UNIVERSIDAD DE GRANADA

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

TRABAJO FIN DE MÁSTER

Aplicación de técnicas de Machine Learning para el estudio, análisis, predicción y extracción de conocimiento del mercado de divisas

Presentado por:

Eilder Jorge García

Tutor:

Carlos Javier Mantas Ruiz

Curso académico 2019 /20

Aplicación de técnicas de Machine Learning para el estudio, análisis, predicción y extracción de conocimiento del mercado de divisas

Eilder Jorge García

Palabras clave: Series temporales, aprendizaje de máquina, agrupamiento, segmentación, reglas secuenciales recurrentes.

Resumen

El objetivo de este trabajo consiste en aplicar diferentes técnicas de aprendizaje de máquina para buscar reglas recurrentes predictivas en el mercado de divisas.

Inicialmente se realiza un análisis y un preprocesamiento de los datos obtenidos de Dukascopy para los 28 pares de divisas en los intervalos de una hora, diario y semanal. Posteriormente se aplica una técnica de segmentación original basada en ventana deslizante, usando diferentes técnicas de reducción de ruido.

Los segmentos obtenidos en el paso anterior son agrupados y clasificados por medio del algoritmo de las K-medias usando la distancia de Pearson entre la estructura de los mismos. Los resultados obtenidos son comparados con otras alternativas.

Finalmente se utiliza una técnica de minería de reglas secuenciales recurrentes para encontrar reglas en dichas series temporales a partir de los segmentos ya clasificados. Se buscan reglas de alta confianza que puedan predecir movimientos en el mercado una vez que se presente un determinado conjunto de segmentos de forma continua.

Estas reglas son validadas en 3 pares de divisas y con varias condiciones de ejecución para verificar la validez y calidad de las mismas. Al término de dichas validaciones se ofrecen conclusiones relevantes.

Aplicación de técnicas de Machine Learning para el estudio, análisis, predicción y extracción de conocimiento del mercado de divisas

Eilder Jorge García

Keywords: Time series, machine learning, clustering, segmentation, recurring sequential rules.

Abstract

The objective of this Project consists on applying different machine learning techniques to search for recurring predictive rules in the foreign exchange market.

Initially an analysis and a preprocessing of the data obtained from Dukascopy for the 28 forex pairs in intervals of one hour, daily and weekly is done. Later an original segmentation technique based on Sliding Window, using different noise reduction techniques is applied.

The segments obtained in the previous step are grouped and classified using the K-means algorithm and the Pearson distance between the structures of them. The results obtained are compared with other alternatives.

Finally a recurrent sequential mining technique is used to find rules in the aforementioned time series based on the already classified segments. Rules of high confidence are searched in order to predict movements in the market upon occurrence of a specific set of continuous segments.

These rules are validated on 3 forex pairs and with varied execution conditions in order to verify the validity and quality of them; upon completion of said validations I offer relevant conclusions.

D. Carlos Javier Mantas Ruiz, Profesor del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

Informa:

Que el presente trabajo, titulado **Aplicación de técnicas de Machine Learning para el estudio, análisis, predicción y extracción de conocimiento del mercado de divisas,** ha sido realizado bajo su supervisión por Eilder Jorge Garcia, y autorizo la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expide y firma el presente informe en Granada a 30 de Junio de 2020.

El director:

Carlos Javier Mantas Ruiz

AGRADECIMIENTOS

Agradezco primeramente a mi mamá y a mi papá por traerme a este mundo y darme todo su apoyo, dándome la posibilidad de continuar con mis estudios hasta el día de hoy.

A mi tutor, Carlos Javier Mantas Ruiz por todo su apoyo y su interés en esta investigación desde el inicio de la misma.

A mi mejor y más viejo amigo, Raúl, por siempre inspirarme a superarme a mí mismo y llegar a cimas cada vez más altas y por siempre escucharme cuando tengo que quejarme por no tener tiempo.

Muchas gracias

ÍNDICE

Capítulo 1

Introducción

Los mercados financieros han sido objeto de estudio desde su creación en 1602. La sencillez de los datos y el gran número de actores involucrados en su manipulación convierte a estos en un objeto de investigación muy interesante. Numerosas técnicas y estudios se han realizado a lo largo de los años para determinar su naturaleza y como poder predecir futuros eventos económicos basado en valores históricos [1].

De los anteriores mercados, el de divisas es el que más liquidez posee, y donde mayor movimiento hay. Los datos de dicho mercado están abiertos al público y todo el mundo tiene acceso a la información ofrecida por los diferentes gobiernos y bancos centrales internacionales al mismo tiempo. Es un mercado muy estable donde la información histórica tiene mucho peso y donde se espera que los países y su moneda cambien de forma similar ante eventos similares.

En este trabajo se hace uso de varias técnicas de Machine Learning sobre estos datos del mercado de divisas, buscando extraer conocimiento del mismo que nos permita definir un modelo predictivo con una precisión suficiente en condiciones determinadas. Para poder explicar los objetivos y la forma en la que se desarrollará el trabajo es necesario hacer una breve introducción a los conceptos de series temporales y de Machine learning.

1.1 Series temporales en el mercado financiero

Una serie temporal es una secuencia de valores observados a lo largo del tiempo y ordenados cronológicamente. Sus mediciones tienden a ser equidistantes, ya sean segundos, minutos, días u otros.

En diferentes sectores se hace uso de series temporales para determinar información relacionada con las características del sector o la estacionalidad de los datos, esta información a su vez es frecuentemente utilizada con algoritmos de Machine Learning para realizar predicciones de eventos futuros.

Las aplicaciones son numerosas, desde la detección de fenómenos naturales y catástrofes hasta la determinación del posible nivel de tráfico en las autopistas de una ciudad a lo largo

del día y en fechas concretas. El estudio de estas series en el mercado financiero para predecir precios futuros es un campo de investigación que ha tomado mucha fuerza en las últimas décadas con la introducción del internet y la capacidad de los ordenadores de realizar el manejo de activos de forma automática. El aumento de los recursos de cómputo disponibles en la actualidad permite la utilización de técnicas de Machine Learning a las series temporales de los diferentes mercados como se observa en [2]

1.2 Aprendizaje automático (Machine Learning)

Por otra parte, el aprendizaje automático tiene uso en la resolución de muchos problemas complejos en la actualidad donde es imposible o poco práctico definir un conjunto de reglas y valores exactos como solución a un problema, pero que es intuitivo de explicar y que ocurre con frecuencia, como es el reconocimiento de números o de caras.

Consiste en el uso de algoritmos que son capaces de aprender de los datos, detectando patrones, secuencias, reglas y otros que permiten clasificar, predecir, agrupar o realizar un conjunto de acciones según el problema a resolver.

En la Figura 1 se observa el proceso tradicional de Machine learning.

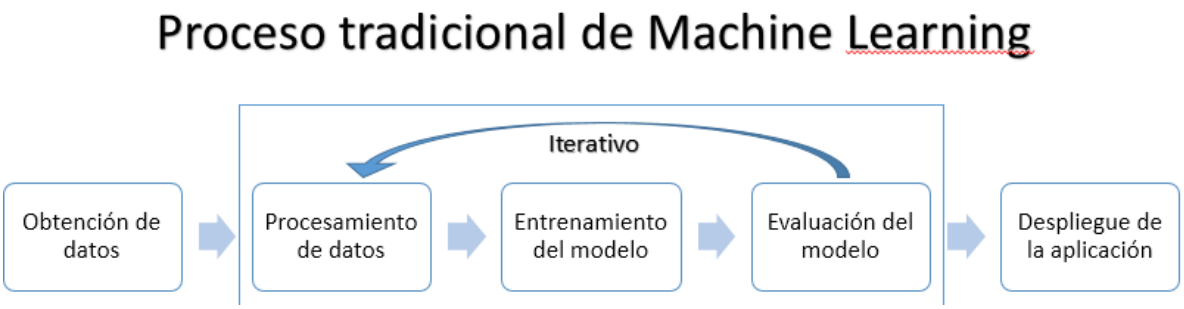

Figura 1: Proceso tradicional de Machine Learning

En las series temporales del mercado de divisas es muy poco práctico definir un conjunto de reglas fijo ya que el mercado se encuentra en constante cambio y desarrollo, por lo que la utilización de técnicas de Machine learning para la predicción de valores futuros ha aumentado significativamente con el paso de los años ya que estas técnicas permiten crear algoritmos que aprenden de los datos históricos del mercado y se van ajustando al mismo en tiempo real.

Por ende, ya que disponemos de gran cantidad de datos de los mercados financieros en el

formato de series temporales y de la gran cantidad de técnicas que actualmente nos ofrece el aprendizaje automático, se ve factible el aplicar estas herramientas de "Machine learning" para analizar el comportamiento de los mercados financieros, por lo que los objetivos del trabajo estarán enfocados en esta idea.

1.3 Objetivos

Se plantea en este trabajo extraer información predictiva del movimiento de los pares en el mercado de divisas haciendo uso de técnicas de Aprendizaje Automático.

Se pretende utilizar y comparar diferentes técnicas de agrupamiento de segmentos y minería de reglas secuenciales recurrentes para buscar la que mejor se adapte a este tipo de mercado y observar si es posible tener un margen de predicción con una precisión de al menos un 60%.

Se trabajará sobre los 28 pares de divisas que mayor flujo monetario tienen y que son los cruces entre el euro, la libra esterlina, el franco suizo, el yen japonés, el dólar americano, canadiense, australiano y el zelandés.

1.4 Desarrollo del Trabajo

Para la realización del trabajo es necesario obtener los datos de los pares de divisas con suficiente longitud para ser representativos. Posteriormente es necesario analizar la estructura de los datos y realizar una limpieza de los mismos, garantizando consistencia y calidad en los mismos.

Llegados a este punto es cuando comienza la parte investigativa del trabajo puesto que es necesario buscar una forma de segmentar la serie temporal para reducir el ruido y el número de datos a procesar. Una vez segmentada la serie temporal se va a estudiar las diferentes formas de agrupamiento que se pueden aplicar sobre dichos segmentos para clasificarlos según diferentes características estructurales.

Una vez clasificados los segmentos se va a buscar hacer uso de reglas secuenciales recurrentes aplicadas a las series temporales simbólicas para obtener reglas de alta confianza que permitan predecir movimientos posteriores a un conjunto de acciones. Finalmente se analizarán dichas reglas y los resultados obtenidos al aplicarlos sobre dichos pares de divisas separados para la validación de los mismos.

1.5 Estructura de la memoria

En este capítulo 1 se ha realizado una breve introducción al mercado de divisas y las series temporales, así como algunos elementos de inteligencia artificial y ciencia de datos necesarios para su estudio.

Capítulo 2: Expone los fundamentos necesarios para los temas estudiados en esta memoria.

Capítulo 3: Presenta un estudio de los datos, su fuente, y características inherentes a los mismos, buscando un entendimiento de cuales técnicas y algoritmos serán de mayor uso en la investigación.

Capítulo 4: Contiene varias técnicas de preprocesamiento de los datos, incluido la limpieza y preparación de los mismos. Además se crean nuevos datos representativos a partir de los datos originales y se aplican diferentes técnicas de suavizado para la eliminación del ruido.

Capítulo 5: En este capítulo se analizan algunas estrategias para segmentar las series temporales para reducir el ruido y la cantidad de datos, creando series más representativas para el estudio.

Capítulo 6: Contiene un estudio de aprendizaje no supervisado sobre los segmentos usando métricas de distancia euclídea y la distancia de correlación de Pearson para su clasificación usando el algoritmo de clustering K-Means.

Capítulo 7: Se analiza la minería de reglas secuenciales y recurrentes y su aplicación sobre diferentes grupos de segmentos clasificados y sus resultados al aplicarse sobre los pares de divisa AUDCAD, EURNZD y USDJPY.

Capítulo 8: El capítulo final expone las conclusiones de la investigación así como posibles extensiones a la investigación y notas importantes obtenidas durante la misma.

Capítulo 2

Fundamentos

En este capítulo se describen las características de los elementos usados en la investigación y los conocimientos básicos necesarios para poder abordar los elementos estudiados en la misma.

2.1 Mercado Financiero

Según [3], se define el mercado financiero como el marco de negociación, determinación del precio y contratación entre demandantes y oferentes de recursos financieros instrumentados por medio de activos financieros.

Además, todo mercado financiero debe desempeñar las siguientes funciones (Figura 2):

- Poner en contacto a los distintos agentes económicos.
- Servir de mecanismo para la fijación del precio de los activos.
- Proporcionar liquidez a los activos que se negocian en éste.
- Reducir los plazos y costes de intermediación

Figura 2: Funciones del mercado financiero

Los mercados financieros son clasificados en diferentes tipos según los activos que se negocien o la forma de negociación. Estos tipos de mercados financieros los podemos apreciar en la Figura 3.

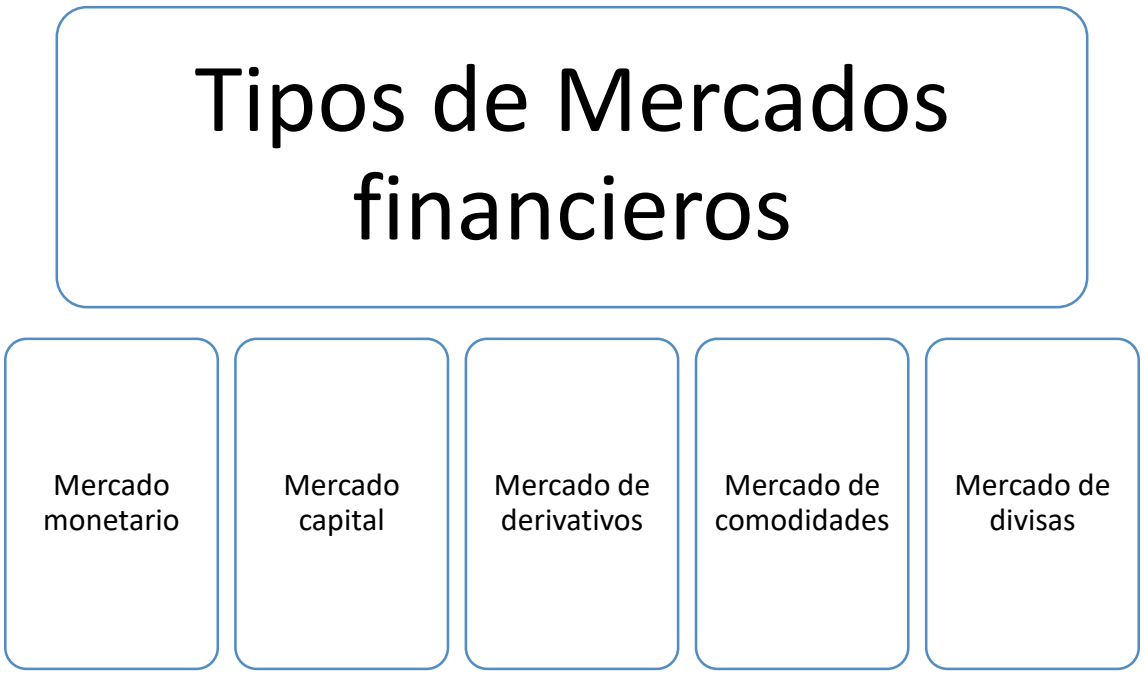

Figura 3: Tipos de Mercado financiero

2.1.1 Mercado de divisas

Son aquéllos en los que se realizan los intercambios de unas monedas por otras.

Salvo el segmento correspondiente al mercado oficial de divisas, si es que existe, el resto del mercado no tiene una ubicación física concreta, pues tanto el mercado interbancario de divisas, que corresponde al contenido sustancial del mercado, como las transacciones de moneda con la clientela carecen normalmente de una localización determinada.

Los mercados de divisas son muy interesantes puesto que son los mercados donde más participantes existen, llegándose a intercambiar la cantidad equivalente a 5 trillones de dólares americanos a diario en la actualidad. Esto causa que el mercado sea muy orgánico y difícil de manipular, puesto que es muy difícil para una institución o un grupo de instituciones mover esa cantidad de dinero.

2.1.2 Representación del mercado de divisas

En la representación gráfica de este mercado, se miden en ordenadas el precio en moneda nacional de una unidad de moneda extranjera ($, €, ¥, £, etc.) y en abscisas la fecha en la que el precio corresponde a la cantidad intercambiada de moneda extranjera por moneda nacional [4]. Esto se puede observar en la Figura 4: donde vemos que para

el día 11 de febrero el precio de venta de un Euro era de 1.09228 dólares americanos

Figura 4: Típica representación de una serie temporal del mercado de divisas usando líneas.

Un par de divisas en el mercado internacional se representa usualmente con la siguiente notación:

FX:EURUSD ASK: 1.09229 BID 1.09228

La primera sección define el organismo o mercado donde se encuentra la oferta, en este caso el mercado sería FX, una abreviatura de FXCM, una firma que ofrece servicios de compra y venta de activos financieros y divisas.

La segunda sección se compone de un código de 6 letras que define el par monetario del que se ofrece información. El primer elemento del par monetario es la moneda nativa, mientras que el segundo elemento es la moneda extranjera. En este ejemplo se ofrece información de cambio de Euros por Dólares Americanos.

La tercera sección muestra el valor actual mínimo a la que alguna entidad en el mercado ofrece su moneda nativa por moneda extranjera; a su vez contiene el valor máximo que una entidad está dispuesta a pagar de moneda extranjera por una moneda nativa. Se puede observar que tanto el vendedor como el comprador tienen un 0.00001 de diferencia entre los precios.

En el mercado de divisas es muy natural tener estas separaciones diminutas, evitando saltos bruscos en los precios durante los horarios de mercado, una característica que no se observa en otros mercados. Esta característica se conoce como liquidez, y es debido a que hay tantos participantes en la compra y venta que la diferencia entre ofertas es mínima, creando lo que se conoce como un mercado profundo y líquido. Esta característica es fundamental para un estudio de Machine Learning puesto que nos asegura que las fluctuaciones en los datos son naturales y pueden representar patrones psicológicos o reacciones recurrentes a eventos internacionales y no manipulación de precios por entidades interesadas, algo que ocurre frecuentemente en mercados carentes de liquidez. Como se observa en la Figura 4 anterior, en ocasiones hay tantas operaciones de compra y venta que la diferencia es literalmente de 0, hay alguien vendiendo al mismo precio que otra persona comprando, pero su orden aún no ha sido procesada.

El mercado de divisas tiene una característica principal que no comparte con otros mercados: la dualidad de información representada en los datos. En un mercado de divisas las operaciones se realizan siempre vendiendo una moneda y comprando otra, con un valor en constante cambio según los principios de suministro y demanda. Esto implica que siempre se observe el precio relativo a las dos monedas asociadas, creando dificultades a la hora de poder hacer análisis sobre estos datos. En una acción de empresa, el precio siempre se observa con respecto al valor de la empresa, si la empresa se encuentra en ascenso, su valor también debe ir en ascenso; en las divisas no es posible conocer a partir de un par de divisas si una de las dos divisas está en alza, si ambas están teniendo poco movimiento, pero una es ligeramente mejor que la otra, entre otros elementos.

2.2 Propiedades de una serie temporal

Una serie temporal o cronológica es una sucesión de valores que adopta una variable (y): [5]

$$y1, y2, y3, \ldots, yn$$

En distintos instantes de tiempo (t):

$$t1, t2, t3, \ldots, tn$$

Utilizando la Figura 5 del mercado financiero para mostrar una representación gráfica de una serie temporal:

Figura 5: Típica representación de una serie temporal del mercado de divisas usando líneas.

En la serie temporal mostrada en la Figura 5 los intervalos son de una hora, por lo que cada punto representa el precio de cierre de la hora en cuestión.

Las mediciones en series temporales se realizan sobre intervalos de tiempo equidistantes definidos a priori, ya sean días, años o segundos. La continuidad de estos intervalos es fundamental a la hora de analizar una serie temporal e intervalos faltantes en la serie deben ser tratados con sumo cuidado.

El tratamiento de valores faltantes es muy importante y debe realizarse con técnicas que permiten conservar las características de la serie, típicamente usando técnicas de imputación y aproximación, esto no es necesario en el estudio debido a la naturaleza del mercado de divisas donde la serie temporal es totalmente continua y equidistante, pero es importante considerarlo puesto que la fuente de datos utilizada podría tener intervalos faltantes, como en el caso de secuencias de los agentes de bolsa o bancos que ofrecen servicios de este tipo y cuyos intervalos podrían no ser una representación fiel del precio exacto del mercado internacional y contener valores faltantes.

2.2.1 Series temporales en diferentes intervalos

En los mercados financieros es habitual trabajar con diferentes intervalos de tiempo,

representando series temporales únicas asociadas a la misma información subyacente. Según el tipo de estrategia que se use es habitual usar intervalos más largos o más cortos. Como se define en [6], [7], [8], el análisis de múltiples intervalos de la misma serie es capaz de mejorar mucho la precisión y la calidad de las operaciones y estrategias.

En principio la separación en diferentes intervalos sirve como una forma de disminuir el ruido y suavizar los datos, reduciendo el margen de error a la hora de realizar el análisis.

Es recomendable utilizar 3 intervalos de tiempo para las operaciones:

- Un intervalo superior, que sirve para determinar la tendencia de la serie y donde es habitual utilizar algoritmos de regresión para buscar una dirección aproximada.
- El intervalo de análisis, donde se realiza un estudio de la tendencia actual, posibles puntos extremos, noticias internacionales, y demás, con el objetivo de buscar un momento donde exista poco riesgo en tomar una posición, basándose en la tendencia definida por el intervalo superior.
- El intervalo de ejecución, el nivel más bajo donde se define el punto de entrada, típicamente buscando valores anómalos o extremos en sentido opuesto a la tendencia superior y que coincida con lo establecido en el intervalo de análisis.

Un ejemplo tradicional del funcionamiento de esta técnica se puede observar en la Figura 6:

Figura 6: Análisis de múltiples intervalos de tiempo en una gráfica

La serie temporal del precio segmentada en intervalos de una hora se puede observar en la parte superior, mientras que la parte inferior contiene la información del intervalo de tiempo semanal, diario y por hora.

Las líneas representan el aumento o descenso de precio de la serie con respecto a su media y a su desviación estándar, siendo 0 la base, este concepto se conoce típicamente como Z-score y será detallado posteriormente. La idea es que mientras más los valores se alejan de 0, más fuerza existe hacia una dirección.

En esta imagen se representa el Z-score del intervalo semanal con la línea señalada con (1), el intervalo diario con la línea (2) y el intervalo actual de una hora, con la línea (3). En esencia se busca establecer una tendencia usando la línea semanal y luego buscar con la línea diaria y la horaria un punto extremo en contra de esta tendencia.

En el primer ejemplo se observa una tendencia un poco bajista de la línea semanal, mientras que la línea diaria parece estar perdiendo impulso y la línea horaria ha hecho un pico hacia el lado alcista, en este caso se puede considerar abrir una posición bajista de riesgo medio ya que aún la línea diaria no ha comenzado una tendencia bajista.

En el segundo se observa una clara tendencia bajista de la serie con la línea semanal y

diaria por debajo de 0 y avanzado a valores cada vez más bajos, pero con un valor extremo positivo de 4 en la línea horaria, indicando una muy fuerte señal de venta de poco riesgo.

El tercer ejemplo es muy similar al primer ejemplo, pero esta vez se observa más claramente el pico en la línea horaria.

El cuarto representa una venta casi ideal, al observar que la línea semanal sigue descendiendo y la línea diaria ha alcanzado un pico y ha comenzado a descender también. El precio después llegó a bajar hasta 1.09 en el transcurso de una semana.

2.3 Machine Learning

El aprendizaje se puede definir como el proceso de adquisición de conocimiento o de habilidades mediante estudios, instrucciones o experiencia. [9]

Una maquina podría considerarse que está aprendiendo si es capaz de modificar su estructura, sus datos o código para mejorar su rendimiento en el futuro.

A partir de estas simples definiciones se podría definir el Machine Learning como la capacidad de sistemas de inteligencia artificial de adquirir conocimientos o habilidades mediante instrucciones o cambios en su ambiente de ejecución, modificando la estructura de los mismos para obtener un mejor rendimiento en el futuro. En la práctica es habitual que el aprendizaje sea limitado al ajuste de parámetros definidos a priori con el objetivo de mejorar el rendimiento del sistema.

Según el tipo de problema y la solución deseada los sistemas de Machine Learning se pueden dividir en tres categorías:

- Algoritmos supervisados. Son algoritmos en los cuales expertos utilizan técnicas de refuerzo y validación para enseñarle al modelo cuales son los resultados correctos y cuáles son los resultados incorrectos. Un ejemplo tradicional es la clasificación de objetos en una imagen, donde el experto define de antemano que objeto es el que se encuentra en la imagen, con el objetivo de que el sistema aprenda las características de ese objeto.
- Algoritmos semi-supervisados. Son algoritmos en los cuales se tiene parte de la información necesaria y se conoce lo que se quiere obtener a partir de esa información, pero no se conoce el valor correcto de todas las clasificaciones. Un caso habitual es el estudio de registros de ataques informáticos, en los cuales se poseen un conjunto de registros y se sabe que algunos son ataques, pero se desconoce si el resto de los casos fueron ataques o no y se busca que el sistema

determine que registros representan un ataque y cuáles no.

- Algoritmos no supervisados. Los algoritmos no supervisados son algoritmos que funcionan sin recibir información de un experto usando solo los datos proporcionados con carencia de una medida de medición de precisión, ya que no se conoce que valores son correctos o incorrectos, o esta información no es aplicable al problema. Estos algoritmos son muy usados en técnicas de búsqueda de conocimiento mediante la detección de grupos o clústeres que contienen características similares, como pueden ser los diferentes tipos de cliente que operan con una empresa.

2.3.1 Minería de datos

La minería de datos es un concepto asociado a la búsqueda y adquisición de conocimientos a partir de un conjunto de datos que representa información. En general se considera que la minería de datos es un paso en particular del proceso, consistiendo en la aplicación de algoritmos específicos para extraer patrones y modelos de los datos [10].

La minería de datos se realiza después de una etapa de procesamiento y limpieza de los datos, y los modelos obtenidos son estudiados y analizados con el objetivo de obtener conocimiento. Tradicionalmente es habitual que se realicen múltiples iteraciones de estos pasos, obteniendo diferentes modelos según el estudio que se desee realizar (Figura 7).

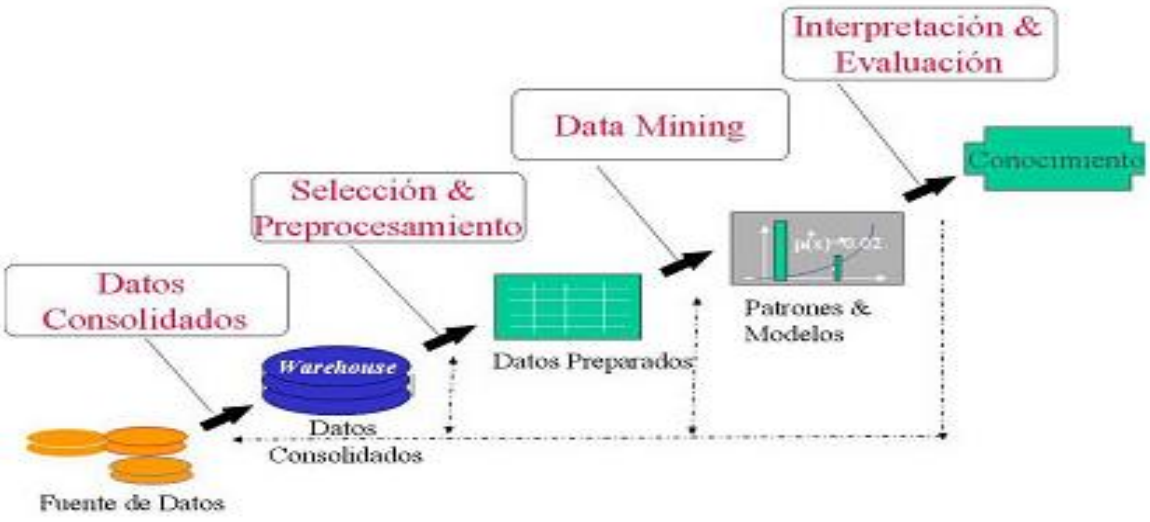

Figura 7: Ciclo de búsqueda de conocimientos y minería de datos

2.3.2 Minería de datos en series temporales

Una serie temporal está compuesta de valores (y):

$$y1, y2, y3, \ldots, yn$$

En distintos instantes de tiempo (t):

$$t1, t2, t3, \ldots, tn$$

En la minería de datos tradicional el orden de los atributos es irrelevante y su relación es independiente de sus posiciones. Pero en una serie temporal la relación entre el valor actual (y) y el tiempo (t) es crucial, y un análisis ordenado de los mismos es imprescindible.

