Número de grupo obtenido con cada método de agrupamiento para las cuatro acciones testigo.

- Múltiple (acciones en cada grupo): Usa los precios de cierre o rendimientos diarios de la acción a predecir y los precios de cierre o rendimientos diarios de todas las acciones en el mismo grupo.

Los dos primeros escenarios se ejecutaron solo con los datos de cada acción, para Simple, con solo los precios de cierre o rendimientos diarios de la acción, o para Múltiple (intradiario), con los datos de apertura, cierre, máximos, mínimos y volumen o cambios diarios de estas variables.

El tercer escenario consistió en utilizar los precios de cierre o rendimientos diarios de todas las acciones del conjunto de datos sin filtrarlos, esperando que el método de predicción pudiera aprovechar aquellos que resulten de utilidad.

El cuarto escenario, y que resulta el centro de esta tesis, es en realidad un conjunto de diferentes escenarios. Se añadieron, además del precio de cierre o rentabilidad diaria de la acción a pronosticar, los datos de cierre o rentabilidad diaria del resto de acciones de un mismo grupo según cada método de agrupamiento. El objetivo fue evaluar si la adición de datos de esas poblaciones (supuestamente relacionadas de acuerdo con el proceso de agrupación) mejoraba la capacidad de pronóstico.

Ejecutamos estos escenarios utilizando dos métodos de predicción habituales en el tratamiento de series temporales; el primero fue ARIMA y el segundo una red neuronal. En ambos casos se entrenaron los modelos para cada uno de los diferentes conjuntos de características.

4.2.1. ARIMA

Generamos un total de 16 predicciones, 8 para los precios y 8 para los rendimientos diarios, correspondiendo 6 de ellos a los modelos sin tener en cuenta los agrupamientos y los restantes 10 a los 5 agrupamientos seleccionados tanto para precios como para rendimientos.

El mejor modelo ARIMA para cada acción se seleccionó utilizando los criterios de información de Akaike, ejecutando modelos con diversos valores de p, d y q, quedándonos con el cual minimizaba el valor de Akaike. Esto fue optimizado mediante la función auto_arima del paquete pmdarima desarrollado por Smith et al. (17) que utiliza el algoritmo stepwise propuesto por Hyndman and Khandakar (2008) para identificar los parámetros dentro del espacio de búsqueda.

Usamos como variable endógena el precio de la acción o rendimiento diario a predecir, y salvo en el caso Simple, donde no se contaba con datos exógenos,

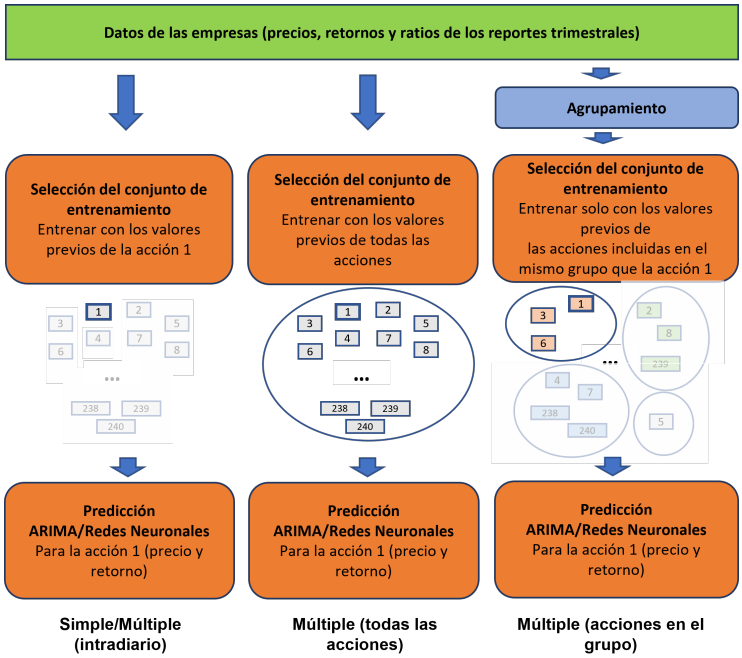

Figura 4.12: Uso de los datos en cada escenario, con el Simple y el Múltiple (intradiario) utilizando solo los datos de cada acción para predecir dicha acción, Múltiple (todas las acciones) usando los precios o rendimientos de todas las acciones para realizar las predicciones de dicha acción y Múltiple (acciones en cada grupo) empleando los datos de cada grupo para predecir los precios y rendimientos de cada acción en dicho grupo.

	Predicción ARIMA			
	Precios		Rendimientos	
Método	sMAPE	MASE	sMAPE	MASE
Simple	1,994	1,009	11,726	6,518
Múltiple (intradiario)	1,864	0,939	12,393	7,199
Múltiple (todas las acciones)	23,058	17,769	11,626	24,239
Múltiple (DTW Ratios)	2,476	1,232	11,834	6,567
Múltiple (EROS Ratios)	3,671	1,936	11,818	6,554
Múltiple (DTW Precios)	2,642	1,340	11,736	6,442
Múltiple (Fourier Precios)	2,236	1,121	11,805	6,561
Múltiple (Fourier Rendimientos)	5,310	3,185	14,618	7,256

Cuadro 4.4: Resultados de los métodos de pronóstico aplicados sobre precios y rendimientos con el valor de sMAPE y MASE en la tendencia de cada pronóstico generado con ARIMA. Los últimos 5 métodos en cada sección de la tabla pertenecen a Múltiple (Acciones en grupo) nombrados como el método de agrupación, indicando la combinación diferente de datos/distancia utilizada.

utilizamos como variables exógenas el resto de la información propuesta en cada caso.

Una vez encontrado el modelo óptimo, ajustamos las predicciones en forma diaria hasta obtener los 61 puntos requeridos correspondientes a cada una de las sesiones de mercado del primer trimestre de 2022.

Con estos pronósticos calculamos los valores de MASE y sMAPE para cada escenario/combinación en la tabla 4.4, Como predicción ingenua en el caso de MASE tomamos los precios o rendimientos del día anterior.

Podemos ver que las predicciones de precios muestran en general mejores valores que las predicciones de rendimientos, pero salvo Múltiple (intradiario) ninguno puede mejorar la predicción ingenua. También se puede observar que el agrupamiento de series de rendimientos, utilizando Fourier, muestra consistentemente los peores índices tanto para la predicción de precios como la de rendimientos, con la excepción del uso de la información de todas las acciones.

4.2.2. Redes Neuronales

Con este método generamos un total de 15 predicciones, 8 para los precios y 7 para los rendimientos diarios, correspondiendo 5 de ellos a los modelos sin tener en cuenta los agrupamientos y los restantes 10 a los 5 agrupamientos seleccionados tanto para precios como para rendimientos. La diferencia con ARIMA es que no ejecutamos el Múltiple (intradiario) para rendimientos, dado que la disponibilidad de los TPUs en la nube era limitada y preferimos concentrarlo en el caso completo.

Para las predicciones con Redes Neuronales utilizamos una red compuesta por capas convolucionales, LSTM y densas para obtener las predicciones. La red tiene un diseño típico compuesto por una capa convolucional para extraer características de los datos de entrada, dos capas LSTM recurrentes para codificar la relación temporal entre los valores de la serie temporal y dos capas densas

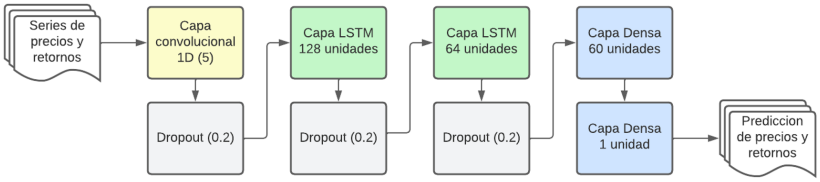

Figura 4.13: Arquitectura de la red neuronal CNN-LSTM utilizada para la predicción de precios y rendimientos.

	Predicción con Redes Neuronales			
	Precios		Rendimientos	
Método	sMAPE	MASE	sMAPE	MASE
Simple	4,725	2,651	12,191	6,623
Múltiple (intradiario)	5,502	3,166	N/D	N/D
Múltiple (todas las acciones)	89,399	167,123	16,381	7,732
Múltiple (DTW Ratios)	9,397	5,180	14,901	7,373
Múltiple (EROS Ratios)	8,358	4,707	15,795	7,710
Múltiple (DTW Precios)	9,469	5,372	14,877	8,161
Múltiple (Fourier Precios)	9,269	5,225	14,710	7,348
Múltiple (Fourier Rendimientos)	5,116	2,615	15,439	7,609

Cuadro 4.5: Resultados de los métodos de pronóstico aplicados sobre precios y rendimientos con el valor de sMAPE y MASE de cada pronóstico generado con una red neuronal. Los últimos 5 métodos en cada sección de la tabla pertenecen a Múltiple (Acciones en grupo) nombrados como el método de agrupación, indicando la combinación diferente de datos/distancia utilizada.

para generar la predicción del precio y el rendimiento de al día siguiente y la arquitectura de la misma puede verse en la figura 4.13.

Usamos una red simple para asegurar que los resultados sean impulsados principalmente por las características y el agrupamiento y, en menor medida, por las capacidades de modelado de una red compleja.

Para ello generamos un conjunto de entrenamiento con todas las variables utilizadas en ventanas de 60 períodos, reservando los últimos 20 grupos de datos para realizar la validación del mismo. Con ese modelo, incrementalmente fuimos prediciendo el día siguiente e incorporándolo al modelo para la predicción del siguiente, tanto para precios como para rendimientos.

En la tabla 4.5 podemos ver los valores de MASE y sMAPE obtenidos en cada predicción.

Como en el caso anterior, podemos ver que las predicciones de precios en general presentan mejores valores que las predicciones de rendimientos, pero sin excepción son peores la predicción ingenua. También se puede observar que el agrupamiento de series de rendimientos utilizando Fourier, y al contrario que con ARIMA, muestra muy buenos resultados con la predicción de precios, mientras que es equivalente al resto en la predicción de rendimientos diarios.

En ambos casos, el alimentar a los modelos de predicción de precios con todos los datos disponibles sin discriminación alguna, muestra el peor nivel de ajuste, mientras que los modelos que discriminaron entre los datos de los grupos mostraron un nivel de ajuste mucho mejor, comparable al de los métodos que solo utilizaban los datos propios de la acción. Este fue el primer indicador de que el método propuesto tenía cierto sentido. Otro punto interesante fue que ARIMA tiene valores de MASE en general mucho mejores que los de Redes Neuronales para precios, pero son de un orden similar para rendimientos.

4.3. Resultado de inversión

Con los 31 pronósticos generados, 16 de ARIMA y 15 de Redes Neuronales, ensayamos distintas estrategias de inversión para ver la utilidad de este método. Si bien los resultados del ajuste no eran favorables, la generación de señales de inversión y por ende el rendimiento de ejecutar los mismos mostraría el resultado final.

Usamos como punto de partida una cartera igualmente ponderada, considerando cada uno de los fondos de cada acción independientes entre sí, reinvirtiendo las ganancias o asumiendo las pérdidas para la próxima transacción y permitiendo inversiones fraccionadas por simplicidad, de forma de ignorar el problema de ecualizar una inversión en acciones de valor por 1000 USD contra una por valor de 1 USD. De esta forma dimos un valor de 1 a cada fondo individual y disminuimos o aumentamos su valor en el rendimiento diario según la ocasión y resultado de la inversión. De esta forma, los métodos de inversión rastreaban los resultados de cada uno de los 264 fondos individuales y promediamos los rendimientos de cada uno al final del período, para obtener el rendimiento general del método. Esto representa una estrategia simple de inversión intradiaria donde se abre la transacción al iniciar el día y se cierra al finalizar la jornada, sin utilizar estrategias de protección de ganancias o reducción de pérdidas como un stop loss.

El primer paso de la estrategia de inversión consistió en interpretar las predicciones y producir señales de compra, venta o pasar y no ejecutar una inversión para cada acción diariamente. Para simplificar las transacciones el período de tenencia se extiende desde la apertura hasta el cierre diario, pero sin conservar la tenencia de acciones de un día para otro.

El algoritmo de búsqueda de oportunidades de inversión activa una señal para cada día de mercado t y cada acción entre el 3 de enero de 2022 y el 31 de marzo de 2022, de acuerdo con las siguientes reglas:

- Señal de Compra si $cierre_{t-1} < predicho_t$ y $apertura_t < predicho_t$.
- Señal de Venta si $cierre_{t-1} > predicho_t$ y $apertura_t > predicho_t$.
- Paso, si ninguna de las dos opciones previas se cumple y se decide no realizar una operación de inversión para ese día en particular

Después de calcular las señales diarias, el algoritmo de inversión ejecuta una de las siguientes transacciones para cada acción de la muestra y calcula el rendimiento de cada operación de la siguiente manera:

- Compra: El algoritmo simula la compra de la acción al comenzar el día y su venta al finalizar la jornada, con un rendimiento resultante del ratio entre los *precios de cierre y apertura* de la acción. Este rendimiento se aplica a la cantidad existente en el fondo individual de dicha acción.
- Venta en Corto: Aquí el algoritmo simula la venta en descubierto de la acción al comienzo del día y su recompra y devolución al finalizarlo. En este caso, el rendimiento entre los *precios de apertura y cierre* se aplica a la cantidad existente en el fondo individual de esa acción.
- Señal de no hacer nada: No se ejecuta ninguna transacción para esa acción en particular ese día. En consecuencia, no hay rendimiento y el fondo individual mantiene su valor.

Para más detalle en las operaciones de compra en largo y venta en corto, ver la sección 2.1.1.

Comparamos estos resultados con dos estrategias comunes de negociación del mercado de valores entre pequeños comerciantes descritos en la sección 2.2 y dos métodos de referencia:

(a) Comprar y mantener (Buy and Hold).

(b) Convergencia/Divergencia de la Media Móvil (MACD).

(c) Compra aleatoria: Esta estrategia crea señales aleatorias de compra y venta para las operaciones diarias (usando un generador de números pseudo-aleatorios), produciendo un 25 % de compras largas, un 25 % de ventas cortas y un 50 % de no hacer nada para cada acción, independientemente del comportamiento real de la misma. Este método se utiliza para encontrar cualquier sesgo que pueda generar el método personalizado al alimentar estas señales generadas en el mismo algoritmo de ejecución comercial.

(d) Pronóstico perfecto: Es la rentabilidad que se obtiene cuando el inversor tiene un conocimiento perfecto del futuro. Así, las previsiones comerciales son 100 % acertadas. Esta estrategia muestra la máxima utilidad que se puede obtener en el período de estudio, con las mismas restricciones y presupuesto inicial.

Finalmente, para simplificar la evaluación, el cálculo de los rendimientos no tiene en cuenta los costos de transacción, que suelen ser pequeños para transacciones de compra, en el orden de los centavos de dólar, pero pueden ser sustanciales en comparación con el rendimiento obtenido en el caso de ventas en corto, ya que pueden llegar al 1 % de la transacción.