Esta característica dificulta mucho el tratamiento de las series temporales y es necesario analizarlas como un todo en vez de tomar los diferentes elementos como un conjunto de valores numéricos. [11]

En el campo de minería de datos en series temporales es habitual realizar estudios que pertenecen a una de las siguientes categorías:

- Indexado: A partir de una serie de interés *s* y una medida de similitud, encontrar la serie más cercana a *s* en una base de datos temporal.
- Descubrimiento de patrones y conglomerados: Consiste en descubrir patrones interesantes que pueden aparecer con frecuencia en las series temporales.
- Clasificación: Tiene como objetivo asignarle etiqueta a una serie a partir de un conjunto de clases previamente definido.
- Segmentación: puede ser considerado como un paso previo de preprocesamiento. Tiene como objetivo, a partir de una serie, obtener un conjunto reducido de segmentos que aproximen la serie original.

En modelos avanzados, como la predicción del futuro basado en las series temporales, es habitual utilizar segmentación, luego clasificación y finalmente un proceso de búsqueda de patrones sobre estos datos con métodos tradicionales.

2.4 Clasificación

La clasificación es el proceso de predecir una clase dando un conjunto de datos. Las clases a veces son llamadas objetivos o categorías. La creación de modelos predictivos de clasificación consiste en la tarea de aproximar una función de relación (f) que a partir de las variables de entrada (X) nos de la clase asociada (y). La clasificación es uno de los objetivos más habituales en el Machine Learning y existen múltiples formas de

realizarse según el problema a tratar, como se ve en la Figura 8:

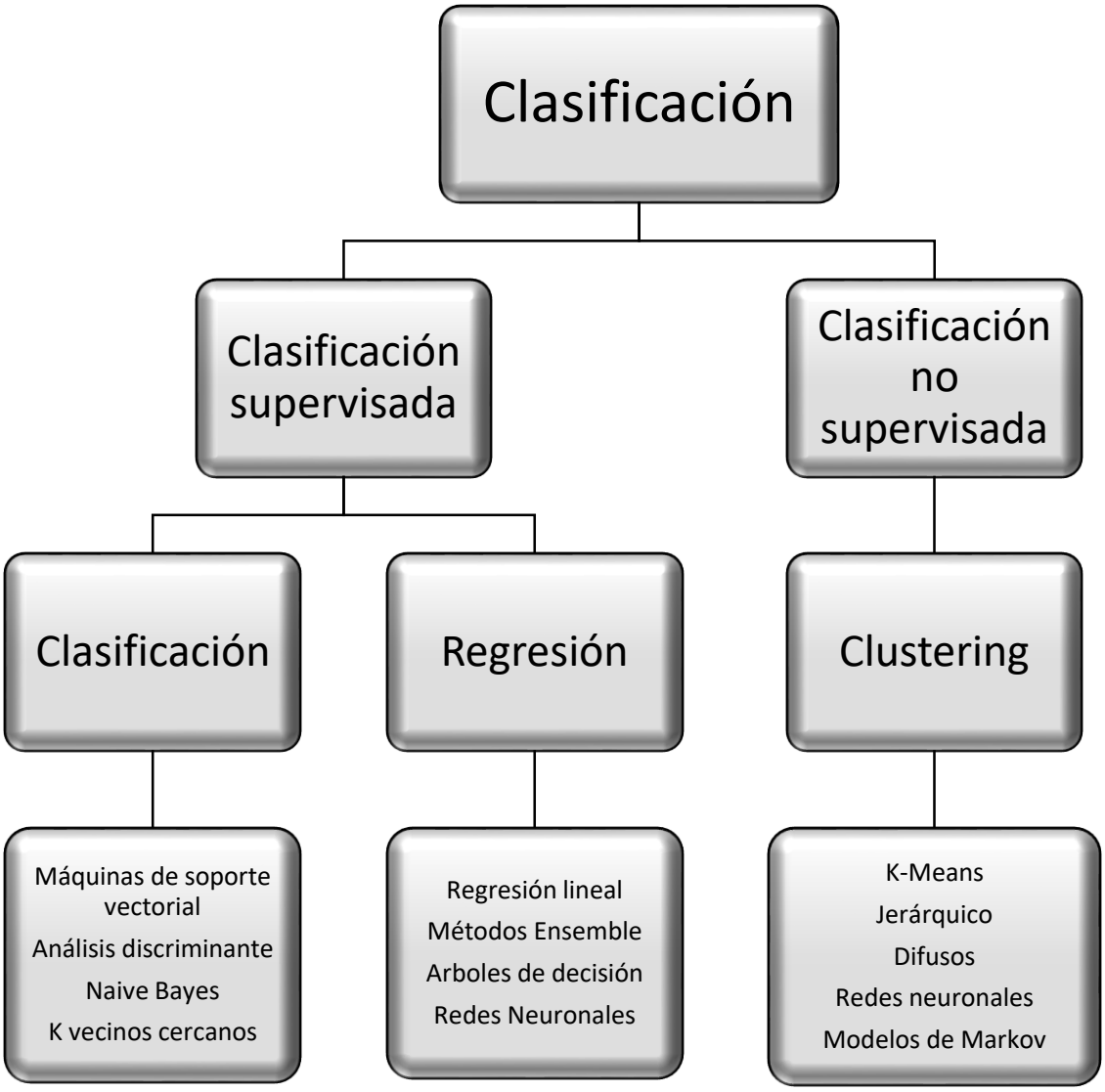

Figura 8: Tipos de Clasificación

El tipo de clasificación a ser utilizado y las técnicas y modelos dependen del problema y el tipo de datos que se traten.

Si se conocen de antemano las clases de los datos, se pueden utilizar técnicas de regresión o técnicas de clasificación para poder definir la función de relación (f) según el modelo utilizado.

En caso de no conocerse las clases de los datos, la clasificación es considerablemente más difícil, y lo que se usa son técnicas de agrupamiento, conocidas típicamente como técnicas de Clustering. El objetivo de estas técnicas es determinar grupos de datos que

compartan características similares, por ejemplo, grupos de amigos en redes sociales, o categorías de clientes en una empresa.

2.5 Z-Score

El Z-Score es un valor estadístico estándar que describe la ubicación de un valor X dentro de su distribución describiendo su distancia a la media en términos de desviación estándar [12] [13].

Su fórmula es muy sencilla:

$$Z = \frac{X - \mu(X)}{\sigma(X)}$$

El Z-score de un valor actual es equivalente a la diferencia del valor con respecto a su media dividido entre su desviación estándar.

En una distribución normal, el Z-Score nos da información que permite comparar variables y modelos de forma intuitiva sin conocer los valores subyacentes de X como se puede observar en la Figura 9:

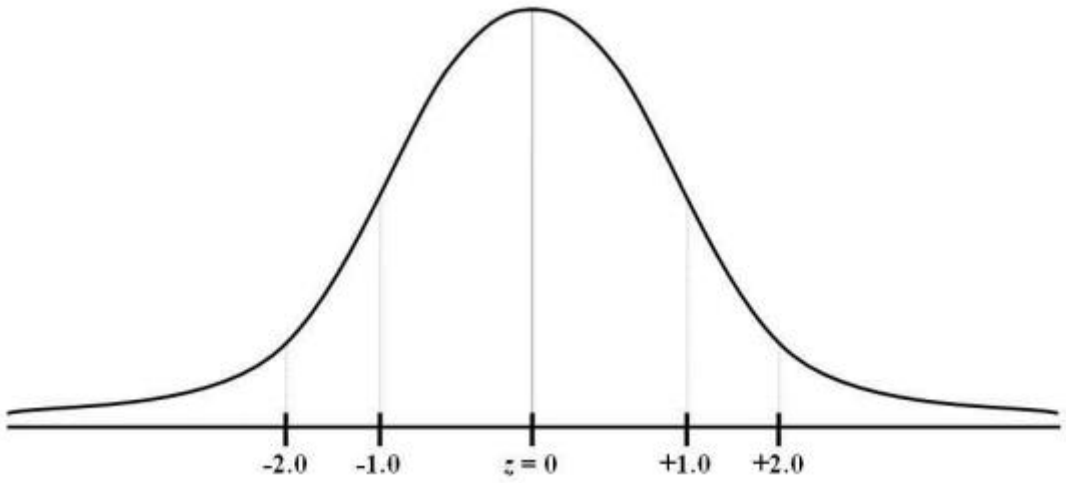

Figura 9: Curva de campana usando Z-Score

Se puede apreciar que con independencia de los valores que tome X, un Z-score de +2 indica un valor muy elevado con respecto a la media y valores incluso mayores, como 5 o 10 indicarían casos muy extremos, ya que el Z-Score toma en consideración no solo la media, sino la desviación estándar de la distribución.

2.5.1 Z-Score en series temporales

El cálculo de Z-Score en series temporales depende totalmente de cuantos intervalos se tomen para el cálculo de la media y la desviación estándar. El Z-score de una serie temporal puede ser representado como una serie temporal en sí, cuyo valor fluctúa en el tiempo, pero nos permite identificar el movimiento y tendencia de la serie independientemente de los valores subyacentes. En la Figura 10 se puede observar una serie temporal de Z-score (inferior) con respecto al precio subyacente asociado (superior) usando como media los últimos 250 intervalos:

Figura 10: Z-Score (parte inferior) en una serie temporal del mercado de divisas: AUDCAD

Es interesante ver como el Z-Score refleja el precio subyacente con respecto a su media y su desviación y no con respecto al valor, se puede observar claramente que el valor -1 cerca del final de la serie se corresponde con -090335, mientras que este mismo -1 al inicio de la serie se corresponde con un 0.98.

El uso de Z-score nos da dos posibilidades principales, la primera es la posibilidad de comparar variables y tendencias con valores y desviaciones diferentes; por ejemplo, en la imagen anterior, se muestra el par AUDCAD, cuyos valores oscilan entre 1.04 y 0.88

durante el 2017 y finales del 2019, pero si se quisiera determinar comparar esta serie temporal con la del AUDJPY, se puede observar la Figura 11:

Figura 11: Z-Score (parte inferior) en una serie temporal del mercado de divisas: AUDJPY

Se tiene que el precio varía entre 90 y 70 durante el mismo intervalo, ¿cómo se podría comparar estas dos series temporales cuando el rango de precio y la variación es totalmente diferente? El Z-Score permite realizar este tipo de comparaciones ya que establece un rango común entre las mismas.

La segunda posibilidad que nos brinda, es la interpolación de un valor aproximado de Z-Score asociado a una tercera variable desconocida a partir de variables conocidas, pero con relaciones diferentes. En los ejemplos anteriores de AUDCAD y AUDJPY, se ve el precio descendiendo sobre el transcurso del tiempo, pero tenemos una relación XY, y XZ, no tenemos una forma de determinar el valor subyacente de X, Y o Z debido a que sus relaciones y rangos son diferentes, pero el Z-Score nos podría decir que hay una probabilidad alta de que si el Z-Score de XY y XZ es negativo, es porque el valor de X es inferior a su media en una determinada cantidad, mientras más relaciones se usen, mejor es la aproximación realizada.

Capítulo 3

Estudio de los datos

Al realizar una investigación de minería de datos y adquisición de conocimientos la calidad y estructura de los datos es tan, o más importante, que los algoritmos y modelos usados para procesar los mismos.

Es de suma importancia verificar la distribución de los datos y el balanceo de las clases, buscando una fuente fiable que proporcione los datos necesarios para el estudio. Según un estudio realizado en el 2016 por Forbes [14] el 19% del tiempo de trabajo en un proceso de adquisición de conocimiento y Machine Learning es la obtención de los datos y un 60% es la preparación de estos datos (Figura 12).

Figura 12: Distribución del tiempo de trabajo durante un proceso de estudio y análisis de datos usando Machine Learning

Otros estudios indican números incluso mayores, a veces llegando a 80%, y la razón fundamental es que este análisis y preparación previa de datos depende mucho del

factor humano, y no es posible utilizar muchos mecanismos automáticos para esta fase.

Este capítulo trata el estudio inicial de los datos del mercado de divisas y las diferentes características de los mismos.

3.1 Obtención de los datos

En esta investigación de Machine Learning sobre el mercado de divisas se realizó una búsqueda de diferentes proveedores de datos, con el objetivo de buscar un proveedor que tuviera toda la información de los 28 pares almacenadas en intervalos mínimos de 1 hora y con varios años de información disponible para hacer el estudio.

De todos los proveedores disponibles, el único que cumplía con las condiciones, además de ser totalmente gratuito, era Dukascopy [15]. Con este proveedor se pudo adquirir los datos de los 28 pares de divisas separados en intervalos de una hora, un día y una semana, generando un total de 3,103,352 datos para el estudio, separados en 84 archivos .csv según el par y el intervalo que contiene, desde el 01.01.2008 hasta el 25.01.2020, fecha final disponible al realizarse esta investigación.

3.2 Características de los datos

Estos datos representan series temporales del precio BID del mercado de divisas y contiene 6 variables, el intervalo de tiempo (t) y 5 valores representando información sobre el precio en ese intervalo de tiempo.

Estas 5 variables son el precio de apertura al inicio del intervalo, el precio máximo alcanzado durante el intervalo, el precio mínimo alcanzado durante el mismo, el precio final de cierre en el último segundo antes de concluir el intervalo y el volumen total de transacciones realizadas en el intervalo, medido en millones. Este volumen de transacciones es el que posee la compañía y solo representa una parte pequeña del volumen real manipulado en todo el mercado, pero sirve como muestra para determinar la proporción del mercado en su totalidad.

Nombre de variable	Tipo de variable	Notas
Local Time	DateTime	En formato día.mes.año hora:minuto:segundo
Open	Numérico real	Decimales variables en dependencia del par

High	Numérico real	Decimales variables en dependencia del par
Low	Numérico real	Decimales variables en dependencia del par
Close	Numérico real	Decimales variables en dependencia del par
Volume	Numérico real	Decimales variables en dependencia del par

A continuación, en la Figura 13 se observa la primera entrada de la tabla de datos de la serie temporal del par de divisas AUDCAD, con inicio el 1 de enero del 2008:

Local time	Open	High	Low	Close	Volume
01.01.2008 00:00:00.000 GMT+0100	0.87258	0.87258	0.87253	0.87257	9.195
01.01.2008 01:00:00.000 GMT+0100	0.87257	0.87257	0.87253	0.87257	7.9589
01.01.2008 02:00:00.000 GMT+0100	0.87257	0.87257	0.87253	0.87257	26.8013
01.01.2008 03:00:00.000 GMT+0100	0.87257	0.87257	0.87253	0.87257	4.5733
01.01.2008 04:00:00.000 GMT+0100	0.87257	0.87257	0.87253	0.87253	12.766
01.01.2008 05:00:00.000 GMT+0100	0.87253	0.87257	0.87253	0.87253	21.6085
01.01.2008 06:00:00.000 GMT+0100	0.87253	0.87257	0.87253	0.87257	13.5261
01.01.2008 07:00:00.000 GMT+0100	0.87253	0.87257	0.87253	0.87253	19.814
01.01.2008 08:00:00.000 GMT+0100	0.87257	0.87257	0.87253	0.87257	11.626
01.01.2008 09:00:00.000 GMT+0100	0.87257	0.87257	0.87253	0.87257	18.6632
01.01.2008 10:00:00.000 GMT+0100	0.87257	0.87257	0.87253	0.87257	2.3922
01.01.2008 11:00:00.000 GMT+0100	0.87257	0.87257	0.87257	0.87257	0.0782
01.01.2008 12:00:00.000 GMT+0100	0.87257	0.87257	0.87253	0.87257	17.7748
01.01.2008 13:00:00.000 GMT+0100	0.87253	0.87257	0.87253	0.87257	15.9809

Figura 13: Primeras entradas del par AUDCAD

En las observaciones de la Figura 13 el par abre con un valor de 0.87258 contra el dólar canadiense, alcanza en valor máximo de 0.87258, un valor mínimo de 0.87253 y finalmente cierra 1 milésima de centavo por debajo del precio inicial a las 0:59:59.999, llegándose a mover dinero equivalente a 9.195 millones entre compras y ventas de dólares australianos con dólares americanos en esta hora inicial.

El mercado de divisas funciona las 24 horas, con excepción de los fines de semana y días festivos internacionales como vísperas de navidad, fin de año y similares; también

es posible que cierren en situaciones de caos internacional o frente a desastres, por lo que hay varios intervalos que no aportan información sobre el desarrollo de la serie y deben ser procesados correctamente. Un ejemplo de fin de semana aparece en la Figura 14.

96	04.01.2008 22:00:00.000	1.47501	1.47516	1.47371	1.47397	24279.5977
97	04.01.2008 23:00:00.000	1.47397	1.47397	1.47397	1.47397	0
98	05.01.2008 00:00:00.000	1.47397	1.47397	1.47397	1.47397	0
99	05.01.2008 01:00:00.000	1.47397	1.47397	1.47397	1.47397	0
100	05.01.2008 02:00:00.000	1.47397	1.47397	1.47397	1.47397	0
101	05.01.2008 03:00:00.000	1.47397	1.47397	1.47397	1.47397	0
102	05.01.2008 04:00:00.000	1.47397	1.47397	1.47397	1.47397	0
103	05.01.2008 05:00:00.000	1.47397	1.47397	1.47397	1.47397	0

Figura 14: Fin de semana en el mercado de divisas

A pesar de no estar funcionando los mercados, ciertas entidades siguen realizando transacciones y modificando el precio, por lo que es habitual que al concluir el fin de semana o las festividades el precio de apertura sea significativamente diferente al precio de cierre (Figura 15).

20.01.2008 22:00:00.000	1.46168	1.46168	1.46168	1.46168	0
20.01.2008 23:00:00.000	1.46029	1.46042	1.45912	1.45957	8405.4502

Figura 15: Inicio de semana en el mercado de divisas

Esto causa que la serie temporal contenga saltos si se analiza en su totalidad, con los elementos de apertura y cierre, esto puede ser útil para ciertos tipos de investigación y de estrategias pero muchos participantes en el mercado prefieren utilizar una representación lineal clásica con solo el precio de cierre, ya que este contiene menos ruido, como se observa en la Figura 16 y la Figura 17.

Figura 16: Salto en el precio de apertura con respecto al cierre de fin de semana

Figura 17: El fin de semana representado usando solo el precio de cierre

3.2.1 Correlación de variables: precios

Siendo una serie temporal donde cuatro variables representan la información subyacente relacionada al valor de la moneda la correlación entre estas variables es muy elevada, y en un elevado número de casos el precio de apertura es equivalente al precio de cierre del intervalo anterior. El precio máximo acota la distribución descartando valores inferiores al máximo, mientras que el precio mínimo la acota descartando valores inferiores al mínimo, pero la distribución es similar.

Para verificar estadísticamente si estas variables tienen una muy alta correlación se utiliza el coeficiente de Pearson, definido por la siguiente ecuación:

$$r(xy) = \frac{\sum Z(x)Z(y)}{N}$$

Esta ecuación hace referencia a la media de los productos cruzados de las puntuaciones estandarizadas de X y de Y. Esta fórmula reúne algunas propiedades que la hacen preferible a otras. Al operar con puntuaciones estandarizadas es un índice libre de escala de medida. Por otro lado, su valor oscila, como ya se ha indicado, en términos absolutos, entre 0 y 1. 1 indica igualdad, con valores cercanos indicando una muy alta correlación, mientras que 0 indica total independencia.

Según el coeficiente de Pearson, la correlación entre el precio de apertura y el precio de cierre es de 0.9997997, indicando una relación muy elevada, esto se puede apreciar en la Figura 18:

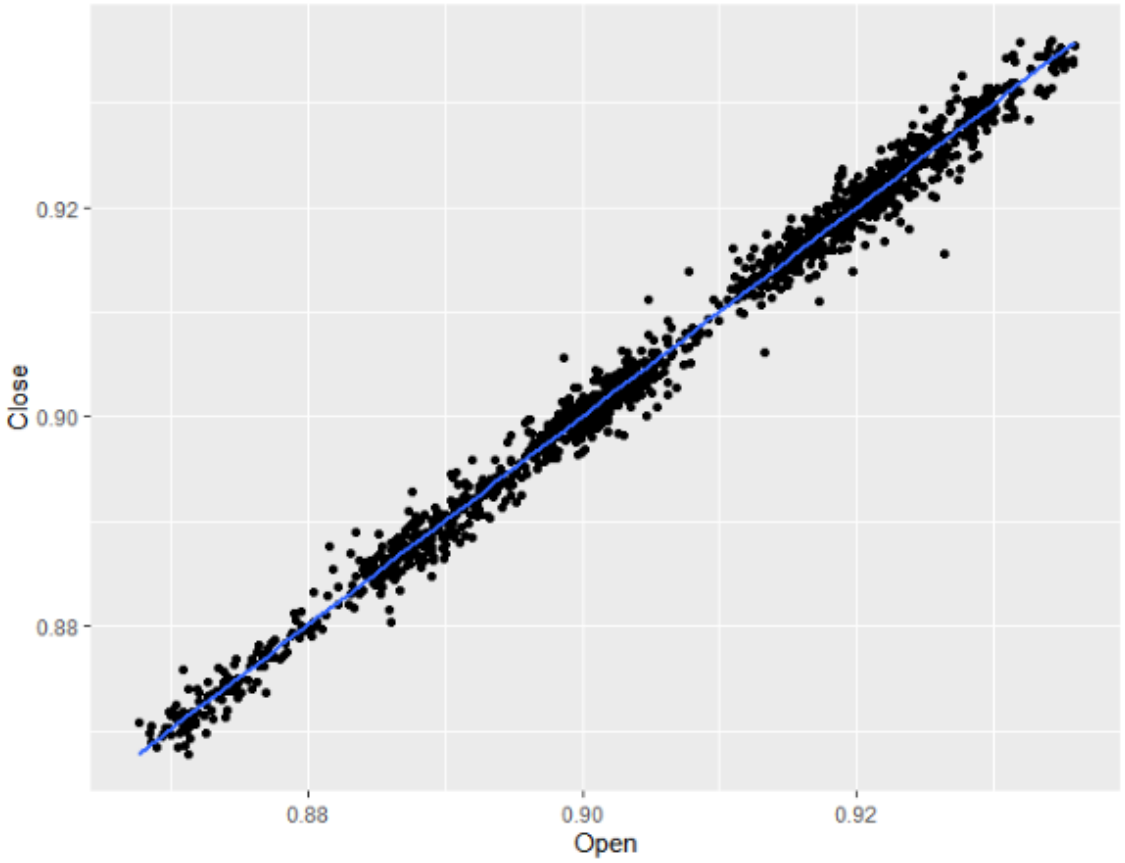

Figura 18: Alta dependencia lineal entre el precio de apertura y el precio de cierre

La correlación con respecto al precio de cierre para el precio máximo y el precio mínimo es superior al 99.98% según este mismo coeficiente de Pearson.

Esta correlación es ligeramente menor en el caso de intervalos mayores donde existe más variación en el transcurso del tiempo, llegando a caer hasta 97% en algunos pares de divisas.

3.2.2 Correlación de variables: volumen

El análisis de volumen es muy interesante en el mercado de divisas y muy usado por profesionales para determinar la fuerza de los movimientos. Existen muchos sitios y lugares donde explican formas de usar el volumen para determinar precios futuros, como en [16].

Sin embargo, ¿existe realmente una relación tan notable entre el precio y el volumen en el mercado de divisas donde se mueven trillones en un día?

En la Figura 19 se puede apreciar que no hay una relación aparente entre estos:

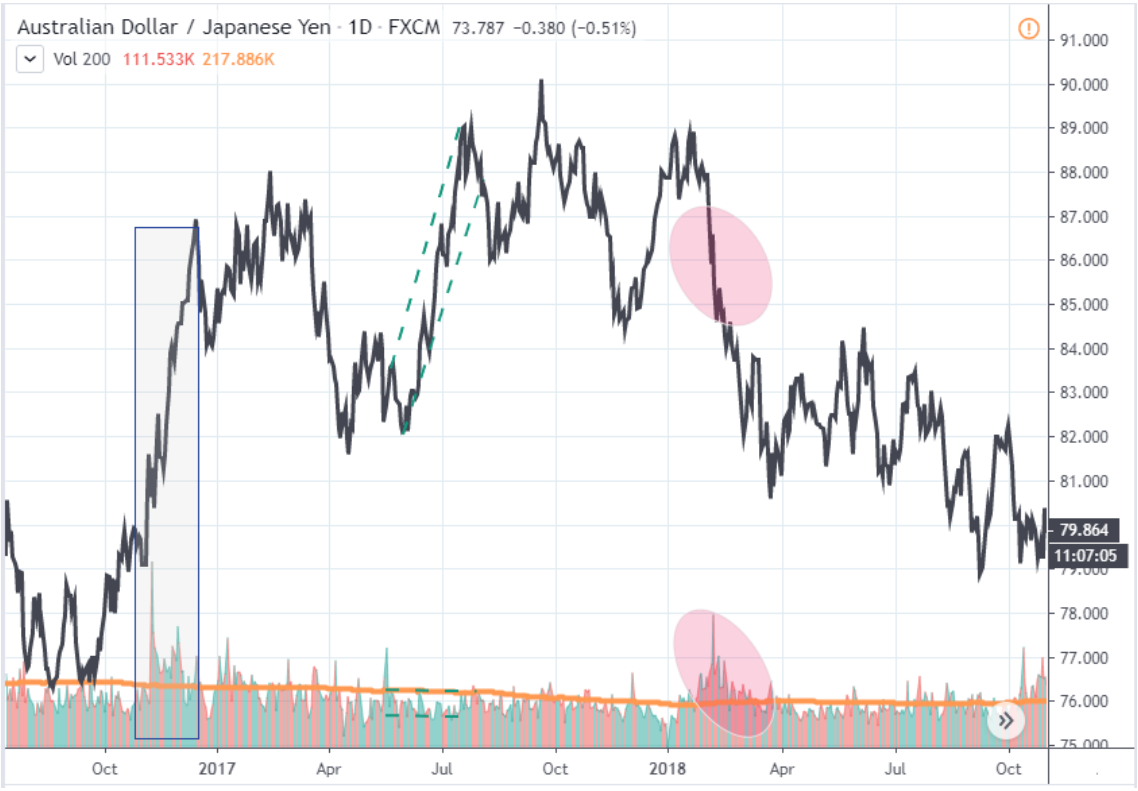

Figura 19: Baja relación entre el volumen y el precio de cierre

En el segmento rectangular vemos un aumento drástico del volumen antes de que realmente el precio aumentara, pero no fue consistente este aumento.

En el medio tenemos líneas troceadas indicando otra subida de precio masiva con un volumen inferior a su media. Y en el caso de la Elipse vemos una caída que se corresponde con un pico de volumen y una reducción en el impulso a medida que el volumen disminuye.