Con la ejecución de los métodos de inversión calculamos la cantidad de transacciones netas generadas, obviando aquellas ocasiones donde el algoritmo decidía no invertir y el porcentaje de precisión, donde el algoritmo acertó la operación correcta para el día, que puede verse en la tabla 4.6. Una vez ejecutado el algoritmo calculamos el rendimiento del fondo respecto al valor inicial, el cual puede verse en la tabla 4.7.

También estudiamos la evolución de los resultados de las inversiones en forma diaria, comparando los mismos con la evolución de los métodos de referencia, los cuales pueden verse en la figura 4.14. Estos indican, para las predicciones

	Predicción con ARIMA			
	Precios		Rendimientos	
Método	Nro. Trans.	% Aciertos	Nro. Trans.	% Aciertos
Simple	6124	39,5 %	14000	48,5 %
Múltiple (intradiario)	8261	59,7 %	13625	48,0 %
Múltiple (todas las acciones)	12861	49,0 %	13968	48,4 %
Múltiple (DTW Ratios)	8466	39,4 %	13990	48,5 %
Múltiple (EROS Ratios)	8890	41,6 %	14002	48,4 %
Múltiple (DTW Precios)	8562	40,2 %	13984	48,5 %
Múltiple (Fourier Precios)	8428	40,0 %	13980	48,5 %
Múltiple (Fourier Rendimientos)	9197	43,0 %	14008	48,4 %
	Predicción con Redes Neuronales			
	Precios		rendimientos	
Método	Nro. Trans.	% Aciertos	Nro. Trans.	% Aciertos
Simple	13473	49,0 %	14385	48,4 %
Múltiple (intradiario)	13553	50,4 %	-	-
Múltiple (todas las acciones)	14343	48,8 %	14387	49,2 %
Múltiple (DTW Ratios)	13913	51,0 %	14390	49,1 %
Múltiple (EROS Ratios)	13858	50,8 %	14387	49,1 %
Múltiple (DTW Precios)	13951	51,2 %	14390	49,2 %
Múltiple (Fourier Precios)	13896	51,0 %	14392	49,1 %
Múltiple (Fourier Rendimientos)	10680	44,3 %	14389	49,3 %

Cuadro 4.6: Resultados de los métodos de pronóstico aplicados sobre precios y rendimientos, mostrando el número de operaciones ejecutadas y el porcentaje de pronósticos exitosos. Los últimos 5 métodos en cada sección de la tabla pertenecen a Múltiple (Acciones en grupo) nombrados como el método de agrupación, indicando la combinación diferente de distancias o métodos y datos utilizados. Las diferentes cantidades de transacciones se debe a la acción de la estrategia de inversión que según el resultado de la predicción puede definir el operar o no en una determinada jornada para una acción.

Métodos de base	Rendimiento			
Predicción perfecta	241.18 %			
Compra al azar	1.38 %			
Buy and Hold	-2.04 %			
MACD	-3.02 %			
Métodos de Predicción	ARIMA	Redes Neuronales Precios	ARIMA	Redes Neuronales Rendimientos
Simple	3,33 %	3,17 %	4,32 %	3,57 %
Múltiple (intradiario)	-1,83 %	6,92 %	3,13 %	-
Múltiple (todas las acciones)	6,19 %	3,82 %	4,21 %	4,09 %
Múltiple (DTW Ratios)	4,64 %	8,13 %	4,42 %	3,35 %
Múltiple (EROS Ratios)	5,73 %	9,09 %	4,75 %	3,31 %
Múltiple (DTW Precios)	4,68 %	10,24 %	4,51 %	3,51 %
Múltiple (Fourier Precios)	4,04 %	8,37 %	4,56 %	5,37 %
Múltiple (Fourier Rendimientos)	5,89 %	3,99 %	4,37 %	4,78 %

Cuadro 4.7: Rendimiento de los tres meses de prueba por método de pronóstico. Los dos primeros métodos corresponden a la inversión perfecta y a la inversión al azar, los dos segundos a métodos de inversión populares y los ocho últimos a las previsiones que generamos a través de ARIMA y Redes Neuronales, utilizando los datos propios de cada acción, el conjunto completo de datos o una selección de datos basados en la selección de conglomerados.

basadas en Redes Neuronales, una lenta pero persistente acumulación de ganancias. ARIMA muestra retrocesos que son compensados a mediados de febrero de 2022 con el aprovechamiento de la caída de los mercados, como consecuencia del inicio de la invasión a Ucrania.

4.4. Consistencia de resultados

Una vez finalizada la exploración principal nos encontramos con dudas respecto a la validez de los resultados y analizamos algunas situaciones de contexto que pudieran influir en dichos resultados positivos. Por un lado, ahondamos sobre el efecto de la inicialización aleatoria de los modelos de Redes Neuronales, por el otro estudiamos si el período no podría haber sido particularmente favorable para estos algoritmos. Considerando estos puntos realizamos algunos experimentos adicionales que detallaremos a continuación.

4.4.1. Redes Neuronales

En el periodo analizado, los resultados obtenidos con los métodos basados en agrupamiento, mostraron un rendimiento mejor que los métodos de referencia. A su vez, aquellos basados en Redes Neuronales, fueron superiores a los calculados con ARIMA. Considerando el componente estocástico de la inicialización de Redes Neuronales, podría haber ocurrido que los resultados fueran buenos simplemente porque las inicializaciones hubieran sido favorables a un resultado positivo y con otro conjunto de inicializaciones, dicho resultado habría sido negativo.

En consecuencia, diseñamos otro experimento que permitiera despejar esta duda. Primero tomamos del conjunto de acciones, 60 acciones al azar y que

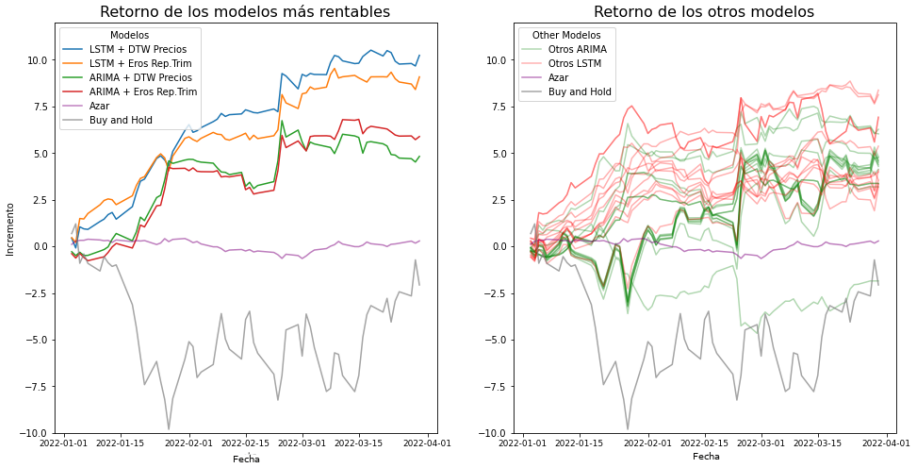

Figura 4.14: Evolución de los rendimientos del conjunto de prueba para el primer trimestre de 2022. El panel izquierdo muestra los resultados de los métodos más rentables, mientras que el panel derecho presenta los métodos restantes coloreados según la categoría del método de pronóstico. En ambos se incluye como referencia la estrategia de compra aleatoria y la de Buy and Hold.

tuvieran un rendimiento promedio similar al de la muestra total. Luego generamos 30 predicciones para el conjunto de acciones con el agrupamiento basado en DTW para Precios en las mismas condiciones, es decir, usando el conjunto de datos de las 264 acciones originales pero solo para las 60 acciones. El resultado final fue una media de 7,3 % de rendimiento con una desviación de 1,5 %. Puede verse en la figura 4.15 el histograma con todos los rendimientos.

Planteando un test de Student con estos valores bajo la H_0 que la media de rendimiento es 0 %, podemos rechazarla confortablemente dado que su valor p es menor a 0,0001 y el intervalo de confianza al 95 % es [6,73 %;7,87 %].

4.4.2. Análisis de las transacciones

Realizamos asimismo el análisis de las transacciones ejecutadas, para ver el rendimiento promedio de las ganancias y el valor promedio de las pérdidas, como puede verse en la tabla 4.8. Observamos que para el caso de Azar, el promedio y cantidad de pérdidas y ganancias es similar, mientras que en todas las predicciones con agrupamiento el promedio de las ganancias es mayor que el de las pérdidas, explicando así la ganancia total del método.

Si realizamos un test de Student para ver si dichas medias son distintas, podemos ver que esta diferencia es estadísticamente significativa para los cuatro métodos ensayados, dado que su $p - valor < 0,05$ que contrasta con el caso de azar donde H_0 no puede ser rechazada.

4.4.3. Otros períodos

Si bien habíamos resuelto con el punto anterior que, a pesar del componente estocástico de la predicción con Redes Neuronales, la ganancia superior a otros métodos tradicionales existía, decidimos estudiar si el período de prueba podría

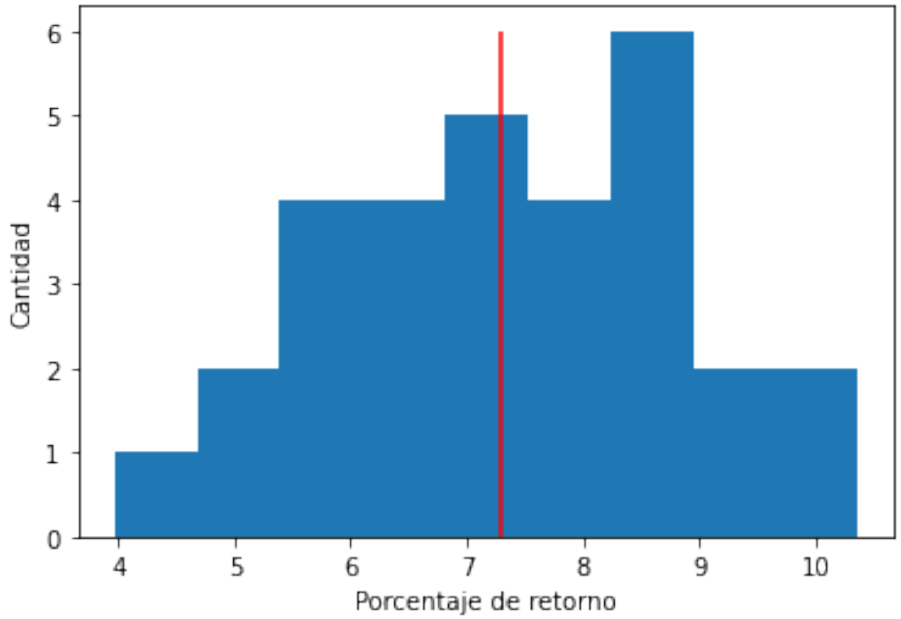

Figura 4.15: Histograma con los rendimientos de las 30 corridas. En rojo la media de los resultados.

Métodos		Ganancias			Pérdidas			p-value
Agrupamiento	Predicción	Media	σ	# Tr.	Media	σ.	# Tr.	
-	Azar	1,87 %	1,90 %	3532	1,88 %	1,92 %	3557	0,8256
DTW Precios	ARIMA	1,96 %	2,11 %	4301	1,81 %	1,80 %	4007	$5{,}15\text{x}10^{-6}$
EROS Ratios	ARIMA	1,98 %	2,11 %	4458	1,76 %	1,74 %	4165	$1{,}46\text{x}10^{-7}$
DTW Precios	R.Neuronales	1,97 %	2,01 %	7078	1,80 %	1,82 %	6498	$2{,}64\text{x}10^{-7}$
EROS Ratios	R.Neuronales	1,96 %	2,03 %	7022	1,81 %	1,82 %	6470	$6{,}37\text{x}10^{-7}$

Cuadro 4.8: Estadísticas descriptivas de las ganancias y pérdidas obtenidas usando los dos métodos de mejor desempeño para cada técnica de pronóstico y para la inversión al azar para el piloto con sus correspondientes valores de p para un t-test.

Métodos de base		rendimiento			
Buy and Hold		-14,59 %			
MACD		-10,63 %			
Métodos de predicción	# Transacciones	% Aciertos	sMAPE	MASE	Rend
Múltiple (DTW Ratios)	13704	51,1 %	10,107	5,233	7,41 %
Múltiple (EROS Ratios)	13728	50 %	10,145	5,382	2,23 %
Múltiple (DTW Precios)	13694	50,5 %	10,035	5,246	5,59 %
Múltiple (Fourier Precios)	13722	50,9 %	10,363	5,589	6,38 %

Cuadro 4.9: Rendimientos y datos de las predicciones para el 2do trimestre de 2022 con Redes Neuronales y cuatro métodos de agrupamiento, contra un par de métodos básicos de inversión

haber inducido favorablemente al resultado. Para evaluar esto, analizamos los modelos con dos períodos y metodologías distintas, en las mismas condiciones de los cálculos anteriores y comparándolos con dos métodos básicos como Buy and Hold y MACD.

Extensión del método

En el primer caso, probamos la extensión con los mismos grupos calculados en el punto anterior para DTW Ratios, EROS Ratios, DTW Precios y Fourier de Precios para la predicción de precios, en el segundo trimestre de 2022. Este período se caracteriza por una baja gradual pero prácticamente constante de los valores de las acciones a lo largo del período.

Los resultados obtenidos, presentados en la tabla 4.9 muestran una mejora notable frente a los métodos tradicionales utilizando Redes Neuronales, ya que estos presentaron fuertes pérdidas, mientras que el algoritmo ensayado muestra una ganancia comparable a la del primer trimestre de 2022. Podemos ver también que la cantidad de predicciones, aciertos, sMAPE y MASE son similares a los del primer trimestre de 2022.

Prueba en el período de comienzo de COVID-19

Quizá el período más complejo de estudio para este método es el que corresponde al comienzo de la pandemia de COVID-19, dado que los mercados venían de un crecimiento casi ininterrumpido desde mediados de 2009, que había multiplicado su valor por 9, para en febrero de 2020 comenzar una caída drástica en el comienzo de la cuarentena causada por la pandemia que llega a su pérdida máxima a fines de marzo con un 30 % de pérdidas y una recuperación parcial durante abril de 2020 para concluir con una pérdida total del 15 % del valor promedio de las acciones. Esto resultaba desafiante dado que el periodo de entrenamiento se componía de dicha etapa de crecimiento y la baja súbita de las acciones podía causar fallas en el pronóstico de precios.

En este caso ejecutamos el método completo, es decir, generamos los grupos para DTW Ratios, EROS Ratios, DTW Precios y Fourier de Precios y utilizamos estos grupos para la predicción de precios con ARIMA y Redes Neuronales en el período febrero a abril de 2020.