Estos ejemplos parecen indicar que no existe correlación directa alguna y usando el coeficiente de Pearson se observa que el precio tiene una correlación de solo 25% con el volumen.

Sin embargo, esta relación depende del intervalo de tiempo utilizado para medir la serie temporal, en intervalos de solo un minuto si se aprecia más notablemente la relación de volumen y precio (Figura 20):

Figura 20: Relación volumen/precio en intervalos pequeños

No obstante, para el momento que se podría detectar un aumento inusual de volumen, ya el precio ha cambiado substancialmente.

Al realizar un estudio de los Z-Score de precio y volumen y compararlos nuevamente con el coeficiente de Pearson obtenemos total independencia entre ambos, esto implica que efectivamente el volumen no nos sirve para establecer modelos predictivos del precio.

El volumen puede quizás proveer información para sistemas y estrategias basadas en la especulación de precio a partir de la reacción a un determinado evento, pero no aporta información a un modelo de predicción de precio basado en valores históricos.

3.3 Distribución de precio de cierre e intervalos

Según lo establecido en los puntos anteriores, solo el precio actual y sus valores anteriores nos pueden ser de utilidad en la determinación de precios futuros, los valores

máximo y mínimo me acotan información del modelo, mientras que el precio de apertura es equivalente en la mayoría de las ocasiones al precio de cierre del intervalo anterior. La independencia del volumen con respecto al precio y la elevada correlación entre apertura, máximo, mínimo y cierre indica que solo es necesario investigar el comportamiento del precio de cierre.

La distribución del precio de cierre es diferente según el mercado, como se puede observar en la Figura 21:

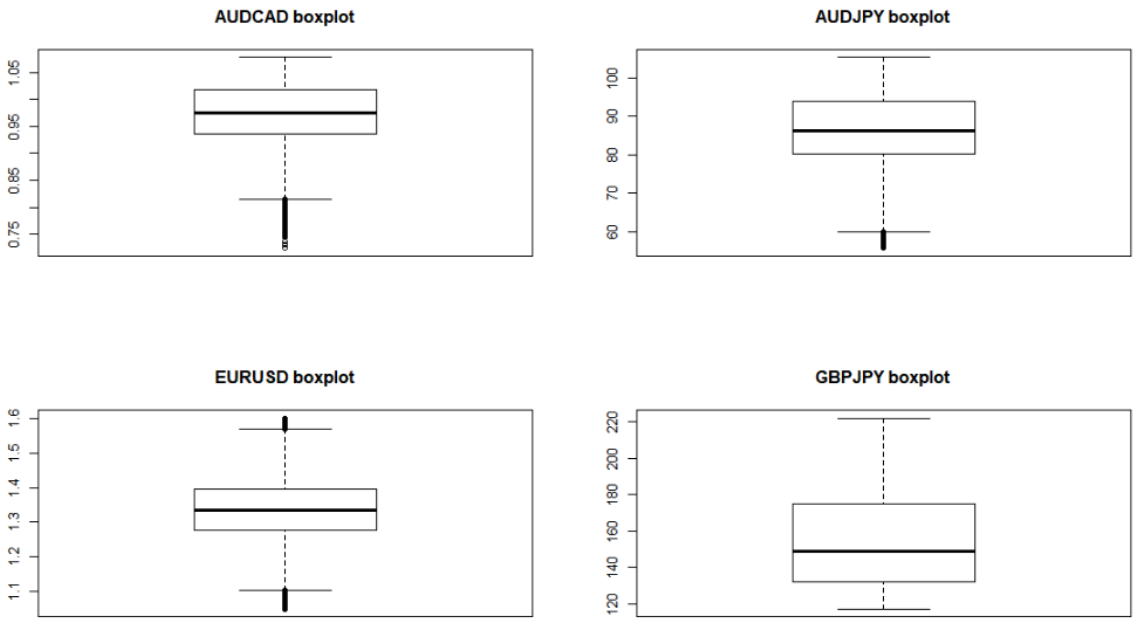

Figura 21: Distribución de precios de cierre según diferentes pares de divisas en el intervalo por horas

Esto indica que un modelo que utilice los rangos de precio de un mercado será inútil en otro mercado, pero cuando se analizan utilizando Z-Score todos muestran una disposición bastante similar (Figura 22):

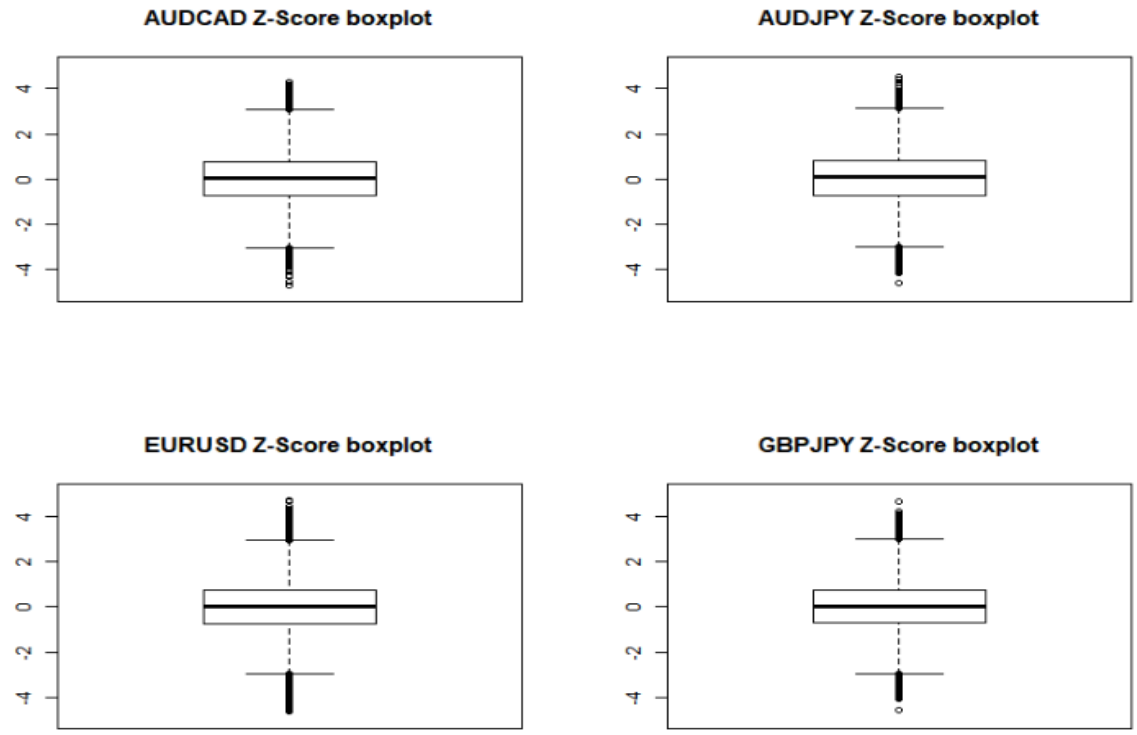

Figura 22: Distribución de Z-Score según diferentes pares de divisas

Un histograma sobre estos Z-Score nos confirma que la distribución de estos mercados es muy similar (Figura 23):

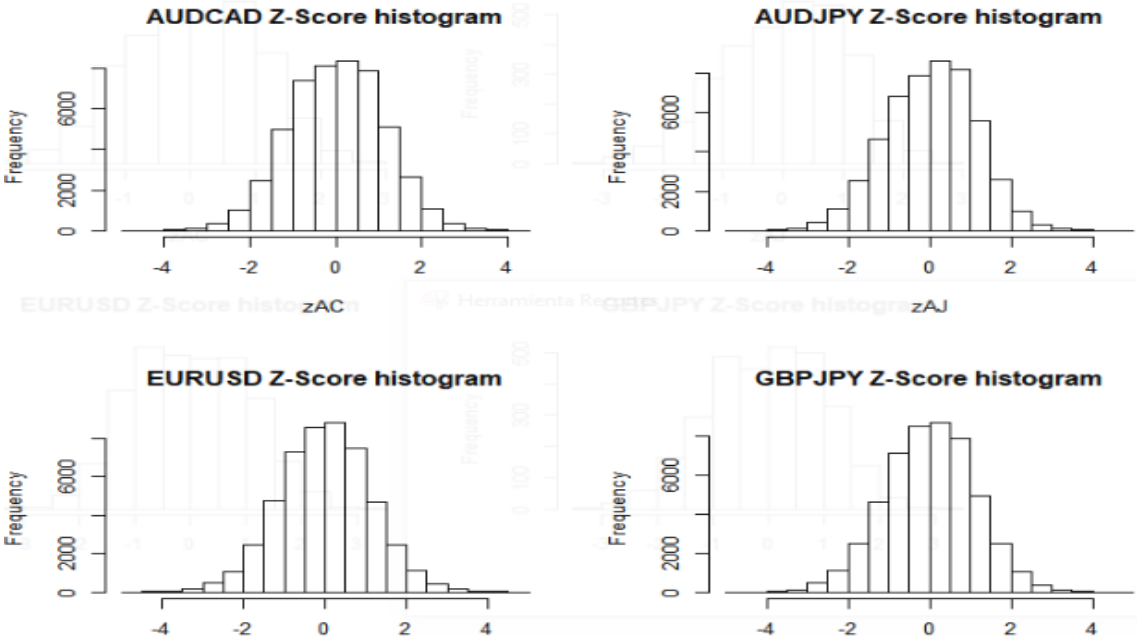

Figura 23: Histogramas sobre la distribución del Z-Score de diferentes pares de divisas en el intervalo por horas

Estas distribuciones se corresponden con el Z-Score del precio de cierre, pero, como se mencionó en el capítulo 2, el Z-Score depende de la media, y en este caso se ha usado la media de las últimas 24 horas para determinar el Z-Score. Esta distribución es un análisis que muestra información interna al día, y con una media de duración de un día. A continuación se procede a analizar estas distribuciones en los intervalos del día y de la semana.

3.3.1 Intervalos diarios

El primer análisis de distribución muestra unas cajas muy similares al del intervalo por horas (Figura 24):

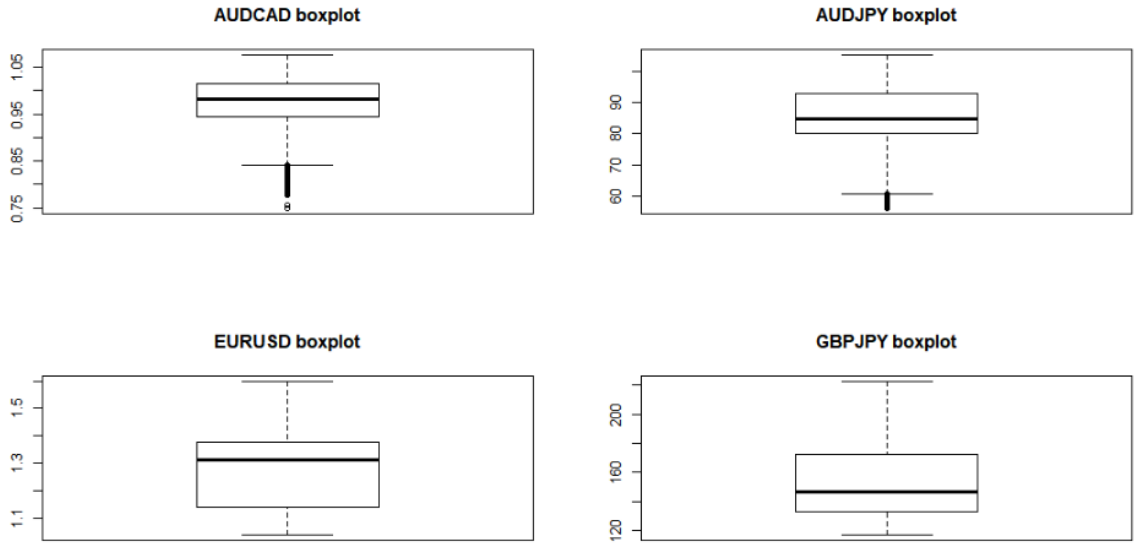

Figura 24: Distribución de precios de cierre según diferentes pares de divisas en el intervalo por días

Esto es de esperarse ya que ambos representan la misma información, pero en diferentes intervalos; pero hay una diferencia fundamental: en el intervalo por horas hay muchos valores extremos (outliers), mientras que en la distribución diaria solo los pares AUDCAD y AUDJPY tienen estos valores extremos.

Esto nos indica que al mirar la serie en intervalos diarios, la cantidad de ruido se disminuye y en efecto, los valores de Z-Score se centran más sobre 0 como se observa en la Figura 25.

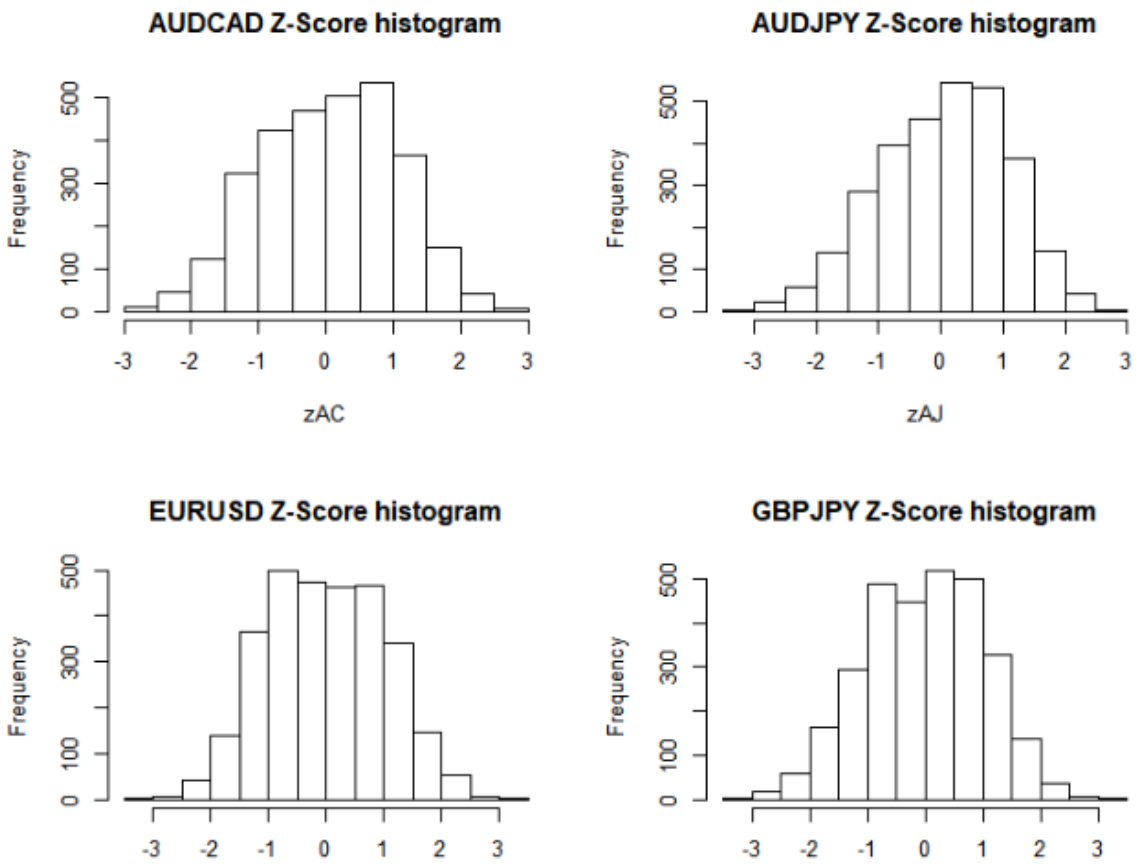

Figura 25: Histogramas sobre la distribución del Z-Score de diferentes pares de divisas en el intervalo por días

3.3.2 Intervalos semanales

Los intervalos diarios reducen mucho el ruido en el análisis, y es de esperarse que los intervalos semanales también lo hagan, pero se cuenta con muchas menos muestras, y las distribuciones ahora muestran valores extremos mucho más notables que en el caso del intervalo diario como se observa en la Figura 26.

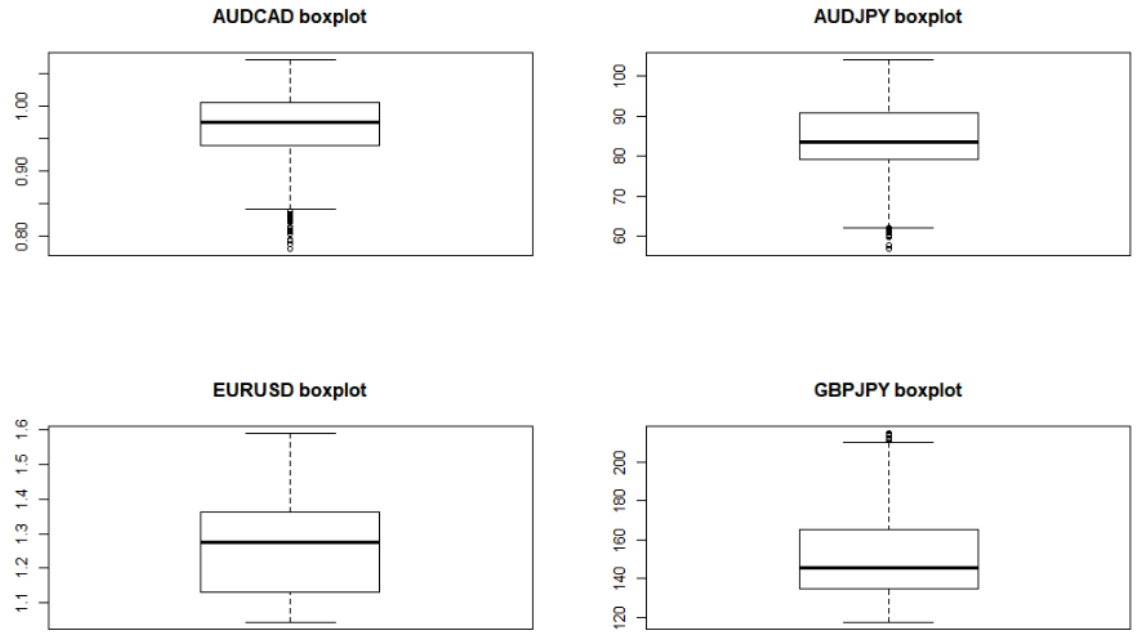

Figura 26: Distribución de precios de cierre según diferentes pares de divisas en el intervalo por semanas

El histograma cambia un poco también puesto que ahora los intervalos se cuentan solo en semanas, pero se puede observar una nueva disminución en los límites del histograma (Figura 27).

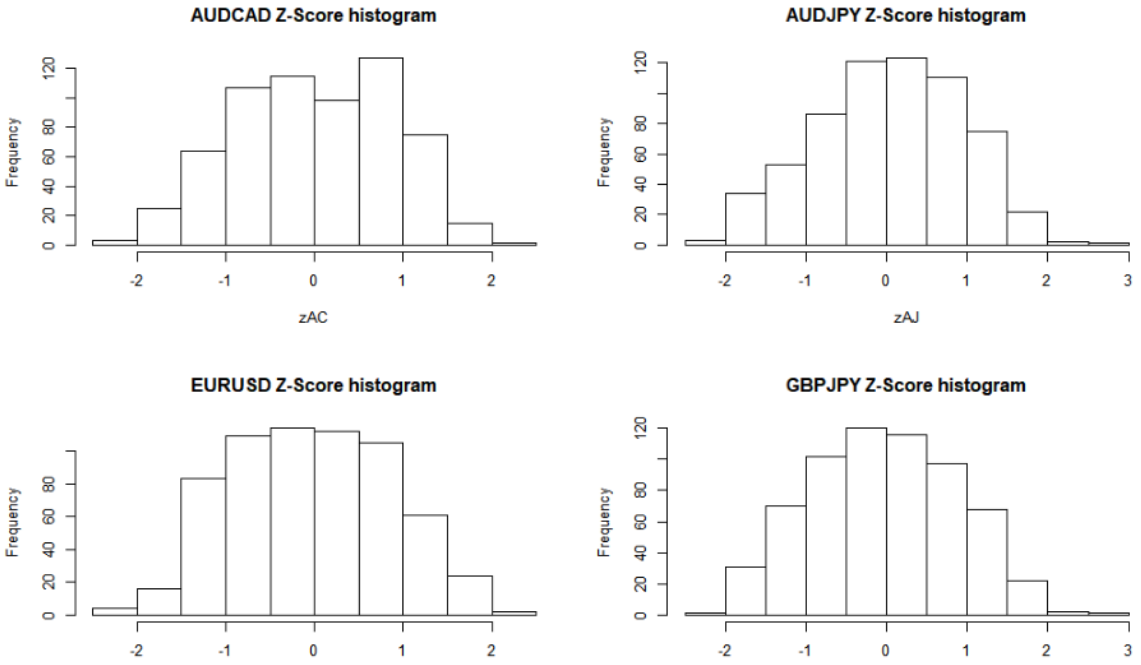

Figura 27: Histogramas sobre la distribución del Z-Score de diferentes pares de divisas en el intervalo por semanas

En el intervalo por hora los valores extremos de Z-Score son equivalentes a 4, en el diario son equivalentes a 3, y aquí en el intervalo semanal el par AUDCAD y EURUSD solo llega hasta 2.5. Esta información implica que dentro del flujo de tiempo, los movimientos horarios pueden llegar a ser muy elevados con respecto a otras horas, pero a medida que analizamos intervalos mayores se observa cada vez precios más cercanos a la media.

3.3.3 Interacción entre intervalos

El análisis de múltiples intervalos sobre la misma serie en el mercado de divisas es muy interesante, como se mencionó en el capítulo 2. Este análisis debe tener en cuenta un detalle importante sobre la forma en que se representan los datos en estos intervalos: el precio de cierre y el precio de apertura.

Si al analizar el precio de cierre de un intervalo por horas, utilizáramos la información histórica del precio de cierre del día o de la semana, efectivamente tendríamos en el análisis información del futuro. Es posible realizar el análisis utilizando el precio de cierre del intervalo anterior, o utilizando el precio de apertura, que en muchos casos será equivalente.

Análisis sobre intervalos mayores contienen menos ruido y son mejores para determinar cálculos de tendencia, pendiente y expectativas a largo plazo, mientras que el análisis sobre el intervalo menor va a poseer mayor ruido, Z-Scores más extremos, pero más datos a estudiar y mayores probabilidades de aplicación práctica en los mercados debido a que el intervalo es menor y tiene menor separación entre el máximo y mínimo.

Capítulo 4

Preprocesamiento de los datos

Una vez obtenidos los datos y realizado un estudio de sus características es importante preparar los datos para los modelos de Machine Learning a utilizar.

En la actualidad existen múltiples técnicas para procesar los datos como se observa en la Figura 28.

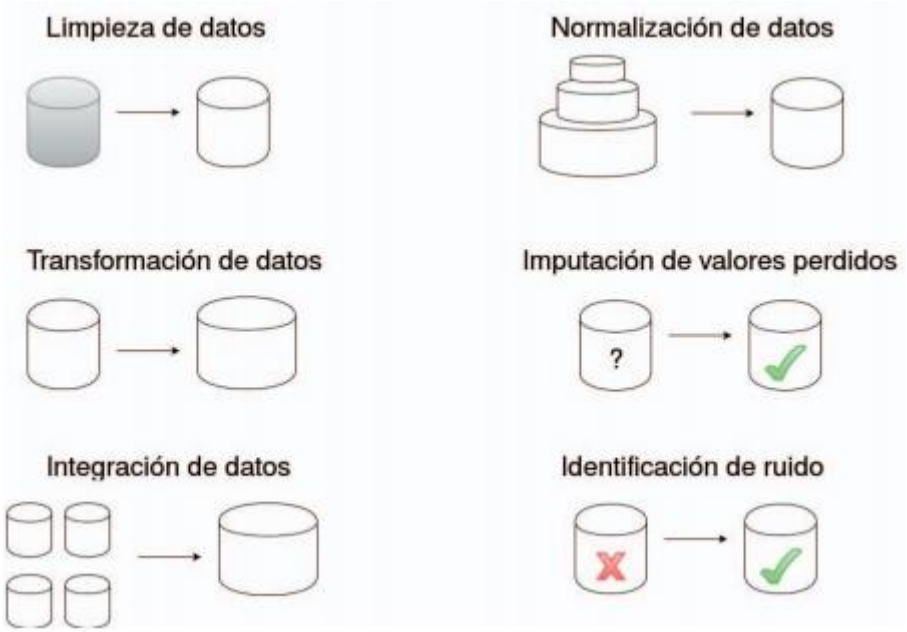

Figura 28: Técnicas habituales de preprocesamiento de datos

El uso de estas técnicas depende del formato de los datos y de las características del problema, y es la parte más difícil del proceso de Machine Learning y donde el factor humano tiene mayor influencia. Para información más detallada sobre todo el conjunto de técnicas de preprocesamiento disponible se recomienda profundizar en [17].

4.1 Limpieza de los datos

El proceso de limpieza de datos es el paso inicial que se realiza al procesar los datos, consiste en asegurar que todos los datos existan y sean compatibles con el problema a resolver. Por ejemplo, si tenemos una variable sexo que admite valores: M y F, cualquier valor diferente es un error y debe ser procesado, incluyendo casos en los que se desconozca el sexo.

El mercado de divisas nunca para, pero no siempre es posible operar con el mismo, más específicamente, en los fines de semana las personas y empresas cesan operaciones sobre el mismo, esto implica que esos datos deben ser eliminados, además existen momentos de crisis internacional o días festivos en los cuales estos mercados cierran. Como ya se analizó en el capítulo 3, en la Figura 14, la cual se muestra nuevamente por conveniencia:

96	04.01.2008 22:00:00.000	1.47501	1.47516	1.47371	1.47397	24279.5977
97	04.01.2008 23:00:00.000	1.47397	1.47397	1.47397	1.47397	0
98	05.01.2008 00:00:00.000	1.47397	1.47397	1.47397	1.47397	0
99	05.01.2008 01:00:00.000	1.47397	1.47397	1.47397	1.47397	0
100	05.01.2008 02:00:00.000	1.47397	1.47397	1.47397	1.47397	0
101	05.01.2008 03:00:00.000	1.47397	1.47397	1.47397	1.47397	0
102	05.01.2008 04:00:00.000	1.47397	1.47397	1.47397	1.47397	0
103	05.01.2008 05:00:00.000	1.47397	1.47397	1.47397	1.47397	0

Figura 14: Fin de semana en el mercado de divisas

Estos datos no aportan al modelo, ya que no hay variación en los mismos y en efecto puede causar problemas en el análisis, ya que al no eliminarlos los cálculos de medias móviles, algoritmos de suavizado, segmentación, Z-Score y otros tendrían un conjunto de valores fijos que influenciarían drásticamente en el valor de los mismos.

Se puede concluir que una fila de datos debe ser eliminada si la variable de apertura es equivalente a la variable de cierre anterior y a su vez es equivalente al precio máximo, mínimo y de cierre, con volumen 0.

Es necesario asegurar todas las condiciones, puesto que es muy raro, pero hay veces de mucha inactividad en el mercado, especialmente cuando se analiza datos anteriores al 2010, donde el volumen es 0, pero el precio llega a variar debido a la influencia de otras entidades en el precio. Es importante recordar que el volumen solo representa la información de volumen de la entidad, por lo que es perfectamente posible que en una hora la entidad haya tenido 0 volumen, pero en otras entidades se hayan realizado operaciones y eso se refleja en el precio.

Este procesamiento es válido si se va a trabajar con una serie individual, pero en esta investigación se realiza operación sobre diferentes intervalos y se busca realizar un estudio comparativo usando los Z-Score, por lo que es imprescindible que todos los mercados tengan la misma longitud, es decir, los mismos días operativos, aunque dentro de estos días algunos intervalos no tengan variación, como se observa en la Figura 29.

31.12.2008 21:00:00.	0.85746	0.86488	0.84516	0.86429	371.456
31.12.2008 22:00:00.	0.86429	0.86518	0.85464	0.85647	280.5556
31.12.2008 23:00:00.	0.85638	0.86032	0.85638	0.86032	8.7201
01.01.2009 00:00:00.	0.86032	0.86032	0.86032	0.86032	0
01.01.2009 01:00:00.	0.86037	0.86037	0.86037	0.86037	0.1241
01.01.2009 02:00:00.	0.86037	0.86037	0.86037	0.86037	0
01.01.2009 03:00:00.	0.86037	0.8604	0.86033	0.86033	18.362
01.01.2009 04:00:00.	0.86037	0.8604	0.86033	0.86037	21.1254

Figura 29: Precios del mercado de divisas sin variación

Estos datos corresponden al fin del 2008, un día donde el mercado estaba abierto, pero existía poca participación, en caso de operar con la serie de forma individual es posible eliminar estos dos intervalos que no aportan información, pero no es posible eliminarlos al hacer un análisis conjunto, puesto que es necesario poseer información sobre cada intervalo, para poder asegurar consistencia.

Debido a esta condición se procede a eliminar solo filas donde el volumen sea 0 en esa fila y la anterior, junto con el precio sin variar en ninguno de los pares de divisas, efectivamente eliminando solo días donde no se haya realizado ninguna operación en el mercado internacional.

A pesar de la simplicidad de este proceso, hay algunas cuestiones importantes que deben ser tratadas:

- La longitud de los días.
- Datos imprecisos.
- Actividad inconsistente en algunos periodos.
- Momentos de operación del día.

La longitud de los días requiere tratamiento por el proceso de horario de verano en algunos países, para que no haya afectaciones en el trading, la hora extra se añade como doble en un fin de semana o se elimina en un fin de semana, como se observa en la Figura 30 y la Figura 31:

2136	29.03.2008 22:00:00.000	0.7917	0.7917	0.7917	0.7917	0
2137	29.03.2008 23:00:00.000	0.7917	0.7917	0.7917	0.7917	0
2138	30.03.2008 00:00:00.000	0.7917	0.7917	0.7917	0.7917	0
2139	30.03.2008 01:00:00.000	0.7917	0.7917	0.7917	0.7917	0
2140	30.03.2008 03:00:00.000	0.7917	0.7917	0.7917	0.7917	0
2141	30.03.2008 04:00:00.000	0.7917	0.7917	0.7917	0.7917	0
2142	30.03.2008 05:00:00.000	0.7917	0.7917	0.7917	0.7917	0
2143	30.03.2008 06:00:00.000	0.7917	0.7917	0.7917	0.7917	0

Figura 30: Horario de verano en el mercado de divisas

7176	25.10.2008 23:00:00.000	0.79315	0.79315	0.79315	0.79315	0
7177	26.10.2008 00:00:00.000	0.79315	0.79315	0.79315	0.79315	0
7178	26.10.2008 01:00:00.000	0.79315	0.79315	0.79315	0.79315	0
7179	26.10.2008 02:00:00.000	0.79315	0.79315	0.79315	0.79315	0
7180	26.10.2008 02:00:00.000	0.79315	0.79315	0.79315	0.79315	0
7181	26.10.2008 03:00:00.000	0.79315	0.79315	0.79315	0.79315	0
7182	26.10.2008 04:00:00.000	0.79315	0.79315	0.79315	0.79315	0
7183	26.10.2008 05:00:00.000	0.79315	0.79315	0.79315	0.79315	0

Figura 31: Horario de verano en el mercado de divisas (2)

La fuente de datos no tiene datos imprecisos, sin embargo como se pudo observar anteriormente existen intervalos inconsistentes, incluso en los fines de semana, existiendo a veces momentos con un volumen extremadamente bajo y una variación mínima de precio, debido a latencia en los sistemas a la hora de cerrar el sistema. Un caso extremo de intervalos inconsistentes es el par AUDUSD anterior al 2011, donde los mercados cerraban por 54 horas durante el fin de semana, en vez de 48, necesitando un tratamiento especial.

La eliminación de días donde no se realizan operaciones depende también de los intervalos donde no ocurren operaciones en sí, el mercado cierra en dependencia del horario de verano a las 10 o a las 11 de la noche del viernes hasta las 9 o las 10 de la noche del domingo (GMT+0100). Esto quiere decir que no se puede eliminar el día en sí, sino los intervalos continuos no operados de estos días, debido al desplazamiento de los mismos.

Al final, al terminar con la limpieza de datos se eliminan todos los datos completamente muertos en el mercado internacional, pero varios pares de divisas van a conservar pequeños intervalos con inactividad por consistencia, esto debe tenerse en cuenta luego en el preprocesamiento y entrenamiento del modelo final.

4.2 Tratamiento y acotamiento de valores extremos y ruido

El mercado de divisas cambia con el tiempo, pero eso no quiere decir que en momentos determinados no existan situaciones inesperadas o extremas, causados normalmente por situaciones internacionales.

El análisis de Z-Score muestra que la gran mayoría de valores están contenidos en 4 desviaciones estándar o menos, pero existen algunos casos en donde el movimiento de

precio es muy drástico, como se observa en la Figura 32.

Figura 32: Movimientos particularmente drásticos en el mercado de divisas

En la imagen se muestra el precio en el medio, con su típica inestabilidad, la media móvil del mismo con una longitud de 20 intervalos y una nube conocida como la banda de bollinger con el grosor de 4 desviaciones estándar. Se ven dos ejemplos donde el movimiento fue muy drástico, efectivamente por encima de 4 desviaciones.

Como es necesario conservar consistencia en todos los intervalos no es posible eliminar los valores extremos del conjunto de datos, pero sería posible acotar el valor. Acotar el valor tiene una consecuencia fundamental, la modificación de valores dependientes del precio subyacentes, como el Z-Score o las medias móviles, además de afectar la precisión del modelo, puesto que al acotar una caída, un intervalo que fuera una caída seguido de una recuperación sería representado por dos intervalos de caída.

No es recomendable modificar los datos de precio original ni de acotar los valores extremos, puesto que estos movimientos drásticos ocurren con naturalidad en el mercado y son importantes para el aprendizaje de modelos y usados en ciertas estrategias de trading.

4.2.1 Suavizado de las series creadas

A pesar de no ser recomendable modificar los valores originales, sí es recomendable

utilizar diferentes técnicas para obtener información adicional sobre los mismos.

La técnica más utilizada en el mundo del trading y de negocio de mercados es el suavizado de la serie temporal usando una, o múltiples, medias móviles para disminuir el ruido y permitir un análisis de pendientes y tendencias mucho más fácil y eficiente.

Las medias móviles son indicadores que aplanan o pulen, en mayor o menor medida, el progreso de los precios, de tal forma que eliminan determinadas oscilaciones, sean a corto, medio o largo plazo [18].

Su forma de cálculo depende del tipo de media móvil, pero la idea es obtener una media del valor actual basado en *n* valores anteriores. El valor y tipo de media móvil depende en muchas ocasiones de este *n*, un *n* muy pequeño causa que se descarte información pasada que podría ser de utilidad en el instante de tiempo que se quiere predecir, mientras que un *n* muy elevado implicaría usar valores que posiblemente no tengan mucha relación con el momento de análisis.

Su eficacia en el mercado financiero ha sido estudiada en muchas ocasiones con resultados generalmente favorables, como se puede observar en [19]. Es una técnica sencilla, pero que ofrece muchas facilidades al análisis posterior de los datos.

Existen múltiples técnicas de suavizado usadas comúnmente:

- Media móvil exponencial (EMA por sus siglas en inglés)
- Media móvil simple (SMA por sus siglas en inglés)
- Media móvil ponderada (WMA por sus siglas en inglés)
- Media móvil de Hull (HMA por sus siglas en inglés)

Estas son las técnicas más comunes porque cubren las variaciones y tienen propiedades deseables para cada tipo de estrategia y análisis. Existen muchos otros tipos de medias móviles usadas, y es posible crear medias móviles usando formulas propias.

La EMA se calcula otorgando un peso exponencial a los elementos más cercanos al momento actual, algo similar a lo que se observa en la Figura 33 donde *n* es 15:

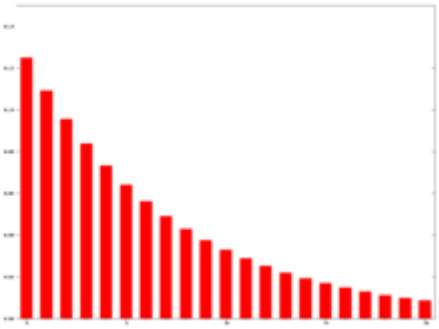

Figura 33: Gráfica de pesos de una media móvil exponencial

Los 5 valores anteriores al intervalo actual tienen la mayor influencia sobre el valor de la media móvil del intervalo actual.

En el caso del SMA, se calcula similar a una media común y corriente, dado n = 15, no sería más que sumar los 15 intervalos anteriores y dividir por 15.

La media móvil ponderada (WMA) es una variación del SMA, donde diferentes intervalos tienen diferentes pesos, típicamente intervalos recientes tienen más peso. La EMA es una variación de la media móvil ponderada.

De forma similar, la HMA es una variación de la media móvil ponderada, donde se utilizan pesos cuadráticos. Esto permite que la media móvil se mueva muy rápido con el precio.

En la Figura 34 se aprecia como la EMA, SMA y HMA interactúan con el precio:

Figura 34: EMA, SMA, HMA con n = 30

Se puede apreciar a simple vista que a pesar de todo el zigzag del precio, las medias móviles mantienen mayormente una pendiente negativa hasta el fin de octubre.

En este trabajo se utilizarán la EMA, SMA y WMA, ya que la HMA y otras medias móviles igualmente volátiles tienen cambios de pendientes muy bruscos y no son de utilidad para el análisis de tendencia, sino que son usados más bien para determinar puntos precisos de compra y venta en sistemas de High Frequency Trading (HFT).

Para la longitud de Z-Score corto, una *n* de 15 es un número bastante bueno en todos los intervalos. La sesión activa del mercado dura 13 horas entre la sesión oficial de Europa y Estados Unidos, típicamente la hora anterior y la hora posterior tienen igualmente actividad, por lo que esta *n* de 15 cubre la sesión en su totalidad en el rango de horas. 15 días es la mitad promedio de un mes, y 15 semanas son 105 días, el cual al eliminar los intervalos muertos por fines de semana nos da casi un trimestre de precisión.

Para la longitud larga, tras varios experimentos se tomó *n* equivalente a 50. Este número no es más que porque es usado comúnmente en la comunidad por ser la mitad de 100, y un número lo suficientemente largo como para almacenar información de eventos pasados e históricos, pero no tan largo como para quitar peso a los eventos actuales. En general se obtienen buenos resultados en los 3 intervalos, pero podría utilizarse una *n* diferente y obtenerse resultados similares.

Luego el suavizado posterior de estos Z-Score fue con una EMA de *n* equivalente a 5 y una *n* de 20 respectivamente, esto da unas curvas limpias y con poco ruido.

Para las SMA, EMA y WMA se usó una *n* de 15 y de 100 para el intervalo corto y largo respectivamente. La longitud 15 fue seleccionada por la misma razón que se explicaba anteriormente, mientras que 100 es una longitud de *n* muy frecuentemente usada en el mercado como un punto de inflexión debido a la pendiente de curvas con una *n* de 100. Un precio que lleva semanas en descenso tendrá una SMA/EMA/WMA muy alta, y al cruzarlo normalmente indica que ya el descenso termina y que puede hacer una recuperación en V o una consolidación antes de la recuperación.

4.3 Creación de series representativas

Al añadir el Z-Score y las diferentes medias móviles con varios *n* a una serie de datos se le añade bastante información técnica útil sobre la misma reduciendo significativamente el ruido en los datos.

Como se observó en 2.2.1, la adición de información de intervalos mayores pueden

ayudar significativamente a determinar precios futuros, y la información de pendiente, media, Z-Score y otros de estos intervalos mayores también aportan mucho a la hora de realizar un estudio de los datos.

En la sección 2.1.2 se analizó la dualidad de información presente en los pares de divisas, ya que el precio inherente a cada una de las dos monedas representadas en el par es desconocido. Poder estimar el valor real y la tendencia de las dos monedas representadas en el par sería de mucha ayuda para determinar precios futuros.

Una posible forma de estimar este valor es mediante el uso de una media de Z-Score entre los 14 pares que contienen las dos monedas. En una gráfica quedaría representado como aparece en la Figura 35.

Figura 35: Uso de Z-Score interpolado del Euro y el dólar (inferior) en el análisis del par EURUSD

Se puede observar como en el inicio de la gráfica el euro se encontraba por encima del dólar americano y sin mucha variación, mientras que el dólar americano descendía. Esto se refleja en el par EURUSD con un aumento de valor del euro contra el dólar americano. Al llegar a un Z-Score inferior a -1.5, poco tiempo después el dólar empieza a recuperarse y el par desciende un poco de valor, mientras que a finales de Abril e inicio de Mayo se observa el dólar americano aumentar drásticamente de valor mientras que el euro cae, y esto se refleja en el par EURUSD de forma inmediata.

Este concepto no es a prueba de fallos ni tiene precisión absoluta, como se observa al final de la gráfica, donde el Euro supera al dólar, pero el par no recibe mucha influencia.

Esto se debe a que el Z-Score se calcula a partir de las medias móviles, y estas son dependientes del valor *n* utilizado, con un valor *n* más elevado se pierde un poco de reacción, pero se observa una tendencia mucho más clara (Figura 36).

Figura 36: Uso de Z-Score interpolado del Euro y el dólar (inferior) en el análisis del par EURUSD con n = 500

La estimación del Z-Score de una divisa se puede inferir a partir de la media de los 7 Z-Score de los pares donde esta divisa sea la base, es decir donde se encuentre a la izquierda del par. En el caso de EURUSD, la divisa base es el euro. En el caso de los pares donde se encuentra a la derecha, se utiliza el complemento del Z-Score, es decir (-1 * Z-Score) del par.

Por ejemplo, para el Z-Score del Euro, la formula sería:

$$\frac{ZScore\ EURGBP\ +\ ZScore\ EURUSD\ +\ \ldots\ +\ ZScore\ EURJPY\ (para\ los\ 7\ pares)}{7}$$

Mientras que en el caso del yen japonés por ejemplo sería:

$$\frac{ZScore\ USDJPY * -1\ +\ ZScore\ AUDJPY * -1\ +\ \ldots\ +\ ZScore\ EURJPY * -1\ (para\ los\ 7\ pares)}{7}$$

Esta estimación al depender solamente del Z-Score no es afectada por la desviación o

el valor real de los pares, sino que es totalmente relativo y consistente con las características del Z-Score.

Existen otras formas de inferir el valor real de una divisa, como son los índices o el mercado de futuros, el DXY, EXY, 6E1!, etc…

El inconveniente de estos métodos es que no son relativos, por ejemplo, en un mismo intervalo de tiempo el DXY tiene 96.49, mientras que el EXY tiene 113.5, para poder compararlos habría que realizar conversiones relativas, y esta información no se encuentra representada en los pares, ni está disponible como datos para su procesamiento. Estos índices además tienen pesos fijos, el DXY tiene 57.6% de su peso en el par EURUSD y solo 3.6% en el par USDCHF, lo que implica que en el caso de un colapso drástico del dólar contra el franco suizo, este índice no se movería casi hasta que dicho colapso fuera replicado en el par EURUSD. Usando la información de precio real relativa a los Z-Score propuesta anteriormente no es necesario el ajuste de pesos, y se puede comparar estos índices directamente para determinar que divisa es más fuerte en un intervalo dado y su tendencia actual.

Finalmente la aplicación de medias móviles sobre estas divisas permite un análisis típicamente imposible sobre los pares de divisas. Si el valor relativo de la divisa base está por encima de su media móvil y con una pendiente positiva mientras que el de la otra divisa del par está por debajo de su media móvil y con una pendiente negativa entonces el par debe aumentar mucho en precio, mientras que en caso contrario el precio del par debe disminuir mucho. En casos mixtos, como que el precio de una divisa caiga por debajo de su media móvil pero siga siendo mayor que el precio de la otra divisa se podría considerar que el impulso inicial ha terminado y que pronto podría ocurrir una reversión, especialmente si el Z-Score está cerca de los extremos. En la Figura 37 se muestra este tipo de análisis y los beneficios que presenta.

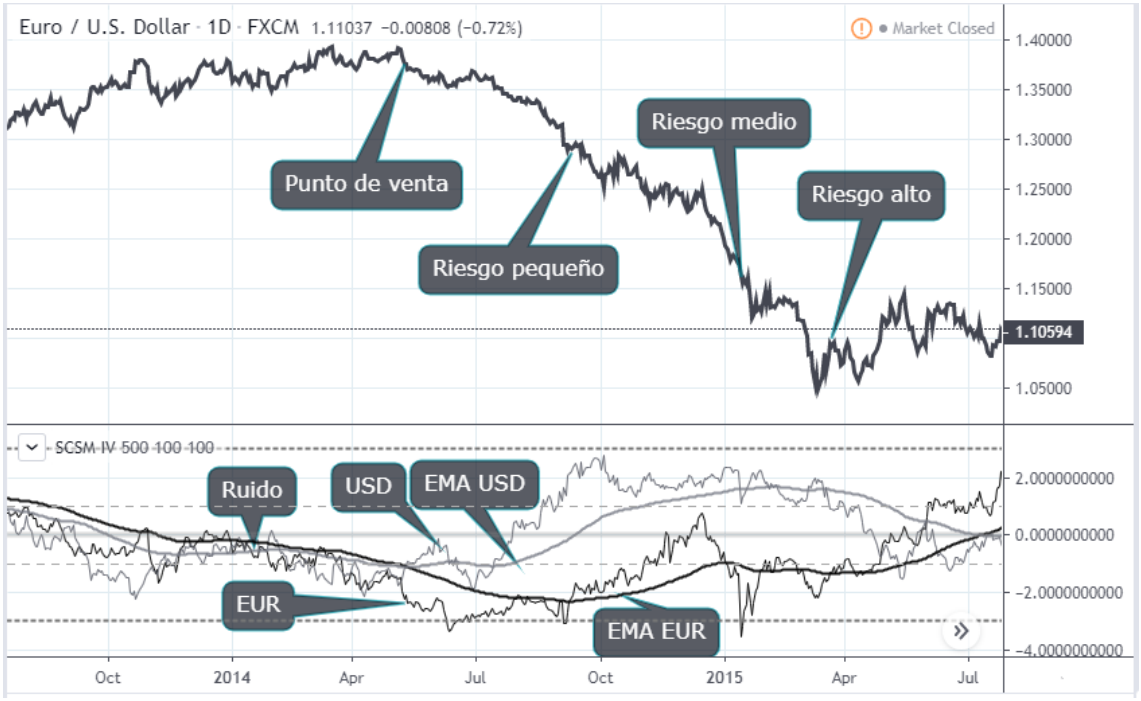

Figura 37: Representación de riesgo de operaciones basado en el uso de Z-Score interpolado y suavizado del par de divisas EURUSD

El punto de venta inicial ocurre al salir de una zona de ruido donde los precios y sus medias móviles se encontraban fuertemente entremezclados; al aumentar el precio del dólar por encima de su media móvil y el del euro caer por debajo de su media móvil, dando una fuerte señal de que el precio caerá en consecuencia. Efectivamente el precio comienza a caer, y al llegar al punto de riesgo pequeño el euro comienza a recuperarse, pero el dólar aún conserva impulso y su pendiente y Z-Score es superior por lo que el precio del par sigue cayendo. Llegado el punto de riesgo medio ya el dólar americano tiene un Z-Score de 1.5, está cayendo por debajo de su media móvil y disminuyendo su impulso, por lo que es probable que el movimiento termine pronto. En el punto de riesgo alto ya el dólar americano está por debajo de su media móvil mientras el euro está por encima de su media móvil y reduciendo distancias indicando que es muy probable que pronto el par comience a aumentar en valor.

4.4 Resumen del preprocesamiento de datos

Con todo lo anterior se considera que para realizar un análisis de la serie por intervalos por horas, es recomendable solo limpiar los datos y asegurar la consistencia de intervalos, además de añadir a los mismos varias columnas adicionales conteniendo la siguiente información:

- SMA, EMA, WMA del par, con al menos 2 *n* diferentes.
- El Z-Score del par, posiblemente con 2 o más *n* diferente de longitud y al menos una media móvil de dichos Z-Score.
- El Z-Score de las divisas que componen el par, y al menos una media móvil de los mismas.
- La información de los dos puntos anteriores, pero de los intervalos superiores. Dígase la SMA, EMA, WMA del par del día y de la semana, así como el Z-Score de las divisas y una media móvil de las mismas del par del día y de la semana. Esta información podría ser resumida en variables categóricas si el modelo a utilizar obtiene mayor precisión de esta forma.

Capítulo 5

Segmentación de las series temporales

La segmentación es una técnica utilizada en el ámbito de la minería de series temporales para trabajar con datos que posean ruido o que sean muy numerosos. Permite la simplificación de la serie en segmentos que la representen con la mayor precisión posible reduciendo la cantidad de datos a ser procesados por algoritmos posteriores. La agrupación de puntos en diferentes segmentos también permite realizar análisis estadísticos sobre dichos segmentos y buscar características comunes en los mismos.

La segmentación no es más que una técnica de preprocesamiento avanzada especializada en serie temporales y solo cambia el formato de datos a ser procesado por pasos posteriores en el flujo de trabajo.

5.1 Tipos de Segmentación

El objetivo de la segmentación, como se mencionaba anteriormente, es dividir la serie temporal en segmentos, para esto es necesario definir los puntos de corte que separan los diferentes segmentos. Existen en la actualidad muchos algoritmos y formas de trabajo para determinar dichos puntos de corte y crear los segmentos.

Las tres formas de trabajo más habituales y que contienen la mayoría de los algoritmos en la literatura son:

- Top Down.
- Bottom Up.
- Sliding Window.

Los algoritmos Top Down están enfocados en la búsqueda de todo el espacio de trabajo desde un punto de vista genérico hacia uno específico. Se parte de un solo segmento que almacena toda la serie temporal en un solo segmento y la va dividiendo en sub-segmentos según criterios determinados por el investigador.

Los algoritmos Bottom Up parten de dos puntos como un solo segmento, y se va extendiendo hasta alcanzar ciertos criterios, a partir del cual comienza otro segmento, y así hasta cubrir todo la serie temporal.

Finalmente los algoritmos basados en Sliding Window, también conocido como algoritmo de ventana deslizante, parten de un punto y van añadiendo puntos del futuro

hasta alcanzar un punto que no cumpla el criterio del segmento, a partir del cual comienza un nuevo segmento.

Cada forma de trabajo tiene ventajas y desventajas, y tienen criterios y formas de trabajo muy estudiadas y con resultados excelentes en diferentes campos.

Los algoritmos Top Down se usan mucho con PIP y con TP, acrónimos para Perceptually Important Points y Turning Points, es decir, puntos que son importantes para la percepción y puntos donde hay un cambio en la tendencia actual. [20] [21] [22]

Este tipo de algoritmo ha mostrado resultados impresionantes al ser combinados con algoritmos genéticos para determinar los puntos de corte ideales para la segmentación, y es excelente para un análisis de eventos pasados.

Los algoritmos Bottom Up tienden a utilizarse con algoritmos metaheurísticos similares a los usados en Top Down, pero su punto de partida es diferente y el procedimiento a seguir es diferente puesto que es necesario realizar muchas uniones y conexiones entre puntos o incluso entre segmentos.

Presentan una precisión excelente al partir desde la base, pero puede tomar un tiempo considerable en segmentar toda la serie temporal, especialmente en el caso de millones de puntos.

Los algoritmos basados en ventana deslizante tienen menos resistencia al ruido que los anteriores y son menos precisos, pero son muy veloces y permiten trabajar con datos en tiempo real.

Como hemos indicado anteriormente la segmentación basada en ventana deslizante es capaz de funcionar en tiempo real ya que el valor del punto actual depende solo de valores pasados y no de valores futuros. Esta característica es muy importante si se quiere crear un modelo y utilizarlo en los mercados de divisas en tiempo real.

Esta investigación utiliza el método de ventana deslizante como base para crear un algoritmo de segmentación basado en fuerza e impulso, y que permite establecer el nivel de ruido permisible y cuanto tiene que disminuir el impulso antes de considerar que ha terminado el segmento y que es necesario comenzar un nuevo segmento.

El algoritmo propuesto en este trabajo para determinar los puntos de corte de cada segmento funciona de forma progresiva, analizando la fuerza de cada punto y determinando si el punto es parte del segmento actual o si forma parte de un nuevo segmento. Veamos en el siguiente apartado como se calcula la fuerza de un punto del segmento en el algoritmo propuesto.

5.2 Fórmula y método de cálculo para la fuerza

La fórmula para determinar la fuerza de un punto que se usa en esta investigación es bastante sencilla, pero comprende varios elementos, por lo que es necesario explicar cada parte por separado:

$$Fuerza = Impulso + F.Medias\ móviles + \frac{F.ZScore\ par + F.ZScore\ divisas}{2}$$

El impulso es equivalente a la diferencia del precio con respecto al precio anterior escalado y centrado a 0 a partir de la desviación estándar. Es decir, siendo x un punto de la serie y x-1 el punto anterior:

$$Impulso[x] = \frac{Valor[x] - Valor[x-1]}{desviación}$$

Donde la desviación va a ser la desviación general de la serie temporal, en esta investigación se usó la desviación general de toda la serie ya que no fluctúa mucho en el tiempo, pero en acciones de empresa u otros donde el valor y, por ende, la desviación cambia considerablemente en el tiempo es necesario establecer un rango para el cálculo de la desviación a utilizar. Esta parte de la formula busca obtener información relativa al movimiento actual, el cual marca siempre los inicios y los fines de las tendencias y de los segmentos.

La fuerza de las medias móviles es un poco más complicada, la idea es buscar cuanta separación hay del precio con respecto a su media, pero que a su vez este valor decrezca rápidamente cuando el precio empiece a perder impulso pero sin esperar a que decaiga considerablemente de su máximo.

La fórmula utilizada está dividida en 3 partes, y se aplica sobre los 3 tipos de media móvil y luego se obtiene la media de las mismas:

$$\begin{aligned} F.Medias\ móviles \\ &= \frac{sepSMA15 + sepSMA100 + sepSMA}{3} \\ &+ \frac{sepEMA15 + sepEMA100 + sepEMA}{3} \\ &+ \frac{sepWMA15 + sepWMA100 + SepWMA}{3} \end{aligned}$$

Donde para cada una de las medias móviles se tiene siendo x un punto de la serie y x-1 el punto anterior:

$$sepMA15[x] = \frac{Precio[x] - MAPrecio15[x]}{desviación} * 0.55$$

$$sepSMA100[x] = \frac{Precio[x] - MAPrecio100[x]}{desviación} * 0.2$$

$$sepSMAs[x] = \frac{MAPrecio15[x] - MAPrecio100[x]}{desviación} * 0.35$$

MAprecio15 se refiere al valor de la media movil de longitude 15, mientras que MA100 se refiere al valor de la media móvil de longitud 100. La desviación usada es la misma a la calculada en el impulso. Se hace un ajuste de pesos buscando que la separación entre el precio y la media móvil más cercana tenga más influencia sobre la fuerza, además que este número va a ser más pequeño en la mayoría de las veces que la obtenida por una media móvil más lejana. Además se incorpora la separación entre estas diferentes media móviles como parte de la fórmula, mientras más separado este el precio y la media móvil de 15 de la media móvil de 100 más impulso y más fuerte es la tendencia. A medida que empiecen a acercarse es habitual que los precios empiecen a estabilizarse en un rango.