Los resultados obtenidos con Redes Neuronales, presentados en la tabla 4.10 muestran una mejora notable frente a Buy and Hold, él que presentó fuertes

Métodos de base		rendimiento			
Buy and Hold		-15,99 %			
MACD		6,25 %			
ARIMA					
Métodos de predicción	# Transac.	% Aciertos	sMAPE	MASE	Rend
Múltiple (DTW Ratios)	9251	32,8 %	4,323	0,982	-1,63 %
Múltiple (EROS Ratios)	9536	33,9	5,404	1,266	-3,73 %
Múltiple (DTW Precios)	9053	32,9 %	4,495	1,022	-0,85 %
Múltiple (Fourier Precios)	8941	32,9 %	4,414	1,005	-3,19 %
Redes Neuronales					
Métodos de predicción	# Transac.	% Aciertos	sMAPE	MASE	Rend
Múltiple (DTW Ratios)	15383	47,2 %	21.170	5,274	6,54 %
Múltiple (EROS Ratios)	15422	47 %	21,211	5,399	4,11 %
Múltiple (DTW Precios)	15381	47,3 %	21,786	5,554	4,99 %
Múltiple (Fourier Precios)	15381	47,6 %	21,827	5,637	5,4 %

Cuadro 4.10: Rendimientos y datos de las predicciones para el período de febrero 2020 a abril 2020 con Redes Neuronales y cuatro métodos de agrupamiento, contra un par de métodos básicos de inversión

pérdidas y un rendimiento similar a MACD en todos los casos positivos. Podemos ver también que la cantidad de predicciones, aciertos, sMAPE y MASE son similares a los del primer trimestre de 2022. Sin embargo, los resultados con ARIMA no son tan favorable, tanto en la cantidad de aciertos como en el rendimiento obtenido. Esta diferencia probablemente se deba a que la fuerte caída registrada a partir de fines de febrero no pudo ser modelizada correctamente por los modelos lineales, a pesar de tener un valor de sMAPE y MASE superiores a las Redes Neuronales, las cuales consiguieron capturar los componentes no lineales del período.

4.5. Hallazgos del piloto

Medimos la bondad de los pronósticos utilizando dos medidas amplias. Una es la precisión, es decir, la relación entre la clasificación correcta de transacciones y el número de todas las transacciones. El otro es el rendimiento promedio, que es el rendimiento obtenido por un hipotético fondo de inversión en su totalidad.

Adicionalmente, calculamos la precisión del pronóstico de cada método para aquellas transacciones efectivamente ejecutadas, ya sean compras o ventas. Las tablas 4.4 y 4.5 muestran el sMAPE y MASE de los pronósticos, mientras que la precisión puede verse en la tabla 4.6.

Con la única excepción del modelo Múltiple (intradiario) para precios en ARIMA, la precisión no es significativamente mayor que la de una predicción aleatoria y, en muchos casos, sustancialmente peor. Los valores de sMAPE y MASE generalmente indican una desviación sustancial del caso naïve, donde el precio o rendimiento del día anterior se toma como predicción para el día siguiente, mejorándolo solo en el caso del modelo Múltiple (intradiario) para precios en ARIMA.

Sin embargo, la situación cambia sustancialmente cuando hacemos backtesting de las recomendaciones de inversión (compra y venta de órdenes) y las medimos en términos de rentabilidad, cuyos resultados se resumen en la Tabla 4.7. Primero, observamos que, independientemente del método de pronóstico, los resultados están muy lejos de ser un pronóstico perfecto. Esto significa que los movimientos del mercado son en su mayoría impredecibles, pero no son totalmente aleatorios. En segundo lugar, las simulaciones empíricas reflejan que todos los métodos de pronóstico (excepto ARIMA múltiplo intradiario para precios) superan a todos los métodos de referencia. En tercer lugar, y lo que es más importante, aquellos métodos que utilizaron información aumentada específica para predecir los precios tuvieron un rendimiento promedio mayor que los métodos que utilizaron solo los datos disponibles para cada acción. Por el contrario, si usamos datos de las 264 acciones de la muestra (Múltiple - todas las acciones en la Tabla 4.7) los resultados son peores que otros métodos (excepto el pronóstico ARIMA con precios).

En cierto modo, esta situación indica que más datos no necesariamente son mejores. Por lo tanto, incluir una técnica de agrupamiento en la preparación de datos puede traer dos beneficios: mejorar los resultados de pronóstico y reducir el tiempo de cómputo.

De acuerdo con parte de la literatura, encontramos que la capacidad de pronóstico de las Redes Neuronales generalmente supera a la de ARIMA, en especial en un escenario como el ocurrido entre febrero y abril de 2020. La razón de tal resultado podría deberse a la mejor idoneidad de las Redes Neuronales para lidiar con las no linealidades y, por lo tanto, para extraer información adicional, como se observa en los rendimientos promedio que se muestran en la tabla 4.11. Finalmente, los pronósticos de precios generalmente dan mejores resultados que los pronósticos de rendimiento, como se muestra en la Tabla 4.7.

La figura 4.14 muestra la evolución dinámica de los rendimientos obtenidos por los diferentes métodos de pronóstico. En el panel izquierdo, seleccionamos los mejores cuatro métodos (los dos mejores de ARIMA y los dos mejores de Redes Neuronales), y mostramos también algunos métodos de referencia para comparar. El panel derecho exhibe las simulaciones restantes con ARIMA y Redes Neuronales como referencia. Podemos observar un dominio de los métodos de Redes Neuronales sobre las contrapartes de ARIMA para el período de espera. Podemos observar que la mayoría de los métodos de pronóstico muestran un claro salto en la última semana de febrero de 2022. Alrededor de esta fecha, Rusia invadió Ucrania, lo que produjo una caída de los mercados notables. Por lo tanto, los métodos de pronóstico más complejos parecen capaces de capturar ese estrés del mercado y aprovecharlo.

Para tener una mejor comprensión del comportamiento de nuestra prueba histórica, agregamos los resultados de todas las transacciones ejecutadas considerando los dos mejores métodos para cada modelo de pronóstico, y lo comparamos con los rendimientos obtenidos a través de una estrategia aleatoria. Como resultado, encontramos un rendimiento promedio pequeño con desviaciones estándar bajas alrededor del rendimiento medio, con un promedio mayor cuando se ejecutan Redes Neuronales que ARIMA. Sin embargo, considerando la gran cantidad de transacciones (obtenemos hasta una señal de negociación para cada acción cada día del período de muestra), estas se acumulan y crean las variaciones observadas en la figura 4.14. Cabe señalar que las Redes Neuronales tienen una desviación estándar más alta que ARIMA. Además, la estrategia

Métodos		Estadísticas descriptivas de algunos métodos			
Agrupamiento	Predicción	Media	σ	Máximo	Mínimo
DTW Precios	RN	0,15 %	2,30 %	21,13 %	-26,74 %
EROS Ratios	RN	0,14 %	2,29 %	22,04 %	-26,74 %
DTW Precios	ARIMA	0,08 %	1,78 %	22,04 %	-16,18 %
EROS Ratios	ARIMA	0,10 %	1,81 %	22,04 %	-13,92 %
-	Azar	-0,01 %	1,65 %	22,04 %	-17,45 %

Cuadro 4.11: Estadísticas descriptivas de los rendimientos obtenidos utilizando los dos métodos de mejor desempeño para cada técnica de pronóstico y para la inversión aleatoria.

aleatoria, como era de esperar, tiene la desviación estándar más baja de todos los métodos y una media muy cercana al 0 %.

Estos resultados nos permitieron definir y optimizar un conjunto de algoritmos y estrategias que, en forma eficiente, pudieran procesar los datos de las 2359 empresas y de esta forma obtener el objetivo deseado de probar estas estrategias sobre la mayoría del mercado de acciones de Estados Unidos. Como veremos en el próximo capítulo, los resultados confirman los ya encontrados en esta etapa, pero con un volumen que nos permite ya reafirmar muchas de los hallazgos encontrados en la etapa de exploración.

Capítulo 5

Conjunto completo

Finalizado el trabajo exploratorio, emprendimos el trabajo con las 2359 acciones del conjunto completo (CD, tal como fuera descrito en 3). Considerando los resultados del piloto, tomamos los cuatro mejores métodos de agrupación utilizados en la etapa de exploración con el piloto con ambos métodos de predicción, ARIMA y Redes Neuronales, con el objetivo de validar los hallazgos del capítulo anterior. Esto significó el descartar las pruebas con todo el conjunto de datos sin discriminar, que habían resultado malas tanto para ARIMA como para Redes Neuronales, descartar las pruebas de predicción de rendimiento, que habían sido inferiores a las de las predicciones de precios y descartar la agrupación de series de rendimientos diarios con Fourier, también dada la pobre performance del mismo. Para este conjunto, entonces solo utilizaremos modelos entrenados a partir del agrupamiento basado en las series de ratios financieros con DTW y EROS como distancias y las series de precios de cierre con DTW y euclidiana usando los coeficientes de las transformadas de Fourier como distancias, en todos los casos usando K-Means o K-Medoids como métodos de agrupamiento.

Aplicamos todas las etapas definidas en la figura 4.12, es decir, el preprocesamiento de los datos, el agrupamiento, la generación de las predicciones (incluyendo el Simple y el Múltiple (intradiario)) con los cuatro métodos de agrupamiento ensayados y por último con estas predicciones ejecutamos el algoritmo de inversión desarrollado en el capítulo anterior.

A continuación describiremos los detalles de cada una de las fases ejecutadas.

5.1. Procesamiento de datos y agrupamiento

Con el conjunto de datos generamos las series de ratios financieros y los agrupamos de cuatro formas distintas, dos para las series de ratios financieros, utilizando DTW+K-Means y EROS+K-Medoids. En el caso de las series de precios, tomamos solo los precios de cierre y los agrupamos mediante DTW+K-Means y Fourier+K-Means. Como en el caso anterior, hallamos los óptimos mediante el mismo método que en el paso anterior, calculando el coeficiente de Silhouette y seleccionando un valor de k que muestre un máximo local y una cantidad adecuada de grupos, en este caso aumentando la cantidad de los mismos.

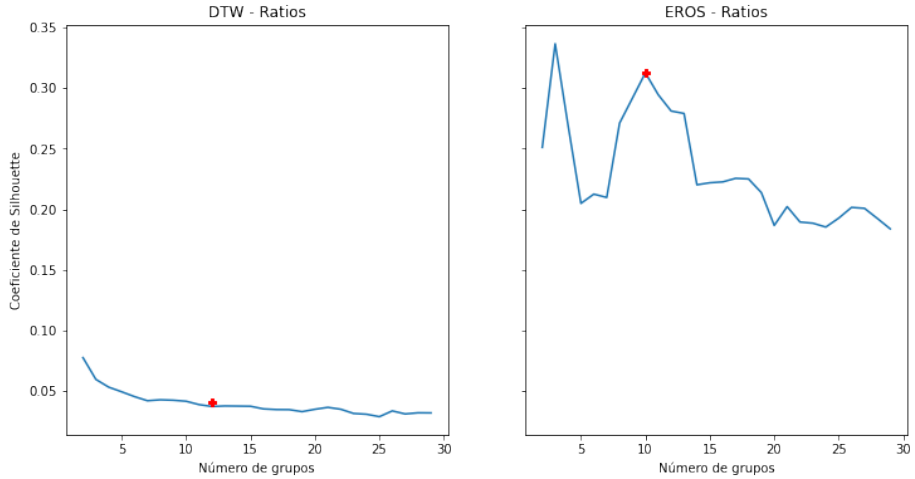

Figura 5.1: Valores de Silhouette para DTW+K-Means y EROS+K-Medoids para distintas cantidades de grupos para el agrupamiento de los Ratios Financieros.

En el caso de las series de ratios financieros encontramos dichos valores adecuados en 12 grupos para DTW y en 10 grupos para EROS, como puede verse en la figura 5.1.

Podemos ver en la matriz de correspondencia entre los grupos de ambos métodos, que nuevamente hallan estructuras distintas entre ellas, como se muestra en la figura 5.2. La medida de Rand ajustada entre ambos agrupamientos es de 0,022, indicando que no hay una correlación alta entre ellas, como podemos ver en la matriz.

Para el caso de DTW, en las series de precios, basados en los experimentos previos, tomamos un muestreo semanal para reducir los tiempos de búsqueda de los coeficientes de Silhouette que indicaran un valor del coeficiente y de número de grupos adecuado. Dicho valor fue tomado en 8 grupos y el proceso tomo poco más de 8 hs. En comparación, el proceso completo hubiera llevado casi tres semanas para ejecutarse.

La ejecución con Fourier no fue tan clara dado que los máximos interesantes se encontraban con valores bajos de grupos. Por ello optamos por un máximo local en 7 grupos, y estudiar si la cantidad de grupos afectaba el rendimiento al particionar aún más la base de datos en las predicciones.

La matriz de correspondencia entre ambos métodos de agrupamiento muestran también una selección de características de las series distinta, permitiendo ensayar diferentes puntos de vista, como puede verse en la figura 5.4. Si bien los grupos están más concentrados que en el caso anterior, se observa un comportamiento similar al hallado durante el piloto. Al calcular la medida de Rand ajustada para estos dos conjuntos, obtenemos un valor de 0,261, mayor al obtenido para las series de ratios financieros.

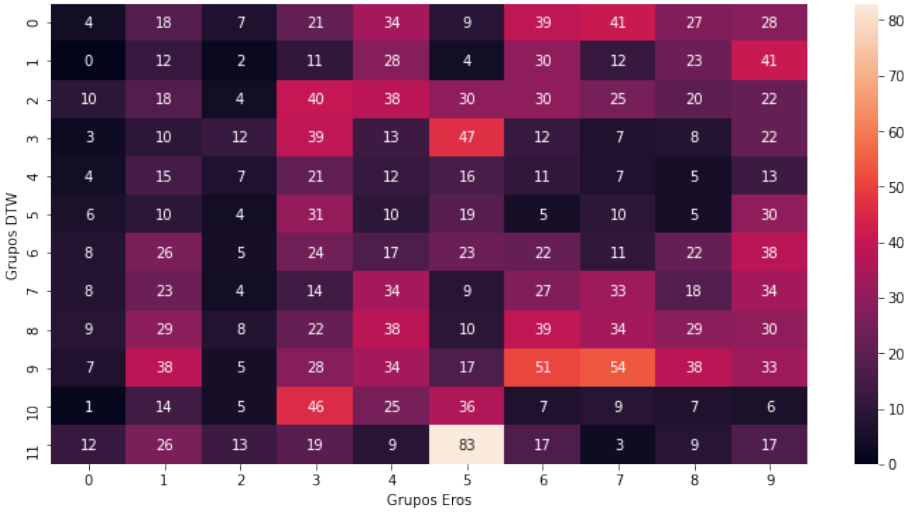

Figura 5.2: Matriz de correspondencia entre las agrupaciones de series de ratios financieros obtenidas mediante DTW+K-Means y EROS+K-Medoids.