Para la determinación de la fuerza de los Z-Score del par se sigue una lógica muy similar a la usada en la media móvil, sin embargo por definición el Z-Score es centrado y escalado en 0, así que no es necesario utilizar la desviación.

Primero la fórmula general:

$$FZscore = \frac{SmoothZScore15 + ZScore15separation + SmoothZscore50 + ZScore50separation}{2} + ZScoresSeparations + impulseZscore50 * 1.25 + impulseZscore15 * 1.4$$

Y los elementos que no tenemos calculados de la fase de preprocesamiento:

$$ZScore15separation[x] = (ZScore15[x] - SmZScore15[x])$$

$$ZScore50separation[x] = (\mathrm{ZScore50}[x] - \mathrm{SmZscore50}[x])$$

$$ZScoresSeparations[x] = (\mathrm{ZScore15}[x] - \mathrm{ZScore50}[x])$$

$$impulseZscore15[x] = (\mathrm{SmoothZScore15}[x] - \mathrm{SmoothZScore15}[x-1])$$

$$impulseZscore50[x] = (\mathrm{SmoothZscore50}[x] - \mathrm{SmoothZscore50}[x-1])$$

Los Zscore15 y ZScore50 son los Z-Score obtenidos durante el capítulo 4.2.1, mientras que SmoothZScore15 y SmoothZscore50 son estos Z-score suavizados usando una media móvil de longitud 5 y 20 respectivamente, que dan resultados bastante buenos

según las observaciones realizadas.

La idea detrás de este segmento de la fórmula es determinar el impulso basado igualmente en la diferencia entre el valor de los ZScore y sus medias móviles, así como la separación entre los mismos y la variación del valor con respecto a su valor anterior. Esta variación recibe un poco más de peso que la separación actual puesto que se busca capturar el cambio de un punto en la serie temporal con respecto a los puntos anteriores y si este punto tiene suficientes características y suficiente impulso para cambiar un flujo ya establecido.

Hasta este punto la fórmula puede ser utilizada para cualquier activo o incluso para cualquier tipo de serie temporal donde se quiera buscar este tipo de análisis. El último segmento incorpora información solo disponible en el mercado de divisas.

$$F.ZScore\ divisas\ = (\mathrm{FZscorepos} - \mathrm{FZscoreneg})$$

El FZscorepos y el FZscoreneg se refieren a la fórmula usada anteriormente, pero aplicada a la divisa positiva y a la divisa negativa. En un par como puede ser el EURAUD, al aumentar la fuerza de la divisa del euro, el precio del par EURAUD aumenta, es decir, tiene una correlación positiva, de forma similar, al aumentar la fuerza de la divisa del dólar australiano disminuye el precio del par EURAUD debido a que tiene una correlación negativa.

5.2.1 Método de segmentación y parámetros

Para la segmentación es necesario determinar los puntos de corte, esto se hace de forma progresiva, analizando cada punto y determinando si es parte del segmento actual o si forma parte de un nuevo segmento.

El código es particularmente extenso, pero la lógica es muy sencilla. Simplemente vamos añadiendo puntos a un segmento y vamos midiendo la distancia al punto de inicio del segmento hasta que cruce un umbral de resistencia. Puntos sucesivos deben ir en la dirección de la tendencia establecida, dentro de un límite establecido por la distancia hasta el punto más extremo dentro del segmento. Por ejemplo, si se tiene un segmento que va de 1 a 5, la distancia recorrida es de 4, y si definimos el límite como 20%, entonces 5-4*0.2=4.2, por lo que una vez un punto caiga por debajo de este valor ya se considera como parte de un nuevo segmento y se cierra el segmento anterior.

Los valores que se usen para la resistencia al ruido y para el límite de tendencia influencian enormemente en el resultado final de la segmentación como se puede

observar en la Figura 38 de un segmento del par de divisas AUDCAD:

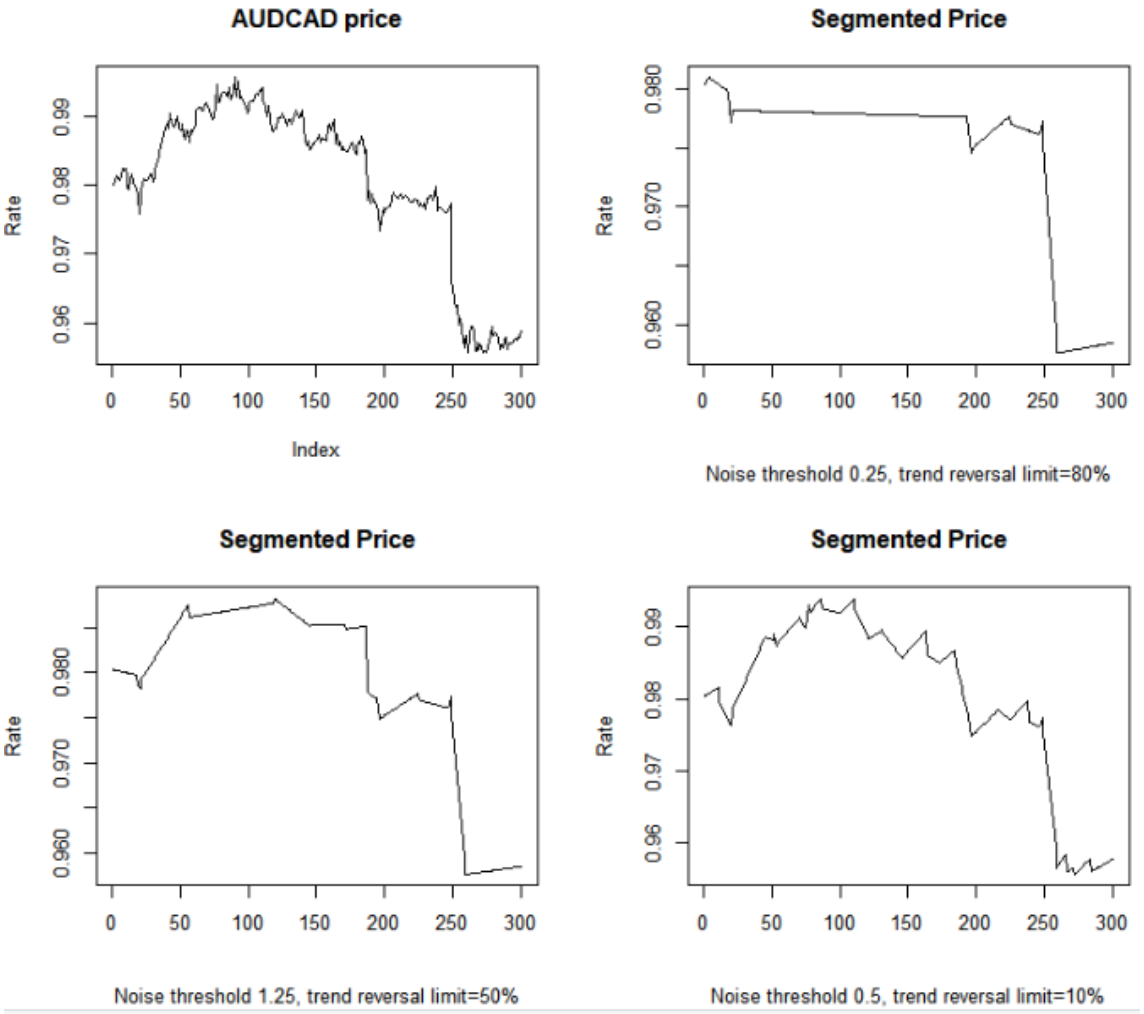

Figura 38: Segmentación de precios con diversos parámetros

Como se puede observar, cambiar los parámetros de resistencia al ruido y de punto límite de la tendencia cambia drásticamente los segmentos obtenidos. Poner una resistencia demasiado baja le da mucha sensibilidad y crea muchos segmentos pequeños con cada pequeño tirón después de un segmento.

Definir el límite demasiado alto, como en la imagen superior derecha de la Figura 38 indica que el precio tiene casi que regresar a su valor inicial antes de crear un nuevo segmento, lo cual no es nada ideal. Dependiendo de la longitud de la serie y el número de segmentos deseados, así como el rango a estudiar, puede ser de interés modificar en mayor o menor medida estos parámetros, pero se recomienda no establecer una resistencia al ruido mayor a dos veces la desviación ni más allá de un 50% el límite de tendencia o se puede perder demasiada información.

En principio los mejores resultados observados fueron con una desviación estándar de

resistencia y 15% de límite. Esto da una compresión entre 10 y 15 de los datos originales y conserva la estructura original de la serie mientras elimina un poco de ruido.

5.2.2 Características, ventajas y desventajas de la segmentación basada en fuerza

La segmentación basada en la fórmula propuesta en la sección anterior, al estar basada en Sliding Window es capaz de funcionar en tiempo real ya que no afecta segmentos pasados y el valor del punto actual depende solo de valores pasados y no de valores futuros. Esta característica es muy importante si se quiere crear un modelo y utilizarlo en los mercados de divisas en tiempo real.

Permite además definir el valor de compresión basado en cuanto ruido se quiere tolerar y cuanto puede caer el precio antes de definir un punto de inflexión, pero estos cálculos no están definidos por el valor exacto de precio, sino por la fuerza subyacente, por lo que es posible que cierre el segmento muy cerca de los valores máximos si el precio se consolida cerca del extremo.

Al estar basado completamente en valores relativos, es fácil ajustar los parámetros y ser usado en otras series temporales y mercados.

Por otra parte, es muy sensible a los valores que se usen como parámetros, rara vez cambia de segmento en el punto extremo ideal, (aunque en valores óptimos cercanos) y, por definición, siempre es necesario analizar un punto para saber si el punto pasado fue el fin del segmento, no tiene forma de determinar si el mismo punto en cuestión es el final del segmento o no.

La otra ventaja es que determina los segmentos de forma automática, ajustando la longitud según sea necesario con cálculos mínimos, en este pequeño segmento de la Figura 39 vemos la comparación entre el precio original, el precio segmentado con una resistencia de 0.8 y límite de 15% y el precio dividido en 5 segmentos de igual longitud.

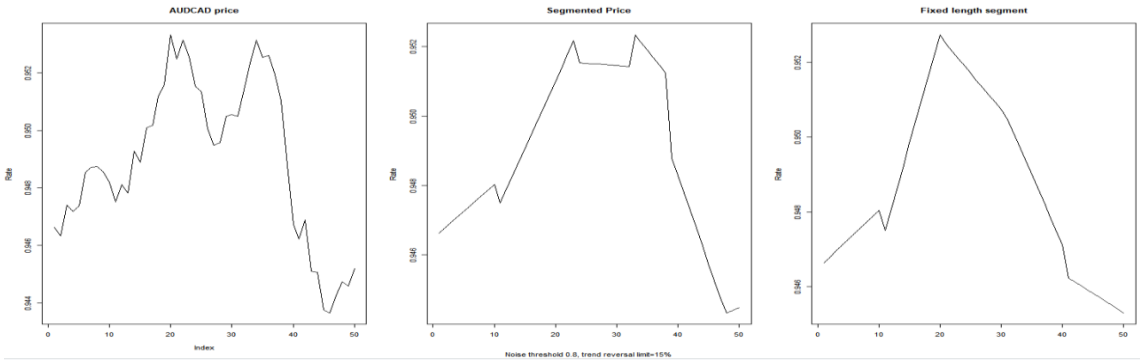

Figura 39: Comparación de segmentación dinámica contra segmentación estática

Se puede observar que la segmentación propuesta es significativamente mejor que la realizada por intervalos fijos. También se observa como toda la curva en el medio queda eliminada en la segmentación, ya que si bien el precio cae, la fuerza no llega a caer lo suficiente como para cruzar el umbral de ruido, tomando toda la curva como un único segmento.

Capítulo 6

Aprendizaje no supervisado. Clustering.

El aprendizaje no supervisado es parte del proceso de Machine Learning (Figura 40) y se enfoca en la detección de características similares entre conjuntos de datos.

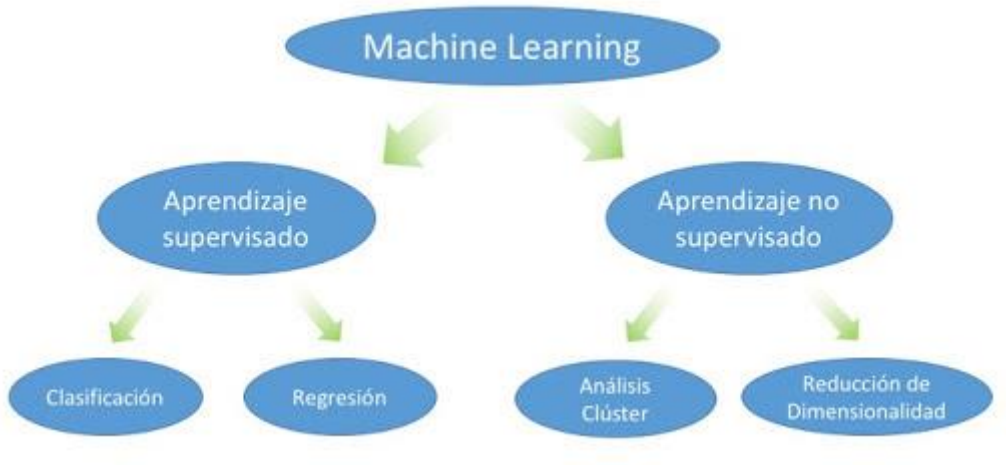

Figura 40: Tipos de aprendizaje de Machine Learning

Estas características similares permiten agrupar datos sin la necesidad de expertos y es muy importante cuando la clasificación de los datos es desconocida.

En el caso de la investigación, al reducir el ruido y el número de datos mediante la segmentación, se puede proceder a analizar los datos de diferentes formas.

Una de las alternativas es la utilización de series temporales simbólicas, para lo cual es necesario clasificar los diferentes segmentos que componen la serie en diferentes grupos [23]

Existen dos formas prácticas de clasificar dichos segmentos y realizar la conversión a series temporales simbólicas:

- Clustering [24]
- Aprendizaje semi-supervisado [25]

El aprendizaje semi-supervisado requiere de expertos y puede introducir bias en el aprendizaje por lo que no se considerará.

6.1 Tipos de clustering

Existen dos tipos fundamentales de clustering:

- Clustering jerárquico
- Clustering particional

En el Clustering jerárquico se van conectando los datos según la distancia entre ellos en un árbol de jerarquía. El método de conexión y la forma de medir la distancia afecta el resultado del algoritmo substancialmente. Se caracteriza por su forma de árbol como se observa en la Figura 41.

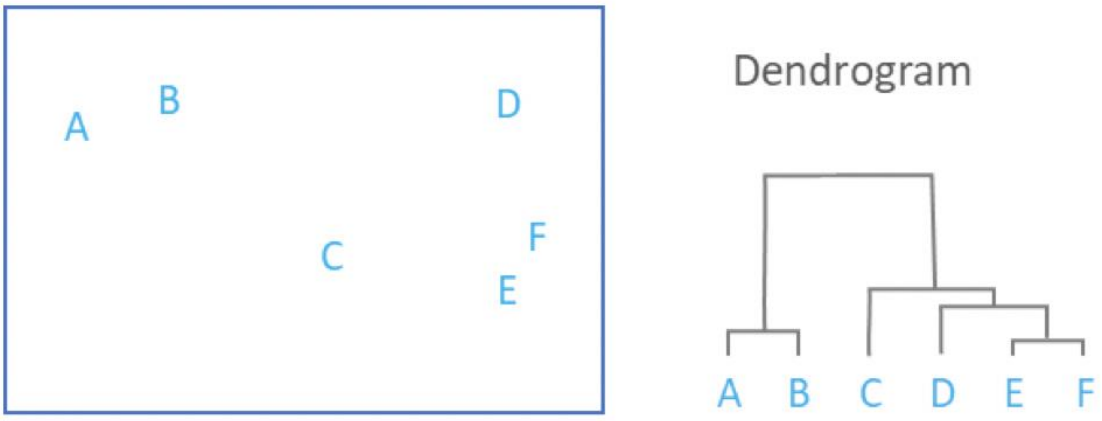

Figura 41: Ejemplo de clustering jerárquico

Por otra parte el clustering particional separa todos los datos en diferentes conjuntos disjuntos según la métrica de distancia y el algoritmo utilizado, como se puede observar en la Figura 42.

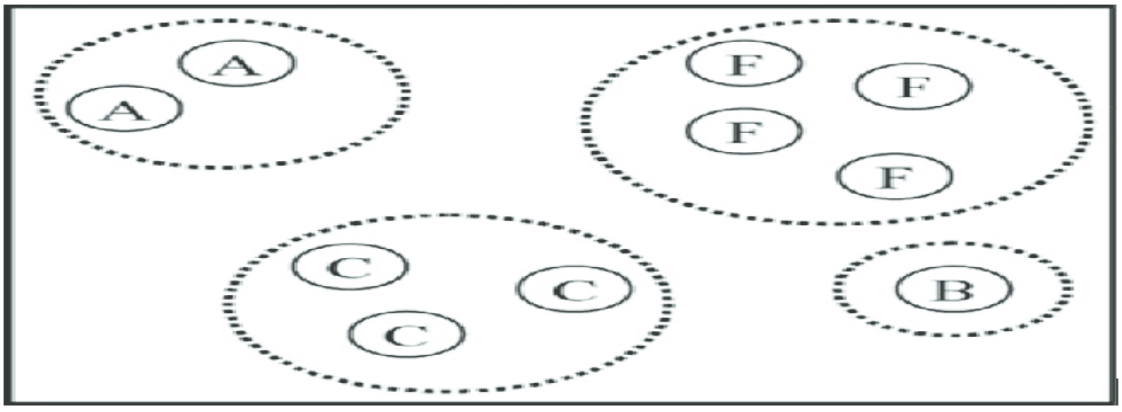

Figura 42: Ejemplo de clustering particional

Uno de los algoritmos más usados en la práctica al realizar clustering particional es el

algoritmo de K-Means, o agrupamiento de las K-medias. En este algoritmo se busca agrupar los datos según la distancia a un número K de centroides. Estos centroides son calculados y ajustados iterativamente hasta que no haya variación o se llegue al límite de iteraciones. En la Figura 43 se puede observar la ejecución del algoritmo sobre un conjunto de datos de ejemplo.

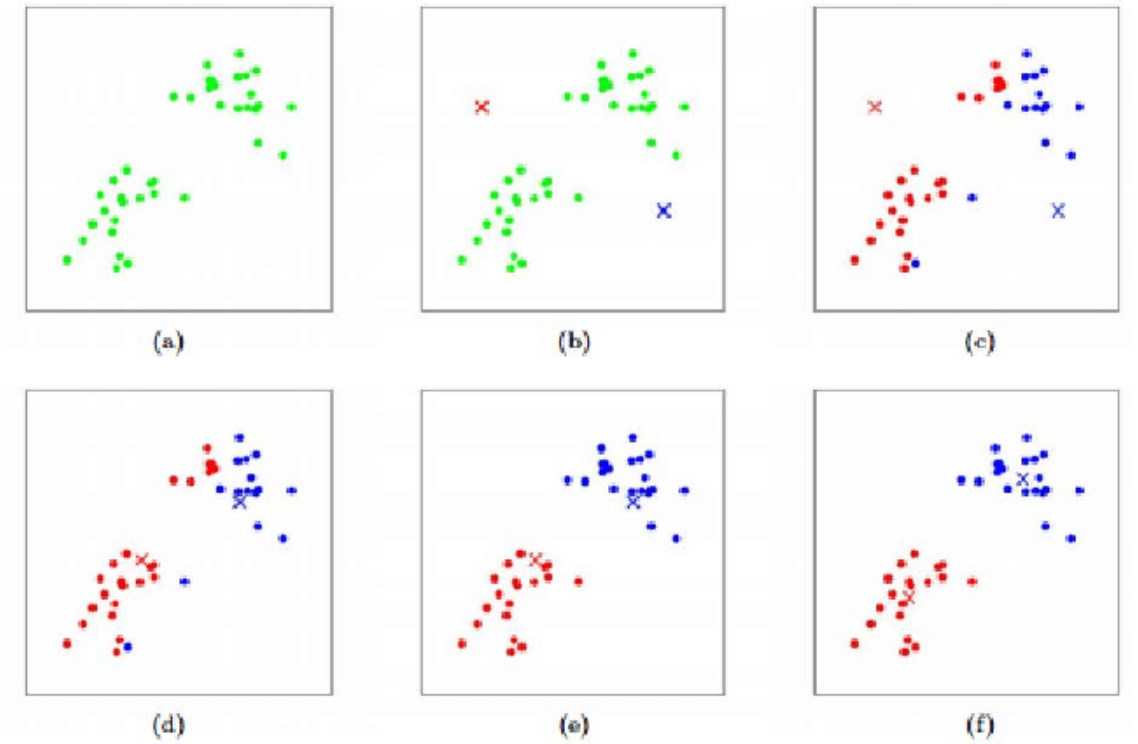

Figura 43: Ejemplo de K-Means con K=2

6.2 Clasificación de los segmentos usando K-Means

Como se mencionó al inicio del capítulo, para crear las series temporales simbólicas es necesario clasificar los segmentos.

Un agrupamiento particional es más práctico en este problema puesto que los segmentos son disjuntos y solo están conectados por el tiempo, está relación es implícita en los datos.

Como ya se mencionaba en la sección anterior, K-Means es un algoritmo muy usado en la práctica para realizar clustering. Funciona creando una matriz de distancia entre los datos y luego a partir de puntos iniciales aleatorios como centroides de los K clústeres se mide la distancia desde los datos hasta los K centroides y se asigna el grupo como el más cercano, tal y como se observa en la Figura 43.

El resultado del algoritmo está muy influenciado por el método usado para medir la distancia entre los datos, el número de clústeres K, y por los puntos iniciales. El problema de los puntos iniciales puede solucionarse realizando varias iteraciones y tomando el mejor resultado.

Con respecto al número de clústeres, depende de los datos, a veces es intuitivo y factible establecer un K a priori, como en el ejemplo de la Figura 43, pero en otras ocasiones realmente no es factible establecer un K a simple vista. En estos casos es habitual utilizar métricas de calidad del clúster, como el coeficiente de silueta o la distancia entre los elementos internos de los clústeres.

Los datos que se usarán para determinar los clústeres creados es la media de la varianza, la curtosis, la asimetría, la desviación y la pendiente para cada segmento obtenido.

La varianza es una medida de dispersión definida como la esperanza del cuadrado de la desviación de dicha variable respecto a su media [26]. Una varianza elevada indica segmentos con puntos dispersos, y está muy relacionada con la desviación.

La curtosis de una variable es una característica de la forma de su distribución de frecuencias [27]. Una curtosis elevada implica una mayor concentración de valores de la variable tanto muy cerca de la media de la distribución como muy lejos de ella, al tiempo que existe una relativamente menor frecuencia de valores intermedios. Segmentos con alta curtosis contienen muchos valores atípicos y extraordinarios.

La asimetría permite establecer el grado de simetría (o asimetría) que presenta una distribución de probabilidad de una variable aleatoria. Existen varias fórmulas para su cálculo, pero en esta investigación se usa

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Que es el coeficiente de momento estandarizado de Fisher-Pearson ajustado y es el más usado en software estadístico y aplicaciones informáticas.

n es el número de elementos en el conjunto.

$\bar{x}$ es la media del conjunto.

s es la desviación típica del conjunto.

La asimetría puede ser positiva o negativa, segmentos con asimetría positiva tienden a tener el mayor movimiento en el inicio del segmento y bajo movimiento en el resto,

mientras que segmentos con asimetría negativa tienen el mayor movimiento cerca del final del segmento.

La desviación estándar de una variable es la raíz cuadrada de su varianza. De forma similar a la varianza, sirve para determinar la dispersión de los datos.

La pendiente de una recta, suele estar representada por la letra *m*, y está definida como la diferencia en el eje Y dividido por la diferencia en el eje X para dos puntos distintos en una recta [28].

Para los segmentos, la pendiente es muy importante, puesto que indica la dirección del precio, una pendiente positiva indica crecimiento en el precio en el tiempo, mientras que una pendiente negativa indica descenso en el precio en el tiempo.

Con estos datos se pasaría de una serie temporal de precios dividida en segmentos a una tabla donde cada fila es un segmento y tienen tantas columnas como datos estadísticos son extraídos de cada segmento. Las filas de esta tabla son los elementos que se van a agrupar con el clustering.

6.2.1 Coeficiente de silueta

El coeficiente de silueta es un coeficiente calculado usando la distancia desde los puntos de un clúster a los otros clústeres y la distancia entre los puntos que componen el clúster en cuestión. El rango del coeficiente es entre -1 y 1, siendo 1 el máximo posible, donde los clústeres están muy separados entre ellos, y los elementos de los mismos están muy unidos.

La media del coeficiente de silueta de los diferentes clústeres obtenidos es una excelente métrica de calidad para determinar que tan bien agrupados están los datos.

6.2.2 Cálculo de distancias

El K-Means depende de la fórmula usada para determinar la distancia y, por ende, la similitud entre los puntos. En nuestro caso los puntos son las filas de la tabla con la información estadística de los segmentos, donde cada fila se corresponde a un segmento y cada columna a sus características.

Dependiendo de la fórmula usada para determinar dicha distancia, puede ser importante escalar los datos y centrarlos. Algunas de las formulas tradicionales para medir distancia son la distancia Euclídea, distancia de Minkowski y la distancia de Mahalanobis. [29]

Además se pueden usar otros tipos de métricas para medir la distancia, como puede ser la distancia de correlación entre segmentos.

Cada una de estas formas obtiene resultados diferentes en la creación de los clúster. En esta investigación se va a usar la distancia euclídea y la distancia por correlación de Pearson: $d_{cor}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$

Donde *x, y* son los dos vectores para los que se quiere calcular la distancia.

n es la longitud de los vectores *x*, *y*.

$\bar{x}, \bar{y}$ Se refiere a la media de los vectores *x*, *y*.

Está distancia correlativa mide el grado de relación lineal entre dos segmentos. Más información sobre el proceso de conversión para obtener distancias métricas a partir de de la correlación de Pearson y Spearman se puede encontrar en [30].

6.2.3 Clúster usando precio y distancia euclídea.

En la Figura 44 se observa el coeficiente de silueta según el número de clústeres para K-Means usando la distancia euclídea.