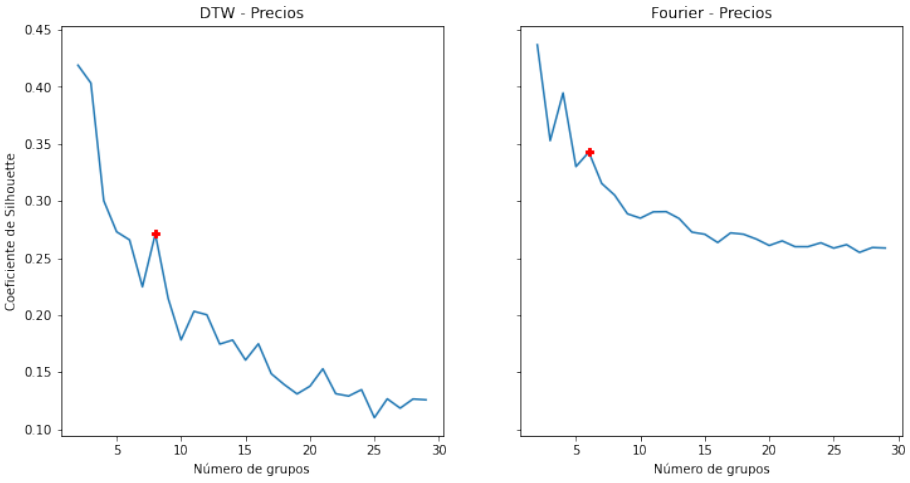

Figura 5.3: Valores de Silhouette para DTW+K-Means y Fourier+K-Means para distintas cantidades de grupos para el agrupamiento de los precios de cierre.

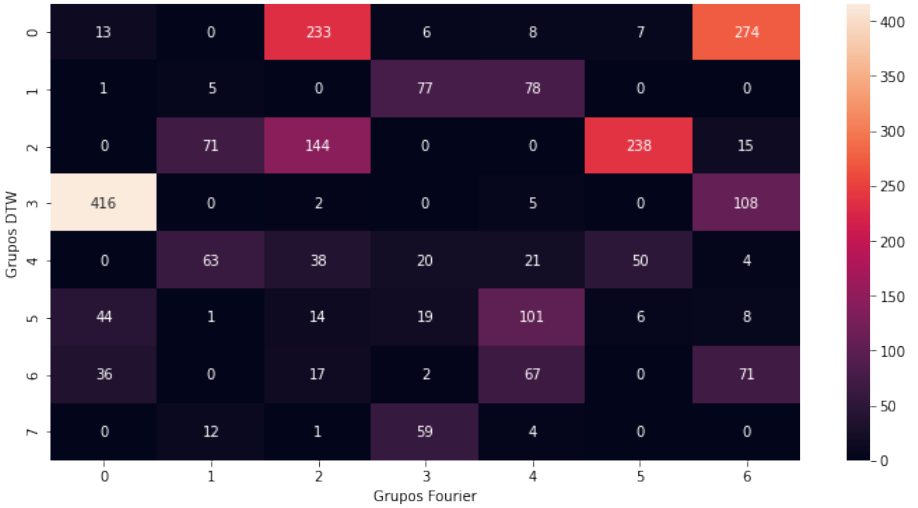

Figura 5.4: Matriz de correspondencia entre la agrupación de series de precios obtenida mediante DTW+K-Means y Fourier+K-Means.

5.2. Predicción

Con estos cuatro agrupamientos obtenidos continuamos con la siguiente tarea, la predicción de los valores finales. Para ello ejecutamos siete modelos en ARIMA y otros seis en Redes Neuronales, representando los casos Simple, Múltiple con los datos propios de la acción y Múltiple con los datos de las acciones de cada grupo con los precios de cierre (solo para ARIMA dada la disponibilidad de equipamiento y el coste computacional). Descartamos, por los resultados obtenidos, los modelos basados en rendimientos o en colocar todos los datos de las acciones.

Un problema que tuvimos para correr estos trece modelos fue el tiempo que llevaba cada corrida, de entre 4 y 6 días cada uno. Esto nos obligó a utilizar, además de la PC que utilizamos en el piloto, 3 máquinas virtuales con CPUs para correr ARIMA en paralelo y 3 máquinas virtuales con TPUs para los modelos basados en Redes Neuronales, ambos en Google Cloud y este último a través del TPU Research Cloud [1] . De esta forma pudimos reducir el tiempo de ejecución de todos los modelos a poco menos de 10 días.

Los resultados respecto al ajuste de todos los métodos pueden verse en la figura 5.1. Al contrario de lo observado con el piloto, el ajuste logrado con Redes Neuronales es mejor que con ARIMA, dado que los valores de sMAPE y MASE son menores en el primero que en el último y para Redes Neuronales aún menores que los obtenidos en el piloto. Una razón posible para este comportamiento pudo ser que una mayor cantidad de datos exógenos perjudica a ARIMA que no pudo procesarlos adecuadamente, mientras que los modelos con Redes Neuronales tuvieron una mayor capacidad para procesar estos datos eficientemente.

[1] https://sites.research.google/trc/

	Predicciones			
	ARIMA		Redes Neuronales	
Método	sMAPE	MASE	sMAPE	MASE
Simple	2,122	1,011	5,205	2,694
Múltiple (intradiario)	29,240	19,818	-	-
Múltiple (DTW Ratios)	17,764	10,228	9,805	4,885
Múltiple (EROS Ratios)	24,585	17,876	9,745	4,844
Múltiple (DTW Precios)	21,191	11,458	9,502	5,330
Múltiple (Fourier Precios)	15,022	8,469	9,458	4,762

Cuadro 5.1: Resultados de los métodos de pronóstico aplicados sobre precios y rendimientos con el valor de sMAPE y MASE de cada pronóstico generado con una red neuronal. Los últimos 4 métodos en cada sección de la tabla pertenecen a Múltiple (Acciones en grupo) nombrados como el método de agrupación, indicando la combinación diferente de datos/distancia utilizada.

5.3. Resultado de inversión

Con estas predicciones, ejecutamos nuevamente el algoritmo de inversión desarrollado en el piloto 4.3 y obtuvimos los resultados que pueden verse en las tablas 5.2 y 5.3. Como puede verse, los porcentajes de acierto están nuevamente en el orden del 50 %, no mejores que los obtenidos mediante el azar. Nuevamente podemos observar un ligero sesgo a favor de Redes Neuronales, así como una mayor cantidad de transacciones ejecutadas en el caso de Redes Neuronales. Sin embargo, los resultados esta vez son muy similares, rondando para todos los métodos en un 7 % para el trimestre, mostrando una baja respecto a lo obtenido en el piloto para Redes Neuronales y una mejora para ARIMA. En ambos casos estos datos siguen siendo positivos respecto a los otros métodos tradicionales. El comportamiento de las ganancias es muy similar al observado en el piloto, con un aprovechamiento de la fuerte caída de los mercados de valores con el comienzo de la guerra entre Rusia y Ucrania para obtener una ganancia de un 3 % adicional.

Podemos ver en la figura 5.5 la evolución día a día del rendimiento de los principales métodos contra el caso simple y una prueba al azar.

5.4. Análisis de los resultados

Como podemos ver en la tabla 5.3 y a diferencia de lo detectado en el piloto, como puede verse en la tabla 4.7, no se notan grandes diferencias entre ARIMA y Redes Neuronales en el promedio de rendimientos, siendo los mismos muy parecidos como puede verse en la tabla 5.4 y un test de Student no puede diferenciar ambos como distintos con un p-valor de 0,3981.

Un resultado más interesante es el que puede verse cuando analizamos los rendimientos por industria o tamaño de la empresa. Para ello dividimos las empresas según alguno de los 11 tipos definidos por Bloomberg o en grupos de 236 empresas ordenadas según su valor de mercado, promediando los rendimientos

	Predicciones			
	ARIMA		Redes Neuronales	
Método	# Transacciones	% Aciertos	# Transacciones	% Aciertos
Simple	67883	40,0 %	136791	49,3 %
Múltiple (intradiario)	126216	49,0 %	-	-
Múltiple (DTW Ratios)	117437	47,3 %	141120	49,8 %
Múltiple (EROS Ratios)	128225	48,9 %	141157	49,9 %
Múltiple (DTW Precios)	123698	48,6 %	140999	50,3 %
Múltiple (Fourier Precios)	115113	47,3 %	140988	50,1 %

Cuadro 5.2: Resultados de los métodos de pronóstico aplicados sobre precios, mostrando el número de operaciones ejecutadas y el porcentaje de pronósticos exitosos. Los últimos 4 métodos en cada sección de la tabla pertenecen a Múltiple (Acciones en grupo) nombrados como el método de agrupación, indicando la combinación diferente de datos/distancia utilizada.

Métodos de base	Rendimiento	
Predicción perfecta	271,73 %	
Compra al azar	1.28 %	
Buy and Hold	-4,13 %	
MACD	-5,42 %	
Métodos de predicción	ARIMA	Redes Neuronales
Simple	3,61 %	5,37 %
Múltiple (intradiario)	7,37 %	-
Múltiple (DTW Ratios)	7,14 %	6,50 %
Múltiple (EROS Ratios)	6,86 %	6,53 %
Múltiple (DTW Precios)	7,34 %	7,41 %
Múltiple (Fourier Precios)	7,58 %	7,37 %

Cuadro 5.3: Rendimiento de tres meses por método de pronóstico. Los dos primeros métodos corresponden a la inversión perfecta y a la inversión al azar, los dos segundos a métodos de inversión populares y los ocho últimos a las previsiones que generamos a través de ARIMA y Redes Neuronales, utilizando los datos propios de cada acción o una selección de datos basados en la selección de conglomerados.

Método	μ	σ	Cantidad
ARIMA	1,0723	0,2185	9436
Redes Neuronales	1,0695	0,2363	9436

Cuadro 5.4: Promedio de rendimientos de cada método de predicción, consolidando todos los métodos de agrupamiento para el mismo.

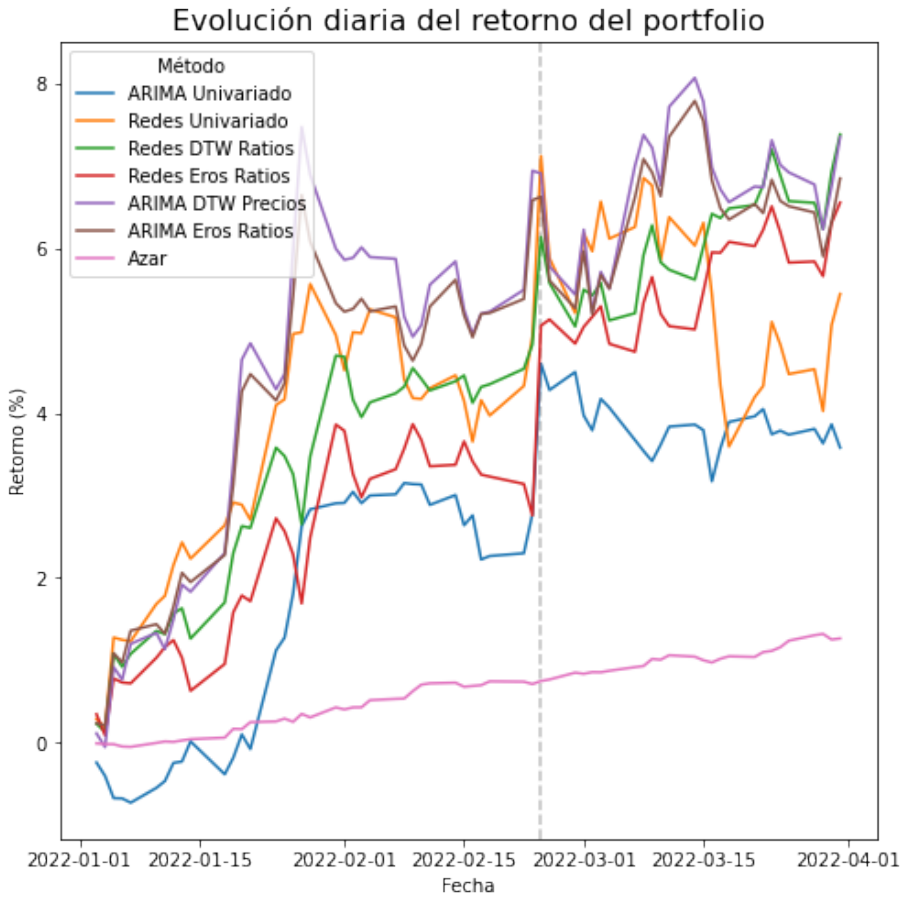

Figura 5.5: Evolución de los rendimientos para dos métodos de agrupamiento seleccionados contra el caso Simple y la inversión al azar. Con una línea vertical punteada se indica la fecha de la invasión rusa a Ucrania.

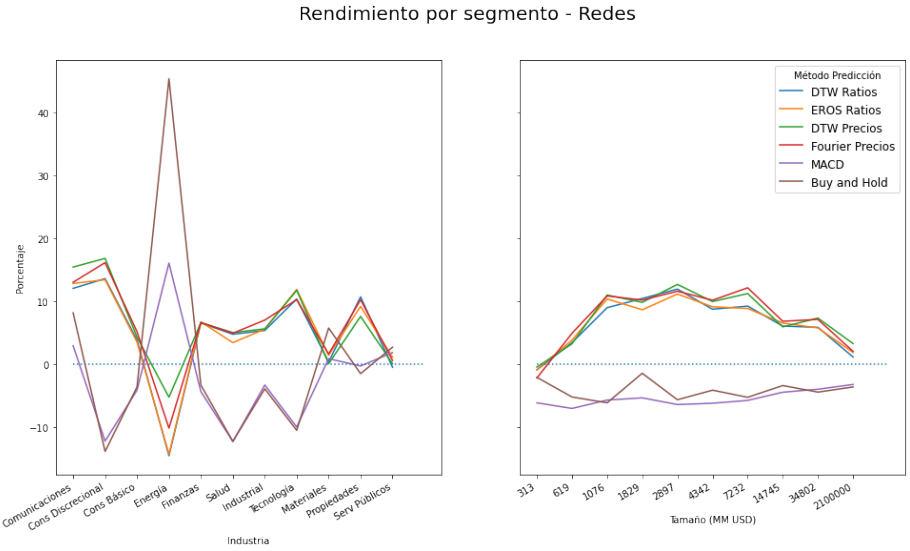

Figura 5.6: Rendimiento promedio para cada método de clustering y datos con predicción a través de Redes Neuronales comparado contra MACD y Buy and Hold para distintos grupos de industria y tamaño.

de cada grupo, como puede verse en las figuras 5.6 y 5.7. El sector de energía fue muy afectado por el inicio de la guerra de Ucrania y el incremento de los precios del petróleo y gas natural, donde Rusia es uno de los mayores productores. Esto obviamente disparó el valor de las empresas y el pico puede verse claramente obteniendo Buy and Hold más de un 40 % de ganancia en dicho segmento, el cual no pudo ser reconocido por Redes Neuronales, pero si por ARIMA.

El otro efecto de interés es la característica de Redes Neuronales de obtener mejores rendimientos en los segmentos medios de valor de las empresas, pero no en los extremos. Creemos que esto se debe a que en el caso de las empresas grandes, existe una gran cantidad de información disponible, recortando la capacidad del método de explotar eficientemente asociaciones o fuentes de datos distintas, mientras que en los segmentos de empresas más pequeñas, tienen una sensibilidad muy grande a eventos inesperados que no pueden ser capturados o modelizados adecuadamente. En el caso de ARIMA, este efecto no es tan notorio, siendo parejo el rendimiento en segmentos de valor medios y bajos disminuyendo en segmentos de valor alto.

.