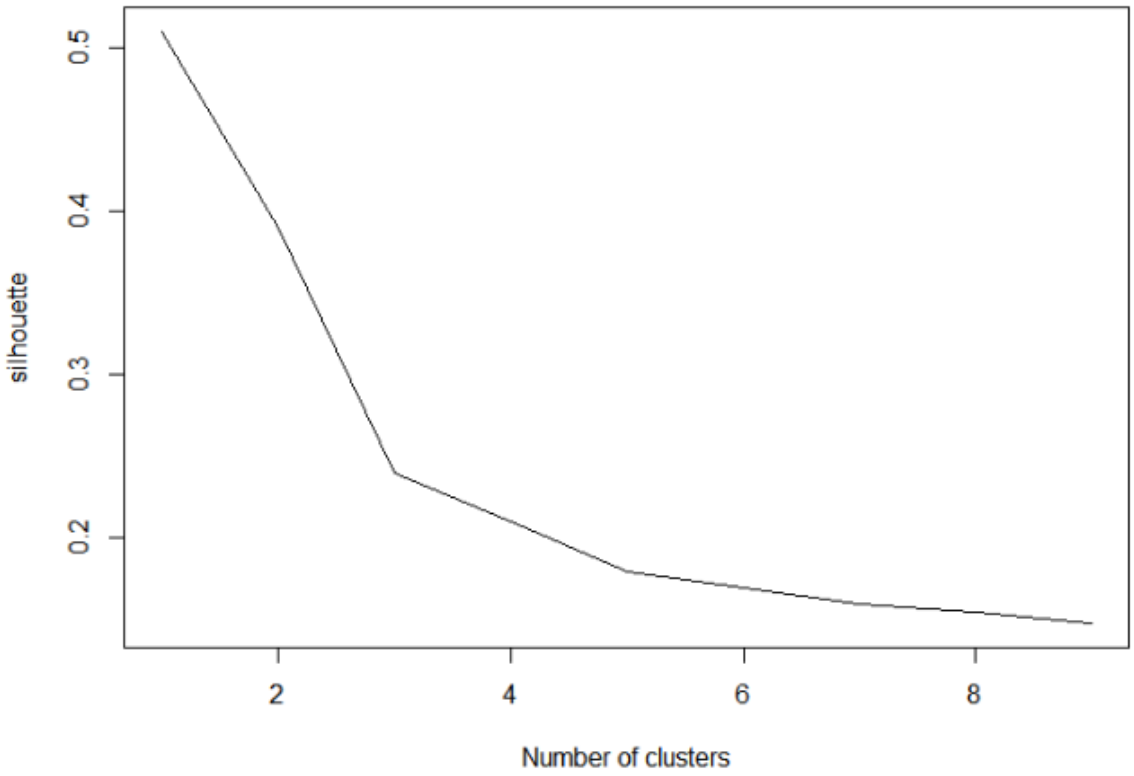

Figura 44: Coeficiente de silueta según número de clústeres para el precio con distancia euclídea

Se observa un coeficiente bastante decente de 0.51 para solo 2 clústeres, disminuyendo rápidamente al aumentar el número de clústeres. Las características de estos clústeres resultantes para el par AUDCAD se puede observar en la Figura 45

Índice de clúster	Variación	Asimetría	Kurtosis	Desviación	Pendiente	Coef. Silueta
clúster 1	0.00001222	-0.07005759	-1.331216	0.00337227	-0.00006353	0.58
clúster 2	0.00000181	-0.05339305	-0.953310	0.00123868	-0.00001422	-0.09

Figura 45: Características de los clústeres creados por K-Means con k=2, distancia euclídea y aplicado sobre el precio

Como se observa en la tabla, solo el coeficiente de silueta de un clúster es elevado, la pendiente es negativa en ambos casos, y no hay demasiada diferencia entre las características, indicando malos resultados en la agrupación. En la Figura 46 se observa la agrupación resultante.

Figura 46: Segmentos agrupados por clúster para K-Means con k=2, distancia euclídea y aplicado sobre el precio

Se observa que la agrupación no es satisfactoria y se descarta el trabajo con distancia euclídea y K=2 enfocado sobre el precio.

A pesar de tener menor coeficiente de silueta, es una posibilidad que aumentar el número de clústeres permita un mejor modelado sobre la serie. Un K-Means con K=5 da resultados igualmente decepcionantes como se puede observar en la Figura 47

Índice de clúster	Variación	Asimetría	Kurtosis	Desviación	Pendiente	Coef. Silueta
clúster 1	0.00001034	0.02559426	-1.45583500	0.00306326	-0.00071444	-0.05
clúster 2	0.00000162	0.67202530	-0.51800140	0.00117274	-0.00002769	0.19
clúster 3	0.00001006	-0.09295046	-1.38329300	0.00302953	0.00051039	-0.1
clúster 4	0.00000153	-0.14737060	-1.31396700	0.00114803	-0.00003229	0.3
clúster 5	0.00000234	-1.04017000	0.52675460	0.00139684	0.00005345	-0.15

Figura 47: Características de los clústeres creados por K-Means con k=5, distancia euclídea y aplicado sobre el precio

La agrupación en un período de tiempo sobre el mismo par de divisas AUDCAD se puede observar en la Figura 48.

Figura 48: Segmentos agrupados por clúster para K-Means con k=5, distancia euclídea y aplicado sobre el precio

Se puede observar que los segmentos azules son descensos, bastante bruscos, pero no se observa mucho más de los otros segmentos, no parecen tener mucho en común entre los que tienen el mismo color.

Se concluye que trabajar directamente con el precio para la distancia Euclídea con K-Means no aporta buenos resultados.

6.2.4 Clúster usando fuerza y distancia euclídea.

Los segmentos fueron creados usando la fórmula de fuerza vista en la sección 5.2, así que es posible que agrupar los segmentos según características relativas a la misma proporcione resultados favorables.

En la Figura 49 se observa el coeficiente de silueta según el número de clústeres creados.

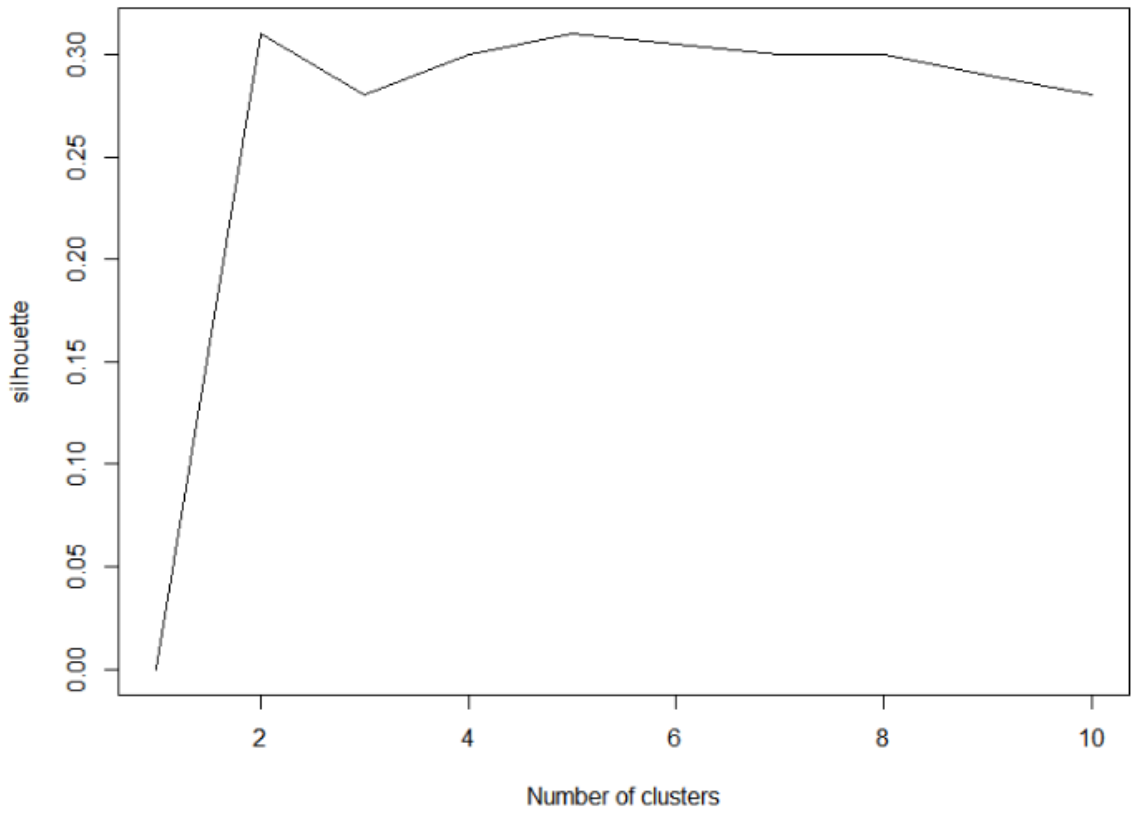

Figura 49: Coeficiente de silueta según número de clústeres para la fuerza con distancia euclídea

El número óptimo de clústeres es 5, y en la Figura 50 se observan las características de los segmentos pertenecientes a los diferentes clústeres. Los valores observados son los relacionados a la fuerza, no al precio.

Índice de clúster	Variación	Asimetría	Kurtosis	Desviación	Pendiente	Coef. Silueta
clúster 1	0.3284356	-0.7642377	-0.3420282	0.5419774	-0.04096702	0.31
clúster 2	0.2745914	0.6559641	-0.6224295	0.4984225	0.02273631	0.26
clúster 3	1.704234	-0.09566083	-1.5103	1.289805	0.108906	0.23
clúster 4	0.5174321	-0.0210592	-1.59312	0.6879398	-0.2454479	0.33
clúster 5	0.4832848	-0.0181517	-1.590407	0.6684387	0.1672737	0.36

Figura 50: Características de los clústeres creados por K-Means con k=5, distancia euclídea y aplicado sobre la fuerza

Los 5 clústeres obtenidos tienen un coeficiente de silueta similar, se observa el clúster 3 como un clúster con alta variación y desviación, mientras que el 4 tiene pendiente negativa muy brusca. En general se observa diferenciación entre los mismos, y en la Figura 51 se observa como la diferenciación estadística entre la fuerza no se refleja en los segmentos usando la distancia euclídea.

Figura 51: Segmentos agrupados por clúster para K-Means con k=5, distancia euclídea y aplicado sobre la fuerza

6.2.5 Clúster usando precio y distancia por correlación de Pearson.

Los resultados finales obtenidos en la sección 6.2.4 descartan la posibilidad de usar la distancia euclídea para agrupar los segmentos, tanto por precio como por fuerza. En esta sección se analiza la posibilidad de agrupar los segmentos según la correlación entre los mismos usando la distancia por correlación de Pearson.

La distancia por correlación de Pearson presenta una mejora significativa en la calidad del coeficiente de silueta con respecto a los obtenidos en las secciones anteriores, como se puede observar en la Figura 52.

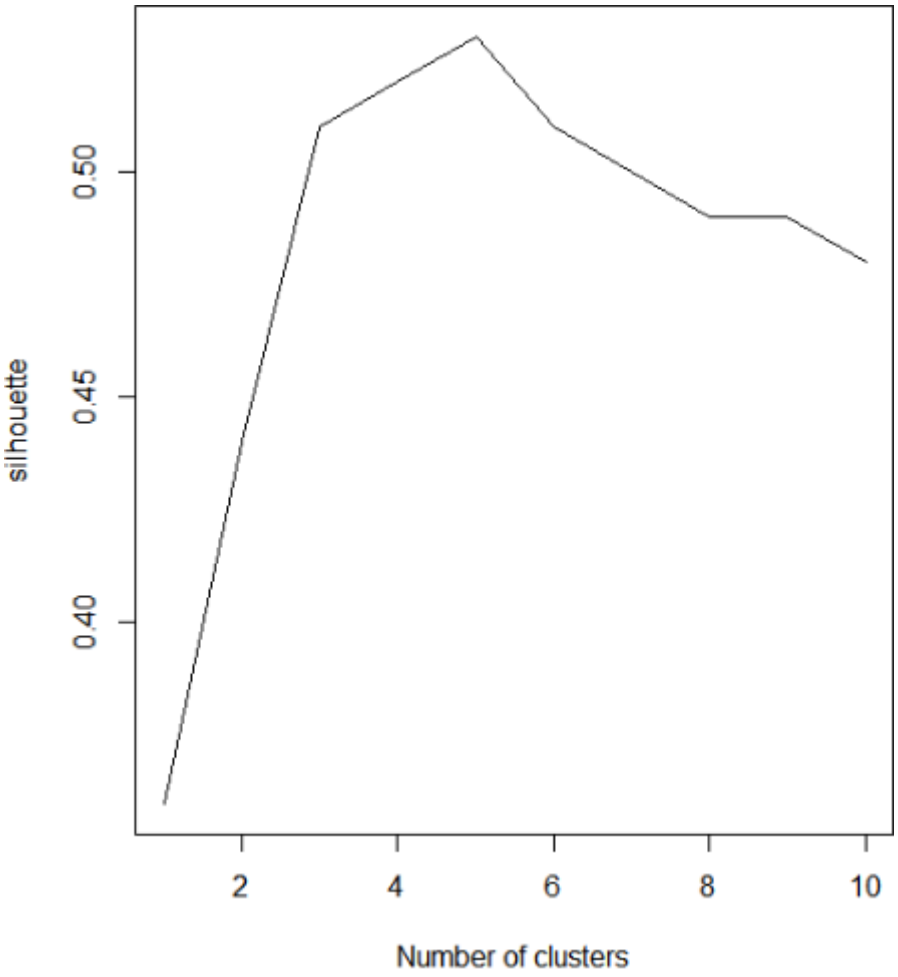

Figura 52: Coeficiente de silueta según número de clústeres para el precio con distancia por correlación de Pearson

Similar a agrupaciones anteriores, el número óptimo de clústeres es de 5. En la Figura 53 se observan las características de los 5 clústeres creados.

Índice de clúster	Variación	Asimetría	Kurtosis	Desviación	Pendiente	Coef. Silueta
clúster 1	0.000006524	-0.025108850	-1.403199000	0.002341894	-0.000483682	0.60
clúster 2	0.000001179	0.090469970	-1.396744000	0.000990856	0.000127725	0.42
clúster 3	0.000001692	-0.708553600	-0.189982600	0.001191222	0.000018752	0.60
clúster 4	0.000001235	0.523362600	-0.790591400	0.001038538	-0.000107948	0.47
clúster 5	0.000007165	-0.242337500	-1.405141000	0.002480269	0.000398103	0.56

Figura 53: Características de los clústeres creados por K-Means con k=5, distancia por correlación de Pearson y aplicado sobre el precio

Los 5 tienen un coeficiente de silueta bastante elevado, y sus características son bastante diferentes entre ellos, la agrupación resultante se puede observar en la Figura 54.

Figura 54: Segmentos agrupados por clúster para K-Means con k=5, distancia por correlación de Pearson y aplicado sobre el precio

Se puede apreciar una mejora notable con respecto a la distancia euclídea, con los segmentos azules marcando un fuerte descenso, con solo una excepción. Los pocos verdes indicando un ascenso y los 3 restantes indicando diferentes estructuras de inestabilidad e inconsistencia con poca variación.

Es una agrupación bastante satisfactoria. No solo las características de los segmentos en los clústeres son diversas, sino que su representación en la serie temporal es representativa de dichos clústeres.

El clúster 1 contiene una pendiente negativa marcada, con alta desviación y una asimetría cercana a 0, indicando que los segmentos pertenecientes a este clúster tienen un descenso estable y prolongado durante todo el segmento.

El clúster 2 se caracteriza por su poca variación con respecto a la pendiente, indicando la mayoría de las veces pequeños aumentos inestables o ruido.

El clúster 3 se caracteriza por su fuerte asimetría negativa y la escasez de pendiente, a pesar de una variación relativamente alta. Es decir, son segmentos con mucha inestabilidad, y en muchas ocasiones, cerca del fin del segmento.

El clúster 4 actúa como una mezcla del 3 y 2, pero en pendiente negativa. Se puede observar como segmentos inestables con ligera tendencia de descenso.

Finalmente el 5to clúster no se observa mucho en la sección mostrada puesto que representa los segmentos con aumento de precio marcado. Es equivalente al primer clúster, pero con pendiente positiva.

6.2.6 Clúster usando fuerza y correlación de Pearson.

Para finalizar la sección de pruebas sobre clústeres con K-Means se realizó la agrupación usando la distancia por correlación de Pearson sobre la fuerza, similar a como se realizó en la sección 6.2.4.

En la Figura 55 se observa el coeficiente de silueta según el número de clústeres, similar al de la sección anterior, pero con un valor máximo de 0.58 con 5 clústeres, ligeramente mejor.

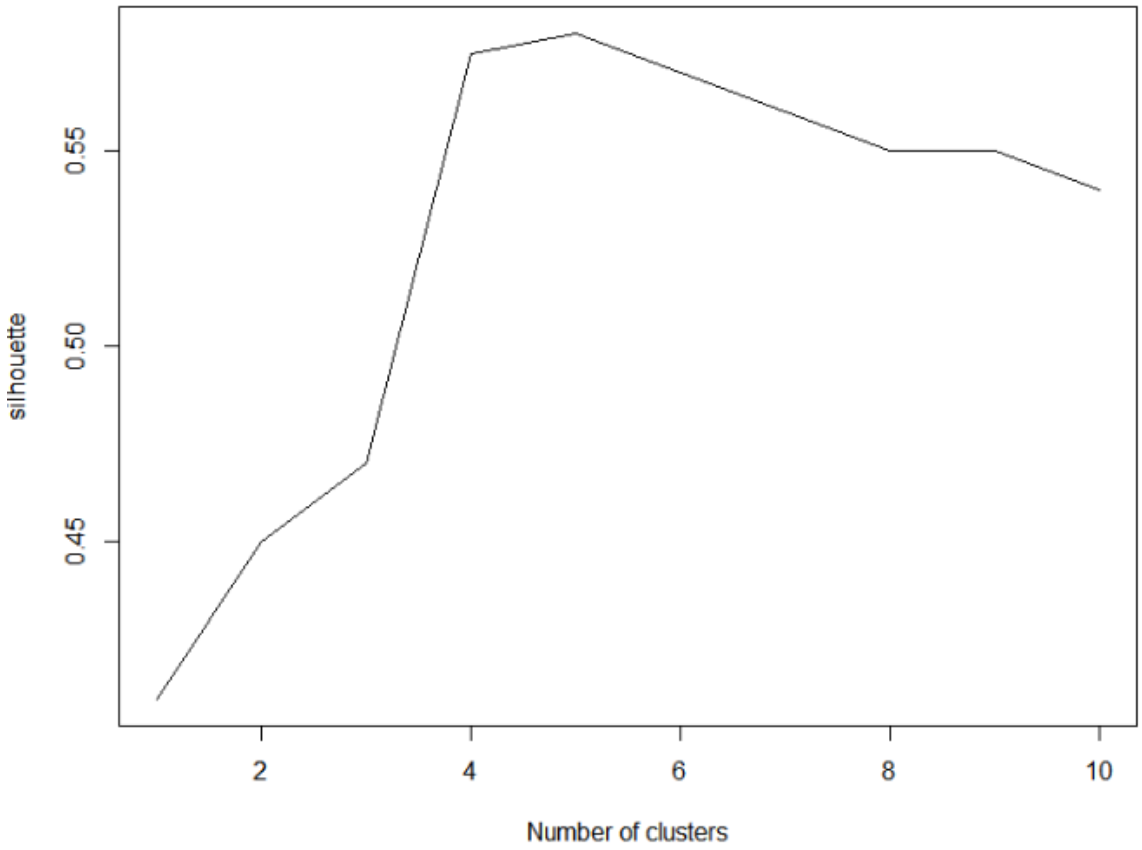

Figura 55: Coeficiente de silueta según número de clústeres para la fuerza con distancia según distancia por correlación de Pearson

Las características de estas agrupaciones se pueden observar en la Figura 56.

Índice de clúster	Variación	Asimetría	Kurtosis	Desviación	Pendiente	Coef. Silueta
clúster 1	0.9608915	-0.05257536	-1.592985	0.965675	-0.3121533	0.61
clúster 2	1.506986	-0.147038	-1.499512	1.209382	0.2126565	0.69
clúster 3	0.2485515	0.677228	-0.6410558	0.4716048	0.0419441	0.53
clúster 4	0.4173792	-0.04193146	-1.617873	0.6216409	0.1779112	0.53
clúster 5	0.2934997	-0.82129	-0.3313431	0.5146446	-0.02706661	0.54

Figura 56: Características de los clústeres creados por K-Means con k=5, distancia según distancia por correlación de Pearson y aplicado sobre la fuerza

Los clústeres tienen mejor coeficiente de silueta en general, con el número 1 y el 2 destacando sobre el resto como los segmentos con más variación, desviación y pendiente, a la vez que contienen pendientes opuestas.

La agrupación resultante se muestra en la Figura 57.

Figura 57: Segmentos agrupados por clúster para K-Means con k=5, distancia según distancia por correlación de Pearson y aplicado sobre la fuerza

Similar a la agrupación según el precio, los resultados usando la distancia por correlación de Pearson sobre la fuerza son bastante satisfactorios aunque los clústeres ahora tienen diferentes códigos. En particular se puede observar que:

El clúster 1 contiene una pendiente negativa marcada, con alta desviación y una asimetría cercana a 0, indicando que los segmentos pertenecientes a este clúster tienen un descenso estable y prolongado durante todo el segmento.

El clúster 2 es el opuesto al primero, posee una fuerte pendiente positiva, con una fuerte desviación y poca asimetría.

El clúster 3 se caracteriza por alta asimetría, y muy baja pendiente, con un nivel medio

de desviación. Se puede observar que son segmentos con mucho ruido y que confunden, normalmente comenzando un movimiento en una dirección, pero terminando en la otra, o muy cerca de donde comenzó.

El clúster 4 actúa muy similar al clúster 1. La principal diferencia es en la varianza, indicando que son segmentos donde la fuerza sube lentamente y con poca variación.

Finalmente el 5to clúster es otra variante de ruido, pero enfocado a una asimetría negativa, no se puede inferir mucho de su estructura.

6.3 Conclusiones sobre el clustering

La utilización de distancia euclídea sobre los segmentos da resultados muy poco prácticos, mientras que la utilización de distancia según la distancia por correlación de Pearson da resultados muy favorables.

La agrupación por las características del precio o por las características de la fuerza presenta resultados igualmente buenos, siendo los clústeres generados usando las características de la fuerza ligeramente mejor en sus características.

Debido a que la fuerza está escalada y centrada sobre 0 para todos los pares de divisas, es posible usar las agrupaciones para el posterior aprendizaje del modelo, y extender el número de datos disponibles.

Capítulo 7

Minería de reglas secuenciales recurrentes.

La minería de reglas secuenciales es una técnica de Machine Learning para obtener reglas que ocurren comúnmente entre secuencias, mientras que la minería de reglas recurrentes es una técnica que busca la obtención de reglas que ocurren recurrentemente en el tiempo. [31]

Podemos definir formalmente una regla como:

{Antecedente} -> {Consecuente}

Donde antecedente es un conjunto de elementos vinculados de alguna manera entre sí, y consecuente es el conjunto obtenido posterior a dicho antecedente. Un ejemplo clásico es el de la compra de supermercado:

{Manzana, Pan} -> {Leche}

Esta regla indica que cuando hay una compra de manzanas y pan, el cliente compra también leche.

Las reglas secuenciales son aquellas que ocurren en múltiples secuencias, por ejemplo:

Secuencia 1: {Pan}, {Leche}, {Manzana}, {Cerveza, Pañales}, {Perfume}

Secuencia 2: {Manzana, Pan}, {Leche}, {Perfume}

Secuencia 3: {Manzana}, {Leche}, {Pan}, {Pañales, Perfume}

Secuencia 4: {Manzana, Pan, Leche, Cerveza}, {Pizza}

Secuencia 5: {Pizza, Pan}, {Manzana, Leche}, {Cerveza}

Aquí se observan 5 secuencias de compras de cliente en un supermercado. Se puede observar que entre los diferentes clientes hay una tendencia a comprar leche después de comprar pan, a veces en la misma compra, a veces en la siguiente, aunque en la secuencia 3 se compra la leche primero que el pan.

Existen diferentes formas de trabajar con reglas secuenciales, según el problema a resolver, dos alternativas muy interesantes se pueden ver en [32] y [33].

Por otra parte, las reglas recurrentes son aquellas que ocurren múltiples veces en una secuencia temporal, donde los elementos del antecedente implican el consecuente en un margen de tiempo determinado [34], por ejemplo:

{Sueño} -> {Dormir}

Siempre que tenemos sueño, vamos a dormir, esto es intuitivo, pero el valor de la regla depende del intervalo, porque quizás antes de llegar a dormir primero hicimos otras cosas:

{Sueño}, {Café}, {Trabajo}, {Cena}, {Dientes} -> {Dormir}

Entonces, se podría decir que sueño implica el dormir, o que el café implica el dormir, o que la combinación de todos ellos en esa secuencia exacta implica el dormir. La forma de trabajar con estas reglas y del intervalo de análisis depende del problema. Algunas alternativas para trabajar con este tipo de reglas con intervalos variables se pueden observar en [35] y [34].

7.1 Minería de reglas secuenciales recurrentes en el mercado de divisas

En el mercado de divisas, para la predicción de tiempo, solo nos interesa el próximo intervalo de precio, una regla que diga que el precio va a bajar “eventualmente” no es de uso alguno. Por ende, en nuestra secuencia cada elemento solo va a contener un segmento, y el consecuente solo va a ser el siguiente segmento, por lo que en la secuencia:
{Aumento de precio}, {Descenso de precio}, {Ruido}, {Aumento de precio}, {Descenso de precio}, {Ruido}

Se podrían inferir que:

{Aumento de precio} -> {Descenso de precio},

{Descenso de precio} -> {Ruido},

{Ruido} -> {Aumento de precio}

De forma similar, se podría decir que:

{Aumento de precio}, {Descenso de precio} -> {Ruido},

{Descenso de precio}, {Ruido} -> {Aumento de precio},

{Ruido}, {Aumento de precio} -> {Descenso de precio}

Es decir, tenemos reglas de un solo elemento en el antecedente, o de dos elementos en el antecedente, y solo un consecuente. Dichos elementos en el antecedente están

conectados directamente en intervalos de tiempo equidistantes.

Estas reglas se pueden generalizar para un intervalo fijo de elementos en el consecuente, pero es importante tener en cuenta que mayor longitud de secuencias en el antecedente ocurre con menos frecuencia de forma recurrente y que, además, crecen exponencialmente con el número de combinaciones de reglas posibles.

Con los clústeres creados anteriormente tendríamos segmentos separados en 5 grupos, y con una longitud de secuencia máxima de hasta 6 elementos en el antecedente tendríamos 5^7 posibles combinaciones de reglas.

7.2 Calidad de las reglas obtenidas

Una vez analizada la secuencia, podemos observar el número de veces que ocurre el antecedente, y en cuántos casos ocurre el consecuente. Si tenemos la regla: {Ascenso de precio}, {Ruido} -> {Descenso de Precio}

Y esta ocurre 3 veces, y además tenemos la regla:

{Ascenso de precio}, {Ruido} -> {Ascenso de Precio}

Que ocurre 2 veces, entonces podemos decir que la primera regla ocurre 3 de cada 5 veces, ya que el antecedente aparece 5 veces, pero el consecuente solo aparece en 3 ocasiones después de dicho antecedente. En este caso se dice que la regla tiene una **confianza** de 0.6, o el equivalente a 60%, mientras que la otra regla tendría solo 0.4 o 40%. Por otra parte, el **soporte** de la regla es equivalente al número de veces que aparece el antecedente, en este caso sería de 5.

Lo ideal es buscar reglas con alta confianza, pero que a su vez tengan un mínimo de soporte, ya que es posible tener reglas con 100% de confianza, pero que solo haya ocurrido una vez en toda la secuencia, esto es bastante común en reglas con antecedentes muy extensos.

7.3 Proceso de obtención de reglas

El proceso de obtención de reglas recurrentes es muy sencillo, a partir de una secuencia, dígase {A},{A},{B},{B},{A},{A},{B},{A}

Se hace un ciclo iterativo a partir del segundo elemento, el elemento actual es el consecuente de la regla, mientras que los elementos anteriores serían los antecedentes, según la longitud máxima que se busca. Si tuviéramos una longitud máxima de 2 en esta secuencia, el proceso sería:

{A},{A},{B},{B},{A},{A},{B},{A}

Regla 1: {A} -> {A}. Ocurre una vez.

{A},{A},{B},{B},{A},{A},{B},{A}

Regla 2: {A} -> {B}. Ocurre una vez.

Regla 3: {A}, {A} -> {B}. Ocurre una vez.

{A},{A},{B},{B},{A},{A},{B},{A}

Regla 4: {B} -> {B}. Ocurre una vez.

Regla 5: {A}, {B} -> {B}. Ocurre una vez.

{A},{A},{B},{B},{A},{A},{B},{A}

Regla 6: {B} -> {A}. Ocurre una vez.

Regla 7: {B}, {B} -> {A}. Ocurre una vez.

{A},{A},{B},{B},{A},{A},{B},{A}

Regla 1: {A} -> {A}. Ocurre dos veces.