Realizamos también sendos análisis sobre la apertura por tamaño y tipo de industria, obteniendo los siguientes resultados para Redes Neuronales (figura 5.8), ARIMA (figura 5.9) y como comparación Buy and Hold (figura 5.10), para ver en más detalle los efectos que se ven en los gráficos anteriores. Para Redes Neuronales notamos que el efecto de la pérdida registrada en energía se concentra más en los segmentos medios a grandes, mientras que en el resto de los segmentos, las ganancias son buenas o muy buenas para Tecnología, Comunicaciones y Consumo Discrecional.

ARIMA, sin embargo, no muestra el problema con energía, siendo más balanceado, pero coincidiendo en que el mejor pronóstico se registra en el segmento

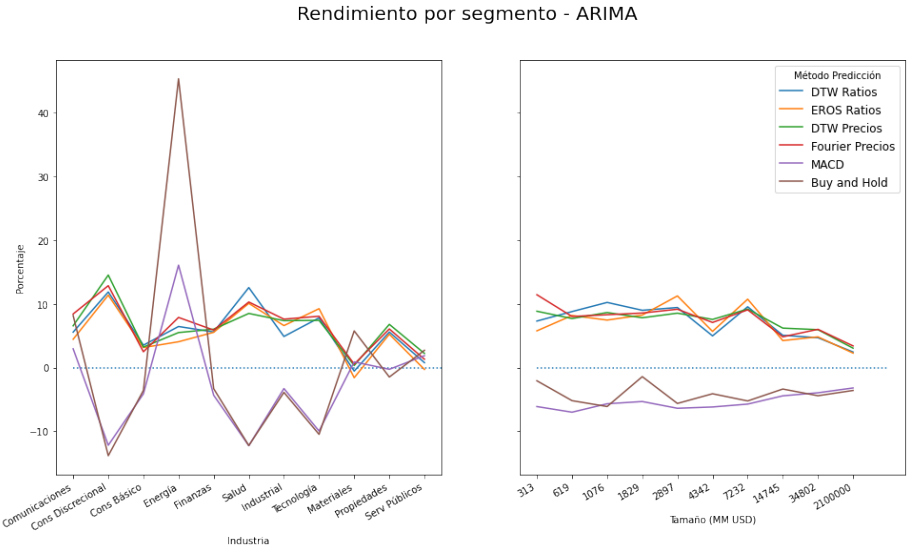

Figura 5.7: Rendimiento promedio para cada método de clustering y datos con predicción a través de ARIMA comparado contra MACD y Buy and Hold para distintos grupos de industria y tamaño.

mediano. Compárese esto con el excelente rendimiento, por las razones antedichas en Buy and Hold donde logra un 45 % de retorno en energía y pérdidas en la mayor parte de las industrias restantes, con la excepción de comunicaciones donde el segmento pequeño tuvo un incremento muy alto o en materiales.

Por último, si vemos un histograma de rendimientos para Redes Neuronales y ARIMA (figuras 5.11 y 5.12), vemos que los mismos se ajustan aproximadamente a una curva normal aunque en el caso de ARIMA tiene un pico en 0 que resulta de un artefacto cuando el orden de ARIMA es (0,0,0) el pronóstico devuelve el precio de cierre del día anterior. En ese caso el algoritmo de inversión aconseja pasar, y por ende el rendimiento en ese caso es 0 %.

5.5. Discusión

Nuestros resultados sugieren algunas implicaciones interesantes. Primero, demasiada información puede ser tan dañina como la falta de ella. Las simulaciones empíricas muestran que cuando la información de una acción dada se complementa con la información del resto del conglomerado, los resultados de pronóstico son mejores en términos de rendimiento. De hecho, superan las previsiones utilizando solo información de acciones individuales. Esto indica que la selección de acciones similares agrega información útil a los métodos de pronóstico. Por el contrario, usar información para todas las acciones agrega más ruido e información contradictoria que perjudica la capacidad de pronóstico. Este hallazgo desafía la sabiduría convencional de que alimentar una red neuronal con más datos ayuda a la predicción.

Por otro lado, los buenos resultados de rendimiento (tabla 5.3) están en aparente conflicto con los resultados de precisión (tabla 5.2). Ofrecemos la siguiente

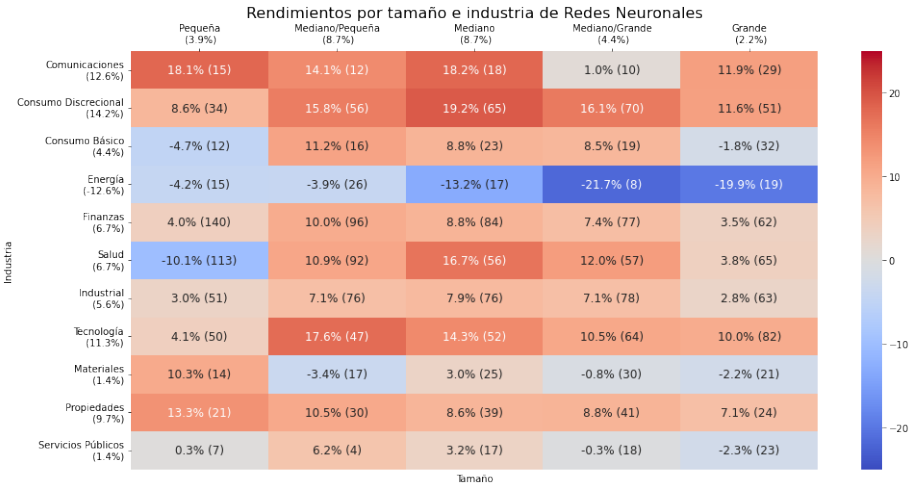

Figura 5.8: Rendimiento promedio consolidando todos los tipos de datos y distancias abiertos por tamaño e industria para la predicción con Redes Neuronales.

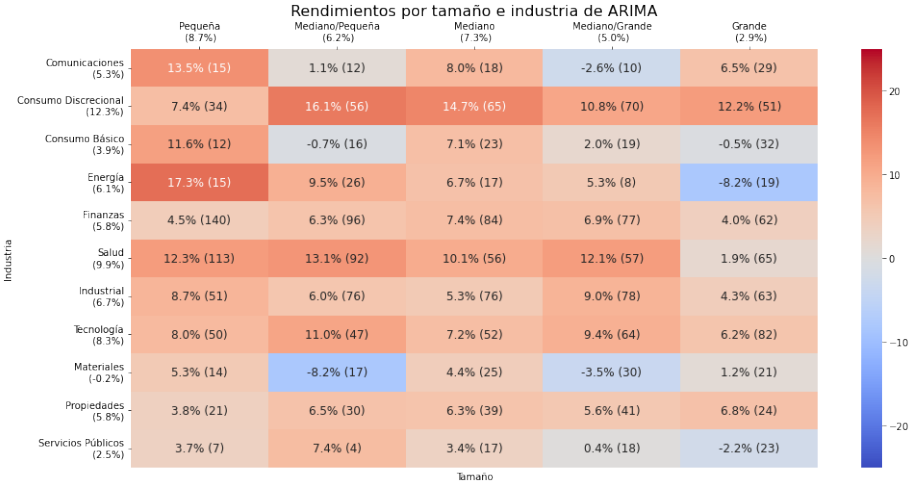

Figura 5.9: Rendimiento promedio consolidando todos los tipos de datos y distancias abiertos por tamaño e industria para la predicción con ARIMA.

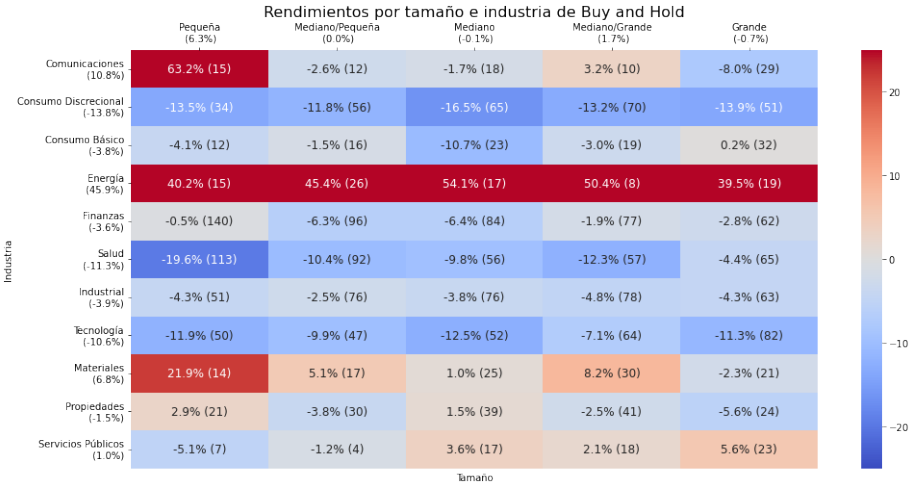

Figura 5.10: Rendimiento promedio por tamaño y tipo de industria para Buy and Hold.

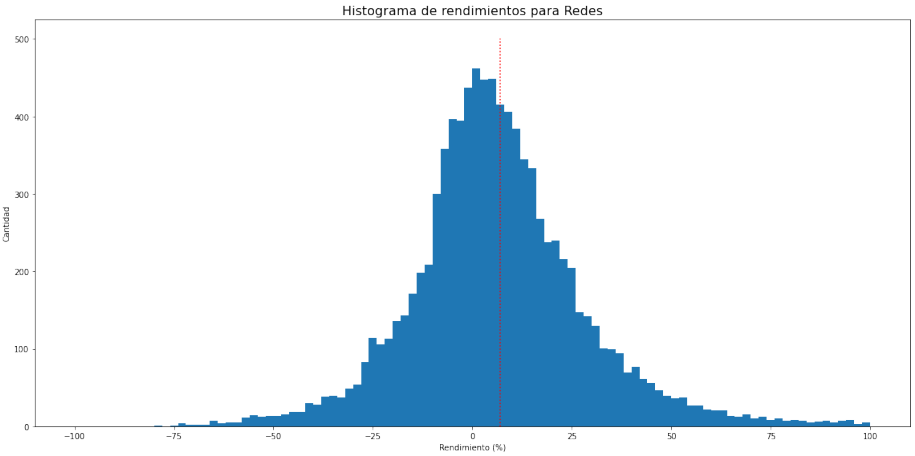

Figura 5.11: Histograma de rendimientos consolidados para Redes Neuronales. En rojo el rendimiento promedio.

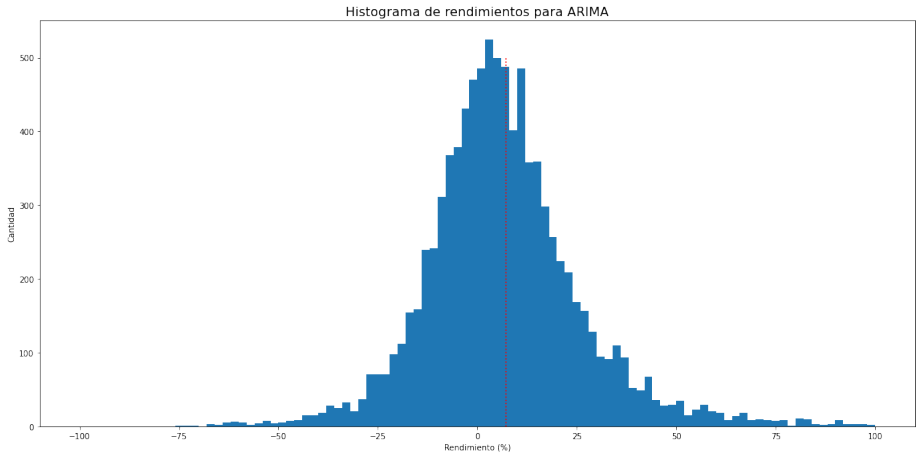

Figura 5.12: Histograma de rendimientos consolidados para ARIMA. En rojo el rendimiento promedio.

explicación para esta paradoja: los métodos de pronóstico (especialmente las Redes Neuronales) son buenos para activar la señal correcta para las transacciones de mayor rendimiento, mientras que en los días con rendimientos bajos, la señal pronosticada suele ser incorrecta. Como se puede observar en la tabla 4.8, el rendimiento medio de las ganancias es mayor que el rendimiento medio de las pérdidas (significativo al nivel del 1 %), excepto en la compra aleatoria, como era esperable.

Una conclusión adicional de los resultados que se muestran en las tablas 4.8 y 5.3 es el número de transacciones ejecutadas por cada método. Se puede observar que los modelos basados en Redes Neuronales inducen más operaciones, mientras que los modelos ARIMA activan señales de compra o venta (venta en corto) con menos frecuencia, dado que la estrategia de inversión no halla tantas oportunidades como en el primer caso.

Con los experimentos realizados tanto en la etapa de exploración, como la de ejecución final con el conjunto completo, podemos presentar una serie de conclusiones que expondremos en el próximo capítulo.

Capítulo 6

Conclusiones

Durante el desarrollo de esta tesis exploramos diversas formas de agrupar datos de precios, rendimientos y ratios financieros para un conjunto de acciones que representa más del 95 % del valor de los mercados de valores de los Estados Unidos. Estas agrupaciones permitieron aumentar los datos de entrenamiento para predecir los precios de las acciones durante 3 trimestres diferentes, utilizando ARIMA y Redes Neuronales. Las predicciones resultantes alimentaron una estrategia de inversión comparando los resultados obtenidos contra lo efectivamente ocurrido y obteniendo en general rendimientos positivos.

Esta tesis proporciona evidencia sobre dos hipótesis importantes. Por un lado, las técnicas de aprendizaje profundo, como Redes Neuronales, generalmente superan (especialmente en precios) a los modelos ARIMA tradicionales, como pudo verse en los períodos ensayados en el piloto. Tanto para el intervalo de febrero a abril de 2020, afectado por la cuarentena debida al COVID-19, o en el período de prueba, entre enero y marzo de 2022, las Redes Neuronales tuvieron mejor desempeño que ARIMA. Esto podría explicarse por la capacidad de las Redes Neuronales para lidiar con las no linealidades en los datos financieros. Además, refleja en un mayor número de señales comerciales desencadenadas por métodos de Redes Neuronales y la ganancia promedio obtenida en cada operación. Por otro lado, las diferentes pruebas indican que no todos los datos son útiles en un método de pronóstico. La agrupación de información bursátil proporciona información relevante que complementa los datos específicos de una acción determinada, mejora el poder de pronóstico y permite obtener mayores rendimientos que utilizando solo datos específicos de la población. En este aspecto, agrupar tanto los datos de los informes contables trimestrales como los precios (o, alternativamente, los rendimientos) agrega conocimiento al modelo de pronóstico. Por el contrario, agregar información (tanto de informes financieros como de precios/rendimientos) de todas las acciones de la muestra perjudica la capacidad de predicción tanto de las Redes Neuronales como de ARIMA. Esto significa que más datos no van necesariamente acompañados de mejores pronósticos. Probablemente, los datos de acciones no relacionadas agregan ruido y signos contradictorios que reducen la precisión del pronóstico y el rendimiento en el conjunto de prueba. En otras palabras, la información relacionada con la acción (proporcionada por el grupo al que pertenece) se desdibuja al agregar datos no relacionados. Como tal, la predicción está más cerca del comportamiento medio del mercado que del comportamiento de una acción específica.

Como era de esperar, por intuición económica y sentido común, solo los datos relacionados brindan información útil para el proceso de pronóstico. Algunos de estos datos relacionados, invisibles para el análisis habitual, emergen a través del proceso de agrupación, lo que demuestra que este método es un paso de preprocesamiento útil. En definitiva, más datos no implican mejores resultados; la calidad en lugar de la cantidad es la clave aquí. Estos puntos son buenas noticias, ya que como consecuencia práctica, menos datos para procesar se traduce en una reducción de los tiempos de proceso computacional.

Estos resultados están en línea con la hipótesis del mercado adaptativo de Lo (2004) porque permiten encontrar algunas de estas desviaciones temporales basadas en el comportamiento disímil de acciones con comportamiento parecido y explotar estas para obtener un beneficio.

El último punto que debe mencionarse es que la explotación de los datos en los informes financieros a través de la creación de series temporales y el posterior agrupamiento proporciona los medios para revelar asociaciones ocultas.

Nuestro estudio tiene algunas limitaciones, principalmente en la selección del conjunto de datos, incluyendo los mercados de valores, los períodos temporales y el conjunto de acciones elegidos, lo que podría hacer que los resultados parezcan específicos del mercado y/o del tiempo.

Teniendo en cuenta tales limitaciones, proponemos en futuros trabajos explorar los siguientes puntos:

- Ampliar el período de prueba para validar que los resultados siguen siendo positivos, utilizando distintos períodos con características disimiles, ejemplo 2020 con el impacto producido por la pandemia con una fuerte caida y recuperación, 2021 como un año de crecimiento y recuperación luego de la pandemia o el primer trimestre de 2023.

- Evaluar la sensibilidad del método con respecto al sector, considerando sectores más estables como los de Consumo Masivo o Finanzas y más volátiles como Energía o Tecnología. Asimismo el impacto de los tamaños de las empresas separándolas en las grandes, medianas y pequeñas, ya que encontramos cierta evidencia de un comportamiento distinto para los métodos ensayados.

- Complementar el método con datos alternativos, como el análisis de sentimientos, o los datos de los informes financieros, con la evolución de los precios de las materias primas o las variables macroeconómicas, para refinar aún más la agrupación.

- Trabajar con otros mercados y otros instrumentos financieros.

- Diseñar modelos ARIMA o de Redes Neuronales mejorados para lidiar con la gran cantidad de información que significa usar todas las acciones.

- Desarrollar aplicaciones que puedan capturar datos online e ir indicando las predicciones en tiempo real.

Apéndice A

Estructura de los Reportes Trimestrales

En esta sección colocaremos aquellos datos que complementan el trabajo realizado o que pueden ser utilizados como referencia a futuro, pero cuya inclusión en el trabajo principal hubiera desviado la atención del lector.

Los siguientes son los datos proporcionados por la API de AlphaVantage [1] con su consiguiente descripción.[2]

A.1. Datos de Balances:

Estos datos reflejan los datos principales del balance.

Depreciación acumulada, agotamiento y amortización de las propiedades, plantas y equipos: Depreciación acumulada, agotamiento y amortización de los activos físicos utilizados en la conducción normal del negocio para producir bienes y servicios.

Arrendamiento financiero, pasivo no corriente: Valor presente de la obligación descontada del arrendatario por los pagos de arrendamiento del arrendamiento financiero, clasificado como no corriente.

Efectivo y equivalentes de efectivo a valor en libros: Cantidad de moneda disponible, así como depósitos a la vista en bancos o instituciones financieras. Incluye otros tipos de cuentas que tienen las características generales de los depósitos a la vista. También incluye inversiones a corto plazo altamente líquidas que son fácilmente convertibles en montos conocidos de efectivo y están tan cerca de su vencimiento que presentan un riesgo insignificante de cambios en el valor debido a cambios en las tasas de interés. Excluye efectivo y equivalentes de efectivo dentro del grupo de disposición y operación discontinuada.

Efectivo, equivalentes de efectivo e inversiones a corto plazo: El efectivo incluye el efectivo en caja, así como los depósitos a la vista en bancos o instituciones financieras. También incluye otros tipos de cuentas que tienen las características generales de los depósitos a la vista en que el cliente puede depositar fondos adicionales en cualquier momento y puede retirar fondos en cualquier momento sin previo aviso o sanción. Los equivalentes de efectivo, excluyendo los elementos clasificados como valores negociables, incluyen inversiones a corto plazo altamente líquidas que son fácilmente convertibles en montos

[1] https://alphavantage.co

[2] https://documentation.alphavantage.co/FundamentalDataDocs/index.html

conocidos de efectivo y están tan cerca de su vencimiento que presentan un riesgo mínimo de cambios en el valor debido a cambios en las tasas de interés. Generalmente, solo las inversiones con vencimientos originales de tres meses o menos califican bajo esa definición. Vencimiento original significa vencimiento original para la entidad que posee la inversión. Por ejemplo, tanto una letra del Tesoro de Estados Unidos a tres meses como una nota del Tesoro a tres años comprada a tres meses del vencimiento califican como equivalentes de efectivo. Sin embargo, una nota del Tesoro comprada hace tres años no se convierte en equivalente de efectivo cuando su vencimiento restante es de tres meses. Las inversiones a corto plazo, excluyendo los equivalentes de efectivo, generalmente consisten en valores negociables destinados a ser vendidos dentro de un año (o el ciclo operativo normal si es más largo) y pueden incluir valores negociables, valores disponibles para la venta o mantenidos hasta el vencimiento o valores (si vencen dentro de un año), según corresponda.

Valor a la par o declarado de las acciones ordinarias: Valor total a la par o declarado de las acciones ordinarias no redimibles emitidas (o acciones ordinarias redimibles únicamente a opción del emisor). Este concepto incluye acciones propias recompradas por la entidad. Nota: los elementos de número de acciones ordinarias no redimibles, valor nominal y otros conceptos de divulgación se encuentran en otra sección dentro del capital contable.

Cantidad estimada de acciones ordinarias en circulación: Mejor estimación de acciones ordinarias en circulación. Si una empresa no informa un valor de fin de período, el monto predeterminado es el valor informado para el promedio ponderado de acciones en circulación trimestral sin dilución.

Cuentas por pagar corrientes: Valor en libros a la fecha del balance general de los pasivos incurridos (y por los que normalmente se han recibido facturas) y pagaderos a proveedores por bienes y servicios recibidos que se utilizan en el negocio de una entidad. Se utiliza para reflejar la parte actual de los pasivos (que vencen dentro de un año o dentro del ciclo operativo normal si es más largo).

Deuda corriente: Cantidad de deuda a corto plazo y vencimiento actual de deuda a largo plazo y obligaciones de arrendamiento de capital con vencimiento dentro de un año o el ciclo operativo normal, si es más largo.

Deuda a largo plazo: Importe, después de la prima no amortizada (descuento) y los costes de emisión de la deuda, de la deuda a largo plazo, clasificada como corriente. Incluye, entre otros, documentos por pagar, bonos por pagar, obligaciones, préstamos hipotecarios y papel comercial. Excluye obligaciones de arrendamiento de capital.

Cuentas por cobrar corriente: El monto total adeudado a la entidad dentro de un año a partir de la fecha del balance general (o un ciclo operativo, si es más largo) de fuentes externas, incluidas las cuentas comerciales por cobrar, pagarés y préstamos por cobrar, así como cualquier otro tipo de cuentas por cobrar, neto de provisiones, establecido con el propósito de reducir dichas cuentas por cobrar a una cantidad que se aproxime a su valor neto realizable.

Total de la deuda de largo y corto plazo: Representa el agregado de la deuda total a largo plazo, incluidos los vencimientos actuales y la deuda a corto plazo.

Ingresos diferidos: Importe de los ingresos diferidos y la obligación de transferir productos y servicios al cliente por los que se ha recibido o se puede recibir una contraprestación.

Valor del fondo de comercio: Importe después de la pérdida por deterioro acumulada de un activo que represente beneficios económicos futuros derivados de otros activos adquiridos en una combinación de negocios que no se identifiquen individualmente y se reconozcan por separado.

Activos intangibles netos: Importe en libros de activos intangibles de vida finita, activos intangibles de vida indefinida y el fondo de comercio. El fondo de comercio es un activo que representa los beneficios económicos futuros que surgen de otros activos adquiridos en una combinación de negocios que no se identifican individualmente ni se reconocen por separado. Los activos intangibles son activos, sin incluir los activos financieros, que carecen de sustancia física.

Activos intangibles netos sin fondo de comercio: Suma de los valores en libros de todos los activos intangibles, excluyendo el crédito mercantil, a la fecha del balance general, neto de la amortización acumulada y los cargos por deterioro.

Inventario neto: Importe después de la valoración y las reservas LIFO del inventario que se espera vender o consumir dentro de un año o ciclo operativo, si es más largo.

Inversiones: Suma de los valores en libros a la fecha del balance general de todas las inversiones.

Deuda a largo plazo: Monto, después de la prima no amortizada (descuento) y los costos de emisión de deuda, de la deuda a largo plazo. Incluye, entre otros, documentos por pagar, bonos por pagar, obligaciones, préstamos hipotecarios y papel comercial. Excluye obligaciones de arrendamiento de capital.

Deuda a largo plazo, excluyendo vencimientos corrientes: Monto después de la prima no amortizada (descuento) y los costos de emisión de deuda de la deuda a largo plazo clasificada como no corriente y excluyendo los montos a pagar dentro de un año o el ciclo operativo normal, si es más largo. Incluye, entre otros, documentos por pagar, bonos por pagar, obligaciones, préstamos hipotecarios y papel comercial. Excluye obligaciones de arrendamiento de capital.

Inversiones a largo plazo: La cantidad total de inversiones que se pretende mantener durante un período de tiempo prolongado (más de un ciclo operativo).

Otros activos corrientes: Importe de los activos circulantes clasificados como otros.

Otros pasivos corrientes: Importe de los pasivos clasificados como otros, con vencimiento dentro de un año o del ciclo normal de explotación, si fuera superior.

Otros pasivos no corrientes: Importe de los pasivos clasificados como otros, con vencimiento posterior a un año o al ciclo normal de explotación, si fuera superior.

Otros activos no corrientes: Importe de los activos no corrientes clasificados como otros.

Propiedades, plantas y equipos: Monto después de la depreciación acumulada, el agotamiento y la amortización de los activos físicos utilizados en la conducción normal del negocio para producir bienes y servicios y no destinados a la reventa. Los ejemplos incluyen, entre otros, terrenos, edificios, maquinaria y equipo, equipo de oficina y muebles y accesorios.

Ganancias retenidas (déficit acumulado): El importe acumulado de las ganancias o el déficit no distribuidos de la entidad que informa.

Deuda a corto plazo: Refleja el valor en libros total a la fecha del balance general de la deuda que tiene plazos iniciales menores a un año o el ciclo normal de operación, si es mayor.

Inversiones a corto plazo: Importe de las inversiones, incluidos valores negociables, valores disponibles para la venta, valores mantenidos hasta el vencimiento e inversiones a corto plazo clasificadas como otras y corrientes.

Activos totales: Suma de los valores en libros a la fecha del balance general de todos los activos que se reconocen. Los activos son beneficios económicos futuros probables obtenidos o controlados por una entidad como resultado de transacciones o eventos pasados.

Activos corrientes: Suma de los valores en libros a la fecha del balance general de todos los activos que se espera realizar en efectivo, vender o consumir dentro de un año (o el ciclo operativo normal, si es más largo). Los activos son beneficios económicos futuros probables obtenidos o controlados por una entidad como resultado de transacciones o eventos pasados.

Pasivos corrientes: Obligaciones totales incurridas como parte de las operaciones normales que se espera pagar durante los siguientes doce meses o dentro de un ciclo comercial, si es más largo.

Pasivos totales: Suma de los valores en libros a la fecha del balance general de todos los pasivos que se reconocen. Los pasivos son sacrificios futuros probables de beneficios económicos que surgen de las obligaciones presentes de una entidad para transferir activos o prestar servicios a otras entidades en el futuro.

Activos no corrientes: Suma de los valores en libros a la fecha del balance general de todos los activos que se espera realizar en efectivo, vender o consumir después de un año o más allá del ciclo normal de operación, si es más largo.

Pasivos no corrientes: Monto de la obligación adeudada después de un año o más allá del ciclo normal de operación, si es mayor.

Capital de los accionistas: Total de todas las partidas (déficit) del capital contable, neto de las cuentas por cobrar de funcionarios, directores, propietarios y afiliadas de la entidad que son atribuibles a la matriz. El monto del capital contable de la entidad económica atribuible a la controladora excluye el monto del capital contable que es asignable a esa participación en el capital contable de la subsidiaria que no es atribuible a la controladora (participación no controladora, participación minoritaria). Esto excluye el patrimonio temporal y, a veces, se denomina patrimonio permanente.

Valor de las acciones en tesorería: El importe destinado a acciones de tesorería. Las acciones en tesorería son acciones ordinarias y preferentes de una entidad que fueron emitidas, recompradas por la entidad y que se mantienen en su tesorería.

A.2. Flujo de Caja

Estos datos reflejan los flujos de caja anuales y trimestrales.

Gastos de capital: La salida de efectivo por compras y mejoras de capital en propiedad, planta y equipo (gastos de capital), software y otros activos intangibles.

Efectivo y equivalentes de efectivo a valor libros: Efectivo y equivalentes de efectivo, a valor en libros, saldo final Cantidad de moneda disponible, así como depósitos a la vista en bancos o instituciones financieras. Incluye otros tipos de cuentas que tienen las características generales de los depósitos a la vista. También incluye inversiones a corto plazo altamente líquidas que son fácilmente convertibles en montos conocidos de efectivo y están tan cerca de su vencimiento que presentan un riesgo insignificante de cambios en el valor debido a cambios en las tasas de interés. Excluye efectivo y equivalentes de efectivo dentro del grupo de disposición y operación discontinuada.

Efectivo neto provisto por actividades de financiamiento: Importe de la entrada (salida) de efectivo por actividades de financiación, incluidas las operaciones discontinuadas. Los flujos de efectivo de la actividad de financiamiento incluyen obtener recursos de los propietarios y proporcionarles un rendimiento y un rendimiento de su inversión; pedir dinero prestado y devolver las cantidades prestadas, o liquidar la obligación; y la obtención y pago de otros recursos obtenidos de los acreedores a crédito a largo plazo.

Flujo de efectivo de la inversión: Importe de la entrada (salida) de efectivo de las actividades de inversión, incluidas las operaciones discontinuadas. Los flujos de efectivo de las actividades de inversión incluyen otorgar y cobrar préstamos y adquirir y enajenar instrumentos de deuda o de capital y propiedad, planta y equipo y otros activos productivos.