Regla 8: {B}, {A} -> {A}. Ocurre una vez.

{A},{A},{B},{B},{A},{A},{B},{A}

Regla 2: {A} -> {B}. Ocurre dos veces.

Regla 3: {A}, {A} -> {B}. Ocurre dos veces.

{A},{A},{B},{B},{A},{A},{B},{A}

Regla 6: {A} -> {B}. Ocurre dos veces.

Regla 3: {A}, {B} -> {A}. Ocurre una vez.

Una vez finalizado el proceso se obtiene un conjunto de reglas y su número de ocurrencias, lo que nos permite calcular el **soporte** y la **confianza** de las mismas.

Debido a las características del mercado de divisas, se propone la posibilidad de extender la base de reglas usando como secuencias los 28 pares de divisas y crear una base de reglas únicas.

En este trabajo se generaron las bases de reglas para 9 grupos con diferentes condiciones. En cada grupo se generaron las reglas para el par de divisas de AUDCAD, EURNZD y un conjunto de reglas unificado. Dichos grupos son:

- Fuerza5 y Precio5. En estos grupos se encuentran las reglas correspondientes a los clústeres creados a partir de la fuerza y el precio vistos en el capítulo de

clustering.

- Pendiente3 y PendienteVar3. Estos son grupos creados usando la misma lógica del capítulo de clustering, pero separando en solo 3 clústeres, uno usando puramente la pendiente para separación y otro usando solo pendiente y varianza.
- Supervisado3_20, Supervisado3_15, Supervisado3_10, Supervisado3_05. Estos son grupos de clasificación supervisada simple. En estos grupos los segmentos están agrupados por pendiente positiva, negativa o ruido. El número indica la inclinación de la pendiente necesaria para no considerarse como ruido.
- SupervisadoBin. Este es un grupo donde los segmentos están calificados directamente como ascendentes o descendentes, sin ningún otro tipo de separación.

En estos grupos, los segmentos fueron asignados un grupo según las descripciones mencionadas anteriormente, y luego fueron reordenados según la pendiente, de forma que el grupo 1 tiene la máxima pendiente positiva, y el grupo 5 tiene la máxima pendiente negativa. Este proceso es necesario para poder realizar la agrupación común de reglas con diferentes pares de divisas, y luego para poder realizar una validación de la calidad de las reglas obtenidas.

7.4 Evaluación de las reglas obtenidas

El proceso de evaluación utilizado está dividido en varios pasos con el objetivo de realizar una evaluación diversa y poder llegar a conclusiones fehacientes.

Primero se separa los datos de los pares de divisa en un conjunto de entrenamiento y de test. Al ser una serie temporal es necesario conservar la relación temporal entre los mismos, por lo que si dividimos los datos en 80% para el entrenamiento y 20% para el test; nos quedaríamos con los datos hasta el 29 de septiembre del 2017 como entrenamiento y el resto de los datos desde esa fecha hasta el 24 de enero del 2020 como test.

Con ese conjunto de entrenamiento generamos las reglas para los 9 grupos mencionados anteriormente. Posteriormente usamos estas reglas para predecir los segmentos siguientes en el conjunto de test y verificar la calidad de las mismas. Para este proceso ordenamos las reglas según su confianza, usando como predicción el consecuente de la regla que mayor confianza tenga según el antecedente. Además, se realizan pruebas solo con reglas de una confianza superior al 60%, y pruebas donde además de una confianza superior al 60% exista al menos una cantidad de ocurrencias de la regla mayor a 10.

Este proceso de predicción se realiza sobre 3 pares de divisas para cada grupo con características muy diferentes, AUDCAD, EURNZD y USDJPY. Este proceso además de predecir el grupo del segmento del test y compararlo con el grupo real del test realiza una simulación de fondos operando sobre los precios al inicio y final de cada segmento. Es decir, si predecimos un segmento 1, que son los que tienen una pendiente positiva, tomamos una posición de compra, y vendemos al precio al final del segmento, de forma similar, al tener un segmento 5 de pendiente negativa iríamos corto y cubriríamos la posición al final del segmento. No se realizan operaciones si el segmento predicho es uno intermedio, que equivaldría a un segmento de ruido o inestabilidad.

En este proceso de compra y venta no se incluye costos de comisiones, slippage ni similares, no hay operaciones de control de posición como trailing stops y similares; es decir, es un proceso muy simple de prueba usando solo el precio inicial y final. En la práctica es posible mejorar los resultados substancialmente con operaciones de control de posición.

Estos procesos de compra y venta se realizan sobre 2 fondos monetarios con diferentes formas de operación, en uno invertimos todo el capital en cada operación, mientras en el otro se invierte una cantidad fija de 10.000 euros en cada operación.

Finalmente, estas operaciones se realizan sobre 3 conjuntos de límites diferentes:

- Sin límites de pérdida o ganancia
- Un límite del 0.4% de pérdida, que se corresponde a movimientos bastante bruscos en el intervalo de estudio en el mercado de divisas.
- Un límite del 0.4% de pérdida y de ganancia.

Además, para el caso de Precio5 y Fuerza5, se considera realizar operaciones tomando solo 3 como ruido, y tomando 2,3 y 4 como ruido.

En resumen se realizan las siguientes pruebas:

Para los 9 grupos y los 3 conjuntos de divisas AUDCAD, EURNZD, USDJPY, usando su propia base de reglas y la base de reglas conjuntas, usando todas las reglas, solo reglas con más de 60% de confianza y reglas con más de 60% de confianza y un total superior a 10; con fondos sin límites, con límite de pérdida y con límite de ganancia. Como excepción se realizan dos conjuntos de prueba adicionales para Precio5 y Fuerza5 tomando diferentes rangos de segmentos como ruido.

El total de pruebas realizada asciende a 1188.

7.5 Resultados notables de las pruebas de reglas

En principio los resultados son muy interesantes, aunque algunos llegan a ser bastante confusos o erráticos por la carencia de suficientes elementos de prueba. Las tablas con todos los resultados exactos se pueden encontrar en el Anexo 1.

En la Figura 58 se puede observar la precisión promedio de las predicciones de las reglas propias del par de divisas determinando el próximo segmento.

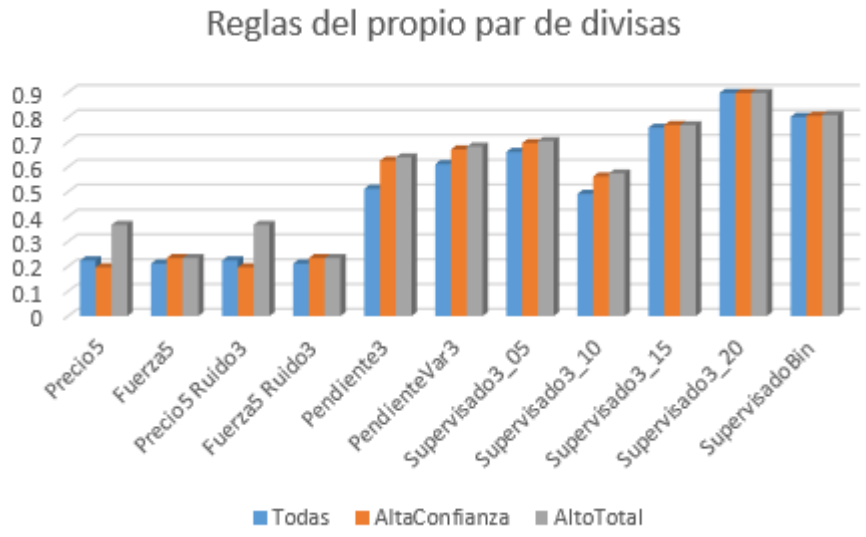

Figura 58: Precisión promedio de las reglas aprendidas del par de divisas aplicadas sobre el mismo par de divisas.

En la Figura 59 se puede observar la precisión promedio de las predicciones de las reglas conjuntas de todos los pares de divisas determinando el próximo segmento.

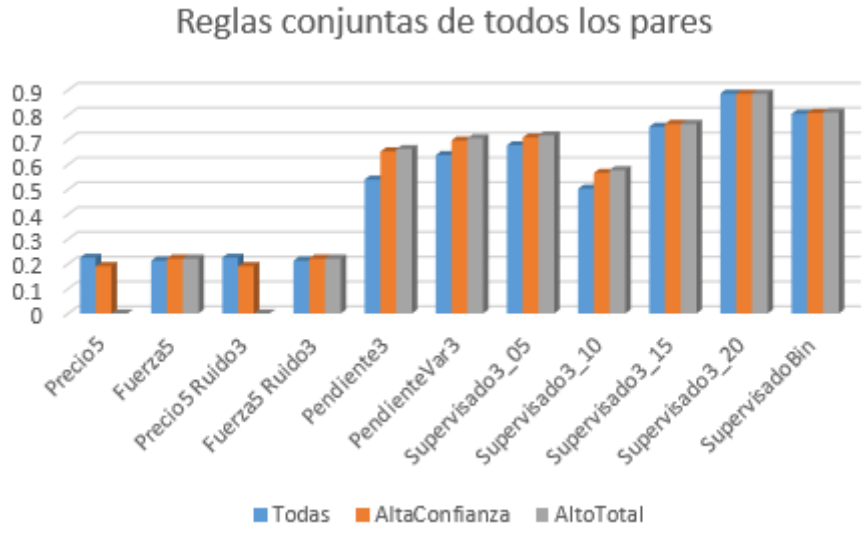

Figura 59: Precisión promedio de las reglas conjuntas aprendidas de todos los pares de

divisas aplicadas sobre los pares de divisas de prueba.

Finalmente en la Figura 60 se puede observar la diferencia entre las reglas conjuntas y las reglas obtenidas de un solo segmento.

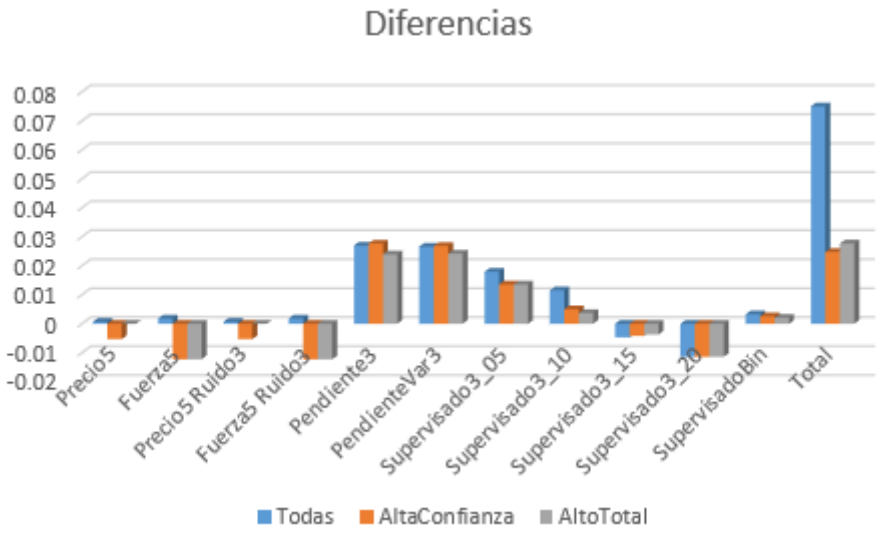

Figura 60: Diferencia de precisión entre las reglas conjuntas y las reglas propias aplicadas sobre los pares de divisas de prueba.

Independiente del resultado de las operaciones sobre las finanzas, se puede observar que, en general, usar una base de reglas conjuntas mejora las predicciones sobre los pares de divisas. Esto muestra un dato interesante: que los pares de divisas, a pesar de ser completamente diferentes funcionan bajo las mismas reglas, y esto a su vez nos permite extender nuestro conocimiento de las reglas analizando todos los pares de divisas, aumentando considerablemente el conjunto de datos disponibles para entrenamiento.

En general aumentar la confianza mínima de las reglas y el total de ocurrencias aumenta la precisión, al coste de realizar menos predicciones.

Los resultados de los fondos son bastante más extensos, en la Figura 61 se observan los fondos finales tras realizar múltiples operaciones de compra y venta según las predicciones de segmentos a partir de un capital inicial de 10.000 euros. Las primera y segunda columnas representa el valor final de los fondos al operar sin límites de ganancia o pérdida con todo el capital o con una cantidad fija de 10.000 euros por operación. Luego en la tercera y cuarta columnas se puede observar los resultados tras limitar las pérdidas en segmentos incorrectamente clasificados. Finalmente en la quinta y sexta columna se observa lo que sucede al limitar ambas ganancia y pérdida en la misma cantidad.

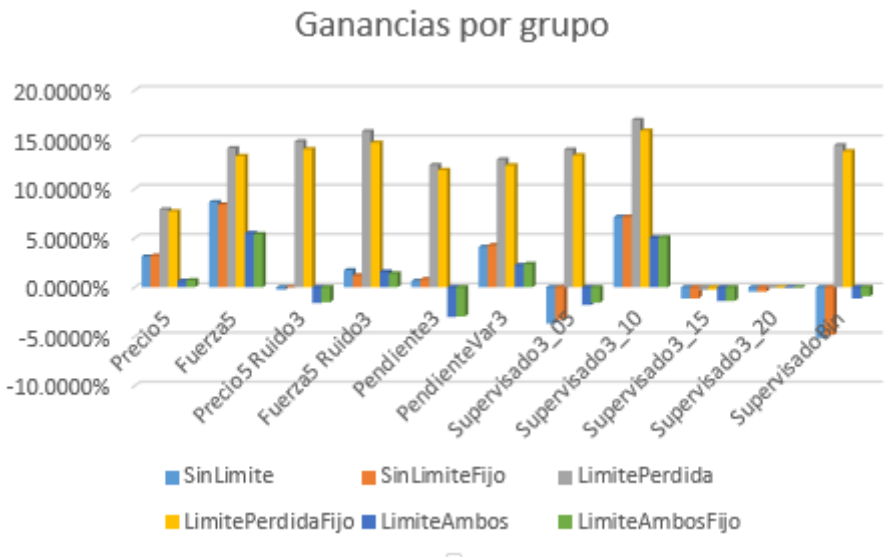

Figura 61: Ganancia promedio de fondos finales aplicando todas las reglas aprendidas de un solo par de divisas

La primera observación aparente es que limitar las pérdidas aumenta significativamente las ganancias finales. Los supervisados con pendiente muy pequeña, o muy elevada tienen pérdidas, igualmente que el binario. La clasificación con el supervisado equilibrado da buenos resultados, igual que la separación con pendiente y Varianza, y ambos grupos de 5 clústeres. Esto es interesante puesto que estos grupos tienen una precisión no muy elevada en general, pero los segmentos que predicen se corresponden correctamente con el precio subyacente en la mayoría de los casos.

En la Figura 62 se observan los resultados al aplicar solo reglas de alta confianza.

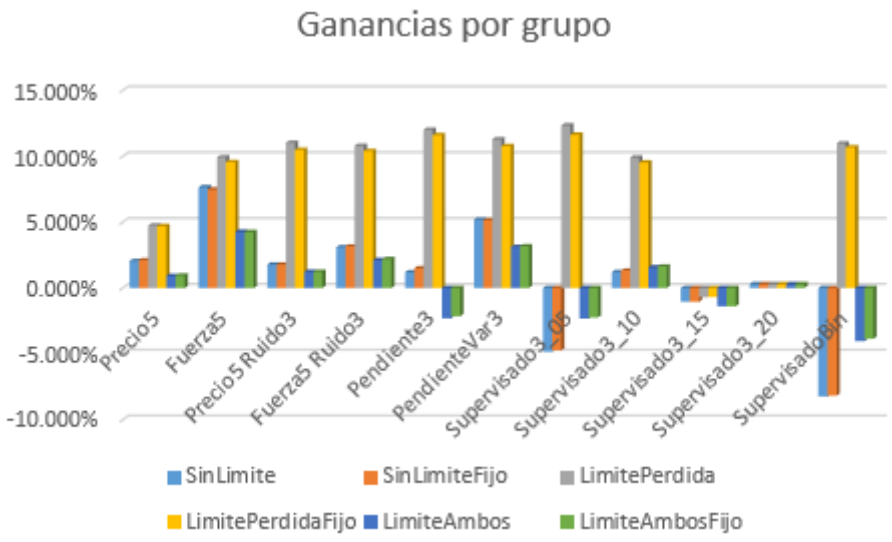

Figura 62: Ganancia promedio de fondos finales aplicando solo reglas de alta confianza aprendidas de un solo par de divisas.

No hay mucha diferencia con respecto a aplicar todas las reglas, pero al realizarse menos operaciones en general los números finales son un poco más pequeños. En esta sección se marcan como claros vencedores la Fuerza5 y PendienteVar3.

En la Figura 63 se muestran los resultados de aplicar solo reglas de alta confianza y con alto total de apariencia.

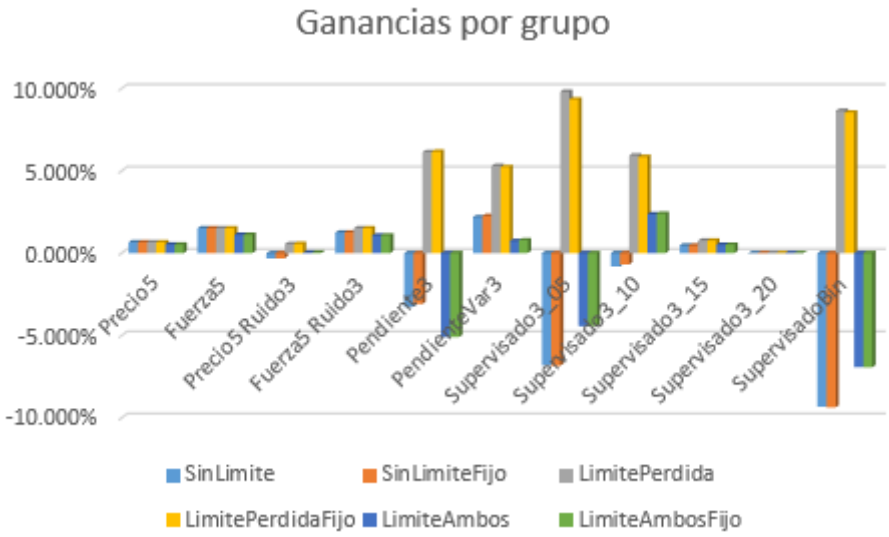

Figura 63: Ganancia promedio de fondos finales aplicando solo reglas de alta confianza y con alto total de apariencia aprendidas de un solo par de divisas.

Similar a los resultados anteriores, Fuerza5 y PendienteVar3 son los mejores en general, sin incluir límite de pérdida, claro. El caso del grupo de supervisado binario es interesante puesto que las pérdidas son bastante significativas, y es una supervisión binaria, por lo que simplemente se podrían invertir las predicciones para lograr grandes ganancias.

En general estas reglas tienen muy pocas predicciones, por lo que el test puede carecer de suficientes ejemplos para ser representativo.

Por otra parte, en la Figura 64 se muestran los resultados usando la base de reglas conjunta.

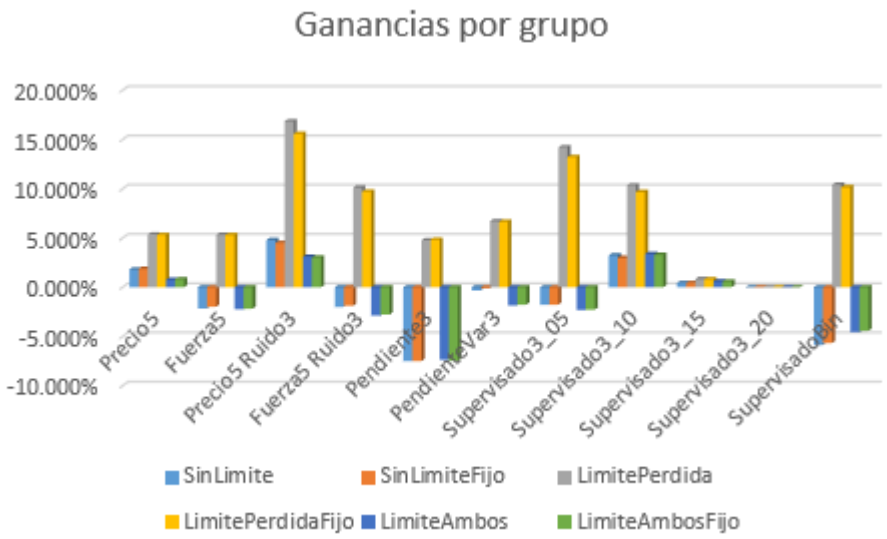

Figura 64: Ganancia promedio de fondos finales aplicando todas las reglas aprendidas de todos los pares de divisas.

En general la base de reglas conjunta tiene mejor precisión prediciendo los segmentos como se observó anteriormente, pero esta calidad de predicciones en realidad es para segmentos poco representativos en la mayoría de los casos.

En contraste con los resultados obtenidos usando reglas propias, Precio5 Ruido3, es decir, el grupo conformado por clústeres separados según las características de precio y usando solo el valor 3 como ruido, realizando compras con 1 y 2, y ventas con 4 y 5; es el que mayor ganancia tiene.

La Figura 65 muestra los resultados de aplicar solo reglas de alta confianza con todo el conjunto de reglas extendido.

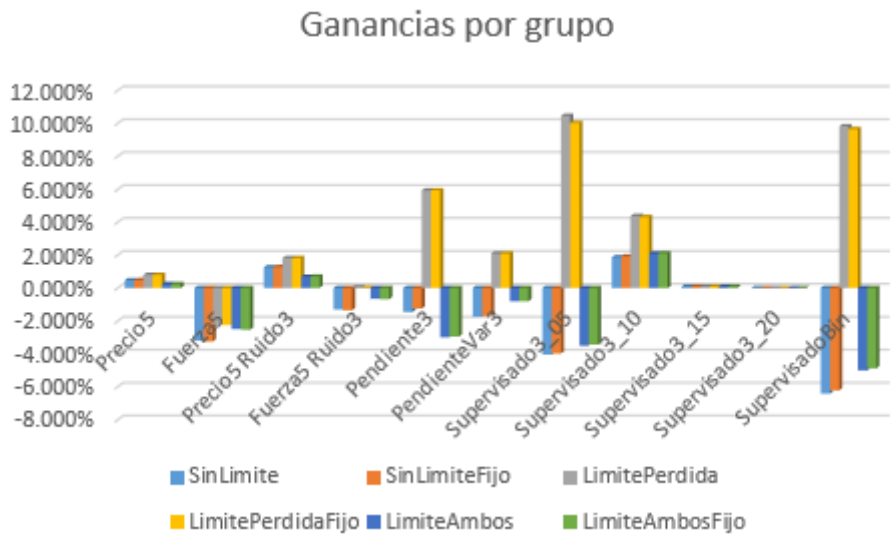

Figura 65: Ganancia promedio de fondos finales aplicando solo reglas de alta confianza aprendidas de todos los pares de divisas.

Los resultados son bastante malos en general, y además se realizan pocas operaciones en general, por lo que en general están acotados y podrían ser mucho peores con más datos de prueba.

Esto se observa de forma incluso más drástica en la Figura 66, donde solo se usan las reglas de alta confianza y alto total de la base de reglas conjunta.

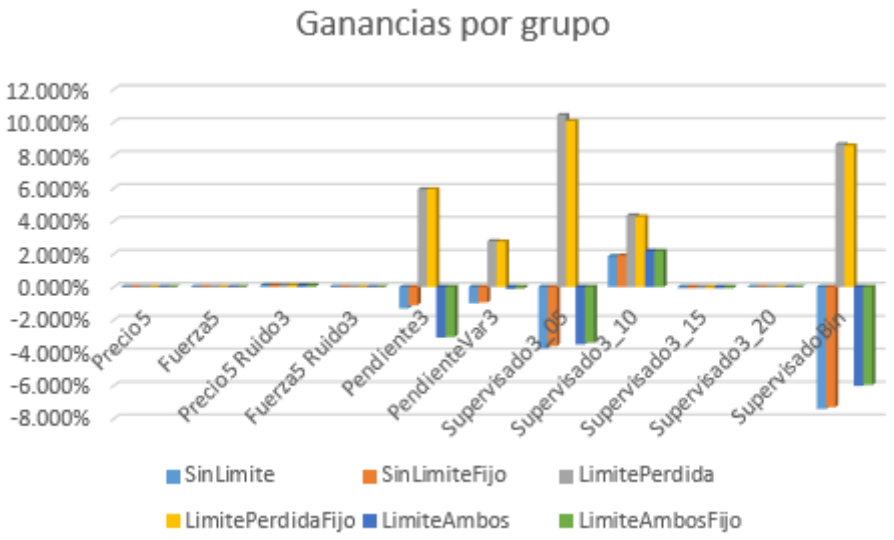

Figura 66: Ganancia promedio de fondos finales aplicando solo reglas de alta confianza y con alto total de apariencia aprendidas de todos los pares de divisas.

En 4 de los grupos no se realiza ninguna operación, y en la mayoría solo se realizan unas pocas, con excepción del grupo binario que siempre realiza predicción, y que, de forma similar a las reglas propias, pierde una cantidad de fondos significativa si no se

limitan las pérdidas.

Las próximas secciones incluyen un análisis más detallado de los resultados más significativos de cada grupo y que no se aprecian a simple vista en los resultados promedio observados en las gráficas.

7.5.1 SupervisadoBin

El grupo más sencillo está compuesto por segmentos separados en ascendente o descendente, sin nada más. En general las reglas creadas tienden a equilibrar las tendencias, es decir, si existen muchos segmentos descendentes en el antecedente, es más común que el consecuente sea ascendente.

La precisión de las reglas es muy elevado, llegando a 89.7% sobre el mismo par de divisas y 88.5% sobre las reglas conjuntas.

A pesar de esta precisión muy elevada, limitando las pérdidas no lo hace tan mal, pero sin limitarlas pierde una cantidad significativa de fondos. Podría simplemente usarse las predicciones inversas en la práctica. Curiosamente esto implicaría que los segmentos predichos no se corresponden con el precio subyacente, o que los segmentos donde fallan las predicciones son mucho más significativos que los que no falla.

Es un grupo interesante si se considera la posibilidad de invertir los consecuentes de las reglas.

7.5.2 Supervisado3_05

En este grupo, incluso una pequeña pendiente es suficiente para declarar un segmento como ascendente o descendente. Como consecuencia muy pocas reglas predicen ruido, llegando a resultados bastante similares a los que presenta la clasificación binaria a la hora de predecir, aunque con una precisión menor en general puesto que incluye segmentos de ruido entre sus predicciones.

La precisión de las reglas es bastante elevado, llegando a 66 % sobre el mismo par de divisas y 67.8% sobre las reglas conjuntas.

Este grupo se comporta de forma bastante parecida al binario, ya que muy pocas predicciones son de ruido. En general la única diferencia es que pierde un poco menos de fondos al realizar menos predicciones.

7.5.3 Supervisado3_10

En este grupo, la pendiente necesaria es suficiente como para crear un equilibrio entre

las reglas creadas.

La precisión de las reglas es decente, y balanceado sobre los grupos, llegando a 49.1 % sobre el mismo par de divisas y 50.3% sobre las reglas conjuntas. El aumento de precisión al solo aceptar reglas de alta confianza es bastante significante, llegando a 7% en algunos casos.

Este grupo es uno de los más interesantes para la gestión de fondos, con una base de reglas equilibradas, y con sus reglas efectivamente prediciendo los segmentos que se corresponden correctamente con el precio.

Usando todas sus propias reglas tiene resultados muy buenos, e incluso usando todas las reglas conjuntas presenta resultados positivos.

7.5.4 Supervisado3_15

En este grupo, la pendiente necesaria es bastante elevada, con el objetivo de reducir predicciones erróneas. En efecto, la precisión de las predicciones aumenta a 75.7% y 75.2%. Este aumento puede ser confuso, ya que la precisión aumenta, pero debido a que la mayoría de las predicciones son de ruido. Esto llega al punto de que solo se realizan 6 predicciones en todo el intervalo de test.

Como resultado sus fondos no tienen muchos cambios, y no es muy representativo ya que en unos casos pierde, y en otros logra obtener ganancias.

7.5.5 Supervisado3_20

En este grupo, la pendiente necesaria es incluso mayor. La precisión de las predicciones aumenta a 89.7% y 88.5%. Ya aquí ocurre que todas las predicciones son de ruido. No hay mucho que se pueda observar en este grupo.

Básicamente no se llega a comprar o vender en muchos de los casos. No es conveniente usar una pendiente demasiado elevada como punto de partida.

7.5.6 Pendiente3

Este grupo es uno de los grupos clasificados automáticamente con K-Means, con k=3, separados únicamente por la pendiente. Sus reglas están bastante balanceadas, con una pequeña mayoría de reglas con consecuente ascendente. La precisión es de 51.2% y 53.9%, aumentando substancialmente a 63.8% y 66.5% en el caso de solo tomar reglas con alta confianza y alto total.

Este grupo es bastante interesante con los resultados de los fondos, ya que de forma similar al binario pierde cantidades significativas con frecuencia, por lo que una inversión de consecuente podría dar buenos resultados.

7.5.7 PendienteVar3

Este grupo es uno de los grupos clasificados automáticamente con K-Means, con *k*=3, separados por pendiente y varianza. Sus reglas están bastante desbalanceadas, con una carencia de reglas de consecuente descendente. La precisión es de 61.1% y 63.7% para todas las reglas.

La separación usando pendiente y varianza obtuvo buenos resultados para reglas propias y no muchas pérdidas para reglas conjuntas. En general sus reglas son bastante buenas, y funcionan de forma consistente en los pares de divisas. A diferencia de otros grupos que poseen excelentes resultados en un par, y malos en otros.

7.5.8 Precio5 y Precio5Ruido3

Estos grupos pertenecen al grupo original clasificado con distancia de Pearson en el capítulo 6 para precio con K-Means, con *k*=5. Sus reglas están bastante balanceadas, con la mayoría ubicadas con consecuente 2, 3 y 4. El consecuente 5, el más negativo, aparece en minoría. La diferencia entre Precio5 y Precio5Ruido3 es solo con respecto a los fondos, la precisión de las predicciones realizadas son las mismas y se corresponden con 22.4% y 22.5%. Precio5 se corresponde a las operaciones realizadas tomando como ruido, y por ende sin operación, las predicciones 2,3 y 4. Mientras que Precio5Ruido3 solo toma como ruido los segmentos predichos como 3, realizando compras con segmentos 1 y 2, y ventas con segmentos 4 y 5.

Sin embargo, la confianza de las reglas en general no es muy elevada y son dispersas, y por ende, no hay suficientes reglas con alta confianza y alto total como para realizar muchas predicciones.

Precio5 en general es bastante estable en los resultados, aunque es recomendable usar todas las reglas, o simplemente la cantidad de reglas es demasiado pequeña.

7.5.9 Fuerza5 y Fuerza5Ruido3

Estos grupos pertenecen al grupo original clasificado con distancia de Pearson en el capítulo 6 para fuerza con K-Means, con *k*=5. Sus reglas están muy bien balanceadas y las predicciones correctas y erróneas están equitativamente distribuidas. Similar al anterior, la diferencia entre Fuerza5 y Fuerza5Ruido3 es solo con respecto a los fondos,

la precisión de las predicciones se corresponden con 21.1% y 21.3%.

Fuerza5 tiene los mejores resultados usando reglas propias, lo cual es muy interesante, aunque está un poco desequilibrado, con el par de divisas AUDCAD obteniendo ganancias mucho mayores a los pares EURNZD y USDJPY, aunque en los 3 casos obtiene ganancias.

Sin embargo, al utilizar las reglas conjuntas sus resultados descienden considerablemente.

7.6 Conclusiones sobre las reglas

La precisión de las reglas llega a ser bastante elevada, pero los fondos no mejoran considerablemente si no se usan límites sobre las pérdidas. Un par de razones para esto es que los clústeres creados no siempre logran representar el precio subyacente de forma perfecta, y que muchas predicciones correctas son para pequeñas cantidades, pero grandes pérdidas eliminan esas ganancias.

Traders profesionales pudieran apoyarse en estas reglas para mejorar sus sistemas de trading, pero de forma independiente carecen de suficiente impacto como para generar ganancias superiores a simplemente comprar un índice.

La segmentación también fue realizada usando un método de Sliding Window, con el objetivo de poder ser aplicado en tiempo real, y se utilizó los valores de apertura en vez del valor de cierre de los intervalos de una hora; esto nos permite realizar un test fiel de forma retroactiva, pero que es un poco más difícil para obtener buenos resultados.

Pese a estos inconvenientes, algunos grupos obtuvieron buenos resultados, y, otros obtuvieron resultados consistentemente negativos debido a que las reglas aprendidas son precisamente las opuestas de lo que ocurre con el precio subyacente.

La utilización de reglas conjuntas mejora la precisión a la hora de predecir segmentos, pero no siempre mejoran los fondos subyacentes, una mejor separación de los clústeres usados para generar las reglas podría mejorar los resultados. La separación en muchos clústeres diferentes disminuye considerablemente el número de reglas, disminuyendo la cantidad de operaciones realizadas, y por ende, el margen de riesgo y ganancia.

A pesar de esta disminución, la calidad de las predicciones de Fuerza5 y Precio5 es bastante elevada usando sus reglas propias, aunque cae al utilizar las reglas conjuntas.

En general usar reglas conjuntas mejora el número de segmentos predichos correctamente, pero disminuye los fondos, ya que los segmentos predichos no se

corresponden directamente con el precio subyacente. Una forma diferente de segmentación, o solo usar subconjuntos de pares de divisas para realizar la base de reglas conjuntas podría mejorar estos resultados.

Conclusiones

Entre las características más distintivas del mercado de divisas se encuentran la independencia del volumen con la dirección del precio a largo plazo. También se observa una distribución cada vez más centrada sobre la media al aumentar la longitud del intervalo. Con esto podemos concluir que el mercado de divisas tiene una tendencia a regresar sobre la media por lo que un desplazamiento demasiado largo de la misma tiende a ser corregido en el transcurso del tiempo.

A partir de la característica anterior, para poder trabajar con las series temporales de dichos mercados es importante considerar el preprocesamiento del ruido utilizando medias móviles y otras técnicas como el Z-Score, además de prestar atención a elementos más sutiles que pudieran existir como el horario de verano o días sin transacciones. La utilización de series representativas de las divisas ayuda a establecer una comparación donde se puede observar el desarrollo de las dos divisas que participan en definir el valor del par de forma independiente.

Esta información fue usada para crear una segmentación basada en Sliding Window utilizando una fórmula basada en la fuerza del par de divisas y que crea los segmentos de forma progresiva y puede ser aplicada en tiempo real. Dicha segmentación propuesta permite la clasificación en tiempo real reduciendo el número de datos a procesar y además reduciendo el ruido de forma significativa mientras conserva la estructura principal de la serie. Es posible ajustar los parámetros de dicha segmentación según la serie temporal y los intereses de la investigación.

Para la creación de las series temporales simbólicas se clasificaron los segmentos usando el algoritmo de agrupamiento K-means. Se determinó que utilizar la distancia de Pearson y 5 clústeres usando la estructura del precio y la estructura de la fuerza ofrecía buenos resultados para dicha clasificación.

La minería de reglas secuenciales recurrentes es un campo que ha sido poco explorado, con mucho énfasis en reglas secuenciales o en reglas recurrentes, pero con pocas investigaciones sobre la mezcla de ambas. Se mostró que la combinación de ambas aumenta la calidad de las predicciones y la confianza de las reglas, pero que ese aumento de calidad disminuye en la clasificación correcta del precio subyacente en los segmentos al ser evaluados. Es recomendable utilizar las reglas del mismo par de

divisas en vez de utilizar reglas extendidas de todos los pares de divisas para maximizar las ganancias. También se concluye que es posible obtener reglas de alta confianza, pero las ganancias generadas son muy pequeñas si no existe un control sobre las pérdidas permitidas en un segmento.

Posibles mejoras podrían obtenerse ajustando el proceso de segmentación con otra fórmula, o cambiando la forma de clasificar los segmentos usando otras distancias. Sería de interés también observar el comportamiento de las reglas aplicadas en un entorno profesional de trading con control de posición y límite de pérdidas, además de incluir los costos de comisiones y otros gastos operativos.

Bibliografía

[1] O. Hegazy, O. S. Soliman y M. Abdul Salam, «A Machine Learning Model for Stock Market Prediction,» *International Journal of Computer Science and Telecommunications,* vol. 4, pp. 17-23, Diciembre 2013.

[2] J. Patel, S. Shah, P. Thakkar y K. Kotecha, «Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques,» *Expert systems with applications,* vol. 42, nº 1, pp. 259-268, 2015.

[3] E. Ruiz Dotras y L. Sust, «Mercados Financieros,» Universitat Oberta de Catalunya, Febrero 2013. [En línea]. Available: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/68509/5/Mercados%20financieros_M%C3%B3dulo%201_Introducci%C3%B3n%20al%20sistema%20financiero.pdf. [Último acceso: 22 Febrero 2020].

[4] Universidad Centroamericana "José Simeón Cañas", «Mercado de divisas,» 5 Julio 2013. [En línea]. Available: http://www.uca.edu.sv/facultad/clases/maestrias/made/m230054/06Cap-2-1-MERCADO-DIVISAS.pdf. [Último acceso: 22 Febrero 2020].

[5] Departamento de Estadística e Investigación Operativa. Universidad de Granada., Universidad de Granada, 16 Marzo 2010. [En línea]. Available: https://www.ugr.es/~fabad/desestacionalizacion.pdf. [Último acceso: 22 Febrero 2020].

[6] CornèrTrader, «Chapter 2.4 Multiple Time Frames,» Cornèr Bank, 28 Febrero 2019. [En línea]. Available: https://www.cornertrader.com/export/sites/cornertraderCOM/.content/.galleries/downloads/website/tutorials/2-4-multiple-time-frames.pdf. [Último acceso: 22 Febrero 2020].

[7] Rolf, «How to perform a multiple time frame analysis,» Tradeciety Academy, 6 Mayo 2019. [En línea]. Available: https://www.tradeciety.com/how-to-perform-a-multiple-time-frame-analysis/. [Último acceso: 22 Febrero 2020].

[8] J. Fundora, «Multiple Time Frames Can Multiply Returns,» Investopedia, 4 Diciembre 2019. [En línea]. Available: https://www.investopedia.com/articles/trading/07/timeframes.asp. [Último acceso: 22 Febrero 2020].

[9] N. J. Nilsson, «Introduction To Machine Learning,» 3 Noviembre 1998. [En línea]. Available: https://ai.stanford.edu/~nilsson/MLBOOK.pdf. [Último acceso: 22 Febrero 2020].

[10] J. C. Riquelme, R. Ruiz y K. Gilbert, «Minería de Datos: Conceptos y Tendencias,» *Revista Iberoamericana de Inteligencia Artificial,* vol. 10, nº 29, pp. 11-18, 2006.

[11] M. González Castellanos, *Clasificación semi-supervisada de series temporales,* Granada, Granada: Universidad de Granada, 2016, pp. 25-27.

[12] StatisticsHowTo, «Z-Score: Definition, Formula and Calculation,» Statisticshowto, 2020. [En línea]. Available: https://www.statisticshowto.com/probability-and-statistics/z-score/. [Último acceso: 22 Febrero 2020].

[13] S. McLeod, «Z-Score: Definition, Calculation and Interpretation,» 17 Mayo 2019. [En línea]. Available: https://www.simplypsychology.org/z-score.html. [Último acceso: 22 Febrero 2020].

[14] G. Press, «Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says,» 23 Marzo 2016. [En línea]. Available: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/. [Último acceso: 22 Febrero 2020].

[15] Dukascopy, «Historical Data Feed,» 1998. [En línea]. Available: https://www.dukascopy.com/swiss/english/marketwatch/historical/. [Último acceso: 22 Febrero 2020].

[16] R. Canessa C., «Estrategias de Trading con el Indicador de Volumen en el Forex,» Mayo 2019. [En línea]. Available: https://www.tecnicasdetrading.com/2019/05/estrategias-trading-indicador-de-volumen-forex.html. [Último acceso: 22 Febrero 2020].

[17] S. García, J. Luengo y F. Herrera, Data Preprocessing in Data Mining, Springer, 2015.

[18] J. E. Díaz Pinzón, «Modelos media móvil simple y media móvil exponencial como pronóstico de la acción de ISA,» vol. 18, nº 1, pp. 44-52, 17 Julio 2018.

[19] V. Ruiz Herrán, M. A. Pérez Martínez y A. Olasolo Sogorb, «Análisis de la eficacia de las medias móviles en el mercado intradiario de renta variable español,» *Universidad, Sociedad y Mercados Globales,* pp. 56-68, 1 Enero 2008.

[20] J. Yin, Y.-W. Si y Z. Gong, «Financial Time Series Segmentation Based On Turning Points,» de *International Conference on System Science and Engineering*, Macau, China, 2011.

[21] J. Jiang, Z. Zhang y H. Wang, «A New Segmentation Algorithm to Stock Time Series Based on PIP Approach,» *2007 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2007,* pp. 5609 - 5612, 25 10 2007.

[22] C. Chen, V. Tseng y H. Yu, «Time series pattern discovery by a PIP-based evolutionary approach,» *Soft Comput 17,* p. 1699–1710, 26 Enero 2013.

[23] «Modelo para el descubrimiento de patrones en series temporales simbólicas,» Universidad Politécnica de Madrid, Madrid, 2017.

[24] R. Xu y D. C. Wunsch, Clustering, New Jersey: John Wiley & Sons, 2009.

[25] J. D. Rodríguez Morales, «Métodos de clasificación semi-supervisada para series temporales,» Universidad Central “Marta Abreu” de Las Villas, Santa Clara, 2015.

[26] J. Francisco López, «Economipedia,» 18 Noviembre 2017. [En línea]. Available: https://economipedia.com/definiciones/varianza.html. [Último acceso: 8 Junio 2020].

[27] F. J. Marco Sanjuán, «Economipedia,» 2 Octubre 2017. [En línea]. Available: https://economipedia.com/definiciones/curtosis.html. [Último acceso: 8 Junio 2020].

[28] K. Academy, «Khan Academy,» 23 Mayo 2017. [En línea]. Available: https://es.khanacademy.org/math/cc-eighth-grade-math/cc-8th-linear-equations-functions/8th-slope/a/slope-formula. [Último acceso: 8 Junio 2020].

[29] C. M. CUADRAS, «Distancias Estadísticas,» *ESTADISTICA ESPAÑOLA,* vol. 30, nº 119, pp. 295-378, 1989.

[30] S. VAN DONGEN y A. ENRIGHT J., «METRIC DISTANCES DERIVED FROM COSINE SIMILARITY AND PEARSON AND SPEARMAN CORRELATIONS,» 16 Agosto 2012. [En línea]. Available: https://arxiv.org/pdf/1208.3145.pdf. [Último acceso: Mayo 8 2020].

[31] S. Harms y J. Deogun, «Sequential Association Rule Mining with Time Lags,» *Journal of Intelligent Information Systems,* nº 22, pp. 7-22, Enero 2004.

[32] P. Fournier Viger, U. Faghihi, R. Nkambou y E. Mephu Nguifo, «CMRules: Mining sequential rules common to several sequences,» *Knowl.-Based Syst.,* vol. 25, pp. 63-76, Febrero 2012.

[33] P. Fournier Viger, C.-W. Wu, V. Tseng, L. Cao y R. Nkambou, «Mining Partially-Ordered Sequential Rules Common to Multiple Sequences,» *IEEE Transactions on Knowledge and Data Engineering,* vol. 27, pp. 1-1, 1 Agosto 2015.

[34] D. Lo, S.-C. Khoo y C. Liu, «Efficient Mining of Recurrent Rules from a Sequence Database,» Department of Computer Science, National University of Singapore, 2010.

[35] I. Kamehkhosh, D. Jannach y M. Ludewig, «A Comparison of Frequent Pattern Techniques and a Deep Learning Method for Session-Based Recommendation,» *Workshop on Temporal Reasoning in Recommender Systems, collocated with ACM RecSys'17, ,* 2017.

[36] S. Asiri, «Machine Learning Classifiers,» 11 Junio 2018. [En línea]. Available: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623. [Último acceso: 22 Febrero 2020].

Anexo 1: Tablas de resultados de las pruebas realizadas sobre las reglas secuenciales.

	Reglas del propio par de divisas		
	Todas	AltaConfianza	AltoTotal
Precio5	0.223925086	0.196190997	0.366666667
Fuerza5	0.211102236	0.233363581	0.233363581
Precio5 Ruido3	0.223925086	0.196190997	0.366666667
Fuerza5 Ruido3	0.211102236	0.233363581	0.233363581
Pendiente3	0.512475943	0.625381005	0.638670431
PendienteVar3	0.611141125	0.670334748	0.681699939
Supervisado3_05	0.660115761	0.69571854	0.704160705
Supervisado3_10	0.491675892	0.561742637	0.573755139
Supervisado3_15	0.757361276	0.76833854	0.767673509
Supervisado3_20	0.89718906	0.89718906	0.89718906
SupervisadoBin	0.801400222	0.805621107	0.808257157

Tabla 1: Precisión promedio de las reglas aprendidas del par de divisas aplicadas sobre el mismo par de divisas.

	Reglas conjuntas de todos los pares		
	Todas	AltaConfianza	AltoTotal
Precio5	0.224597198	0.190903288	NA
Fuerza5	0.21292166	0.221102786	0.221102786
Precio5 Ruido3	0.224597198	0.190903288	NA
Fuerza5 Ruido3	0.21292166	0.221102786	0.221102786
Pendiente3	0.53941443	0.653005123	0.662588524
PendienteVar3	0.637685977	0.697143125	0.705866117
Supervisado3_05	0.678022667	0.709225204	0.717700331
Supervisado3_10	0.503185745	0.566749586	0.577443902
Supervisado3_15	0.752636378	0.76413409	0.763965745
Supervisado3_20	0.885728755	0.885728755	0.885728755
SupervisadoBin	0.804660626	0.808177784	0.810286769

Tabla 2: Precisión promedio de las reglas aprendidas de todos los pares de divisas.

	Diferencias		
	Todas	AltaConfianza	AltoTotal
Precio5	0.000672112	-0.005287709	NA
Fuerza5	0.001819424	-0.012260795	-0.012260795
Precio5 Ruido3	0.000672112	-0.005287709	NA
Fuerza5 Ruido3	0.001819424	-0.012260795	-0.012260795
Pendiente3	0.026938487	0.027624118	0.023918093
PendienteVar3	0.026544852	0.026808377	0.024166178
Supervisado3_05	0.017906906	0.013506664	0.013539626
Supervisado3_10	0.011509853	0.005006949	0.003688763
Supervisado3_15	-0.004724898	-0.00420445	-0.003707764
Supervisado3_20	-0.011460305	-0.011460305	-0.011460305
SupervisadoBin	0.003260404	0.002556677	0.002029612
Total	0.074958371	0.024741022	0.027652613

Tabla 3: Diferencia en la precisión entre las reglas propias y las conjuntas.

	SinLimite	SinLimiteFijo	LimitePerdida	LimitePerdidaFijo	LimiteAmbos	LimiteAmbosFijo
Precio5	10310.735	10320.90309	10787.42654	10770.4407	10064.08434	10071.49143
Fuerza5	10863.6068	10838.16285	11411.03201	11332.77684	10549.16679	10539.76561
Precio5 Ruido3	9973.9157	9997.539936	11478.01865	11401.41862	9840.294197	9848.434763
Fuerza5 Ruido3	10171.1926	10119.74657	11581.72685	11467.17978	10157.82624	10140.14803
Pendiente3	10062.7702	10080.94843	11240.21358	11189.30652	9697.769933	9702.728457
PendienteVar3	10409.4221	10425.86123	11295.39837	11236.55507	10224.4744	10234.7304
Supervisado3_05	9632.65402	9668.876707	11396.01316	11338.19257	9820.702014	9842.994186
Supervisado3_10	10713.067	10715.1401	11697.42509	11589.59057	10503.54016	10505.39805
Supervisado3_15	9885.28089	9886.997284	9972.906593	9974.964476	9859.403519	9860.205372
Supervisado3_20	9955.4378	9955.643868	9994.082708	9994.221902	9994.082708	9994.221902
SupervisadoBin	9483.08528	9514.973735	11442.25507	11381.02476	9887.142187	9908.297289

Tabla 4: Ganancia promedio de fondos finales aplicando todas las reglas propias sobre sus propios pares de divisas

	SinLimite	SinLimiteFijo	LimitePerdida	LimitePerdidaFijo	LimiteAmbos	LimiteAmbosFijo
Precio5	10205.4075	10211.15628	10474.72733	10470.23595	10091.11133	10094.58161
Fuerza5	10764.5672	10747.43423	10992.27067	10955.84703	10426.51245	10424.25627
Precio5 Ruido3	10177.7269	10177.62041	11102.21768	11048.96701	10123.56871	10123.57567
Fuerza5 Ruido3	10309.3878	10315.32048	11080.03395	11039.8844	10211.30892	10219.29482
Pendiente3	10118.6807	10149.06985	11201.10009	11160.67974	9770.964555	9784.832803
PendienteVar3	10518.5394	10514.89059	11128.30998	11077.70585	10311.92043	10315.9501
Supervisado3_05	9515.9058	9527.463235	11234.73172	11164.09186	9767.876385	9776.186558
Supervisado3_10	10121.1469	10132.52581	10988.95259	10952.90748	10156.85543	10163.05233
Supervisado3_15	9895.10885	9896.217707	9929.997354	9931.312794	9860.831497	9860.992829
Supervisado3_20	10028.8937	10028.79413	10028.89368	10028.79413	10028.89368	10028.79413
SupervisadoBin	9176.81478	9183.030284	11096.51949	11068.13785	9597.431428	9614.466922

Tabla 5: Ganancia promedio de fondos finales aplicando las reglas de alta confianza propias sobre sus propios pares de divisas

	SinLimite	SinLimiteFijo	LimitePerdida	LimitePerdidaFijo	LimiteAmbos	LimiteAmbosFijo
Precio5	10065.9392	10065.83259	10065.93917	10065.83259	10051.60656	10051.54798
Fuerza5	10151.4019	10151.20688	10151.40193	10151.20688	10110.84064	10110.92979
Precio5 Ruido3	9966.94731	9968.248102	10055.071	10055.8429	10003.10717	10003.73937
Fuerza5 Ruido3	10124.8978	10125.42566	10150.99541	10151.03986	10106.6157	10106.96665
Pendiente3	9673.15798	9690.975309	10613.39762	10615.59697	9490.601118	9490.571764
PendienteVar3	10217.9746	10224.69344	10528.90042	10522.11499	10073.11287	10078.18154
Supervisado3_05	9318.22401	9314.655538	10979.60273	10933.02418	9555.104196	9556.445095
Supervisado3_10	9918.83256	9930.371111	10592.53492	10584.23541	10235.63438	10238.91645
Supervisado3_15	10045.6208	10045.50086	10075.54917	10075.44372	10050.29532	10050.40705
Supervisado3_20	10000	10000	10000	10000	10000	10000
SupervisadoBin	9067.1635	9063.050263	10863.88395	10854.04469	9307.19101	9306.260615

Tabla 6: Ganancia promedio de fondos finales aplicando las reglas de alta confianza y alta apariencia propias sobre sus propios pares de divisas

	SinLimite	SinLimiteFijo	LimitePerdida	LimitePerdidaFijo	LimiteAmbos	LimiteAmbosFijo
Precio5	10179.7223	10187.23686	10536.06178	10530.64609	10075.66722	10080.70056
Fuerza5	9784.02205	9801.280214	10532.4303	10529.68429	9775.406613	9780.961127
Precio5 Ruido3	10479.4381	10449.59502	11684.03155	11559.10325	10310.21493	10302.75412
Fuerza5 Ruido3	9800.53274	9811.998119	11010.60401	10971.19027	9713.667098	9723.376101
Pendiente3	9249.02204	9250.534398	10476.21588	10482.75928	9262.40662	9252.556991
PendienteVar3	9968.25361	9988.942468	10670.0285	10665.51299	9814.920324	9822.403607
Supervisado3_05	9821.41912	9822.115158	11421.88904	11322.41242	9768.027261	9771.361162
Supervisado3_10	10323.8255	10298.12066	11034.85487	10967.35763	10338.61501	10328.96496
Supervisado3_15	10043.5034	10042.6239	10082.81863	10081.82459	10059.46509	10058.94833
Supervisado3_20	10000	10000	10000	10000	10000	10000
SupervisadoBin	9413.68177	9434.240295	11041.22427	11015.90194	9544.082226	9556.631404

Tabla 7: Ganancia promedio de fondos finales aplicando todas las reglas conjuntas.

	SinLimite	SinLimiteFijo	LimitePerdida	LimitePerdidaFijo	LimiteAmbos	LimiteAmbosFijo
Precio5	10046.6566	10047.26879	10079.66482	10079.78927	10023.82618	10024.21455
Fuerza5	9680.77044	9675.026067	9778.94524	9774.829305	9750.287258	9746.152897
Precio5 Ruido3	10126.0964	10126.56111	10182.14202	10181.7282	10068.1748	10069.25481
Fuerza5 Ruido3	9869.34499	9865.359888	10004.10825	10001.17437	9934.121629	9932.271864
Pendiente3	9854.55312	9874.749398	10591.9735	10592.81524	9700.465826	9704.379348
PendienteVar3	9825.01016	9829.598792	10208.6727	10209.00745	9919.434555	9919.962951
Supervisado3_05	9595.19855	9601.472722	11045.20285	11002.31919	9649.416629	9656.183323
Supervisado3_10	10186.7784	10190.7941	10437.41699	10430.88678	10208.98916	10210.20797
Supervisado3_15	10008.1647	10008.08783	10008.16473	10008.08783	10008.16473	10008.08783
Supervisado3_20	10000	10000	10000	10000	10000	10000
SupervisadoBin	9358.35638	9377.415473	10982.4484	10964.84589	9500.076548	9511.34413

Tabla 8: Ganancia promedio de fondos finales aplicando las reglas de alta confianza conjuntas.

	SinLimite	SinLimiteFijo	LimitePerdida	LimitePerdidaFijo	LimiteAmbos	LimiteAmbosFijo
Precio5	10000	10000	10000	10000	10000	10000
Fuerza5	10000	10000	10000	10000	10000	10000
Precio5 Ruido3	10009.5827	10009.57863	10009.58266	10009.57863	10009.2978	10009.29391
Fuerza5 Ruido3	10000	10000	10000	10000	10000	10000
Pendiente3	9868.68763	9889.709954	10592.17947	10592.97767	9692.398514	9695.684929
PendienteVar3	9900.98333	9905.321328	10277.49933	10275.10563	9988.795256	9988.981988
Supervisado3_05	9625.402	9641.848245	11043.29975	11006.92912	9651.597077	9661.48533
Supervisado3_10	10185.91	10188.76859	10434.16559	10426.24814	10217.79948	10218.00766
Supervisado3_15	9992.44662	9992.446623	9992.446623	9992.446623	9992.446623	9992.446623
Supervisado3_20	10000	10000	10000	10000	10000	10000
SupervisadoBin	9258.7347	9270.560948	10865.27314	10858.10662	9399.214124	9404.720118

Tabla 9: Ganancia promedio de fondos finales aplicando las reglas de alta confianza y alta apariencia conjuntas.