Variación en el período del efectivo y equivalentes de efectivo: Monto del aumento (disminución) en efectivo y equivalentes de efectivo. El efectivo y los equivalentes de efectivo son la cantidad de efectivo disponible, así como los depósitos a la vista en bancos o instituciones financieras. Incluye otros tipos de cuentas que tienen las características generales de los depósitos a la vista. También incluye inversiones a corto plazo altamente líquidas que son fácilmente convertibles en montos conocidos de efectivo y están tan cerca de su vencimiento que presentan un riesgo insignificante de cambios en el valor debido a cambios en las tasas de interés. Incluye el efecto de las variaciones del tipo de cambio.

Cambios en el efectivo por la variación de tipo de cambio: Importe del aumento (disminución) del efecto de las variaciones del tipo de cambio sobre los saldos de efectivo y equivalentes de efectivo mantenidos en moneda extranjera.

Cambios en el inventario: El aumento (disminución) durante el período sobre el que se informa en el valor agregado de todo el inventario mantenido por la entidad que informa, asociado con las transacciones subyacentes que se clasifican como actividades de operación.

Cambios en activos operativos: El aumento (disminución) durante el período sobre el que se informa en la cantidad total de activos utilizados para generar ingresos operativos.

Cambios en los pasivos operativos: El aumento (disminución) durante el período sobre el que se informa en la cantidad total de pasivos que resultan de actividades que generan ingresos operativos.

Cambios en cuentas por cobrar: El aumento (disminución) durante el período sobre el que se informa en el monto total adeudado dentro de un año (o

un ciclo operativo) de todas las partes, asociado con transacciones subyacentes que se clasifican como actividades operativas.

Depreciación, agotamiento y amortización: El gasto total reconocido en el período actual que asigna el costo de los activos tangibles, activos intangibles o activos agotados a períodos que se benefician del uso de los activos.

Pago de dividendos: Salida de efectivo en forma de distribuciones de capital y dividendos a accionistas comunes, accionistas preferenciales y participaciones no controladoras.

Pago de dividendos ordinarios: Importe de la salida de efectivo en forma de dividendos ordinarios a los accionistas ordinarios de la entidad matriz.

Pago de dividendos preferidos: Importe de la salida de efectivo en forma de dividendos ordinarios a los accionistas preferentes de la entidad matriz.

Ingresos netos: La porción de la utilidad o pérdida del período, neta de impuestos a las ganancias, que es atribuible a la controladora.

Flujo de caja operativo: Importe de la entrada (salida) de efectivo de las actividades operativas, incluidas las operaciones discontinuadas. Los flujos de efectivo de la actividad operativa incluyen transacciones, ajustes y cambios en el valor no definidos como actividades de inversión o financiación.

Pagos por actividades operativas: Monto total de efectivo pagado por actividades operativas durante el período actual.

Pagos por recompra de acciones comunes: La salida de efectivo para readquirir acciones ordinarias durante el período.

Pagos por recompra de capital: La salida de efectivo para readquirir acciones ordinarias y preferidas.

Pagos por recompra de acciones preferidas: Pagos por recompra de acciones preferentes y acciones preferentes La salida de efectivo para readquirir acciones preferentes durante el período.

Ingresos por la emisión de acciones comunes: La entrada de efectivo por la aportación adicional de capital a la entidad.

Ingresos por la emisión de deuda a largo plazo y valores netos de capital: La entrada de efectivo asociada con un instrumento de valor que representa un acreedor o una relación de propiedad con el tenedor del valor de inversión con un vencimiento superior a un año o un ciclo operativo normal, si es mayor. Incluye los ingresos de (a) deuda, (b) obligaciones de arrendamiento de capital, (c) valores de capital redimibles obligatorios y (d) cualquier combinación de (a), (b) o (c).

Ingresos por la emisión de acciones preferidas: Producto de la emisión de acciones de capital que establece un dividendo específico que se paga a los accionistas antes de cualquier dividendo al accionista común, que tiene prioridad sobre los accionistas comunes en caso de liquidación y de la emisión de derechos para comprar acciones comunes a un precio predeterminado.

Ingresos por las actividades operativas: Importe total de efectivo recibido de las actividades operativas durante el período actual.

Ingresos por los reembolsos de la deuda a corto plazo: La entrada o salida neta de efectivo por préstamos que tienen un plazo inicial de reembolso dentro de un año o el ciclo operativo normal, si es más largo.

Ingresos por la recompra de acciones: La entrada o salida neta de efectivo resultante de la transacción de acciones de la entidad.

Ingresos por venta de acciones en tesorería: La entrada de efectivo procedente de la emisión de una acción de capital que ha sido previamente readquirida por la entidad.

Utilidad o pérdida neta: La utilidad o pérdida consolidada del período, neta de impuestos a la utilidad, incluyendo la porción atribuible a la participación no controladora.

A.3. Datos de Ingresos

Estos datos reflejan los ingresos anuales y trimestrales.

Ingresos netos de impuestos: Monto después de impuestos del aumento (disminución) en el patrimonio por transacciones y otros eventos y circunstancias de la utilidad neta y otro resultado integral, atribuible a la entidad controladora. Excluye los cambios en el patrimonio resultantes de las inversiones de los propietarios y las distribuciones a los propietarios.

Costo de los ingresos: El costo total de los bienes producidos y vendidos y los servicios prestados durante el período sobre el que se informa.

Costo de bienes y servicios vendidos: Los costos agregados relacionados con los bienes producidos y vendidos y los servicios prestados por una entidad durante el período sobre el que se informa. Esto excluye los costos incurridos durante el período de informe relacionados con los servicios financieros prestados y otras actividades generadoras de ingresos.

Depreciación y amortización: El gasto del período actual con cargo a las ganancias de los activos físicos de larga duración que no se utilizan en la producción y que no están destinados a la reventa, para distribuir o reconocer el costo de dichos activos durante su vida útil; o para registrar la reducción en el valor en libros de un activo intangible durante el período de beneficio de dicho activo; o para reflejar el consumo durante el período de un activo que no se utiliza en la producción.

Depreciación: El monto del gasto reconocido en el período actual que refleja la distribución del costo de los activos tangibles a lo largo de la vida útil de los activos. Incluye depreciación relacionada con la producción y no relacionada con la producción.

Ganancias antes de intereses e impuestos: La porción de la utilidad o pérdida del período, antes de impuestos a la utilidad y gastos por intereses, que es atribuible a la controladora.

Ganancias antes de intereses, impuestos, depreciación y amortización: La porción de la utilidad o pérdida del período, antes de impuestos a la utilidad, gastos por intereses y depreciación y amortización, que es atribuible a la controladora.

Utilidad bruta: Ingresos agregados menos el costo de los bienes y servicios vendidos o los gastos operativos directamente atribuibles a la actividad de generación de ingresos.

Utilidad antes de pagar impuestos: Utilidad (Pérdida) Atribuible a la Matriz antes de impuestos, Total La porción de la utilidad o pérdida del período, antes de impuestos a la utilidad, que es atribuible a la controladora.

Gasto/beneficio por el impuesto a las ganancias: Importe del gasto (beneficio) por impuesto a las ganancias corriente y gasto (beneficio) por impuesto a las ganancias diferido correspondiente a las operaciones continuadas.

Intereses y gastos de deuda: Gastos relacionados con intereses y deuda asociados con actividades financieras no operativas de la entidad.

Gastos por intereses: Monto del costo de los fondos prestados contabilizados como gastos por intereses.

Ingresos por intereses: El monto de los ingresos por intereses.

Ingresos de inversiones: Importe después de la acumulación (amortización) del descuento (prima) y los gastos de inversión de los ingresos por intereses y los ingresos por dividendos sobre valores no operativos.

Ingresos netos: La porción de la utilidad o pérdida del período, neta de impuestos a las ganancias, que es atribuible a la controladora.

Utilidad de operaciones continuas neta de impuestos: Importe después de impuestos de la ganancia (pérdida) de operaciones continuadas atribuible a la controladora.

Ingresos o gastos por intereses operativos: El importe neto de los ingresos (gastos) por intereses operativos.

Ingresos no relacionados con intereses: La cantidad total de ingresos no relacionados con intereses que pueden derivarse de: (1) honorarios y comisiones; (2) primas ganadas; (3) cargos de la póliza de seguro; (4) la venta o disposición de activos; y (5) otras fuentes no especificadas.

Gastos operativos: Costos generalmente recurrentes asociados con operaciones normales, excepto por la porción de estos gastos que pueden estar claramente relacionados con la producción e incluidos en el costo de ventas o servicios. Incluye gastos de venta, generales y administrativos.

Resultados operativos netos: El resultado neto del período de deducir los gastos operativos de los ingresos operativos.

Otros ingresos no operativos: Importe de los ingresos (gastos) relacionados con actividades no operativas, clasificados como otros.

Gastos en investigación y desarrollo: Los costos agregados incurridos (1) en una búsqueda planificada o investigación crítica dirigida al descubrimiento de nuevos conocimientos con la esperanza de que dichos conocimientos sean útiles para desarrollar un nuevo producto o servicio, un nuevo proceso o técnica, o para lograr una mejora significativa a un producto o proceso existente; o (2) para traducir los resultados de la investigación u otro conocimiento en un plan o diseño para un nuevo producto o proceso o para una mejora significativa de un producto o proceso existente, ya sea que esté destinado a la venta o al uso de la entidad, durante el período de presentación de informes imputado a la investigación y proyectos de desarrollo, incluidos los costos de desarrollo de software informático hasta el momento en que se logre la viabilidad tecnológica, y los costos asignados en la contabilización de una combinación de negocios a proyectos en proceso que se considera que no tienen un uso futuro alternativo.

Gastos de ventas, generales y administrativos: Los costos totales agregados relacionados con la venta de productos y servicios de una empresa, así como todos los demás gastos generales y administrativos. Los gastos de venta directa (por ejemplo, crédito, garantía y publicidad) son gastos que pueden vincularse directamente con la venta de productos específicos. Los gastos indirectos de venta son gastos que no pueden vincularse directamente con la venta

de productos específicos, por ejemplo, gastos de teléfono, Internet y gastos postales. Los gastos generales y administrativos incluyen salarios del personal no comercial, alquiler, servicios públicos, comunicación, etc.

Ingresos totales: Importe de los ingresos reconocidos por los bienes vendidos, los servicios prestados, las primas de seguros u otras actividades que constituyen un proceso de obtención de ingresos. Incluye, pero no se limita a, los ingresos por inversiones y por intereses antes de la deducción de los gastos por intereses cuando se reconocen como un componente de los ingresos y las ganancias (pérdidas) por ventas y operaciones.

Apéndice B

Grupos obtenidos con el piloto

Los siguientes gráficos muestran la agrupación de las series de precios con cada uno de los métodos usados.

Puede verse entonces las estructuras de grupos obtenidas en las figuras siguientes:

- Grupos obtenidos aplicando DTW y K-Means sobre las series de precios de cierre diario en la figura B.1.
- Grupos obtenidos extrayendo los coeficientes reales e imaginarios mediante transformadas rápidas de Fourier a las series de precios de cierre diarios y aplicando K-Means sobre los mismos en la figura B.2.
- Grupos obtenidos extrayendo los coeficientes reales e imaginarios mediante transformadas rápidas de Fourier a las series de rendimientos diarios y aplicando K-Means sobre los mismos en la figura B.3. Puede notarse que algunos de estos grupos están compuestos por muy pocos ejemplos.

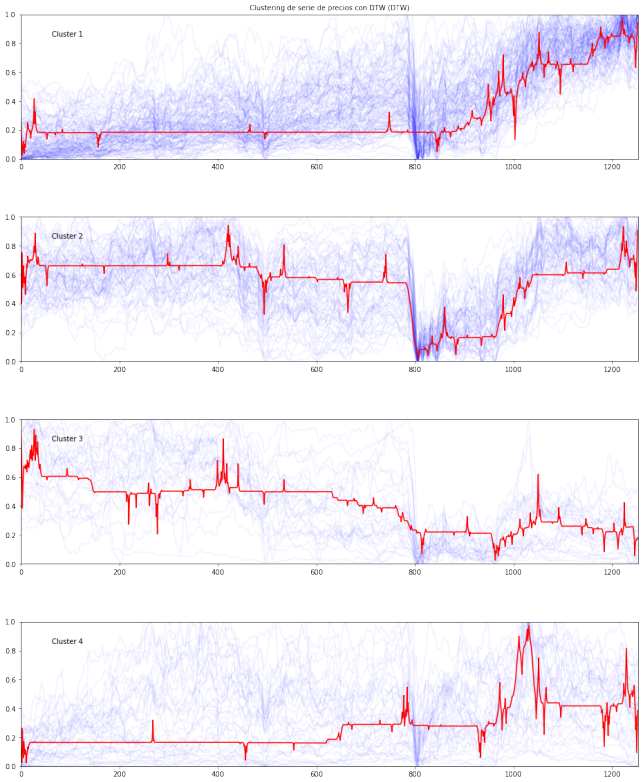

Figura B.1: Grupos obtenidos mediante DTW y K-Means de las series de precios univariadas. En rojo la curva centroide del grupo.

Figura B.2: Grupos obtenidos mediante los coeficientes de series de Fourier y K-Means de las series de precios univariadas. En rojo la serie representativa del grupo mediante el promedio de los coeficientes de Fourier de las series del grupo.

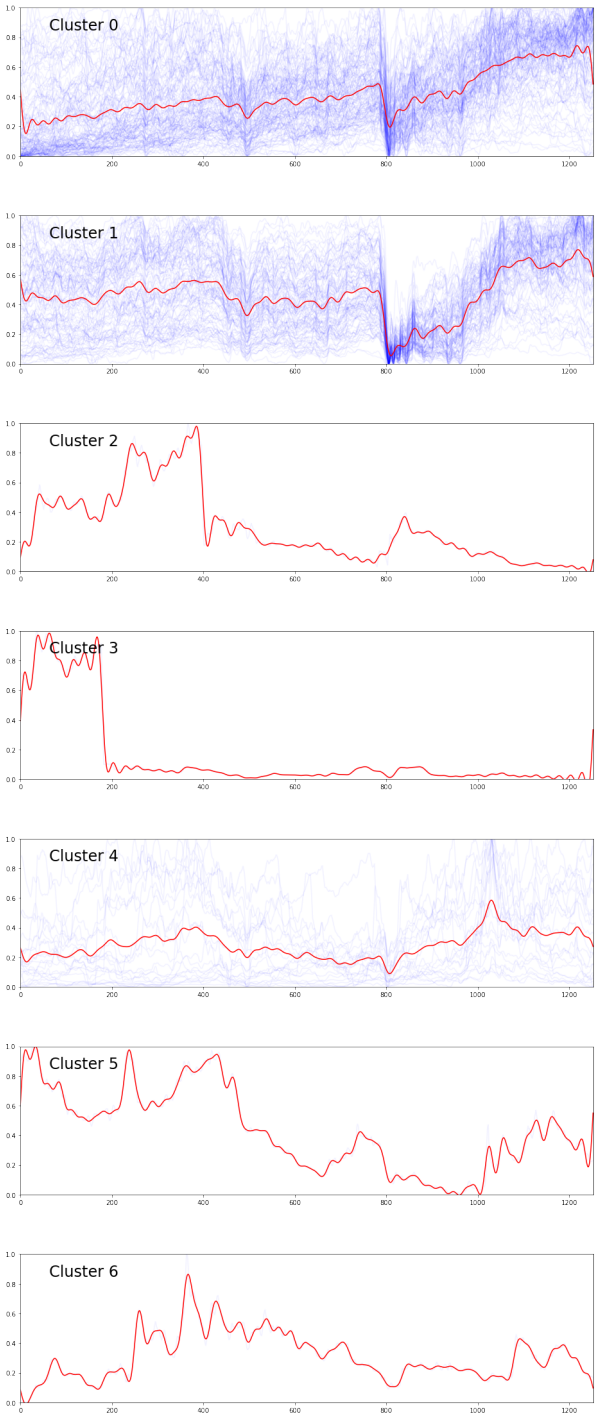

Figura B.3: Grupos obtenidos mediante los coeficientes de Fourier y K-Means de los rendimientos diarios. En rojo la serie representativa del grupo mediante el promedio de los coeficientes de Fourier de las series del grupo.

Apéndice C

Grupos obtenidos con el conjunto completo

Los siguientes gráficos muestran la agrupación de las series de precios con cada uno de los métodos usados.

Puede verse entonces las estructuras de grupos obtenidas en las figuras siguientes,

- Grupos obtenidos aplicando DTW y K-Means sobre las series de precios de cierre diario en la figura C.1.
- Grupos obtenidos extrayendo los coeficientes reales e imaginarios mediante transformadas rápidas de Fourier a las series de precios de cierre diarios y aplicando K-Means sobre los mismos en la figura C.2.
- Grupos obtenidos extrayendo los coeficientes reales e imaginarios mediante transformadas rápidas de Fourier a las series de rendimientos diarios y aplicando K-Means sobre los mismos en la figura B.3. Puede notarse que algunos de estos grupos están compuestos por muy pocos ejemplos.

Por último, si tomamos componentes principales y proyectamos los grupos obtenidos con Fourier sobre las primeras dos dimensiones, obtenemos la separación que puede verse en la figura C.3.

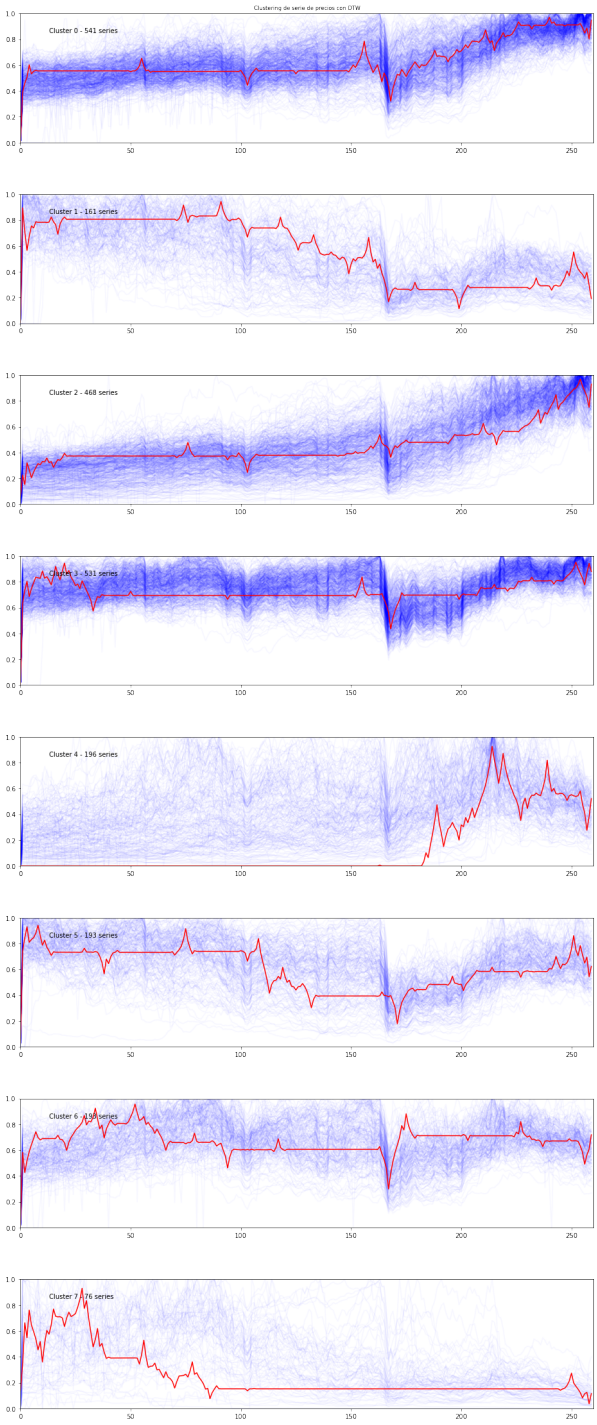

Figura C.1: Grupos obtenidos mediante DTW y K-Means de las series de precios univariadas con la cantidad de series en cada grupo. En rojo la curva centroide del grupo.

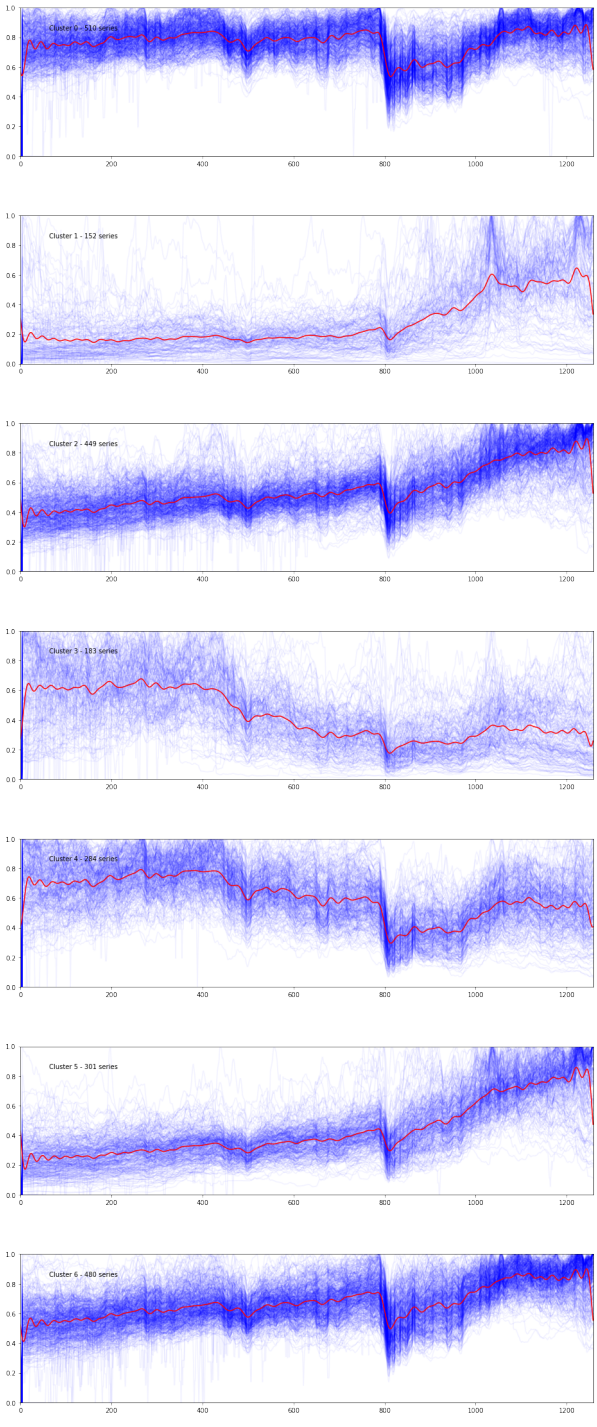

Figura C.2: Grupos obtenidos mediante los coeficientes de series de Fourier y K-Means de las series de precios univariadas con la cantidad de series en cada grupo. En rojo la serie representativa del grupo mediante el promedio de los coeficientes de Fourier de las series del grupo.

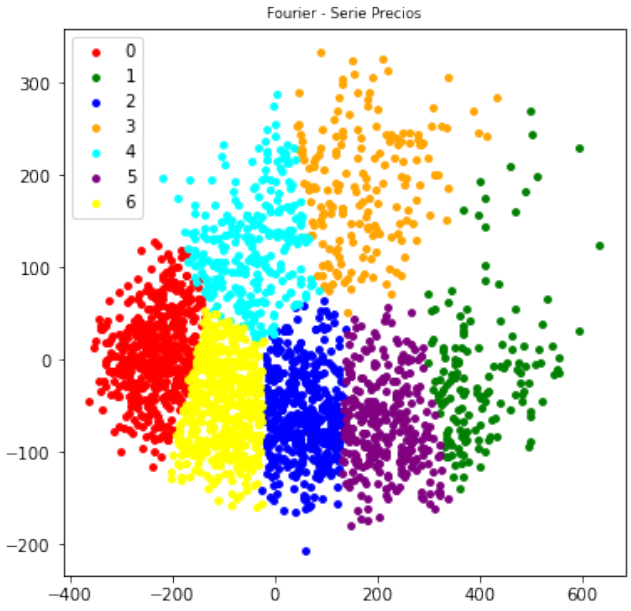

Figura C.3: Grupos obtenidos mediante los coeficientes de series de Fourier y K-Means de las series de precios univariadas proyectados sobre las primeras dos dimensiones de componentes principales.

Bibliografía

Adebiyi, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Comparison of arima and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*, 2014.

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, 53:16–38.

Agrawal, R., Faloutsos, C., and Swami, A. (1993). Efficient similarity search in sequence databases. In *International conference on foundations of data organization and algorithms*, pages 69–84. Springer.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60.

Appel, G. and Dobson, E. (2007). *Understanding MACD*, volume 34. Traders Press.

Babu, M. S., Geethanjali, N., and Satyanarayana, B. (2012). Clustering approach to stock market prediction. *International Journal of Advanced Networking and Applications*, 3(4):1281.

Bachelier, L. (1900). *Théorie de la spéculation*. Annales scientifiques de l'École Normale Supérieure, Paris.

Bagnall, A., Lines, J., Bostrom, A., Large, J., and Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660.

Bandara, K., Bergmeir, C., and Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert systems with applications*, 140:112896.

Basalto, N., Bellotti, R., De Carlo, F., Facchi, P., and Pascazio, S. (2005). Clustering stock market companies via chaotic map synchronization. *Physica A: Statistical Mechanics and its Applications*, 345(1-2):196–206.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:.

Chen, K. H. and Shimerda, T. A. (1981). An empirical analysis of useful financial ratios. *Financial management*, pages 51–60.

Dhakal, S. (2019). Reliability of clustering in forecasting stock prices of companies traded on the stock exchanges. *Merge*, 3(1):5.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2):383–417.

Fang, W.-X., Lan, P.-C., Lin, W.-R., Chang, H.-C., Chang, H.-Y., and Wang, Y.-H. (2019). Combine facebook prophet and lstm with bpnn forecasting financial markets: the morgan taiwan index. In *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 1–2. IEEE.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769.

Goodwin, P. and Lawton, R. (1999). On the asymmetry of the symmetric mape. *International journal of forecasting*, 15(4):405–408.

Hadavandi, E., Shavandi, H., and Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems*, 23(8):800–808.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):190–195.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2:193–218.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for r. *Journal of statistical software*, 27:1–22.

Kaufman, L. and Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 344:68–125.

Krauss, C., Do, X. A., and Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, 259(2):689–702.

Lee, A. J., Lin, M.-C., Kao, R.-T., and Chen, K.-T. (2010). An effective clustering approach to stock market prediction. In *PACIS*, page 54.

Li, M., Zhu, Y., Shen, Y., and Angelova, M. (2022). Clustering-enhanced stock price prediction using deep learning. *World Wide Web*, pages 1–26.

Li, X. and Wu, P. (2021). Stock price prediction incorporating market style clustering. *Cognitive Computation*, pages 1–18.

Lin, Y.-L., Lai, C.-J., and Pai, P.-F. (2022). Using Deep Learning Techniques in Forecasting Stock Markets by Hybrid Data with Multilingual Sentiment Analysis. *Electronics*, 11(21):3513.

Liu, S. and Motani, M. (2021). Towards better long-range time series forecasting using generative adversarial networks. *arXiv preprint arXiv:2110.08770*.

Lloyd, S. P. (1957). Least squares quantitization in pcm. *IEEE Transactions on Information*.

Lo, A. W. (2004). The adaptive markets hypothesis. *The Journal of Portfolio Management*, 30:15–29.

Long, J., Chen, Z., He, W., Wu, T., and Ren, J. (2020). An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in chinese stock exchange market. *Applied Soft Computing*, 91:106205.

Ma, Q. (2020). Comparison of arima, ann and lstm for stock price prediction. In *E3S Web of Conferences*, volume 218, page 01026. EDP Sciences.

Marvin, K. (2015). Creating diversified portfolios using cluster analysis. *Princeton University*.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Nair, B. B., Kumar, P. S., Sakthivel, N., and Vipin, U. (2017). Clustering stock price time series data to generate stock trading recommendations: An empirical study. *Expert Systems with Applications*, 70:20–36.

O'Connor, M. C. (1973). On the usefulness of financial ratios to investors in common stock. *The Accounting Review*, 48(2):339–352.

Osborne, M. F. M. (1959). Brownian motion in the stock market. *Operations Research*, 7(2):145–173.

Osborne, M. F. M. (1962). Periodic structure in the brownian motion of stock prices. *Operations Research*, 10(3):345–379.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Samal, K. K. R., Babu, K. S., Das, S. K., and Acharaya, A. (2019). Time series based air pollution forecasting using sarima and prophet model. In *proceedings of the 2019 international conference on information technology and computer communications*, pages 80–85.

Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6(2):41–49.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Siami-Namini, S., Tavakoli, N., and Namin, A. S. (2018). A comparison of arima and lstm in forecasting time series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1394–1401. IEEE.

Smith, T. G. et al. (2017–). pmdarima: Arima estimators for Python. [Online; accessed ¡today¿].

Tadayon, M. and Iwashita, Y. (2020). A clustering approach to time series forecasting using neural networks: A comparative study on distance-based vs. feature-based clustering methods. *arXiv preprint arXiv:2001.09547*.

Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.

Tupe, T. (2014). Financial ratio analysis for stock price movement prediction using hybrid clustering.

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics*, 4(1):52–57.

Wang, X., Yang, K., and Liu, T. (2021). Stock price prediction based on morphological similarity clustering and hierarchical temporal memory. *IEEE Access*, 9:67241–67248.

Wang, Y.-J. and Lee, H.-S. (2008). A clustering method to identify representative financial ratios. *Information Sciences*, 178(4):1087–1097.

Yamak, P. T., Yujian, L., and Gadosey, P. K. (2019). A comparison between arima, lstm, and gru for time series forecasting. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 49–55.

Yang, K. and Shahabi, C. (2004). A pca-based similarity measure for multivariate time series. In *Proceedings of the 2nd ACM international workshop on Multimedia databases*, pages 65–74.

Yenidoğan, I., Çayir, A., Kozan, O., Dağ, T., and Arslan, Ç. (2018). Bitcoin forecasting using arima and prophet. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pages 621–624. IEEE.

Índice de cuadros

Índice de figuras